"ANALYZING THE PERFORMANCE OF STUDENTS BY VARYING THE PATTERN OF EXAM”

Dissertation submitted in fulfilment of the requirements for the Degree of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

Ramanpreet kaur

11511166

Supervisor

Mr. Kewal krishan

(ASSISTANT PROFESSOR)

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

January-May, 2017

INDIA'S LARGEST UNIVERSITY
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 :** CSERGD0230

Supervisor Name : Kewal Krishan **UID :** 11179 **Designation :** Assistant Professor

Qualification : ________________________ **Research Experience :** ________________________

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Ramanpreet Kaur	11511166	2015	K1519	9465269709

SPECIALIZATION AREA : Database Systems **Supervisor Signature:** ____________________

PROPOSED TOPIC : Data Mining in Education sector.

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	7.25
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	6.75
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.	7.00
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	7.75
5	Social Applicability: Project work intends to solve a practical problem.	7.00
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	7.25

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: **Analyzing the performance of students by varying the pattern of Exam.**

Overall Remarks: Approved (with major changes)

PAC CHAIRPERSON Name: 11011::Dr. Rajeev Sobti **Approval Date:** 22 Nov 2016

ABSTRACT

In this report I have studied about data mining which is the computational process of discovering patterns in large data sets involving methods at database systems.. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

The work depends on expectation examination to foresee the understudy execution on subjective and objective type of exams. Next, I have discussed about the dataset in which how many students passed the subjective exam and how many students passed from objective exam. Two techniques are passed the first technique of clustering and second technique of classification.

In the technique of clustering density based clustering can be applied which calculate the most dense region from the dataset and on the basis of EPS, Euclidian distance final results of dissimilar and similar data get generated in the form of clusters. The final generated clusters will be given as input to classifier to the classified result is generated as pass and failed students.

Next I have discussed that, In the Density based clustering, the EPS values is calculated statically which reduce accuracy of clustering and classification..

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation entitled ""**ANALYZING THE PERFORMANCE OF STUDENTS BY VARYING THE PATTERN OF EXAMINATION**”” 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 Mr. kewal krishan. 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

Ramanpreet Kaur

11511166

SUPERVISOR’S CERTIFICATE

This is to certify that the work reported in the M.Tech Dissertation entitled “"**ANALYZING THE PERFORMANCE OF STUDENTS BY VARYING THE PATTERN OF EXAMINATION”**”, submitted by **Ramanpreet kaur** 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

Mr. KEWAL KRISHAN

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 feel immense happiness for the completion of the Dissertation-II under the stipulated time duration to fulfill requirements and specifications under the guidance of guideposts which acted as lightening pillars to enlighten the way for carrying out Dissertation-II completion steps and due to which I was able to reach the heights of its completion and study.

I feel really grateful to my Mentor Mr. kewal krishan for providing me for all guidance and valuable time, care, support, sincere operation and involvement during all the phases of the Dissertation-II. I thank for the continuous efforts for completing my Dissertation-II within a fixed time.

Without thanking all who have been supporting and helping me throughout the overall reviewing and searching for the Research papers and the sites related to it, this acknowledgement could not be completed. I thank them for their support in ideas, creation and their help in maintaining and supporting. I am really thankful to my University, my Faculty Member’s guidance, without their help the Dissertation-II could not be taken to a feasible state.

TABLE OF CONTENTS

CONTENTS **PAGE NO.**

LIST OF ABBREVIATIONS

S. No.	Abbreviation	Expansion
1.	DBMS	Database Management System
2.	KDD	Knowledge Discovery in Database
3.	DBSCAN	Density- Based Clustering of Application With Noise
4.	MATLAB	Matrix Lab

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Data Mining

Information mining, the extraction of concealed prescient data from extensive databases, is a serious new innovation with the great potential to help organizations focus on the most basic data in their information distribution centers. Information mining apparatuses foresee future patterns and work on, allowing organizations to make proactive, learning driven choices. Data mining strategies are the deferred result of a long methodology of research and thing headway. This improvement started at the point when business information was at first secured on PCs, continued with updates in information get to, and all the all the additionally starting late, made advances that allow customers to research through their information legitimately. Information mining takes this developmental strategy past review information get to and course to arranged and proactive information movement. Data mining methodologies can produce the advantages of mechanization on present programming and hardware organizes, and can be completed on new structures as existing stages are redesigned and new things made. Right when data mining instruments are executed on tip best equivalent get ready structures, they can separate huge records in minutes. Speedier get ready proposes that clients can, therefore, investigate distinctive roads with respect to more models to comprehend complex data. Speedy makes it valuable for clients to separate tremendous measures of data. Greater databases, in this way, yield overhauled desires [1].

The philosophy that is utilized to play out these achievements in data mining is called illustrating. Showing is just the display of building a model in one condition where you know the appropriate response and after that applying it to another situation that you don't. This show of model building is in this way something that people have been completing for quite a while, clearly before the approach of PCs or data mining advancement. What occurs on PCs, in any case, is alongside no not the same as the way people collect models. PCs are stacked up with heaps of information around a grouping of conditions where an answer is known and a while later the data mining programming on the PC must experience that data and distil the characteristics of the data that ought to go into the model. Once the model is produced it can then be utilized as a bit of comparable conditions where you haven't the faintest thought regarding the answer.

While extensive plate guideline headway has been moving unmistakable exchange and symptomatic structures, information mining gives the relationship among both. Information mining programming takes a gander at affiliations and cases in setting without end exchange information in light of open-finished customer ask. A few sorts of educational written work PC projects are available: quantifiable, machine learning, and neural structures.

For the most part, any of four sorts of affiliations are hunt down:-

- **Classes:** Secured record is used to find information in foreordained social affairs. For example, a cafeteria framework could mine customer acquire proof to pick when purchaser visit and what they generally arrange. This piece of information could be used to collect development by having very much requested specials.
- **Clusters**: Information things are gathered by affiliations or buyer inclines. For example, information can be mined to see propel zones or purchaser affinities.
- **Associations**: Information can be mined to perceive affiliations. The blend diaper case is an event of natural mining.
- **Sequential patterns**: Information is mined to theorize coordinate outlines and cases. For instance, an outside outfit retailer could imagine the probability of a rucksack being acquired in context of a buyer's buy of resting sacks and climbing shoes.

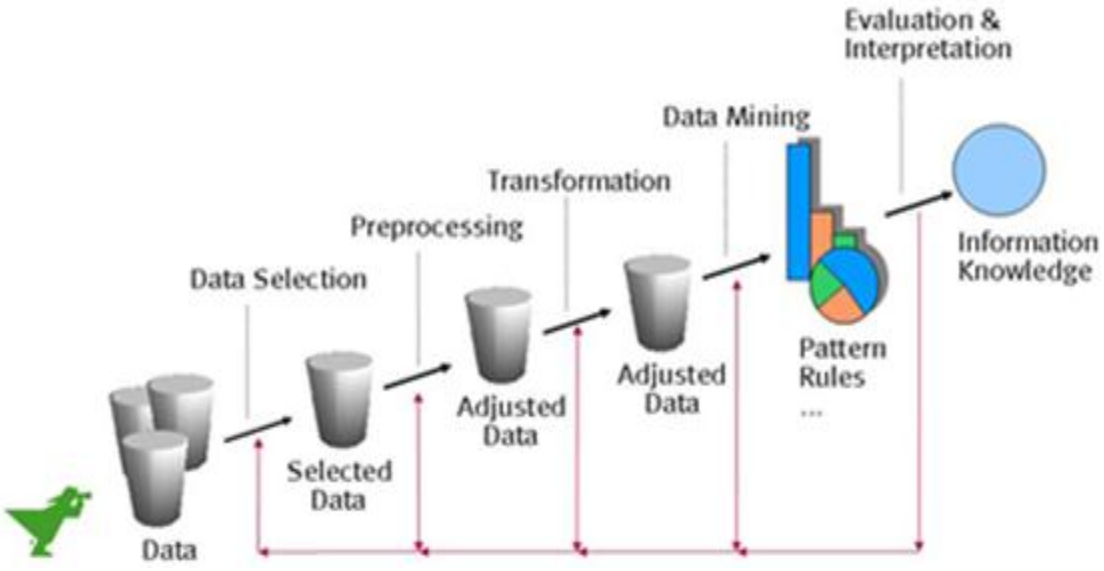


Fig1.1:-Data mining model

1.2 Data Clustering

There are various computing applications which involve the data analysis either within the design phase or within the on-line operations involved in the processing. On the basis of the types of models which are available for the data sources, the data analysis can be categorized The main aim is to make sure that the data is grouped or classifying the measurements on the basis of certain factors such as:

i. How good it is for the postulated model.

ii. Natural groupings did on the basis of analysis

The organizing of patterns which are gathered to form clusters on the basis of similarity of objects is known as cluster analysis. These patterns are mostly in the form of a direction of dimensions or a idea in the multidimensional area. The patterns which belong to the similar cluster have similar properties and are much alike as compared to the ones that are from the different clusters. There have been a lot of techniques proposed for the representation of data, measuring the proximity of the data elements as well as grouping the data elements which have further resulted in forming various clustering methods [1].

There is a lot of difference between the clustering which also comes under the unsupervised classification as well as the discriminant analysis which refers to as the supervised classification. The collection of labeled or pre-classified patterns is done within the supervised classification. Here, the main focus is to highlight the newly entered pattern which has no label. The descriptors of classes are learned from the given labeled patterns and thus the newly entered patterns are further labeled. The main objective of bundling is to collection the gathered unlabeled designs within meaningful clusters. The clusters are related to the labels. However, these type labels are data driven which means that they are directly achieved from the data. There are various fields in which the clustering method is used such as the design study, bunching, conclusion making as well as machine-learning conditions which involve information mining, image separation, design arrangement and countless others. There is, however a lack of knowledge regarding the data. There are very fewer assumptions proposed by the decision-makers as well. The relationships amongst these data points are made through the clustering methodology for making a proper assessment of their design.

For the purpose of grouping the unlabeled data, the clustering method is used in various applications. There are different terminologies as well as assumptions related to the modules of the clustering route as well as the environment in which the clustering has to be utilized. There are hence, a lot of predictions being made related to this.

1.2.1 Components of a Clustering Task

Typical pattern clustering activity involves the following steps:-

a. Pattern representation (optionally including feature extraction and/or selection):-

There are numerous classes, patterns as well as features of different numbers, types, and scales which are involved in the clustering algorithms. These all are included within the pattern representation of data. The practitioner might not be able to control all such information alone. The procedure of recognizing the furthermost current subsection amongst the unique structures for utilizing in the bundling is known as the feature selection. For obtaining an suitable set of types in clustering, these techniques can be used [2].

b. Definition of a pattern proximity measure appropriate to the data domain: -

The distance function which is sharp on the couples of configurations is measured with the help of pattern proximity. At different locations, different distance measures are utilized. The dissimilarity amongst two patterns can be determined with the help of an easy space measure such as Euclidean distance. However, the conceptual likeness amongst the patterns can be determined with the help of other various similarity measures.

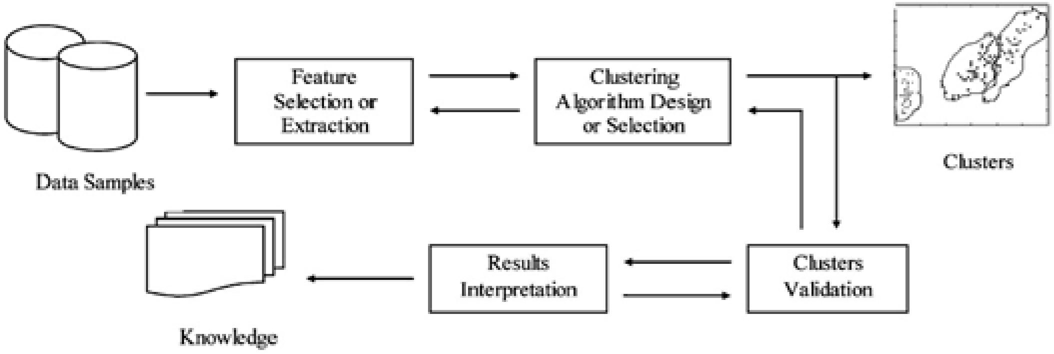

Data Samples
Feature Selection or Extraction
Clustering Algorithm Design or Selection
Clusters
Knowledge
Results Interpretation
Clusters Validation

Fig no 1.2.1:- components of clustering classification

c. Clustering or grouping:-

There are two ways in which the grouping task can be outperformed. The first is the one in which the data is partitioned into groups and the output achieved here is the clustering. The second is fuzzy in which every one form has a different unit of relationship for respectively of the collections achieved. The merging or splitting of clusters on the basis of similarity is done which helps in producing a nested series of partitions in the case of hierarchical clustering. The area which optimizes the clustering criterion is recognized with the help of partition clustering algorithms. There are various techniques which also help in grouping the data such as the probabilistic and graph-theoretic grouping techniques.

d. Abstraction of data:-

The procedure in which simple and compacte symbol of record set are extracted is known as data abstraction. In view of the automatic analysis or the human-oriented method, this method terms to be simple. The details of each cluster are held within the data abstraction in clustering. The description might be in positions of bunch originals or representative decorations, for instance, centroids.

e. Assessment of output:-

There are some better clusters than the rest of the clusters within the data. There are various facets for assessing the output of the clustering technique. Instead of providing the clustering algorithm itself the data domain is assessed. The clustering algorithm should not be used for the data which does not have any clusters in it. The input data is examined to check for any advantages related to cluster analysis which is currently being processed. This is however not that important. The assessment of the output achieved from clustering method is known as the cluster validity analysis. Optimality is mostly used here for the analysis purposes for the related data [3].

1.2.2 Common application domains for clustering:

Intermediary Phase aimed at further vital data mining harms: The solution to most of the data mining issues for instance organization 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.

I. **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.

II. 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.

III. 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 are easier.

IV. 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.

V. Multimedia Data Analysis:-

The multimedia data involves the images, audio, video and various types of documents. There are huge solicitations for example recognition of comparable leftovers of melody, 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 [4].

