



**COMPARATIVE ANALYSIS OF K-MEAN AND
HYBRID K-MEAN CLUSTERING ALGORITHM FOR
PREDICTION OF DIABETES MELL_EH_TISS**

A Dissertation report

Submitted

By

Sonal Arora

(11301491)

To

Department of Computer Science & Engineering

In partial fulfillment of the requirement for the

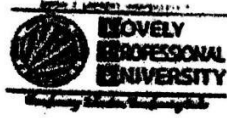
Award of the degree of

Master of Technology in Computer Science & Engineering

**Under the guidance of
Kundan Munjal (16806)**

(May 2015)

PAC APPROVAL FORM



School of: Computer Science and Engineering

DISSERTATION TOPIC APPROVAL PERFORMA

Name of the student : SONAL ARORA
Batch : 2013-2015
Session : 2014-2015

Registration No : 11301491
Roll No : RK2306A27
Parent Section : K2306

Details of Supervisor:

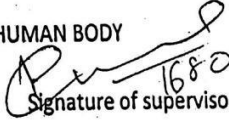
Name : KUNDAN MUNJAL
UID : 16806

Designation : Assistant Professor
Qualification : M.Tech
Research Exp. : 2.5 year

Specialization Area: DATABASE AND INFORMATION SYSTEMS (pick from list of provided specialization areas by DAA)


Proposed Topics:-

1. PREDICTION OF DIABETES USING UNSUPERVISED LEARNING METHOD
2. VARIOUS CLUSTERING ALGORITHM DIFFERENTIATION
3. EFFICIENT DATA MINING APPROACH FOR DETECTION OF DIABETES IN HUMAN BODY


Signature of supervisor

PAC Remarks:

Topic 1 is approved


11/6/14

APPROVAL OF PAC CHAIRMAN

Signature: 

Date:

- *Supervision should finally encircle one topic out of three proposed topics and put up for an approval before Project Approval Committee (PAC).
- *Original copy of this format after PAC approval will be retained by the student and must be attached in the Project/ Dissertation final report.
- *One copy to be submitted to supervisor.

ABSTRACT

In the medical environment, data is “information rich” but “knowledge poor” because the data which is available is not “mined” as it cannot discover the hidden patterns and information which is used for reaching at the effective decision. Also, there is a lack of effective tools for analysis which can discover the hidden relationships and trends in data. This problem can be solved by using advanced data mining techniques like CANFIS (neuro-fuzzy inference system), rule based, decision tree, Naïve Bayes and artificial neural network. Data mining have numerous applications in field of business and scientific domains. Valuable knowledge can be discovered by the application of data mining techniques in the healthcare systems. Diabetes is a disease which has attacked most of the people around the world. Almost all people are reporting their problems of diabetes whether occurred by hereditary or any other problem. People around the world are using glucometer for testing their diabetes unknown of the fact that their results vary if another test is conducted after the first test. This leads to the error in detection of glucose with ± 10 mg/dl. The data which is hidden can be mined and extracted so as to obtain the fruitful results out of that data. The basic proposal of this research study is to extract the hidden knowledge and patterns from the diabetes dataset and preprocessing of that data takes place. K-means Clustering algorithm to obtain the accuracy of the prediction is applied and also the new algorithm will be designed which will be the improved version of K-means to increase the accuracy and see the effect on the prediction of diabetes. This research study is very beneficial for the doctors and patients. It will provide the user with a user friendly environment without the need of doctor or any hospital staff and will help to mine the data using new technique being designed. These algorithms being designed will define the effectiveness and efficiency of the method used to predict the diabetes mellitus.

CERTIFICATE

This is to certify that **Sonal Arora** has completed M.Tech (Computer Science and Engineering) Dissertation titled “**Comparative Analysis of K-means and Hybrid K-means Clustering Algorithm for the prediction of Diabetes Mell-EH-Tiss**” under my guidance and supervision. To the best of my knowledge the present work is the result of his original investigation and study. No part of the dissertation has ever been submitted for any other degree or diploma.

The dissertation is fit for the submission and the partial fulfillment of the conditions award of M.Tech (Computer Science and Engineering) degree.

Date: _____

Signature of Advisor

Name: _____

UID: _____

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose constant guidance crowned our efforts with success.

I sincerely express our deep gratitude to the management of our college for giving us liberty to choose and to work on the most relevant project i.e. “**Comparative Analysis of K-means and Hybrid K-means Clustering Algorithm for the prediction of Diabetes Mell-EH-Tiss**”. I am thankful to **Dr. Dalwinder Singh** (HOD, CSE DEPT.) for ensuring that we have a smooth environment in the university by providing us with the best suitable mentors according to our field. I would also like to thank the Research and Development department (R&D department) for providing the opportunity to conduct the research work.

I would like to thank my guide **Kundan Munjal, Assistant Professor, CSE Department**, who encouraged and insisted us in the formulation of problem definition & without her valuable guidance and constant inspiration it would have been difficult for us to prepare this project report.

DECLARATION BY STUDENT

I hereby declare that the dissertation proposal entitled, “**Comparative Analysis of K-means and Hybrid K-means Clustering Algorithm for the prediction of Diabetes Mell-EH-Tiss**” submitted for the M.Tech. Degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date: _____

Sonal Arora
Regn. No. 11301491

TABLE OF CONTENTS

| Sr. No | TOPIC | PAGE |
|-----------|---|-----------|
| 1. | INTRODUCTION | 1 |
| | 1. Introduction..... | 1 |
| | 1.1 Data Warehousing and Data Mining..... | 3 |
| | 1.2 Learning Strategies..... | 6 |
| | 1.2.1 Supervised Learning Strategy | 6 |
| | 1.2.2 Unsupervised Learning Strategy..... | 7 |
| | 1.3 Diabetes..... | 8 |
| | 1.3.1 Types of Diabetes..... | 8 |
| | 1.4 K-means Clustering Algorithm..... | 9 |
| 2. | LITERATURE SURVEY | 11 |
| 3. | PRESENT WORK | 19 |
| | 3. Present work..... | 19 |
| | 3.1 Scope of the study | 19 |
| | 3.2 Problem formulation | 20 |
| | 3.3 Objectives of the study | 21 |
| | 3.4 Research Methodology..... | 22 |
| | 3.4.1 K-means Clustering Algorithm..... | 24 |
| | 3.4.2 Proposed Algorithm..... | 25 |
| | 3.4.3 Dataset Used..... | 27 |
| 4. | RESULT AND DISCUSSION | 31 |
| | 4. Result and discussion..... | 31 |
| | 4.1 Pre-processing of Dataset..... | 31 |
| | 4.2 Time comparison of Algorithms..... | 36 |
| | 4.3 Analysis of the Algorithm..... | 40 |
| 5. | CONCLUSION AND FUTURE SCOPE | 43 |
| 6. | REFERENCES | 44 |
| 7. | APPENDIX | 46 |
| | 7.1 Questionnaires | 46 |
| | 7.2 List of Abbreviations..... | 47 |

LIST OF FIGURES

| FIGURE.NO | TOPIC | PAGE |
|-----------|---|------|
| 1. | Data Warehouse Architecture | 4 |
| 2. | KDD Process of Data Mining | 6 |
| 3. | Formation of Clusters | 7 |
| 4. | General formation of clusters In K-means algorithm..... | 9 |
| 5. | Flow Chart of K-means Clustering Algorithm..... | 10 |
| 6. | Pictorial Process K-means Clustering Algorithm..... | 24 |
| 7. | Flow Chart describing Methodology used in the Research..... | 30 |
| 8. | Weka Tool..... | 31 |
| 9. | Pre-processing filters in Weka | 32 |
| 10. | Attributes of Dataset | 33 |
| 11. | Dataset before Pre-processing..... | 34 |
| 12. | Dataset after Normalization..... | 34 |
| 13. | Dataset after Replacing Missing Values..... | 35 |
| 14. | Simple K-means Clustering Algorithm..... | 36 |
| 15. | K-means Clustering Algorithm with reduced iteration..... | 37 |
| 16. | K-means Clustering Algorithm with reduced iteration having subsamples..... | 37 |
| 17. | Setting parameters for the Algorithm..... | 38 |
| 18. | K-means Clustering Algorithm with reduced iteration for numcluster=10..... | 39 |
| 19. | K-means Clustering Algorithm with reduced iteration having subsamples for numcluster=10..... | 39 |

| | | |
|-----|---|----|
| 20. | K-means Clustering Algorithm with reduced iteration for numcluster=20..... | 40 |
| 21. | K-means Clustering Algorithm with reduced iteration having subsamples for numcluster=20..... | 41 |
| 22. | Graph depicting the time of all algorithms..... | 41 |
| 23. | Graph depicting the time of all algorithms with 10 clusters.... | 42 |
| 24. | Graph depicting the time of all algorithms with 20 clusters.... | 42 |

LIST OF TABLES

| TABLE NO | TOPIC | PAGE |
|----------|---|------|
| 1. | Diabetes dataset attributes..... | 28 |
| 2. | Values of the dataset for the attributes..... | 29 |
| 3. | Description of the attributes taken in dataset..... | 41 |

CHAPTER 1

INTRODUCTION

Introduction

The major challenge faced by the various healthcare organizations like high technology hospitals and many medical centres is the delivery of standard services at cheaper costs which can be afforded by ever individual [10]. The decisions made in medical environment that give positive result are totally depends doctor's perception and knowledgeable information present in clinical databases. Some sorts of decisions are necessarily taken which may be in fact not a good unsympathetic decision and can lead to catastrophic results which are intolerable [11]. Most of sanatoriums use their clinical data to diagnosis patients and producing positive results on patients. The knowledge hidden in the clinical databases procures us to making a poor decision in patient diagnosis. Clinical decision support integrated with computer generated patient records could enhance patient safety, diminish medical errors, improve victim outcome and reduce unfavourable practice variation [13]. This hidden information may utilize to diagnosis a patient who is suffering from particular diseases e.g. Heart Diseases and also predict the disease on the basis of symptoms.

Any doctor can use hidden information to treat his patient. Finding of evidence from knowledge extract from clinical databases is great job in front medical persons. EBM uses a data mining technology that makes it possible to automatically analyze huge clinical Databases and to discover patterns behind them [1]. The integration of evidence-based medicine rules into clinical decision-support systems would both improve quality and reduce costs of care, by recommending guidelines for only the most efficient treatments and medications. Internal clinical experience in integration with the external clinical expertise must be accessible to the healthcare specialists at the appropriate time and in the appropriate manner.

Data warehousing and Data Mining offers a comprehensive support for gathering, analyzing and presenting medical data. Clinical decisions are often made using the doctor's prescription and experience in his or her field rather than the knowledge base which is rich in data hidden in the database [11]. Usually such

practices results in errors, wrong advice to the patients in case if the doctors are in fatigue and stress, unwanted biases and also leads to the extravagant medical price which directly affects the quality of services provided to patients [11]. The benefits of such system would be as follows:

- i. **Quality of care:** EBGGM relies entirely on proven medical experience. Incorporating EBM into the clinical decision-making process ensures that only impartial and scientifically verified knowledge will be offered to physicians.
- ii. **Interoperability:** In order to give the best possible care to the patient, physicians need insight into the patient's complete health record. During their lifetime, patients receive healthcare from diverse medical institutions that keep the records of the individual treatments for a mandatory period of time. It creates a unique patient's health record which is then made available to the decision makers in a user-friendly manner.
- iii. **Decentralized Data Storage:** An electronic health record (EHR) is not necessarily stored as a single physical entity in a centralized system. Instead, it can be aggregated into a single coherent record from data stored at various geographical locations, when required.
- iv. **Scalability:** A federated DWH utilizes a component-based architecture. Each new data source can be easily included into the federation without redesign of the existing system.
- v. **Adoption of International Standards:** Our approach recommends the application of internationally adopted standards in order to enable seamless transmission and understanding of healthcare data among health providers.

Some hospitals deploy decision support systems, but they are generally limited in number. They can only provide the answer to the simple queries like "What is the age of person who is having a particular disease", "How many surgeries have taken place in the hospital in the last few days" or "How many patients are admitted in the hospital in the last 10 days" [1]. But, the existing decision support system cannot answer complex queries like "Whether the patient is suffering from the cancer or not", "If the patient is suffering from cancer then which medicine should be prescribed for the particular patient's cancer stage and case" or "Whether the person suffering from the cancer should be given with the treatment of chemotherapy" [10].

1.1. Data Warehousing and Data Mining

The term data warehouse was introduced by W. H. Immon as: "A Data Warehouse is a subject-oriented, time-variant, integrated, and non-volatile collection of data in support of management's decision making process" [3].

In contrast to traditional On-Line Transaction Processing (OLTP) applications, decision support usually places some different essentiality on database technology. Data warehouses provide storage, a sense of functionality and responsiveness to queries that are beyond the capacity of OLTP databases [5]. They contain historical, summarized, and consolidated data over probably extensive duration of time. Their size can be from hundreds of gigabytes to terabytes. There is a great need to provide decision-makers at all levels of management with information at the desired level of detail, to support their decision-making. Apart from performing regular, predefined reporting activities, a number of parallel users are submitting ad hoc and complicated queries. These queries require access to huge amount of records and cause numerous scan, join, and aggregate operations across the warehouse and results in query throughput and response times which are main issues in multi-user decision-support systems [4].

Figure 1 given below describes the architecture of Data Warehouse, ranging from source data collection to data delivery and then to the decision- makers. The source data is usually reserved in different source system having different formats.

- During the *extraction phase*, source data is collected from operational systems or external sources. The external data is often stored into spreadsheets, personal databases, web logs etc. It can be accessed directly or indirectly (in case of recovery systems).
- In the *transformation phase*, data which is collected is cleaned and converted into the specific format and made structure compatible with the DWH. Several Syntactic and semantic distinctions between operational sources of data are adjusted and then local logical models are portrayed and integrated into the global DWH data model. The mapping characteristics are captured and stored in the DWH as metadata.
- In the *data storage phase*, new data which is extracted and transformed is loaded into the data warehouse and integrated with the existing stored data. During this phase, data is usually restructured for optimized querying. Data loads need to be run on a regular basis in order to keep warehouse data accurate.

