HUMAN DISEASE PREDICTION SYSTEM USING DATA MINING TECHNIQUES

Dissertation submitted in fulfilment of the requirements for the Degree of

MASTER OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

By

MANISHA SINGH

11502545

Supervisor

JASPREET KAUR SAHIWAL

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

May 2017

LOVELY PROFESSIONAL UNIVERSITY

Transforming Education, Transforming India

TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering

Program : P172::M.Tech. (Computer Science and Engineering) [Full Time]

COURSE CODE : CSE546 **REGULAR/BACKLOG :** Regular **GROUP NUMBER :** CSERGD0292

Supervisor Name : Jaspreet Kaur Sahiwal **UID :** 14752 **Designation :** Assistant Professor

Qualification : _____ **Research Experience :** _____

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Manisha Singh	11502545	2015	K1519	9915121514

SPECIALIZATION AREA : Programming-I **Supervisor Signature:** _____

PROPOSED TOPIC : Human Disease Prediction System using Data Mining Techniques

Qualitative Assessment of Proposed Topic by PAC		
Sr.No.	Parameter	Rating (out of 10)
1	Project Novelty: Potential of the project to create new knowledge	7.50
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	7.00
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	7.25
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	8.50
5	Social Applicability: Project work intends to solve a practical problem.	7.50
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	8.50

PAC Committee Members		
PAC Member 1 Name: Janpreet Singh	UID: 11266	Recommended (Y/N): Yes
PAC Member 2 Name: Harjeet Kaur	UID: 12427	Recommended (Y/N): Yes
PAC Member 3 Name: Sawal Tandon	UID: 14770	Recommended (Y/N): Yes
PAC Member 4 Name: Raj Karan Singh	UID: 14307	Recommended (Y/N): NA
DAA Nominee Name: Kanwar Preet Singh	UID: 15367	Recommended (Y/N): Yes

Final Topic Approved by PAC: Human Disease Prediction System using Data Mining Techniques

Overall Remarks: Approved

PAC CHAIRPERSON Name: 11011::Dr. Rajeev Sobti **Approval Date:** 25 Oct 2016

4/26/2017 11:13:05 AM

ABSTRACT

Human Diseases are increasing rapidly in today's generation mainly due to the life style of people like poor diet, lack of exercises, drugs and alcohol consumption etc. But the most spreading disease that is commonly occurring in people and causing 80% of death in country is heart disease. It is expected that by 2030, around 25 million people may die because of heart diseases. Though many researchers have suggested and proposed methods for diagnosing the heart diseases from the enormous amount of heart disease data, there is no proper effective techniques and are not properly mined. In field of Medicine, a large amount of information is generated each and every day which is stored in medical database. This database contains raw dataset which consist of inconsistent and redundant data. The health care system is no doubt very rich in aspect of storing data but at the same time very poor in fetching knowledge. Data mining methods can help in extracting a valuable knowledge by applying data mining techniques like classification, regression, clustering etc. After the collection of data when the dataset becomes more large and complex then data mining algorithms (here considering Decision tree, Naive Bayes, Neural Network, K-Nearest Neighbor) are used. To get accuracy and efficiency in result a new approach called improved k-mean algorithm is proposed in this paper. The dataset used for prediction is obtained and utilized from UCI machine learning repositories. The research work is based on prediction analysis for heart diseases detection.

Keywords- Data mining, Classification, Clustering, Regression, Heart Disease, Naïve Bayes, Neural Network, k- Nearest Neighbor, K-mean, Decision Tree Algorithm.

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation entitled "HUMAN DISEASE PREDICTION SYSTEM USING DATA MINING TECHNIQUES" in partial fulfillment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Mrs. Jaspreet Kaur Sahiwal. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Signature of Candidate

Manisha Singh

11502545

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.TECH Dissertation entitled "**HUMAN DISEASE PREDICTION SYSYTEM USING DATA MINING TECHNIQUES"**, submitted by **Manisha Singh** at **Lovely Professional University, Phagwara, India** is a bonafide record of her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

(Jaspreet Kaur Sahiwal)

Date: 26/04/2017

Counter Signed by:

1) **Concerned HOD:**
   HoD's Signature: _____

   HoD Name: _____

   Date: _____

2) **Neutral Examiners:**

   External Examiner

   Signature: _____

   Name: _____

   Affiliation: _____

   Date: _____

   Internal Examiner

   Signature: _____

   Name: _____

   Date: _____

ACKNOWLEDGEMENT

I am using this opportunity to express my gratitude to everyone who supported me throughout the course of this dissertation. I am thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the thesis work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to this research.

I express my warm thanks to Assistant professor Jaspreet Kaur Sahiwal for her support and guidance.

I would also like to thank all the people who provided me with the facilities being required and conductive conductions.

Thank you,

Manisha Singh

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Data Mining

Nowadays everyone is facing problem such as the government, corporates and industrial communities due to constantly increasing number of databases. These databases need to be managed as well as explored. The distributed heterogeneous multimedia database requires secure access with rich metadata and to meet the time constraints. The exploratory tools which supports the identification of domain and also the mission critical elements, for example, patterns in data access (e.g., security breach determinations), patterns in data (e.g., marketing and clustering), or for patterns in transactions (e.g., data compression), to site a couple, is the second thing which it requires. Discovery of knowledge is moderately a new work in database which employs many tools to explore and identify structure and patterns in [1]. Regularly these huge databases the data is preprocessed to facilitate such computations (data warehousing). The data is then mined for particular rules that are constructed incrementally and frequently steered by clients on account of a particular set of goals. Data mining is fairly reminiscent of approaches utilized as a part of statistical analysis throughout the most recent 20 years. Projection interest is a technique that endeavors to identify clusters in n-dimensional data spaces. Projections in increasing dimensional spaces are utilized to develop to locally optimal clustering of the data. Recently, these techniques are client steered. What has been lacking in the statistical field, and clearly is so in the data mining field, also is the involvement of the client in courses other than simply as a primitive input device. Human creatures have rich capabilities in interpreting complex situations still unequaled to any machine. As a part of computation, visualization technologies have been utilized to direct the analysis, also used in the query interfaces to complex the multimedia data in databases, and are also utilized in navigation as well as presentation of information [2]

1.1.1 Knowledge Data Discovery (KDD)

The main reason to increase the capabilities of human analysis is to handle extensive number of bytes which are collected from both scientific and economic. All business utilizes the data mainly to obtain the advantage of competitions and increment the efforts to give useful services which can be helpful for the customers. Data which are captured from the surrounding are said to be the fundamental proof one uses for constructing various models as well as theories. Since PCs have allowed humans to easily access large amount of information that one can accumulate, it is normal to convert into computational techniques just to derive a valuable structures and patterns from large volume of data. Subsequently, KDD is mainly used to express an issue that is made by digital information which is an unacceptable fact for every individual: data overload [3].

KDD alludes the entire process of extracting valuable information from dataset and the data mining alludes only a specific step in KDD process. For extracting the useful patterns from the data, data mining is considered as the most commonly used application of particular algorithm. The difference between the data mining steps and KDD process is a main point which is being discussed in this paper. The steps that are involved additionally in the KDD process are data cleaning, data selection, data readiness, incorporating prior knowledge which is appropriate and legitimate mining results interpretations, which are necessary to make sure that helpful information or knowledge from huge amount of data is obtained. We must be careful while considering the applications of data-mining methods because applications taken blindly may be a very risky, effectively gives invalid patterns and meaningless information. KDD has developed and it is keep on developing day by day by combining all the research fields all together, such as the design recognition, databases, machine learning, statistics, knowledge procurement for master systems, AI, data visualization, and elite registering. The main objective is to extract the abnormal state of information from the low-level data from extensive data sets. The component of data mining in KDD depends totally on known techniques which include design recognition, machine learning, and statistics to discover useful patterns from the datasets in the process of KDD. KDD are not quite the same as example recognition or machine learning (and related fields). These fields give a portion of the data-mining methods as a part of data mining steps that are

utilized in KDD process. KDD mainly concentrates on the entire process of extracting useful or valuable knowledge from the data in KDD which includes how the data are accessed as well as stored, how in the massive datasets the algorithms are applied and then also it works proficiently, how we can interpret and envision the results, and how the interaction between the man-machine may be supported and modeled. The KDD process contains multiple activities that use techniques for a specific discipline such as machine learning. In this paper, to contribute to KDD many opportunities have been found in the fields of different AI (except for machine learning). KDD gives an uncommon accentuation which finds the patterns in understandable form that are represented as helpful and intriguing information. Data warehousing is a related field that comes out from databases, for decision support and online analysis the well-known trends of businesses like collecting and cleaning transactional data is made accessible. The platform of KDD is set by the data warehousing in two ways such as data cleaning and the data access [4].

(1) Data Cleaning : It is mainly used when the database contains all irrelevant, noise and errors, consistent data as well as missing data. To make this irrelevant data as relevant data, data cleaning is performed to obtain useful information.

(2) Data Access : It is capable enough to access the data from the database, and well known methods should be made just for retrieving the data from the database and provides a path to access the data that are very hard to obtain (for example offline storing data).

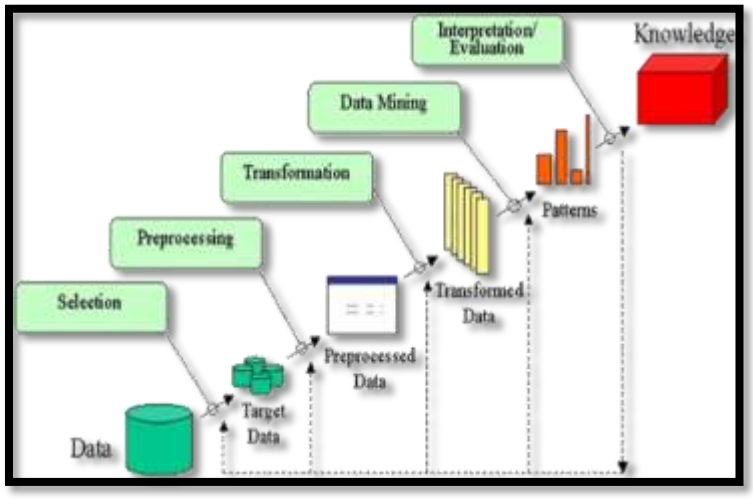

Figure 1: Knowledge Data Discovery Process

The KDD process includes various steps which includes the decision of many clients they are as follows:

Step1: In this step, it basically improves the understanding level of the applications for each and every individual, and also it determines the main goals of KDD process from the views of the customer.

Step 2: In this step, the discovery is mainly performed on the targeted data sets in which we first select the required data set and then we concentrate more on the data samples.

Step 3: In this step, it basically discuss about two things that is data cleaning and data preprocessing. It eliminates all the noise and missing data that may be found from the database through data cleaning. After that it performs data preprocessing in which it converts the raw data into easy an understandable form [5].

Step 4: In this step, data reduction and projection are mainly discussed. Depending on the main objectives of the task, it describes useful characteristics to represent data. It generally uses reduction method to reduce the attributes by decreasing the data in the database.

Step 5: This step deals with selecting the appropriate tasks of data mining. It is quite similar with the goals of (step 1), it recognizes the methods like clustering, classification and regression which helps in deciding whether these methods are the main goals of KDD process or not.

Step 6: This step is responsible for choosing the algorithms of data mining. For searching a useful pattern in a data set, few methods would be selected. It also tells that which parameters and models usage would be more appropriate. It also compares the entire norms of KDD process with the specific data mining methods [6].

Step 7: In this step, data mining plays a major role in searching of patterns of enthusiasm for a specific presentation frame including classification rules or trees, clustering and regression etc.

Step 8: This step is mainly used for clarifying the patterns which are already mined. This step includes the models as well as the extracted patterns of visualization.

Step 9: In this step, it mainly deals with enhancing the discovered knowledge. In this process, with the help of already accepted information it additionally incorporates inspecting and settling potential conflicts.

1.1.3 Data Mining Methods

The data mining task is generally divided into two types one is predictive and the other is descriptive. In predictive task it utilizes the variables which can be used to predict the unknown values of the other variables of interest. The predictive task includes the methods within it such as classification, regression, time series analysis and prediction. The description focuses more on describing the data patterns which can be easily understood by human. The descriptive task includes clustering, summarization, association rules and sequence discovery. The boundaries in between the predictive and descriptive models are found not so strong. For a particular data mining applications, the importance of the prediction and description may differ some times. The goals of description and prediction are accomplished by using variety of mining methods of data [7].

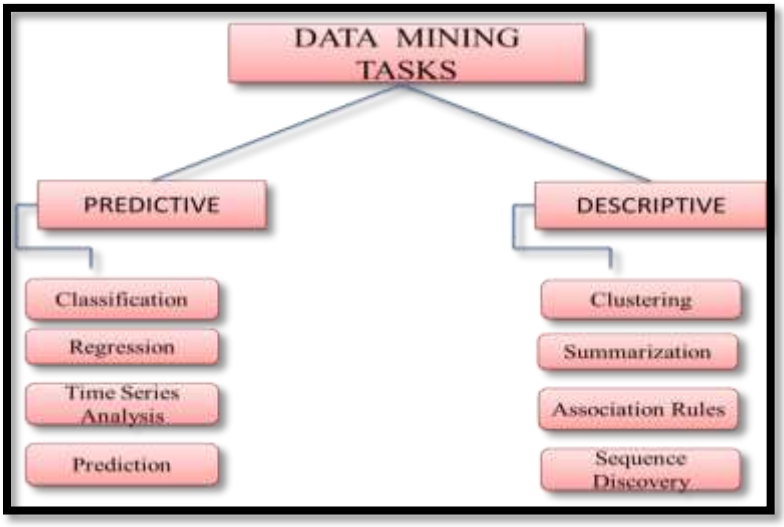

Figure 2: Data Mining Tasks

a. Classification: Classification is one kind of function in the field of data mining. In classification we generally predict the result of anything in the form of discrete outputs. In a simple way we can just say that it is nothing but mapping the input variables in the discrete form. The target class in the data for each case is easily predicted with full accuracy, which is referred as one of the important goal of classification. For example, in giving loans to the applicants the bank keeps the records of high, medium and low risk of credit by classifying all the aspect loan is given to the deserving applicant. So in this way the classification is used everywhere to get a successful result.

b. Regression: Regression is the function of data mining which predicts the result based on continuous outputs. It is also used for the numerical predictions such as age, weight, distance and time etc. It mainly starts with the known target values of data set. For example, the heights of the children can be predicted numerically by observing data over the period of time for all the children. Regression always comes under the supervised learning.

c. Clustering: Clustering can be defined as grouping of similar data into a single group is said to be called as a clusters. Each clusters has a different groups based on the similarities of the objects. Basically clustering is done to reduce data size by reducing its complexity based on some attributes. For example, we are having thousands of essay writing with us and if we need to assemble them in some order then surely it would become a very complicated task for us. By the uses of clustering this problem can be solved easily by using some attributes like word frequency, page count and length of sentence, based on these attributes the essays can be made into different cluster forms and hence the problem is resolved quickly.

d. Summarization: Summarization is one of the concepts found in data mining which is mainly used to represent the mean and standard deviation in the tabular form for the analysis of data. One of the data mining techniques named as clustering always summarizes the large sets of data. Summarization comes under the descriptive models in the methods of data mining.

e. Dependency modeling: Dependency modeling includes the model that determines remarkable dependencies between the variables. A dependency model occurs in two levels:

(1) Level 1: The model of structural level represents the certain variables which are dependent on each other.

(2) Level 2: The model of quantitative level which determines the dependencies of the strengths by utilizing some numeric measures.

f. Change and deviation detection: The most notable changes in the data that would be measured beforehand and it will be discovered by focusing more on it with the help of deviation detection and change [9].

1.2 Cluster Analysis

On the basis of the properties of a specific set of objects, the division of these objects on the basis of their similarities is known as clustering process. A specific joint algorithm that can be applicable to almost all required information analysis is done with the help of the technique which partitions the data. An object is made to be a part of the cluster or is assigned to it with the help of the clustering analysis method. The type of grouping involved here is the hard partitioning. Within the determination degree each object is related to a cluster. This is known as the soft partitioning method. There are varieties of divisions proposed within the clusters that are based on the distinct objects present. The partitioning method also depends on the various models present. On the basis of the relationship amongst the objects and the organization present there are various models proposed [11].The cluster analysis is mainly used to identify group of objects (like customer) which are very much similar to each other and based on some attributes like price consciousness and brand loyalty the similar object will be assign into a cluster. It is necessary to decide the clustering procedure to form group of objects only after deciding the clustering variables like price consciousness and brand loyalty. As different procedures require different decisions prior to analysis, therefore it is considered as a crucial step for the analysis. In market research, we discuss the popular approaches which can be easily computed using SPSS. These approaches includes the following methods like hierarchical methods, partitioning methods (more precisely, k-means) and two-step clustering which is generally formed by combination of first two methods. To group the most similar objects into a cluster and to determine each object's cluster membership, each of these procedures follows different approaches for

grouping objects. In other words we can say that, the object in a certain cluster should be similar to the objects of the same cluster and it should be different from objects of another cluster [12].

