iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VI **Month of publication:** June 2026

DOI: https://doi.org/10.22214/ijraset.2026.84062

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

A Comparative Study of Data Mining Clustering Methods for Diabetes Prediction in Healthcare Data Set

Kamalpreet Kaur[1], Er. Bharti[2], Dr. Naveen Dhillion[3]
[1]M.tech Scholar at RIET Phagwara
[2]Head of Department (ECE)
[3]Principal Riet. Phagwara

Abtstract: Diabetes is a disease which is affecting many people now-a-days. Diabetes is a chronic disease caused due to the expanded level of sugar addiction in the blood. Various automated information systems were outlined utilizing various classifiers for anticipate and diagnose the diabetes. Due to its continuously increasing rate, more and more families are unfair by diabetes mellitus. Most diabetics know little about their risk factor they face prior to diagnosis. Data mining approach helps to diagnose patient's diseases. It has played an important role in diabetes research. It would be a valuable asset for diabetes researchers because it can unearth hidden knowledge from a huge amount of diabetes-related data. Various data mining techniques help diabetes research and ultimately improve the quality of health care for diabetes patients. The primary target of this examination is to assemble Intelligent Diabetes Disease Prediction System that gives analysis of diabetes malady utilizing diabetes patient's database .Clustering is the process of partitioning the data or objects into the same class and data in one class is more similar to each other than to those in other cluster. The process of partitioning data objects into subclasses is called as cluster. The quality of cluster depends on the method used. In this research we present the comparison of different clustering techniques using Waikato Environment for Knowledge Analysis or in short (WEKA) by using diabetes data set. The algorithm or methods tested are Density Based Clustering, filtered Cluster and K-MEANS clustering algorithms. This research present a comparative analysis for various clustering techniques on diabetes datasets.
Keywords: Diabetes, Clustering, Weka, K-Means.

I. INTRODUCTION

Data mining is the most important step for discovering knowledge from the dataset. Data mining provides us useful patterns or models to discover important and useful data from whole database. We use different algorithms to extract the valuable data [1] Clustering is use to describe the data and categories it into similar objects belongs to one group. Classification is an important task in data mining. Classification is use to classify the data items into the predefined classes and find the models to analysis. Its purpose is to set up a classifier model and map all the samples to a certain class which can provide much convenience for people to analysis data. Furthermore classification belongs to direct learning and the main methods include Decision Tree, Bayesian Classification, Neural Network and Genetic Algorithm. Clustering and Classification are two of the mostly used methods of data mining which provide us more convenience in our research.

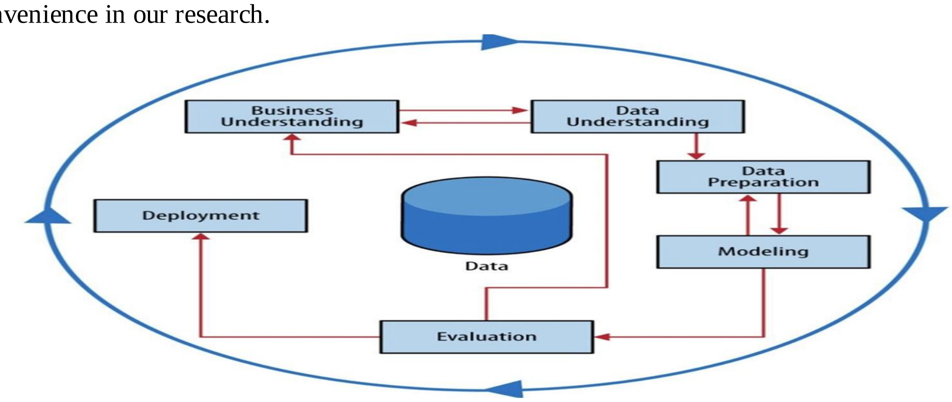


Figure 1.1: Process of Data Mining

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

A. Process of Data Mining

The whole process of data mining cannot be completed in a single step. In other words, you cannot get the required information from the large volumes of data as simple as that. It is a very complex process than we think involving a number of processes. The processes including data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge representation are to be completed in the given order.