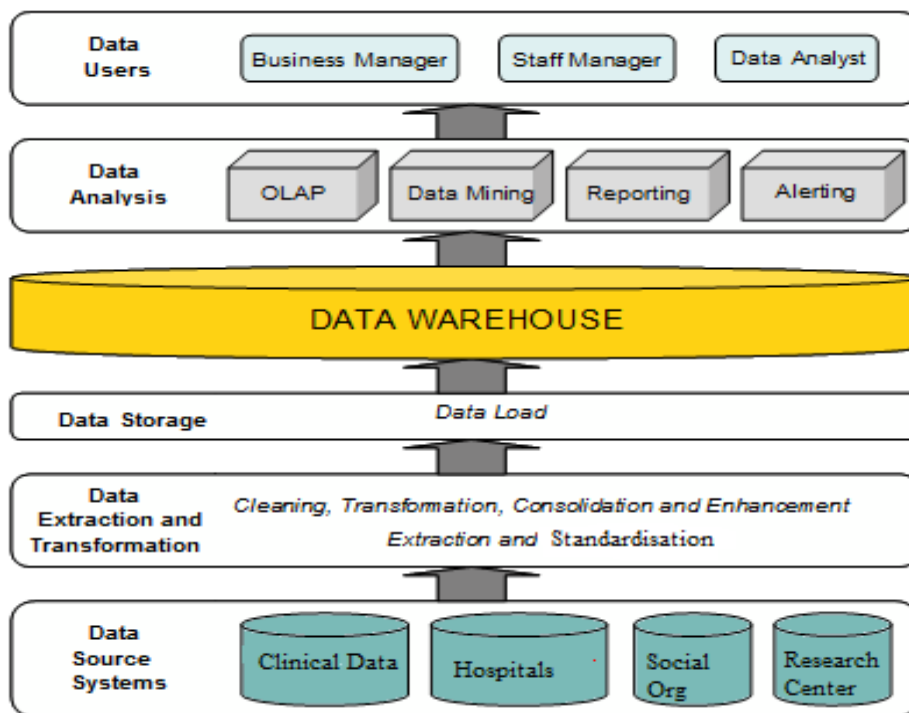


Figure 1: Data Warehouse Architecture

The term data mining generally refers to the process of automatically examining large databases to extract useful and hidden patterns. The same way knowledge discovery in broad area artificial intelligence which is also known as statistical analysis or machine learning, the concept of data mining evolves to discover hidden rules and patterns from the huge amount of data. The term Data mining is different from machine learning and statistics in the way that it manages large volumes of data, which is primarily stored on disk. That is, data mining deals with “knowledge discovery in databases.”

The extraction of non trivial, implied, hidden and actionable knowledge from large volume of datasets is performed by data mining. Discovery of the useful knowledge that can improve proficiency of processes cannot be handled manually.

The aim of Data warehouses is to collect and archive important operational data and pure data. Data Warehouses are used to predict similar trends instances in the data and is used for decision support and performing the analysis on historical data and legacy data. One of the major tasks in data warehousing is to perform data cleansing on the data collected from various input sources. Schemas in data warehousing tends to be in

multidimensional form, which involves one or a few some large fact tables and some amount of smaller dimension tables.

The coined term warehousing is evolving in the area of EBM, thereby, creating evidence-based rules and based on that rules it delivers evidence-based guidelines to the patients and alerts at the point where care of patient is needed and to support clinical and business strategies, thereby, providing generation of either standing or ad-hoc reports for healthcare expertise like clinicians, physicians, nurses and other decision makers.

An online analytical processing (OLAP) tool provides the help to analysts so that they can view data in a summarized form in different ways and formats and with that view they can peep and view the proper functioning of an organization.

- OLAP tools work on data present in multidimensional form which is characterized by several attributes such as dimension attributes and measure attributes.
- The data cube comprises of data in multidimensional form which provides summarized information in different ways. Pre-computing of that data cube helps to speed up queries performed on the summaries of data.
- Cross-tab displays allow the users to view at a time only the two dimensions out of multidimensional data, along with the summaries of that data.
- There are the several basic operations that users perform with OLAP tools like Slicing and Dicing, Drill down and Roll up operations.

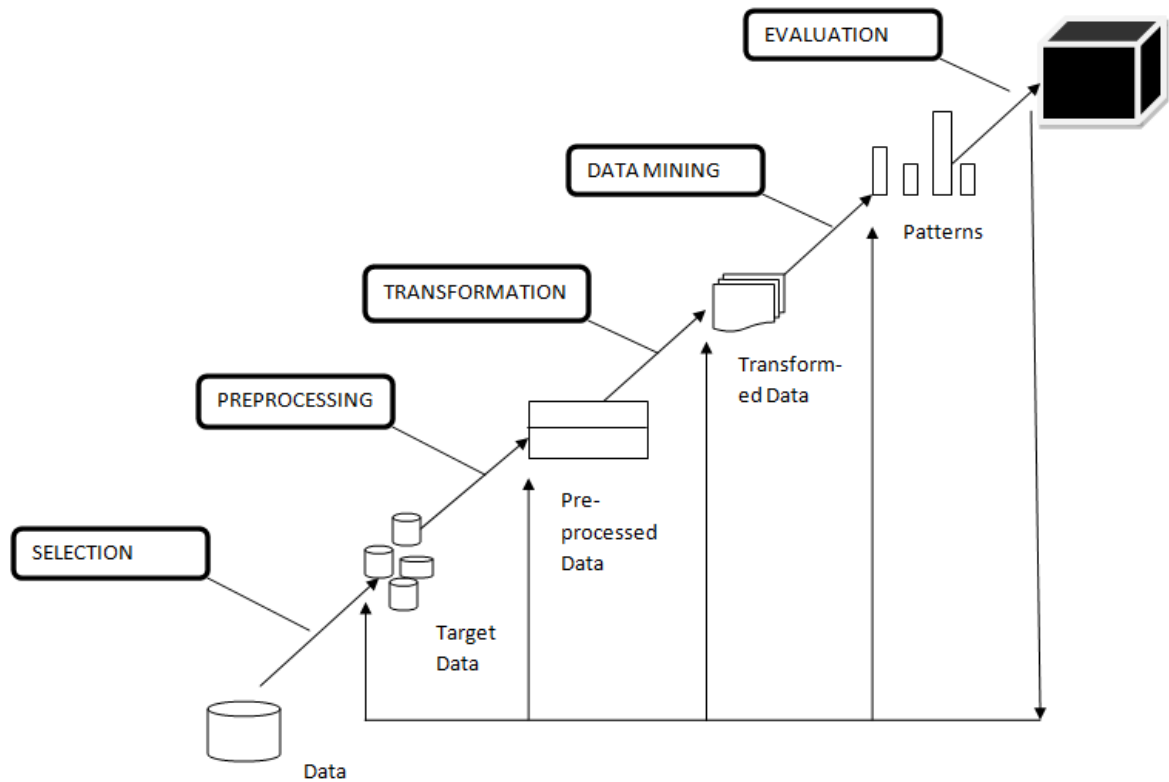


Figure 2: KDD Process of Data Mining

1.2. Learning Strategies

Data mining usually implements two types of strategies:

- Supervised Learning Strategy
- Unsupervised Learning Strategy

1.2.1 Supervised Learning Strategy

In a supervised learning method, a training set is already available which is used to learn parameters. Classification algorithm uses supervised learning method approach. Each of these data mining techniques uses a different approach depending upon the purpose of modeling objective [10]. There are usually two common modeling objectives viz. Classification and Prediction. Classification model predicts the categorical data that is in discrete and unordered form whereas prediction model predicts the continuous valued data [10]. Decision Trees and neural networks uses Classification algorithm. Decision Tree algorithms includes CART (Classification and Regression Trees), ID3 (Iterative

Dichotomized 3) and C4.5 whereas prediction algorithm uses Regression, Association rules and Clustering algorithms [10]. Even though Decision Trees handle discrete data but they continuous data can also be handled provided that data must be converted to categorical data. Neural network consists of three layers- input layer to taken the input, hidden layer for processing and output layer to determine the output and there is a weight assigned on each of the particular input unit.

1.2.2 Unsupervised Learning Strategy

In unsupervised learning method, no training set is available to learn the parameters. Clustering algorithm uses unsupervised learning method approach. There are various clustering algorithms like K-mean clustering algorithm, K-mediod algorithm, DBSCAN and OPTICS, hidden markov algorithm. Unsupervised learning provides the capability to learn more larger and complex models. In unsupervised learning strategy, the learning can be preceded in hierarchical fashion from the observations resulting into ever more deeper and abstract levels of representation.

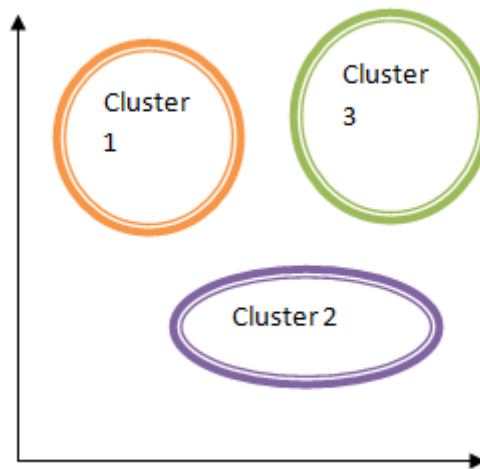


Figure 3: Formation of Clusters

1.3. Diabetes

Diabetes is known disease since ancient times. Diabetes is a disease that is a metabolic disorder in which person has high traces of glucose in the blood caused by inadequate production of glucose content in the body because the body cells do not respond the way

they have to actually respond to insulin [16]. If the trace level of glucose increases in the blood then it will be specified by the various symptoms such as heavy thirst, frequent urination, unexplained weight loss etc. [16].

1.3.1 Types of Diabetes Mell-EH-Tiss

Three types of Diabetes are there in humans:

a) TYPE-1 DIABETES

Type-1 diabetes is caused when the human cells do not produce insulin. Usually Type-1 diabetes affects the people in early adulthood or before the age 40. About 5-10% of the people around the world have been affected with type-1 diabetes [16]. This type of diabetes is also known as juvenile diabetes, insulin-dependent diabetes or early-onset diabetes [17].

b) TYPE-2 DIABETES

Type-2 diabetes occurs when human body cells do not in response with insulin or insulin confrontation. Around 90% of all cases of diabetes all over the world are of this type. Normally, type-2 diabetes affects order individuals. They are usually found in people having more age [16].

c) GESTATIONAL DIABETES

This kind of diabetes normally affects the females during pregnancy [16]. This is caused when some women is determined having very high levels of glucose content in their blood, and their bodies are not capable to produce enough insulin which is required to transport all the glucose into their cells which results in constantly rise in the level of glucose and their detection is made during pregnancy [17]. The gestational diabetes can be controlled by a majority of people by taking the adequate amount of diet and regularly doing exercise. From the overall gestational diabetic patients around 10% to 20% of them will need to take some kind of glucose controlling medications. If the gestational diabetes is treated as undiagnosed or uncontrolled then it can raise the risk of complications during childbirth [17].

1.4. K-means Clustering Algorithm

K-means clustering algorithm is an unsupervised method of learning. By the term unsupervised learning we mean that even though if the input is known, output is unknown. The basic terminology used in K-means clustering algorithm is that it partitions n observations that are n data items into k clusters in which each data item belongs to the cluster having the nearest mean. According to this, in the beginning we will determine “K” that is, the number of clusters and we will assume the centroid or center of these clusters using the distance between two objects and taking their mean value.



Figure 4: General formation of clusters

In K-means algorithm

At the initial phase we can take any random objects and then apply the K means clustering algorithm which will do the three steps given below until convergence occurs. The convergence will occur until STABLE = no object in the group:

- Determine the centroid point.
- Compute the distance of each object to the centroid to form the groups.
- Group the object based on minimum distance and form the cluster.

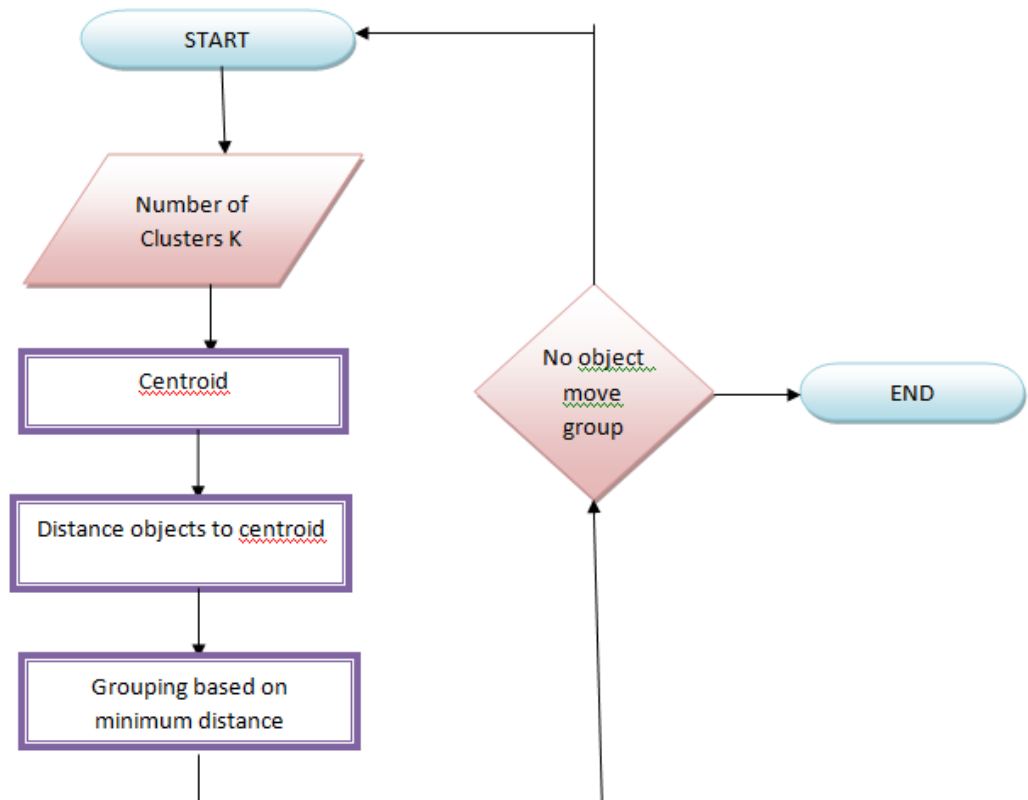


Figure 5: Flow Chart of K-means Clustering Algorithm

There are several disadvantages of K-means Clustering algorithm. Some of them are:

- How we are going to initialize the means in K-means clustering algorithm is not specified. The basic terminology used is to randomly choose the K of the samples.
- The result is dependent on the value of K.
- The results produced will be dependent on the initial value of the means which are calculated.

