Comparative analysis of K-means and a time efficient Hybrid K-means for Prediction of Diabetes Mell-EH-Tiss

Sakshi, Student,
Computer Science and Technology,
SIET, Aliyaspur Ambala,

Er. Anuradha Saini, Assistant Professor
Computer Science and Technology,
SIET, Aliyaspur Ambala

***Abstract*—There is a lot of information that is present in today's world which is not efficiently managed and utilized so that it can be useful for the various purposes like for the prediction and forecasting of the data. For solving this problem, the concept of big data is introduced so that the huge amount of data can be managed but the question still comes How to use that information? For solving this problem role of data mining comes into play which extracts the hidden knowledge so as to evaluate the hidden patterns and find the information. Clustering is a unsupervised learning technique which is used to find the clusters and group the data objects in similar clusters. We have used the K-means clustering as a data mining approach and enhanced the earlier approach by decreasing the time of the K-means clustering algorithm. The accuracy of both the algorithms are measured in terms of time efficiency which is time taken to form the cluster by including instances. The results of both the algorithms are then analyzed and used for the prediction of diabetes mell-eh-tiss.**

***Index Terms*—Disease Management, Data Warehousing, Data Mining, Medical Services, Diabetes, Diseases, Medical Diagnosis, Hospitals, Evidence Based Medicine**

I. INTRODUCTION

Data Mining is an emerging area of research which is used to extract hidden patterns from large amount of data and that patterns are used to evaluate the results. The results are being analyzed on the basis of behavior of the data mining approach used. Data mining is combined with various disciplines like networking, artificial intelligence, and statistics of data etc. The goal of data mining is to process the input data that can be in the form of numbers, facts and rules or data in the form of datasets into knowledge which can be used for the evaluation purposes. 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].

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.

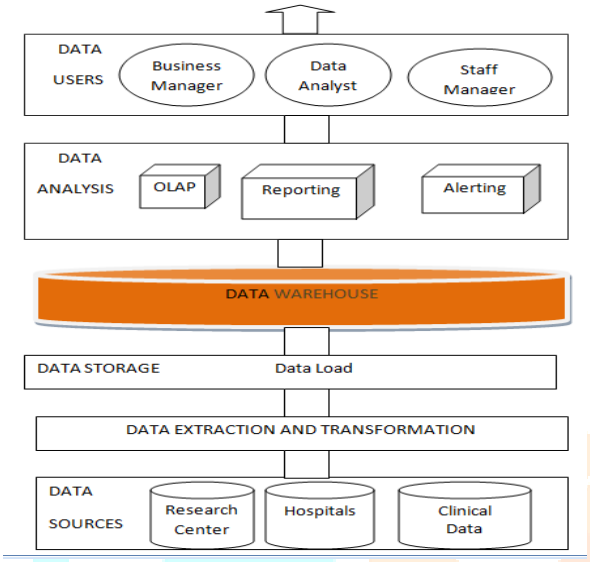


Figure 1 DATA WAREHOUSE ARCHITECTURE

II. 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].

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].

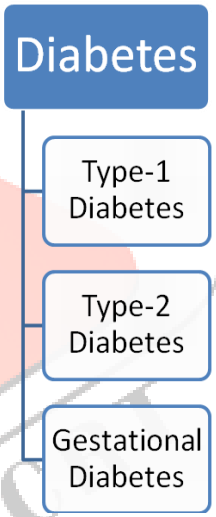


Figure 2 TYPES OF DIABETES MELL_EH_TISS

III. OBJECTIVES 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.

IV. 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. Earlier this dataset was used by:-

Smith ,~J.~W., Everhart ,~J.~E., Dickson ,~W.~C., Knowler ,~W.~C., \& Johannes ,~R.~S. (1988).

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. The total of eight attributes was taken and one attribute is for class that predicts that whether the patient is suffering from diabetes or not. The total of 768 instances was taken.

	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	

diabetes

Table 1: VALUES OF THE DATASET FOR THE ATTRIBUTES

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 2: DIABETES DATASET ATTRIBUTES

The real time dataset for the diabetes patients is also being taken from the Metro Hospital, Faridabad.

