CLASSIFICATION AND PREDICTION IN BREAST CANCER

Dissertation submitted in fulfilment of the requirements for the Degree of

MASTER OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

By

Liza Dhingra

11201766

Supervisor

Anita Sharma

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

May 2017

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 cancer. It is expected that by 2030, around 25 million people may die because of cancer. Though many researchers have suggested and proposed methods for diagnosing cancer from the enormous amount of cancer 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 cancer detection and prediction.

Keywords- Data mining, Classification, Clustering, Regression, Breast Cancer, 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 "CLASSIFICATION AND PREDICTION IN BREAST CANCER" in partial fulfilment 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 Ms. Anita Sharma 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

Liza Dhingra

11201766

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.TECH Dissertation entitled "**Classification and Prediction in Breast Cancer"**, submitted by **Liza Dhingra** 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

(Anita Sharma)

Date: 27/04/2017

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 Anita Sharma 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,

Liza Dhingra

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Data mining

There are large amount of databases available with the increase in the Information Technology. Data is present in huge amount which comprises of various fields. For the purpose of future decision makings, the data needs to be stored and manipulated. For this, various databases have been developed and researches have been carried out for their managements. The procedure of extraction of valuable data and patterns from vast measure of stored information is known as information mining. There are other names for this process as well, such as knowledge discovery process, knowledge mining from information, knowledge extraction pattern analysis. Various types of data is analyzed with the help of certain data mining tools. There are certain applications such as the customer retention, education system, production control, healthcare, manufacturing engineering, decision making, and so on related to this technique [1].

There are various functionalities such as information collection and information creation, data management and advanced data analysis which are evolved through database as well as data management. The development of information collection and information creation mechanisms has been providing the later development of new methods which can help in information storage and retrieval which also include inquiry and exchange preparing. Data warehousing is proper evolving data repository architecture. Multiple heterogeneous data sources are organized for the purpose of defining unified schemes to provide single site which will help in facilitating the management for decision making. There are various other steps involved in data warehousing which are data cleaning, data integration as well as online analytical processing (OLAP). There are other functionalities, for example, synopsis, union and total. It likewise includes the usefulness of review data from different edges. It is a very challenging task to analyze the data effectively and efficiently which is of various types and is integrated using data retrieval, data mining and data analysis technologies. There are various interesting patterns through

which the huge data can be stored in an efficient manner within the databases, data warehouses and other repositories. The famous techniques are known as the knowledge discovery in databases (KDD). The integration of techniques is done from different aspects such as statistics, database technology, machine learning, neural networks, high performance computing, pattern matching, and information recovery and so on.

The useful information is fetched from the database, data warehouse, information repository and the servers. The fetching of data is based on the client's data mining request. The search is conducted with the help of knowledge gathered from parts and further calculations are done on the basis of interesting patterns. The information mining modules interact with the pattern evaluation which can help in focusing the depiction of interesting patterns and graphical user interface. The client as well as data mining systems allows the user interface which interacts between the client and the data mining systems for providing a user interface within the system. Following are the categories for the data mining tasks [2]:

1. **Class description**: The individual classes and concepts in summarized, concise and precise terms are done with the help of description of the class.

2. **Association analysis:** The revelation of affiliation principles demonstrating quality esteem conditions which happen as of late together which the given arrangement of information.

3. **Classification:** The strategy of decision the arrangement of models which clarify and separate the information classes, their ideas, for which the reason of having the capacity to utilize the model which can gauge the class of objects of which the class names is not known is known as the grouping procedure. The examination of set of preparing information which can speak to in types of tenets and directions, choice trees, division standards, et cetera.

4. **Cluster analysis**: The information objects are analyzed without consulting the known class in the clustering process. The class labels are not present in which the training data due to the reason that the initial point in not known. On the basis of rule

of expanding the intra-class closeness and diminishing the between class likeness the articles are gathered and bunched.

5. **Outlier analysis**: The broad-spectrum behavior of model of data is not confirmed by the outliers of data objects. The statistical tests of the distance measures can be utilized for detecting the outliers.

6. **Evolution analysis**: The objects whose behaviors are changed after some duration are described and their trends are modeled. There are various things included in it which involve time-series data analysis, succession also known as periodicity pattern matching as well as the similarity-based data analysis [3].

The large amount of data which needs certain powerful data analysis tools are thus put for the here which is also known as the data rich but information poor condition. There is an increase in the growth of data, its gathering as well as storing it in huge databases. It is no more in the hands of humans to do it easily or without the help of analysis tools. There are certain data archives created here which can be visited when the data is required. The insightful, interesting and novel patterns of data are discovered from large-scale data sets using the data mining. The knowledge discovery in databases process is a very important step in data mining. The data mining and KDD are often termed as synonyms. There are databases, data warehouses, internet, information repositories involved within the data sources. There are various fields in computer science such as databases, machine learning, pattern recognition, statistics, artificial intelligence, uncertainty, expert systems, information retrieval, statistics, computing and networking in which the KDD processes are involved. The end goal here is to extract information from the information set and transfer it into an understandable structure which can be helpful in further use which can help provide the data mining process to evolve. Any kind of data repository can be presented in the through this technique. There are various types of algorithms and techniques which are utilized for various typed of data. There are different databases in which the data mining can be used. The object-relational databases, relational database, data warehouses and multimedia data bases and so on which can be involved here.

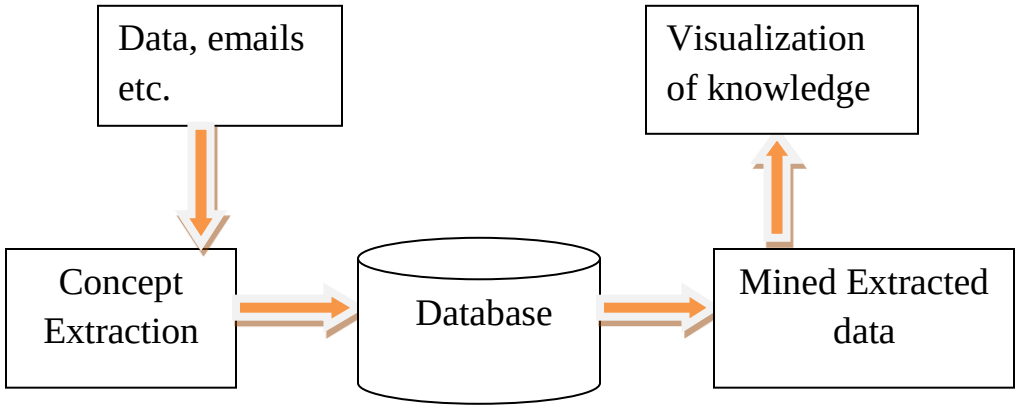


Figure1: Data mining concept

There are special kinds of functionalities within the data mining. These are utilized for specifying certain kind of patterns which can help in identifying various tasks of the data mining process. There are two categories in which the data mining tasks can be classified. They are the descriptive as well as the predictive. The tasks which characterize the general purpose properties of the data within the database are known as the descriptive mining tasks. The tasks which perform the inference on the present data for making predictions are known as the predictive mining tasks. The various data mining functionalities are given below [4]:

1. **Characterization and Discrimination:** Data characterization is a summarization of data of the class under study and data discrimination is a comparison of the target class with one or a set of comparative classes. Class/concept descriptions are derived using these two functionalities.

2. **The Mining of Frequent Patterns, Associations and Correlations:-** Frequent examples are the examples that happen as often as possible in information. Affiliation control mining is the way toward finding fascinating connections, visit examples or relationship among sets of things in the exchange databases, social databases or other data storehouses.

3. **Classification and Regression:-** Classification is an information mining (machine learning) method used to anticipate assemble enrolment for information

occasions. Regression is a technique that is essentially utilized for numeric expectation.

4. **Cluster Analysis:**-Cluster analysis is the assignment of collection an arrangement of items such that articles in a similar gathering (called a bunch) are more like each other than to those in different gatherings (groups).

5. **Outlier Analysis:**- A few questions in an informational index don't agree to the general conduct or model of the information. These information items are anomalies and examination of exception information is known as exception investigation.

1.2 Knowledge Discovery from Data (KDD) Process

There are certain steps which are to be followed for the KDD process. The term KDD mainly refers to the process of extracting knowledge from the data present and further emphasizing it on high-level applications for achieving particular data mining methods. There are various fields in which the data mining is very essential. The most important role for presenting the frequently utilized object sets is utilized with the help of determining the correlations between the various types of fields which are present in the data base. The association rule is another important factor which can be utilized here for the identification of frequently used object sets in KDD process. The retail stores utilize this association rule concept for the purpose of managing the marketing, advertising and handling problems which are present in this field [5].

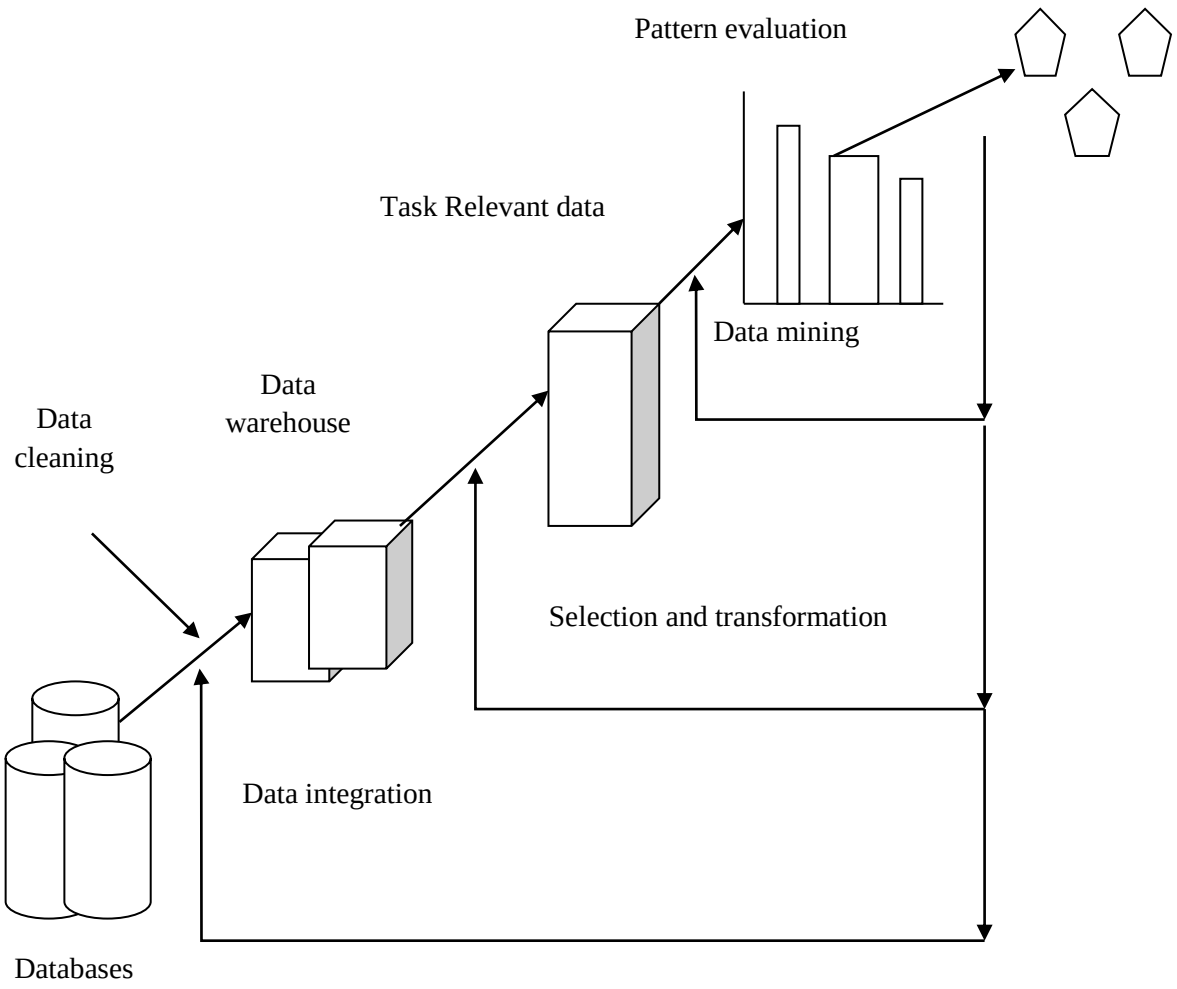


Figure2: Data Mining in an Organization

As it is already known that there is a huge growth in the information technology on daily basis and the databases are being created by organizations for managing the data. There are organizations which belong to the telecommunications, banking, marketing, transportation, and so on. It is vital to explore the complete databases efficiently for defining the valuable data. For identifying the information in huge databases, the data mining method is used. The data which shows well defined relationship between the variables is created using the KDD process. The useful patterns are discovered from the database with the help of automatic discovery process which is mainly the KDD method. It is also known as discovering required data from huge data bases. The association rule mining is one of the important rules which are developed for data mining. For the purpose of decision making, these rules are proposed on applications such as market based, banking, and so on.

The main objective of the KDD process is to gain the knowledge from the information from within the huge databases. The various steps involved in this process are [6]:

1. **Data Cleaning:** The noise and irrelevant data is removed here in the initial step.
2. **Data Integration:** The data from many data sources is then combined.
3. **Data Selection:** Further, the information which is relevant to the analysis task is gained from the database.
4. **Data Transformation:** The total or outline operations are performed to change and merge information into structures proper for mining.
5. **Data Mining:** This progression is basic in which the information examples are extricated by applying shrewd techniques.
6. **Pattern Evaluation:** The genuinely intriguing examples are recognized for speaking to learning in view of intriguing quality measures.
7. **Knowledge Presentation:** In the last stride the perception and learning portrayal methods are utilized to show mined information to clients.