CHAPTER 2

LITERATURE SURVEY

Literature survey

Andrea Bei, Stefano De Luca, Giancarlo Ruscitti, Diego Salamon (2005), “*Health Mining: a Disease Management Support Service based on Data Mining and Rule Extraction*” defines the disease management as a process which is used to control the health care services so as to provide the better quality services to the patients thereby reducing the cost of the treatments used for the procedures. Author in his paper describes the health mining as a decision support system for controlling the harmful use of hospitalization and therapies so that we can effectively make use of standard guidelines and identify the guidelines which are emerged from the various medical practices such as Evidence based Medicine. Author defines the disease management as an integration of various components: the disease which is to be detected, guidelines procedure to be followed for the disease, records selection and adaption, data mining technique used or the extraction of rules and the evaluation of results using the evidence based medicine. The main aim of this study is to extract the rules and then identify the rules in terms of fast evolving technologies and medicines. They have tested the Health-mining using a real test case for an Italian region, Veneto, on the installation of pacemaker and ICD. The author future work in this paper tends to extend the Health mining with other diseases like diabetes mell-EH-tiss.

Latha Parthiban and R. Subramanian (2007), “*Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm*” implements the prediction of Intelligent Heart Disease which is built with coactive neuro-fuzzy inference system (CANFIS) and Genetic algorithm for the prediction of heart disease. They generally use this approach by combining the three terminologies that is, neural network adaptive capabilities and the fuzzy logic qualitative approach which is then further combined with genetic algorithm so as to diagnose whether the disease is present or not. They have taken the heart-disease dataset from the University of California, Irvine which is also available from UCI Machine Learning Repository. Several methodologies are applied on the dataset to pre-process it and cleaned out from noise and missing values to prepare it for classification

process. The performance achieved by implementing CANFIS model was evaluated in terms of training performances achieved and classification accuracies and the results from the conducted research showed that the proposed CANFIS model has great potential in predicting the heart disease.

Sellappan Palaniappan and Rafiah Awang (2008), “*Intelligent Heart Disease Prediction System Using Data Mining Techniques*” developed an Intelligent Heart disease Prediction System (IHDP). The author in this paper has used the various techniques in data mining like Decision Trees algorithm, Naïve Bayes classification algorithm and neural network for the prediction of heart disease that is whether a patient is suffering from heart disease or not . Normal and traditional decision support system cannot answer the complex queries while this system can answer the complex queries like ‘what if’ type queries. The author reviews that this system can predict if or not the patient is suffering from the heart disease on the basis of few medical parameters taken as an input like patient’s age, sex, body mass index, blood pressure and blood sugar. The researcher took dataset from Cleveland Heart Disease database and 909 records were taken and 15 medical attributes were taken into consideration. Dataset was divided into two: Training dataset and testing dataset which was made with 455 records and 454 records respectively. The result of this research shows that if the processing is done on half of the total population then the neural network outcomes with the highest accuracy i.e. 49.34% proceeded by other techniques, Naïve Bayes having 47.58% and Decision Trees having 41.85% but if the processing is done on entire population then Naïve Bayes model has highest accuracy of 86.12% proceeded by Neural Network (85.68%) and Decision Trees (80.4%) accuracy. Finally, the author concluded Naïve Bayes as the most effective model followed by neural network and decision trees approach although all the three techniques extract the patterns on the basis of the predictable state. In the future work of this paper we can also incorporate other techniques in data mining such as the use of association rules and the various clustering algorithms like K-means Clustering Algorithm, K-mediod Clustering Algorithm, OPTICS and DBSCAN algorithm.

Candice MacDougall, Jennifer Percival and Carolyn McGregor (2009), “*Integrating Health Information Technology into Clinical Guidelines*” brings the use of research

based evidence into practice so as to develop clinical guidelines into practice. This paper provides a review of current research on the integration of Health Information Technology (HIT) into clinical guidelines so as to achieve more accurate results.

K.Srinivas, B.Kavihta Rani and Dr. A.Govrdhan (2010), “Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks” uses the application of techniques for data mining in the Healthcare and the prediction of Heart Attacks. The author has deeply examined the use of data mining techniques in classification such as Rule based technique, Decision Tree technique, Naïve Bayes Classification technique and Artificial Neural Network technique for the extraction of huge amount of patterns from the abundant data which is not mined so as to discover the hidden information from the data. The author implemented the data preprocessing tasks and decision making one dependency augmented Naïve Bayes classifier (ODANB) and Naïve credal classifier 2 (NCC2) and then the ODANB is equated with the existing methods that improve the Naïve Bayes with the Naïve Bayes itself. It can predict whether the patient is suffering from a heart disease or not using several medical parameters such as patient’s age, sex, blood pressure and blood sugar etc. In the rule based technique the rules were piled in the database in the form of IF-THEN rules that is the antecedent part resulting in the conclusion part. Decision Tree includes Classification and Regression Tree (CART), Iterative Dichotomized 3 (ID3) and C4.5. The results when compared show that the ODANB is better than the other methods likes Naïve Bayes. The author also concluded that there are several problems and constraints of using different algorithms of data mining. The hidden patterns can be extracted regarding the prediction of heart attack from data warehouses.

Mai Shouman, Tim Turner, Rob Stocker (2012), “Using Data Mining Techniques in Heart Disease Diagnosis and Treatment”, determines the gap resulting from the previous theories of research on diagnosis and treatment of heart disease and introduces a model to systematically close those occurring gaps to discover if we apply techniques of data mining to the heart disease treatment data then it can provide a reliable performance than it is achieved in diagnosing heart disease. In this paper author has used hybrid data mining techniques which includes naïve density, bagging algorithm and support vector machine. In this literature survey, the different datasets which are being used in the

previous year papers using different-different techniques are being defined and their accuracy is measured so as to determine the various algorithms of data mining used in the diagnosis of heart disease. The author uses single type algorithm like Naïve Bayes, Decision Tree, Bagging algorithm and hybrid type algorithm like Fuzzy-AIRS-K-nearest neighbor, Neural network ensembles and then determine the accuracy. The results show that the hybrid type approach is having more accuracy than the single type as the maximum accuracy achieved by using single data mining technique is 84.14% by naïve bayes while the accuracy achieved by using hybrid data mining technique is 89.01% by neural network ensemble. This paper results in the output that heart disease can be predicted with higher accuracy with hybrid data mining techniques.

Bata Sundar V, T Devi and N Saravanan (2012), “*Development of data Clustering Algorithm for Predicting Heart*” finds out the accuracy of the result by using the K-means Clustering Algorithmic Technique for the prediction and diagnosis of Heart disease. It uses two datasets – real and artificial datasets. The real dataset is the dataset taken from the real life patients of hospitals and patients of laboratory tests whereas the artificial dataset is the dataset taken from the UCI machine learning Databases, 2004. The author in this research mainly focuses on the prediction of Heart disease using K-means Clustering in context of Data Mining. The author first pre-processes the dataset which includes the various steps like eviction of duplicate records, finding out the missing values from the dataset, removing the outliers and noise and normalizing the various values which are used to portray the information in the databases. Then the preprocessed heart disease data is taken and clustered using the K-means algorithm. 13 attributes were taken in the dataset and the performance and analysis of the various algorithms like Decision Trees, Naïve Bayes , Neural network and K-means is done. Each algorithm is compared with another algorithm in terms of its accuracy and time taken to predict .According to this research been taken place the highest accuracy is of the K-means with 66.00% and the time taken is 8 second. The second highest accuracy is 39.96% with the shortest time taken that is of

4sec by the neural network. The lowest accuracy is of Decision Tree with 24.73% with time taken of 10 seconds.

Ms. Nishigandha V. Wankhade , Mrs. Madhuri A. Potey (2013), “Transfer Learning Approach for Learning of Unstructured Data from Structured Data in Medical Domain” focuses on the approach to deal with unstructured data. This literature utilizes the approach of bisecting K-means algorithm for the purpose of disease prediction. The author uses a patient pathology report lab report as a dataset which stores different lab test names like hemoglobin, sugar etc. of the four diseases like Diabetes, Lipid profile cholesterol, Liver profile and kidney profile. This research aims to improve the performance of system by transferring knowledge which is learned in one or more multiple tasks and uses this way to improve learning in the related target task. Bisecting K-means Algorithm is a type of Divisive Hierarchical Clustering which follows the procedure to start with one cluster i.e. the root cluster and recursively splits each cluster into a required K number of clusters. The bisecting K-means is an improved version of K-means algorithm since it produces clusters that are healthier than the clusters formed by normal K-means. They provide a solution for identifying the disease to which a patient belongs by processing the patient’s previous as well as current pathology test reports. The author describes the model by taking the input as Disease database and extracts the tests from database thereby creating a single cluster and thereby applying the bisecting K-means Clustering algorithm to obtain four clusters for four diseases. The features are extracted from the pathology reports database and then mapped with the created clusters. Then it will display the disease to which patient belongs. Unstructured data like .doc-lab test reports is successfully dealt with this research study. This work can be extended by doing the classification of videos, social networking challenges and logical inferences.

Nirmala Devi M., Appavu alias Balamarugan. S, Swathi U.V (2013), “An Amalgam ANN to predict Diabetes Mellitus” presents the development of an amalgam model for classifying Pima Indian diabetic database (PIDD). This amalgam model combines K-means with K-Nearest Neighbour (KNN). They compare the results of simple KNN with cascaded K-means and KNN for the same k-values. It uses the following algorithms: KNN Classifier, K-means partitioning and Amalgam KNN. The dataset is taken from UCI Machine learning data repository for diabetes mellitus that is PIDD. This dataset is from Indian Pima Diabetes Datasets. The results are then compared by measuring the statistical measures such as accuracy, sensitivity and specificity and calculated using WEKA tool. For k=5, K-means and KNN has accuracy of 97% while the simple KNN has the accuracy of 73.17% and Amalgam KNN has accuracy of 97.4%. For k=3,

Amalgam KNN has accuracy of 96.87% and simple KNN has accuracy of 72.65%.The author concluded that performance of the algorithm increases if the value of K increases.

M. Durairaj and C. Vijitha (2014), “*Educational Data Mining for Prediction of Student Performance Using Clustering Algorithms*” defines that the biggest challenge is in the educational institutions that are facing the eruptive growth in an educational data and they also uses this data to refine the quality of managerial decisions for more efficient decision making. Since, the educational database contains the useful information which is used for predicting the student’s performance, their rank factor and details. The author conducts the analysis of student details and applies the data mining methods to get vital information. The author in this research develops a model using data mining techniques like K-means and naïve bayes classification algorithm. The researcher collected the real time data from department of Computer Science, Engineering and Applications, Bharathidasan University, Tiruchirappali, India that describes the relationships between learning behavior of students and their academic performance. The data contains the student’s detail of different subject marks in semester wise which is then subjected to data mining process. The author uses WEKA 3.7.9 to implement data mining approach which contains several in-built features like pre-processing of data, classification of data, various clustering techniques and association for extraction of rules. It uses various parameters for evaluating performance like FP rate, TP rate, Precision, Recall F-measure and ROC area. It then compares each parameter stated above for decision tree and naïve bayes and calculates the accuracy of both resulting in the naïve bayes algorithm giving more accurate results than decision tree. Using the K-means clustering algorithm, the author predicted the pass percentage and fail percentage of the overall students appeared for a particular examination.

Gunasekar Thangarasu, Assoc. Prof. Dr. P.D.D. Dominic (2014), “*Prediction of Hidden Knowledge from Clinical Database using Data mining Techniques*” predicts the diabetic disease from clinical database by using Neural Network algorithm. This research was being conducted with the various objectives like it identifies the various complications that cause diabetes from clinical databases through Fuzzy logic Techniques. It develops a Hybrid Genetic Algorithm that computes the best fitness value which is used for evaluating the prediction accuracy of diabetes from clinical databases. It

also identifies the type of diabetes the patient is suffering from through data clustering algorithms from the clinical database. Then, it will analyse the performance of projected algorithms. It uses hybrid system model to identify diabetes mellitus, its types and complications which uses certain algorithms like Neural-network algorithm, Fuzzy logic techniques, Hybrid Genetic Algorithm and Data clustering algorithms. The dataset is collected via questionnaire distribution with participants. All the dataset was organized and analysed using a computer program SPSS 20. This model was successfully implemented with input as symptoms that may appear in an individual during the early stages of diabetes and also based on the physical condition of the individual. This research study avoids the patients from undertaking certain blood tests, checking the diastolic and systolic blood pressure etc. Thereby creating a user friendly interface and environment for the patient's without any requirement of a doctor or hospital staff.