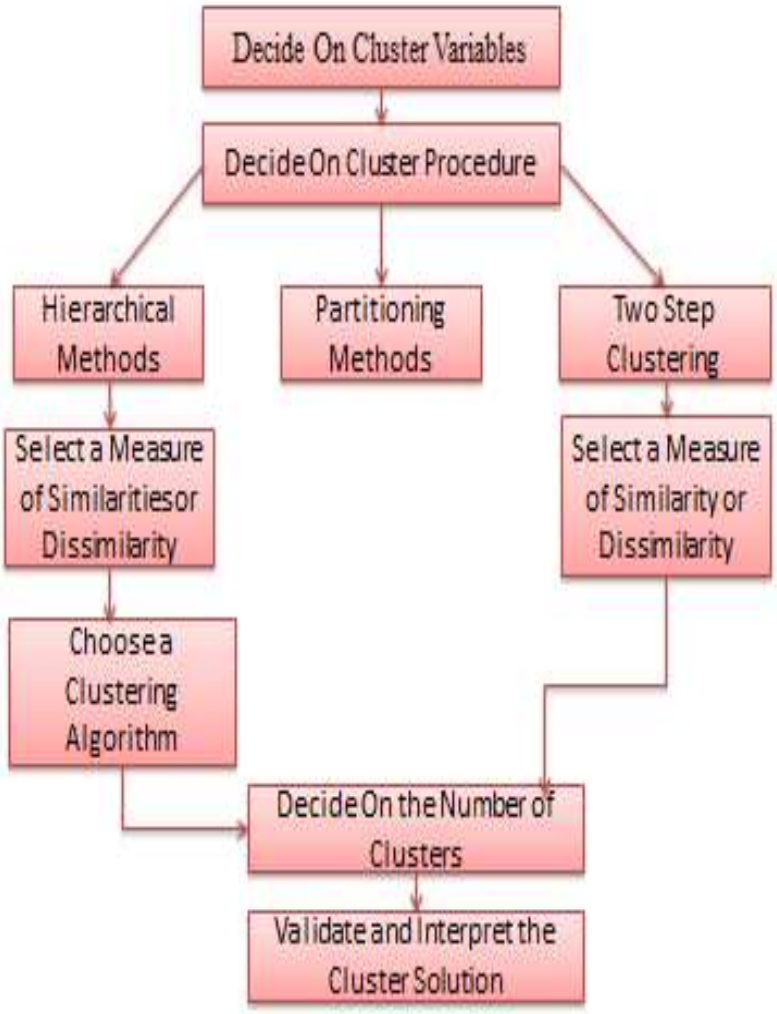

Figure 3: Steps in a Cluster Analysis

1.2.1 Clustering Analysis Techniques

i. Hierarchical: The representation of clusters in these methods is done at different levels of granularity, using dendogram. The system representation can be either agglomerative or divisive on the basis of hierarchical representation. It can be either top-down or bottom-up designed.

- **Agglomerative:** There is a bottom-up approach utilized within these methods where the individual data points are to be started with and further the clusters are merged for creating a tree-like structure. On the basis of the merging of clusters, various choices are possible. On the basis of quality and efficiency, the various tradeoffs are provided. There are various examples such all-pairs linkage, single-linkage, centroid-linkage, and sampled-linkage clustering in which these methods are utilized. There is a utilization of the shortest distance amongst a pair of points within the single-linkage clustering. The average of all pairs is utilized in the all-pairs linkage. However, a sampling of data points amongst two clusters is utilized in sampled linkage. This is used to calculate the average distance of the two clusters. The centroid-linkage process utilizes the distance between the centroids [13].
- **Divisive:** For the purpose of partitioning the data points into a tree-like structure, a top-down mechanism is utilized within these methods. The partitioning at each step can be done by utilizing any flat clustering algorithm. In the terms of hierarchical structure of the tree as well as the level of balance within various clusters, the divisive partitioning is allows flexibility. In the terms of the depths of various nodes or a tree there is no special requirement of having a perfect balanced tree where the degree of each branch is exactly two. Various tradeoffs are provided for the balancing of node depths and node weights to construct a tree structure.

ii. Density- and Grid-Based Methods

The two closely related classes are the density and grid based techniques. Here, the data space is explored at higher levels of granularity. In terms of number of data points in certain defined volume of its locality or in terms of smother kernel density estimate, the density at a specific point within the data space is defined. At certain level of granularity and the post-

processing phase, the data space is explored. The dense regions of the data space are put together within an arbitrary shape. A grid-like structure is formed using the individual regions of the data space within the grid-based techniques of the specific class of density-based methods. As it is easy to put the various dense blocks within the post-processing phase, the grid-based structures are easy to be implemented. Within the high-dimensional methods, those grid-like techniques are also utilized as the lower dimensional grids help in defining the clusters on the subsets of dimensions. The data space is explored at higher level of granularity within these methods which proves to be beneficial. Thus, the complete shape of data distribution is utilized for reconstruction. DBSCAN and STING are the two classical techniques utilized for the density-based and grid-based techniques. Amongst the data points within a continuous space, the density-based methods are naturally defined. This is a very tough task for the density-based methods. Thus, within the discrete or non-Euclidean space, there cannot be a meaningful utilization. For this purpose, an embedded approach is to be utilized. Without any specialized transformations, it is tough to utilize the various arbitrary data types such as the time-series data within the density-based methods. In case of higher dimensionality, the density computations are extremely difficult to define due to the higher number of cells within the grid structure along with presence of sparse data in the grid available [14].

iii. Flat: The data division is done and various clusters are formed with the help of certain partitioning representatives in this case. It is important to partition the representative and distance functions and regulate the behavior of the algorithm. Towards the nearest representatives the data points are assigned on each iteration. Further, as the data points are assigned to the cluster, the representative is adjusted. The iterative nature of EM algorithm is compared with this technique as there are soft tasks performed in each E-step and model parameters are adjusted within the M-step. There are various methods which help in creating the partitions which are described in the section below:

- **k-Means:** The mean of each cluster is related to the partitioning representatives within these techniques. There is no original data set from which the partitioning representative is drawn. It is designed as the function of the data present. For the purpose of computing the distances, the Euclidean distance is utilized. One of the

simplest techniques for the purpose of clustering data is the k-means method. Due to its simple nature this method is used most widely in the practical implementations.

- **k-Medians:** For the purpose of creating the partitioning representative, the median within each dimension is utilized within these techniques instead of using the mean. From the original data set the partitioning representatives are not drawn in k-means technique. There is high sensitivity of the median of a set of values due to the extreme values present in the data in the case of k-medians approach. Hence, this technique is more stable to noise as well as outliers. The partitioning representatives are drawn from the original data in the case of k-Median technique which is otherwise also known as k-Medoids technique. However, both these techniques are not the same and have variations within them [15].
- **k-Medoids:** From the original data present, these methods sample the partitioning representative. The cases which involve the clustering of data points which are arbitrary objects, these techniques are involved. The functions of these objects are not be much discussed here. For instance, discussing about the mean and median of a set of network or discrete sequential objects is not meaningful. In these situations, from within the data the partitioning representatives are achieved. The iterative methods help in enhancing the quality of those representatives. From within the representatives, one representative is replaced from within the current data of each iteration. This helps in determining whether the quality of clustering is enhanced or not. This method is thus considered to be as of the hill climbing method. As compared to the k-means and k-medoids techniques, these methods need more iteration. The situations in which the discussion of means or medians of data objects is not meaningful, this method can be utilized. This is however, not possible in the case of the other two methods.

1.2.2 Clustering applications:

- Intermediate Step for other fundamental data mining problems: The solution to most of the data mining issues for instance classification is done through the summarization of data which is mainly known as the clustering. For various types of application-specific organizations, the less information related to data is helpful.

- Collaborative Filtering: The summarization of closely related users is done through the collaborative filtering techniques. The collaborative filtering is done using the ratings which are given by the various users towards each other. This helps in providing certain recommendations as per the requirements to enhance them.
- Customer Segmentation: The collaborative filtering is similar to this method as there are groups which involve similar clusters within the data. The only difference here is that the arbitrary attributes related to the objects are utilized here for clustering rather than the rating information.
- Data Summarization: There are various dimensionality reduction methods which provide the clustering techniques. These techniques help in providing data summarization which further helps in providing compact data representations. These representations help in providing usage in various applications which is easier.
- Dynamic Trend Detection: There are various dynamic as well as streaming algorithms which are utilized in order to detect data in various applications which involve dynamically clustered data. Various patterns of changes are performed here. For instance, the multidimensional data, text streams, trajectory data, etc. With the help of clustering methods, the key trends as well as events in data are identified.
- Multimedia Data Analysis: The multimedia data involves the images, audio, video and various types of documents. There are huge applications such as recognition of similar snippets of music, or pictures are involved for the recognition of similar segments. There are various types of data and it might also involve the multimodal representation in various instances [16].
- Biological Data Analysis: Due to the evolvement of human genome as well as various kinds of gene expression data, the biological data is very important. The sequences or networks can be formed for the purpose of structuring the biological data. Better ideas for providing new trends related to data are done using the clustering algorithms.
- Social Network Analysis: For determining the important communities within the network, the structure of social network is utilized. Within the community detection there is a better understanding of the community structure within the network, which helps to introduce it in the social network analysis. The social network summarization

also utilizes the clustering technique which is used in various applications. There are also applications related to clustering within the social network summarization [17].

1.3 Predicting Heart Diseases using techniques of data mining

Data mining is utilized for extracting the hidden patterns in the enormous datasets to obtain valuable information. Vast amount of data is generated day by day by utilizing diverse symptoms of patients as well as the clinical reports in the medical fields. Therefore data mining is utilized vigorously for obtaining valuable information from the huge database. For the clinical diagnosis, the explored hidden pattern from the database is utilized. Notwithstanding, now a days the medical datasets are widely increasing, which are heterogeneous in nature. With the help of the hospital management these datasets should be combined and sorted out.

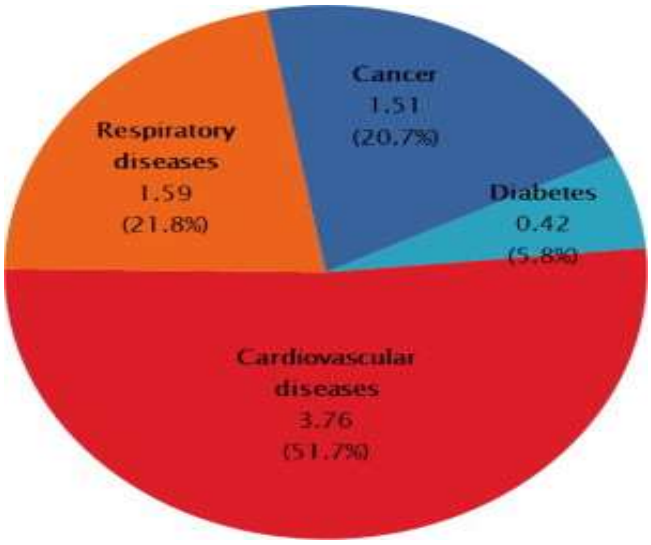

Figure 4: Phi Chart of Heart Diseases

Heart diseases are most common disease that is spreading rapidly in the today's generation. As per world health association around 12 million people died due to this heart problem. This disease is occurring among any age of people due to their poor diet, hereditary and life style which means that people's life is at risk. As it is increasing day by day it needs to be

analyzed correctly and also accurately. Regularly, it can be analyzed utilizing the help of medical specialist. In health care organization if the technique gets combine with the medical system then it would provide us with advantageous information. The risk factors can be described in the form of this chart:

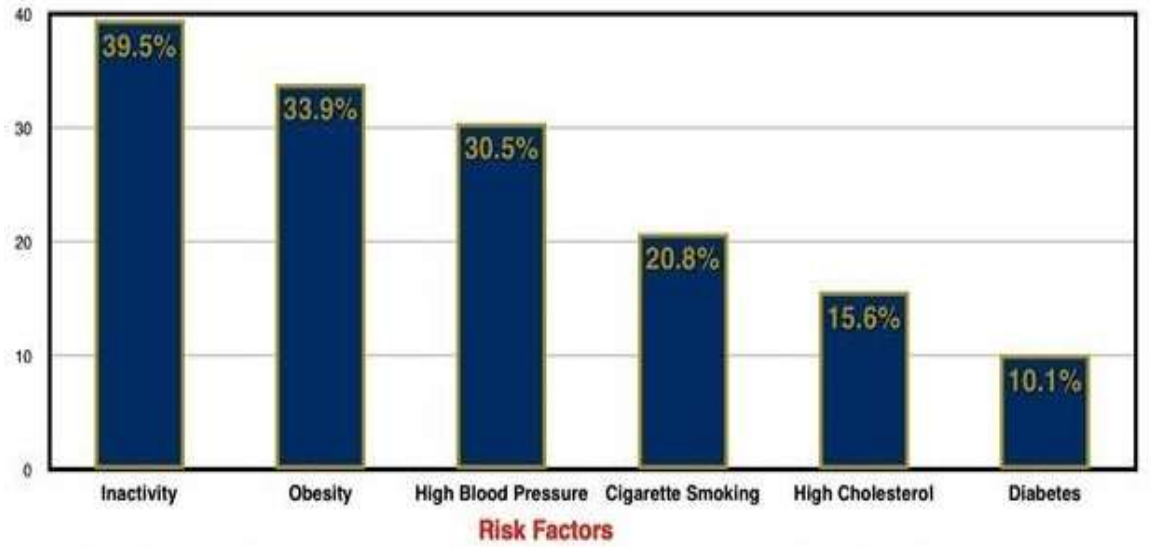

Figure 5: Risk Factors of Heart Disease

It can facilitate accurate diagnosis at low cost based on the information of pc and decision support system. This integration of fluctuated data mining techniques with the current medical decision support system requires comparison of all the disease of heart. The main reasons of suffering from heart diseases are utilization of tobacco, poor diet, hereditary, life style and also by alcohol utilization. To diagnosis the heart disease, the researchers utilize the statistical and data mining tools to help the professionals of healthcare organization [10].

From the past experience it has been noticed that algorithms and complex data are of great benefits with the existing package and software. This technique is commonly utilized in the various fields like crime analysis, engineering, medicines, prediction of expert, portable computing and mining in field of web. Medical diagnosis should be predicted accurately because it is very complicated as well as critical task. Data mining is always considered to be an important as well as necessary step of knowledge discovery. To explore the relationships

and also the hidden patterns from the given database, the data mining had to join with the machine learning, database innovation and statistical analysis.

Data generally uses two kinds of learning such as unsupervised as well as supervised learning. Training set is required in supervised learning to learn the parameters of the models while no training set is needed in unsupervised learning. Every technique of the data mining serves as an alternate based on the objectives of the modeling. Prediction and classification are the common objectives of modeling. Prediction of categorical marks (discrete, unordered) is done in classification models while prediction of continuous-valued functions is done in prediction models. Many techniques like Naive Bayes, networks of neural, decision tree and k mean algorithm shows different levels of accuracy which can be utilized as a part of heart infection treatment.