V. RESEARH 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*:

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

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects, and q is a positive integer

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

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

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

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

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:

$$centroid = {x1+x2}/{2}$$

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:

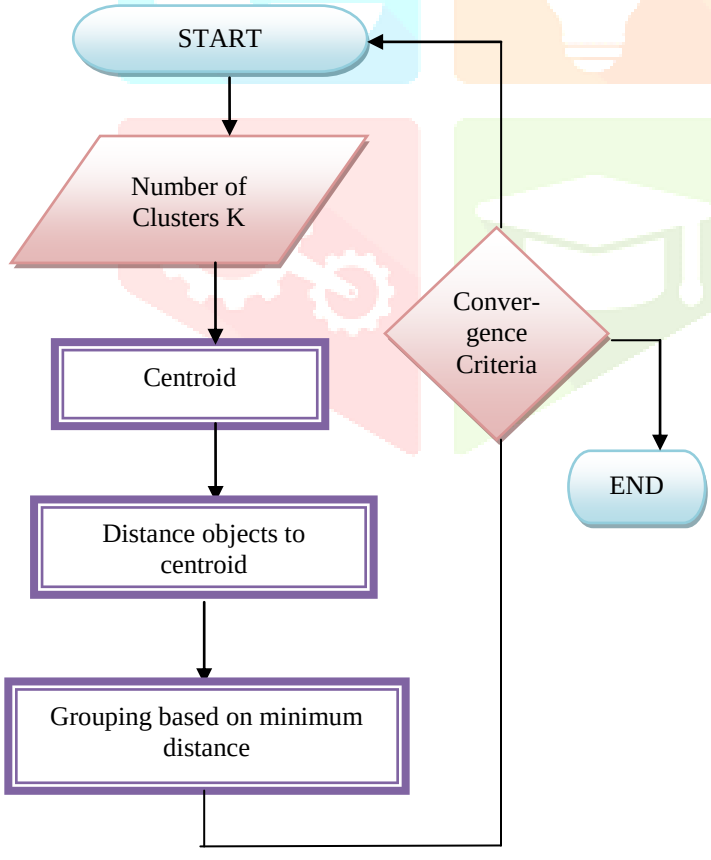


Figure 3 K-MEANS CLUSTERING ALGORITHM

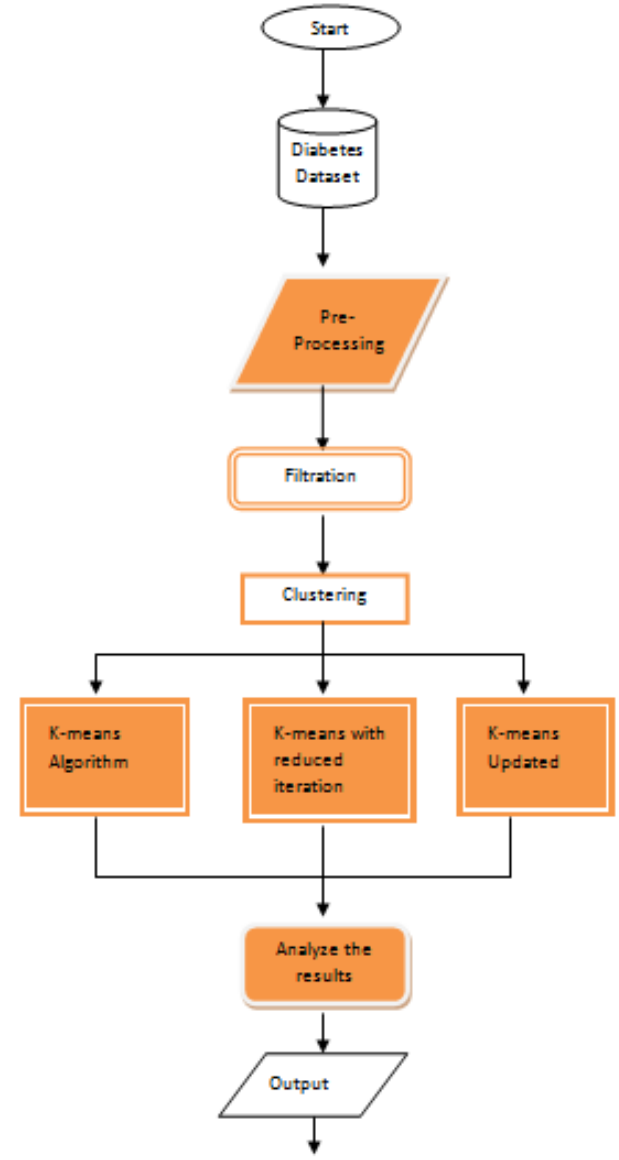


Figure 4 FLOW CHART DESCRIBING METHODOLOGY USED IN THE RESEARCH

K-means Clustering algorithm with reduced iterations approach removes the disadvantage of calculating the distance of each data object to every cluster center which possibly takes a lot of execution time and is removed by storing that each instance is present in which cluster and what is the distance of that instance with the previous cluster center. By this suppose distance earlier calculated was 200 but now when again its centroid are computed the distance comes out to be 250, then the earlier distance will be taken into account which removes the difficulty of K-means clustering algorithm. The reduced iterative approach saves the time taken to build the