VI. Biological Data Analysis:-

Due to the evolvement of the human genome as well as various manners of genetic factor look records, the genetic documents 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.

VII. Social Network Analysis:-

For determining the important communities within the network, the structure of the social network is utilized. Within the community detection, there is a improved 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.

1.2.3 Data clustering categories

I. **Technique-centered:-**

The techniques such as feasibility methods, space-based techniques, spectral techniques, density-based techniques, and dimensionality-reduction founded procedures utilize clustering mechanism within them. There are various merits and demerits as per different situations as well as problem domains. There are some specialized techniques to be involved within the various record forms including extraordinary dimensional record, large record or streaming record as they have various challenges of their own.

II. **Data-Type Centered:-**

There are various data types animated to different applications which also have their own distinct properties respectively. For instance, it can be seen that the ECG machine provides time series data points which are related to each other on differing closer base. However, the jumble of article and fundamental records is given by the social network. The instances of this are a resounding record, time series data, discrete sequences, network data, and probabilistic data. The selection of which type of methodology is to be used for clustering completely depends on the type of data present. Due to the departure among various types of features for example behavior or contextual features, there are data types which are not easy to be performed.

III. **Additional Insights from Clustering Variations:-**

For the kinds of clustering variations, the different insights are designed. For instance, visual analysis, supervised analysis, ensemble-analysis, or multiview analysis can be recycled in demand to increase supplementary visions. From the view of providing certain insights related to the performance of clustering, the problem of cluster validation is necessary [5].

1.3 Common Techniques Used in Cluster Analysis

There are various types of methods to be proposed for solving the issues related to clustering. There are specific techniques to be proposed for the pre-processing phase as well. The commonly utilized techniques are explained below:

a. Feature Selection Methods:-

For the purpose of enhancing the quality of the clustering method, the most important step involves in pre-processing is the feature selection phase. All of the features do not contribute equally to determining the clusters. There are some clusters which are noisier as compared to others. The utilization of a preprocessing phase is thus important where the noisy, as well as irrelevant features, are extracted from the contention. There is a close relation between the feature selection as well as dimensionality reduction. The selection of original subsets related to the features is done through the feature selection. There might be the usage of linear combinations of features in the case of dimensionality reduction such as principal component analysis for enhancing the feature selection effect. The greater interpretability is the foremost advantage which is followed by the reduced amount of converted instructions which are utilized for the presentation method. For the purpose of achieving better locality specific insights, the feature selection can be integrated into clustering algorithm directly. In a case where various features are related to the various localities of data, this method can be useful. The failure of global feature selection algorithms results in a formation of high dimensional subspace clustering algorithms. Related to the various set of dimensions within the real data examples, there are various points which are correlated. The pruning of a large number of dimensions at same instant is not feasible and can also result in the loss of information. Thus, the best way of achieving this objective is to utilize local feature selection through the integration of feature selection process within an algorithm. The extension of such local features can be done through the dimensionality reduction issue and can be named as local dimensionality reduction [6].

b. Probabilistic and Generative Models:-

The main objective of modeling the data with a generative process is involved within the probabilistic models. In the beginning, the assumption of a specific form of the generative model is done which is followed by the estimation of parameters by the support of Expectation-Maximization (EM) algorithm. The estimation of parameters is done with the help of present data set in such a manner that the maximum likelihood is appropriate for the generative model. The estimation of underlying data points is done with the help of generative probabilities within this model. The high fit probabilities are provided by the data points which fit the distribution in a proper manner. However, low fit probabilities are given by the anomalies. On the basis of whether the prior probabilities are determined related to the problem setting or whether in a component of a mixture, the inter-attribute correlations are predicted, the models of various flexibilities are designed. There is a direct circular dependency of the model parameters as well as the chance of mission of documents opinions to the related bunch. For the purpose of resolving the circularity, the iterative techniques are required. The EM technique is used for solving the generative models. This process begins with the random of heuristic initialization and is followed by the iterative steps for resolving the circularity which is mentioned below on the basis of their properties:

- (E-Step): This step determined the estimated chance of project of record ideas to a bunch. The recent classical factors are utilized for this.

- (M-Step): This step limits the ideal classic limitations of every combination. The assignment probabilities are utilized as bulks for this.

For different kinds of data, there can be an easy generation of the EM-models. This is a very good property of these models as long as there is a proper selection of the reproductive model for every module for the specific jumble element. The most fundamental models amongst the various clustering models are the generative models as the attempt to understand the basic process which can help in generation of a cluster. The clustering methods, as well as the generative models, have various connections with each other which involve special cases related to the prior probabilities of the mixture parameters [7].

c. Distance-Based Algorithms:-

For the purpose of reducing the algorithm to distance-based algorithms, there are various forms of generative algorithms formed. The separation task within the possibility sharing is utilized for the mixture components within the generative models. For instance, the data generation chances in terms of Euclidian distance from the nominate combination are represented using the Gaussian distribution. Therefore, there is a very close relationship amongst the Gaussian distribution conjointly the k-means algorithm. The reduction or simplification of various kinds of generative models can be done through the distance-based algorithms. There are two types of distance-based algorithms which are explained below:

i. 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 record ideas are assigned to the cluster, the representative is adjusted. The iterative nature of EM algorithm is compared with this technique as there are soft 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 procedures for the purpose of grouping 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 dataset, the partitioning representatives are not drawn in k-means technique. There is the 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 subdividing agents are drained from the unique record in the case of k-Median technique which is otherwise also known as k-Medoid technique. However, both these techniques are not the same and have variations within them [8].

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 the mean and median of a fixed of the web or distinct 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.

ii. Hierarchical:- The representation of clusters in these methods is done at various levels f granularity, using the dendrogram. The representation of the system can be either agglomerative or divisive on the basis of hierarchical representation. It can be either bottom-up or top-down 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 as sole-linkage, each-pairs linkage, centroid-linkage, and tested-linkage grouping 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 examine of files themes amongst dual bunch is utilized in sampled linkage. This is used to calculate the average distance of the two clusters. The distance between the centroids is utilized for the centroid-linkage process [9].

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 classified form of the tree as well as the elevation of stability within various clusters, the divisive partitioning is allowed flexibility. In the expression of the depths of various intersection 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.

iii. Density- and Grid-Based Methods:-

The dual closely linked classes are the density and grid-based techniques. Here, the data space is explored at higher levels of granularity. In relations of an amount of record points in certain defined bulk of its section or in terms of suffocate kernel density estimate, the compactness at a specific opinion within the facts interplanetary is well-defined. At a 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 areas of the documents 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 support in significant the clusters on the subsets of dimensions. The record space is explored at the 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 random records categories such as the time-series data within the density-based methods. In a case of higher dimensionality, the density computations are extremely difficult to define due to the higher digit of cells within the grid formation along with the presence of the sparse record in the grid available [10].

1.4 Density-based spatial clustering of applications with noise (DBSCAN)

The various types of grouping methods developed here are partitioning, hierarchical, density, grid, model, and constraint established. On the basis of the notion of density, the density based method works. There is a difference between the clusters formed in thick regions as well as thin regions. The objective here is to increase the recognized clusters until the density in the neighborhood is higher than the threshold value.

For the purpose of finding the arbitrary shaped clusters and differentiating the noise from huge spatial databases, the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm is utilized. There are two parameters in this algorithm. They are Eps (radius) and the MinPts (minimum points-a threshold). On the basis of center-based approach, this method is based, Here, the density is assessed for a specific point within the information set. The numbers of points present within a specific radius, Eps are counted. Further, the points are classified into categories such as core point, border point as well as a noise point. The objective here is to attain less minimum number of points (MinPts) by the neighborhood of a given radius (Eps) of each point of a cluster.

Algorithm Description

1. Choose a random point p

2. Attain all points which are density-reachable from p with respect to Eps and MinPts

3. A group is formed if p is a core point

4. Visit the next point of the information set, if p is a border point and none of the points are density-reachable from p

5. Recurrence the overhead process until all of the points have been observed [11].

Spatial data mining works on spatial data. The recognition of interesting similarities of characteristics and patterns within the huge spatial data sets is known as spatial data mining. As per the given nature of data sets, the various possible trends and clusters are to be gathered in the spatial clustering technique. The measurement of density is done in the form of points by totaling the quantity of themes within a region of determined radius surrounding the point in this technique. The clusters are formed with the help of points which have certain threshold vale and densities. The selection of clustering attributes, detection f noise with various densities as well as huge difference of values of border objects within the opposite directions of similar clusters are important concerns within the DBSCAN algorithm. At least once, the point of any object is to be visited. If the point is a candidate of various clusters only then it can be visited multiple times. There are various efforts being made by the government agencies, scientific fields as well as other private sectors for gathering the large data sets of the spatial features as the collection of data is a very important task. New and proceed techniques of mining and knowledge discovery are required for the moving objects as well as the dynamic data present in the multidimensional data.

1.4.1 Classification of DBSCAN algorithm

There are various categories in which the DBSCAN clustering algorithm can be categorized:

a. Partitioning based DBSCAN clustering;

b. Grid-based DBSCAN clustering;

c. Hierarchical DBSCAN clustering;

d. Detection Based DBSCAN clustering,

e. Incremental DBSCAN clustering

f. Spatial–temporal DBSCAN clustering.

a. Partition based DBSCAN clustering:-

There are M clusters generated within this technique where each object belongs to one cluster. A centroid or a bunch descriptive is used for representing every cluster. A small description of various objects present in a cluster is provided here. On the basis of the type of object being clustered, the relative description is providing here. The arithmetic mean of the feature paths for various things in a gathering is given where real-cost record is present. An appropriate representative is given through this. In other cases, the alternate types of centroids might be involved. For instance, a list of keywords which can be seen in minimum number of documents in a cluster in case where there are clusters of documents. When there are large numbers of clusters, the centroids are to be clustered for producing hierarchy in a dataset.

b. Grid-based DBSCAN clustering technique:-

The data set is quantized into numerous cells and then further work along with the objects that belong to those respective cells. The points are not relocated here however the various hierarchical levels of groups of objects are created. The distance measure is not a factor which affects the merging of grids and further the clusters. A predefined parameter is used for the determination here [12].

c. Hierarchical DBSCAN clustering:-

A hierarchy of clusters is built using this method. The Lance-Williams formula is utilized for the basics of hierarchical clustering. It also involves the idea of conceptual clustering. Gradually the hierarchical algorithms help in building the clusters. There are two types of strategies for the hierarchical clustering. A particular cluster is not used for the partitioning the data in single step in the hierarchical clustering. There is a sequence of dividers which are performed from one only gathering which involves very substances to n gatherings each of which has a solo object. There is a subdivision of hierarchical clustering into agglomerative techniques. Here, the succession of combinations of n articles into clutches and troublesome systems is done which distinguishes the n objects into certain smaller groups. The more prominently utilized techniques are the agglomerative techniques. The dendrogram is used for representing the two-dimensional diagram of the hierarchical clustering.

d. Detection Based DBSCAN clustering technique:-

On the discredited likelihood function, the simplified detection issue is resolved. The maximum likelihood number of sides as well as orientation at most likely polygons is done with the help of this efficient algorithm. A discrete Hough-based algorithm is represented as the initial step for the detection. The approximation of complete likelihood function for recovering the orientation and number of slides is performed in the second stage.

e. Incremental DBSCN clustering algorithm:-

The dynamic databases are handled using this algorithm which also can change the radius threshold value in dynamic manner. The number of final clusters can be restricted and an original dataset can be read only once at a time. Also using this algorithm the frequency information of the attribute values is provided which can further be utilized for the categorical data. Not only the impact of the inadequate memory can be overcome here but also the accuracy can be improved for reflecting the properties of the dataset present.

f. Spatial-Temporal DBSCAN clustering:-

For the purpose of storing and clustering huge spatial-temporal data, a new clustering algorithm is proposed. An integration of environmental data from various sources as coverages, grids, tables etc is done here. For the purpose of developing special functions for data integration, data conversion, management etc. there are some distinct functions developed. The imperfection users were allowed for developing user-friendly interfaces for the operation of a system [13].

g. Spatial–temporal DBSCAN algorithm:-

The indexing and retrieval of spatial-temporal data is done as per the spatial and time dimensions. The validity or the time duration in which it is stored within the database is known from the time dated committed to the spatial data. The effective period, transaction time or both can be supported by the temporal database. In accordance with the real world, the valid time is denoted with the time period that is true. The time during which the fact is stored in the databank is known as the operation period. The lawful time for the temporal data is the highlight here.

1.4.2 Applications of DBSCAN algorithm

The WEKA is a software program which utilizes the DBSCAN algorithm. The practical application so DBSCAN algorithm are given below:

a. **Satellites images:-**

   From the satellites present all around the world, there is huge amount of data gathered. The data is to be converted into efficient information. For example, the satellite images taken from the forests, water as well as mountains are to be converted. There is some work which is to be performed in image processing at first. Only then it can be classified into three elements

b. **X-ray crystallography:-**

All the atoms within the crystal are located using the X-ray crystallography. This further generates huge amount of data. The atoms can be found and classified within the data with the help of DBSCAN algorithm.

c. **Anomaly Detection in Temperature Data:-**

In cases such as credit fraud, health condition, etc. the pattern anomalies within the data is focused upon. The anomalies within the temperatures is measured which is related to the environmental changes. The errors within the equipment can also be identified here. The situation can control and detected using these unusual patterns. The patterns are discovered in the data in the DBSCAN algorithm [14].

1.5 Classifiers

a. Support Vector Machine (SVM) Classifier

For regression, classification as well as general pattern recognition, the SVM classifier is proposed. Due to its high generalization performance without requiring any priori knowledge to add in it, this classifier is considered to be good in comparisons to other classifiers. The performance is even better when the measurement of the input space is extremely great. In order to differentiate between the dual classes of the training data, the SVM requires identifying the top classification task. The best classification function metric can be represented geometrically as well. The hyper plane f(x) is separated from the linear classification function for the linearly separable dataset. This hyper plane permits through the center of dual classes which can be said to separating them. The newly record instance xn is categorized by testing the sign function function f(xn); xn which fits to the positive class if f(xn)> 0. This is done after the determination of a new function.