1.3 Data Mining Methods

The two abnormal state essential objectives of information mining by and by tend to be expectation and portrayal. As communicated some time recently, forecast includes using a couple of factors or fields as a piece of the database to foresee obscure or future estimations of various factors of intrigue, and depiction concentrates on finding human-interpretable examples portraying the information. Notwithstanding the way that the limits among expectation and depiction are not sharp (a bit of the prescient models can be unmistakable, to the extent that they are justifiable, and the a different way), is significant for understanding the general disclosure objective. The relative significance of forecast and depiction for specific information mining applications can contrast extensively. The objectives of expectation and depiction can be refined using an assortment of specific information mining strategies [7].

a. Classification: Classification is taking in a capacity that maps (groups) an information thing into one of a few predefined classes. Cases of characterization strategies used as a component of learning disclosure applications consolidate the arranging of patterns in money related markets and the robotized distinguishing proof of objects of enthusiasm for immense picture databases. It is illogical to isolate the classes sublimely using a direct choice limit. The bank may need to use the arrangement areas to naturally pick whether future advance candidates will be given an advance or not.

b. Regression: Regression is taking in a capacity that maps an information thing to a genuine esteemed expectation variable. Relapse applications are numerous, for instance, anticipating the measure of biomass present in a woodlands given remotely detected microwave estimations, evaluating the likelihood that a patient will survive given the aftereffects of an arrangement of symptomatic tests, foreseeing customer interest for another item as a component of publicizing consumption, and foreseeing time arrangement where the information factors can be time-slacked variants of the expectation variable.

c. Clustering: Bunching is a typical illustrative errand where one tries to distinguish a limited arrangement of classifications or groups to depict the information. The classes can be totally unrelated and comprehensive or include a wealthier portrayal, for instance, various leveled or covering classifications. Cases of bunching applications in a learning revelation setting consolidate finding homogeneous subpopulations for purchasers in showcasing databases and distinguishing subcategories of spectra from infrared sky estimations. The groups cover, permitting information focuses to have a place with more than one bunch. Immovably identified with bunching is the undertaking of likelihood thickness estimation, which includes systems for assessing from information the joint multivariate likelihood thickness capacity of the extensive number of factors or fields in the database [8].

d. Summarization: Synopsis includes strategies for finding a reduced depiction for a subset of information. A straightforward illustration would group the mean and standard deviations for all fields. More complex strategies include the deduction of rundown standards, multivariate perception systems, and the disclosure of utilitarian connections

between factors. Plot procedures are every now and again associated with intuitive exploratory information examination and robotized report era.

e. Dependency modeling: Dependency modelling comprises of finding a model that portrays huge conditions between factors. Reliance models exist at two levels:

Dependency modelling contains finding a model that depicts noteworthy conditions between factors. Dependency models exist at two levels:

(1) The basic level of the model demonstrates (often in realistic casing) which factors are locally subject to each other and

(2) The quantitative level of the model decides the qualities of the conditions using some numeric scale.

For instance, probabilistic reliance systems use contingent autonomy to show the auxiliary piece of the model and probabilities or connections to decide the qualities of the conditions. Probabilistic reliance systems are progressively finding applications in extents as various as the headway of probabilistic therapeutic master frameworks from databases, data recovery, and demonstrating of the human genome.

f. Change and deviation detection: Change and deviation detection focuses on discovering the most significant changes in the data from beforehand measured or normative values [9].

1.4 Clustering in Data Mining

There are various applications which involve the usage of clustering analysis in them. Applications such as market research, pattern recognition, data analysis and image processing are some of them. On the basis of purchasing patterns and characterizing the groups of customers, the clustering can provide help to the marketers as per the interest of the customers in the business fields. For the purpose of making plant and animal taxonomies, the cluster analysis is used in the fields of biology. In helps to categorize the genes which are of similar functionality and provide outlook to the structures which have huge populations. The clustering can be employed by the specialists for identifying the

lands which have similarities, similar houses within a city and other properties which come used the geology filed. For the purpose of information discovery, the data clustering can provide help in documents classification on the Internet [10].

The data clustering is an unsupervised classification method. Its main objective is to create group of objects or clusters in such a manner that the objects which have similar properties can be grouped together. Here the distinct objects are thus present in different groups as per their properties. Within the data mining research area, cluster analysis a very old and efficient area for study. For the purpose of knowledge discovery, this step is the starting point in this direction. The data objects are grouped within a set of disjoint classes called clusters using the clustering method. There is a higher resemblance of objects which are present within a same class as compared to the two objects which belong to separate classes.

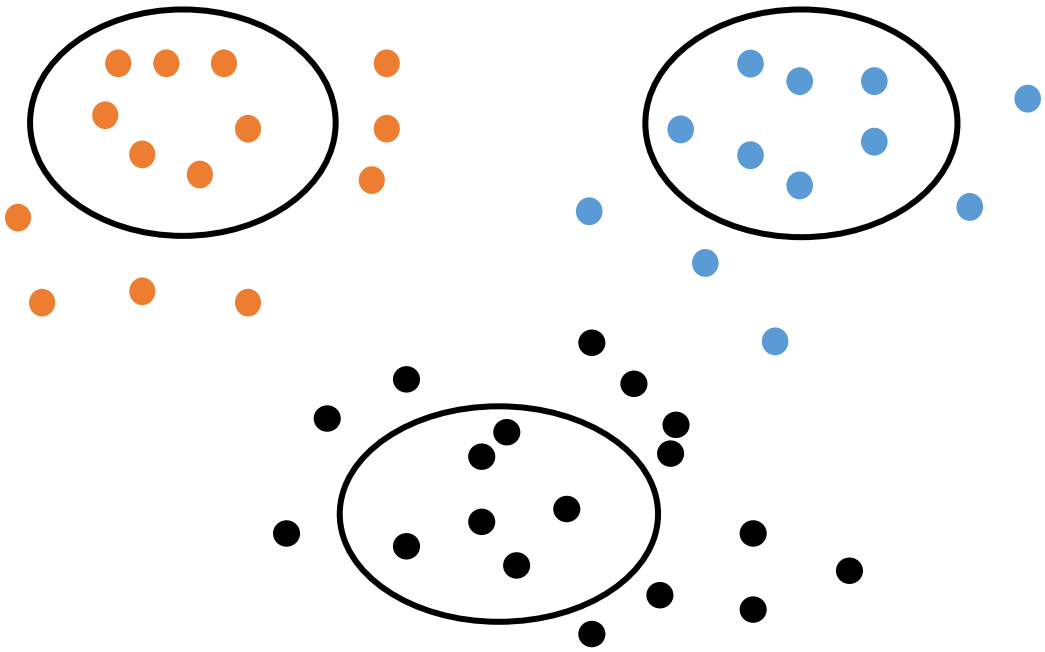


Figure3: Clustering

The similar objects are grouped together in the forms of clusters. This method is known as the clustering technique. The documents however are categorized on various basis and the predefined methods are not utilized. There can be various subtopics in which the

objects can appear which helps in preserving them at some place within the database so that they can be retrieved when urgently required or lost. For every document, the various topics are formed as a vector and the weights are measured which determine how healthy the document is as per the cluster. The clustering mechanism is involved in the unsupervised classification. Classification technique is defined as the method which assigns the data objects into certain set of classes. The clustering is not dependent on the predefined classes as well as the training. The pattern recognition is not similar to the unsupervised clustering. The discriminate analysis and decision analysis are the basis for providing statistics which classify objects from a given set of object.

There are many clustering algorithms used for clustering. The major fundamental clustering methods can be classified into following categories [11]:

1. **Partitioning Methods:** The highly similar samples present within a cluster are combined within this partitioning technique. The different clusters formed have high dissimilarity among themselves. The various partitioning methods are distance-based. In a system, if k is the given number of partitions for construction, the partitioning method helps in creating an initial partitioning. Further, the iterative relocation method is used for improving the partitioning technique with the help of moving the objects from one group to another. The objects which are present in the same cluster are closer and are related to each other in the case of a good partitioning. The objects which are present in different clusters have different properties. There are various greedy techniques such as k-means and k-mediod algorithms which are utilized in applications as some heuristic methods. They are used mainly for improving the quality of clustering and also provide a local optimum. In order to find the spherical-shaped clusters within the small to medium size databases, the clustering techniques are very efficient.

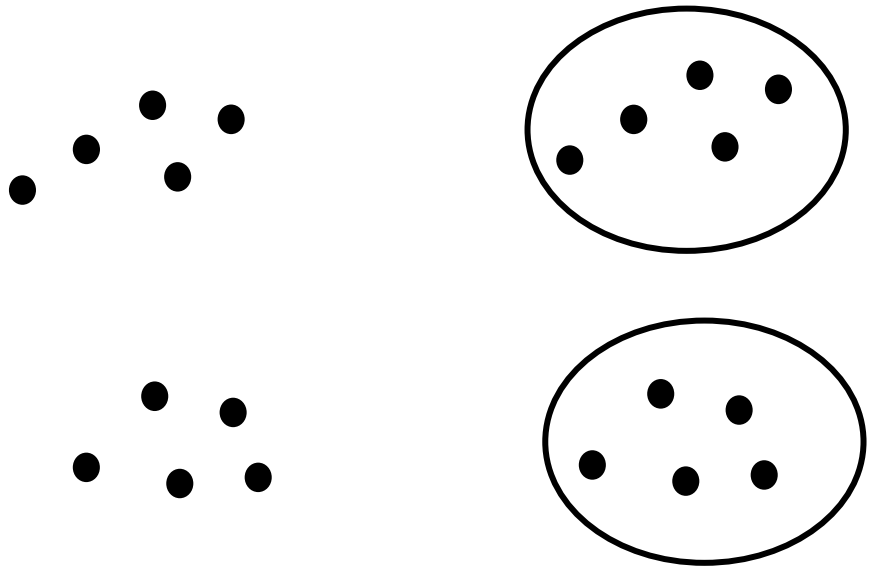


Figure4: Partitioning Clustering

2. **Hierarchical Methods:** The hierarchical decomposition of given set of data objects in known as the hierarchical clustering technique. There are two broader classifications involved here which are the agglomerative and the divisive based methods. The classifications are made on the basis of how the hierarchical decomposition is formed. The bottom-up approach which begins which the each object forming a different group is known as the agglomerative approach. Further, the groups which are close to each other are merged till there is only one group left [12]. Another approach in which the clusters are present within the similar clusters is separated after each iteration step is known as a divisive approach. It is a top-bottom approach. Here the result provides every object to form a different cluster. A hierarchical decomposition of the provided data set of the objects is formed using the hierarchical algorithm. A tree structure also known as a dendogram is used to represent the hierarchical decomposition here. There are no clusters to be provided here as inputs. At various levels of granularities, different partitions are also possible using various types of K.

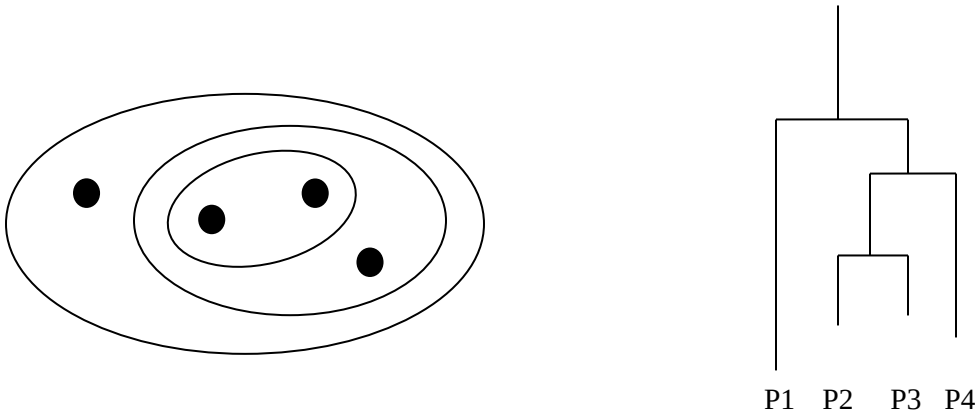


Figure5: Hierarchal Clustering

3. **Density Based Methods:** There are various distances between the objects on the basis of which the partitioning methods are proposed. Using this method the spherical shaped clusters can be discovered. The clusters which are of various arbitrary shapes can be encountered with certain difficulties. Hence, there are certain density-based methods which are utilized for the arbitrary shapes for the notion of density. The clusters keep growing as long as the density of the neighborhood exceeds certain threshold. The notion of density is the base of this method. The provided cluster has to keep growing according to the density of the neighborhood exceeds some threshold. At least six minimum numbers of points are to be provided for the radius of a given cluster for each data point within a given cluster. The arbitrary shaped clusters are discovered using this method. Also the noise present in the data can be removed using this technique within one time scan.. These properties of this method have made it possible to utilize this method in various applications [13].

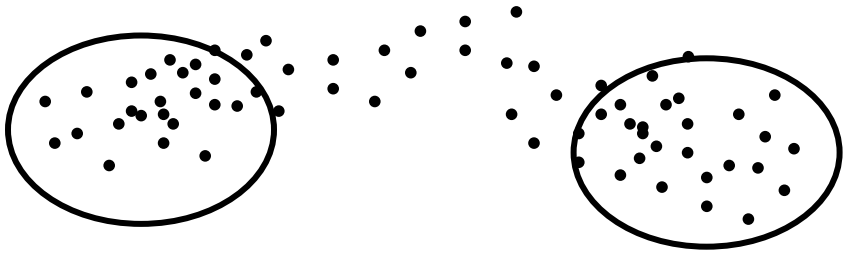


Figure6: Density based clustering

4. **Grid Based Methods:** The object space is quantized into a finite number of cells which create a grid structure. This method is known as the grid based method. This method is really of high speed and does not depend on the number of data objects present. The only factor on which this method is dependent is the number of cells present within each dimension of the quantized space. The objects gather together and form a grid. The object grid cells are assigned here and the density of every cell is calculated. The clusters which have density below certain threshold are removed here. As per the group of dense clusters, the clusters are crated. The process has higher speed due to the fact that there are no distant computations present. The neighboring clusters present here are also easy to determine. The union shapes are only provided here as a limit. On the basis of the grouping of certain cells, the complexity of the clustering is decided. Finite numbers of grids are placed here for quantizing the space of the grid based algorithms. Further, the operations are performed on this quantized space. Only on the basis of number of segments present within each dimension of the quantized space, the approaches provide high speed for processing the time independent data sets [14].

1.5 K-Mean Clustering Algorithm

The K-Means calculation uses a recursive framework. Along these lines, its functionality it is known like k-means calculation; it is described from the centre calculation as Lloyd's calculation, particularly in the Data mining group. K-implies grouping is a technique for vector quantization, at first from banner preparing, that is prevalent for bunch examination in data mining. K-implies bunching hopes to parcel n fragments into k groups in which each perception has a place with the gathering with the closest mean, filling in as a model of the group. This result in a parceling of the data space into Voronoi cells. The calculation has a free relationship to the k-closest neighbor classifier, a prevalent machine learning procedure for game plan that is as often as possible mistaken for k-implies in light of the k in the name. One can apply the 1-closest neighbor classifier on the gathering focuses acquired by k-intends to organize new data into the current groups. This is known as closest centroid classifier or Rocchio estimation [15].

Given an underlying arrangement of k means m1… , mk, the figuring proceeds by two stages:

Task (step-I): It allows each Data indicate the group in light of whose mean the slightest in the comparative bunch figured as aggregate of squares. Along these lines, the Euclidean separation is processed as squared of total squares, this is likewise speaks to the "closest" mean.

Refresh (step-II): It processes the new intends to be the centroids of the data focuses in the new groups.

The calculation joins when there is no modification in assignments, since every one of the means pick the goal of estimation, and there will be just a set number of such divisions, the approach meet to a range perfect. The count is acquainted as allotting articles with the closest bunch by registering their separation. The estimation goes for constraining the objective, and thusly it designates by "minimum total of squares," which is accurately proportionate to assigning by the slightest Euclidean separation. k-medoids have been proposed to permit utilizing other separation measures.

1.6 Classifiers

a. PCA: Principal component analysis (PCA) is a measurable technique that uses an orthogonal change to change over an arrangement of perceptions of perhaps related factors into an arrangement of estimations of directly uncorrelated factors called central parts. The quantity of primary parts is not exactly or equivalent to the quantity of unique factors [16].