V. Veena Vijayan and Ravikumar Aswathy (2014), “*Study of Data mining Algorithms for prediction and Diagnosis of Diabetes Mellitus*” studies the various data mining algorithms for the prediction and diagnosis of diabetes Mellitus. According to the author several traditional methods are available for diagnosing diabetes based on the several physical and chemical tests which are being performed. The author has chosen the dataset from the Pima Indian Diabetic Set from the University of California for classification and experimental simulation. They took 8 attributes into the consideration. The author surveyed various algorithms and analyzed that Expectation Maximization Algorithm is the simplest algorithm which can be performed by using two steps but the results of this algorithm is less than 70% and author defined that this algorithm is not very accurate for the higher dimensional data sets due to imprecision. The author also used the K-nearest Neighbor Algorithm which is one of the simplest and also named as lazy learning algorithm used for the classification. The accuracy of this algorithm comes out be 73.17% because of certain drawbacks in this algorithm. The more efficient approach used by the author is K-means Clustering algorithm but the accuracy of this approach is also 66-77%. For solving the disadvantages of the K-means clustering algorithm, the author combined it with KNN algorithm and forms the Amalgam KNN which improves the accuracy even for the large datasets. The author observed that the value of K is very crucial as the value of K decreases the accuracy becomes very less and with the increase of value of K, the accuracy is increased. The author then uses the Adaptive Neuro Fuzzy

Inference System (ANFIS) algorithm combined with the adaptive KNN which leads to the accuracy of 80%.

ShravanKumar Uppin and M A Anusuya (2014), “*Expert System Design to Predict Heart and Diabetes Diseases*” designs an expert system that predicts the heart disease and diabetes disease. The author uses reduced number of attributes and then uses data mining technique in which he applied C4.5 classification algorithm so that there is more accuracy and less run time. The author also applies decision tree algorithm for the prediction of heart disease and diabetes and foretells that whether disease is present or not. According to the author the existing method takes 0.05 sec whereas the proposed method took around 0.025 sec and the accuracy is also increased from 84.35% to 85.96%.

CHAPTER 3

PRESENT WORK

Present work

- i. The work has completed till the development of K-means Clustering algorithm. We worked on the development of new efficient algorithm and got the efficient results in those processes. This process has implemented in the Weka 3.6.9.
- ii. The preprocessing of the data takes place which includes several processes like Removing the noise, removing the missing values and normalization of the data.
- iii. We calculated the time taken by the simple K-means Clustering algorithm and then we developed the reduced iterative approach for simple K-means by reducing the number of iterations so as to save time.
- iv. When we used that reduced iterative approach to further enhance the K-means by taking the subsamples of the whole data by going through in this fashion:

Dataset → Samples → K-means (reduced iteration) → Updated K-means

- v. The time taken by all the algorithms was being measured and results were analysed.

3.1 Scope of the Study

- i. The proposed algorithm presents the two new methods such as K-means clustering algorithm and Hybrid K-means Clustering Algorithm.
- ii. This proposed algorithm provides the efficient results on the prediction of diabetes mell-eh-tiss. This proposed method is useful in many areas such as medical fields, supermarket stores or in any real time applications such as student's analysis of report.
- iii. The proposed algorithm presents the time saving methods the advantages of using these methods is that there is time taken for the computation of results is very less.
- iv. The proposed system provides a very economical medical treatment if it is used with the design of Evidence based Guideline System.

- v. Since data is stored electronically, patient freed from keeping prescriptions for hospitals for further treatment.
- vi. Improves the efficiency and effectiveness of antiquated healthcare system.

3.2 Problem formulation

Data Mining has been used in many areas. Data Mining has been popular topic for both researches and applications in the last 10 years.

The problem formulation used for attaining the objectives of this research study is done by the various approaches:

- **Qualitative Approach**

In the qualitative approach we study the various papers related to the clinical databases and how that data is not useful in analyzing or predicting any disease although we have several data mining techniques available like classification techniques and clustering techniques for converting that raw data into useful information that is hidden and to mine that data into relevant data. Generally mostly diseases like Heart disease, Fluoride disease is being predicted using K-means algorithm. Some real time datasets like educational databases for predicting the pass or fail marks of students and analyzing the student's performance are also being designed. Diabetes is also predicted but using amalgam KNN that is a type of classification algorithm and we are achieving the same using clustering algorithm. Hence from the above qualitative approach, we came to know that we can predict the diabetes mellitus which is till now in the recent study.

- **Quantitative Approach**

In the quantitative analysis we are going to design and define the algorithm we are using for prediction using diabetes dataset. From the recent study and approach which is used till now in the literature survey is the several classification techniques like amalgam KNN, Naïve Bayes algorithm, Decision Trees etc. and they are proving the better results and help in mining of the data efficiently and more accurately than any other algorithm can do because they are providing the results with approximately 97.4 % accuracy which is even more harder to improve in the research study. But uncertainly if we look at the clustering algorithm that is being used for the prediction of Heart disease gives the accuracy of about 66% when compared to the above algorithms being implemented. The clustering algorithm that will be implemented is K-means Clustering Algorithm to test if K-means algorithm

when applied on diabetes dataset produces the accuracy of how much (in terms of percentage). Also, we will enhance the accuracy of K-means Clustering Algorithm by improving and comparing the performance from the existing k-means clustering algorithm by using our research methodology in which we are using this technique as we are dividing the dataset into sub samples and then apply clustering with reduced iterations that will take less time as compared to existing algorithm and improves its accuracy.

- **Mixed Approach**

This is basic approach that we are going to implement which includes both the qualitative approach and quantitative approach. The basic terminology that we are going to use by combining both the approaches is predict the diabetes by implementing the simple k-mean clustering algorithm and hybrid K-mean clustering algorithm and then matches the accuracy of both the algorithms in terms of time and efficient computations. In the mixed approach, we will enhance the K-means by modifying the original algorithm so that it can produce better results by using our methodology, that is, dividing the dataset into sub samples and then apply clustering with reduced iterations that will take less time as compared to existing algorithm and improves its accuracy.

3.3Objective of the Study

Objectives of the study provides the researcher the real view point to study and acts as a motto in developing the purpose for doing the job. This research study is done with several objectives to attain and serve the various purposes like:

- i. Collect the artificial dataset or real time dataset for diabetes patients from university or hospital which is to be taken as dataset.
- ii. Preprocessing of the data which is collected will be done which includes the various steps like removing the noise, removing the missing values and normalizing the data.
- iii. Apply the simple K-means Clustering Algorithm and measure its accuracy or efficiency.
- iv. Apply hybrid algorithm for K-means Clustering Algorithm to improve its speed or increase its efficiency by using our approach that will be the proposed methodology.

- v. Analyze the performance of simple K-means and hybrid K-means algorithm that is being designed.

3.4 Research Methodology

The basic terminology used in the development of the simple K-means Clustering Algorithm is that we are randomly taking the data items and assuming that data items itself as a individual clusters .Taking the third data item and we have to see that whether that data item is included in the first cluster or in the second cluster by using the formula of the Euclidian distance and then formulating the clusters by seeing the data item which is having less distance out of the both. Similarity and Dissimilarity between Objects is calculated by:

- Distances are normally used to measure the similarity or dissimilarity between two data objects.
- Some popular ones include: *Minkowski distance*:

$$d(i, j) = \sqrt[q]{(|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + \dots + |x_{i_p} - x_{j_p}|^q)}$$

where $i = (x_{i_1}, x_{i_2}, \dots, x_{i_p})$ and $j = (x_{j_1}, x_{j_2}, \dots, x_{j_p})$ are two p -dimensional data objects, and q is a positive integer

- If $q = 1$, d is Manhattan distance

$$d(i, j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \dots + |x_{i_p} - x_{j_p}|$$

- If $q = 2$, d is Euclidean distance:

$$d(i, j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \dots + |x_{i_p} - x_{j_p}|^2)}$$

- Properties

- $d(i, j) \geq 0$
- $d(i, i) = 0$

- $d(i,j) = d(j,i)$, *symmetry*
- $d(i,j) \leq d(i,k) + d(k,j)$, *triangular inequality*

If there are two data items $c1(a, b)$ and $c2(x, y)$ then the Euclidian distance for the above can be calculated as:

$$dist = \sqrt{(x - a)^2 + (y - b)^2}$$

The data item which is having smaller distance will form one cluster and that new one data items will be included in that cluster and the centroid of both the data items is calculated by using the formula:

$$centroid = \frac{x1 + x2}{2}$$

Where $x1$ and $x2$ are the two data items.

This whole process will be repeated until the convergence criterion is met. This pictorial process of the K-means clustering algorithm can be shown diagrammatically as:

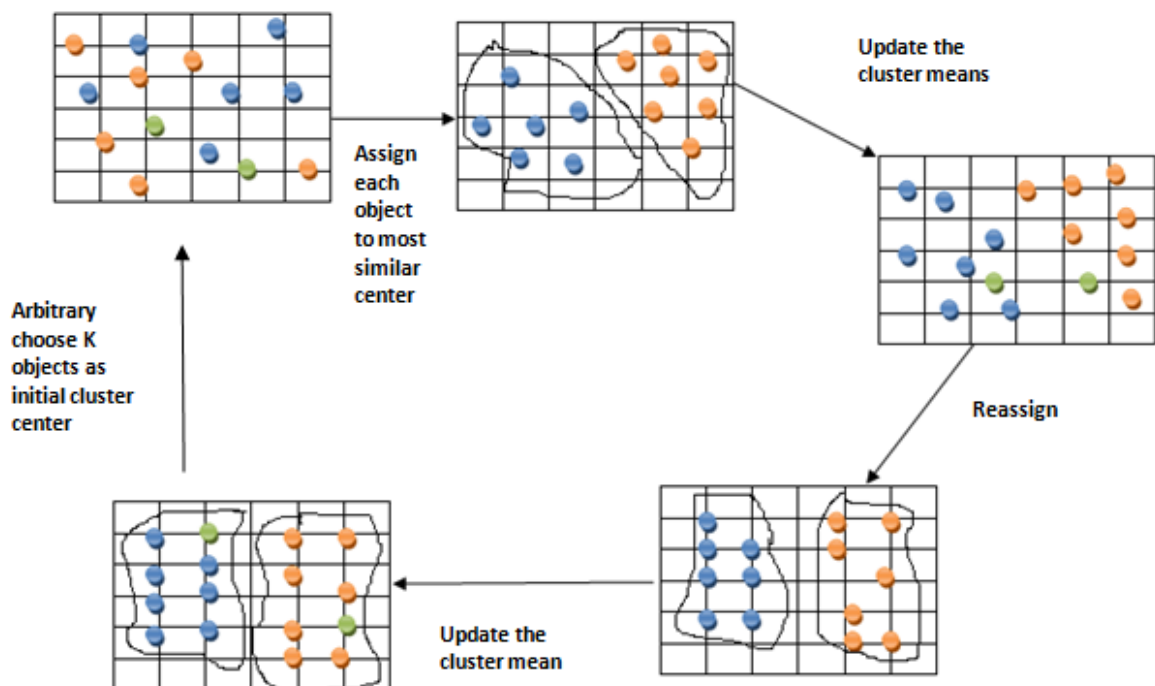


Figure 6: Pictorial Process K-means Clustering Algorithm

I. K-means Clustering Algorithm

The algorithm for the above is:

Input- D that is data set

Output-K that is number of clusters

Step 1: Initialize the data centers as D.

Step 2: Randomly choose K objects from D.

Step 3: Repeat until no change in cluster takes place that is convergence criteria is met.

Step 4: Compare the mean value of the data objects in the clusters for the purpose of initialization.

Step 5: Initialize the object within one of the clusters K which is having the most similar values or having shorter mean value.

Step 6: Take the new mean value of the objects for each of the cluster K.

Step 7: Update the cluster mean with respect to the value which is being generated.

Now, this simple K-means Clustering algorithm has two disadvantages namely:

- k- Means algorithm will calculate the distance from each data object to all the centers of cluster in its each iteration. This will take lot of execution time for large datasets.
- K- Means will take whole dataset at once and make clusters.

The new proposed algorithm will overcome the above two disadvantages of K-means.

The approaches used are:

Approach is used to solve the first disadvantage is to take two data structures **one for cluster and other for distance of data object from every cluster during each iteration**. By doing so, when we calculate the distance between the current data object and the new cluster center, if the computed distance is smaller than or equal to the distance to the old center, the data object will remain in its cluster that was assigned to in previous iteration. Therefore no need to calculate the distance from this data object to other cluster centers will save lot of time. This will reduce the no. of iterations.

Approach to solve the second disadvantage is **to divide the dataset into subsample** then apply the k- means algorithm for each subsample this will reduce the probability of dividing one big cluster into two or more ones owing to the adoption of cluster error criteria.

We have combined two approaches i.e. first dividing the dataset into sub samples and then applying the reduced iteration approach to each subsample to find the clusters. This will save lot of time and improve the performance for the large datasets.

- 1) Split the whole data into various subsamples.
- 2) Then apply the algorithm that will reduce the number of iterations.

II. Proposed Algorithm

The proposed algorithm for the above approach is:

Step 1: Extract subsamples from the dataset

For each extracted subsample

Step 2: Select randomly k objects from the subsamples as initial clusters centers

Step 3: Calculate distance between each data object then assign data object to the nearest cluster.

Step 4: For each data object find the closest center and assign data object to that cluster center.

Step 5:

Store the label of cluster center in which the data object is and the distance of data object to the nearest cluster and store them in array $Cluster[]$ and the $Dist[]$ separately.

Set $Cluster[i] = j$, j is the label of nearest cluster.

Set $Dist[i] = d(d_i, c_j)$, $d(d_i, c_j)$ is the nearest Euclidean distance to the closest center.

Step 6: For each cluster j , recalculate the cluster center;

Repeat

Step 7:

For each data object calculate its distance to the center of the present nearest cluster;

- a) If this distance is less than $Dist[i]$, the data object stays in the initial cluster;
- b) Else

For every cluster center $c_j (1 \leq j \leq k)$, compute the distance $d(d_i, c_j)$ of each data object to all the center, assign the data object d_i to the nearest center c_j .