It is the main objective of SVM to determine the best task by exploiting the margin among the two periods. This is due to the fact that there are various such linear hyper planes. The amount of space or distance amongst two classes is known as the hyper plane. The shortest among the near information points to a point on the hyper plane is known as margin. This can further help us in defining the way to extend the margin which can help in selecting only a few hyper planes for the solution to SVM even when so many hyper planes are available [15].

The objective of SVM is to produce a linear function which can help in identifying the target function. This can further help in extending the SVM for performing regression analysis. The error models are of hushed help here for the SVRs. In a case when the alterations among the predicted and actual values are within an epsilon cost, the mistake is to be defined as zero. In the off chance, there is a linear growth in the epsilon- insensitive mistake. Through the reduction of Lagrangian, the support vectors can be studied. The insensitivity to the outliers can be of benefit for the support vector regression. The demerit of SVM is that the computations are not efficient enough. There are many solutions proposed for this. The breakage of one big problem into numerous numbers of smaller problems is one way to solve this issue. There are only some selected variables for the efficient optimization for each problem. Until all the problems are solved eventually, this process keeps working in iterative nature. The problem of learning SVM is to be solved also by recognizing the approximate minimum encircling droplet of a usual of occurrences in the program.

b. Naïve Bayes Classifier

The objective here it to propose a rule which allows assigning the future objects to a session when a set of objects is given for each class. The future objects are described here by the given vector of variables only. These types of problems are also known as the difficulty of direct classification and various technique has been proposed for developing rules for them. Pone very chief single is the naïve Bayes process—known as idiot's Bayes, simple Bayes, and independence Bayes. There are various reasons for this. The construction of this method is very easy and also does not require iterative parameter estimation methods. This makes it applicable to huge data sets. The classification made by it is understood as the interruption is easy here are so the users which are unskilled can clearly justify its causes.

c. Decision Tree Classifier

Choice Trees (DTs) are a non-parametric managed adapting path used for order and relapse. The territory is to make a graph that figures the charge of an neutral variable by taking in basic choice rules accumulated from the records highlights. It is a procedure for resembling separate-valued objective function, in which the learned capacity is spoken to by a decision tree. Decision trees categorize instances by arrangement they downcast the tree from the starting point to some needle node, which gives the arrangement of the occasion.

Every clot in tree indicates a trial of a certain attribute of the occurrence, and each appendage descending from that coagulation connection to a solitary of the plausible qualities for this property. An event is characterized by starting at the seed hub of the tree, analyze the point recognized by this hub, then moving ground floor the tree limb identifying with the cost of the trait. This strategy is then rehashed for the subtree established at the new hub [16].

Choice tree learning is a technique for the most part utilized as a bit of information extraction. The motive is to make a graph that expects the approximation of an objective variable in perspective of a couple input factors. Each inner hub connections to one of the info factors; there are limits to youngsters for each of the possible standards of that information variable. Each leaf signifies a cost of the objective variable given the estimations of the information factors addressed by the course from the assurance to the leaf. A decision tree is an essential depiction for masterminding cases. For this section, the bigger parts of the segments have constrained discrete regions, and there is a solitary target highlight called the request. Every segment of the space of the portrayal is known as a class. A decision tree or a portrayal tree is a tree in which each inside (non-leaf) center point is separate with an information incorporate. The indirect bits beginning from a center point named with a component are separate with each of the possible estimations of the segment. Each leaf of the tree is separate with a class or a probability apportionment over the classes.

d. K-Nearest neighbor

K-Nearest neighbor classifiers be established on learning by comparability. The arrangement tests are portrayed by n-dimensional numeric attributes. Every example play out a point in a n-dimensional space. In this manner, the higher bit of the planning tests is secured in a n-dimensional illustration space. Precisely when given a dark illustration, a k-nearest neighbor classifier scans the case space for the k get ready tests that are closest to the dark example. "Closeness" is characterized as far as Euclidean separation. Not under any condition like choice have tree acceptance and back spread, closest neighbor classifiers relegated equal the initial investment with weight to every property. This may accomplish perplexity when there are different inconsequential components in the records. Nearest neighbor classifiers can comparably be utilized for figure, that is, to give back a really regarded fancy for a given new delineation. For this situation, the classifier gives back the standard estimation of the genuine blame related for the k nearest neighbors of the odd model. The k-nearest neighbors' figuring is among the minimum troublesome of all machine learning counts [17].

CHAPTER 2

LITERATURE REVIEW

Huan Yu, et.al," DBSCAN Data Clustering Algorithm for Video Stabilizing System", 2013

This paper proposed a technique in perspective of DBSCAN information bunching calculation to balance out the jitter of advanced video with moving articles in it. Remembering the ultimate objective to perceive the corners on moving articles with those on foundation, in the wake of removing the sides of each edge, DBSCAN calculation was used to group each one of the corners by bunching their movement vectors' lengths and bearings [18]. By then, the scattering of each bunch is contrasted and avowing whether the corners in each group were had a place with moving articles or foundation. Recreation exploratory outcomes exhibited that the proposed strategy had great adjustment effects to balance out jitter in a video grouping with moving articles in it. The trials had exhibited that the calculation proposed by this paper is better than the calculations proposed by various scientists. It can precisely perceive corners on moving items and foundation. Besides, it furthermore can be associated in conditions that few moving items exist in the meantime or the moving articles' ranges are expensive.

Nagaraju S, et.al," A Variant of DBSCAN Algorithm to Find Embedded and Nested Adjacent Clusters", 2016

In this paper, a proficient approach for clustering examination is proposed to distinguish inserted and settled contiguous groups using the possibility of thickness based thought of bunches and neighborhood distinction [19]. In a general sense our proposed calculation is enhanced adaptation basic DBSCAN calculation, proposed to address the bunching issue with the usage worldwide thickness parameters in basic DBSCAN calculation and issue of distinguishing settled neighboring groups in an EnDBSCAN calculation. Our test comes about that suggested that proposed calculation is more viable in distinguishing implanted and settled neighboring groups pondered both DBSCAN and EnDBSCAN without including any extra computational unpredictability. Also, we have the preset strategy to assess the worldwide thickness parameters using sorted k-separate plot and first demand subsidiary. Through this paper, the idea of thickness based methodologies for information bunching and considered neighborhood distinction is used viably distinguish installed and settled adjoining groups. Our trial comes about suggested that proposed calculation compelling in recognizing settled nearby groups diverged from DBSCAN and EnDBSCAN calculation with computational many-sided quality as same as DBSCAN calculation.

Jianbing Shen, et.al," Real-time Superpixel Segmentation by DBSCAN Clustering Algorithm", 2016

In this paper, a constant picture super-pixel division strategy is proposed with 50fps by using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) calculation. Remembering the true objective to diminish the computational expenses of super-pixel calculations, we receive a snappy two-arrange system [20]. In the essential grouping stage, the DBSCAN calculation with shading similitude and geometric limitations is used to rapidly bunch the pixels, and a while later little groups are converged into super-pixels by their neighborhood through a separation estimation characterized by shading and spatial elements in the second consolidating stage. A hearty and clear separation limit is characterized for improving super-pixels in these two stages. The test comes about show that our steady super-pixel calculation (50fps) by the DBSCAN grouping outflanks the condition of-the-craftsmanship super-pixel division strategies to the extent both precision and effectiveness. In future work, we will show signs of improvement conservativeness of superpixels by working up another DBSCAN calculation that has the worldwide ideal property. We moreover plan to extend the current superpixel system to realtime video super voxel division for keeping up spatiotemporal conservative shapes.

Saefia Beri, et.al," Hybrid Framework for DBSCAN Algorithm Using Fuzzy Logic", 2015

Information mining procedure is to get data from an informational index and after that change over it into a reasonable and vital data for further use. DBSCAN, a thickness based bunching calculation, perceives groups of moving shape and anomalies. DBSCAN relies on upon bivalent rationale. Therefore, it can simply identify protests as thoroughly having a place with a specific group or not totally having a place with it [21]. In this paper, a system of a philosophy of DBSCAN calculation with the mix of fluffy rationale is proposed. The degree to which a protest has a place with a specific bunch will be made plans to use participation values. The enhanced variant of DBSCAN calculation will be the hybridization of DBSCAN calculation with fluffy if-then guidelines. Information mining procedure is to get data from an informational collection and after that change over it into reasonable and imperative data for further use. Along these lines, it can simply identify protests as absolutely having a place with a specific bunch or not by any stretch of the imagination having a place with it. In this paper, a structure of the approach of DBSCAN calculation with the incorporation of fluffy rationale is proposed. The degree to which a protest has a place with a specific bunch will be set out to use enrollment values. The enhanced form of DBSCAN calculation will be the hybridization of DBSCAN calculation with fluffy if-then standards.

Bharathi S, et.al," Automatic Land Use/Land Cover Classification using Texture and Data Mining Classifier", 2013

Nowadays wherever remote detecting pictures are used for a wide arrangement of employments, making of mapping things for military and regular applications, appraisal of characteristic damage, checking of land use, radiation watching, urban orchestrating, improvement control, soil assessment, and item yield examination. A couple number of picture order calculations have demonstrated great exactness in arranging remote detecting information. A productive classifier is required to order the remote detecting symbolisms to concentrate data [22]. The surface based managed characterization is used here. Here the technique dissected particular grouping strategies. KNN, SVM and Neural system are used. All the three classifier gives great outcome yet neural system classifier takes quite a while, the time intricacy is high. Arrive use mapping has been done by contrasting the pictures and region of the land used is computed. Neural Network gives great order exactness, nonetheless, time unpredictability is high. Arrive use mapping has been done by taking pictures of two unmistakable days and age and moreover zone of land use is ascertained. The confirmation is done by field study. Real range of land used is 192310 m2. Street extraction is bad with this calculation and in instances of land use mapping if the determination of the picture changes arrive used zone is not exact. The paper is attempting to improve these features.

Sneha Chandra, et.al," Enhancement of Classification Accuracy of our Adaptive Classifier using Image Processing Techniques in the Field of Medical Data Mining", 2015

Medical Data Mining is a standout amongst the most difficult fields of Data Mining. The best test lies in classifying the diseases with high classification accuracy. In this exploration work, image processing techniques have been utilized on the advanced version of our Adaptive Classifier, to produce categories for the attributes of sample medical datasets. The propelled variant of our Adaptive Classifier has been produced using the systems of Clustering Data Mining in conjunction with Classification Data Mining [23]. The proposed approach works upon the example therapeutic datasets and thinks about the aftereffects of our Adaptive Classifier with the consequences of its constituent classifiers. The trial comes about created exhibited higher grouping precision for our Adaptive Classifier, which has properly stirred the interest required for further examination. Our Adaptive Classifier could fulfill over 90% characterization precision on the specimen therapeutic datasets which have fittingly stimulated the interest required for further examination. The Adaptive Classifier aims to give a substitute method of disease detection, since the diagnostic tests are costlier, and a large portion of the times, require surgical intervention and post-operative care (here, biopsy in the cases of Liver Cirrhosis and Lymph Node Tuberculosis).

Yomna M. El Barawy, et.al," Improving Social Network Community Detection Using DBSCAN Algorithm", 2014

Informal organizations depict the coordinated efforts between individuals or components and are spoken to by a diagram of interconnected hubs. The investigation of such diagrams prompts cognizance of this information and completing different groups. Among the different bunching calculations, DBSCAN is a viable unsupervised grouping calculation which is actualized in this work to accentuate group discovery in the informal organization [24]. The outcomes decide the amount of high impact individuals spoken to by center, less impact spoken to by fringe and individuals with no impact in the gatherings spoken to by anomalies. By dispensing with the exceptions the dataset will be without commotion to manage it. The DBSCAN calculation was intended to find the anomalies in datasets. In this examination, the DB SCAN calculation is used to choose the irregularity hubs by changing the range (epsilon) of the bunch. Informal community examination most focuses on centers as they have the impact on various individuals. So that wiped out peculiarity individuals prompts an exact bunching result that helps with the group location issue in the informal community investigation field. In future, applying the proposed strategy with different group calculations and contrasting it and the aftereffect of Grievant-Newman bunching Algorithm is arranged.

Dominik Fisch, et.al," Knowledge Fusion for Probabilistic Generative Classifiers with Data Mining Applications", 2013

If information, for instance, characterization tenets are removed from test information distributedly, it may be critical to join or circuit these guidelines. In an ordinary approach this would frequently be done either by joining the classifiers' yields (e.g., in type of a classifier group) or by consolidating the arrangements of order guidelines (e.g., by weighting them solely). In this article, another method for combining classifiers is presented at the level of parameters of order guidelines [25]. This system depends on the usage of probabilistic generative classifiers using multinomial dispersions for clear cut info measurements and multivariate run of the mill disseminations for the ceaseless ones. That infers, we have conveyances, for instance, Dirichlet or run of the mill Wishart disseminations over parameters of the classifier. We suggest these disseminations as hyper appropriations or second-organize dispersions. We show that combining (no less than two) classifiers ought to be conceivable by expanding the hyper-disseminations of the parameters and decide clear equations for that undertaking. Properties of this new approach are exhibited with several examinations. The principle preferred standpoint of this combination approach is that the hyper-conveyances are held all through the combination procedure. In this way, the combined segments may, e.g., be used as a piece of taking after preparing steps (internet preparing).