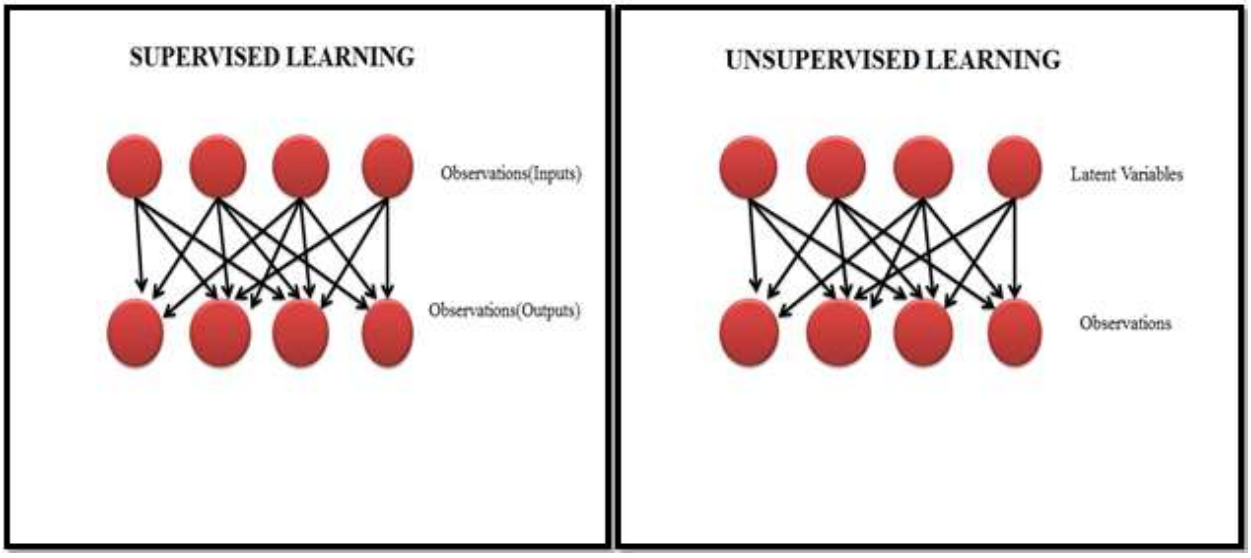

Figure 6: Supervised & Unsupervised Learning

1.4 Algorithms used in Heart Disease Prediction

Distinctive algorithms of supervised learning that is Neural Network, Naive Bayes, Apriori calculation of association, calculating the decision for the dataset have been utilized in this study. Weka 3.6.6 which is one of the data mining tools is utilized in testing. In performing

the data mining tasks weka is considered as a collection of machine learning algorithms. Either the algorithms can be straightforwardly connected to the dataset or from own specific code of java it can be called easily. For performing classification, data pre-processing, clustering, regression, visualization and association rules weka contains certain tools. It is likewise appropriate for developing new machine learning schemes.

a)Decision Tree: Decision tree learning is a method generally used as a piece of data mining. The goal is to make a model that predicts the value of a target variable in view of a couple input variables. Each interior node relates to one of the input variables; there are edges to children for each of the conceivable values of that input variable. Each leaf represents a value of the target variable given the values of the input variables spoke to by the route from the root to the leaf. A decision tree is a simple representation for classifying cases. For this segment, the larger parts of the features have finite discrete domains, and there is a solitary target feature called the classification. Each element of the domain of the classification is known as a class. A decision tree or a classification tree is a tree in which each internal (non-leaf) node is marked with an input feature. The circular segments starting from a node named with a feature are marked with each of the conceivable values of the feature. Each leaf of the tree is marked with a class or a probability distribution over the classes. There is no requirement of domain knowledge or parameter setting and can high dimensional data can be handled. It produces results which are simpler to read and interpret. The drill through feature to access nitty gritty patients' profiles is just accessible in Decision Trees [11].

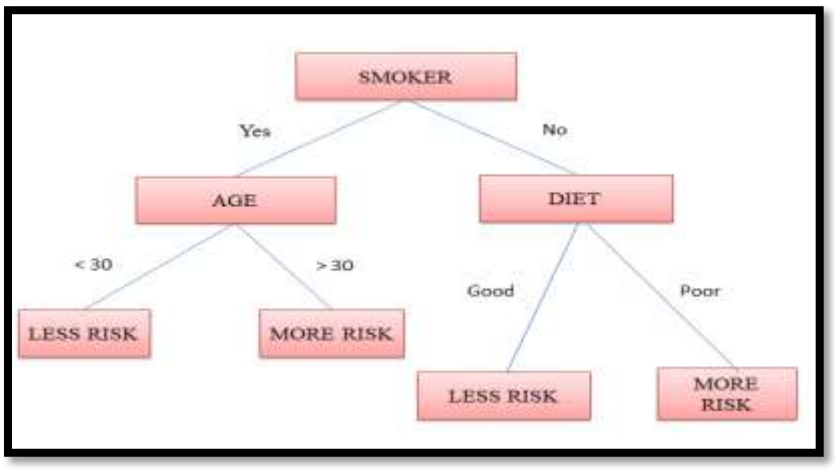

Figure 7: Decision Tree for Heart Disease Prediction

b. Naive Bayes: This classifier calculation utilizes conditional independence; means it expects that an attribute value on a given class is independent of the values of different attributes. The advantage of utilizing credulous Bayes is that one can work with the Naïve Bayes model without utilizing any Bayesian methods. The performance of the algorithms like Decision Tree, Naïve Bayes, Neural Network and K- Mean for accuracy level is shown below in the figure.

c. Neural Networks: It is a mathematical model or computational model in light of biological neural network. At the end of the day, it is an emulation of biological neural system. In nourish forward neural networks the neurons of the main layer forward their yield to the neurons of the second layer, in a unidirectional fashion, which explains that the neurons are never obtained from the directions in reverse. In between the every layer the connection is built and to every connection weights are also given. The main function of input layers of neurons is generally to break the input xi in hidden layers into many neurons. In hidden layer the neurons includes input xi along with weights wji of single connections from the input layer. The yield Yj is a function which is represented as

$$Yj = f(\sum wji\ xi)$$

Here, f is nothing but a simple threshold function like hyperbolic tangent function.

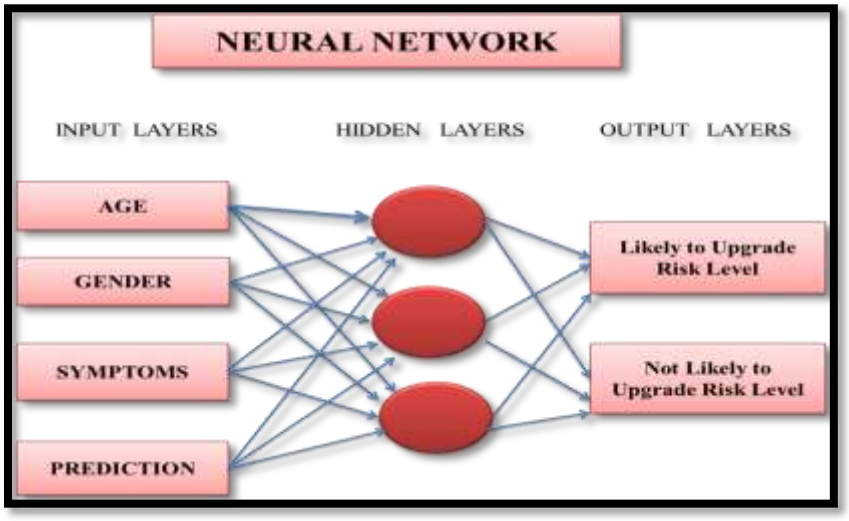

Figure 8: Neural Network

d. K-Nearest Neighbor

K-Nearest neighbor classifiers depend on learning by analogy. The training samples are depicted by n dimensional numeric attributes. Every sample represents a point in an n-dimensional space. Along these lines, the greater part of the training samples is stored in an n-dimensional pattern space. At the point when given an unknown sample, a k-nearest neighbor classifier looks the pattern space for the k training samples that are closest to the unknown sample. "Closeness" is defined in terms of Euclidean distance. Not at all like decision have tree induction and back propagation, nearest neighbor classifiers assigned break even with weight to every attribute. This may bring about confusion when there are numerous irrelevant attributes in the data. Nearest neighbor classifiers can likewise be utilized for prediction, that is, to give back a genuine valued prediction for a given unknown sample. For this situation, the classifier gives back the average value of the genuine valued associated with the k nearest neighbors of the unknown sample. The k-nearest neighbors' algorithm is among the simplest of all machine learning algorithms [20].

CHAPTER 2

LITERATURE REVIEW

1. Ranganatha.S et.al "Analysis For Heart Disease And Medical Data Mining Dataset Using Techniques Of Classification, 2013":

A modern medicine in today's world generates a lot of information that are stored in database. It is important to extract the valuable information from the database to diagnosis and provide treatment for diseases based on the decisions. So data mining was used to solve the problem of heart disease. The management of hospital information can be improved very easily. Primarily the patient care activity was managed by the medical fields and it is also directed to the resources of the [12]. The main motive of gathering these data from the field of medicines is to get helpful for each and every patient. The study mainly stores the information of the patients who are suffering from heart disease and need to be hospitalized. The algorithms of data mining such as naive bayes and ID3 will be utilized when data sets are large or complex. ID3 outputs the outcome as decision tree which can be effectively caught on. The chances of heart diseases are predicted by utilizing the Bayesian for a given condition. Because of the immeasurable volume of information that is generated, people developed few algorithms that deliver output for a given query. The objective of this paper was to store large chunks of data and it gives Smooth workflow.

2. Syed Umar Amin, et.al "Data Mining Based On Genetic Neural Network In Prediction Of Heart Disease By Utilizing Risk Factors,2013":

Data mining techniques have been broadly utilized as a part of decision support systems that helps in the prediction of different heart diseases accurately. These ideas are exceptionally working smoothly for designing the clinical support systems [13]. A standout amongst the most critical usage of these systems is for treating the heart diseases since everywhere throughout the world heart disease is found as main cause of deaths. There is no such system that predicts heart diseases in aspect of risk factors like, age, diabetes, obesity, physical inactivity, taking tobacco, having high cholesterol, intake of alcohol etc. Patients suffering

from heart disease are found with obvious factors of risk which is utilized for the treatment. Such systems which is found with some of this risk factors helps the experts to predict the chances of geting heart disease before the patient goes for a costly checkup. Subsequently, for the heart disease prediction this study presents a method by utilizing real factors of risk. The best data mining tools are included in this method such as genetic algorithm and the networks of neural. The learning is more quick, more stable and accurate when it is compared with the learning of back propagation. The research includes prediction of risk levels of patients suffering from heart disease and finally the system will be implemented in Matlab.

3.Theresa Princy et.al "Prediction Of Human Heart Disease System Using Techniques Of Data Mining,2016":

Health diseases are increasing every day because of different ways of living life and family history. Particularly, disease of heart has turned out to be more normal nowadays which indicates the level of risk of the patients. The high blood pressure, rate of pulses and high cholesterol differs from one person to another person in every individual. In this paper the survey about different techniques utilized for predicting the risk level of every person is given in view of age, blood pressure, cholesterol, gender, and pulse rate [14]. The level of risk for each patient can be classified utilizing data mining arrangement techniques, for example, KNN, Naive Bayes, Decision Tree, Neural Network algorithm and so forth. Due to more attributes in the data sets the accuracy level of the risk is very high. The fundamental motive of this study is to give overview predicting risk rate of diseases of heart by using the techniques of data mining. Different classifiers and techniques of data mining are utilized for efficient diagnosis of heart disease. According to the mode of analysis, it is found that different authors using attributes that are totally different and technologies for their research study. Therefore, each and every technology gives different results based on the attributes. By applying K Nearest Neighbor and ID3 algorithms the heart disease was detected and for different number of attributes the accuracy level is also provided. In future, the accuracy would be increased and the attributes in number will be reduced by utilizing some different algorithms.

4. K.Srinivas et.al "Prediction And Analysis Of Coronary Heart Disease In Coal Mining Regions Using The Techniques Of Data Mining,2010":

Diagnosis of Heart disease medically is very confusing task that ought to be accurately performed. The survey helps in analyzing the study in aspect of risk rates among the different age of people [15]. Subordinate several incorporate measures that are self- reported of being determined to have disease of cardiovascular such as (1) stroke (2) chest pain (3) heart attack. The morbidity is predicted by the study of heart indicating many attributes. The system that is automated enhances care provided by medical centers and reduces cost for medical diagnosis. In this research work, well known techniques of data mining specifically, making decision tree, using Naive Bayes and Networks of Neural are utilized in the treatment of heart disease. This can advance enhanced and further expanded. Fundamentally the list of attributes mentioned in database of medical center is utilized for predicting heart attack. Other than this rundown, we need to combine different types of attributes that has an impacts on the output like person's financial status, pollution, tensions and past history of any medical record. Other than the techniques of data mining, clustering, series in aspect of time and rules of association are additionally utilized to judge the behavior of the patients. It is therefore used to study the condition of patient's morbidity in the health care centers.

5. R. Kavitha et.al "Framework Used In Data Mining For Heart Disease Classification Using Feature Selection And Feature Extraction Technique,2016":

In order of high dimensional data set in the database of heart diseases is utilized as a part of the data mining stage of pre-processing. The crude database comprises of irrelevant and conflicting information subsequently incrementing the pursuit data storage data. One needs to remove the irrelevant and redundant data that are present, to achieve the characterization accuracy. The strategy of reduction is utilized to combine the data from higher to lower dimensional data with a few variables. For the simple heart disease prediction a structure is integrated [16]. The structure made by utilizing the essential segments to extract the features and numerical model is processed to choose the pertinent characteristics utilizing the significant requirement. The work proposed, can be used to enhance the accuracy,

effectiveness, and process having speed .The study may be used in image processing, recovery of data, and coordinating patterns. The subnet of component is selected and utilized by utilizing wrapped channel classifier to give appropriate output. When compared with other scoring function like Pearson correlation coefficients and Euclidean distance, improving the performance of the system. It facilitates later on efforts the exception will be addressed by demonstrating better level of accuracy.

6. Eman Abu Khousa eta.al "Predictive Data Mining To Support Clinical Decisions: An Overview Of Heart Disease Prediction Systems,2012":

The Organization of health care is facing with many problems to give high quality of care to the patients in less cost. The volume of data that are accessible from the healthcare information systems in the database will be analyzed by both clinicians and administrators with a specific end goal to invent learning and settle down based on some decisions. It is mainly used to improve the diagnosing level of disease and its prevention is also improved [17]. It is the fate if there arises an occurrence of diseases related to heart which will be viewed as a major cause of death in individual. It utilizes analysis tools to discover the valuable patterns as well as hidden relationships in medical data. In this study, it surveys the five models that are developed to support decision of single and combined data mining techniques in clinics for diagnosis and prediction in heart disease. Every systems display some strengths as well as drawbacks regarding the kind of data it handles, accuracy, interpretation, simplicity of, speculation ability and reliability. Poor speculation is still a noteworthy problem in healthcare basically as a result of shortage of information. In this research work, we actually studied regarding the systems that include five several techniques of data mining as well as the several types of medical data to predict the patients suffering from heart diseases. Each effective models from those of five takes out a valuable patterns for the prediction such as drawbacks and strengths, interpretation of simplicity, data accuracy and speculation ability and reliability.

7. Monika Gandhi et.al "Predictions In Heart Disease Using Techniques Of Data Mining":

A vast amount of information is delivered in medical affiliations but data is not appropriately utilized. The healthcare systems are fully fledged in aspects of data storing but however poor in matter of fetching data. To discover connections and patterns, the successful analysis methods are found to be absent in healthcare systems. Several methods are used in data mining which acts as a solution in this circumstance. Hence, several techniques of data mining can be utilized [18]. The aim of the present effort is to discover the parts of utilization for aid of people in healthcare systems by strategy for machine adapting moreover procedures of data mining. The primary points basically recommend the system which is automated for diagnosing heart diseases by considering prior information and data. This paper gives information regarding the different knowledge abstraction techniques by utilizing methods of data mining which are being utilized for prediction of heart disease in today's research. In this research work, methods of data mining to be specific by using algorithm of decision tree, Naive Bayes and Networks of Neural will be analyzed on medical data sets. Each and every data mining techniques have advantages as well as disadvantages of classification of data and extraction of knowledge. Moreover, decision tree, Networks of Neural or Naive Bayes can be researched in deep, which can be used in future in healthcare organizations.

8. Heon Gyu Lee et.al "A Data Mining Approach For Coronary Heart Disease Prediction Using HRV Features And Carotid Arterial Wall Thickness,2008":

The fundamental goal of this work is to build and after that intend another and methodology that are unique and are used in building the different characteristics of heart rate variability in treating the patients suffering from heart diseases. The people likewise intend appropriate prediction model to improve the treatments for cardiovascular disease [19]. For three recumbent postures the HVR is analyzed. The collaboration impacts in between the groups of normal people and the recumbent postures and patients suffering from heart diseases were seen in light of Heart Rate Variability lists. The arteries is measured and utilized by the estimations of blood vessel which are very thick enough as different characteristics. Heart

disease patients experienced scanning of vein utilizing ultrasound devised of high resolution as part of a past study. Keeping in mind the end goal to extract different features, six classification methods are tested. For classification, the proposed different features are employed permitting to pick from a large pool of classifier all around concentrated on classification methods. Researchers propose the likelihood that features of multi-parametric, contemplating entire conceivable HRV features as a demonstrative tool might use in treatment of cardiovascular disease. Finally, a few methods of the supervised learning are considered which includes the Bayesian classifiers which are extended, CMAR, SVM and MDA. From the output of experiments, the methods of classifications such as SVM and CPAR outperform.

9. Limia Abed et.al "Using Data Mining Technique To Diagnosis Heart Disease,2012":

Diagnosing the medical problems is a guarantee to makes use of the techniques of data mining. An expert helps in the treatment, which generally involves one human expert which may even commit some mistake. Interestingly the knowledge is extracted from chunks of clinical data by data mining and delivers the models of prediction and also utilizes the task of classification to get the treatment. Several methods are presented in to deliver the classifiers in the field of classification [20]. Naive Bayes is one of them. In this research work, previously executed experiment were discussed with the help of Naive Bayes technique keeping in mind the end goal to manufacture models of prediction which helps in an artificial treatment for the people suffering from heart disease in view of set of parameters contained in dataset which were measured already for the individuals. At that point compare the results with different techniques as indicated by utilizing similar to UCI repository data. The outcome that is derived considered as good from practical work, while every one of the answers were right. So the model is accurately achieved by the ratio (100%). The outcome was matched with different performance similar to the dataset. Furthermore, another problem arises in the field of treatment, where every heart disease contains own parameters which combine together in diagnosing, and diseases that are totally different from each other. Therefore in this research work, very good outcomes were presented in the diagnose of heart

utilizing the images of Single Proton Emission Computed Tomography (SPECT) images with classifiers of naive bayes are compared with the other classifiers.