Algorithm:

- Mean focus the information (discretionary)
- Figure the covariance grid of the measurements
- Discover eigenvectors of covariance lattice
- Sort eigenvectors in diminishing request of eigen qualities
- Extend onto eigenvectors all together

- Expect information framework B is of size m×n
- For each measurement, figure mean μi
- Mean focus B by subtracting μi from every segment i to get A
- Figure covariance grid C of size n×n
- In the event that mean focused, C=A^T A
- Find eigenvectors and comparing eigen values (V,E) of C
- Sort eigen values with the end goal that e_1≥e_2≥⋯≥e_n
- Extend well ordered onto the essential parts (v_1,) ⃗(v_2) ⃗,… and so forth**.**

b. SVM: A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. The steps are explained below:

1. **Set up the training data:** The preparation information of this activity is shaped by an arrangement of marked 2D-indicates that have a place one of two distinct classes; one of the classes comprises of one point and the other of three focuses.
2. **Set up SVM's parameters:** In this instructional exercise we have presented the hypothesis of SVMs in the most basic case, when the preparation cases are spread into two classes that are straightly distinguishable. Be that as it may, SVMs can be utilized as a part of a wide assortment of issues (e.g. issues with non-directly distinguishable information, a SVM utilizing a part capacity to raise the dimensionality of the illustrations, and so on). As a result of this, we need to characterize a few parameters before preparing the SVM. These parameters are put away in a protest of the class CvSVMParams [17].
3. **Train the SVM:** We call the method CvSVM::train to build the SVM model
4. **Regions classified by the SVM:** The strategy CvSVM::predict is utilized to order an info test utilizing a prepared SVM. In this case we have utilized this strategy keeping in mind the end goal to shading the space contingent upon the forecast done by the SVM. As such, a picture is navigated deciphering its pixels as purposes of the Cartesian plane. Each of the focuses is hued relying upon the class anticipated by the SVM; in green on the off chance that it is the class with name 1 and in blue on the off chance that it is the class with mark - 1.

5. **Support vectors:** We use here a few techniques to get data about the support vectors. The technique CvSVM::get_support_vector_count yields the aggregate number of support vectors utilized as a part of the issue and with the strategy CvSVM::get_support_vector we get each of the support vectors utilizing a list. We have utilized these techniques here to discover the preparation illustrations that are support vectors and highlight them**.**

   c. Bayesian classifier: A Bayesian classifier relies on upon the likelihood that the piece of a (characteristic) class is to foresee the estimations of components for individuals from that class. Representations are gathered in classes since they have ordinary qualities for the elements. Such classes are frequently called characteristic sorts. Around there, the objective component looks at to a discrete class, which is not as is normally done double. The idea behind a Bayesian classifier is that, if a specialist knows the class, it can anticipate the estimations of interchange components. In case it doesn't know the class, Bayes' govern can be used to foresee the class given (some of) the component values. In a Bayesian classifier, the learning specialist produces a probabilistic model of the elements and utilizations that model to anticipate the characterization of another case. A dormant variable is a probabilistic variable that is not viewed. A Bayesian classifier is a probabilistic model where the arrangement is a dormant variable that is probabilistically related to the watched factors. Arrangement then gets the chance to be found in the probabilistic model [18].

1.7 Algorithms used in Prediction of data in Data Mining

Distinctive supervised machine learning calculations i.e. Guileless Bayes, Neural Network, close by weighted association Apriori calculation, Decision calculation have been used for analysing the dataset. The information mining device Weka 3.6.6 is used for test. Weka is a gathering of machine learning calculations for information mining tasks. The calculations can either be associated direct to a dataset or called from your own specific Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is moreover suitable for developing new machine learning plans.

Decision Tree: Decision Tree is a notable classifier which is basic and simple to actualize. Choice tree learning is a procedure for the most part utilized as a snippet of data mining. The goal is to make a model that predicts the estimation of a target variable in perspective of a couple input factors. Each inside hub relates to one of the information factors; there are edges to youngsters for each of the conceivable estimations of that info variable. Each leaf addresses an estimation of the target variable given the estimations of the data elements tended to by the course from the root to the leaf. A decision tree is a clear depiction for requesting cases. For this piece, the greater parts of the components have constrained discrete spaces, and there is a solitary target highlight called the portrayal. Each part of the territory of the request is known as a class. A decision tree or a gathering tree is a tree in which every inside (non-leaf) center point is separate with an information highlight. The round areas starting from a center point named with a segment are separate with each of the conceivable estimations of the component. Each leaf of the tree is separate with a class or a probability movement over the classes. There is no need of territory data or parameter setting and can high dimensional data can be managed. It produces comes to fruition which are less unpredictable to examine and interpret. The enter through component to get to snappy patients' profiles is as of late open in Decision Trees

Neural Networks: A reproduced neural system (ANN), much of the time just called a "neural system" (NN), is a scientific model or computational model in light of organic neural system. By the day's end, it is a copying of organic neural framework. In sustain forward neural systems the neurons of the primary layer forward their respect the neurons of the second layer, in a unidirectional manner, which clarifies that the neurons are not got from the turn around course. There is association between each layer and weights are doled out to each association. The essential capacity of neurons of info layer is to gap enter xi into neurons in concealed layer. Neuron of concealed layer incorporates input signal xi with weights wji of individual associations from info layer. The yield Yj is capacity of

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

Where, f is a simple threshold function such as sigmoid or hyperbolic tangent function.

CHAPTER 2
REVIEW OF LITERATURE

Doreswamy, et.al," BAT-ELM: A Bio Inspired Model for Prediction of Breast Cancer Data", 2015 Medical informatics primarily deals with finding solutions for the issues identified with the diagnosis and prognosis of different deadly diseases utilizing machine learning and data mining approaches. One such disease is breast cancer, killing millions of people, particularly women. In this paper we propose a bio inspired model called BATELM which is a mix of Bat algorithm (BAT) and Extreme Learning Machines (ELM) which is first of its kind in the study of non image breast cancer data analysis. The idea of BAT and ELM which has many advantages when compared to the existing algorithms of their genre has motivated us to build a model that can predict the medical data with high accuracy and minimal error [20]. Here we make utilization of BAT to optimize the parameters of ELM so that the prediction task is completed efficiently. The fundamental aim of ELM is to predict the data with least error. For accomplishing a minimal error we have tested Wisconsin Breast Cancer Prognostic (WBCP) dataset upon three diverse learning functions (sigmoid, sin and tanh) and the function which produces the best result has been considered as the final. We completed two case studies to support our model. In case study I the goal was to predict whether the breast cancer is recurrent or non-recurrent. The accuracy obtained for this case is observed to be 95.7% with a RMSE of 0.32. In case study II our goal was to predict the season of repeat, the result obtained for this case were observed to be 93.75% accurate with a RMSE of 0.30. In both the cases tanh function performed better.

R. Karakis, et.al," A genetic algorithm model based on artificial neural network for prediction of the axillary lymph node status in breast cancer", 2013 Axillary Lymph Node (ALN) status is a basic part to audit metastatic chest development. Surgical operations which might be critical and cause some adversarial effects are performed in confirmation ALN status [21]. The motivation driving this survey is to expect ALN status by systems for picking chest development patient's essential clinical and histological feature(s) that can be procured in each retouching center. 270 chest sickness patients' data

are assembled from Ankara Numune Educational and Research Hospital and Ankara Oncology Educational and Research Hospital. It is done up from LR and GA based MLP, that menopause status and lymphatic interruption are the most basic components for choosing ALN status. GA accommodates pick best segments as MLP wellsprings of data. It moreover streamlines the weights of back expansion figuring in MLP. The estimations of backslide and precision of the GA based MLP with 9 highlights (numerical age, unmitigated age, menopause status, tumor evaluate, tumor sort, tumor region, T organizing, tumor audit and lymphatic interruption) are found as 0.96 and 98.0% with independently. As exhibited by results, proposed GA based MLP classifier can be utilized to predict the ALN status of chest tumor without surgical operations.

Ahmed Iqbal Pritom, et.al," Predicting Breast Cancer Recurrence using effective Classification and Feature Selection technique". This paper goes for discovering bosom malignancy repeat likelihood utilizing diverse information mining systems. They additionally give a respectable approach keeping in mind the end goal to enhance the precision of those models. They have gathered Cancer patient's information from Wisconsin dataset of UCI machine learning Repository. This dataset contained aggregate 35 traits in which we connected Naive Bayes, C4.5 Choice Tree and Support Vector Machine (SVM) arrangement calculations and figured their forecast precision. An effective component determination calculation made a difference them to enhance the precision of each model by lessening some lower positioned characteristics. Not just the commitments of these characteristics are less, however their expansion too misinforms the arrangement calculations. After a cautious determination of upper positioned traits they found a much enhanced exactness rate for each of the three calculations. In their work they found Support Vector Machine giving much better output both before and after attribute selection

Marjia Sultana, et.al," Analysis of Data Mining Techniques for Heart Disease Prediction", 2016 Heart disease is considered as one of the significant purposes behind death all through the world. It can't be successfully anticipated by the therapeutic authorities as it is a troublesome errand which requests aptitude and higher learning for expectation. The coronary illness transforms into a torment all through the world. It can't

be easily anticipated as it is a troublesome errand that requests ability and higher information for expectation. Information mining separates concealed data that expect a vital part in settling on decision [22]. This paper addresses the issue of expectation of coronary illness according to info properties on the introduce of information mining methodologies. We have examined the coronary illness forecast using KStar, J48, SMO, Bayes Net and Multilayer Perceptron through Weka programming. The execution of these information mining strategies is measured by consolidating the consequences of prescient exactness, ROC bend and AUC esteem using a standard informational collection and a gathered informational index. In view of the execution elements, SMO and Bayes Net frameworks show ideal exhibitions than the exhibitions of KStar, Multilayer Perceptron and J48 methodologies.

Kamaljit Kaur et.al," Analyzing the Effect of Difficulty Level of a Course on Students Performance Prediction using Data Mining", 2015 The new system, called the Credit Based Continuous Evaluation and Grading System (CBCEGS), assesses a student on the premise of her persistent evaluation during the semester, joined with her performance at last semester examination. This multistage examination design gives a chance to students to improve their performance. In the event that a student can't perform well in tests during the semester, she can improve her performance at last semester test. In any case, it doesn't appear to be so natural [23]. In specific courses, because of their difficulty level, for example, mathematics, a student will most likely be unable to improve her knowledge at last despite hard work. However, it might be conceivable in case of courses that are relatively simple, for example, System Analysis and Design. This paper breaks down and predicts student's performance utilizing data mining strategies for two data sets of 1000 students every one for Mathematics, and the other for System Analysis, and Design. This study can assist the education community with understanding learning conduct of students to the extent courses of differing difficulty are concerned. It is observed that Classification and Regression Tree (CART) supplemented by AdaBoost is the best classifiers for the prediction of students' grades for both subjects.

Monali Paul, et.al," Analysis of Soil Behaviour and Prediction of Crop Yield using Data Mining Approach", 2015 Yield expectation is astoundingly famous among

agriculturists these days, which especially adds to the correct choice of harvests for sowing. This makes the issue of foreseeing the yielding of harvests an intriguing test. Earlier yield expectation was performed by considering the agriculturist's involvement on a specific field and product [24]. This work shows a framework, which uses information mining techniques remembering the true objective to foresee the class of the separated soil datasets. The class, in this way anticipated will show the yielding of products. The issue of anticipating the product yield is formalized as an order run, where Naive Bayes and K-Nearest Neighbor techniques are used. In this work, characterization of soil into low, medium and high classes are done by embracing information mining methods remembering the ultimate objective to anticipate the harvest yield using available dataset. This review can help the dirt experts and ranchers to choose sowing in which land may bring about better yield creation. The future work may mean to make more proficient models using other information mining grouping strategies, for instance, bolster vector machine, primary segment investigation, et cetera. This examination uses a little dataset as a result of the occasion of a couple of complexities. Henceforth a bigger dataset of no less than 1gb may be used as a piece of the later work.

J. Refonaa, et.al," Analysis and Prediction of natural disaster using spatial data mining technique", 2015 Goal of the information mining system is to focus data from an informational collection and change it to a customer justifiable structure. Information mining uses the data from the past to research the aftereffect of a specific issue or situation that may emerge. The examination techniques depend on mining framework, for instance, MapReduce structure. MapReduce is a programming model for get ready huge arrangements of information with a parallel, conveyed calculation on a group [25]. It can be pertinent to more mind boggling spatial issues. Promote the usage ought to be conceivable being developed and use of novel computational methodology for the investigation of substantial spatial databases. Issue Direction depends on the application using spatial dataset takes after atmosphere information, volcanic ejections, seismic tremor, collocation designs, spatial affiliation mining. In this manner the further work will be prepared with CMR system which will be executed in cloud framework to perceive the tempest influenced shoreline front zones. It is watched that contrasting with various systems usage in cloud framework is turned out to be more productive one.

Richa Sharma, et.al," Medical Data Mining Using Different Classification and Clustering Techniques: A Critical Survey", 2016 This paper gives review on two distinctive complex maladies which consolidates the coronary ailment and Cancer infection, paper on a very basic level viewed the ebb and flow writing work to find noteworthy learning around there and outlined diverse methodologies used as a piece of illness diagnosing, advance discussed the instruments available for preparing and grouping of information. In this overview paper the abstract works of various scholars are assessed in field of medicinal information mining using diverse characterization and bunching strategies propel it is discussed that different devices are available for information preprocessing and order. This paper packs different methodologies, calculations associated around there would be helpful for scientists in restorative analysis and therapeutic practioners to build up a choice emotionally supportive network incorporating grouping and bunching strategies [26]. The choice of information mining methodologies is not same for all it really relies on upon the dataset sort, if open dataset is named then the best approach is to apply arrangement calculations while if there ought to emerge an event of unlabelled dataset it is perfect to apply grouping strategy which is most suitable for instance affirmation. This overview contemplate uncovers the significance of research in range of life crippling illness conclusion. Advance the example of prosperity is discussed that one needs to accomplish for the exactness of penny percent different explores roughly goes to their objective yet ailment conclusion experiences high false caution so we need to propose novel way to deal with decrease this false alert rate which would help in early analysis of sickness.

Sonali Shankar, et.al," Performance Analysis of Student Learning Metric using K-Mean Clustering Approach", 2016 The vast volume of information in all fields over the globe must be overseen and is used by the choice makers to receive something profitable in return. The gigantic information of 14000x5 of Harvard University online course is destitute down to find the execution measurements of enrolled understudies from various nations by methods for K-mean grouping strategy. The execution of the understudy relies on upon number of elements and grades are insufficient to speak to the all around information of an understudy. The paper hopes to separate the execution of the understudies in light of various ascribes regarding their nation [27]. The normal

execution of the understudies having a place with various nations is examined in light of various properties, for instance, created events, areas learned and number of days they communicated with the course. The characteristics are in this way contrasted and the normal evaluations of understudies of particular nations and it is inferred that the evaluations are by all record not by any means the only element to speak to the most ideal comprehension of the course. The examination can similarly be stretched out to consider exchange characteristics, for instance, 'confirmed', "investigated" et cetera.