Dianwei Han, et.al," A novel scalable DBSCAN algorithm with Spark", 2016

In this paper, another parallel DBSCAN algorithm is shown utilizing the new big data framework Spark. With a particular true objective to lessen seek time, KD-tree is connected in this calculation [26]. More especially, we propose a novel way to deal with avoiding correspondence between agents so we can locally gain incomplete bunches more proficiently. In view of Java API, proper information structures are chosen exactly: Utilizing Queue to contain neighbors of the data point, and utilizing Hashtable while checking the status of and setting up the data centers. What's more, other propelled components are used from Spark to make our usage more compelling. The calculation is executed in Java and assesses its versatility by using an unmistakable number of preparing centers. Our investigations show that the calculation we propose scales up to a great degree well. Utilizing instructive accumulations containing up to 1 million high-dimensional concentrations, it is demonstrated that our proposed figuring completes speedups up to 6 utilizing 8 focuses (10k), 10 utilizing 32 focuses (100k), and 137 utilizing 512 focuses (1m). Another examination utilizing 10k data centers is driven and the result shows that the count with Map Reduce satisfies speedups to 1.3 utilizing 2 focuses, 2.0 utilizing 4 focuses, and 3.2 utilizing 8 focuses.

Md. Rejaul Karim, et.al," An Adaptive Ensemble Classifier for Mining Complex Noisy Instances in Data Streams", 2014

Ongoing information streams arrangement is a trying information mining errand. Progressively spilling conditions, thoughts of occasions may change at whatever point, for instance, climate forecasts, enormous and interference identification et cetera. To address this issue, we present a versatile troupe classifier for information streams order, which uses an arrangement of choice trees for mining complex uproarious occasions in information streams [27]. The group display refreshes actually with the objective that it speaks to the most recent thoughts in information streams. In each accentuation, the gathering model creates another preparation information from the first preparing dataset, then structures a choice tree using new preparing information and doles out a weight to the tree in light of its order precision on unique preparing examples. In like manner, it revives the largeness of preparing examples in preparing dataset. We tried the execution of the proposed gathering classifier against that of existing C4.5 choice tree classifier using genuine benchmark datasets from UCI (University of California, Irvine) machine learning storehouse. The test comes about exhibit that the proposed troupe classifier demonstrates magnificent adaptability and heartiness in information streams order.

Karlina Khiyarin Nisa, et.al," Hotspot Clustering Using DBSCAN Algorithm and Shiny Web Framework", 2014

Woodlands bursts are an imperative issue that happens more than once in Indonesia. Spirit occasions can be normal by viewing the record set of hotspots which are verified through isolated distinguishing cable. This audit expects to make a net tender that accomplishes gathering on the hotspots documents. They claim executes DBSCAN count utilizing Glossy trap structure for R software design vernacular. Gathering is completed on a dataset of hotspots on Kalimantan Island and South Sumatra Province in 2002-2003 [28]. The banquet case of hotspots come to fruition by this gathering can be utilized as a canny prototypical of timberland flares occasion and can be gotten to complete the web program. This examination developed an online application gathering with DBSCAN computation utilizing the R programming vernacular with a Shiny framework. DBSCAN needs minPts and Eps parameter. The greater estimations of minPts will convey less, yet added the amount of upheaval. While the greater estimation of Eps will realize less clusters. MinPts parameter confirmation is finished by observing estimations of the record and strategy the outline of minPts and the amount of packs in addition tumult. While Eps parameter affirmation is gotten from k-plate graph perception and the grade refinement cunnings.

Negar Riazifar, et.al," Retinal Vessel Segmentation Using System Fuzzy and DBSCAN Algorithm", 2015

Retinal vessel division utilized aimed at the primary assurance of retinal ailments, for example, hypertension, diabetes, and glaucoma. There occur a couple of procedures for dividing veins from retinal pictures. The motivation behind the paper is to isolate the retinal container division in perspective of the packing figuring DBSCAN depending upon a thickness based considered gatherings which is expected to discover groups of optional shape. DBSCAN requires just a single data parameter and a motivating force for this parameter is proposed to the client [29]. The execution of a computation is considered and eviscerated utilizing different measures which join affectability and specificity. The specificity and affectability of this system are ۹5.36 and ۷3.82 autonomously. The DBSCAN figuring deals with every one of the issues when utilizing gathering systems, finds the correct information parameters, limits packs of optional figures and does the entire methodology in a sensible time. The execution of the estimation in this paper is best by and large of the past ones. The new game plan diminishes the ideal time without moving time and augments the rapidity and precision of division.

Ilias K. Savvas, et.al," Parallelizing DBSCAN Algorithm Using MPI", 2016

The most recent years, gigantic groups of material are extracted by computational arrangements and electrical gadgets. To misuse the determined measure of information, new inventive calculations obligation be utilized or the developed ones can be different. A champion among the most entrancing and beneficial strategies, remembering the ultimate objective to find and concentrate data from information archives is bunching, and DBSCAN is a popular thickness created calculation which groups information concurring its qualities [30]. In any case, its principle impediment is its outrageous computational multifaceted nature which exhibits the strategy especially missing to apply on enormous datasets. Despite the fact that DBSCAN is an amazingly especially mulled over method, a totally working equivalent form of it, has not been acknowledged yet by built up scientists. In this phase, a triple phase equivalent variant of DBSCAN is exhibited. The trial grades are to a great degree empowering and exhibit the rightness, the versatility, and the adequacy of the method. The future work consolidates three fundamental objectives, confirmation of rightness, figuring of the time many-sided quality, and examinations. Initially, the strategy will exhibit theoretically the rightness of the proposed system while one needs to assess precisely its time unpredictability as an element of the partaking hubs and the extent of the informational index. Finally, the strategy will contrast our method and other near calculations paying little heed to the likelihood this is not sensible since the proposed procedure is the only a solitary found in the written work where the informational collection is traded by the hubs instead of exchanging the calculations to the informational collections.

Shaohua Teng, et.al," A Cooperative Multi-Classifier Method for Local Area Meteorological Data Mining", 2014

Cataclysmic events can provoke outrageous misfortunes in humanoid lifetime and possessions. Since various components join in a catastrophe, such actions are tough to figure precisely. A helpful multi-classifier technique is suggested there paper to coalface neighborhood information. The proposed technique is checked by the usage of both base and joining classifiers [31]. Knn-based agreeable grouping technique has preferences when managing the issue of climate estimating. Since the information don't contain works, the strategy can clearly use Euclidean separations to determine the separations of the information histories. In addition, it everything rapidly and is definitely not hard to regulator. The label traits are immovably connected to the order belongings. Meanwhile, there are a couple of inadequacies in the planned method, for instance, an instability around the K esteem and the quantity of improper classifiers. Simply finished experiences and rehashed tests would we have the capacity to secure a high exactness esteem. This should be researched later on. What's more, we will consider extra components on climate anticipating, for instance, lightning strike expectation. Through such work, we may enhance farsighted determining to decrease the effects of catastrophic events. Exploratory outcomes demonstrate that our proposed technique has higher arrangement exactness and speedier social event limit contrasted and ordinary classifiers.

Sudeep D. et.al," Extended Performance Appraise of Bayes, Function, Lazy, Rule, Tree Data Mining Classifier in Novel Transformed Fractional Content Based Image Classification", 2015

Picture order has ended up being one of the imperative research fields as several pictures are created conventional which induces the need to collect the grouping framework. To develop quicker and straightforward characterization framework, the visual substance of pictures is utilized.the exactness of order depends on the component extraction which is a champion among the most critical walk in picture grouping [32]. The paper demonstrates the execution of extra four orthogonal changes using changed fragmentary substance as the component for picture arrange where the Kekre, Hartle, Slant and Haar change are utilized as a bit of the development to prior proposed utilization of sine, cosine and Walsh changes. Twelve sorted out classifiers transversely over more than five information mining classifier family (Bayes, Function, Lazy, Rule and Tree) are used. Here 504 number of groupings for proposed picture organize method are taken a stab at using twelve classifiers, seven orthogonal changes and six sections of changed substance. The Simple Logistic classifiers with Kekre change gives better picture orchestrate enthusiastically took after by Simple Logistic with sine change and Simple Logistic with Hartley change. The Kekre change with the Simple Logistic classifier of Function family gives the best execution as appeared by higher rate gathering exactness for proposed picture game plan framework. Kstar classifier is not by any stretch out of the inventive capacity appropriate for picture arrange for all progressions.

Yumian Yang, et.al," Application of E-commerce Sites Evaluation based on Factor Analysis and Improved DBSCAN Algorithm", 2014

With the quick change of E-trade, how to appraise the E-business areas precisely has transformed into a main matter. Regardless, evaluation record of E-trade positions has components of extraordinary extents and unpredictable thickness, which prompts repulsive introduction of the estimation result. To separate 100 E-trade affirmation undertakings in 2013-2014 termed by the Department of the Business Working class Republic of China, this paper contracts dimensionality by highlight examination prepare initially, then apparatuses an overhauled DBSCAN calculation to strategy the sporadic thickness, all in all, offers to these 100 E-trade ventures in view of exploring results [33]. The customary DBSCAN is moved up to segment the record with dissimilar thicknesses and pack these destinations. This paper arranges frontward another handling thought on E-business destinations estimation: another DBSCAN calculation blending component examination with dissimilar solidities. Related with the customary DBSCAN calculation, the outcomes of surveying sites are extra useful and interpretable with the created DBSCAN calculation. In the anticipated work, the extent of the evaluated question will be additional drawn out and encourage examination ought to be finished.

XiaoqingYu, et.al," Explore Hot Spots of City Based on DBSCAN Algorithm", 2014

Spatial bunching is one of the principle strategies for information mining and learning revelation. DBSCAN calculation can be found in space with "clamor" database bunching of discretionary shape, is a sort of good grouping calculation. The general protest of information mining is the traditional social database. Information set away in the database by and large is value-based or social information. In any case, the attributes and exceptional stockpiling structure of spatial information and intriguing stockpiling structure doesn't allow us to use the customary information mining strategies in a spatial database for information mining and learning revelation. This paper presents the basic idea and rule of DBSCAN calculation and applies this calculation to perform grouping investigation circulations of web area data [34]. The article contrasts k-implies calculation and DBSCAN calculation remembering the ultimate objective to demonstrate the viability of DBSCAN calculation. This paper examines the crucial standard of DBSCAN calculation and its usage procedure. It applies this calculation in the field of city wanting to find the hot region in the city. Furthermore, it looked at DBSCAN calculation and k-implies calculation, and demonstrate its adequacy. Later on, it can be used to tear down city open offices or metropolitan open offices to give a logical commence and direction for city arranging.

Dr. R. Geetha Ramani, et.al," Data Mining Method of Evaluating Classifier Prediction Accuracy in Retinal Data", 2012

The examination as of late underlines the use of computational methods in the arena of ophthalmology. Diabetic Retinopathy, a retinal infection is a critical purpose behind visual impairment. Early location can help in behavior, notwithstanding, consistent transmission intended for premature recognition has been an extremely work - and resource serious assignment. Thus, programmed recognition of the infections through computational procedures would be a marvelous social cause. In this paper, the classifiers used for the programmed discovery of the malady are assessed using the information mining strategies [35]. The forecast precision of the significant number of classifiers assessed using diverse assessment strategies is exhibited. This work especially puts focus on the arrangement procedures to precisely arrange the ailment related with the retina in light of the elements separated from retinal pictures through picture handling strategies. Moreover, a point by point correlation of order calculations and the forecast assessment unrushed from many assessment techniques has been performed. A preparation precision of penny percent is expert by two or three classifiers although the expectation exactness stays at 76.67.Our outcomes show that a preparation precision of 100% can be proficient by two or three classifiers and a forecast precision of 76.67%.

Wilfried Segretier, et.al," An evolutionary data mining approach on hydrological data with classifier juries", 2012

In this paper, a developmental approach is appeared for separating a model of surge forecast from hydrological information watched opportune on river statures in a stream turning point [36]. Later, sort of information chronicled by radars on stream bowls is significantly uncommon and in a perfect world massively unbalanced between instances of surges and non-surges, creators have received the thought of total factors which qualities are enrolled as totals on unrefined information. A developmental calculation is included in allowing picking the greatest groups, boards of classifiers, of such factors as prescient factors. Binary genuine hydrological informational collections are prepared and they together exhibit the productivity of the technique contrasted with customary answers for expectation. From the information mining perspective, forthcoming effort on a dataset advanced by recently chronicled qualities, must permit to healthier comprehend the systems that deliver such contrasts among knowledge and exam exhibitions and to lessen them. The examinations have exhibited that this distinction could be imperative, both with our stochastic technique or with established arrangement strategies. Obviously, genuine positive rate, genuine negative rate, and earliness are foe destinations. It is fascinating to examine multi-target procedures to streamline them simultaneously.

Geeta Yadav, et.al," Predication of Parkinson's disease using Data Mining Methods: a comparative analysis of tree, statistical and support vector machine classifiers", 2012

The forecast of Parkinson's sickness in early age has been testing errand among specialists in light of the way that the indications of illness show up in center and late middle age. There is a piece of the manifestations that prompts Parkinson's malady. However, this paper focus on the discourse verbalization inconvenience indications of PD affected individuals and endeavor to define the model for the sake of three information mining techniques. These three information mining techniques are taken from three unmistakable areas of information mining i.e. from tree classifier, factual classifier and bolster vector machine classifier [37]. The execution of these three classifiers is measured with three execution frameworks i.e. exactness, affectability, and specificity. Thusly, the primary undertaking of this paper is endeavored to find which demonstrate distinguished the PD impacted individuals more precisely. We, finally, infer that LR show distinguished individuals with PD more successfully then Tree and SVM classifiers for the benefit of talked about execution frameworks. So in the investigation of Parkinson's infection, simply voice estimations of individuals have been considered to distinguish the individual with PD. There is a piece of manifestations that leads the Parkinson's sickness, for instance, age figure, natural variable, trembling in the hands, arms, legs, blocked discourse creation and discourse explanation inconveniences. In any case, in a bad position of PD affected individuals is considered for a development of the model and examination the talked about model on this indication of Parkinson's infection.

Gao Hua," Customer Relationship Management Based on Data Mining Technique-Naive Bayesian classifier", 2011

With the wild contention in the local and worldwide business, the Customer Relationship Management (CRM) has ended up being one of the matters of worry to the venture [38]. CRM takes the clients as the inside, it gives another life to the undertaking association framework and upgrades the business procedure. With a true objective to help endeavors comprehend their clients' shopping conduct and the approaches to hold esteemed clients, we propose information mining methods. As a rising subject, information mining is accepting an irrefutably vital piece of the choice bolster development of each stroll of life. This paper for the most part centered around the examination of the customer order and forecast in business banks in light of a Naive Bayesian classifier that suits the vulnerability intrinsic in foreseeing customer direct. The review will help the organization to separate and gauge customer's example of utilization, and introduce of customized showcasing services and administration. Although the paper concentrates mainly on the banking industry, the issues and applications discussed are applicable to different businesses, for example, protection industry, retail industry, produce enterprises, etc.