clusters. If we use this reduced iterative approach on the subsamples rather than taking the complete dataset as a input it again reduces the probability of dividing one big cluster into several subsamples which further removes the second disadvantage of K-means clustering algorithm. We form the Hybrid approach by merging both the above approaches and

computes the results. The flow chart so formed for the Hybrid and time efficient K-means clustering algorithm which is named as updated K-means is as follows:

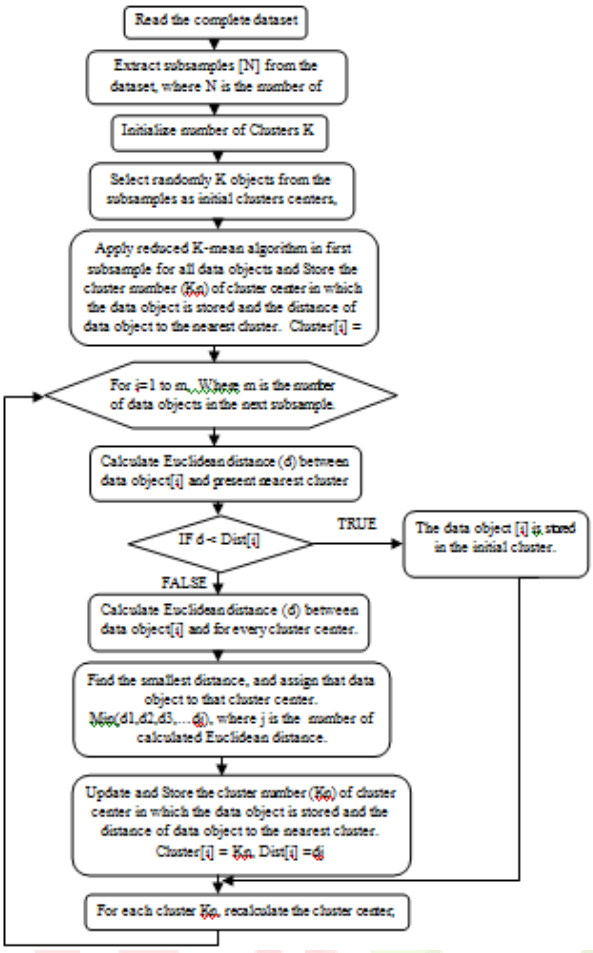


Figure 5 FLOW CHART FOR HYBRID K-MEANS CLUSTERING ALGORITHM

VI. EXPERIMENTAL RESULTS

The experimental results of all the algorithms being developed were calculated and the analysis by changing the certain parameters was done and their effect on the results was seen. The dataset before pre-processing contains certain anomalies, missing values and there is a need to normalize that data. Weka provides certain filters for supervised and unsupervised learning algorithms. We have used 2 filters namely replace missing value and normalize filter to convert dirty data to pure data. Replace missing value filter converts the missing value to the value by taking mean or median of that attribute values and normalize filter normalizes the data by removing the redundant data and removes the noise. The dataset before any pre-processing of data is as depicted below in figure 6:

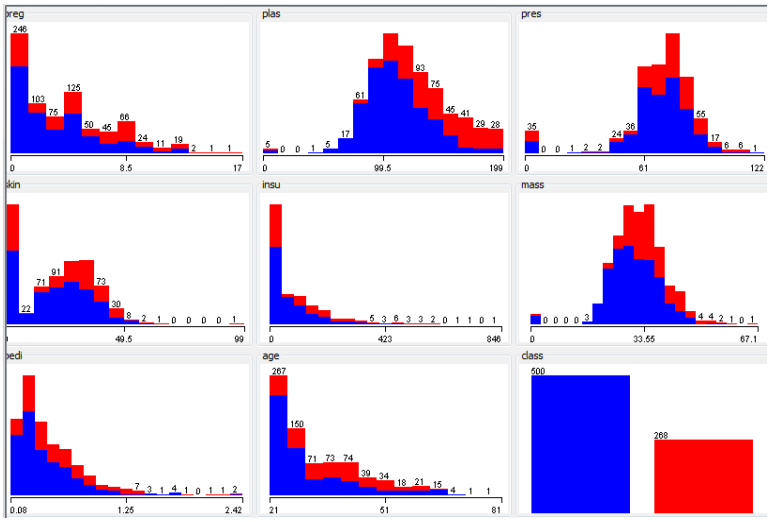


Figure 6 DATASET BEFORE PRE-PROCESSING

The dataset after pre-processing that is after normalization and replacing missing values is as shown in figure 7 and 8:

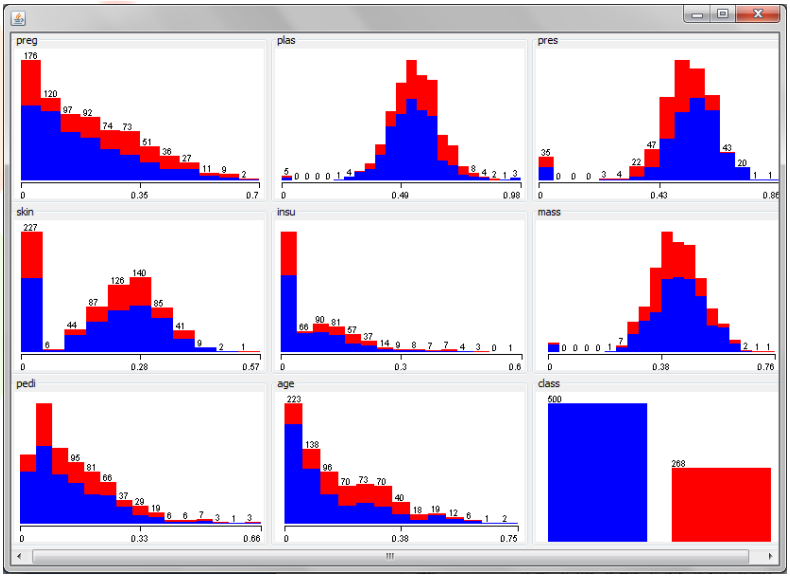


Figure 7 DATASET AFTER NORMALIZATION

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].

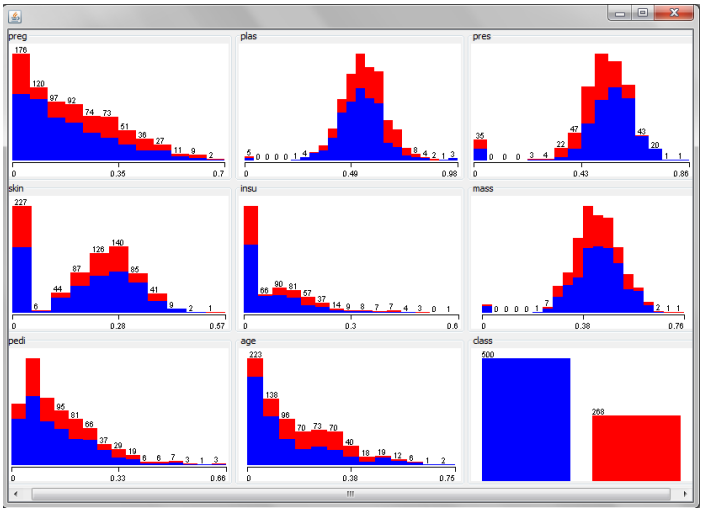
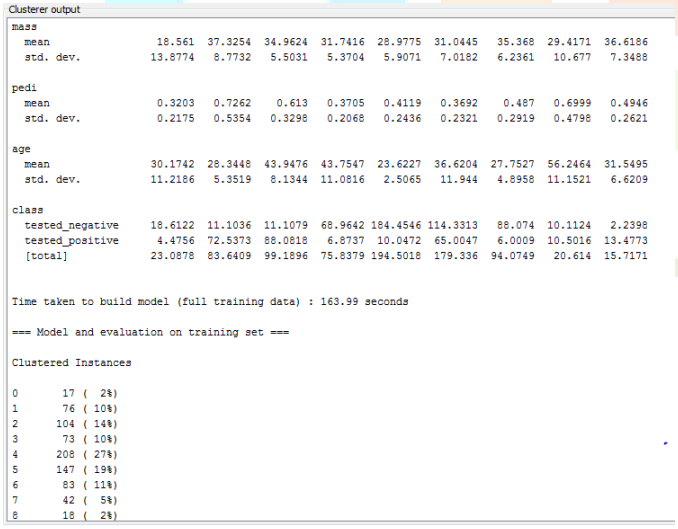