Set Cluster $[i] = j$;

Set Dist $[i] = d(d_i, c_j)$;

Step 8: For each cluster center $j (1 \leq j \leq k)$, recalculate the centers;

Step 9: Until the convergence criteria is met.

Step 10: Output the clustering results;

Step 11: Combine two nearest clusters into one cluster and recalculate the new cluster center for the combined cluster until the number of clusters reduces into k .

This algorithm is the new updated algorithm for K-means Clustering algorithm.

III. Dataset Used

The dataset used in this research work is collected from National Institute of Diabetes and Digestive and Kidney Diseases and is based on Pima Indian Diabetic Set from University of California, Irvine (UCI) Repository of machine learning databases. The Pima Indian diabetes database, donated by Vincent Sigillito, is a collection of medical diagnostic reports of 768 examples. Before 1694, they referred to themselves as OTAMA. The name PIMA is thought to have derived from communication problems between Europeans and members of the Otama tribe. Earlier this dataset was used by:-

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). The diagnostic, binary-valued variable investigated is whether the patient shows signs of diabetes according to World Health Organization criteria (i.e., if the 2 hour post-load plasma glucose was at least 200 mg/dl at any survey examination or if found during routine medical care). The population lives near Phoenix, Arizona, USA.

Results: Using 576 training instances, the sensitivity and specificity of their algorithm was 76% on the remaining 192 instances.

Relevant Information:

Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. The patients whose data is being taken are from Pima Indian Population living in Arizona, USA.

Number of Instances: 768

| sno | attribute | type | mean | Standard deviation |
|-----|--|------|-------|--------------------|
| 1 | Number of times pregnant | Real | 3.8 | 3.4 |
| 2 | Plasma glucose concentration a 2 hours in an oral glucose tolerance test | Real | 120.9 | 32.0 |
| 3 | Diastolic blood pressure (mm Hg) | Real | 69.1 | 19.4 |
| 4 | Triceps skin fold thickness (mm) | Real | 20.5 | 16.0 |
| 5 | 2-Hour serum insulin (mu U/ml) | Real | 79.8 | 115.2 |
| 6 | Body mass index (weight in kg/(height in m) ²) | Real | 32.0 | 7.9 |
| 7 | Diabetes pedigree function | Real | 0.5 | 0.3 |
| 8 | Age (years) | real | 33.2 | 11.8 |

Table 1: Diabetes dataset attributes

The real time dataset for the diabetes patients is also being taken from the Metro Hospital, Faridabad. We have taken eight parameters and one parameter to check whether the patient is tested positive or negative.

| | A | B | C | D | E | F | G | H | I | J |
|----|------|------|------|------|------|------|-------|-----|-------|-----------------|
| 1 | preg | plas | pres | skin | insu | mass | pedi | age | class | |
| 2 | 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | | 50 | tested_positive |
| 3 | 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | | 31 | tested_negative |
| 4 | 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | | 32 | tested_positive |
| 5 | 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | | 21 | tested_negative |
| 6 | 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | | 33 | tested_positive |
| 7 | 5 | 116 | 74 | 0 | 0 | 25.6 | 0.201 | | 30 | tested_negative |
| 8 | 3 | 78 | 50 | 32 | 88 | 31 | 0.248 | | 26 | tested_positive |
| 9 | 10 | 115 | 0 | 0 | 0 | 35.3 | 0.134 | | 29 | tested_negative |
| 10 | 2 | 197 | 70 | 45 | 543 | 30.5 | 0.158 | | 53 | tested_positive |
| 11 | 8 | 125 | 96 | 0 | 0 | 0 | 0.232 | | 54 | tested_positive |
| 12 | 4 | 110 | 92 | 0 | 0 | 37.6 | 0.191 | | 30 | tested_negative |
| 13 | 10 | 168 | 74 | 0 | 0 | 38 | 0.537 | | 34 | tested_positive |
| 14 | 10 | 139 | 80 | 0 | 0 | 27.1 | 1.441 | | 57 | tested_negative |
| 15 | 1 | 189 | 60 | 23 | 846 | 30.1 | 0.398 | | 59 | tested_positive |
| 16 | 5 | 166 | 72 | 19 | 175 | 25.8 | 0.587 | | 51 | tested_positive |
| 17 | 7 | 100 | 0 | 0 | 0 | 30 | 0.484 | | 32 | tested_positive |
| 18 | 0 | 118 | 84 | 47 | 230 | 45.8 | 0.551 | | 31 | tested_positive |
| 19 | 7 | 107 | 74 | 0 | 0 | 29.6 | 0.254 | | 31 | tested_positive |
| 20 | 1 | 103 | 30 | 38 | 83 | 43.3 | 0.183 | | 33 | tested_negative |
| 21 | 1 | 115 | 70 | 30 | 96 | 34.6 | 0.529 | | 32 | tested_positive |
| 22 | 3 | 126 | 88 | 41 | 235 | 39.3 | 0.704 | | 27 | tested_negative |
| 23 | 8 | 99 | 84 | 0 | 0 | 35.4 | 0.388 | | 50 | tested_negative |
| 24 | 7 | 196 | 90 | 0 | 0 | 39.8 | 0.451 | | 41 | tested_positive |
| 25 | 9 | 119 | 80 | 35 | 0 | 29 | 0.263 | | 29 | tested_positive |

Table 2: Values of the dataset for the attributes

The terminology we have used is that we have collected the diabetes patient's dataset and pre-processing of that dataset takes place through filtration which includes removal of noise and removal of missing values. After that K-means clustering algorithm is applied and the results are seen. When the methodology of K-means clustering algorithm with reduced iterations and K-means clustering algorithm with reduced iterations and subsamples are applied then also result is analyzed. The efficiency of the algorithms is measured in terms of time taken in seconds.

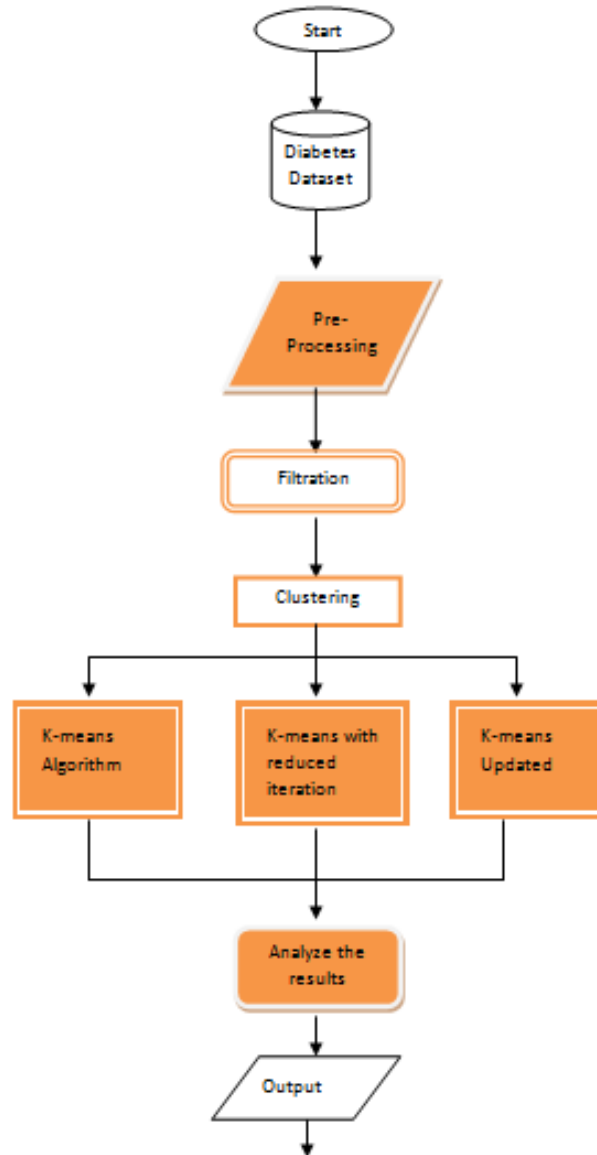


Figure 7: Flow Chart describing Methodology used in the Research

Result and Discussion

This proposed algorithm has implemented in the Weka 3.6.9. Weka tool provides the machine level support for performing various operations like pre-processing of data, clustering and classification of data.

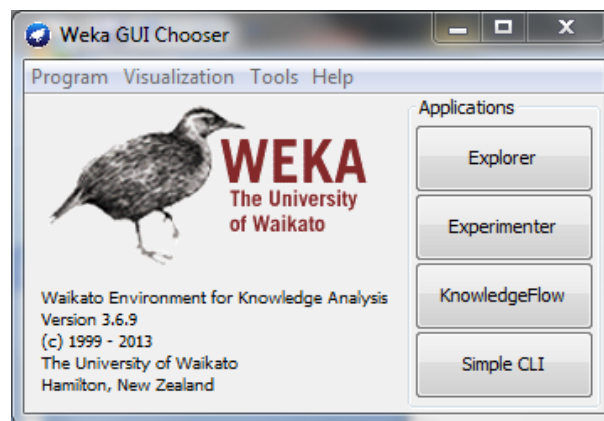


Figure 8: Weka Tool

4.1 Pre-Processing of dataset

While performing the data mining pre-processing of the data items is the first step. Since, the data in the real world is dirty data that is it contains certain missing values, certain anomalies are there in the data or the data which is present contains noisy data due to which there are errors in the result. Hence, for that purpose normalization of the data takes place. We have to remove the missing values and clean that data so that it can be used for mining. Weka provides certain filters like remove noise, normalize, replace missing values, string to binary conversion, string to literal conversion and many more. All these are filtration methods provided by the weka both under the category of supervised and unsupervised.

The way to select the filter and the various types of filters are shown as below:

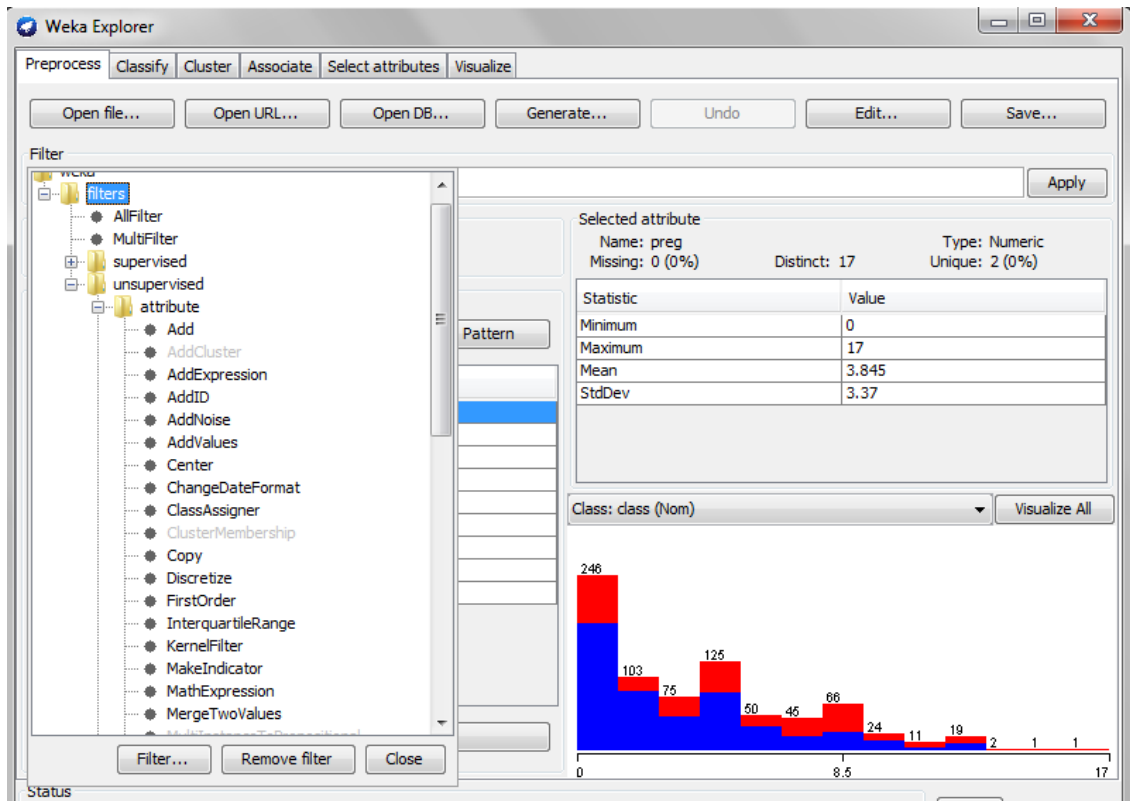


Figure 9: Pre-processing filters in Weka

The data set taken for the diabetic patients and description of the attributes which is listed is shown in the table no. 3:

| | |
|----------|--|
| pregnant | Number of times pregnant |
| glucose | Plasma glucose concentration (glucose tolerance test) |
| pressure | Diastolic blood pressure (mm Hg) |
| triceps | Triceps skin fold thickness (mm) |
| insulin | 2-Hour serum insulin (mu U/ml) |
| mass | Body mass index (weight in kg/(height in m) ²) |
| pedigree | Diabetes pedigree function |

| | |
|----------|------------------------------------|
| age | Age (years) |
| diabetes | Class variable (test for diabetes) |

**Table 3: Description of the attributes
Taken in dataset**

The graph depicting the various attributes of the dataset is shown below:

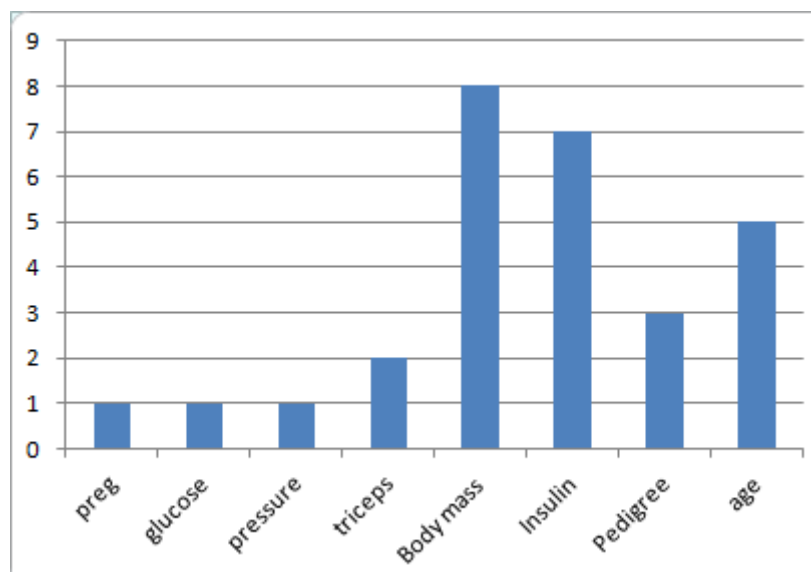


Figure 10: Attributes of Dataset

Now, the pre-processing of dataset takes place which includes normalization and replacing the missing values.