10. Sellappan palaniappan et.al "Intelligent Heart Disease Prediction System Using Data Mining Techniques,2008":

The healthcare organization collects bulk of raw data from the healthcare system to extract hidden patterns to obtain valuable information that may be helpful in making standard decisions. Inventing of invisible patterns as well as relationships sometimes gets unexploited. To solve this problem, data mining techniques helps a lot to rectify this problem [21]. A prototype has been developed in this research papers utilizing the techniques of data, in particular, algorithm of decision Trees, Bayesian and Network of Neural. It demonstrates the outcome for every technique in understanding the defined mining to derive the objectives. It basically answers the complex "imagine a scenario where" which traditional decision support systems queries can't do. Medical profiles of the patients are utilized from the database like gender, patient's age, circulatory strain, and the glucose level which detects probability of occurring heart diseases among the patients. It gives the sufficient knowledge such as pattern recognition, deriving relationships between the factors of medical data identified with diseases related to heart which need to be set up. The IHDPS is considered as user-friendly, based on web, expandable as well as reliable which can run on .NET platform. IHDPS is improved and expanded in future. For instance, it puts together the attributes of other medical data. Therefore, it incorporate other techniques of data mining like clustering, Series of time and Rules of Association. Instead of using simply categorical data, continuous data is utilized. Mining of unstructured data that are available in the database from large chunks of data, text mining is utilized. Another great challenge is to integrate text mining as well data mining together.

11. Mai Shouman et.al "Using Data Mining Techniques In Heart Disease Diagnosis And Treatment,2012":

The accessibility of huge chunks and chunks of data from the medical data set leads to the requirement for tools of data analysis which is very powerful in extracting valuable information. The statistical and data mining tools are utilized by the researchers which are of

great concerned for improving the analysis of data on large sets of data. Treatment for diseases is one of the applications in which data mining tools shows successful results beyond doubt [22].From the past few years the main cause of death being reported is heart disease. In order to detect to various forms of heart diseases, various experts are using techniques like the statistical and the data mining ones. The highest level of perfection and approximation was being achieved by using single data mining techniques in order to detect the various heart diseases. By the process of hybridization, the researchers hybridize one technique with another in order to tackle the problems of heart disease. As, by using this data mining techniques, the main identification criteria for the various heart diseases has attained less attention. As in this research, the main focus is to identify the gaps that arise during the treatment process and close and reduce these kinds of gaps. These are mainly used by the process of the data mining in this research paper in order to get a better performance.

12. Sivagowry.S et.al "An Empirical Study On Applying Data Mining Techniques For The Analysis and Prediction Of Heart Disease,2013":

In this research paper, the main focus was done in order to get good information with the help of data mining technique. When talking about health care centers, they are best at providing information, but the main problem arises in extracting this information and by data mining it is extracted. The main reason behind about extracting this is due to shortage of effective analyzing kind of tools. For extracting the information, here the data mining was used to give a good information. As the problem of heart and its disease is one of the biggest problem that is of great concern as various deaths are also caused by this. So for this researches used various hybrid techniques and ways to tackle this problem. In this paper the mining techniques as used to detect the heart problems. Here the technique of mining used was of classification type. In this, it played an important role and was such beneficial than the methods like the clustering, Regression and also the association type. In classification, the decision tree is used and was of greater much beneficial than Neural type of network and the Naive Bayes in various cases. The conclusion was that by the data mining techniques, it was one of the best methods in order to deal with the heart problems with lesser number of variables.

13.K. Prasanna Lakshmi, et.al,"Fast Rule-Based Heart Disease Prediction using Associative Classification Mining", 2015

Associative classification is one of the approach which consolidate classification and associative rule mining. This technique is pulled in numerous researches as it determines the classifier accurately with powerful rules. The associative classifiers are valuable for the application where the maximum predictive accuracy is fancied [33]. Healthcare organization collects huge volume of data which are not mined to discover hidden knowledge for taking choice. It is required to develop a choice support system for anticipating the large datasets that are generated by healthcare industry. In this work, we technique used for heart disease prediction is very efficient. The classifier with prediction of high intriguing quality values are built by associative classification which are mostly used by many researchers. The result helps the doctors to make a better decision in predicting diseases.

14.Elena Baralis, et.al,"A Lazy approach To Associative Classification", 2008

Associative classification is a technique which is responsible for building accurate classifiers accurate classifiers. The rule mining may yield huge rule sets for large or correlated datasets or huge, affiliation [34]. To choose small subset of high quality rules a few pruning techniques have been proposed. The accuracy of the classifier may improve by the accessibility of a "rich" rule set. The lazy pruning technique build the L^3 associative classifier that discards solely rules that misclassify the training data. The classification of data that are unlabeled is performed in two stages. It initially considers a small subset of high-quality rules. A large rule set is exploited when the data set is not ready for classification. To adapt to the need of mining large rule sets and to efficiently utilize them for classification, a compact form is proposed to represent a complete rule set in a space-efficient manner and without information loss. In the result the L^3 improves the classification accuracy with respect to previous approaches by an extensive experimental evaluation on real and synthetic data sets demonstration.

15.K. Prasanna Lakshmi, et.al," Compact Tree for Associative Classification of Data Stream Mining", 2012

The data streams have emerged late to address the problems of continuous data. The process of mining with data streams helps in extracting knowledge structures from continuous, rapid data records by mining with data streams. An important goal in data stream mining is era of compact representation of data [35]. For further basic leadership process it helps in making less time and space. In this work another scheme called Prefix Stream Tree (PST) for associative classification has been proposed. We utilized a PS Tree that was constructed utilizing the idea of prefix tree and to handle the stream data it was restructured. The memory consumption is reduced by constructed tree which is a compact tree. It helps in finding exact set of late incessant item sets and predicts the class label for the requested tuple. It helps in compact storage of data streams. In a single scan the PS Tree is generated. The exact set of pattern is discovered by this tree from data streams using sliding window.

16.Bing Liu, et.al," Integrating Classification and Association Rule Mining", 2012

Classification and association rule mining aims to discover set of rules which are small in volume in the database to form accurate classifier. Every rules existing in the database that satisfy some minimum support and minimum confidence constraints are discovered by affiliation rule mining. The target of discovery is not pre-decided for affiliation rule mining, while there is only single foreordained target for classification rule mining [36]. In this work, we coordinate these two techniques of mining. The integration is finished by concentrating on mining a special subset of affiliation rules, called class affiliation rules (CARs). For building a classifier in light of the set of discovered CARs an efficient calculation is likewise given. The experimental results demonstrate that the classifier created by state of the art classification system C4.5 is not so accurate whereas the classifier built as a rule is more accurate. The numbers of problems that exist in the current classification systems are solved with the help of integration.

CHAPTER 3

PRESENT WORK

3.1 PROBLEM FORMULATION

This work is based on prediction analysis for heart diseases detection. The prediction analysis contains two phases in which the first phase is clustering and second phase is of classification. In the first phase of clustering technique of k-mean clustering is applied in which dataset is given as input and arithmetic mean of the input dataset is calculated which will be the centroid point, from the centroid point Euclidian distance is calculated. The data points which are similar are clustered in one cluster and other are in the second cluster. The clustered data is classified by applying classification technique. The performance of data classification depends upon the accuracy of clustering. In this work, the difficult task is to obtain string relationship between the attributes of the dataset. In this work, in k-mean clustering the improvement will be proposed to derive the relationship between the attributes. This directly leads to improve clustering and classification accuracy and reduce execution time.

3.2 OBJECTIVES OF THE STUDY:

1. To propose improvement in k-mean clustering algorithm to increase accuracy in clustering.

2. To obtain the string relationship between the attributes of the dataset.

3. To get accuracy and efficiency in result using improved k-mean algorithm.

4. To reduce the execution time in result to save time of any individual.

5. To compare results in terms of accuracy, time in both proposed and existing techniques.

3.3 RESEARCH METHODOLOGY:

The k-mean clustering algorithm is used to cluster the similar type of data for prediction analysis. In k-mean clustering algorithm, the functions are clustered using Euclidian distance formula. To improve the cluster quality, we will improve the Euclidian distance formula. The improvement will be done depending upon the normalization. Additional features will be added in the enhancement. The first point is to calculate normal distance matric on the basis of normalization. The second point is the functions will be clustered on the basis of majority voting. The proposed technique will be implemented in MATLAB.

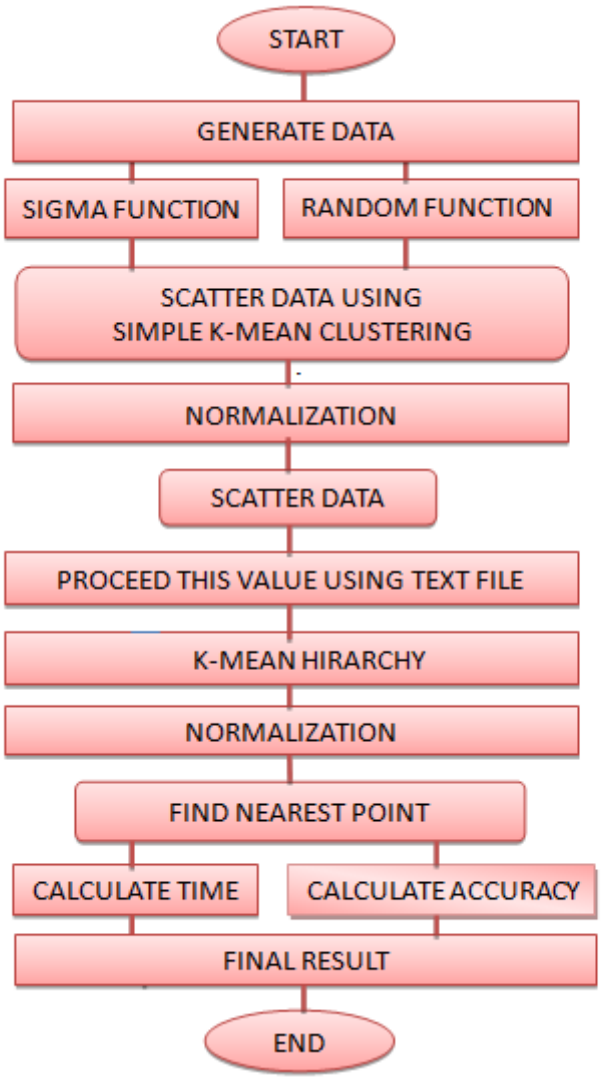

Figure 9: Flowchart of research methodology

First of all we have to start process in which at initial stage we generate data from user end in which we give number of data, that are generated by sigma and random functions. When all data has been generated then Simple k-means apply and getting result in subplot.

After apply normalization on that data we give scatter data in second subplot, now we have to apply normalization in proceed in which we read text file data of that generate data after that apply hierarchy k-means before normalization in which we get result in different form rather than first subplot. Apply normalization on that process in which iterations process start.

This process is continue until we don't get a nearest point to accurate position with generating data, at last calculate their total time in which we get result and then accuracy.

HYBRID ALGORITHM

INPUT : Dataset

OUTPUT: Clustered Data

Start ( )

1. Read dataset and dataset has number of rows "r" and number of columns "m"
2. For (i=0 ;i=r; i++) /// selection of medoid point
    1. For (j=0; j=m; j++)
    2. Select k=data (i, j);

       End
3. Calculation of Euclidian distance()
    1. For (i=0;i=r;i++)
    2. For (j=9;j=m;j++)
    3. A(i)=data(i);
    4. B(i)=data(j);

5. Distance =squt[(A(i+1)-A(i)^2) –(B(j+1)-B(j)^2);

   End

4. Normalization ()
    1. For (k=0;k=data;k++)
    2. Swap k(i+1) and k(i);

       end

5. Repeat step 3 to 4 until all points get clustered.

CHAPTER-4

IMPLEMENTATION

4.1 EXPERMENTAL RESULT

In the first phase of clustering technique of k-mean clustering is applied in which dataset is given as input and arithmetic mean of the input dataset is calculated which will be the centroid point, from the centroid point Euclidian distance is calculated. The similar data points are clustered in a single cluster and other are in the second cluster.

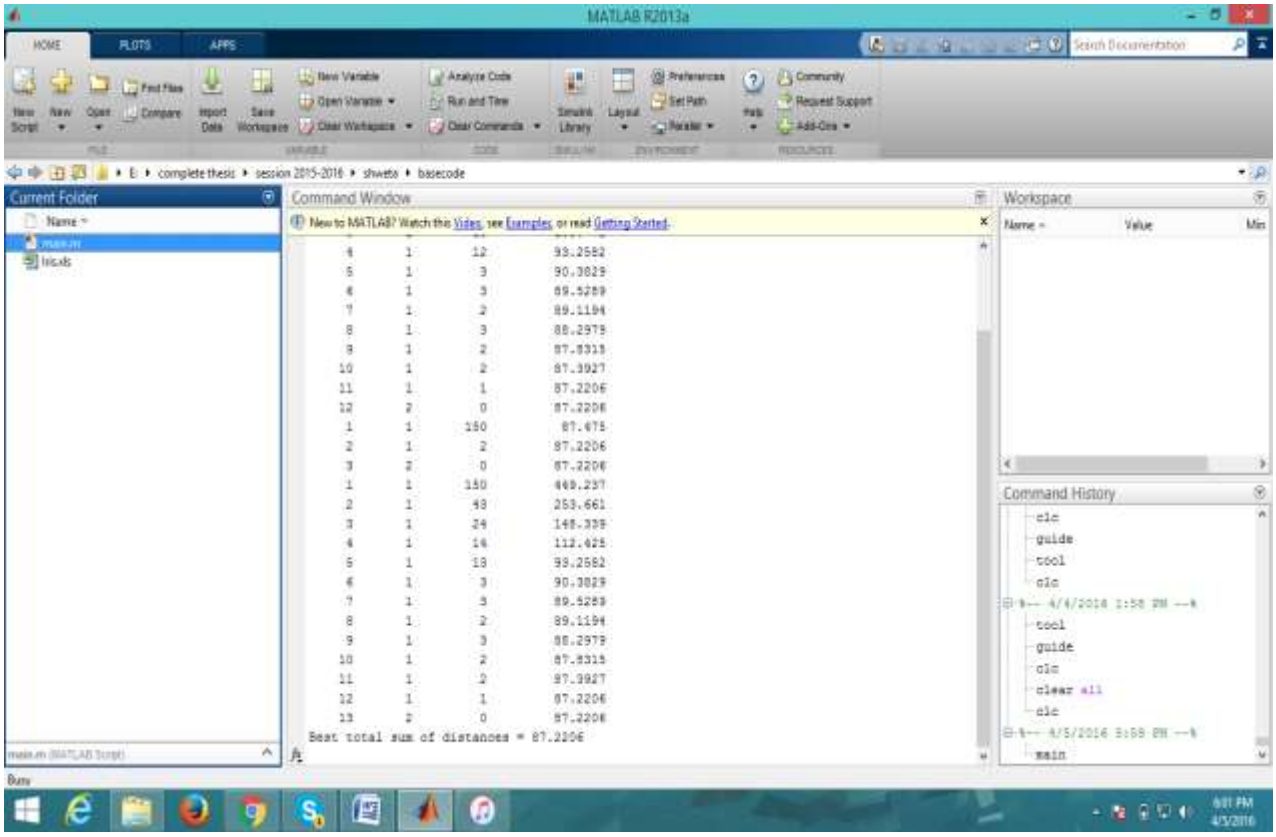

Figure 10: K-mean clustering :- As shown in the figure, the k-mean is the clustering algorithm which is used to cluster the dataset. In k-mean clustering algorithm central points are selected according to the number of clusters. The Euclidian distance is calculated to cluster the dataset. In the figure, the Euclidian distance is calculated to cluster the dataset.