Vadlana Baby, et.al," Distributed threshold k-means clustering for privacy preserving data mining", 2016 Protection saving is imperative in wherein information mining changes into an agreeable task among people. In information mining, a champion among the most able and as often as possible used frameworks is k-implies bunching. The information is at first grouped and after that explored to find designs. To get more exact information designs, associations share their information, which can trade off the protection of clients and their information. There are various procedures are produced to guarantee security and protection of information. In this paper, a proficient dispersed edge security saving k-implies grouping calculation is suggested that uses the code based limit mystery sharing as a protection safeguarding instrument [28]. Improvement incorporates code based approach which allows the information to be parceled into various shares and handled independently at various servers. This convention takes less number of cycles contrast and existing conventions and it don't require any trust among the servers or clients. The trial results are in like manner outfitted close by examination and security investigation of the proposed plot. It grants get-togethers to cooperatively perform grouping and henceforth evading put stock in untouchable. The convention is contrasted and CRT based bunching proposed. This calculation does not require any trust among the servers or clients and it give glorify security safeguarding of customer information.

Cheng-Fa Tsai, et.al," A New Data Clustering Approach for Data Mining in Large Databases", 2002 Bunching is the unsupervised characterization of examples (information things, highlight vectors, or discernments) into gatherings (groups). Grouping in information mining is to a great degree important to find appropriation

designs in the essential information. Bunching calculations as a general rule utilize a separation metric based likeness measure remembering the ultimate objective to allocate database with the true objective that information focuses in a comparable portion are more similar than focuses in various segments. In this paper, another information bunching technique is exhibited for information mining in extensive databases [29]. These reenactment comes about exhibit that the proposed novel grouping technique performs better than anything the Fast SOM joins K-implies approach (FSOM+K-implies) and Genetic K-Means Algorithm (GKA). Moreover, in each one of the cases concentrated, this technique makes substantially littler mistakes than both the FSOM+K-implies approach and GKA. In this paper, a novel calculation called subterranean insect settlement advancement with various support (ACODF) for information grouping is proposed. The ACODF calculation has the accompanying three vital appealing techniques: (a) using ACO with various support to deal with the bunching issue, (b) grasping mimicked treating thought for ants to decreasingly visit the measure of urban zones to get neighborhood ideal arrangements, (c) utilizing competition determination methodology to pick a way. The ACODF technique is contrasted and the FSOM+K-implies approach and GKA. Through trials, it is exhibited that ACODF productively finds exact bunches in huge high dimensional datasets.

Steve Russell, et.al," Fuzzy Clustering in Data Mining for Telco Database Marketing Campaigns", 1999 Fuzzy techniques have been associated with information mining and to databases of client data for promoting. This paper investigates fluffy bunching ways to deal with media communications database showcasing [30]. Fluffy bunching techniques can be used to mine Telco client and prospect databases to expand private and business client bit of the pie. Four key fluffy upgrades to customary database showcasing are created in this paper. To begin with, clients open have critical enrollment values in more than one particular fluffy bunch and can be considered in a characteristic way for half and half or different contacts in a given showcasing effort. Second, fluffy bunching results are seemed, by all accounts, to be reliant on the particular offer or promoting message. Third, there are differences in bunching results after some time as different offers and medicines are progressively displayed to buyers, and as items and tastes change. This development of fluffy groups can be used to grasp client devotion and

to concentrate more ideal lifetime financial relationship esteem. Fourth, in the more broadened run, formal systems can be proposed including instinctive fluffy based grouping measurements for consistent process change, to bolster dynamically adaptable and crafty crusade administration.

Vaibhav Kumar, et.al," K-mean Clustering based Cooperative Spectrum Sensing in Generalized k-μ Fading Channels", 2016 Machine learning based methodologies for range detecting and range inhabitance forecast in intellectual radio applications appear to have pulled in sufficient enthusiasm for the present writing. In this paper, K-mean grouping based unsupervised learning strategy has been embraced for the execution upgrade of helpful range detecting in summed up k-μ blurring channels. Broad recreation has been accomplished for various framework parameter exchange off in describing the collector working attributes [31]. Unsupervised learning based approach using K−mean grouping has been associated for the execution assessment of agreeable range detecting in summed up k-μ blurring channels. The learning based strategy brings about critical execution get in the ROC contrasted with the established vitality location based CSS with ideal OR-blend run the show. The accomplished outcomes assert that K = 2 gives the best execution trademark contrasted with the bunch size of K = 4 and K = 7. Energize focuses on have starting at now been started to misuse other learning strategies, for instance, diagram discriminant examination on multi-unpredictable and limited Boltzmann machines (RBM) for enhanced execution get in helpful range detecting and range inhabitance forecast in CR systems.

R. Kumari, Sheetanshu, et.al," Anomaly Detection in Network Traffic using K-mean clustering", 2016 With the headway of mechanized age and web propels advanced assaults continuously have been inciting the news features. These assaults misuse a framework's shortcomings to increment unapproved access to the touchy data or now and again essentially make a surge to shield genuine clients from getting to it. In any case interruption of the framework accept a key part before the execution of any attack. In this paper we will inspect a how these interruptions can be related to k-implies grouping based machine learning approach using huge information investigative procedures and put forth the trial results to keep away from assaults at its particularly center [32]. This

paper used the oversimplified Euclidiean remove based approach since it was expeditiously available in the mlib library. Other separation strategies, for instance, Mahalonoan separation can be use to relate the separation between various components of the informational collection more almost. More measurements, for instance, sillhoute coefficient can be used to portray the bunch quality which decides the closeness of information focuses to one group and in addition to various groups too. At last we could in like manner substitute the k implies iterative model with Gaussian blend model or DBSCAN display which may get the chance to be open in start's mlib in not all that removed future. Again with a ultimate conclusion we might need to close is that bunching and recognizing variations from the norm is not by any stretch of the imagination for framework interruption, it can in like manner be reached out to concentrate budgetary information, conduct of clients, market container examination et cetera.

Kaustubh S. Chaturbhuj, et.a;," Parallel Clustering of large data set on Hadoop using Data mining techniques", 2016 Traditional data processing techniques are not enough to handle rapidly developing data. Hadoop can be utilized for processing such large data. K-means is the traditional clustering method which is simple, scalable and can undoubtedly implement yet Kmeans converges to local minima from beginning position and sensitive to initial centers [33]. K-means required number of clusters in advance. Molecule Swarm Optimization i.e PSO is impersonate behavior based algorithm used to introduce the connectivity principle in the centroid based clustering algorithm that will gives optimum centroid and thus discover better clusters. We utilized PSO for discovering initial centroids and K-means to discover better clusters. Hadoop is utilized for quick and parallel processing of large datasets. Rapidly generating Big data is hard to handle with traditional data mining techniques. Traditional k-means clustering is sensitive to a few issues. These issues are conquer utilizing parallel k-means clustering with Particle Swarm Optimization (PSO) and generate globally optimal clusters. Hadoop and MapReduce help us handle large data parallel thus time required for clustering is reduced. Multi hubs (data hubs) help us to parallel processing and henceforth increment the scalability of method.

Daniele Casagrande, et.al," Hamiltonian-Based Clustering Algorithms for static and dynamic clustering in data mining and image processing", 2012 Gigantic measure of data is open for examination and organization raises the need for portraying, choosing, and expelling imperative information from the data. Thusly in intelligent, building, and money related matters thinks about, the demonstration of collection data rises really when sets of data must be isolated into subgroups with the motivation behind conceivably deducting normal components for data having a place with a similar subgroup. Level limit can be utilized to group data demonstrates inside a level line [34]. Level lines can be settled as bearings of a Hamiltonian structure. All the more unequivocally, the level limit is interpreted as a Hamiltonian work, and the relating Hamiltonian structure is fused. The essential static figuring can be abused to depict dynamical gathering, both in the discrete- and constant time, cases. In perspective of the differing strategy for the two time scales, unmistakable responses for the issue of dynamic gathering are depicted. The development of the method to the bundling of n-dimensional data centers is clear. Truly it fundamentally contains an iterative utilization of the two dimensional type of the computation and to the intersection purpose of the delayed consequences of every accentuation. The applications delineated in the last scope of the article demonstrate the sufficiency of the method.

Manish Kumar Sharma, et.al," Design & Analysis of K-means Algorithm for Cognitive Fatigue Detection in Vehicular Driver using Oximetry Pulse Signal", 2015 The majority of the fatal wounds and the loss of lives happen because of absence of opportune and snappy move to be made by the vehicular drivers. The trouble in determining the rate of fatigue-related accidents is because of the trouble in distinguishing fatigue as a causal or causative factor in accidents. In many instances, at least one indirect or circumstantial pieces of evidence are utilized to put forth the defense that fatigue was a factor in the accidents. In this paper an unconventional approach is proposed for fatigue detection in vehicular drivers utilizing Oximetry Pulse (OP) signal to distinguish cognitive fatigue of the driver along these lines reducing the loss of the lives and vehicular accidents [35]. This method incorporates implementation of fundamental K-means and modified K-means for detection of fatigue condition of drivers. Oximetry Pulse signal has been recorded from vehicular drivers for Pre and Post driving states and

were processed utilizing various wavelet functions to extract run of the mill set of features. The K-means classifiers were trained and tested for these datasets. Each of the features extracted was treated as single decision making parameter. From the test outcomes it could be found that a portion of the wavelet features could fetch 100 % classification accuracy with modified K means while few others with essential K-means classifier. After top to bottom analysis of the results, it could be finally conclude that the proposed algorithm could perform extremely well in accordance with the current methodologies with some selected features of OximetryPulse signal.

Aimin Yang, et.al," A Constructing Method of Fuzzy Classifier Using Kernel K-means Clustering Algorithm", 2009 A developing technique for fluffy classifier using piece k-implies grouping calculation is presented in this paper. This developing technique is partitioned into three stages, to be particular bunching stage, fluffy manage made stage and parameters adjusted stage. Right off the bat, the primary case space is mapped into a high dimensional element space by choosing fitting piece work. In the element space, preparing tests are assembled into a couple groups by part k-implies bunching calculation [36]. By then for each made bunch, a fluffy lead is described with the suitable participation work. At long last, a couple of parameters of fluffy classifier are picked by GAs. The analysis comes about show the proposed fluffy classifier has high arrangement exactness by the correlation comes about with the relative approach, and has the better associated values. In this paper, we present a building technique for fluffy classifier using KKMC calculation. For each made bunch, a fluffy grouping guideline is made. CK and ∂ parameters are adjusted by GAs. The exactness of the built classifier by our proposed strategy are similar to the most outrageous precision of reference, and the preparation time is considerably shorter. Later on, other bit capacities and enrollment capacities will be chosen to finish more examination.

CHAPTER 3
PRESENT WORK

3.1 PROBLEM FORMULATION

Following are the various research gaps of existing work which is fulfilled in this research

1. In the existing algorithm two step clustering is used to cluster the data of input dataset. In the two step clustering k-mean algorithm is used in the first step in which arithmetic mean is calculated of the whole dataset which will be the central point to calculate similarity between the member of the dataset. To calculate similarity between the points technique of Euclidian distance is applied and in second step of two step clustering decision tree classifier is applied which create final clusters. It is been analyzed that when the dataset is complex means difficult to drive relation between members then accuracy of the clustering is reduced which need to be improved in the future work
2. In the basepaper work, technique of two step clustering is not able to drive exact relationship between member of the dataset due to which accuracy of clustering is reduced. The accuracy is reduced because some data points are remained uncluttered or wrongly clustered. The relationship between the member functions depends upon the Euclidian distance if it is not calculated appropriately then accuracy is reduced. In future technique will be applied which will calculate Euclidian distance in the iterative manner and exact distance is assumed which gave maximum accuracy.

3.2 Objectives

1. To study and analyze various prediction based technique for Data mining

2. To propose improvement in K-mean and SVM based prediction techniques for data classification

3. The proposed improvement will be based on back propagation algorithm to increase accuracy of classification

4. To implement proposed technique and compare with existing in terms of accuracy, execution time

3.3 Research Methodology

The k-mean clustering is the clustering technique in which similar and dissimilar data is clustered together on the basis of their similarity. In the k-mean clustering, the dataset is considered and from that dataset arithmetic mean is calculated which will be the central point of the dataset. The Euclidian distance from the central point is calculated and points which are similar and dissimilar are clustered into different clusters. The Euclidian distance is calculated dynamically in this work to increase accuracy of clustering. The Euclidian distance is calculated dynamically using back propagation algorithm using cluster the uncluttered points and increase accuracy of clustering. The back propagation algorithm is the algorithm which learns from the previous experiences and drive new values. The formula given below is used to drive values from the input dataset. In the formula given the x is the each point in the dataset and w is the value of the data point from which the actual output is taken and bias the value which is used to change the final value of output. The output of each iteration is compared with the output of next iteration and iteration at which error is minimum is the final value of Euclidian distance. When the error is reduced , the accuracy of clustering is increased and execution time is reduced.

$$\text{Output: } \sum_{\substack{w=0 \\ x=0}}^{\substack{w=n \\ x=n}} x_n w_n + bias$$

$$Error = Desired\ Output - Actual\ Output$$

HYBRID ALGORITHM

INPUT : Dataset

OUTPUT: Clustered Data

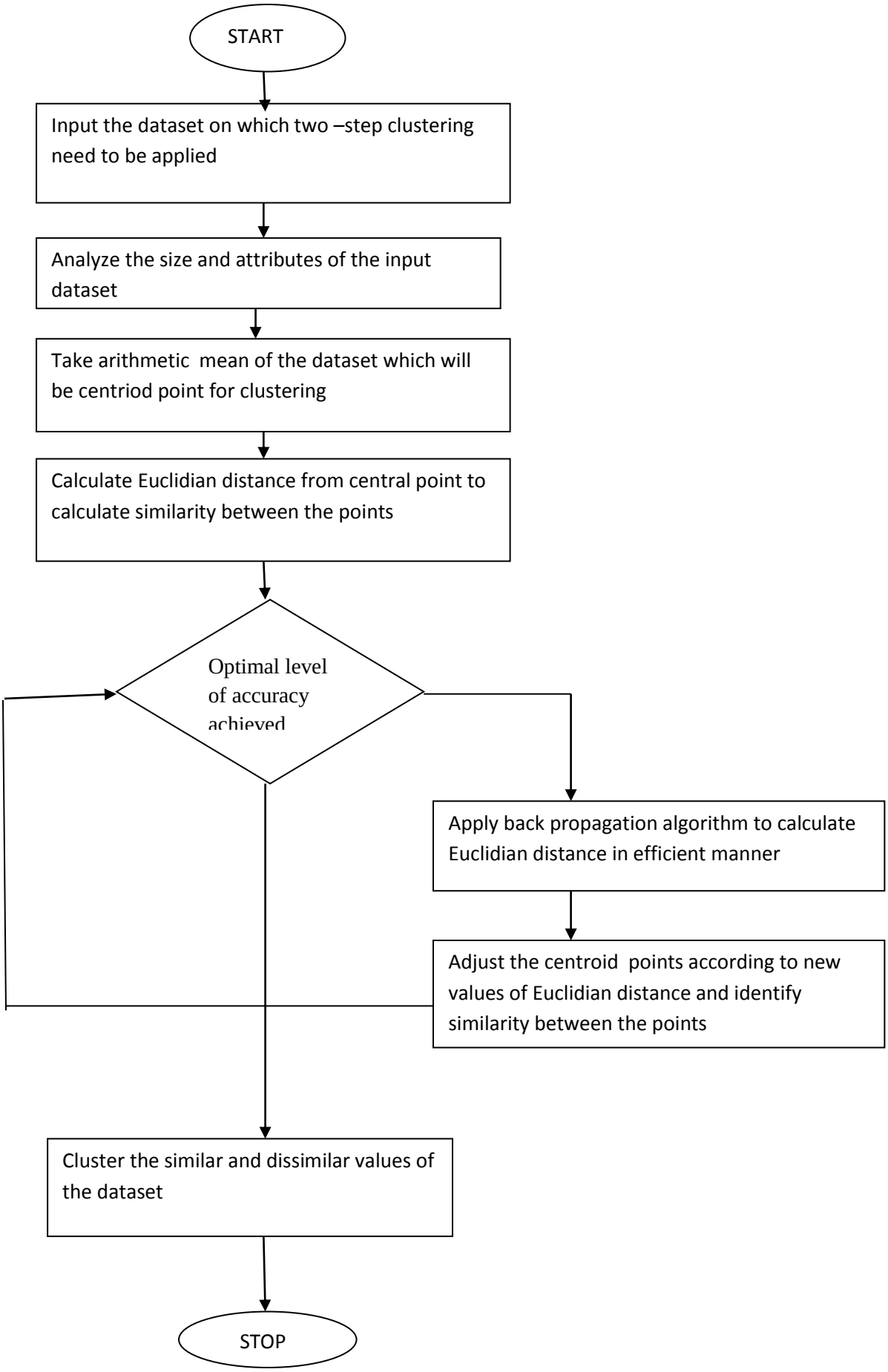
Start ( )

1. Read dataset and dataset has number of rows “r” and number of coloums “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.