AI-Radaideh, et.al," A Study on Student Data Analysis Using Data Mining Techniques", 2006

Information mining procedure has a gigantic commitment for analysts to separate the shrouded learning and data which have been acquired in the information utilized by scientists. It is a preparing technique of removing dependable, novel, compelling and justifiable examples from database. This paper is utilized to classify the understudies into review arrange in all their training studies and it helps in meeting circumstance [39]. This study investigates the socio-demographic factors (age, sexual orientation, name, bring down class review, higher class review, degree capability and additional information or aptitude, and so forth). It looks at to what degree these elements sorts' understudies in rank request to organize the enlistment procedure. Because of this, all understudies get profited and it additionally diminishes the short postings. Here, bunching, affiliation tenets, order and exception identification has been utilized to assess the understudies' execution. One of the information mining procedures that is grouping, precisely characterizes the information for sorting understudy in light of the levels. As one vital capacity of information mining, bunching investigation either as a different apparatus to find information sources conveyance of data, and in addition other information mining calculation as a preprocessing step, the group examination has been into the field of information mining is an essential research subject. Bunching is utilized to the gathering the understudies as indicated by their review and capability. This goes far to help how characterize the enrollment procedure in a less demanding way.

Y. Freund, et.al," Experiments with a new boosting algorithm", 1996

Neural system gathering is a learning worldview where numerous neural systems are mutually used to take care of an issue. In this paper, the relationship between the troupe and its part neural systems is dissected from the setting of both relapse and grouping, which uncovers that it might be ideal to gathering numerous rather than the majority of the neural systems within reach. This outcome is intriguing in light of the fact that at present, most methodologies group all the accessible neural systems for expectation. At that point, keeping in mind the end goal to demonstrate that the fitting neural systems for creating a gathering can be viably chosen from an arrangement of accessible neural systems, an approach named GASEN is exhibited. GASEN trains various neural systems at first. At that point it allocates arbitrary weights to those systems and utilizes hereditary calculation to advance the weights with the goal that they can describe to some degree the wellness of the neural systems in constituting an outfit. At last it chooses some neural systems in view of the advanced weights to make up the group. A huge observational review demonstrates that, contrasted and some prominent outfit methodologies, for example, Bagging and Boosting, GASEN can create neural system troupes with far littler sizes yet more grounded speculation capacity. Moreover, so as to comprehend the working instrument of GASEN, the predisposition fluctuation deterioration of the mistake is given in this paper, which demonstrates that the accomplishment of GASEN may lie in that it can fundamentally diminish the inclination and also the change.

CHAPTER-3

PRESENT WORK

3.1. Problem Formulation

This work is based on prediction analysis to predict the student performance on a subjective and objective type of exams. In this work, the dataset is considered in which how many students passed the subjective exam and how many students passed from an objective exam. In the work, two techniques are passed the first technique of clustering and second technique of classification. In the technique of clustering density based clustering can be applied which calculate the most dense region from the dataset and on the basis of EPS, Euclidian distance final results of dissimilar and similar data get generated in the form of clusters. The final generated clusters will be given as input to the classifier to the classified result is generated as pass and failed students. In the Density-based clustering, the EPS values are calculated statically which reduce the accuracy of clustering and classification.

3.2. Objectives

1. To study and analysis various prediction analysis technique of Data mining

2. To proposed improvement in density based clustering and classification for data mining

3. The proposed improvement is based on to calculate EPS value dynamically to increase accuracy of clustering

4. To implement proposed and existing techniques and compare results in terms of accuracy and execution time

3.3. Research Methodology

This work is based on to predict the student performance in terms of students passed the subjective and objective type of exams. The dataset of the students gets prepared through the questioner designed. The dataset is given as input to density based clustering which calculates the dense region and from the dense region, EPS value is calculated which is centered point of the dense region. To calculate the similarity between the points technique of Euclidian distance is applied which is given a similar and dissimilar type of data. The output of density based clustering is input to classification algorithm which will classify the data points. To improve the performance of prediction analysis neural networks will be applied with density-based clustering which clusters the unclustered data points and improves an accuracy of clustering, reduces execution time. The SVM classifier is the used which will classify the data according to the input dataset. In the proposed technique the euclidian distance is calculated dynamically using the technique of back propagation. The back propagation returns the euclidian distance at which the error is minimum means the accuracy is maximum. The back propagation algorithm is one of the most utilized Neural Network algorithms. This method is used for training the artificial neural networks and also utilizes the two-phase cycle which involves the propagation and weight updates. When an input network enters the network, it is spread forward through the network across each layer until it reaches the output layer. The comparisons are made using the output achieved as well as the desired output. This is done utilizing a loss function. For every neuron in the output layer, an error value is calculated. The propagation of the error values is then done in the backward manner which starts from the output. Here, each neuron has its own error value which also shows its contribution to the originally achieved output.

There are mainly four steps in which this algorithm can be executed. The required corrections are to be computed only once the weights of the network are selected randomly. The following are the steps in which the algorithm is decomposed:

i) Feed-forward computation

ii) Back propagation to the output layer

iii) Back propagation to the hidden layer

iv) Weight updates

At the time when the values of error function become small, the algorithm is stopped. This is just an overview of the basic BP algorithm. However, various changes are proposed by researchers with time. The algorithm for back propagation is mentioned below:

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

Error=Desired Output-Actual Output

The formula of error is applied which will calculate the error at each iteration which leads to defining the accuracy of clustering.

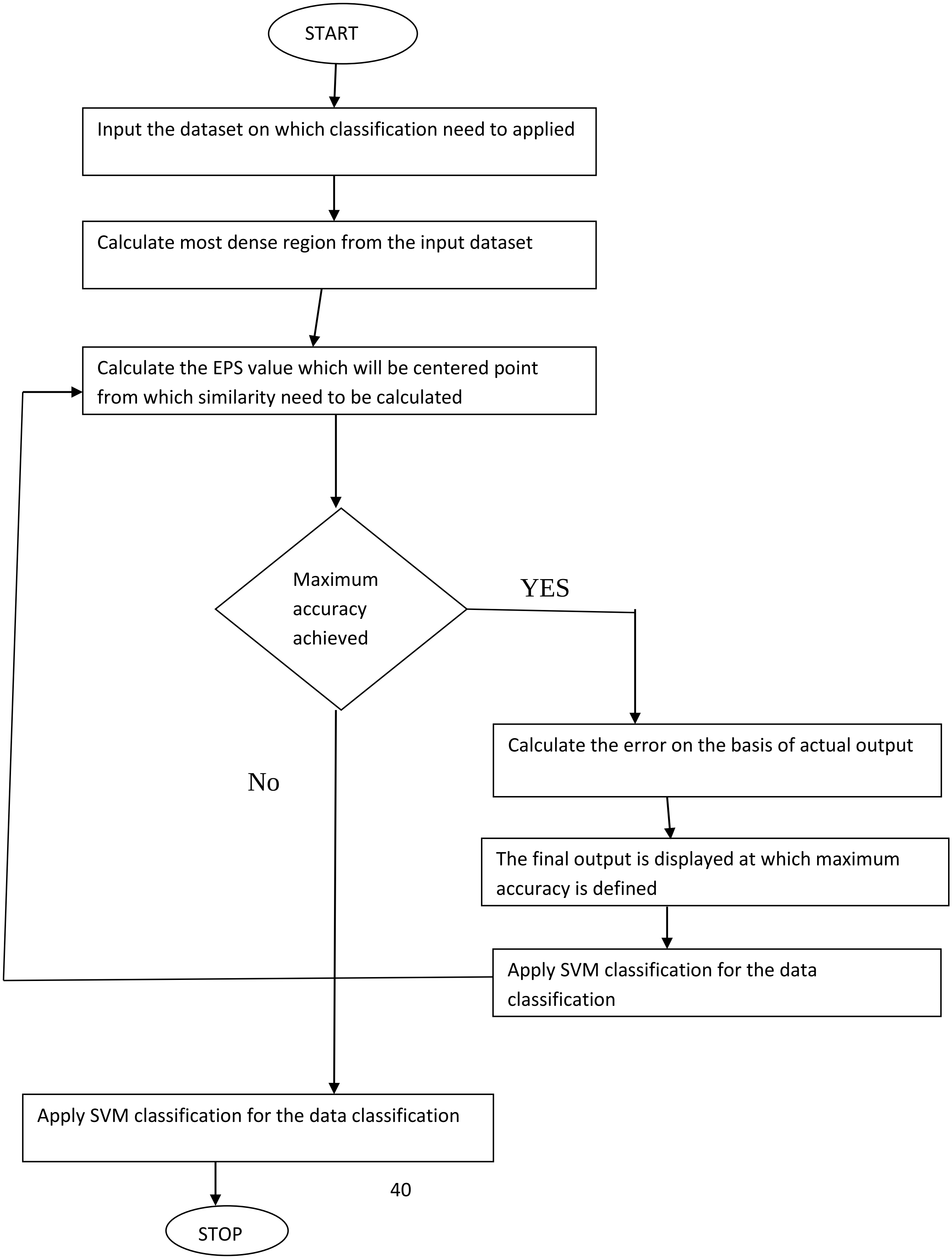
Pseudo code of proposed work

Input: The dataset for the classification

Output: Classified Data

```
1. Input the dataset and store in the variable k

2. Calculate most dense region from the dataset and store dense region in variable d

3. calculate Eps value from the dataset

4. for (i=0;i=n;i++)

            For(j=0;j=n;j++)

                  If(distance (k(i,j)<k(i+1,j+1))

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

                       Calculate error= Desired Output-Actual Output

                           Cluster=if (error (K (i,j)>error(K(i+1,j+1);

                                       Cluster=K(i,j);
            }
        }
    }
5. Apply SVM classifier for the data classification
6. Return classified data
```

FLOWCHART


START
Input the dataset on which classification need to applied
Calculate most dense region from the input dataset
Calculate the EPS value which will be centered point from which similarity need to be calculated
Maximum accuracy achieved
YES
No
Calculate the error on the basis of actual output
The final output is displayed at which maximum accuracy is defined
Apply SVM classification for the data classification
Apply SVM classification for the data classification
STOP

CHAPTER-4

RESULT AND DISCUSSION

4.1. Tool Description

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 process elements in the form of MATRIXs and various other languages like JAVA, PYTHON, and FORTRAN are used in MATLAB. The MATLAB default interface has following parts

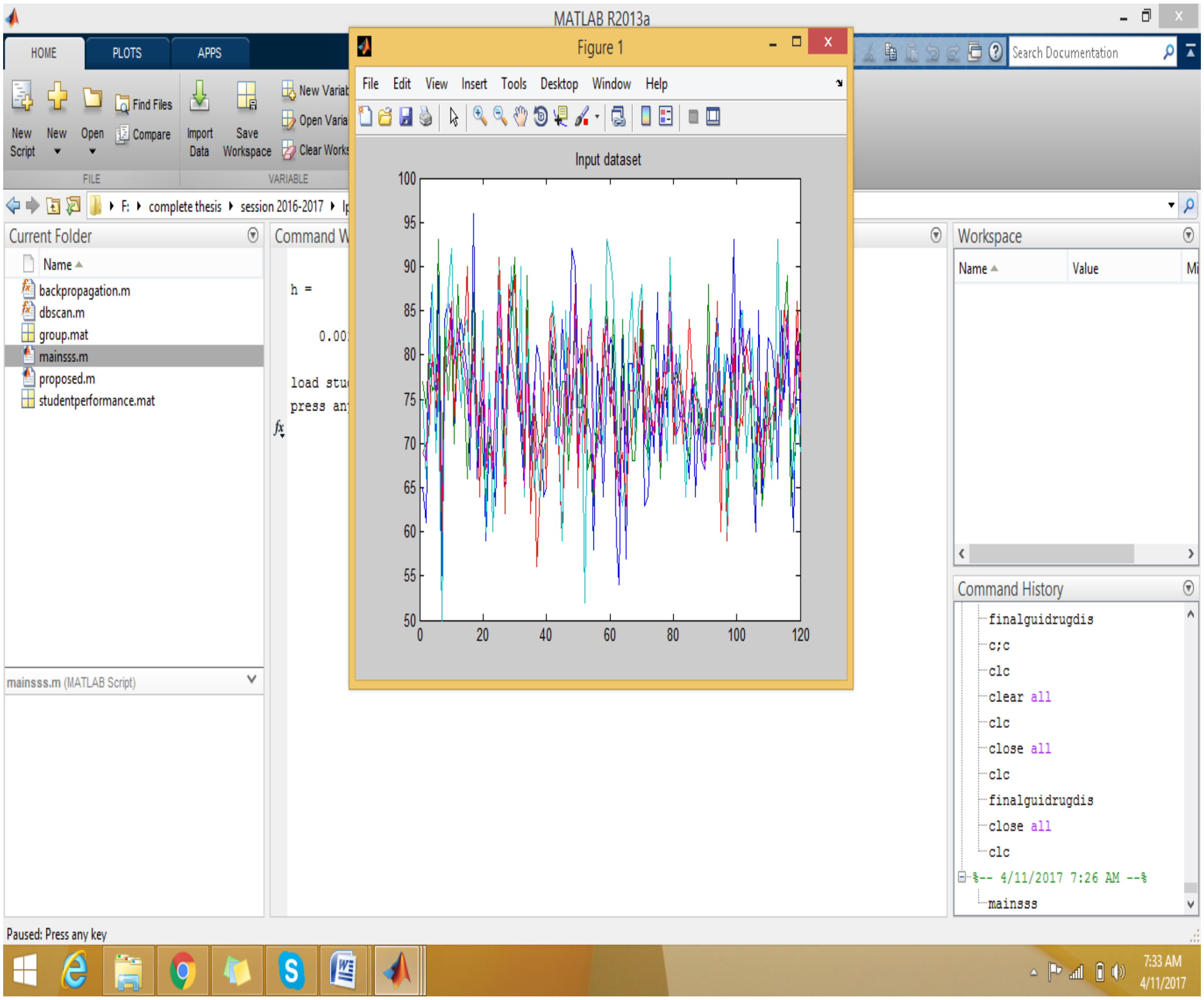
I. **Command Window**:- The Command Window is the first importance part of MATLAB which is used to show output of already saved code and to execute MATLAB codes temporarily