Figure 8 DATASET AFTER REPLACE MISSING VALUE FILTER

Figure 9 shows the results when we are performing the simple K-means Clustering algorithm. The time taken to build the model is 163.99 seconds. The results compared when are performing K-means with the reduced iterations and the K-means with merged reduced iterations and subsamples are as depicted below in figure 10 and 11:



```
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 9 K-MEANS CLUSTERING ALGORITHM

The below 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. The graph comparing the predicted time is as shown in figure 14.

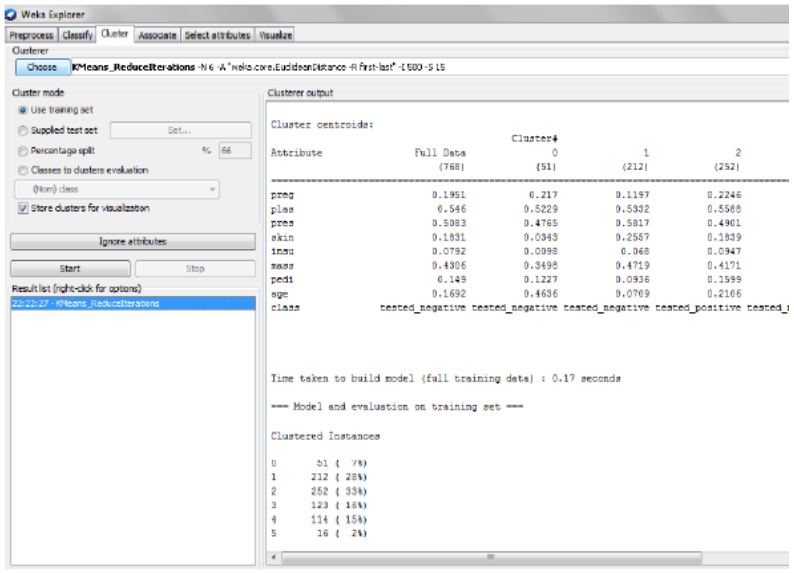


Figure 10 K-MEANS WITH REDUCED ITERATIONS

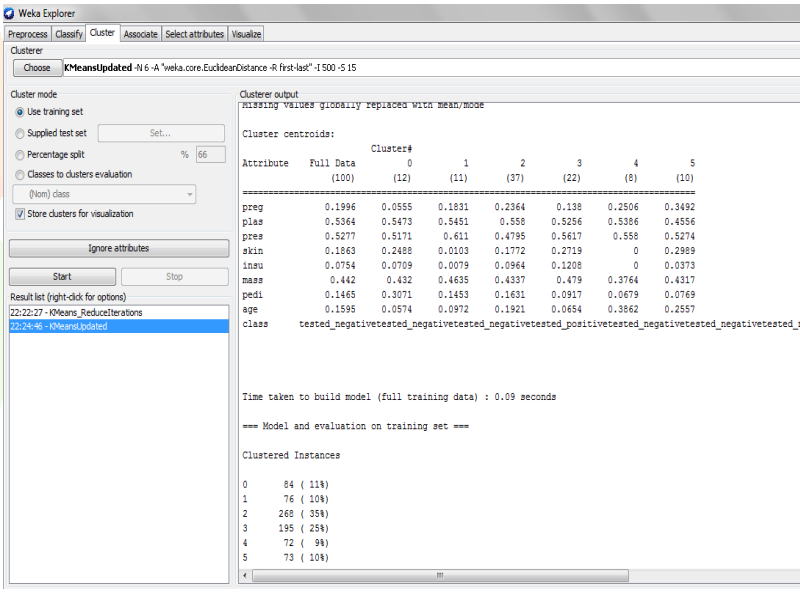


Figure 11 K-MEANS WITH REDUCED ITERATIONS AND SUBSAMPLES