START
Input the dataset on which two –step clustering need to be applied
Analyze the size and attributes of the input dataset
Take arithmetic mean of the dataset which will be centriod point for clustering
Calculate Euclidian distance from central point to calculate similarity between the points
Optimal level of accuracy achieved
Apply back propagation algorithm to calculate Euclidian distance in efficient manner
Adjust the centroid points according to new values of Euclidian distance and identify similarity between the points
Cluster the similar and dissimilar values of the dataset
STOP

CHAPTER 4
RESULT AND DISCUSSION

4.1 Tool Description

The matlab is the tool which is used to perform mathematical complex computations. In this MATLAB simplified C is used as the programming language. The MATLAB has various inbuilt toolboxes and these toolboxes are mathematical toolbox, drag and drop based GUI, Image processing, Neural networks etc. The MATLAB is generally used to implement algorithms, plotting graphs and design user interfaces. The MATLAB has high graphics due to which it is used to simulate networks. The MATLAB has various versions by current MATLAB version is 2015. The MATLAB default interface has following parts

1. **Command Window:-** The Command Window is the first importance part of MATLAB which is used to show output of already saved code and to execute MATLAB codes temporarily
2. **WorkSpace:-**The workspace is the second part of MATLAB which is used to show allocation and deallocation of MATLAB variables. The workspace is divided into three parts. The first part is MATLAB variable,variable type and third part is variable value
3. **Command History:-** The command history is the third part of MATLAB in which MATLAB commands are shown which are executed previously
4. **Current Folder Path:-** The current Folder path shows that path of the folder in which MATLAB codes are saved
5. **Current Folder Data:-** The Current Folder Data shows that data which is in the folders whose path is given in Current Folder Path

The MATLAB has three Command which are used frequently and these commands are :-

1. CLC:- The ‘clc’ stands for clear command window
2. Clear all:- The ‘clear all’ command is used to de-allocate the variable from the workspace

3. Close all:- The close all is the command which is used to close all the interfaces and return you to default MATLAB interface

4.2 Experimental Results

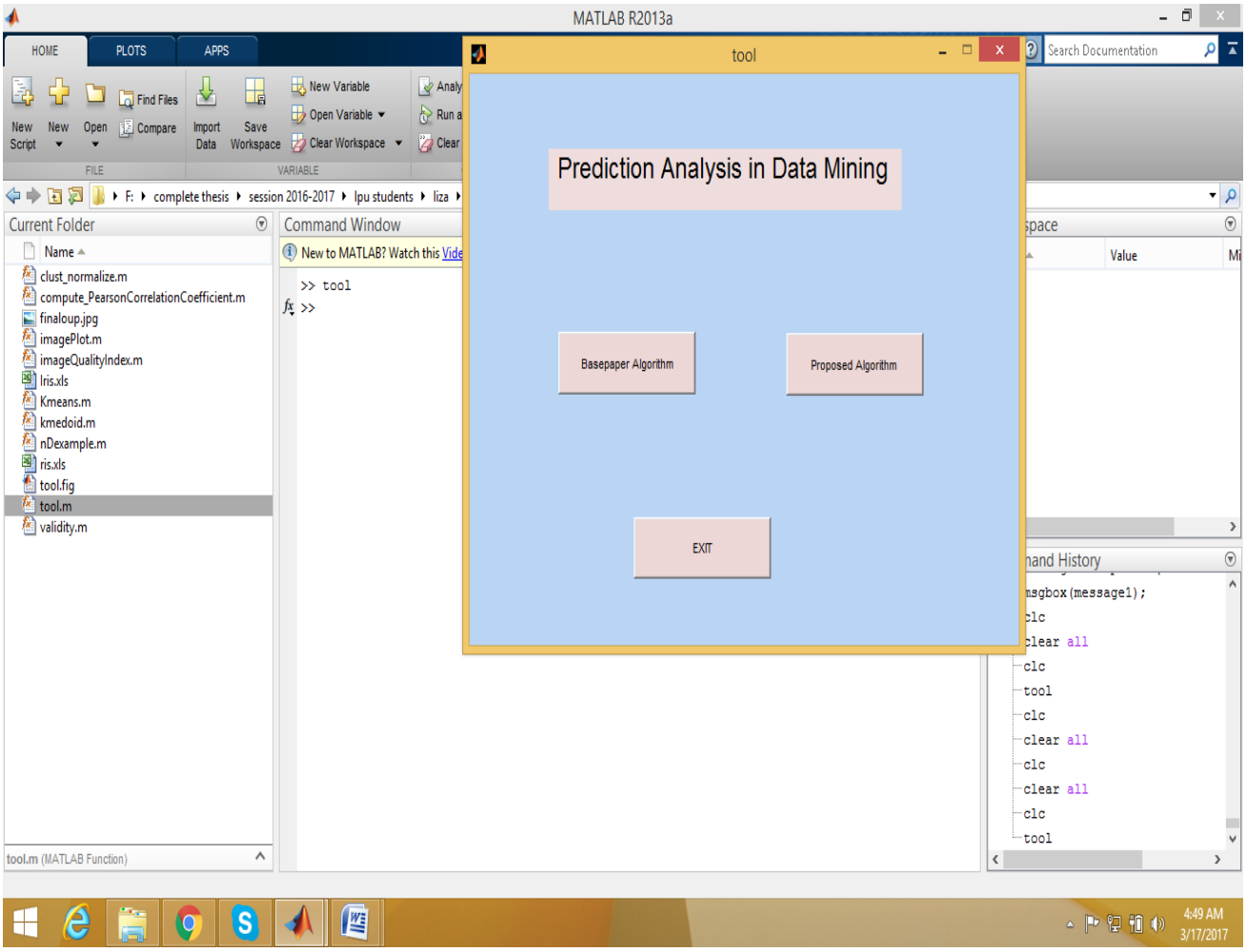


Figure7: Interface of code execution

As shown in figure 1, the interface is designed in the MATLAB using the guide toolbox and in the interface three buttons are their the first button is of basepaper execution, second of proposed work and last of exit.

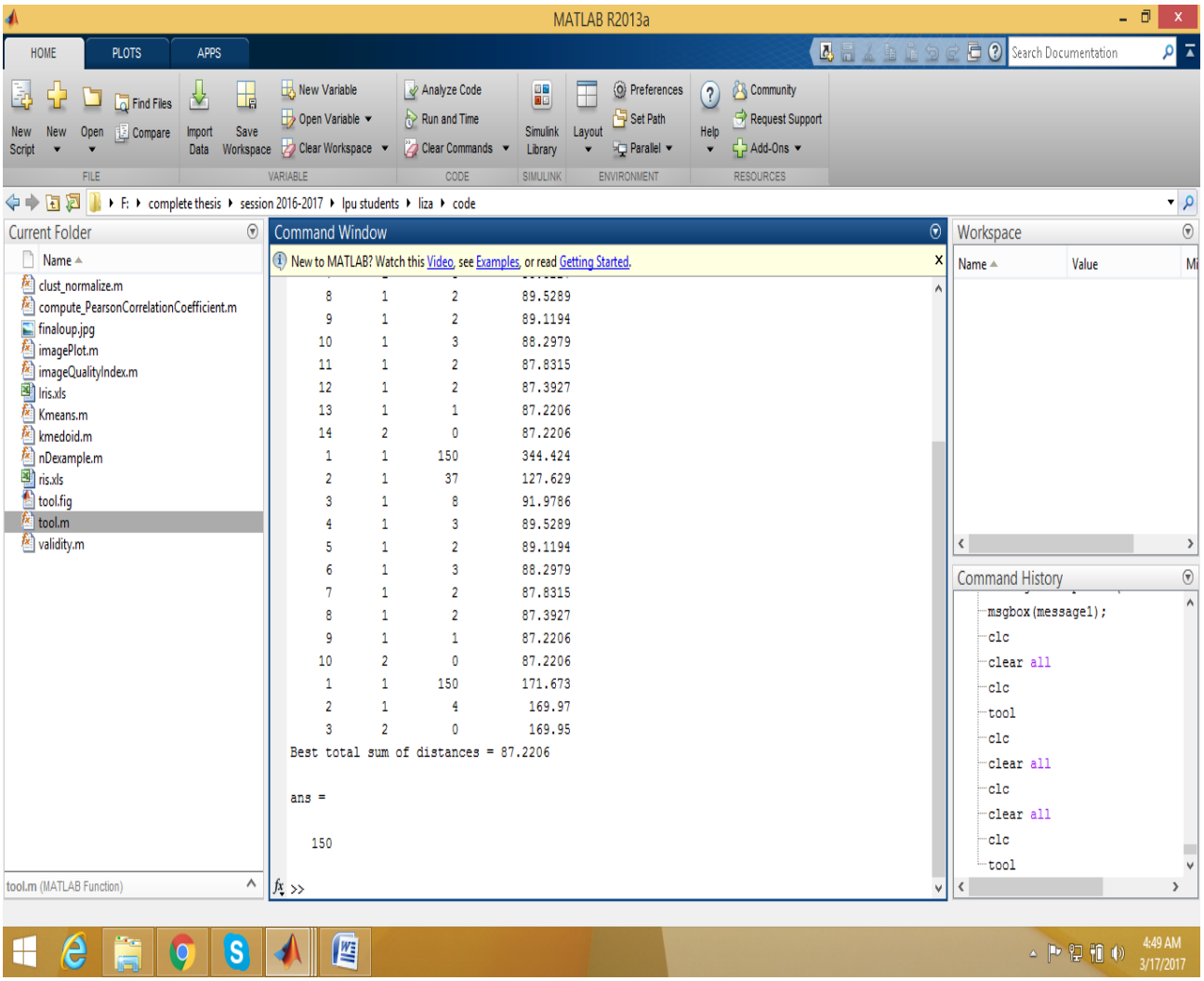


Figure8: Euclidian distance calculation

As shown in figure 2, the datatset is loaded and loaded dataset is used for the clustering. The Euclidian distance calculated to analyze similarity and dissimilarity between the data points.

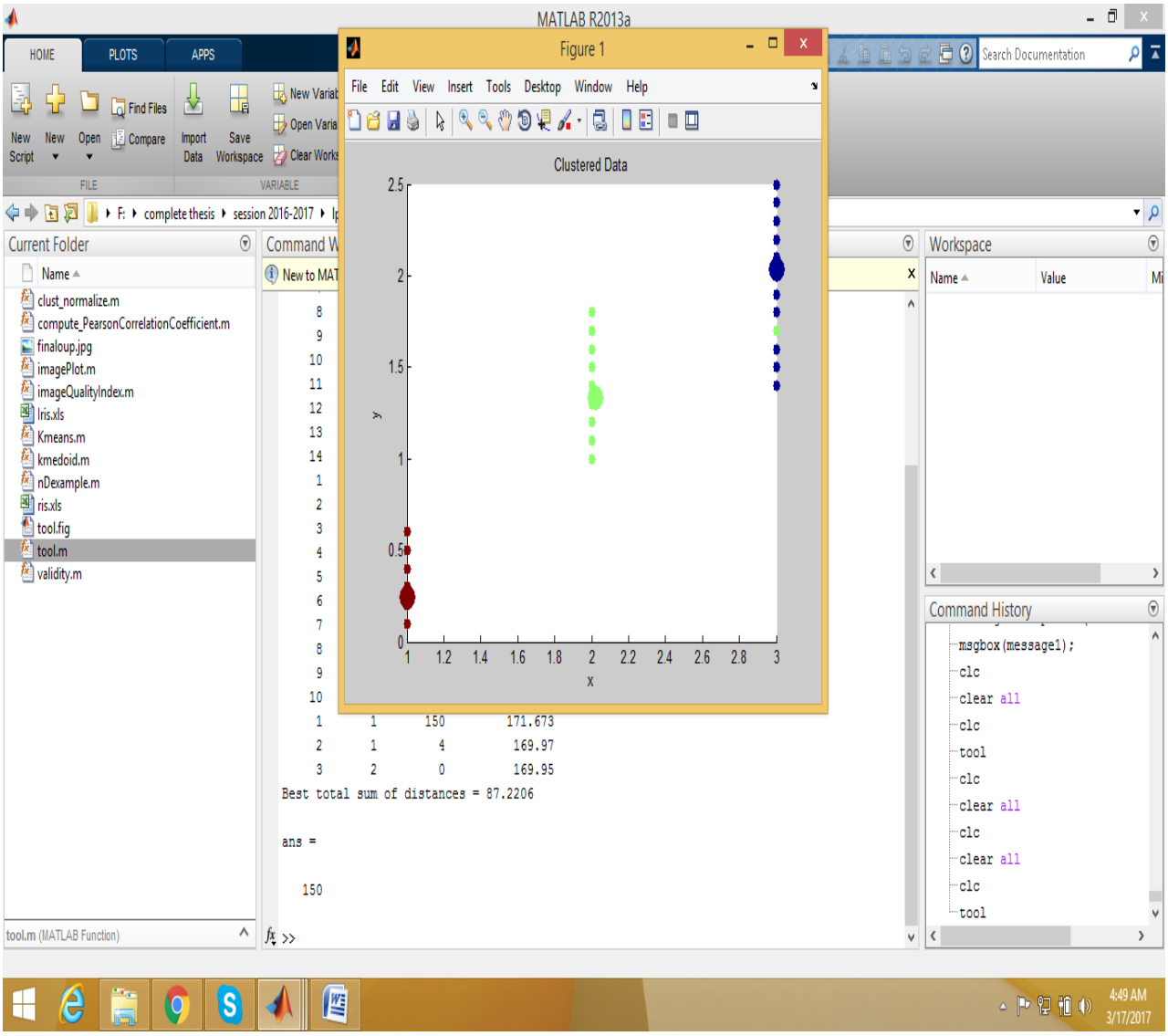


Figure9: Clustered output

As shown in figure 3, the Euclidian distance is checked to analyze similarity and dissimilarity between the data points. The final output is shown in the form of clusters