The graphical representation of the attributes before the pre-processing is as shown in figure:

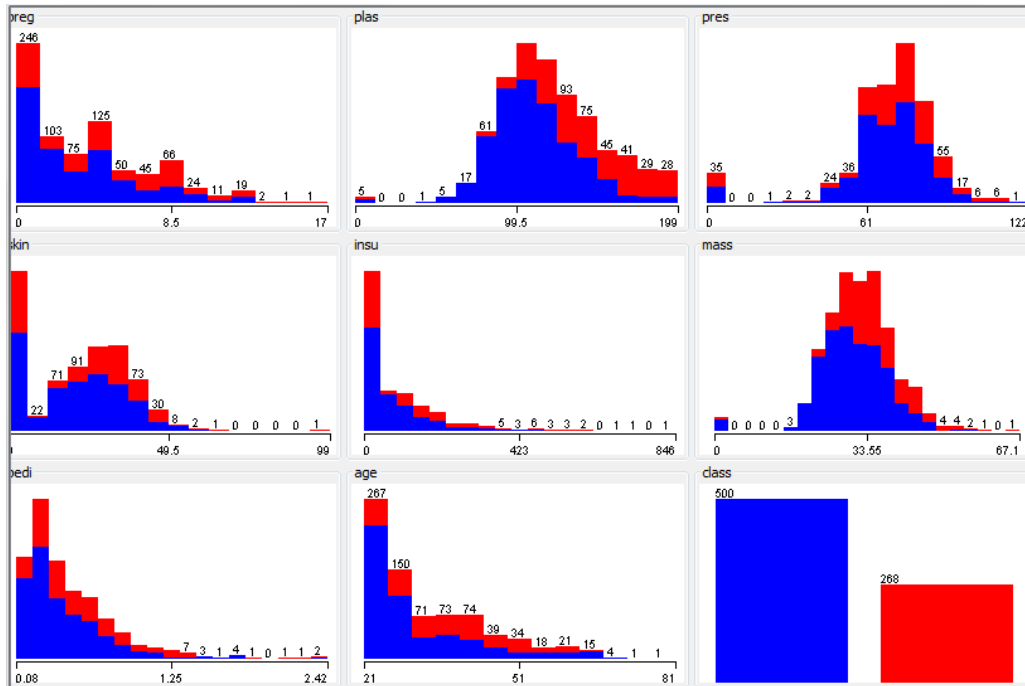


Figure 11: Dataset before Pre-processing

The pre-processing of data takes place which includes various steps like:

a) Normalization

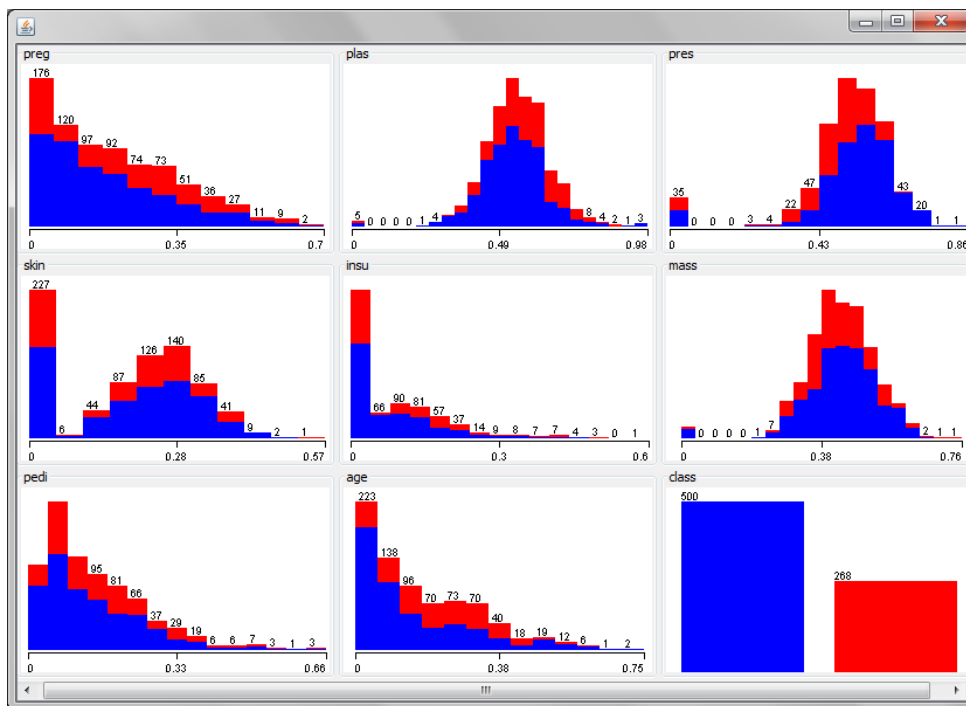


Figure 12: Dataset after Normalization

Normalizes all numeric values in the given dataset (apart from the class attribute, if set). The resulting values are by default in $[0, 1]$ for the data used to compute the

normalization intervals. But with the scale and translation parameters one can change that, e.g., with scale = 2.0 and translation = -1.0 you get values in the range [-1, +1].

b) Replace Missing values

This filter will replace all missing values for nominal and numeric attributes in a dataset with the modes and means from the training data.

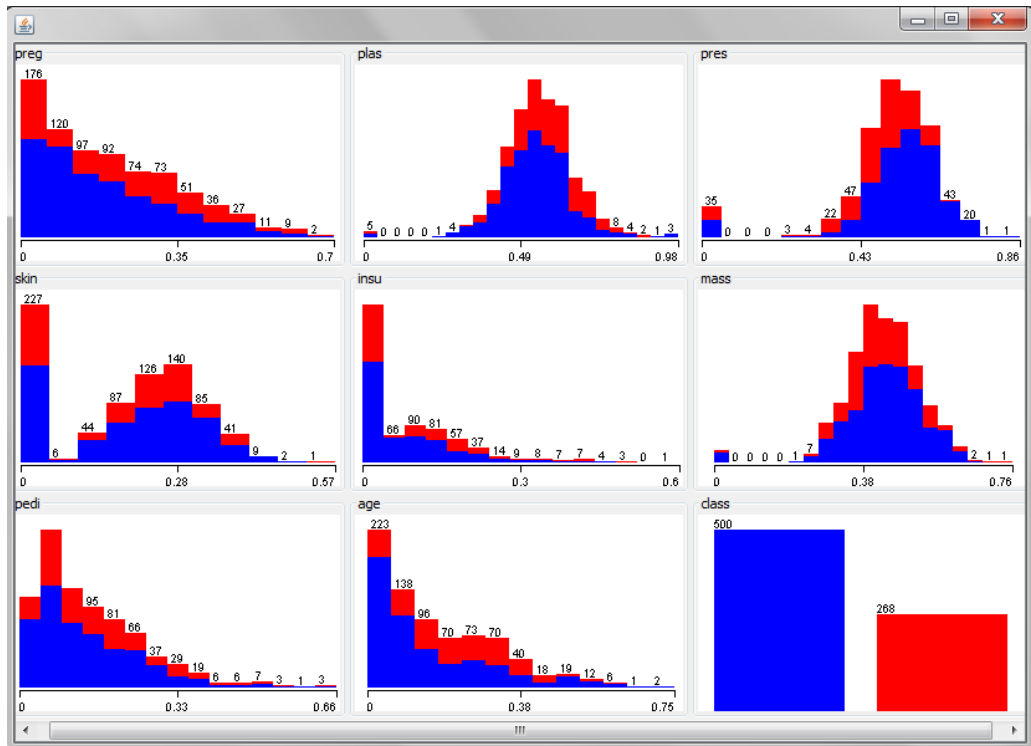


Figure 13: Dataset after Replacing Missing Values

4.2 Time Comparison of the Algorithms

Various algorithms were designed and the time taken by them for the computation of the algorithm was taken into consideration for evaluating the dataset of the diabetes patients given as the input. The simple K-means Clustering algorithm took around 163.99 seconds to evaluate the model.

```

Clusterer output
mass
  mean      18.561  37.3254  34.9624  31.7416  28.9775  31.0445  35.368  29.4171  36.6186
  std. dev. 13.8774  8.7732  5.5031  5.3704  5.9071  7.0182  6.2361  10.677  7.3488

pedi
  mean      0.3203  0.7262  0.613  0.3705  0.4119  0.3692  0.487  0.6999  0.4946
  std. dev. 0.2175  0.5354  0.3298  0.2068  0.2436  0.2321  0.2919  0.4798  0.2621

age
  mean      30.1742  28.3448  43.9476  43.7547  23.6227  36.6204  27.7527  56.2464  31.5495
  std. dev. 11.2186  5.3519  8.1344  11.0816  2.5065  11.944  4.8958  11.1521  6.6209

class
  tested_negative 18.6122 11.1036 11.1079 68.9642 184.4546 114.3313 88.074 10.1124 2.2398
  tested_positive 4.4756 72.5373 88.0818 6.8737 10.0472 65.0047 6.0009 10.5016 13.4773
  [total]        23.0878 83.6409 99.1896 75.8379 194.5018 179.336 94.0749 20.614 15.7171

Time taken to build model (full training data) : 163.99 seconds

=== Model and evaluation on training set ===

Clustered Instances
0      17 ( 2%)
1      76 (10%)
2     104 (14%)
3      73 (10%)
4     208 (27%)
5     147 (19%)
6      83 (11%)
7      42 ( 5%)
8      18 ( 2%)

```

Figure 14: Simple K-means Clustering Algorithm

The results when compared if we are performing K-means with the reduced iterations approach and the K-means with merged reduced iterations and subsamples that is the updated K-means is as depicted below in figure 15 and figure 16:

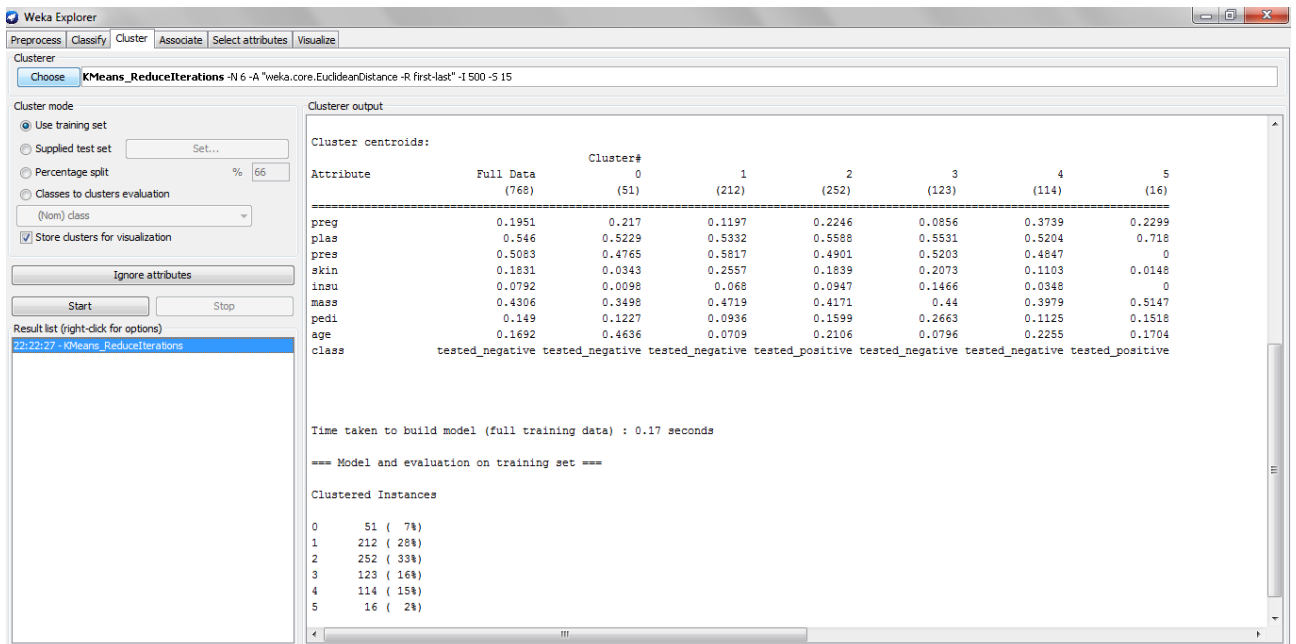


Figure 15: K-means Clustering Algorithm with reduced iteration

The depicted results when compared shows that the time taken by K-means with reduced iterations is 0.17 seconds and that of K-means with merged reduced iterations and subsamples is 0.09 seconds.