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 12 while for K-means with reduced iterations and subsamples having 10 clusters is 0.13 sec as shown in figure 13:

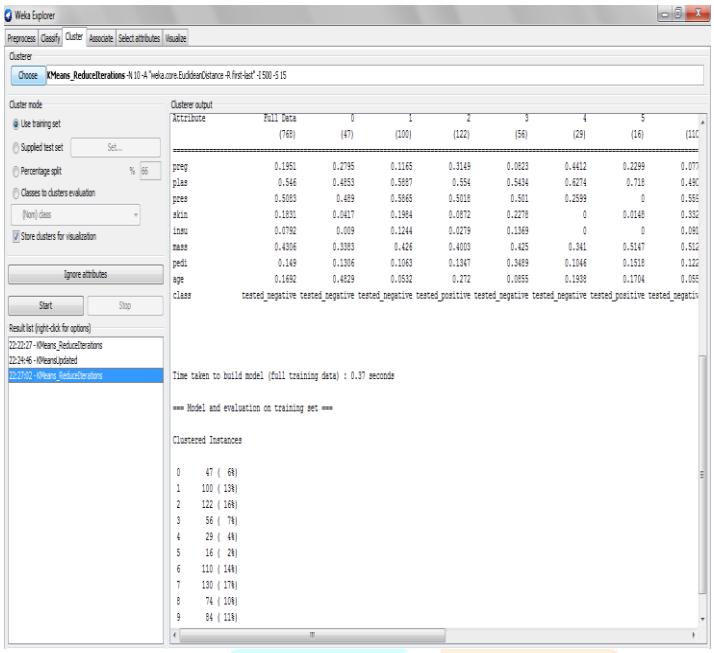


Figure 12 K-MEANS WITH REDUCED ITERATIONS AND 10 CLUSTERS

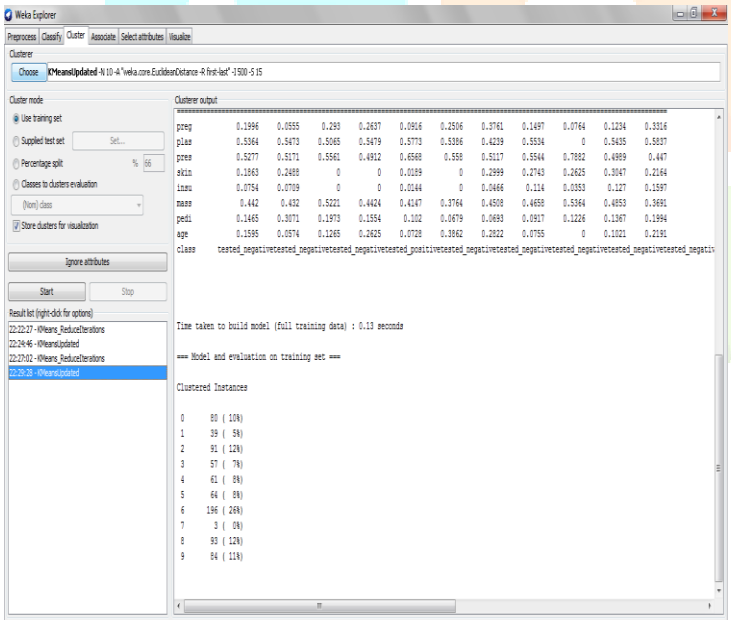


Figure 13 K-MEANS WITH REDUCED ITERATIONS AND SUBSAMPLES WITH 10 CLUSTERS

The graph comparing the predicted time for the number of clusters=10 is depicted in figure 15.

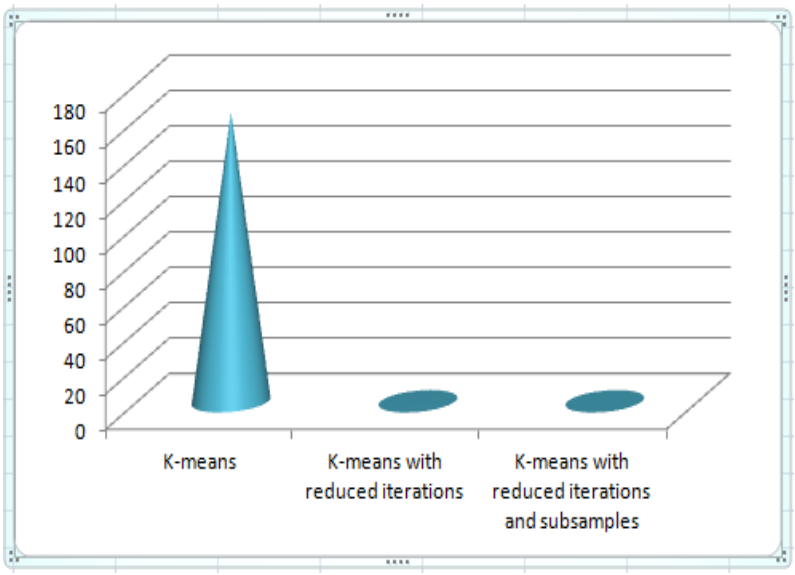


Figure 14 GRAPH DEPICTING THE TIME OF ALL ALGORITHMS

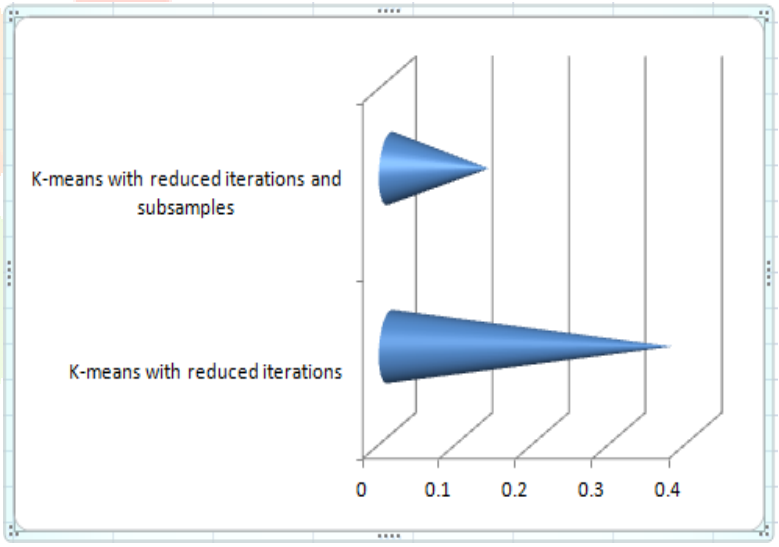


Figure 15 GRAPH DEPICTING THE TIME OF ALL ALGORITHMS WITH 10 CLUSTERS

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 as shown in figure 16 and figure 17.

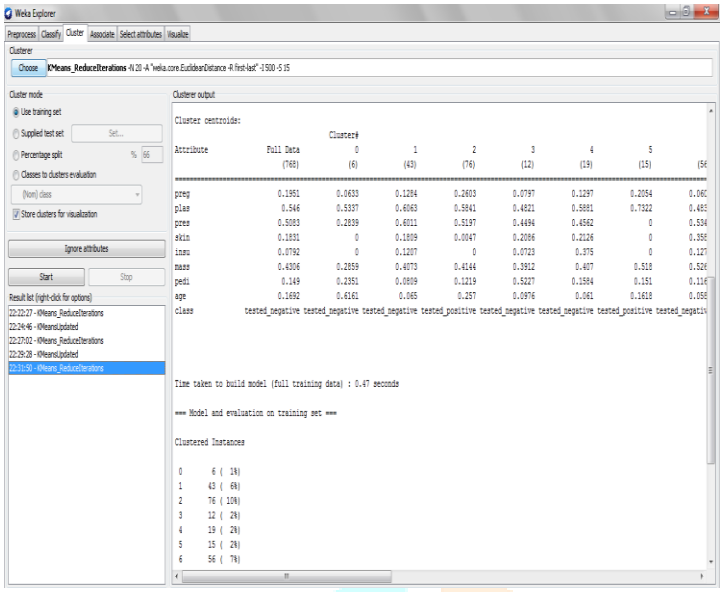