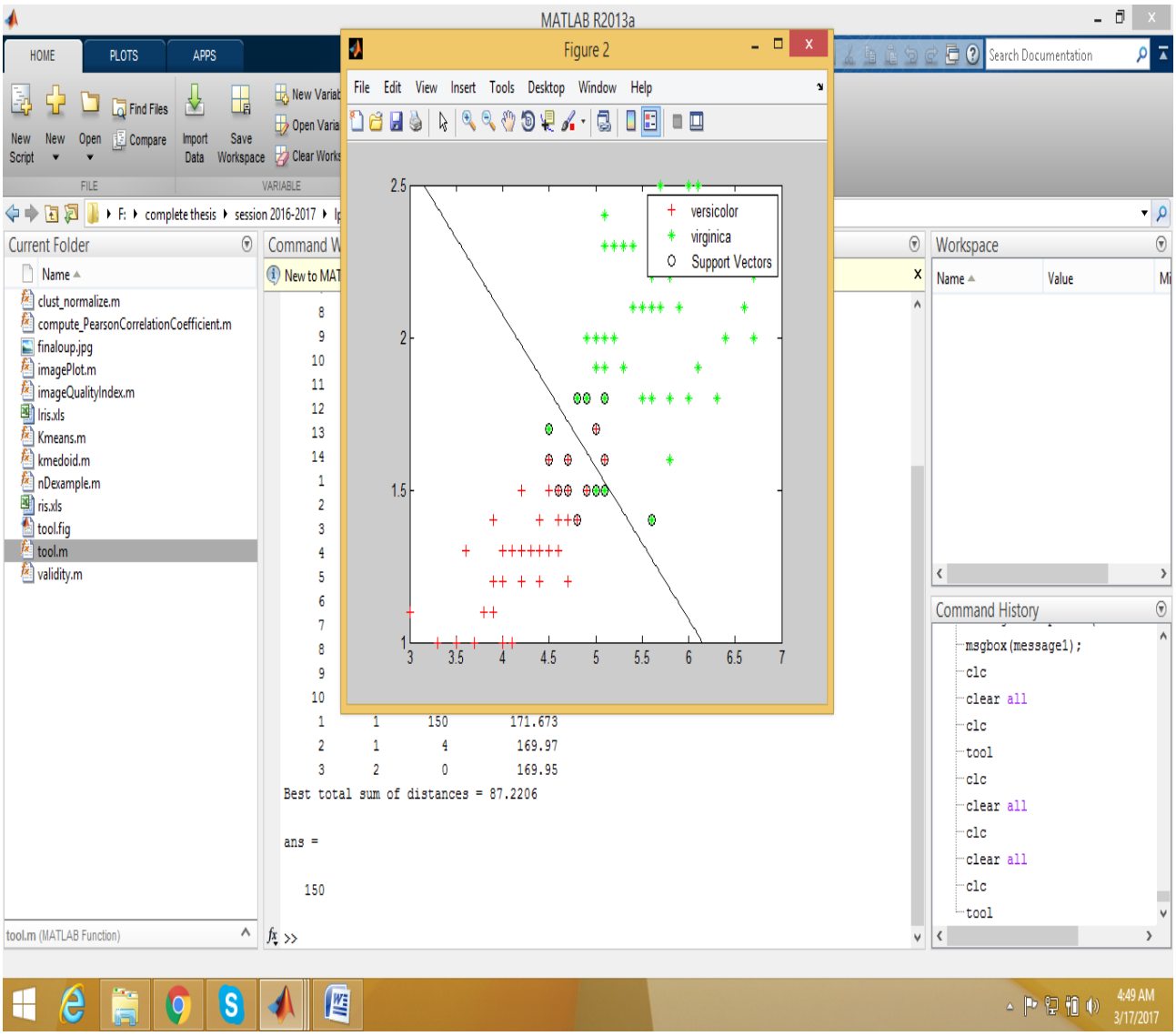


Figure10: SVM classification

As shown in the figure 4, the clustered result is the clustering of similar and dissimilar data points

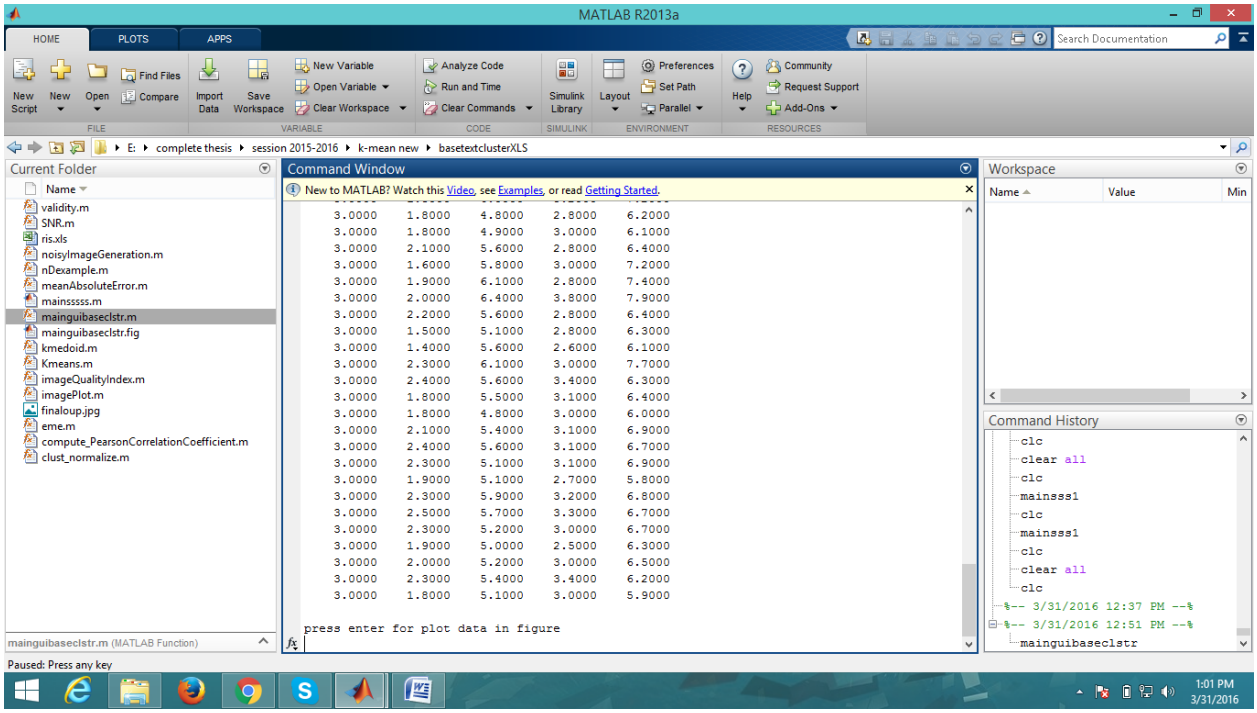


Figure11: DATASET Clustered

As appeared in the figure 5, As clarified in the beforehand the K-mean is the calculation in which the information will be grouped by Euclidian separation. The irregular focus focuses had been chosen from the information. The Euclidian separation will be ascertained from the server farms to different focuses and focuses will be bunched as needs be. The yield of the bunched will be appeared in the 2D plane. At the point when the information will be appeared in 2D arrange, a few focuses which are near each other can't be indicated which lessen the bunch quality.

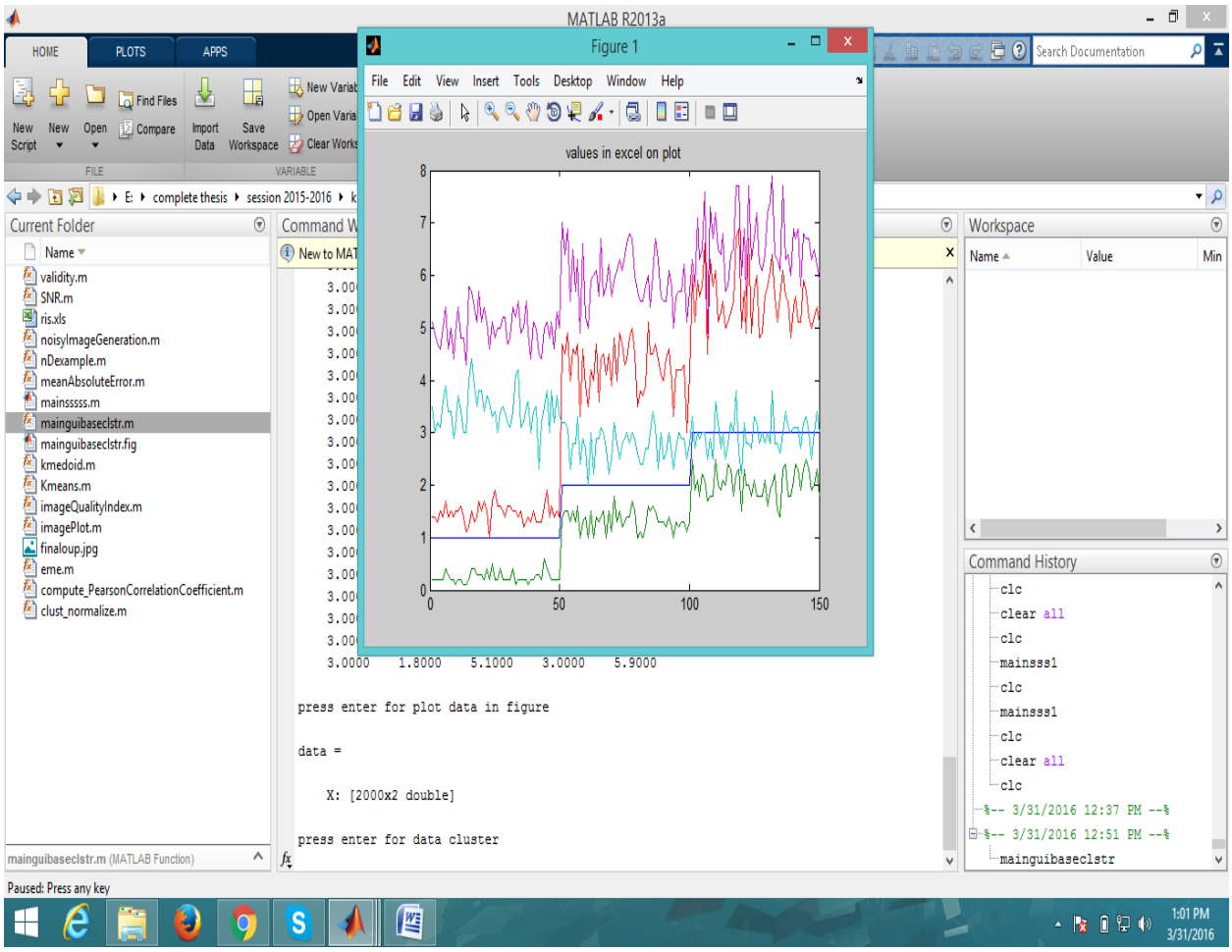


Figure12: Plot of Data

As illustrated in figure 6, the dataset which is read with xlsread commands of MATLAB will be plotted on the 2-D plane

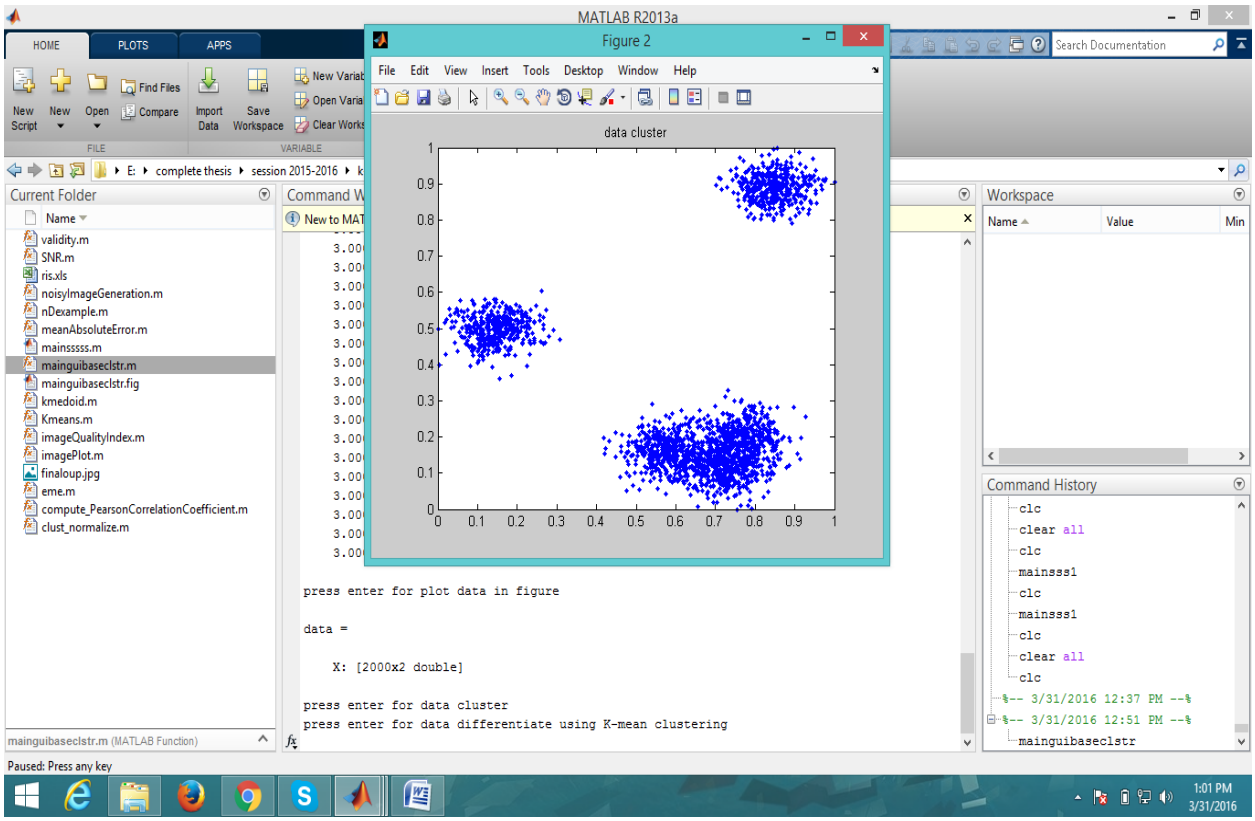


Figure13: First iteration of clustering

As appeared in the figure 7, To dissect the execution of the calculation the K-mean calculation will be connected on another dataset. In this Dataset, different figures have been appeared for information bunching. In the figure3, the information focuses have been plotted which we need to group and first focused point will be chosen and as per first chose focuses information will bunched.