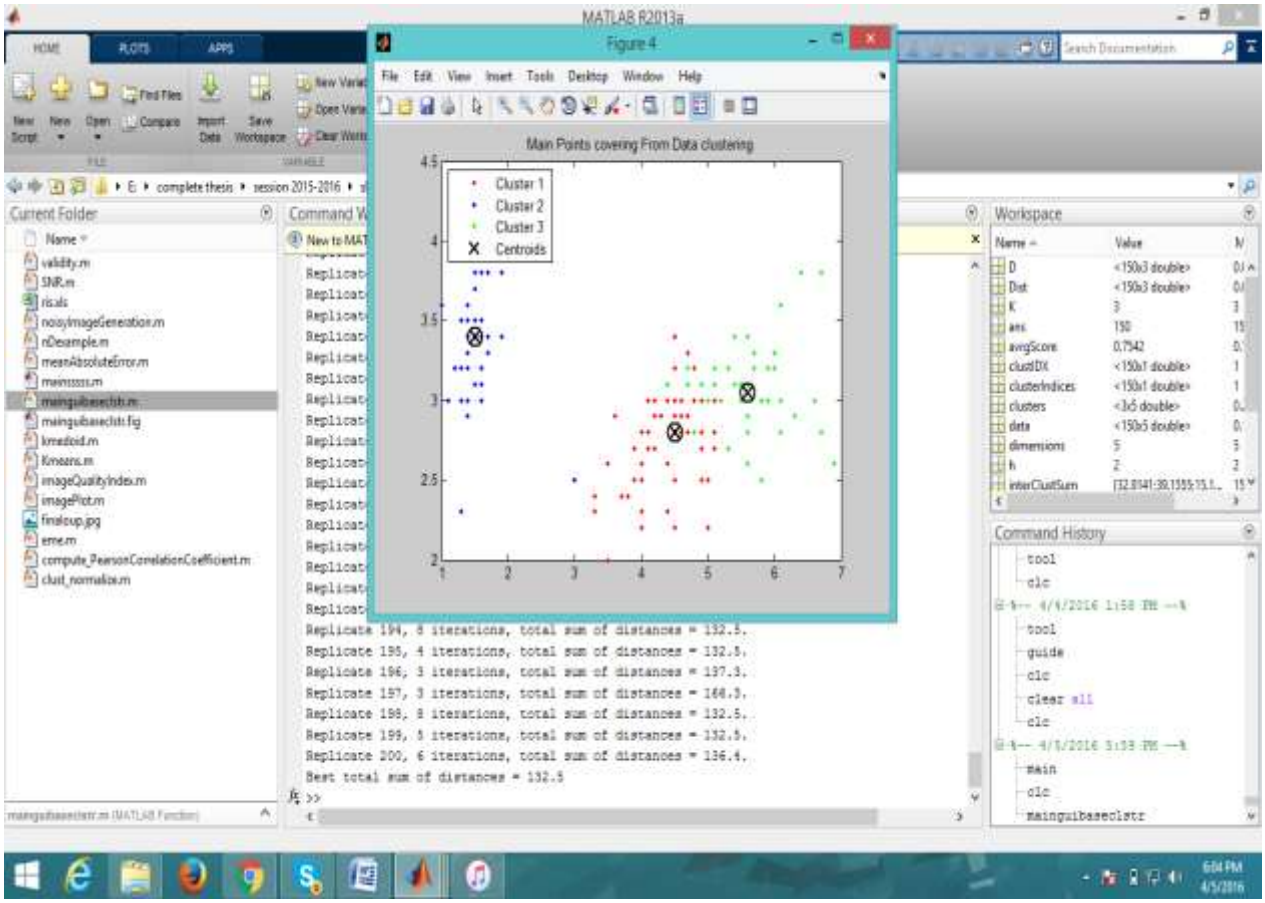

Figure 11: Clusters marked with centroid

As shown in figure, the k-mean clustering is improvement to improve cluster quality using the technique of normalization. The dataset is loaded is and it is shown on the command window. The dataset which is loaded is plotted on the 2-D plane for analysis. The plotted data will be clustered using the algorithm of k-mean clustering. The central points are marked in each cluster. The normalization technique is applied to calculate best distance to make high quality clusters. The final output of the clusters is shown on the 2-D plane.

Final output of the k-mean algorithm

From the above figure, to cluster the dataset k-mean clustering algorithm is used. In the algorithm central point is selected and Euclidian distance is calculated to generate final output. In the final output clusters are formed in which central points are highlighted

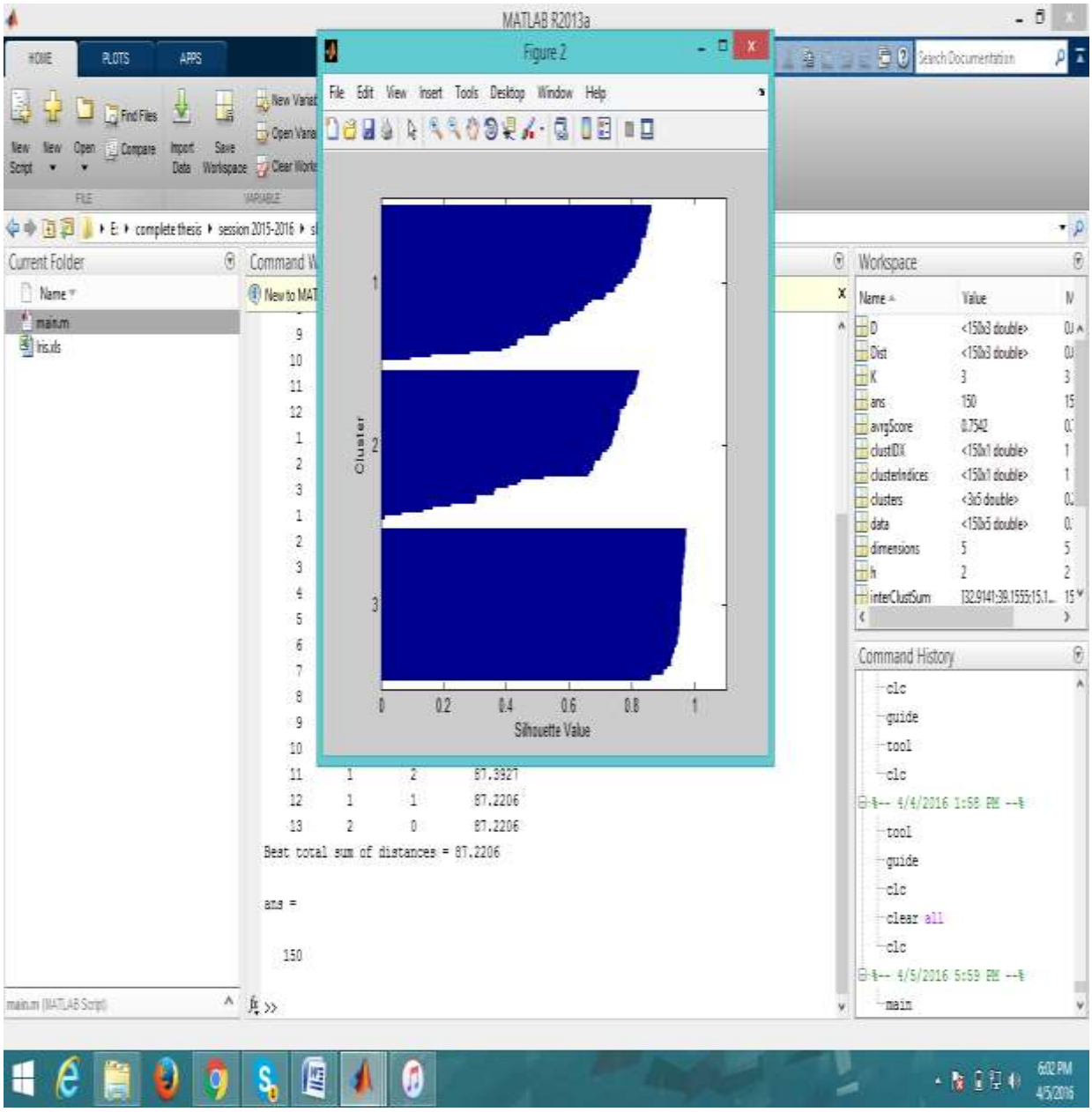

Figure 12: Cluster quality analysis

As shown in figure, the k-mean clustering algorithm is used to cluster the dataset. The central point is selected in the algorithm and Euclidian distance is calculated to generate final output. To analysis the cluster quality, graphs are drawn and it is analyzed that cluster quality of third cluster is low.

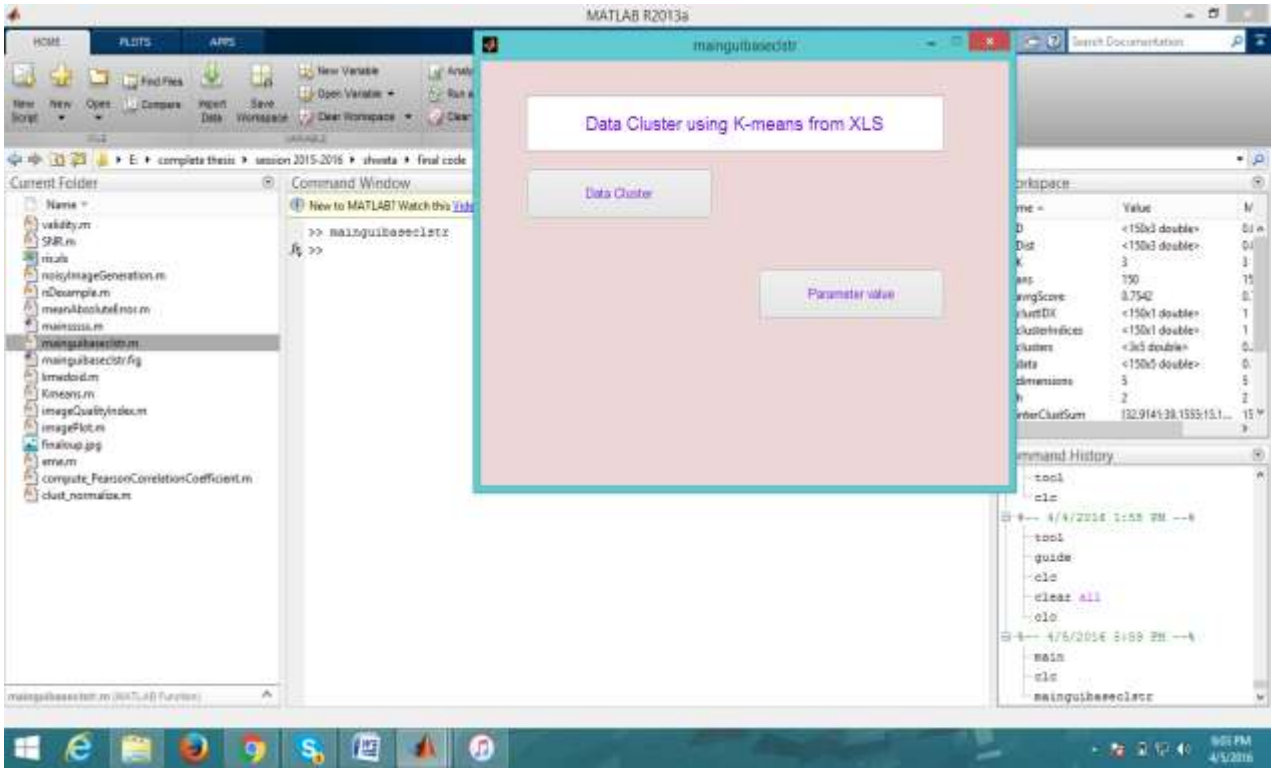

Figure 13: Data Cluster Using K-mean from XLS :- As shown in figure, the improved k-mean clustering is used to improve cluster quality using the technique of normalization.

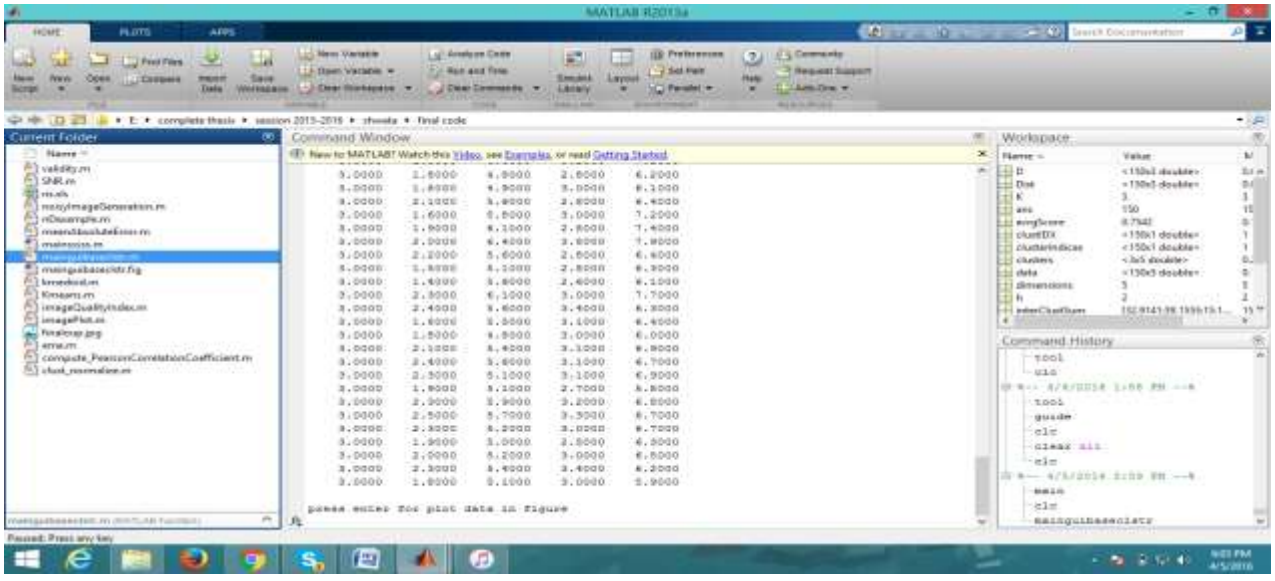

Figure 14: Dataset loaded in K-mean Clustering :- As shown in figure, the improved k-mean clustering is used to improve cluster quality using the technique of Normalization. The dataset is loaded and it is shown on the command window.

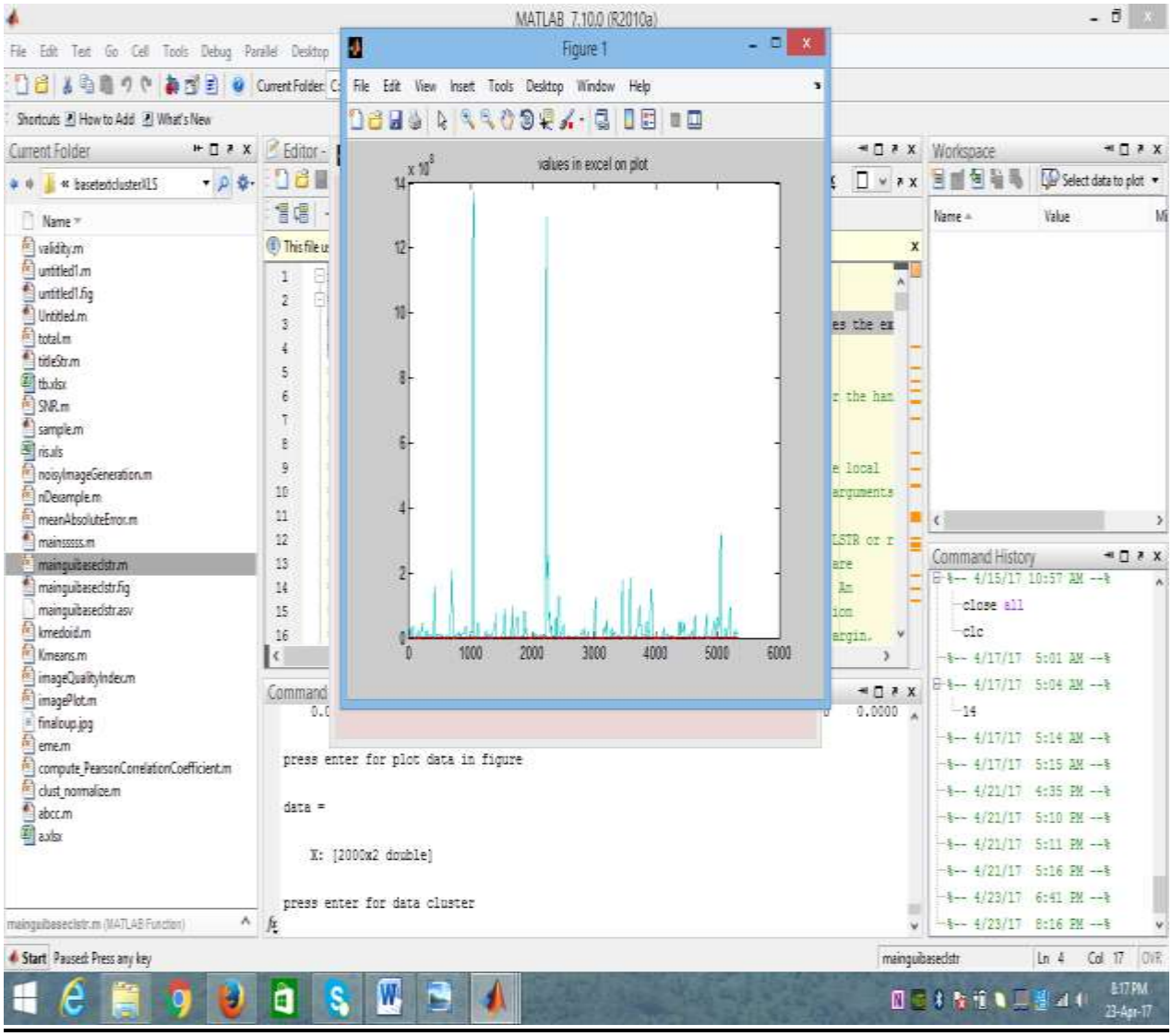

Figure 15: Plotted Dataset

As shown in figure, the improved k-mean clustering is used to improve cluster quality using the technique of normalization. The dataset is loaded and it is shown on the command window. The dataset which is loaded is plotted on the 2-D plane for analysis.