Different data mining processes can be classified into two types: data preparation or data preprocessing and data mining. In fact, the first four processes, that are data cleaning, data integration, data selection and data transformation, are considered as data preparation processes. The last three processes including data mining, pattern evaluation and knowledge representation are integrated into one process called data mining.

1) Data Cleaning

Data cleaning is the process where the data gets cleaned. Data in the real world is normally incomplete, noisy and inconsistent. The data available in data sources might be lacking attribute values, data of interest etc. For example, you want the demographic data of customers and what if the available data does not include attributes for the gender or age of the customers? Then the data is of course incomplete. Sometimes the data might contain errors or outliers. An example is an age attribute with value 200. It is obvious that the age value is wrong in this case. The data could also be inconsistent. For example, the name of an employee might be stored differently in different data tables or documents. Here, the data is inconsistent. If the data is not clean, the data mining results would be neither reliable nor accurate.

Data cleaning involves a number of techniques including filling in the missing values manually, combined computer and human inspection, etc. The output of data cleaning process is adequately cleaned data.

2) Data Integration

Data integration is the process where data from different data sources are integrated into one. Data lies in different formats in different locations. Data could be stored in databases, text files, spreadsheets, documents, data cubes, Internet and so on. Data integration is a really complex and tricky task because data from different sources does not match normally. Suppose a table A contains an entity named customer_id where as another table B contains an entity named number. It is really difficult to ensure that whether both these entities refer to the same value or not. Metadata can be used effectively to reduce errors in the data integration process. Another issue faced is data redundancy. The same data might be available in different tables in the same database or even in different data sources. Data integration tries to reduce redundancy to the maximum possible level without affecting the reliability of data.

3) Data Selection

Data mining process requires large volumes of historical data for analysis. So, usually the data repository with integrated data contains much more data than actually required. From the available data, data of interest needs to be selected and stored. Data selection is the process where the data relevant to the analysis is retrieved from the database.

4) Data Transformation

Data transformation is the process of transforming and consolidating the data into different forms that are suitable for mining. Data transformation normally involves normalization, aggregation, generalization etc. For example, a data set available as "-5, 37, 100, 89, 78" can be transformed as "-0.05, 0.37, 1.00, 0.89, 0.78". Here data becomes more suitable for data mining. After data integration, the available data is ready for data mining.

5) Data Mining

Data mining is the core process where a number of complex and intelligent methods are applied to extract patterns from data. Data mining process includes a number of tasks such as association, classification, prediction, clustering, and time series analysis.

6) Pattern Evaluation

The pattern evaluation identifies the truly interesting patterns representing knowledge based on different types of interestingness measures. A pattern is considered to be interesting if it is potentially useful, easily understandable by humans, validates some hypothesis that someone wants to confirm or valid on new data with some degree of certainty.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

B. Techniques used in data mining

1) Association: Association is classified as the best technique among others in the data mining technique. For the transaction of the similar data from one particular image to other, can be done with the help of association in which a pattern is discovered on the basis of relationship [3]. This association technique has been utilized to predict the presence of diabetes in the body and also provide the analysis about the relationship of different attributes. For the prediction of the disease, all the risk factor in the patients is sorted out.
2) Clustering: In the data mining technique, clustering is a technique in which clustering of objects are identified. An automatic technique has been utilized for this purpose as it has the similar characteristics. This clustering technique defined the classes and objects in order to define the process that how objects are assigned into a predefined classes. The prediction of heart disease becomes feasible with the help of clustering technique in which list of patients which have same risk factor are clustered. A separate list of patients is made using this technique.
3) Classification: In the data mining technique, classification is a method that is based on machine learning. In the classification of the data, each item in the data set is classified into predefined set of groups. There are several mathematical techniques that have been utilized by the classification method.
4) Prediction: It is possible to recognize a relationship among the variables which are dependent and independent using the data mining's another approach named as prediction. In the various fields this techniques can be utilized in order to predict profit for future. Therefore, dependent variable is referred as profit and independent variable is referred as sale. Historical sale and profit data has been utilized for the prediction of profit using a fitted regression curve.
5) Outlier detection: In many cases, simply recognizing the overarching pattern can't give you a clear understanding of your data set. You also need to be able to identify anomalies, or outliers in your data. For example, if your purchasers are almost exclusively male, but during one strange week in July, there's a huge spike in female purchasers, you'll want to investigate the spike and see what drove it, so you can either replicate it or better understand your audience in the process [4].
6) Regression: Regression, used primarily as a form of planning and modeling, is used to identify the likelihood of a certain variable, given the presence of other variables. For example, you could use it to project a certain price, based on other factors like availability, consumer demand, and competition. More specifically, regression's main focus is to help you uncover the exact relationship between two (or more) variables in a given data set.
7) Tracking patterns: One of the most basic techniques in data mining is learning to recognize patterns in your data sets. This is usually recognition of some aberration in your data happening at regular intervals, or an ebb and flow of a certain variable over time. For example, you might see that your sales of a certain product seem to spike just before the holidays, or notice that warmer weather drives more people to your website.
8) Evolution and deviation analysis: Evolution and deviation analysis pertain to the study of time related data that changes in time. Evolution analysis models evolutionary trends in data, which consent to characterizing, comparing, classifying or clustering of time related data. Deviation analysis, on the other hand, considers differences between measured values and expected values, and attempts to find the cause of the deviations from the anticipated values.