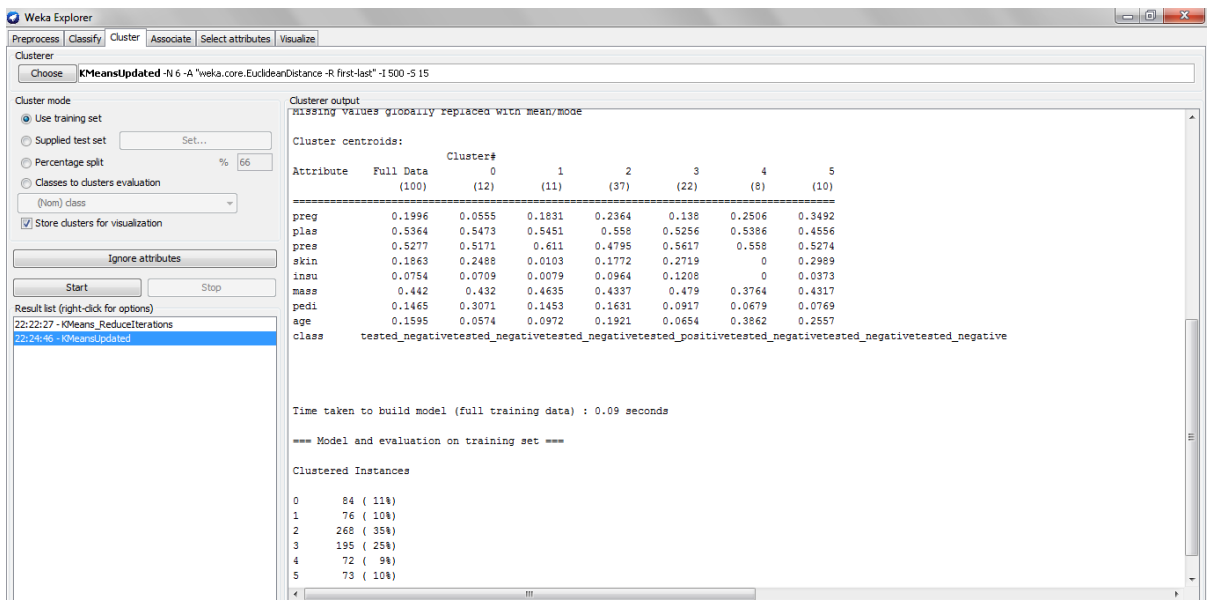


Figure 16: K-means Clustering Algorithm with reduced iteration

Having subsamples

These results were taken with the default setting of the weka as shown below:

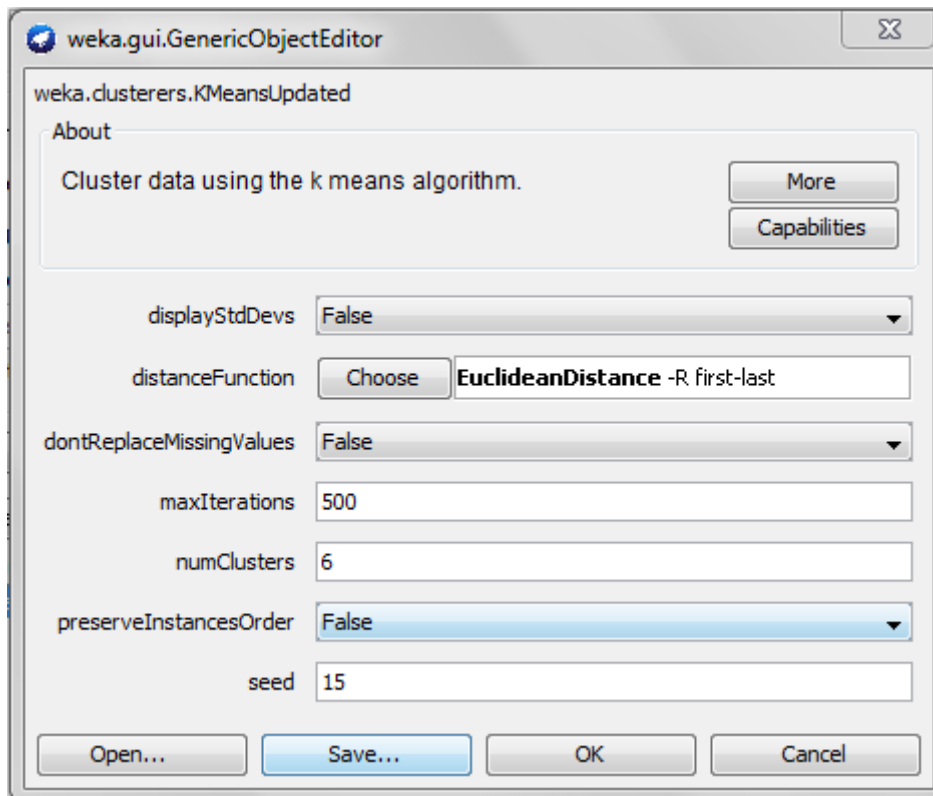


Figure 17: Setting parameters for the Algorithm

The results are taken with the number of clusters 6 and choosing the Euclidean distance as default and the display of standard deviation of these parameters set to off. If we are changing the parameter numClusters then the result for time taken to compute the model varies as shown in figure 18 and figure 19:

Now, if we increase the no. of clusters to 10 and then apply the K-means with reduced iterations and having 10 clusters, the time taken is 0.37 seconds as shown in figure 18 while for K-means with reduced iterations and subsamples having 10 clusters is 0.13 sec as shown in figure 19:

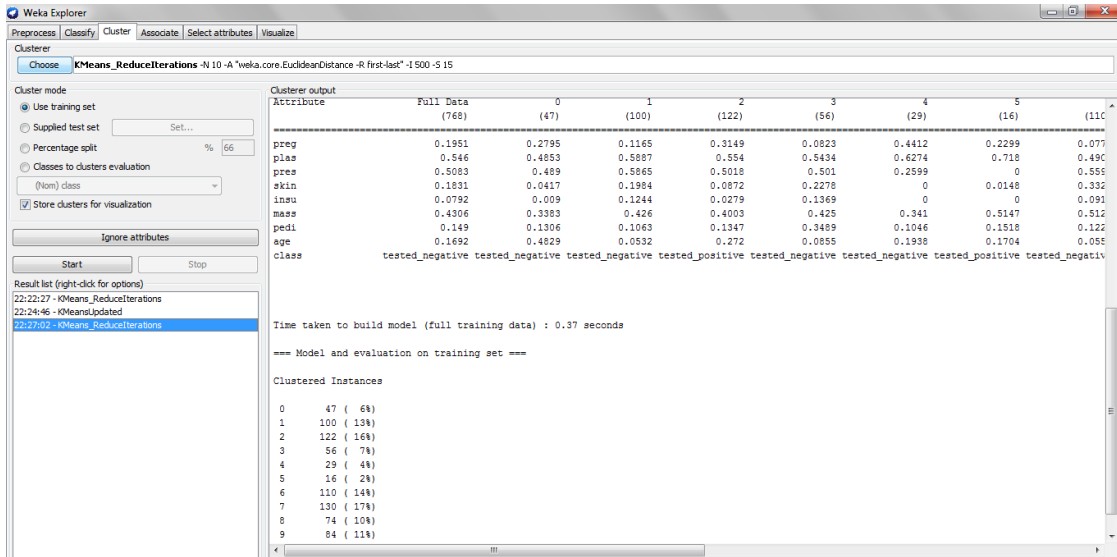


Figure 18: K-means Clustering Algorithm with reduced iteration

For numcluster=10

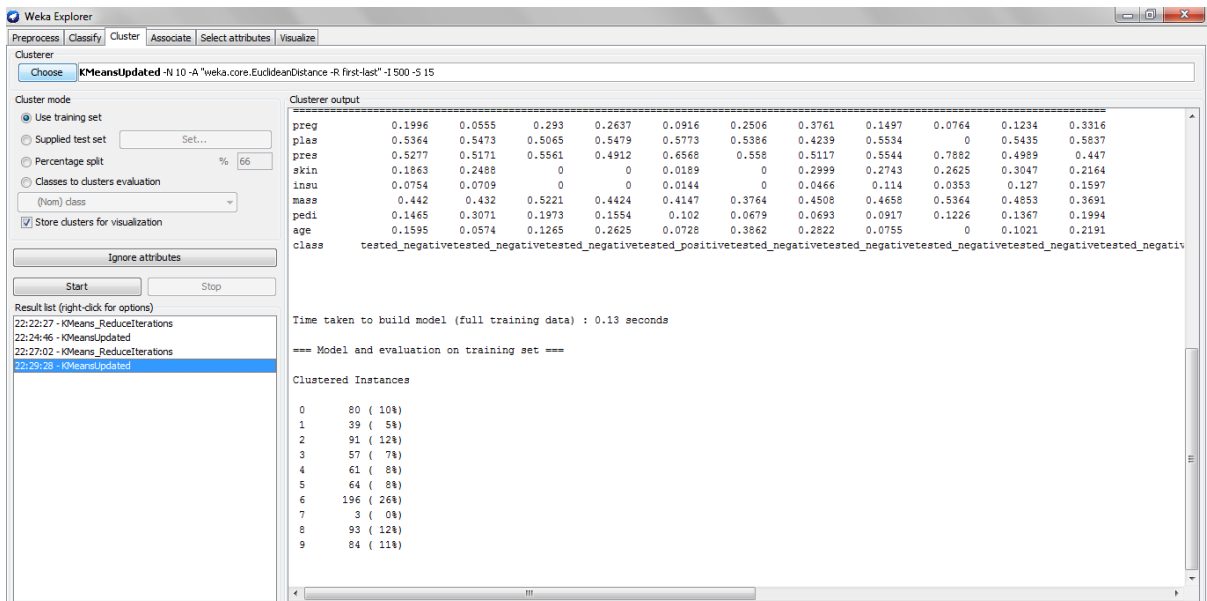


Figure 19: K-means Clustering Algorithm with reduced iteration

Having subsamples for numcluster=10

4.3 Analysis of the algorithm

I. Increasing the clusters to 20

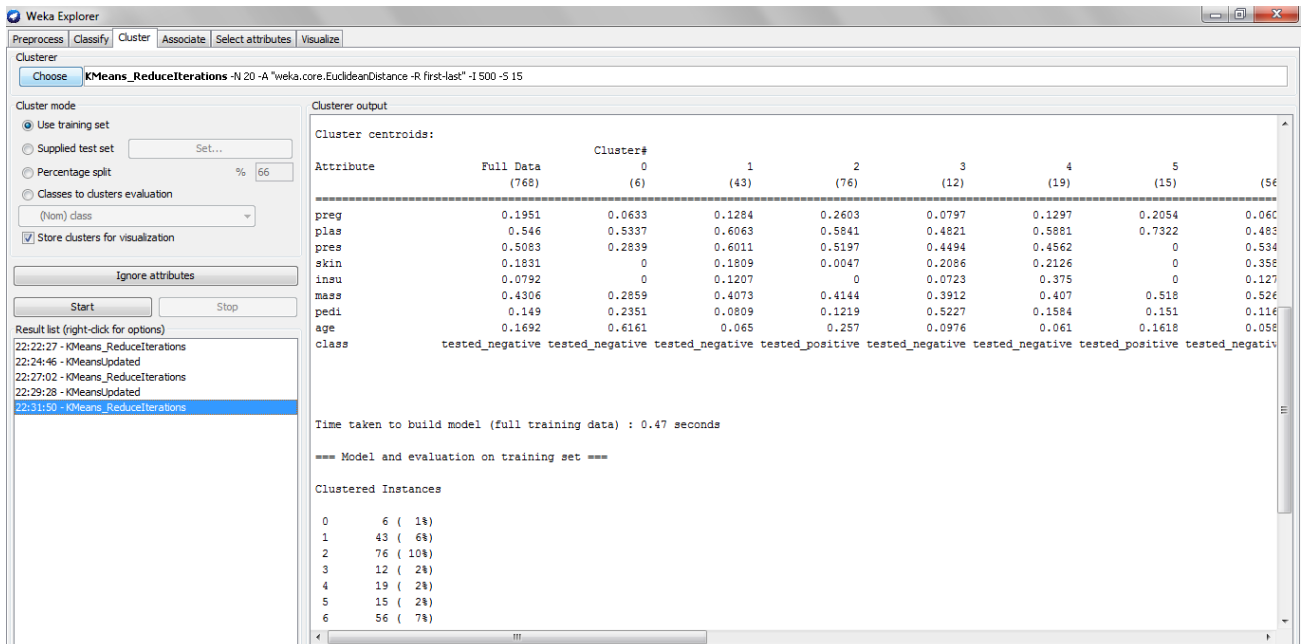


Figure 20: K-means Clustering Algorithm with reduced iteration

For numcluster=20

Again on increasing the no. of clusters to 20, the time taken by K-means with reduced iterations is 0.47 seconds and that of K-means with merged reduced iterations and subsamples is 0.20 seconds.

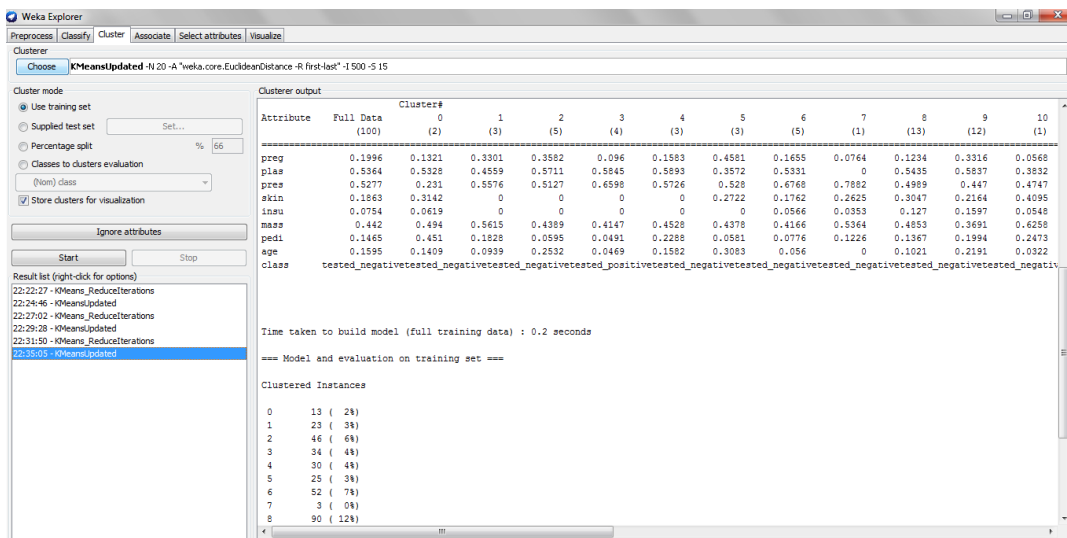


Figure 21: K-means Clustering Algorithm with reduced iteration

Having subsamples for numcluster=20

Thus, increasing the number of clusters also effect the time for computation of results.