II. **Work Space** :-The workspace is the second part of MATLAB which is used to show allocation and de-allocation of MATLAB variables. The workspace is divided into three parts. The first part is MATLAB variable ,variable type and third part is variable value

III. **Command History** :- The command history is the third part of MATLAB in which MATLAB commands are shown which are executed previously

IV. **Current Folder Path** :- The current Folder path shows that path of the folder in which MATLAB codes are saved

Current Folder Data: - The Current Folder Data shows that data which is in the folders whose path is given in Current Folder Path



backpropagation.m
dbscan.m
group.mat
mainsss.m
proposed.m
studentperformance.mat

mainsss.m (MATLAB Script)

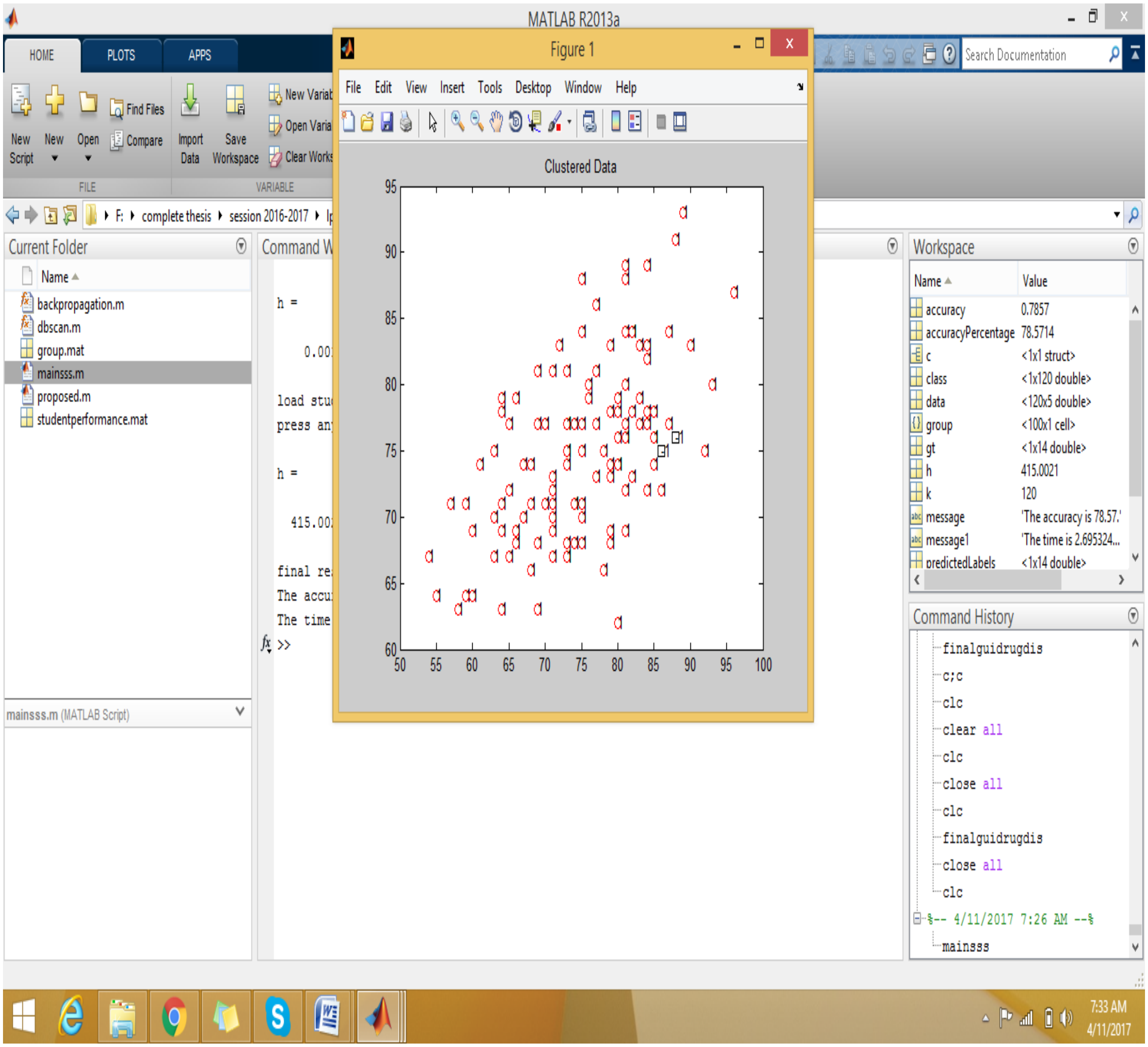
```
h =

    0.00

load stu
press an
```

```
finalguidrugdis
c;c
clc
clear all
clc
close all
clc
finalguidrugdis
close all
clc
%-- 4/11/2017 7:26 AM --%
mainsss
```

Paused: Press any key

Fig 4.1.1: Input dataset

As shown in figure 1, the data is given as input which is of student performance dataset. The dataset is plotted on the 2-D plane and so that x-axis shows the number of attributes and y-axis shows the member values of the dataset.



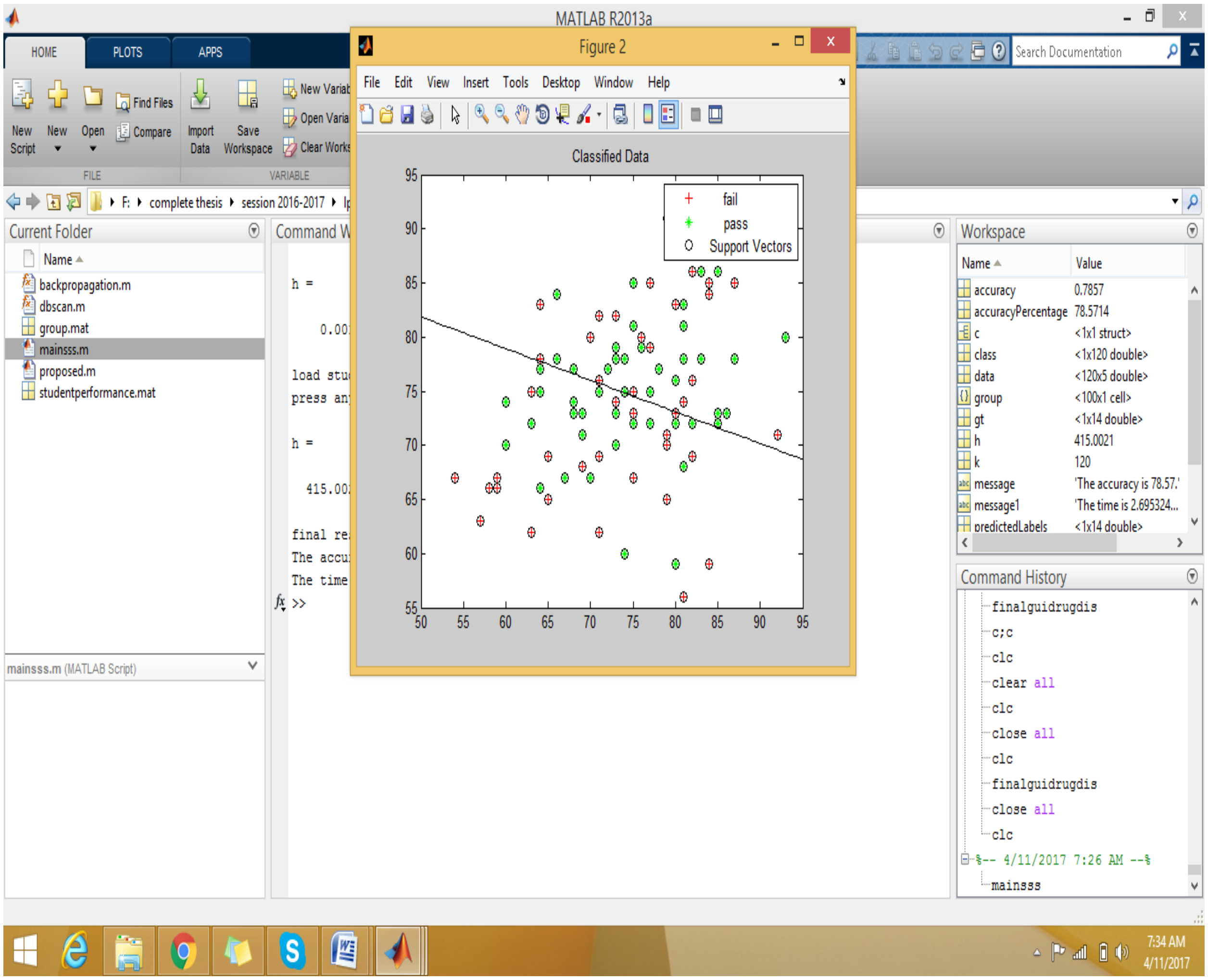
```
h =

    0.00

load stu
press an

h =

  415.00

final re
The accu
The time
>>
```

Name	Value
accuracy	0.7857
accuracyPercentage	78.5714
c	<1x1 struct>
class	<1x120 double>
data	<120x5 double>
group	<100x1 cell>
gt	<1x14 double>
h	415.0021
k	120
message	'The accuracy is 78.57.'
message1	'The time is 2.695324...
predictedLabels	<1x14 double>

```
finalguidrugdis
c;c
clc
clear all
clc
close all
clc
finalguidrugdis
close all
clc
%-- 4/11/2017 7:26 AM --%
mainsss
```

Fig 4.1.2: Result of DBSCAN algorithm

As shown in figure 2, the density based clustering is applied which will cluster the similar and dissimilar data and the dataset is of -1 type according to which clusters are shown in figure



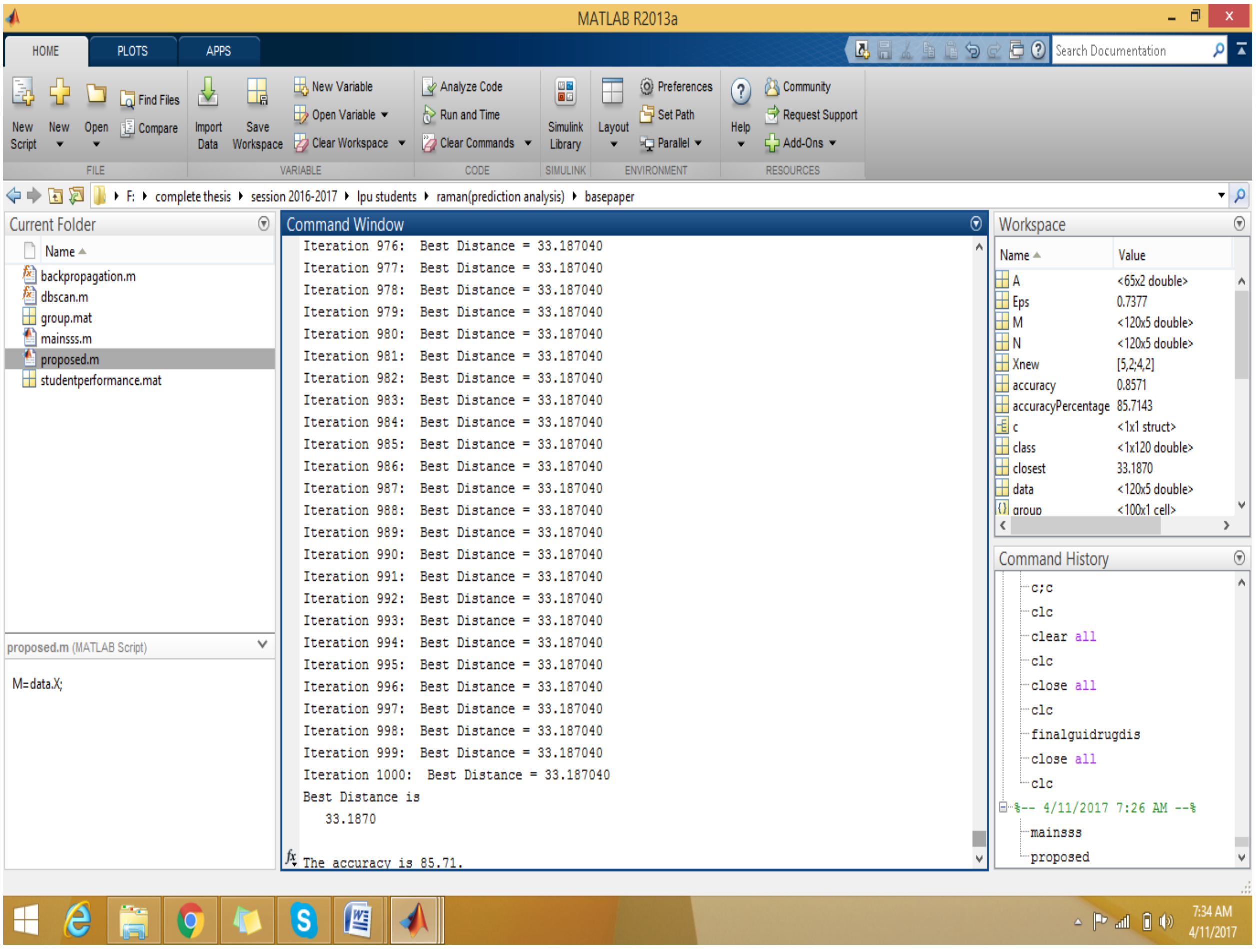
```
h =

    0.00

load stu
press an

h =

  415.00

final re
The accu
The time
>>
```