II. RESEARCH METHODOLOGY

The diabetes prediction problem is complex in nature due to available of large number of attributes in the dataset. The voting is the classification method which is the combination of several classification methods. Several classifiers are collected to generate one ensemble classifier called the voting based classifier. For exploiting the various peculiarities of each algorithm, they are trained and evaluated in parallel individually.

The methodology describes all the steps according to which comparative analysis of clustering algorithms is performed.

1) Step1. Choose the clustering algorithms: To perform the comparative analysis, these clustering algorithms are chosen namely K-means, filtered cluster and Make Density.
2) Step2. Choose the dataset: The various data set has been chosen from specific location where it is stored. The file format is CSV.
3) Step3. Load data on WEKA: Load data file for further analysis.
4) Step4. Normalize data: After loading of the dataset the next step is to normalize the dataset using the WEKA tool through filter tab. Select normalize filter and apply on the same data set. Save the result using save button.
5) Step5. Apply clustering algorithms: Apply the all clustering algorithms on un-normalize as well as normalize dataset.
6) Step6. Store the result: After running all algorithms, results are stored into the tabular forms and based on number of iteration,

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

sum of squared error, time taken to build clusters, correctly clustered data, and comparative analysis is performed.

7) Step7. Plot the graph: Represent results in graphical format.

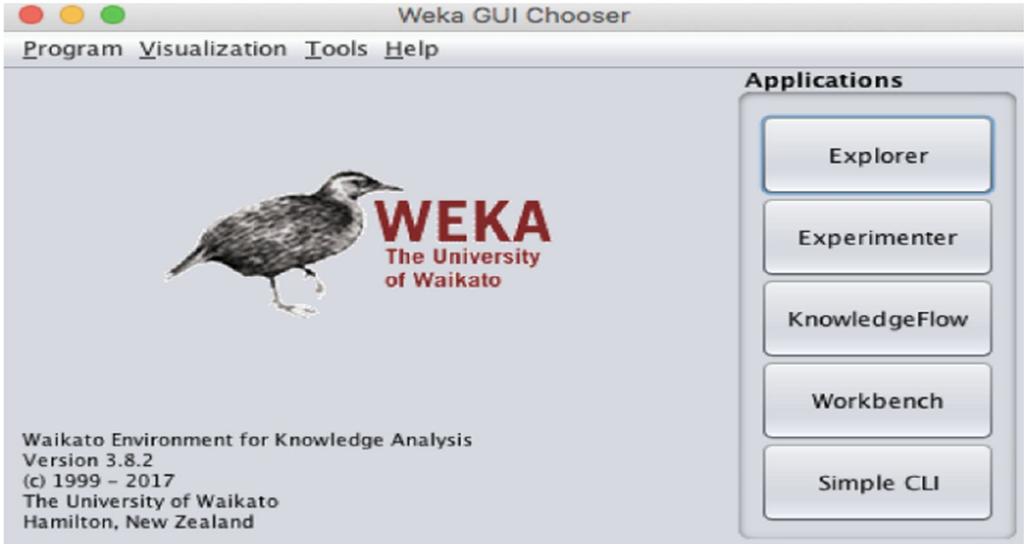


Figure 2.1: Weka Tool

III.RESULT

We are using the WEKA tool for implementation thesis. To import the dataset click on open file and select the data set, which is must be a CSV file.