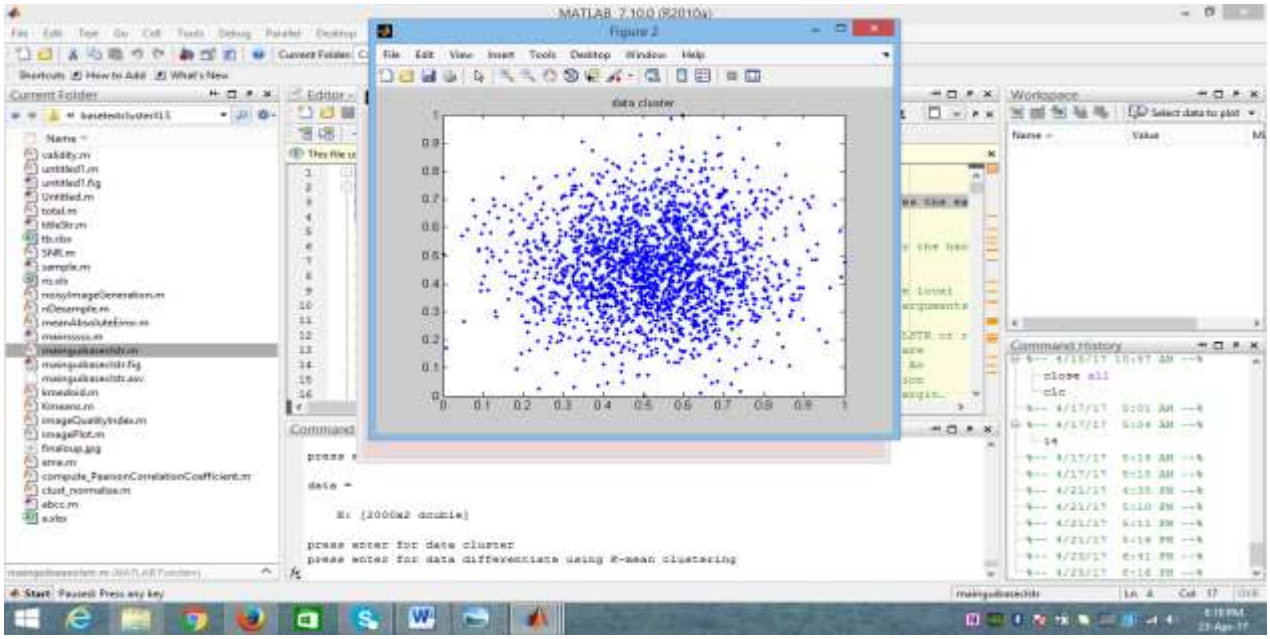

Figure 16: Data in Single Cluster :- As shown in figure**,** the dataset which is loaded is plotted on the 2-D plane for analysis and then the data is combined to form a single Cluster.

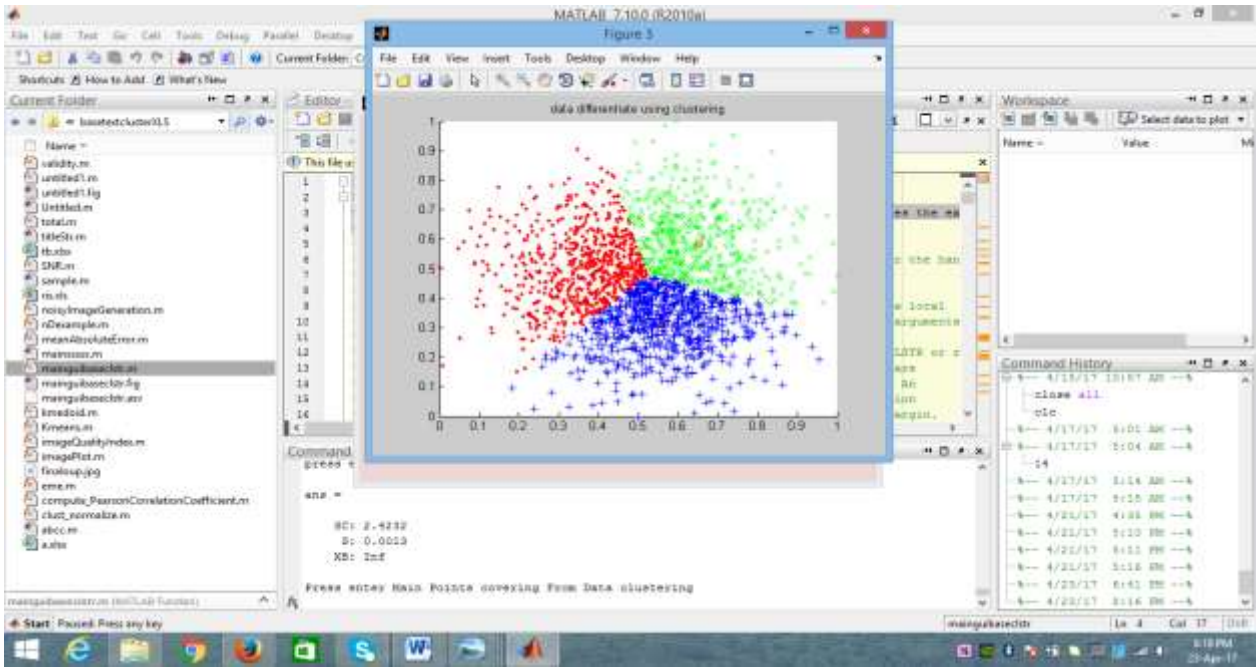

Figure 17: Data Differentiated Using Clustering :- As shown in Figure, the single data cluster is partially divided into three different forms of Cluster.

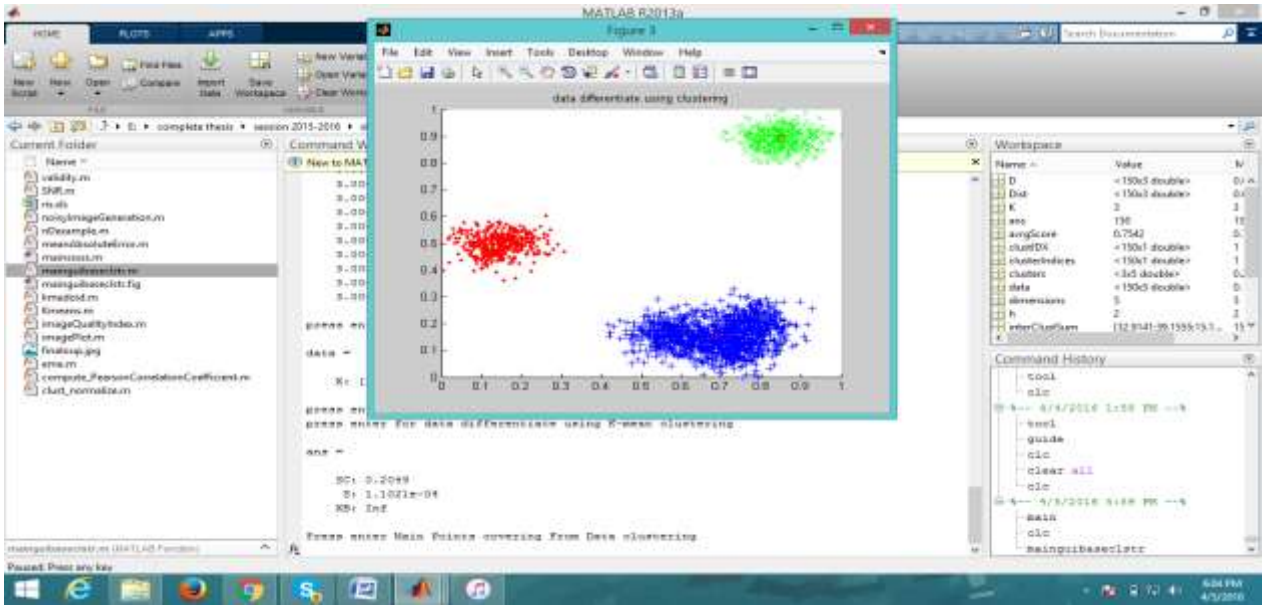

Figure 18: Data Clustered using K-mean clustering :- The plotted data will be completely clustered using the algorithm of k-mean clustering. The central points are marked in each cluster.

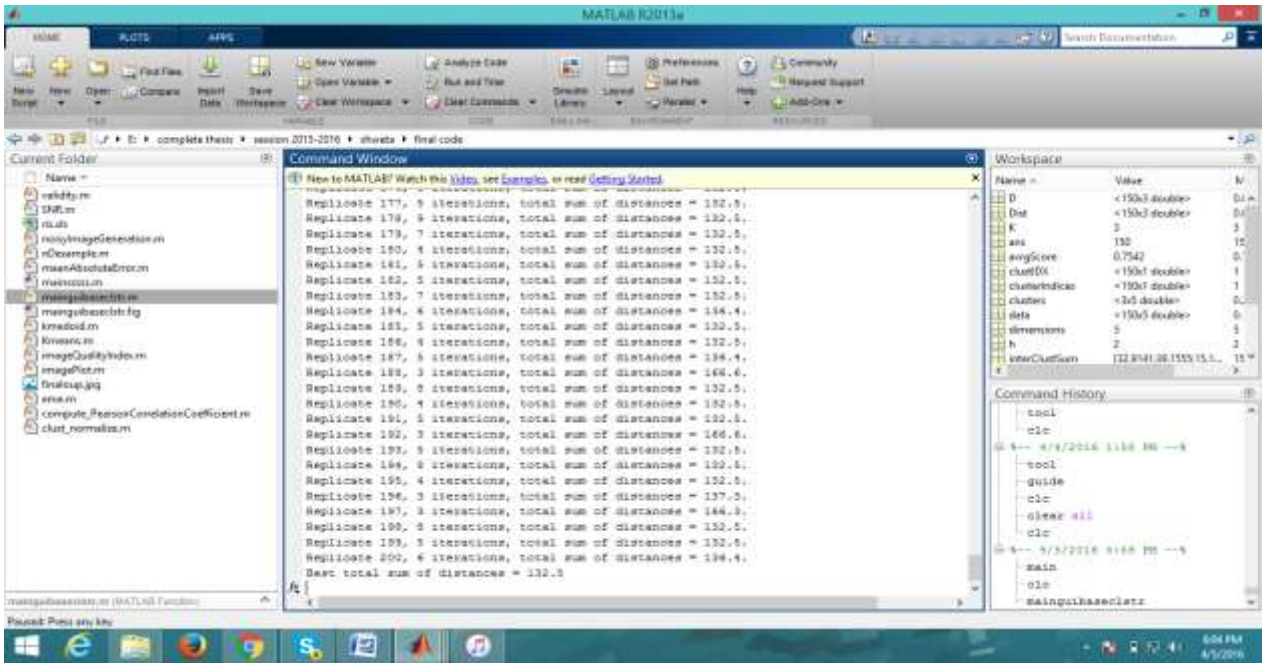

Figure 19: Clustered data plotted :- The clustered data will be plotted in the command window. The central points are marked in each cluster. The normalization technique is applied to calculate best distance to make high quality clusters.

4.2 COMPARISON WITH EXISTING TECHNIQUE

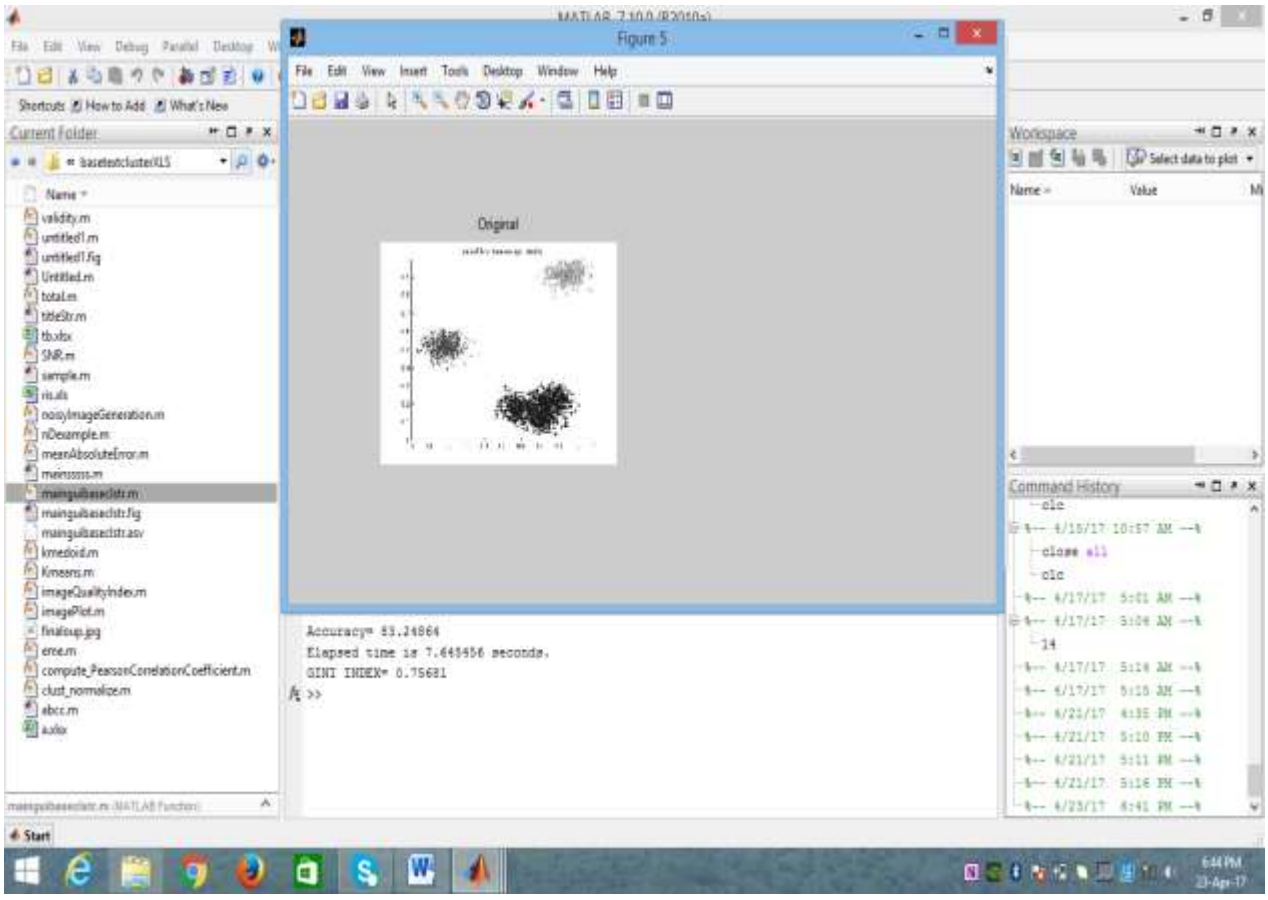

Figure 20: Accuracy & Execute Time of Base paper

As shown in figure, the k-mean clustering is improvement to improve cluster quality using the technique of normalization. The plotted data will be clustered using the algorithm of k-mean clustering. The parameter values which are obtained due to existing techniques are mentioned below:

Accuracy = 83.24864

Time = 7.645456

GINI Index = 0.75681

The Second Phase is Classification. The clustered data is classified by applying classification technique. The performance of data classification depends upon the accuracy of Clustering