Name	Value
accuracy	0.7857
accuracyPercentage	78.5714
c	<1x1 struct>
class	<1x120 double>
data	<120x5 double>
group	<100x1 cell>
gt	<1x14 double>
h	415.0021
k	120
message	'The accuracy is 78.57.'
message1	'The time is 2.695324...
predictedLabels	<1x14 double>

```
finalguidrugdis
c;c
clc
clear all
clc
close all
clc
finalguidrugdis
close all
clc
%-- 4/11/2017 7:26 AM --%
mainsss
```

Fig 4.1.3: SVM classification results

As shown in figure 3, the technique of SVM classifier is applied which classify the similar and dissimilar type of data. As shown in figure, the algorithm will classify the data of the students which get passed and which get failed in the exams



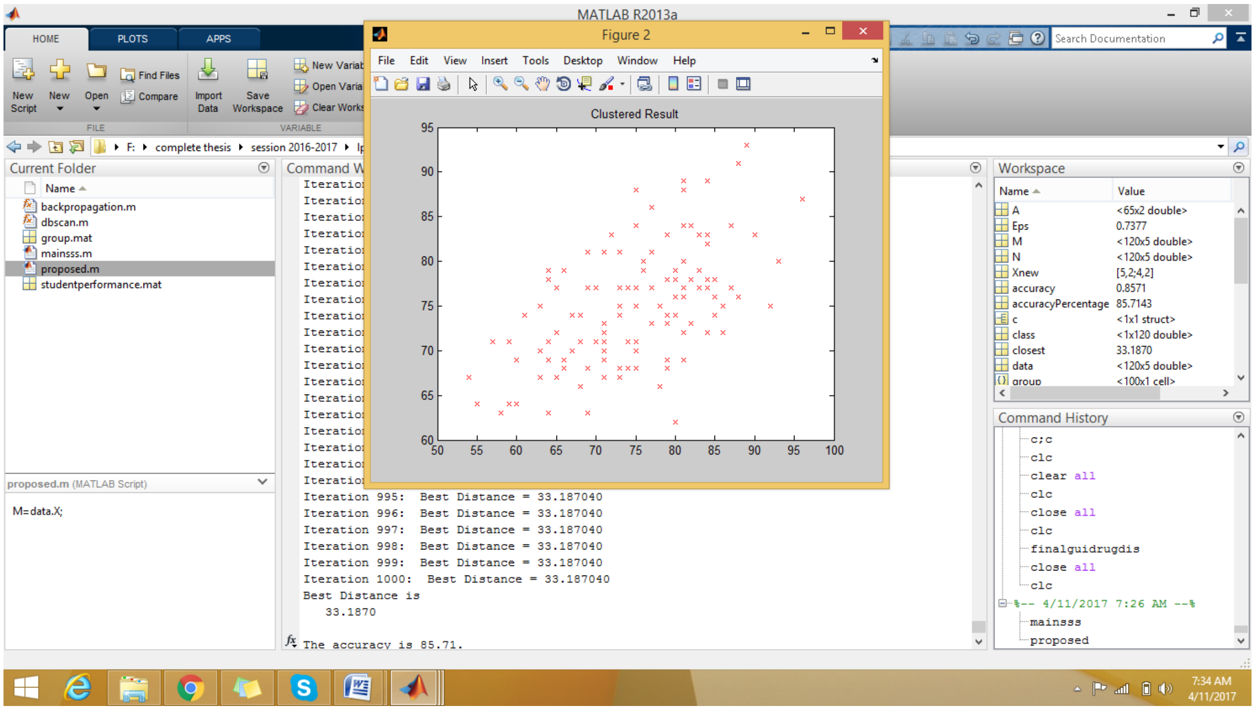
```
Iteration 976:  Best Distance = 33.187040
Iteration 977:  Best Distance = 33.187040
Iteration 978:  Best Distance = 33.187040
Iteration 979:  Best Distance = 33.187040
Iteration 980:  Best Distance = 33.187040
Iteration 981:  Best Distance = 33.187040
Iteration 982:  Best Distance = 33.187040
Iteration 983:  Best Distance = 33.187040
Iteration 984:  Best Distance = 33.187040
Iteration 985:  Best Distance = 33.187040
Iteration 986:  Best Distance = 33.187040
Iteration 987:  Best Distance = 33.187040
Iteration 988:  Best Distance = 33.187040
Iteration 989:  Best Distance = 33.187040
Iteration 990:  Best Distance = 33.187040
Iteration 991:  Best Distance = 33.187040
Iteration 992:  Best Distance = 33.187040
Iteration 993:  Best Distance = 33.187040
Iteration 994:  Best Distance = 33.187040
Iteration 995:  Best Distance = 33.187040
Iteration 996:  Best Distance = 33.187040
Iteration 997:  Best Distance = 33.187040
Iteration 998:  Best Distance = 33.187040
Iteration 999:  Best Distance = 33.187040
Iteration 1000:  Best Distance = 33.187040
Best Distance is
   33.1870

fx The accuracy is 85.71.
```

Name	Value
A	<65x2 double>
Eps	0.7377
M	<120x5 double>
N	<120x5 double>
Xnew	[5,2;4,2]
accuracy	0.8571
accuracyPercentage	85.7143
c	<1x1 struct>
class	<1x120 double>
closest	33.1870
data	<120x5 double>
group	<100x1 cell>

```
c;c
clc
clear all
clc
close all
clc
finalguidrugdis
close all
clc
%-- 4/11/2017 7:26 AM --%
mainsss
proposed
```

Fig 4.1.4: Proposed Algorithm

As shown in figure 4, the data which is given as the input is represented on the 2-D plan and Euclidian distance is calculated dynamically which show the best distance at which maximum accuracy is achieved.



Current Folder

- backpropagation.m
- dbscan.m
- group.mat
- mainsss.m
- proposed.m
- studentperformance.mat

proposed.m (MATLAB Script)

```
M=data.X;
```

```
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteration 995:  Best Distance = 33.187040
Iteration 996:  Best Distance = 33.187040
Iteration 997:  Best Distance = 33.187040
Iteration 998:  Best Distance = 33.187040
Iteration 999:  Best Distance = 33.187040
Iteration 1000:  Best Distance = 33.187040
Best Distance is
   33.1870

The accuracy is 85.71.
```

Name	Value
A	<65x2 double>
Eps	0.7377
M	<120x5 double>
N	<120x5 double>
Xnew	[5,2;4,2]
accuracy	0.8571
accuracyPercentage	85.7143
c	<1x1 struct>
class	<1x120 double>
closest	33.1870
data	<120x5 double>
group	<100x1 cell>

```
c;c
clc
clear all
clc
close all
clc
finalguidrugdis
close all
clc
%-- 4/11/2017 7:26 AM --%
mainsss
proposed
```

Fig 4.1.5: Final Clustering Result

As shown in figure 5, the dataset which is taken as input will be shown on the 2-D plan and from that dataset technique of back propagation will be applied which will calculate distance dynamically and final result of clustering is shown which shows the data of class 1

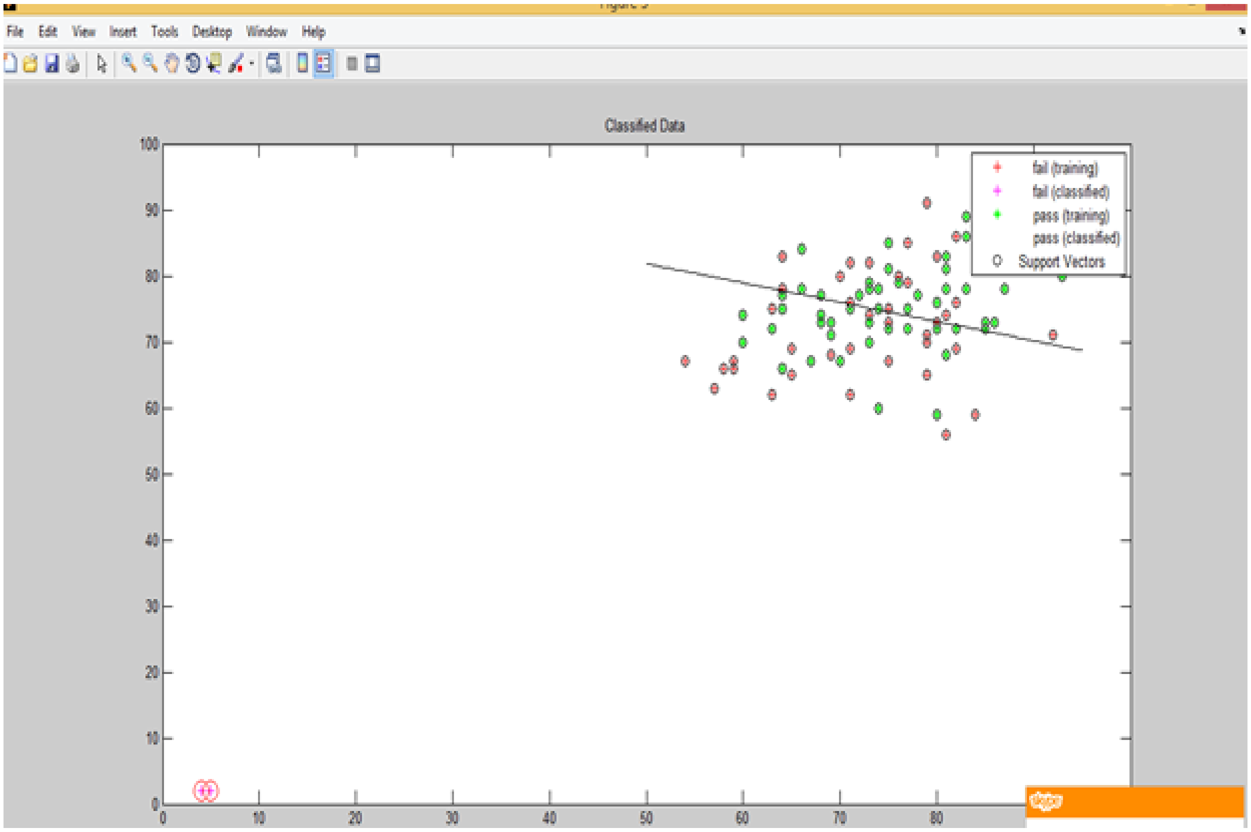


Fig 4.1.6: SVM classifier

As shown in figure 6, the SVM classifier is been applied which will classify the data into the two classed and it will merge the three classes into two.

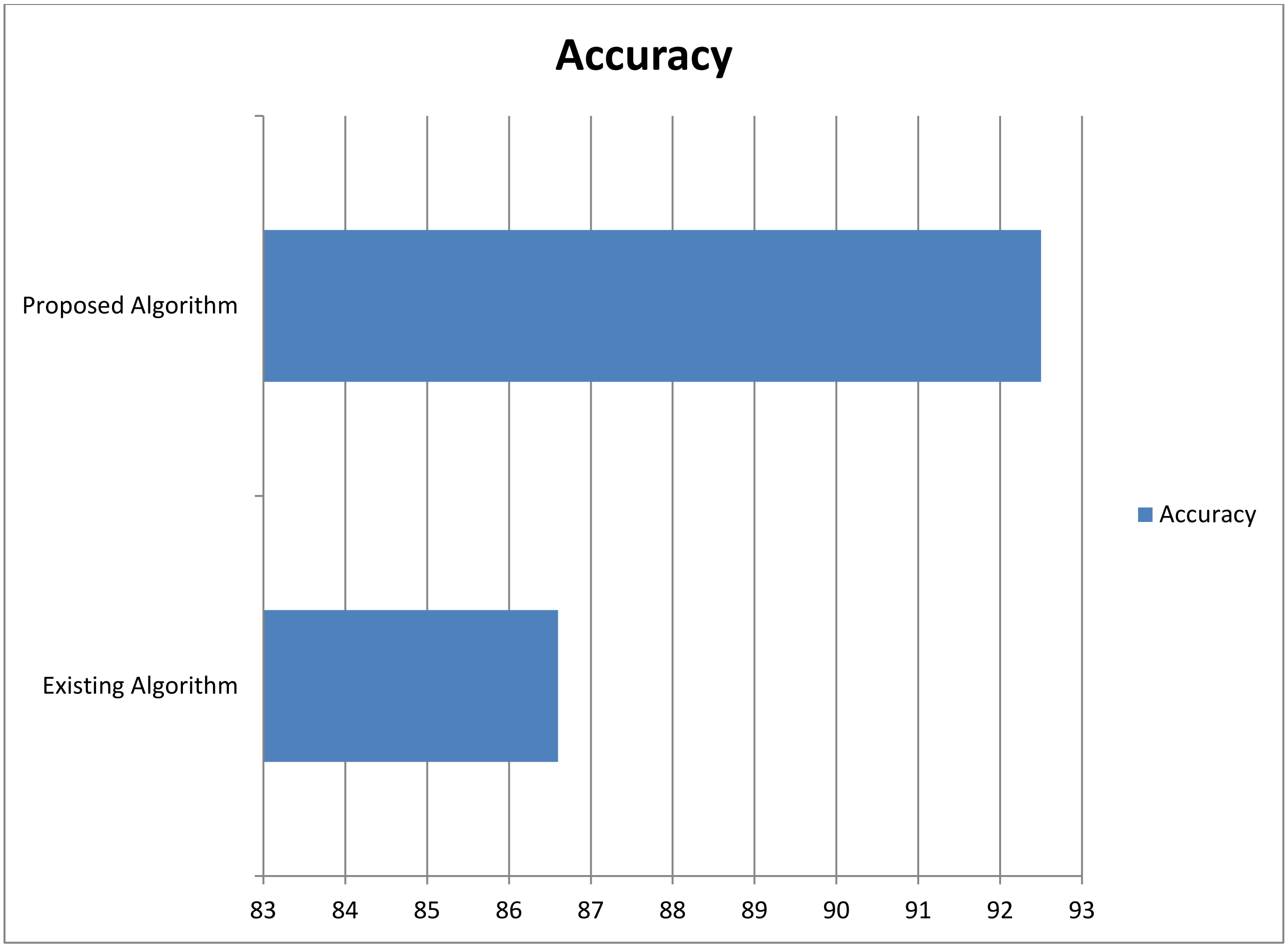


Fig 4.1.7: Accuracy Comparison

As shown in figure 7, the accuracy of proposed and existing algorithm has been compared and it is been analyzed that it is increased to 92 percent from 86 percent