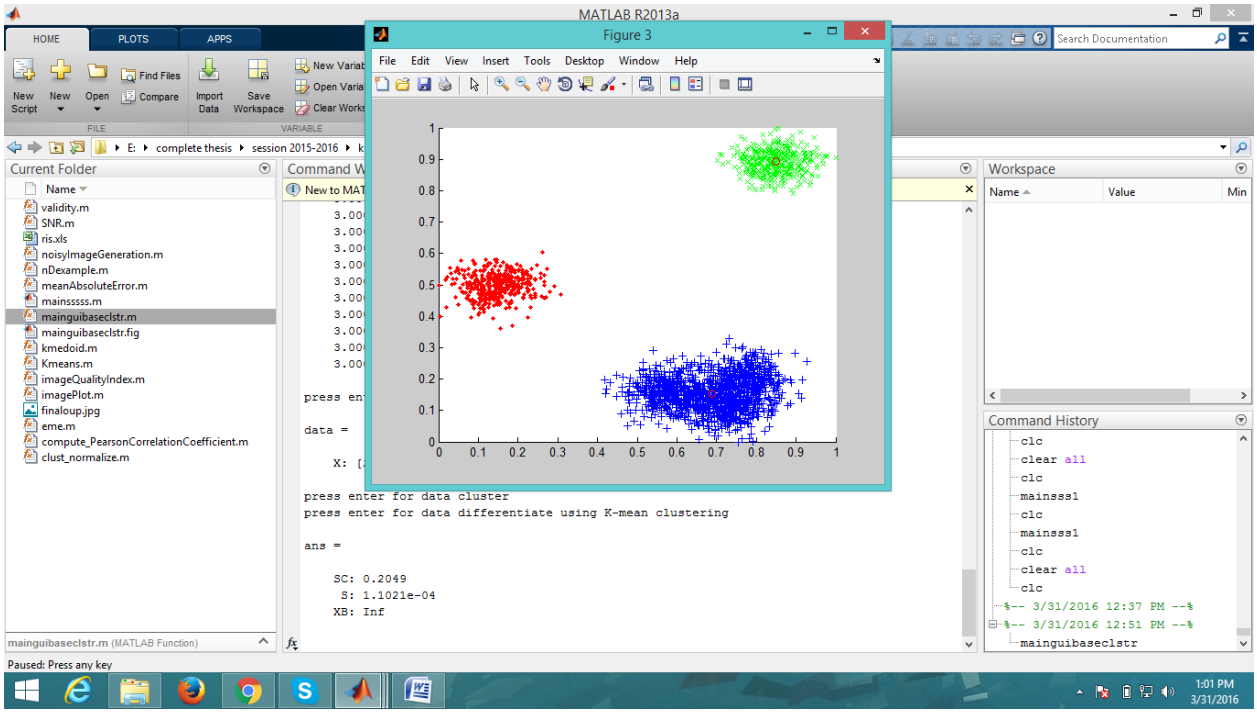


Figure14: Coloring of data points

As appeared in figure8, in figure3, the primary chose focuses are utilized for grouping of information. In this figure, the information will be bunched utilizing Euclidian separation and this information will be appeared with changed hues for better investigation of information

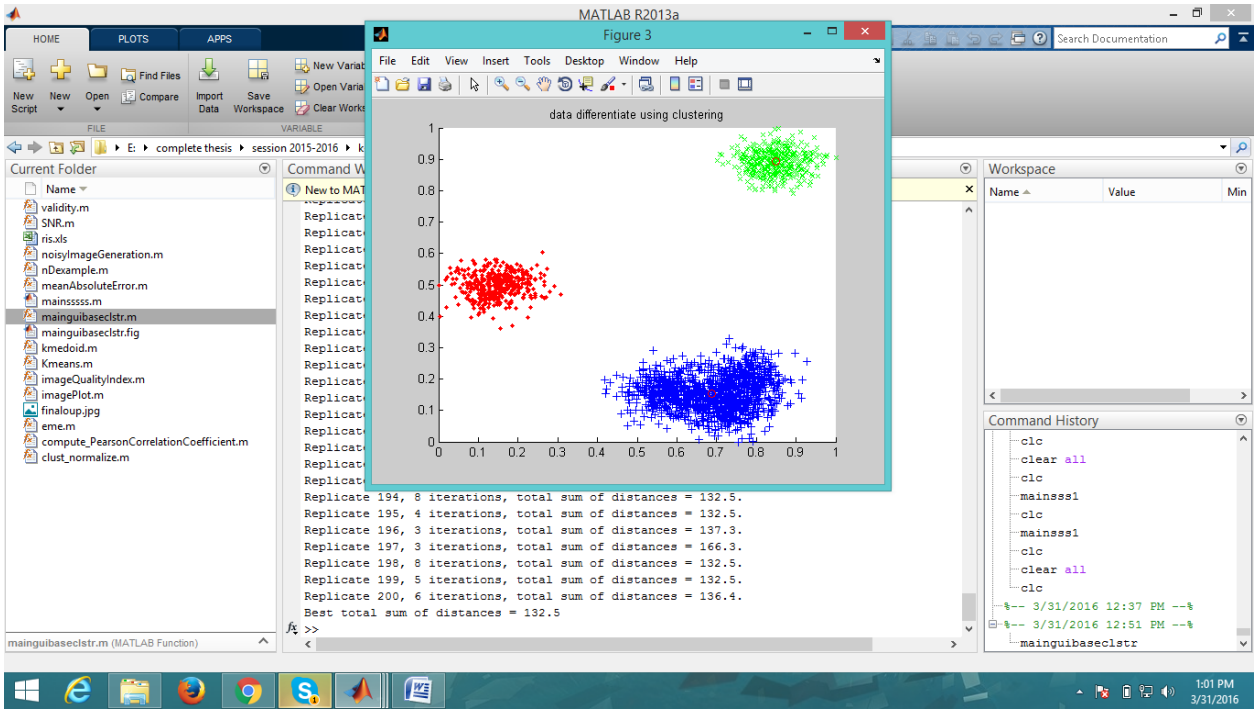


Figure 15: Clustering of Data

As appeared in figure 9, The informational collection which is utilized for bunching is been grouped and each group will be set apart with various hues. In this figure, different emphasis run, implies at each cycle new focused point is chosen and on the premise of that focused point, group task method will be finished.

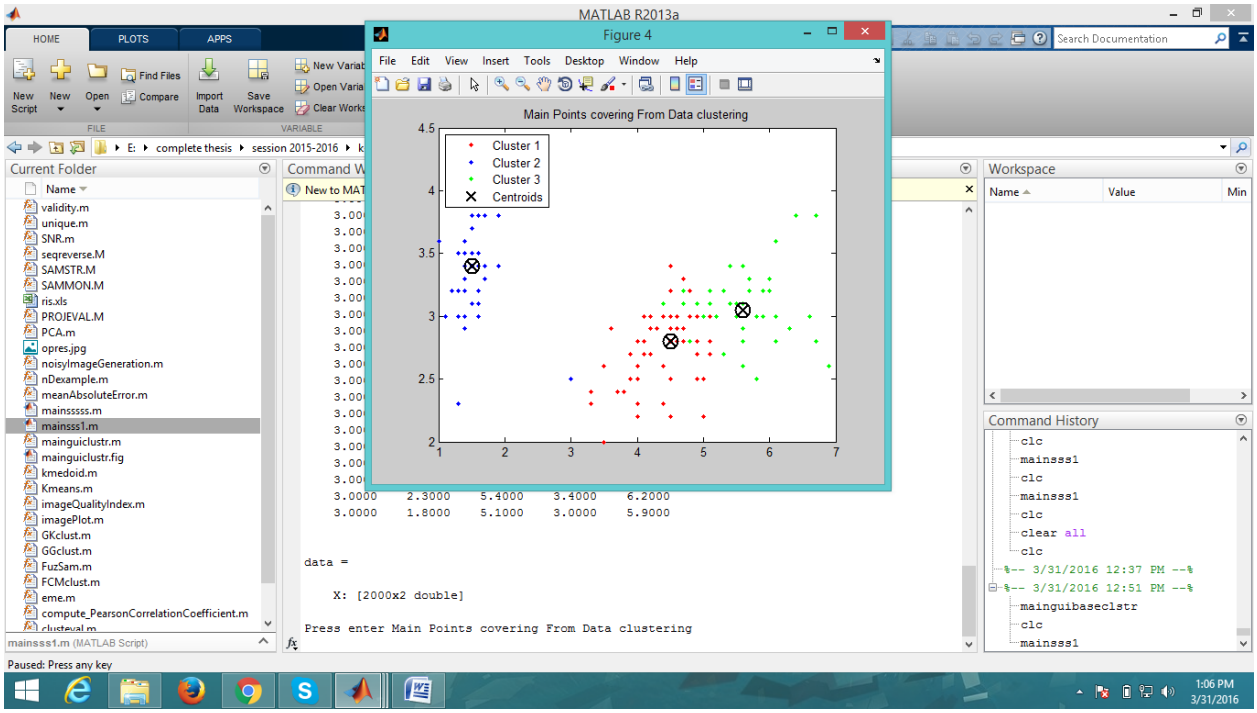


Figure16: Vornoalie Representation

As appeared in figure 10, the dataset which is utilized as a part of the past figure will be grouped utilizing the half and half kind of k-mean bunching calculation. At the point when the dataset will be bunched utilizing half and half calculation group quality will be enhanced and each point in the dataset will be appeared on voronlie plane for better examination of dataset

4.3 Improvement in Result

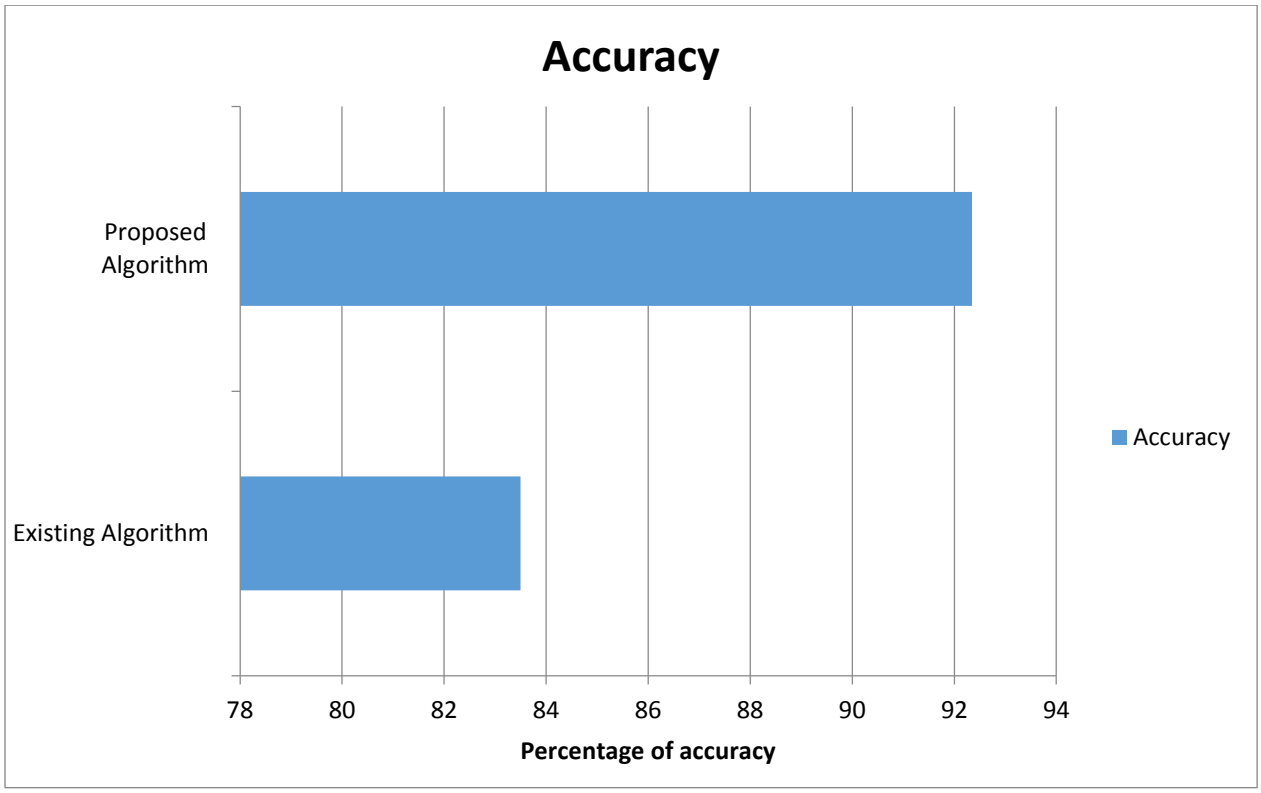


Figure17: Accuracy Comparison

As shown in figure 11, the accuracy of proposed and existing algorithm is been compared and it is been analyzed that proposed algorithm has high accuracy due to clustering of uncluttered points from the dataset

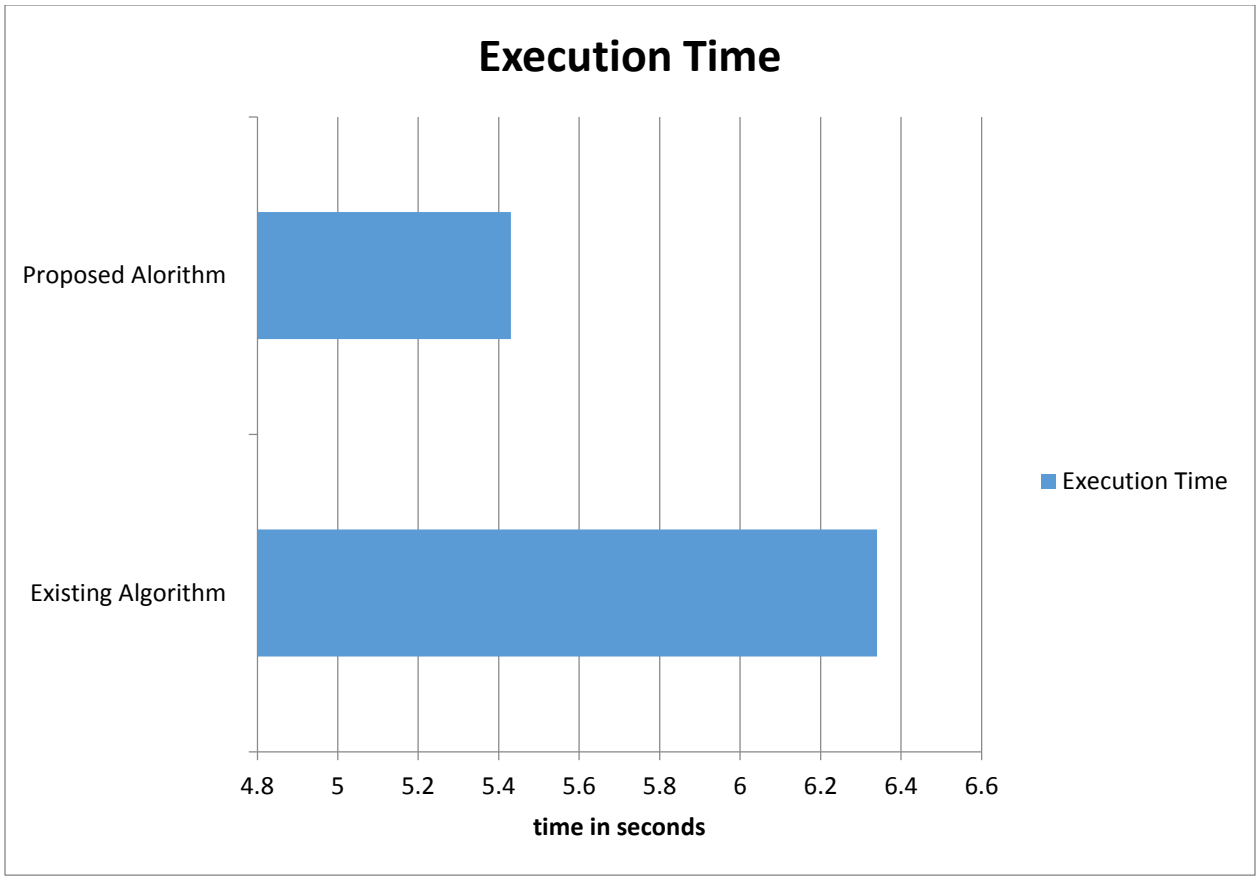


Figure18: Execution time

As illustrated in figure 12, the execution time of proposed and existing algorithm is been compared and due to used of back propagation algorithm execution time is due in the proposed work