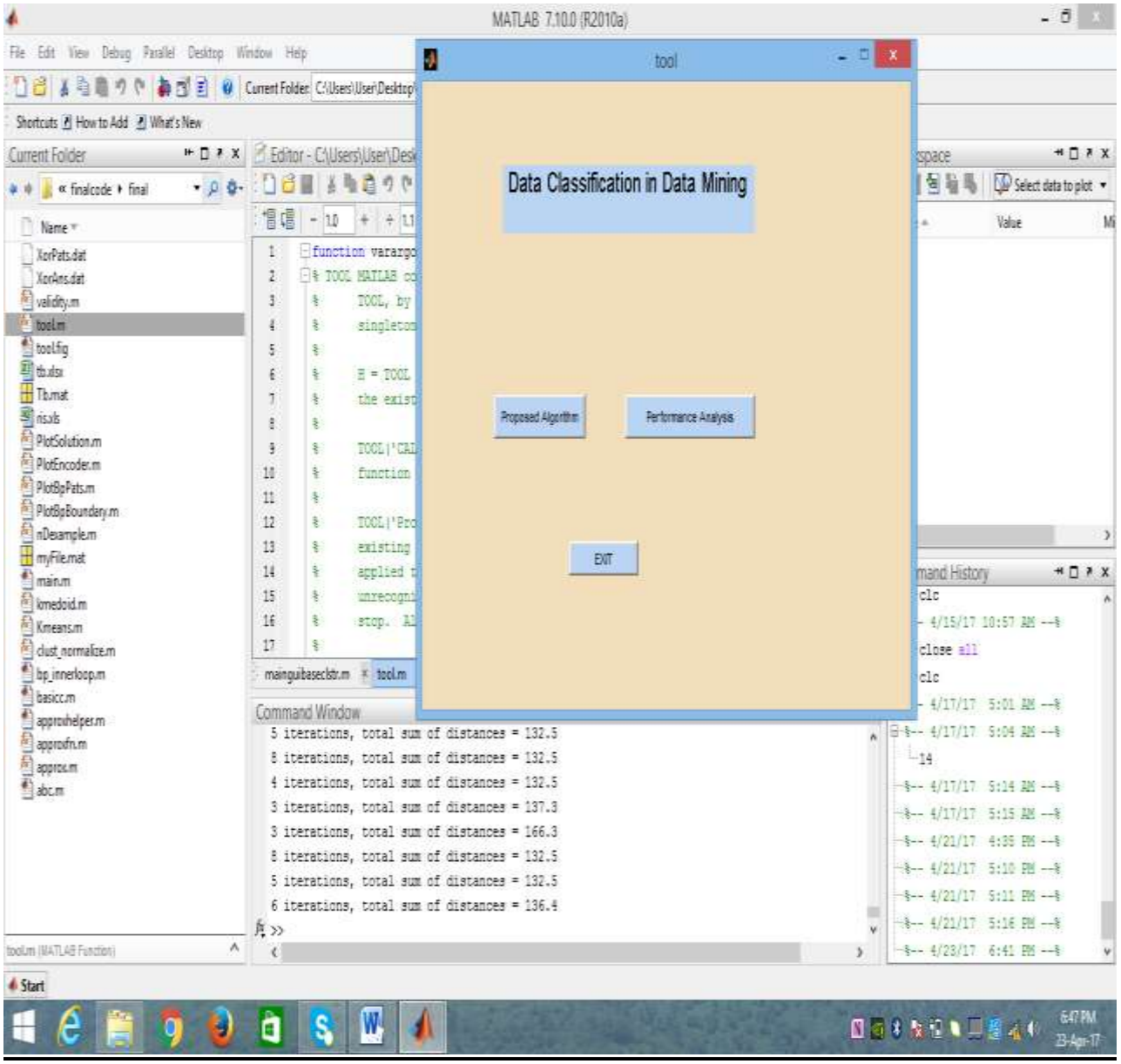

Figure 21: Classification performed on clustered data :- The clustered data is classified by applying classification technique. The performance of data classification depends upon the accuracy of Clustering

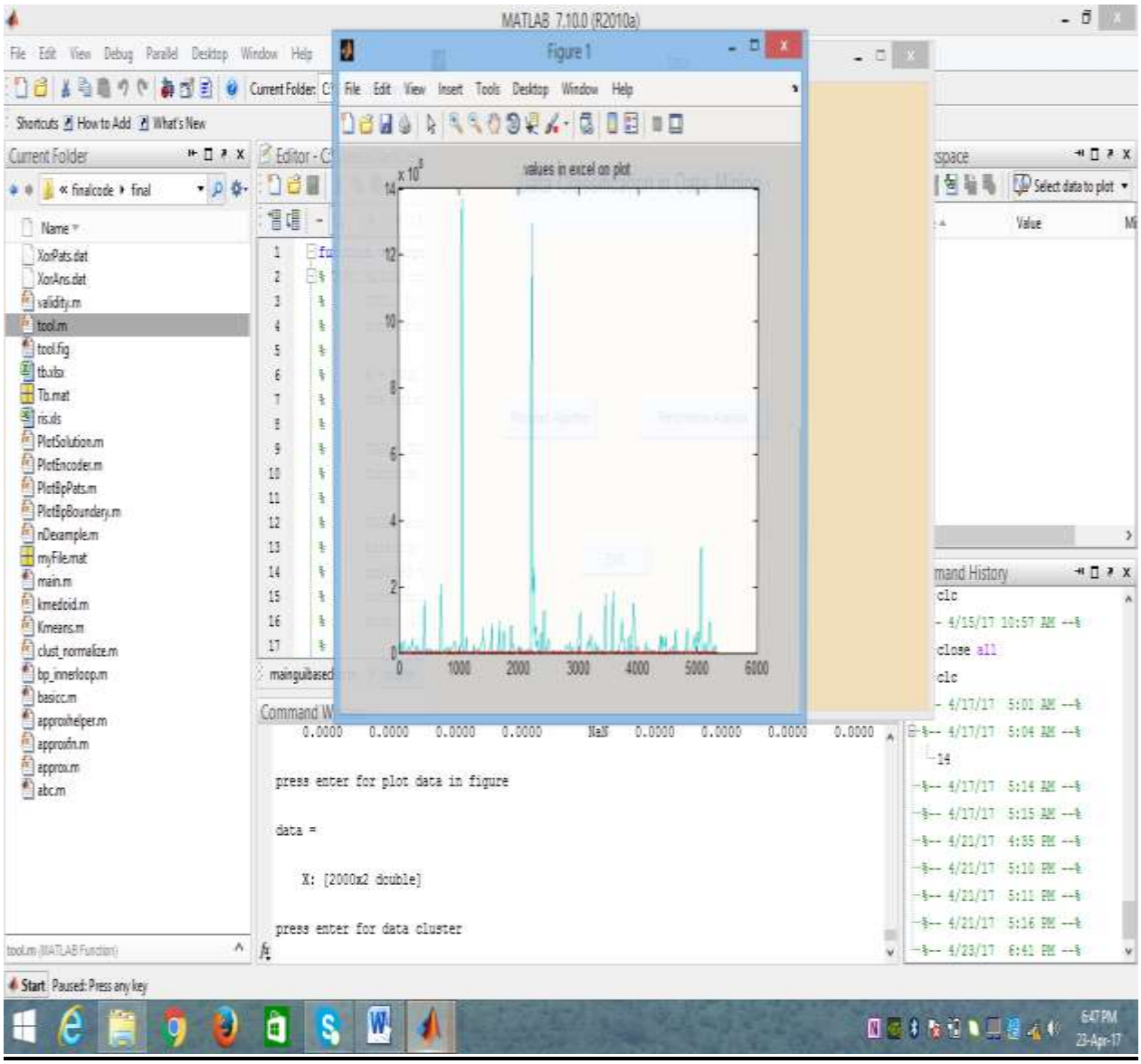

Figure 22: Classified Clustered Dataset Plotting

Classification is used to classify the Clustered data**.** The dataset is loaded and it is shown on the command window. The dataset which is loaded is plotted on the 2-D plane for analysis. The dataset values of classified Clusters are plotted in excel.

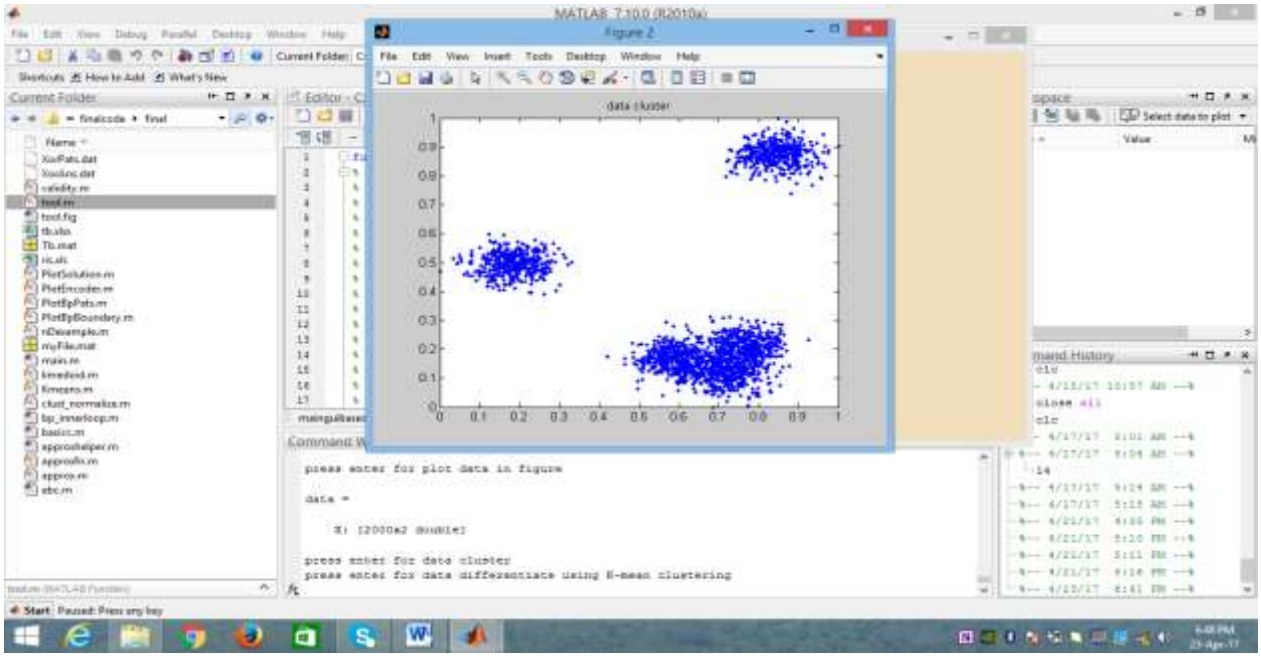

Figure 23: Classification technique applied on clustered data :- In this figure**,** clustered data is classified by classification techniques.

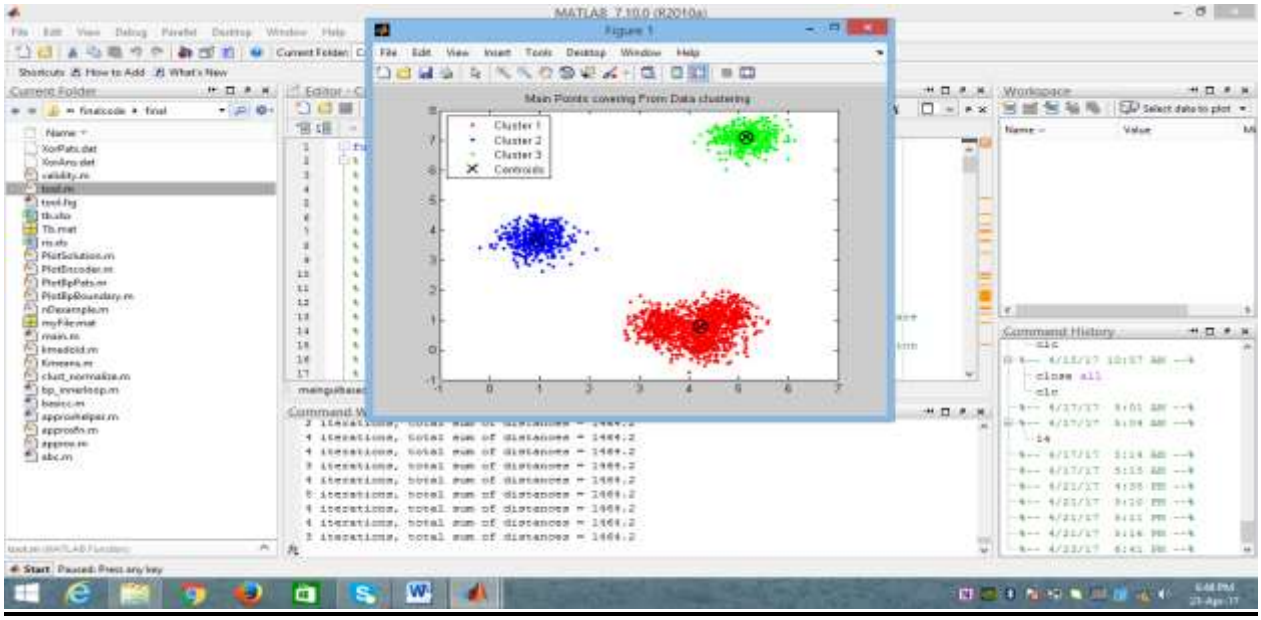

Figure 24: Main points covering Data clustering :- As shown in the figure, three clusters are formed and in each cluster a centroid is marked. Classification techniques are then applied to classify the clusters.

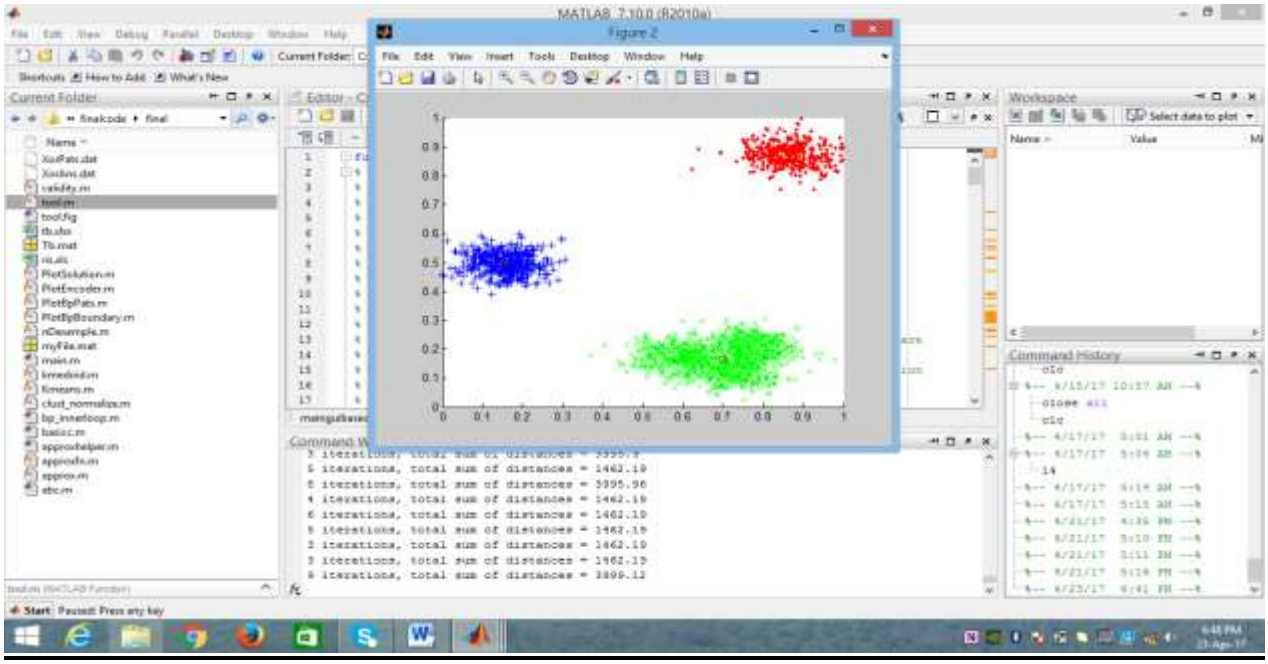

Figure 25: Data Clustered completely by Classification techniques :- This directly leads to improve clustering and classification accuracy and reduce execution time.

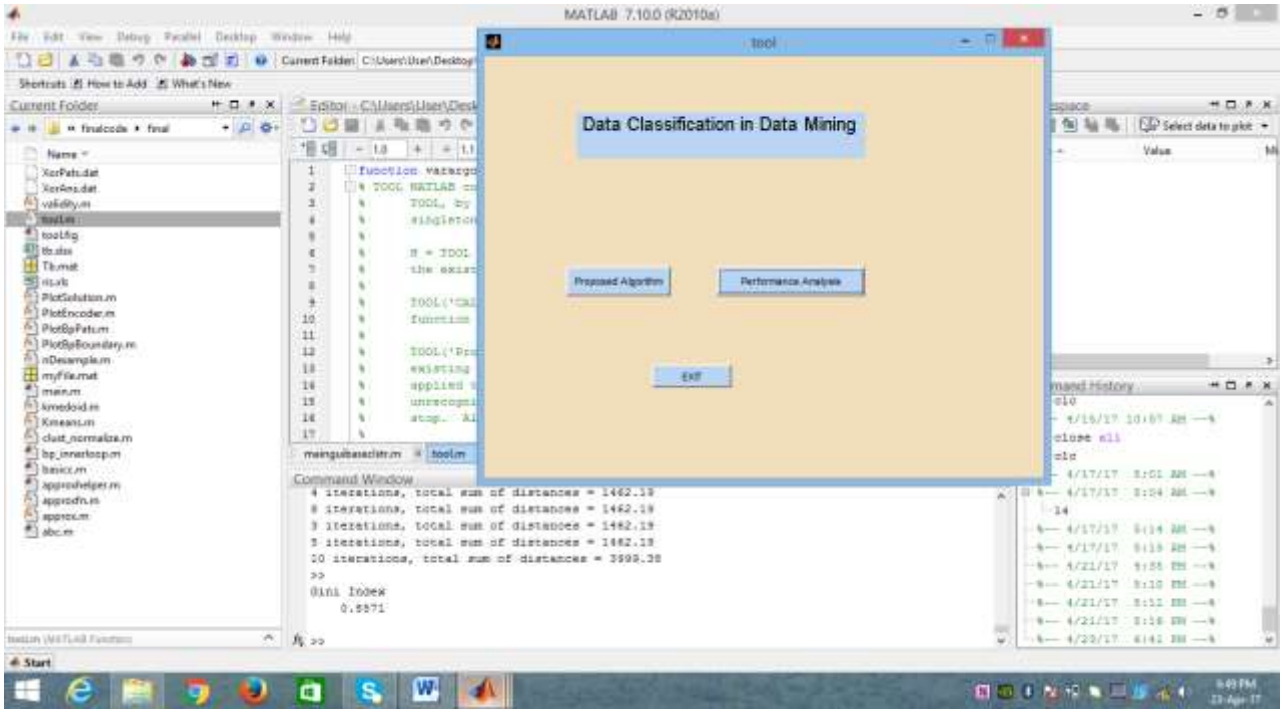

Figure 26: Performance Analysis :- The clustered data is classified by applying classification technique. The performance of data classification depends upon the accuracy of clustering.