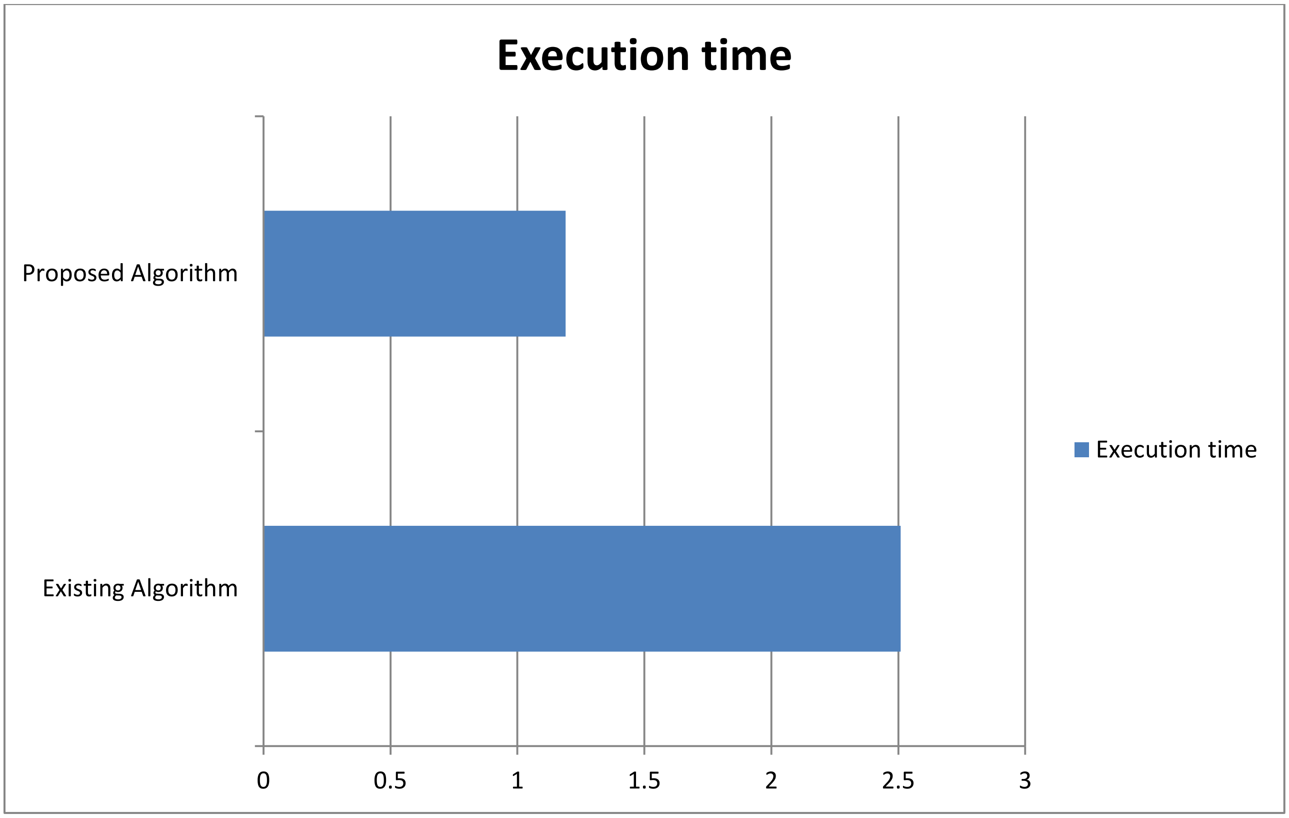


Fig 4.1.8: Execution time comparison

As shown in figure 8, the execution time of proposed and existing algorithm is compared and it is been analyzed that execution time of proposed algorithm is reduced to 1.21 percent

CHAPTER-5

CONCLUSION AND FUTURE SCOPE

5.1 CONCLUSION

The density based clustering is the technique in which similar and dissimilar type of data is clustered together according to data density. The DBSCAN algorithm is applied which will calculate the EPS point which is the central point and from the central point the Euclidian distance will be calculated which will cluster the similar and dissimilar type of data. In the DBSCAN algorithm the Euclidian distance will be calculated statically. In this work, technique of back propagation will be applied which calculate Euclidian distance dynamically. The proposed improvement leads to increase accuracy and reduce execution time. The SVM classifier is applied which will classify the similar and dissimilar type of data. The proposed technique is implemented in MATLAB and it is analyzed that accuracy is increased to 20 percent, execution time is reduced to 10 percent.

5.2 Future Scope

Following are the various future scopes of this research

1. The technique will be applied in future which increase the gini index of classification and also reduce the execution time for the data classification

2. The proposed algorithm can be tested on the other datasets like disease datasets to check its reliability

REFERENCES

[1] Y. He, H. Tan, W. Luo, S. Feng, and J. Fan, “Mr-dbscan: A scalable mapreduce-based dbscan algorithm for heavily skewed data,” 2014, Frontiers of Computer Science, vol. 8, no. 1, pp. 83–99

[2] J. B. MacQueen, “Some methods for classification and analysis of multivariate observations,” in 5th Berkeley Symposium on Mathematical Statistics and Probability, U. of California Press, Ed., 1967, pp. 281–297

[3] M. Goetz, M. Richerzhagen, C. Bodenstein, G. Cavallaro, P. Glock, M. Riedel, and J. Benediktsson, “On scalable data mining techniques for earth science,” 2015, vol. 51, no. 1, pp. 2188–2197

[4] Y. He, H. Tan, W. Luo, H. Mao, D. Ma, S. Feng, and J. Fan, “Mrdbscan: An efficient parallel density-based clustering algorithm using mapreduce,” 2011, Springer, pp. 473–480.

[5] M. M. A. Patwary, D. Palsetia et al., “A new scalable parallel dbscan algorithm using the disjoint-set data structure,” 2012, International Conference for High Performance Computing, Networking, Storage and Analysisi - SC12

[6] P. De Meo, F. Messina, D. Rosaci, and G. M. Sarne, “An agent oriented, trust-aware approach to improve the qos in dynamic grid federations,” 2015, Concurrency and Computation: Practice and Experience, vol. 27, no. 17, pp. 5411–5435

[7] E. Januzaj, H.-P. Kriegel, and M. Pfeifle, “Scalable density-based distributed clustering,” 2004, Lecture Notes in Computer Science, vol. 3202, pp. 231–244

[8] B.-R. Dai and I.-C. Lin, “Efficient map/reduce-based dbscan algorithm with optimized data partition,” 2012, pp. 59–66

[9] R. Giunta, F. Messina, G. Pappalardo, and E. Tramontana, “Providing qos strategies and cloud-integration to web servers by means of aspects,” 2015, Concurrency and Computation: Practice and Experience, vol. 27, no. 6, pp. 1498–1512

[10] F. Messina, G. Pappalardo, D. Rosaci, C. Santoro, and G. Sarne, “A trust-aware, self-organizing system for large-scale federations of utility computing infrastructures,” 2016, Future Generation Computer Systems, vol. 56, pp. 77–94

[11] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” 1996, 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), pp. 226–231

[12] C. P. McQuellin, H. F. Jelinek, and G. Joss, “Characterisation of fluorescein angiograms of retinal fundus using mathematical morphology:a pilot study,” 2002, 5th International Conference on Ophthalmic Photography, Adelaide, p. 152

[13] T. Y. Wong, W. Rosamond, P. P. Chang, D. J. Couper, A. R. Sharrett, L. D. Hubbard, A. R. Folsom, and R. Klein, “Retinopathy and risk of congestive heart failure,” 2005, Journal of the American Medical Association,vol. 293, no. 1, pp. 63–69

[14] M.M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A.R. Rudnicka, C.G. Owen and S.A. Barman “Blood vessel segmentation methodologies in retinal images – A survey,” 2012, computer methods and programs in biomedicine,

[15] K. Buhler, P. Felkel and A.L. Cruz, “Geometric methods for vessel visualization and quantification – a survey,” 2003, Geometric Modelling for Scientific Visualization, pp. 399–421

[14] C. Kirbas and F. Quek, “A review of vessel extraction techniques and algorithms,” 2004, ACM Computing Surveys 36, pp. 81–121

[15] M.S. Mabrouk, N.H. Solouma and Y.M. Kadah, “Survey of retinal image segmentation and registration,” 2006, ICGST International Journal on Graphics, Vision and Image Processing 6, pp. 1–11

[16] Chaudhuri S, Chatterjee S, Katz N, Nelson M and Goldbaum M.,“ Detection of blood vessels in retinal images using two-dimensional matched filters”, 1989, IEEE Trans Med Imaging;8(3), pp. 263–9

[17] Chutatape O, Zheng L and Krishnan SM. “Retinal blood vessel detection and tracking by matched Gaussian and Kalman filters”, 1998 Proceedings of the 20th annual international conference of the IEEE on engineering in medicine and biology, vol. 6, pp. 3144–9

[18] Huan Yu, Wenhui Zhang,” DBSCAN Data Clustering Algorithm for Video Stabilizing System”, 2013, International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC)

[19] Nagaraju S, Manish Kashyap,” A Variant of DBSCAN Algorithm to Find Embedded and Nested Adjacent Clusters”, 2016, 3rd International Conference on Signal Processing and Integrated Networks (SPIN)

[20] Jianbing Shen, Xiaopeng Hao, Zhiyuan Liang, Yu Liu, Wenguan Wang, and Ling Shao,” Real-time Superpixel Segmentation by DBSCAN Clustering Algorithm”, 2016, IEEE

[21] Saefia Beri, Kamaljit Kaur,” Hybrid Framework for DBSCAN Algorithm Using Fuzzy Logic”, 2015 1st International conference on futuristic trend in computational analysis and knowledge management (ABLAZE)

[22] Bharathi S, Manju M, Vasavi Manasa C L, Mallika H M, Maruti M Kurule, P. Deepa Shenoy, Venugopal K R, L M Patnaik,” Automatic Land Use/Land Cover Classification using Texture and Data Mining Classifier”, 2013, IEEE

[23] Sneha Chandra, Maneet Kaur,” Enhancement of Classification Accuracy of our Adaptive Classifier using Image Processing Techniques in the Field of Medical Data Mining”, 2015, IEEE

[24] Yomna M. ElBarawy, Ramadan F. Mohamedt and Neveen I. Ghali,” Improving Social Network Community Detection Using DBSCAN Algorithm”, 2014, IEEE

[25] Dominik Fisch, Edgar Kalkowski, Bernhard Sick,” Knowledge Fusion for Probabilistic Generative Classifiers with Data Mining Applications”, 2013, IEEE

[26] Dianwei Han, Ankit Agrawal, Wei−keng Liao, Alok Choudhary,” A novel scalable DBSCAN algorithm with Spark”, 2016 IEEE International Parallel and Distributed Processing Symposium Workshops

[27] Md. Rejaul Karim, and Dewan Md. Farid,” An Adaptive Ensemble Classifier for Mining Complex Noisy Instances in Data Streams”, 2014, 3rd INTERNATIONAL CONFERENCE ON INFORMATICS, ELECTRONICS & VISION

[28] Karlina Khiyarin Nisa, Hari Agung Andrianto, Rahmah Mardhiyyah,” Hotspot Clustering Using DBSCAN Algorithm and Shiny Web Framework”, 2014, IEEE

[29] Negar Riazifar, Ehsan Saghapour,” Retinal Vessel Segmentation Using System Fuzzy and DBSCAN Algorithm”, 2015 2nd International Conference on Pattern Recognition and Image Analysis (IPRIA)

[30] Ilias K. Savvas, and Dimitrios Tselios,” Parallelizing DBSCAN Algorithm Using MPI”, 2016, 25th IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises

[31] Shaohua Teng, Jihui Fan, Haibin Zhu, Wei Zhang, Dongning Liu, Xiao Chen, Xiufen Fu,” A Cooperative Multi-Classifier Method for Local Area Meteorological Data Mining”, 2014, Proceedings of the IEEE 18th International Conference on Computer Supported Cooperative Work in Design

[32] Sudeep D. Thepade, Madhura M. Kalbhor,” Extended Performance Appraise of Bayes, Function, Lazy, Rule, Tree Data Mining Classifier in Novel Transformed Fractional Content Based Image Classification”, 2015 International Conference on Pervasive Computing (ICPC)

[33] Yumian Yang, Jianhua Jiang,” Application of E-commerce Sites Evaluation based on Factor Analysis and Improved DBSCAN Algorithm”, 2014 International Conference on Management of e-Commerce and e-Government

[34] XiaoqingYu, Yupu Ding, Wanggen Wan, Etienne Thuillier,” Explore Hot Spots of City Based on DBSCAN Algorithm”, 2014, IEEE

[35] Dr.R. Geetha Ramani, Lakshmi.B, Shomona Gracia Jacob,” Data Mining Method of Evaluating Classifier Prediction Accuracy in Retinal Data”, 2012, IEEE

[36] Wilfried Segretier, Manuel Clergue, Martine Collard, Luis Izquierdo,” 2012, WCCIIEEE World Congress on Computational Intelligence

[37] Geeta Yadav, Yugal Kumar, G. Sahoo,” Predication of Parkinson’s disease using Data Mining Methods: a comparative analysis of tree, statistical and support vector machine classifiers”, 2012 National Conference on Computing and Communication Systems (NCCCS)

[38] Gao Hua,” Customer Relationship Management Based on Data Mining Technique-Naive Bayesian classifier”, 2011, IEEE

[39] AI-Radaideh, AI-Shawakfa and AI-Najjar, "Mining Student Data Using Decision Trees”, The 2006 International Arab Conference on Information Technology (ACIT'2006), 2006