Figure 16 K-MEANS CLUSTERING ALGORITHM WITH REDUCED ITERATION FOR NUMCLUSTER=20

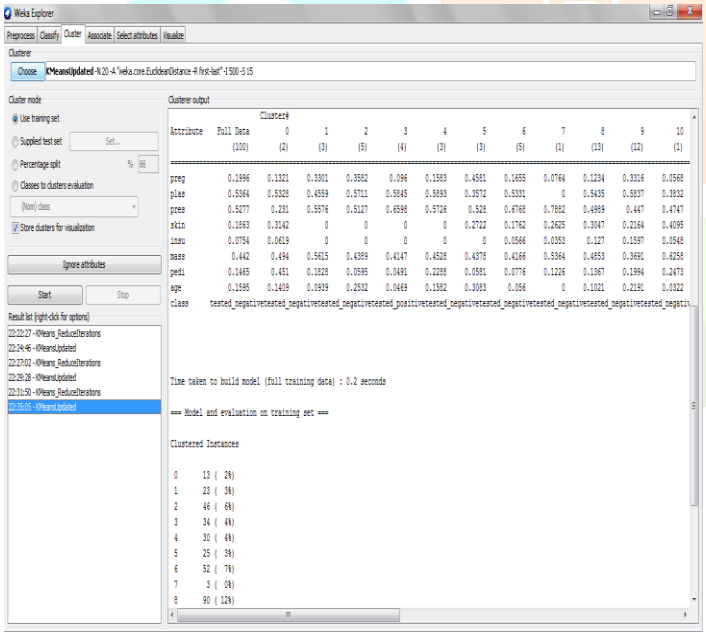


Figure 17 K-MEANS CLUSTERING ALGORITHM WITH REDUCED ITERATION AND HAVING SUBSAMPLES FOR NUMCLUSTER=20

The graph for the above analysis is also depicted in figure 18.

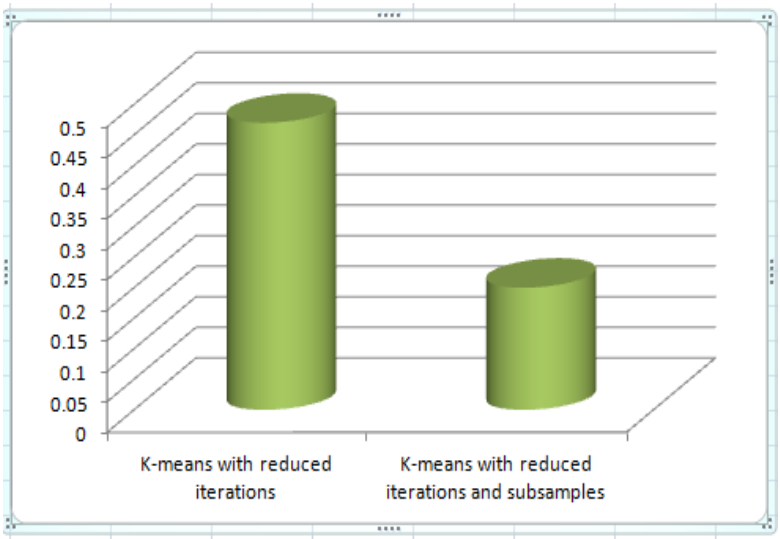


Figure 18 GRAPH DEPICTING THE TIME OF ALL ALGORITHMS WITH 20 CLUSTERS

The above result shows that if we are going to increase the number of clusters then the time taken to build the model is also increasing. Also, if we are increasing the number of subsamples then the Hybrid approach behaves exactly as that of simple K-means Clustering algorithm.

VII. CONCLUSION AND FUTURE SCOPE

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 algorithms being designed will define the effectiveness and efficiency of the method used to predict the diabetes mellitus. 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.

VIII. ACKNOWLEDGEMENT

The author would like to thank Asst. Prof. Kundan Munjal of Lovely Professional University, Phagwara for the important collaboration and ideas which I received on this research. He always provides guidance about the current innovations which can be implemented in this research.

REFERENCES

[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] 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.

[14] 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.

[15] Thuraisingham, B. (2000), "A Primer for understanding and Applying Data mining", IT Professional, 28-31, 2000.

[16] 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.