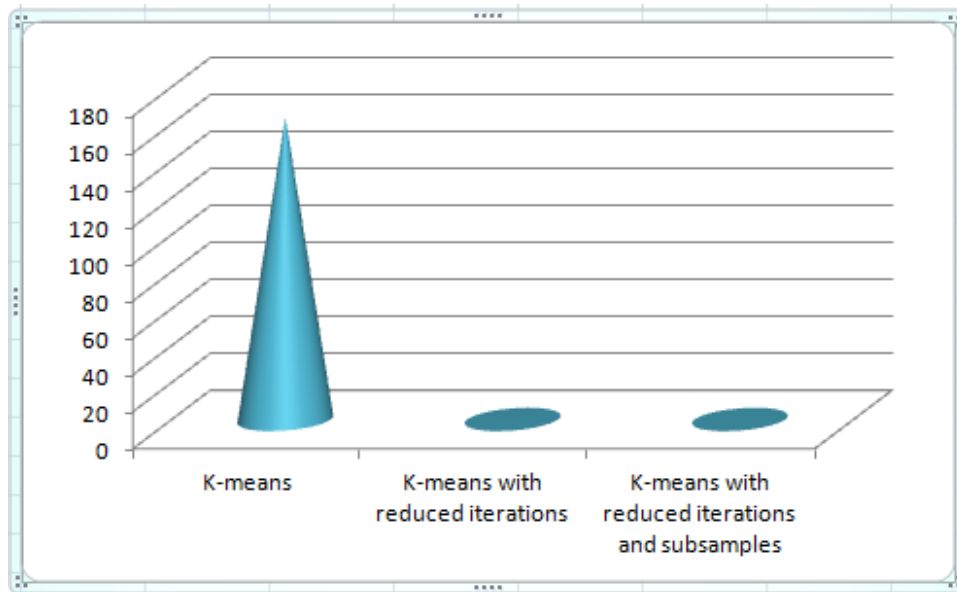
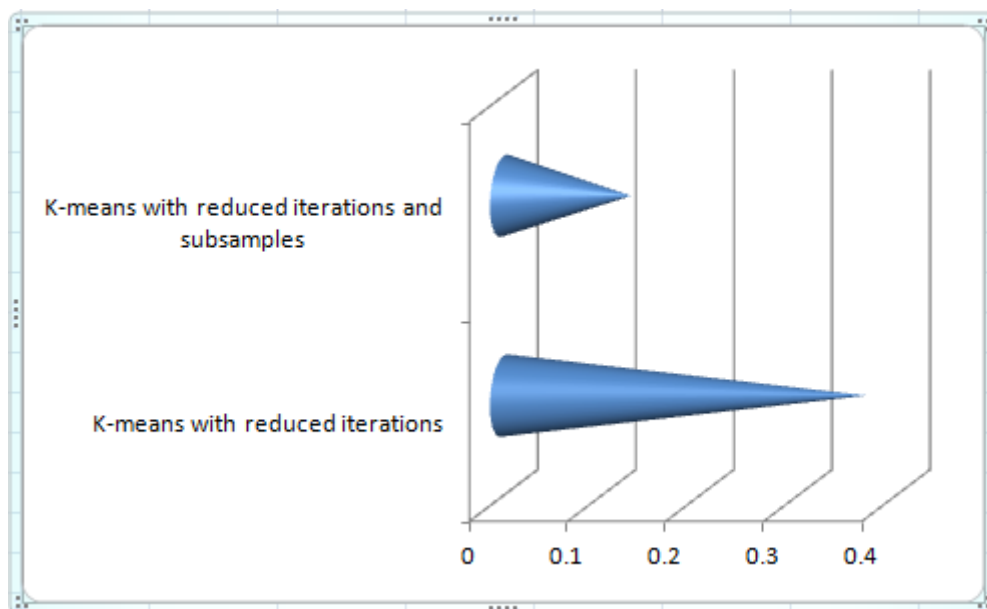
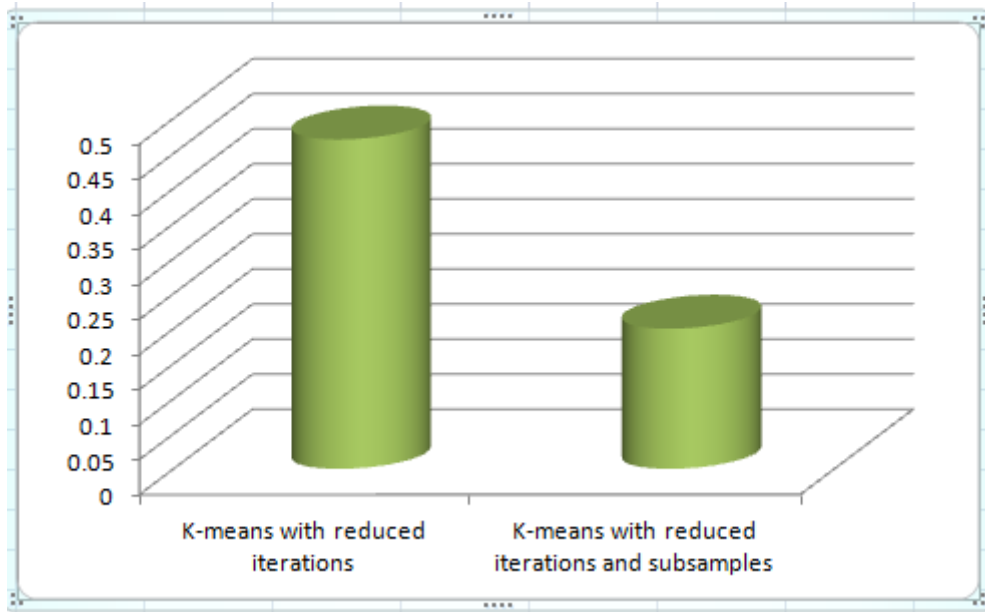


Figure 22: Graph depicting the time of all algorithms



**Figure 23: Graph depicting the time of all algorithms
With 10 clusters**



**Figure 24: Graph depicting the time of all algorithms
With 20 clusters**

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

Conclusion

The conclusion of all the above work is that a new algorithm that is, Hybrid version of K-means clustering algorithm will be designed which is better than the previously used K-means Clustering Algorithm and aids to produce the results than the normal K-means algorithm more accurately and efficiently. The analysis of whole the results is being seen and it is being experimentally shown that the time taken by K-means with reduced iterations is 0.09 seconds and that of K-means with merged reduced iterations and subsamples is 0.05 seconds. The algorithms being designed will define the effectiveness and efficiency of the method used to predict the diabetes mellitus.

Future scope

The future work of all the above work is that the performance of K-means clustering algorithm can be analyzed by overcoming on some another disadvantage like the random selection of the value of k or some other. We have considered only two disadvantages. This hybrid K-means can also be analyzed by using the Classification method like Naïve Bayes Classifier or support vector machine. This Updated K-means can also be used for prediction of any dataset like Super market data or in any other medical diagnosis. The efficiency is also calculated in terms of time however the calculation of accuracy in terms of percentage value can also be taken into consideration for the future work.

CHAPTER 6

REFERENCES

References

Reports and Official Documents

- [1] Bei Andrea, Luca Stefano De , Ruscitti Giancarlo, Salamon Diego (2005), “Health Mining : a Disease Management Support Service based on Data Mining and Rule Extraction”, Proceeding of the 2005 IEEE, Symposium on Computer – Based Medical Systems, 27th Annual Conference September 1-4,2005, Shangai, China.
- [2] Durairaj M. and Vijtha C. (2014), “Educational Data Mining for Prediction of Student Performance Using Clustering Algorithms”, International Journal of Computer Science and Information Technologies, Vol. 5(4), 2014, 5987-5991, Tiruchirappali, India.
- [3] Fayyad, U, “Data Mining and Knowledge Discovery in Databases: Implications from scientific databases”, Proc. of the 9th Int. on the Scientific and Statistical Database Management, Olympia, Washington, USA, 2-11, 1997.
- [4] Giudici P, Wiley John (2003), “Applied Data Mining: Statistical Methods for Business and Industry”, New York.
- [5] Han J., Kamber M. (2006). “Data Mining Concepts and Techniques”, Morgan Kaufmann Publishers.
- [6] Mac Dougall Candice, Percival Jennifer and Mc Gregor Carolyu (2009), “Integrating Health Information Technology into Clinical Guidelines”, Annual International Conference of the IEEE, EMBS Minneapolis, Minnesota, USA, September 2-6, 2009.
- [7] M Nirmala Devi , Balamurugan.S Appavu alias, U.V Swathi (2013), “An amalgam KNN to predict Diabetes Mellitus”, 2013 IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology, Madurai, Tamil Nadu, India.
- [8] Ms. Wankhade Nishigandha V. and Mrs. Potey Madhuri A. (2013), “Transfer Learning Approach for Learning of Unstructured Data from Structured Data in Medical Domain”, 2013 IEEE 978-81-920249-7-4, Pune, India.
- [9] Obenshain, M.K. (2004), “Applications of Data Mining Techniques to Heath care Data”, Infection Control and Hospital Epidemiology, 25(8), 690-695, 2004.

- [10] Palaniappan Sellappan and Awang Rafiah (2008), “Intelligent Heart Disease Prediction System Using Data Mining Techniques “, International Journal of Computer Science and Network Security, VOL. 8 No. 8, August 2008, Selangor, Malaysia.
- [11] Parthiban Latha and Subramanian R. (2008) , “Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm” International Journal of Biological and Life Sciences 3:3 2008, Pondicherry.
- [12] Shouman Mai, Tumer Tim, Stocker Rob (2012), “Using Data Mining Techniques in Heart Disease Diagnosis and Treatment”, International Conference on Electronics, Communications and Computers 2012, IEEE, Northcott Drive, Canberra.
- [13] Uppin ShravanKumar and M A Anusuya (2014), “Expert System design to predict Heart and Diabetes Diseases”, International Journal of Scientific Engineering and Technology Vol. 03,Mysore, India.
- [14] Srinivas K, Kavihta Rani B. and Dr. Govrdhan A. (2010), “Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks”, International Journal on Computer Science and engineering Vol. 02,No. 02, 2010 ,250-255, Jagtial, Karimnagar.
- [15] Sundar V Bata and Tevi T, Saravanan N (2012), “Development of a Data Clustering Algorithm for Predicting Heart”, International Journal of Computer Applications(0975-888) Volume 48- No. 7, June 2012, Coimbatore, India.
- [16] Thuraisingham, B. (2000), “A Primer for understanding and Applying Data mining ”, IT Professional, 28-31, 2000.
- [17] Thangarasu Gunasekar and Assoc. Prof. Dr. Dominic P.D.D. (2014), “Prediction of Hidden Knowledge from Clinical Database using Data mining Techniques”, 2014 IEEE 978-1-4799-0059-6, Tronoh Perak, Malaysia.
- [18] V. Veena Vijayan and RaviKumar Aswathy (2014), “Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus”, International Journal of Computer Applications, Vol. 95-N0. 17, Trivandrum, India.

Website

- [19] <http://www.medicalnewstoday.com/info/diabetes>.

7.1 Questionnaire

i. Why do we need data mining?

The major reason for using the various data mining techniques is as the amount of data is increasing day-by-day there is a need to manage that data. Some technique is required to convert that raw data into useful data patterns which is used for mining of that data. Till now, several data mining techniques have been deployed and used to mine data for various purposes like Educational performance prediction, Heart disease, Fluoride disease and cancer disease prediction. The evaluation of hidden patterns from the results being obtained is being studied. The data mining techniques uses generally classification algorithms like Naïve Bayes, KNN algorithm etc. while clustering algorithms like K-means, K-mediod clustering algorithm. The data which is kept in DWH is in impure form and to make that that in pure form data mining is being done on that data.

ii. What is the problem in the existing method?

The major problem with the existing method is that when clustering algorithm like K-means is being applied on Heart disease gives the accuracy of 66% when compared to other Classification algorithms and takes a lot of time for computation for predicting the particular disease.

iii. How proposed method works?

The proposed method is having the time consuming approach which will be compared with the existing method and which provides the more accuracy. The proposed method will increase the accuracy by dividing the dataset into sub samples and then apply clustering with reduced iterations that will take less time as compared to existing algorithm.

7.2 List of abbreviations

C

CANFIS- “Coactive neuro-fuzzy inference system”

CART – “Classification and Regression Trees”

D

DWH- “Data Warehouse”

DBSCAN- “Density Based Special Clustering of Applications with noise”

E

EBM- “Evidence Based Medicine”

I

ICD- “International Classification of disease”

ID3 – “Iterative Dichotomized 3”

IHDPS- “Intelligent Heart Disease Prediction System”

O

OLAP- “Online analytical Processing”

OLTP- “Online Transactional Processing”

Publications

Paper accepted

- I. Sonal Arora and Kundan Munjal (2015), "Prediction of Diabetes Mell-EH-Tiss using Unsupervised Learning Approach", International Journal of Applied Engineering and Research (IJAER).
- II. Sonal Arora and Kundan Munjal (2015), "A review on Prediction of Diabetes Mell-EH-Tiss using Unsupervised Learning Approach", International Conference of Applied Engineering and Research (ICAAET).