4.3 IMPROVEMENT IN RESULT

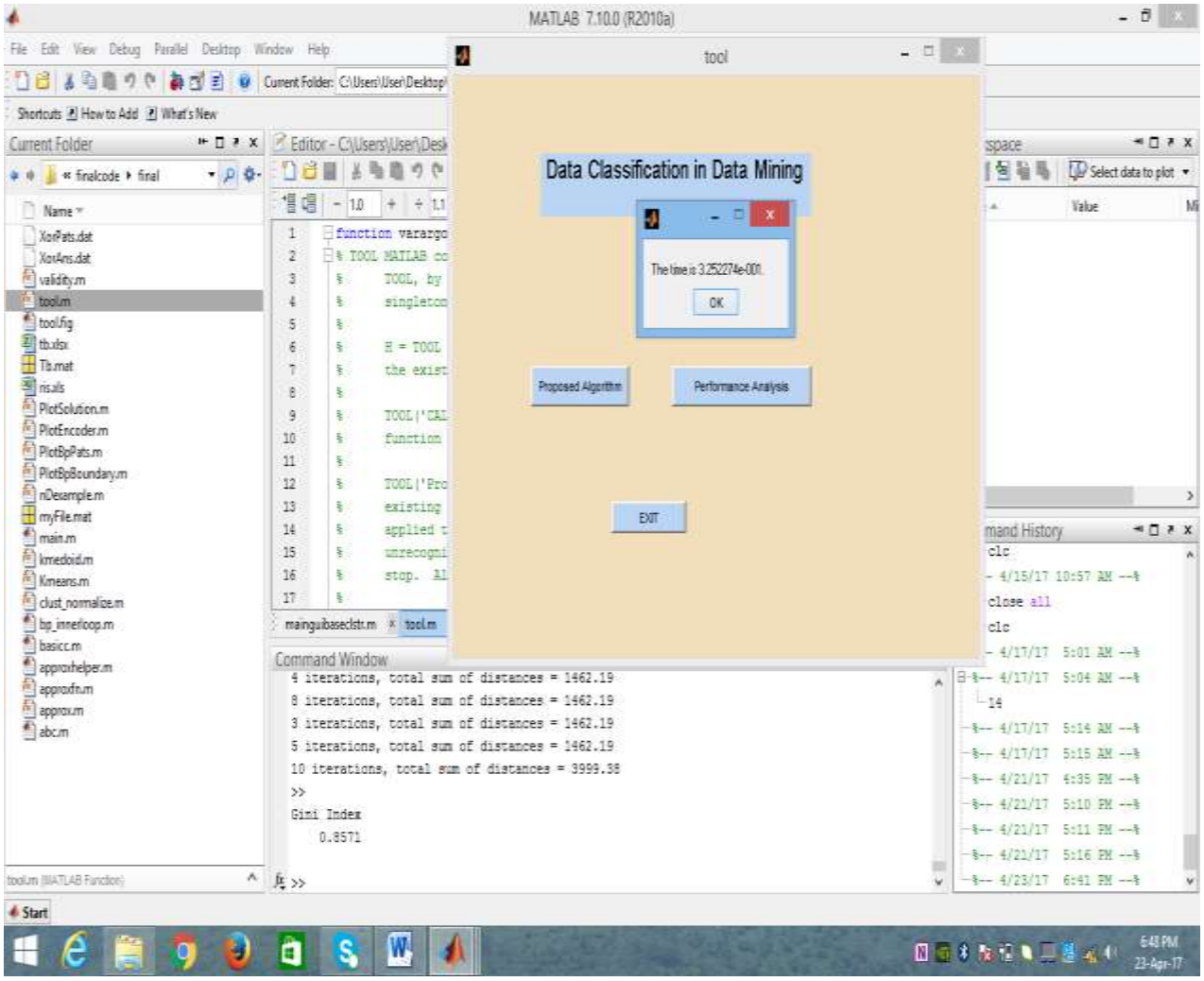

Figure 27: Execution time :- The clustered data is classified by applying classification technique. The performance of data classification depends upon the accuracy of clustering

In the performance Analysis the Execution time = 3.252274 which very less as compared to the previous parameter values and GINI index = 0.8571 which gives high accuracy level as GINI index is also high.

Execution time = 3.252274

GINI index = 0.8571

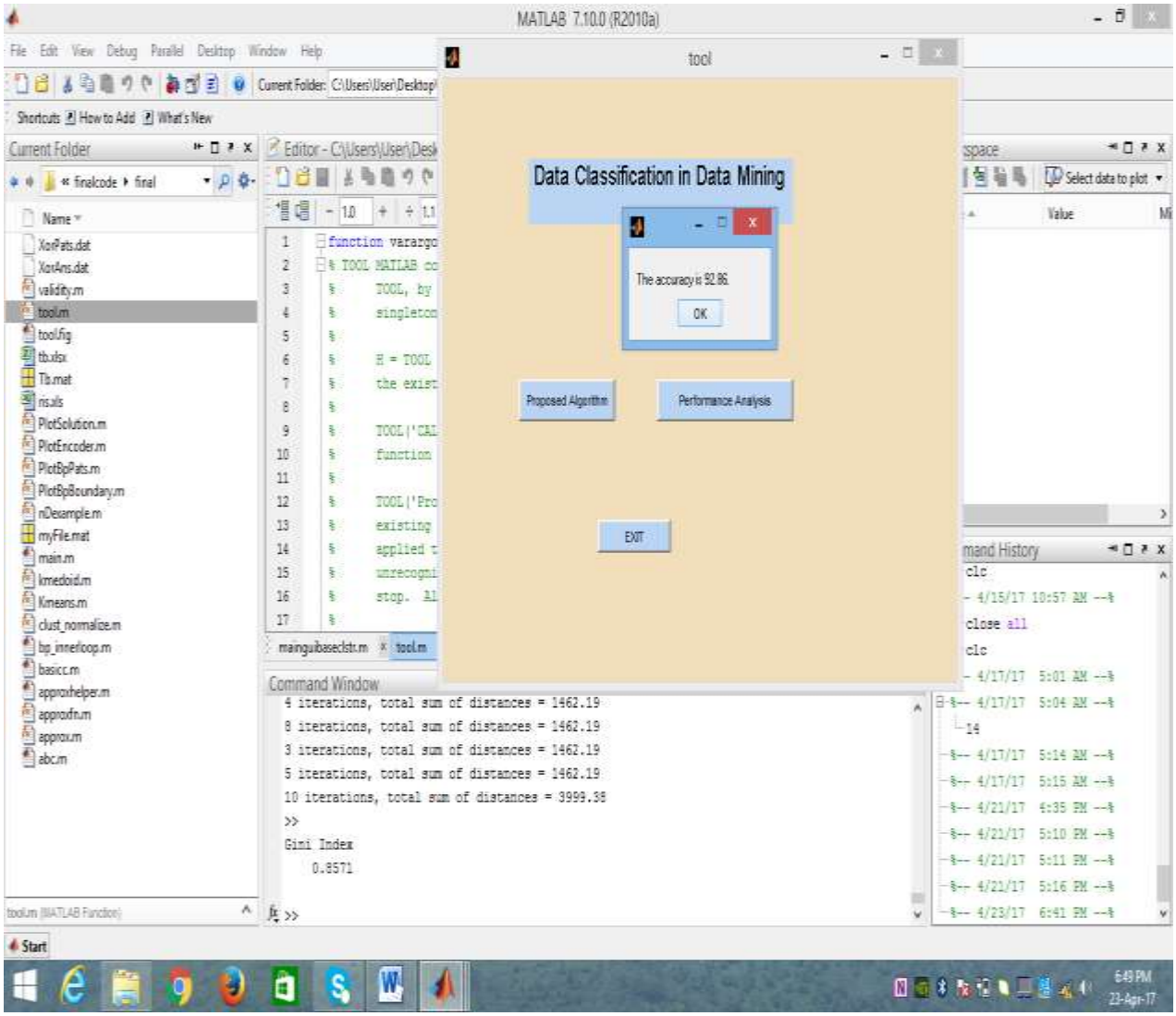

Figure 28: Accuracy Level :- The clustered data is classified by applying classification technique. The performance of data classification depends upon the accuracy of clustering. As the GINI index is very high, therefore the accuracy level will also be high along with GINI index.

The Improvement is shown in the result by comparing this accuracy level, GINI index, execution time as well with the previously obtained parameter values.

Execution time = 3.252274

GINI index = 0.8571

Accuracy Level = 92.86

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

5.1 CONCLUSION

The motive of this paper is to detect the heart diseases by using various data mining techniques by performing prediction analysis. These techniques help in digging meaningful information from large amount of data. In this paper K-mean clustering technique is utilized for the given datasets to calculate the arthematic mean of input dataset from the center point.

5.2 FUTURE SCOPE

The Future scope of this study is as follows

- As the dataset becomes complex, k mean algorithm can help in deriving a relationship between the attributes of the dataset to predict the heart diseases accurately in health care organizations.

- These attributes of the dataset can be reduced in future.

- Accuracy level can be increased by using some of the algorithms of data mining.

- Execution time is also reduced which helps the individual to save their time.

REFERENCES

[1] Andritsos, P., Tsaparas, P., Miller, R.J. and Sevcik, K.C., "Limbo: Scalable Clustering of Categorical Data", Extending Database Technology (EDBT), 2004.

[2] Han, J. and Kamber, M., "Data Mining Concepts and Techniques", Morgan Kaufmann, 2001.

[3] Anil, K. Jain, and Richard, C. Dubes., "Algorithms for Clustering Data", Prentice-Hall International, 1988.

[4] Anjan Goswami, Ruoming Jin, and Gagan Agrawal, "Fast And Exact K-Means Clustering", International Conference on Data Mining (ICDM), 2004.

[5] Huang, Z., "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values", Data Mining and Knowledge Discovery, 1998.

[6] Hartigan, J. A. and Wong, M.A., "A k-means clustering algorithm", Applied Statistics, pp. 100–108, 1979.

[7] Subramanian, D.K., Narasimha Murthy, M., Vijaya, P.A., "Leaders-Subleaders: An Efficient hierarchical clustering algorithm for large data sets", ELSEVIER, PP. 505–513, 2004.

[8] Danyang Cao, Bingru Yang, "An improved k-medoids clustering algorithm", Computer and Automation Engineering (ICCAE), 2010

[9] Chen, H.L., Chuang, K.T and Chen, M.S., "Labeling Un clustered Categorical Data into Clusters Based on the Important Attribute Values", IEEE International Conference on Data Mining (ICDM), 2005.

[10] J.Zhu, H.Wang, M.Zhu, B.K.Tsou, and M.Ma,," Aspect-Based Opinion Polling from Customer Reviews," T. Affective Computing2(1):pp. 37- 49, 2011.

[11] M.Karamibekr,A.A.Ghorbani,"Verb Oriented Sentiment Classification," Processed of the IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT), Vol (1): pp.327-331, 2012.

[12] Yun Chi, Haixun Wang, Philip S. Yu, Richard R. Muntz, "Moment: Maintaining Closed Frequent Itemsets over a Stream Sliding Window," 2004, icdm, pp.59-66, Fourth IEEE International Conference on Data Mining (ICDM'04)

[13] Han, J., Pei, J., and Yin Y.," Mining frequent patterns without candidate generation", 2000, International Conference on Management of Data, 1-12

[14] R. C. Agarwal, C. C. Aggarwal, and V. V. V. Prasad," A tree projection algorithm for generation of frequent item sets", 2001, Journal of Parallel and Distributed Computing, 61(3):350–371

[15] Hand D. J.," Statistics and Data Mining: Intersecting Disciplines", 1999, ACM SIGKDD Explorations, 1, 1, pp. 16- 19

[16] Burl M., Fowlkes C., Roden J., Stechert A., and Mukhtar S.," Diamond Eye: A distributed architecture for image data mining", 1999, SPIE DMKD, Orlando

[17] Kargupta, H., Park, B., Pittie, S., Liu, L., Kushraj, D. and Sarkar, K.," MobiMine: Monitoring the Stock Market from a PDA", 2002, ACM SIGKDD Explorations, Volume 3, Issue 2. Pages 37-46

[18] Kargupta H., Bhargava R., Liu K., Powers M., Blair S., Bushra S., Dull J., Sarkar K., Klein M., Vasa M., and Handy D.," VEDAS: A Mobile and Distributed Data Stream Mining System for Real-Time Vehicle Monitoring", 2004, Proceedings of SIAMInternational Conference on Data Mining

[19] Li Su, Hong-yan Liu, Zhen-Hui," Song: A New Classification Algorithm for Data Streams", 2011, I.J. Modern Education and Computer Science 4, 32-39

[20] Koh, J.-L., and Shieh, S.-F," An efficient approach for maintaining association rules based on adjusting FP-tree structures", 2004, ACm, research proceedings, pp. 53-123

[21] Ranganatha S., Pooja Raj H.R., Anusha C., Vinay S.K.," MEDICAL DATA MINING AND ANALYSIS FOR HEART DISEASE DATASET USING CLASSIFICATION TECHNIQUES", 2013, IEEE, 32987u-45096-45

[22] Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg," Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors", 2013 IEEE Conference on Information and Communication Technologies (ICT) 978-1-4673-5758

[23] Theresa Princy, R, J. Thomas," Human Heart Disease Prediction System using Data Mining Techniques", 2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT], 978-1-5090-1277-0

[24] K.Srinivas, Dr.G.Raghavendra Rao, Dr. A.Govardhan," Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques", 2010, IEEE, 978-1-4244-6005-2

[25] R.Kavitha, E.Kannan," An Efficient Framework for Heart Disease Classification using Feature Extraction and Feature Selection Technique in Data Mining", 2016, IEEE, 29753-48506-65-7

[26] Eman AbuKhousa, Piers Campbell," Predictive Data Mining to Support Clinical Decisions: An Overview of Heart Disease Prediction Systems", 2012 International Conference on Innovations in Information Technology (IIT), 978-1-4673-1101-4

[27] Monika Gandhi, Dr. Shailendra Narayan Singh," Predictions in Heart Disease Using Techniques of Data Mining", 2015 1st International Conference on Futuristic trend in Computational Analysis and Knowledge Management (ABLAZE)

[28] Heon Gyu Lee, Ki Yong Noh, and Keun Ho Ryu," A Data Mining Approach for Coronary Heart Disease Prediction using HRV Features and Carotid Arterial Wall Thickness", 2008 International Conference on BioMedical Engineering and Informatics, 978-0-7695-3118-2

[29] Lamia AbedNoor Muhammed," Using Data Mining technique to diagnosis heart disease", 2012, IEEE, 3545-56445-4

[30] Sellappan Palaniappan, Rafiah Awang," Intelligent Heart Disease Prediction System Using Data Mining Techniques", 2008, IEEE, 978-1-4244-1968-5

[31] Mai Shouman, Tim Turner, Rob Stocker," USING DATA MINING TECHNIQUES IN HEART DISEASE DIAGNOSIS AND TREATMENT", 2012, IEEE, 978-1-4673-0484-9

[32] Sivagowry .S, Dr. Durairaj. M and Persia.A," An Empirical Study on applying Data Mining Techniques for the Analysis and Prediction of Heart Disease", 2013, IEEE, 23543-56-465

[33] K. Prasanna Lakshmi, Dr. C.R.K.Reddy," Fast Rule-Based Heart Disease Prediction using Associative Classification Mining", 2015, IEEE International Conference on Computer, Communication and Control (IC)

[34] Elena Baralis, Silvia Chiusano, and Paolo Garza," A Lazy Approach to Associative Classification", 2008, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Vol. 20, No. 2

[35] K. Prasanna Lakshmi, Dr. C. R. K. Reddy," Compact Tree for Associative Classification of Data Stream Mining", 2012, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 2

[36] Bing Liu, Wynne Hsu, Yiming Ma," Integrating Classification and Association Rule Mining", 2012, Research gate, AAAI organization, pp. 345-753.