CHAPTER 5
CONCLUSION AND FUTURE SCOPE

5.1 CONCLUSION

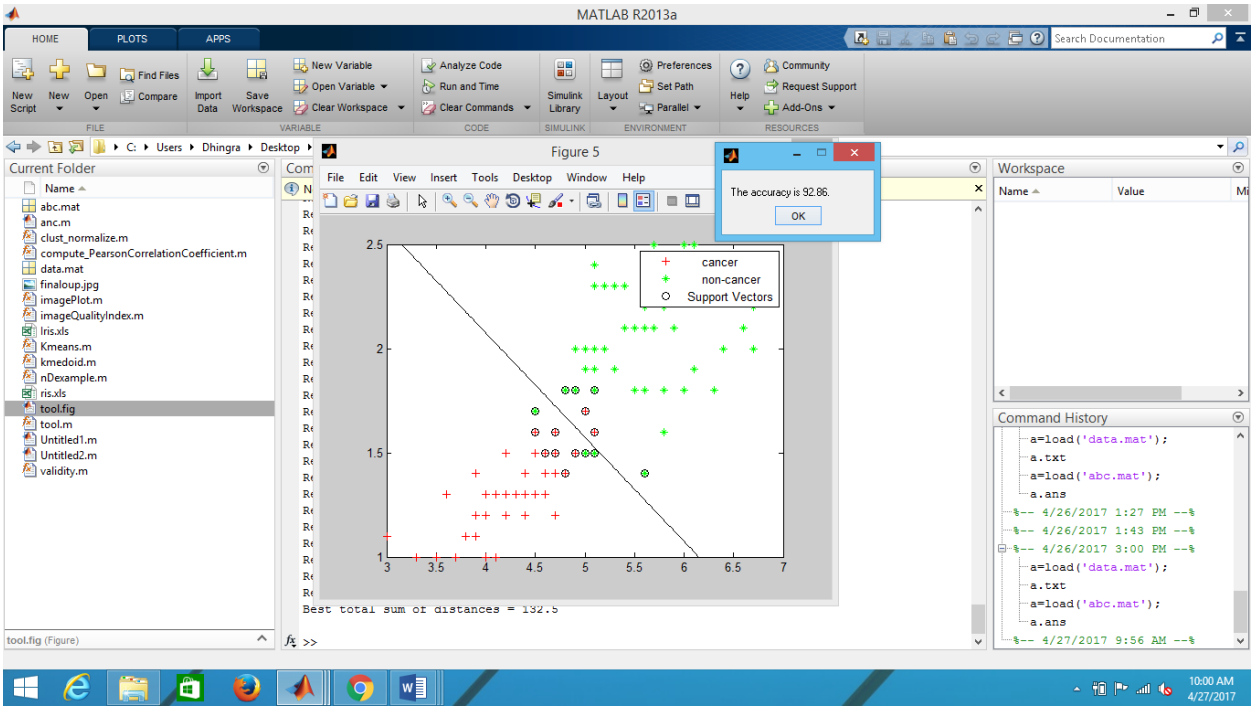


Figure 19: Improved Accuracy
The accuracy is increased to 92.86

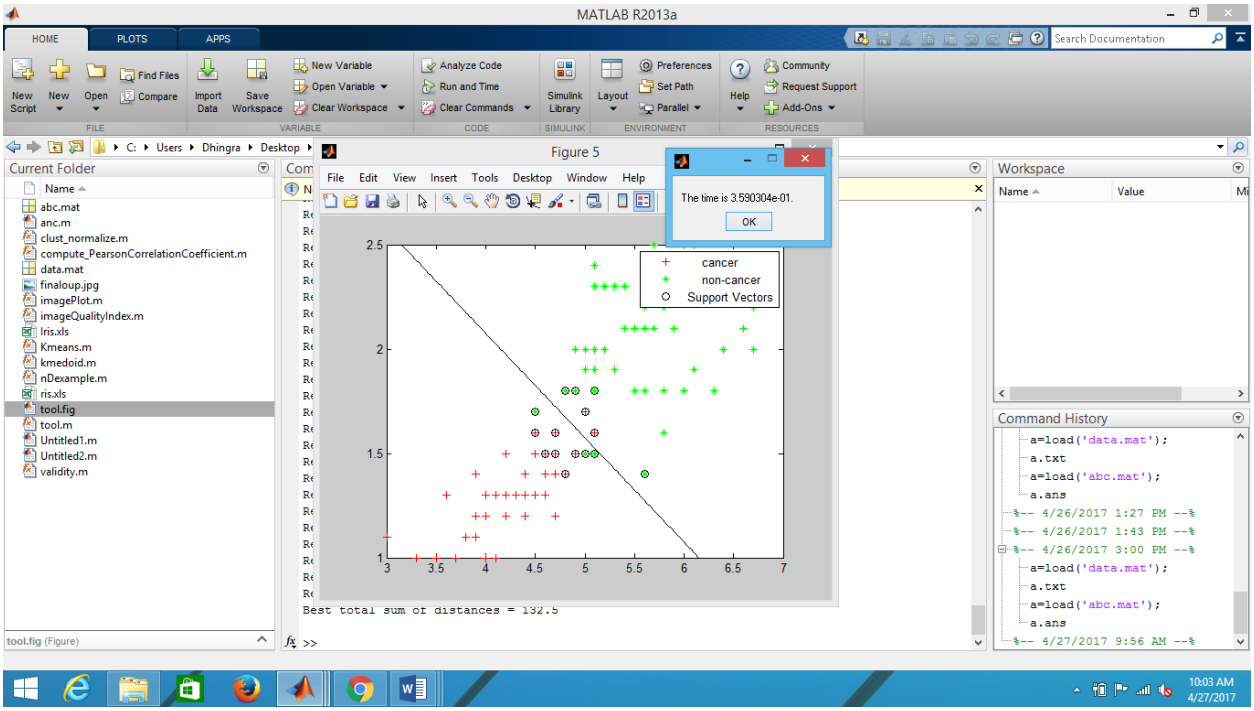


Figure 20: Improved Execution Time

The execution time is reduced to 3.590304e-01

5.2 FUTURE SCOPE

Following are the various future prospective of this research:-

1. The proposed technique can be applied on the other datasets to test the performance of the improved algorithm

2. The proposed prediction analysis technique can be compared

3. We can also create another Hybrid Model by using another algorithm for classification to improve accuracy and decrease execution time

REFERENCES

[1] Micheline Kamber and Jian Pei Jiawei Han,” Data Mining Concepts and Techniques”, 2012, 3rd ed.

[2] Mohammed Abdul Khaled, Sateesh Kumar Pradhan and G.N. Dash,” A survey of data mining techniques on medical data for finding locally frequent diseases,” 2013, International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, pp. 149-153

[3] Abdur Razzak, "A questionnaire survey on infectious disease among hospital patients in Kushtia and Jhenaidah, Bangladesh," 2011, International Journal of Genetics and Molecular Biology, vol. 3, no. 9, pp. 120-xxx

[4] D. P. Shukla, Shamsher Bahadur Patel and Ashish Kumar Sen, "A literature review in health informatics using data mining techniques," 2014, International Journal of Software & Hardware Research in Engineering, vol. 2, no. 2

[5] V. Gayathri, M.Chanda Mona and S.Banu Chitra, "A survey of data mining techniques on medical diagnosis and research," 2014, International Journal of Data Engineering (OOE) Singapore Journal of Scientific Research (SJSR), vol. 6, pp. 301-310

[6] M.Akhil Jabbar, Priti Chandra and B.L Deekshatulu, "Heart disease prediction system using associative classification and genetic algorithm," 2012, International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies-ICECIT

[7] R. Chitra and V.Seenivasagam, "Review of heart disease prediction system using data mining and hybrid intelligent," 2013, ICTACT Journal on Soft Computing, vol. 03, no. 04

[8] Abhishek Taneja, "Heart disease prediction system using data mining techniques," 2013, Oriental Journal of Computer Science & Technology, vol. 6, pp. 457-466

[9] Hlaudi Daniel Masethe and Mosima Anna Masethe, "Prediction of heart disease using classification algorithms," 2014, Proceeding of the World Congress on Engineering and Computer Science, vol. II, San Francisco, USA

[10] Rupali, R.Patil, "Heart disease prediction system using Naive Bayes and Jelinek-mercer smothing," 2014, International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, no. 5

[11] Shamsher Bahadur Patel, Pramod Kumar Yadav and Dr. D. P. Shukla, "Predict the diagnosis of heart disease patients using classification mining Techniques," 2013, IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS), vol. 4, no. 2, pp. 61-64

[12] Jyoti Soni, Ujma Ansari and Dipesh Sharma, "Prediction data mining for medical diagnosis: An overview of heart disease prediction," 2011, International Journal of Computer Applications (0975-8887), vol. 17

[13] John G. Cleary and Leonard E. Trigg," K: An Instance-based learner using an entropic distance measure," 1995, Proc. 12th International Conference on Machine Learning, pp. 108-114

[14] S. Vijayarani and M. Muthulkshmi, "Comparative analysis of Bayes and Lazy classification algorithms," 2013, International Journal of Advanced Research in Computer and Communication Engineering, vol. 2, no. 8

[15] R. Vijaya Kumar Reddy, K. Prudvi Raju, M. Jogendra Kumar, CH. Sujatha, P. Ravi Prakash, "Prediction ofheart disease using decision tree approach," 2016, International Journal of Advanced Research in Computer Science and Engineering, vol. 6, no. 3

[16] Promad Kumar Yadav, K. L. Jaiswal, Shamsher Bahadur Patel, D. P. Shukla, "Intelligent heart disease prediction model using classification algorithms," 2013, UCSMC, vol. 3, no. 08, pp. 102-107

[17] Gaurav Taneja and Ashwini Sethi, "Comparison of classifiers in data mining," 2014, International Journal of Computer Science and Mobile Computing, vol. 3, pp. 102-115

[18] Sheweta Kharya, "Using data minig techniques for diagnosis of cancer disease," 2012, UCSEIT, vol. 2, no. 2

[19] M. Hall, E. Frank, G.Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The weka data mining software: An update," 2009, SIGKDD explorations, vol. 11

[20] Doreswamy, Umme Salma M," BAT-ELM: A Bio Inspired Model for Prediction of Breast Cancer Data", 2015, IEEE

[21] R. Karakis, M. Tez, Y. Kilic, Y. Kuru, and I. Guler," A genetic algorithm model based on artificial neural network for prediction of the axillary lymph node status in breast cancer," 2013, Engineering Applications of Artificial Intelligence, vol. 26, no. 3, pp. 945–950

[22] Marjia Sultana, Afrin Haider and Mohammad Shorif Uddin," Analysis of Data Mining Techniques for Heart Disease Prediction", 2016, IEEE

[23] Kamaljit Kaur and Kuljit Kaur," Analyzing the Effect of Difficulty Level of a Course on Students Performance Prediction using Data Mining", 2015 1st International Conference on Next Generation Computing Technologies (NGCT)

[24] Monali Paul, Santosh K. Vishwakarma, Ashok Verma," Analysis of Soil Behaviour and Prediction of Crop Yield using Data Mining Approach", 2015, IEEE

[25] J. Refonaa, Dr. M. Lakshmi, V.Vivek," ANALYSIS AND PREDICTION OF NATURAL DISASTER USING SPATIAL DATA MINING TECHNIQUE", 2015, International Conference on Circuit, Power and Computing Technologies [ICCPCT]

[26] Richa Sharma, Dr. Shailendra Narayan Singh, Dr. Sujata Khatri," Medical Data Mining Using Different Classification and Clustering Techniques: A Critical Survey", 2016, IEEE, 978-1-5090-0210-8

[27] Sonali Shankar, Bishal Dey Sarkar, Sai Sabitha, Deepti Mehrotra," Performance Analysis of Student Learning Metric using K-Mean Clustering Approach", 2016, IEEE, 978-1-4673-8203-8

[28] Vadlana Baby, Dr. N. Subhash Chandra," Distributed threshold k-means clustering for privacy preserving data mining", 2016, IEEE, 978-1-5090-2029-4

[29] Cheng-Fa Tsai, Han-Chang Wu, and Chun-Wei Tsai," A New Data Clustering Approach for Data Mining in Large Databases", 2002, IEEE, 1087-4089

[30] Steve Russell, Steve Russell," Fuzzy Clustering in Data Mining for Telco Database Marketing Campaigns", 1999, IEEE, 0-7803-521 1 -4

[31] Vaibhav Kumar, Deep Chandra Kandpal, Monika Jain," K-mean Clustering based Cooperative Spectrum Sensing in Generalized k-μ Fading Channels", 2016, IEEE, 978-1-5090-2361-5

[32] R. Kumari, Sheetanshu, M. K. Singh, R. Jha, N.K. Singh," Anomaly Detection in Network Traffic using K-mean clustering", 2016, IEEE, 978-1-4799-8579-1

[33] Kaustubh S. Chaturbhuj, Mrs. Gauri Chaudhary," Parallel Clustering of large data set on Hadoop using Data mining techniques", 2016, IEEE, 978-1-4673-9214-3

[34] Daniele Casagrande, Mario Sassano, and Alessandro Astolfi," Hamiltonian-Based Clustering Algorithms for static and dynamic clustering in data mining and image processing", 2012, IEEE, IEEE CONTROL SYSTEMS MAGAZINE, 1066-033X

[35] Manish Kumar Sharma, Mahesh M. Bundele," Design & Analysis of K-means Algorithm for Cognitive Fatigue Detection in Vehicular Driver using Oximetry Pulse Signal", 2015, IEEE International Conference on Computer, Communication and Control IC4

[36] Aimin Yang, Qing Li, Xinguang Li," A Constructing Method of Fuzzy Classifier Using Kernel K-means Clustering Algorithm", 2009, IEEE, 978-0-7695-3888-4