The parameters for the performance analysis are described below:-

1) Accuracy: The ratio of total number of points that are classified correctly to the total number of points multiplied by 100 is called accuracy.

$$\text{Accuracy} = \frac{\text{Number of points correctly classified}}{\text{Total Number of points}} * 100$$

2) *Execution Time:* Execution time is defined as difference of end time when algorithm stops performing and start time when algorithm starts performing

Execution time = End time of algorithm - start of the algorithm

In the first phase, the data of diabetes is collected from the authorized websites. The collected data is in structured form and technique of pre-processing is applied to transform the data. The method of random sampler is applied for the data transformation.

Sr.	Characteristics	Value
1	Data set Characteristics	Multivariate time series
2	No of instance	768
3	Area	Diabetes
4	Attribute characteristics	Categorical integer
5	No of attribute	09

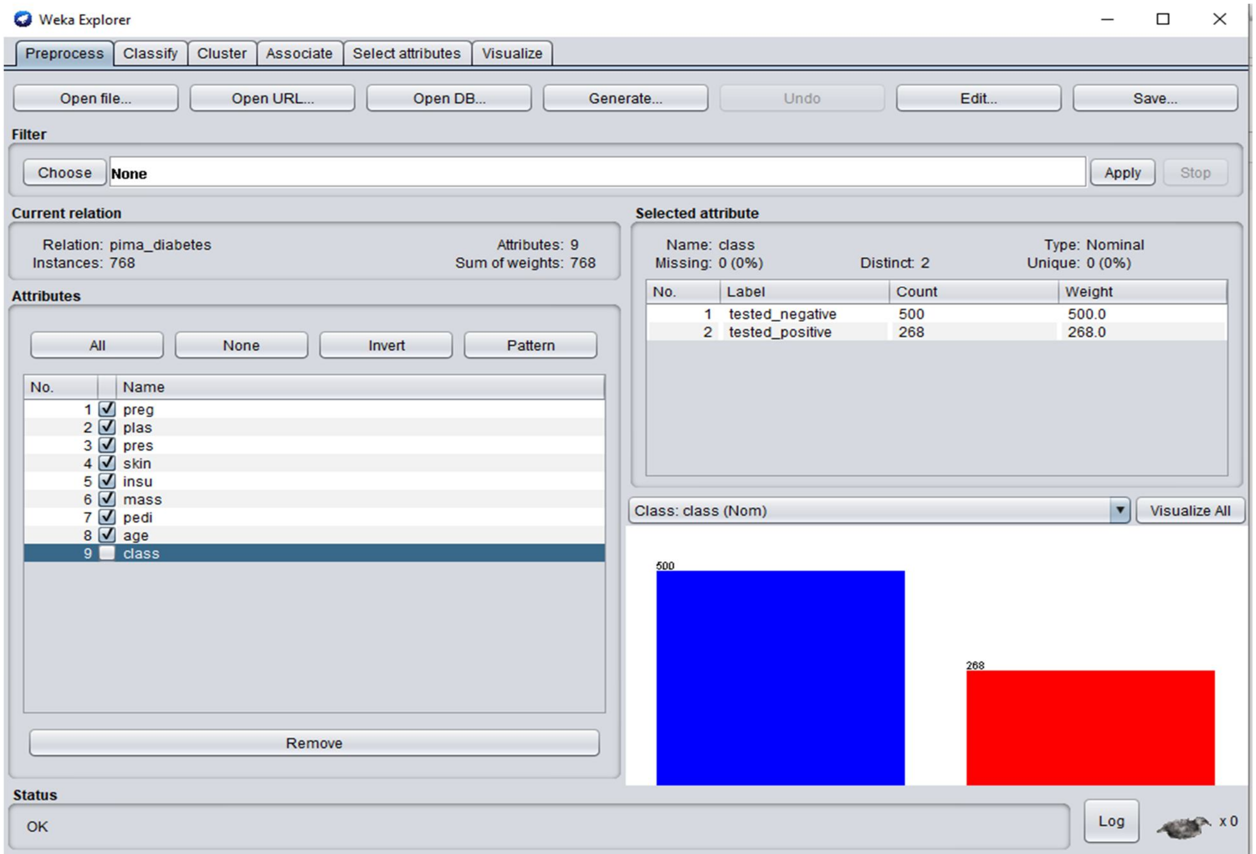


Figure 3.1: Define DIABETES Dataset.

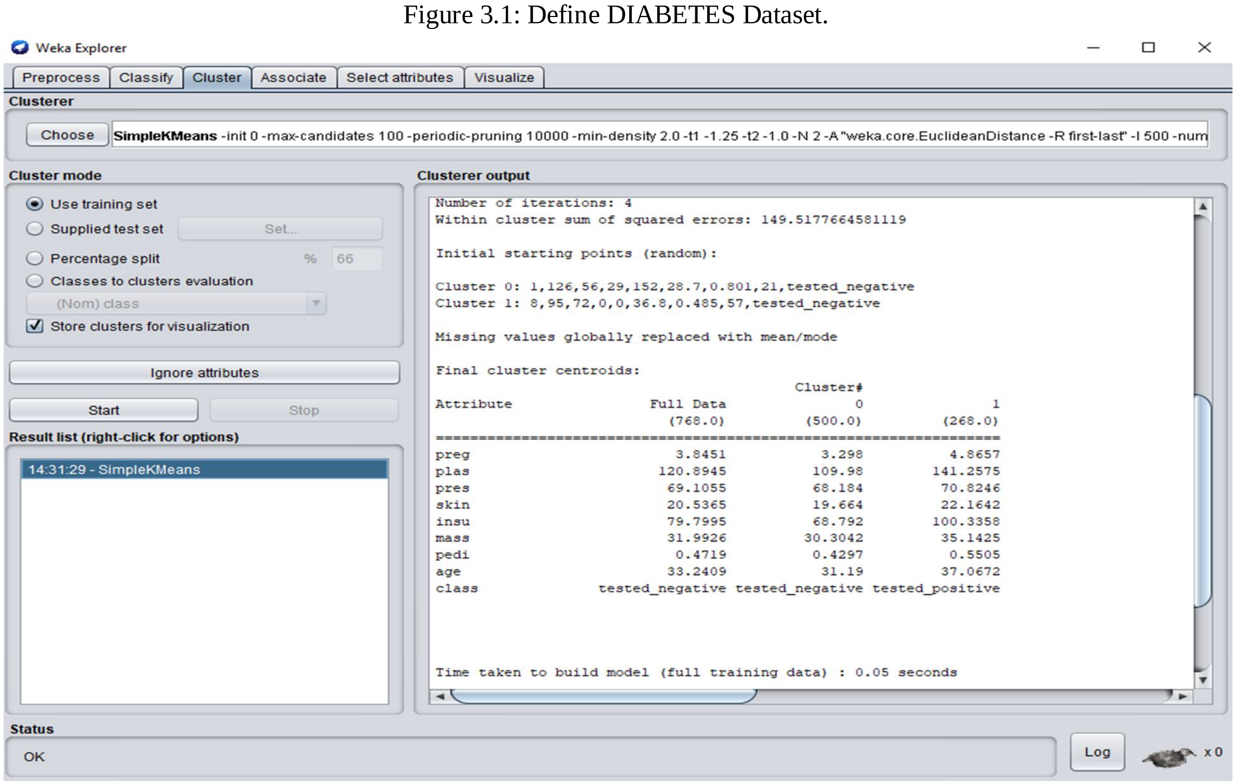


Figure 3.2: In this figure, Firstly click on cluster option from the menu bar then apply the clustering algorithm from choose button and apply the simple K-means algorithms on Diabetes data set. In this figure we can see the two clusters that is cluster 0 and cluster 1 on the attributes.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

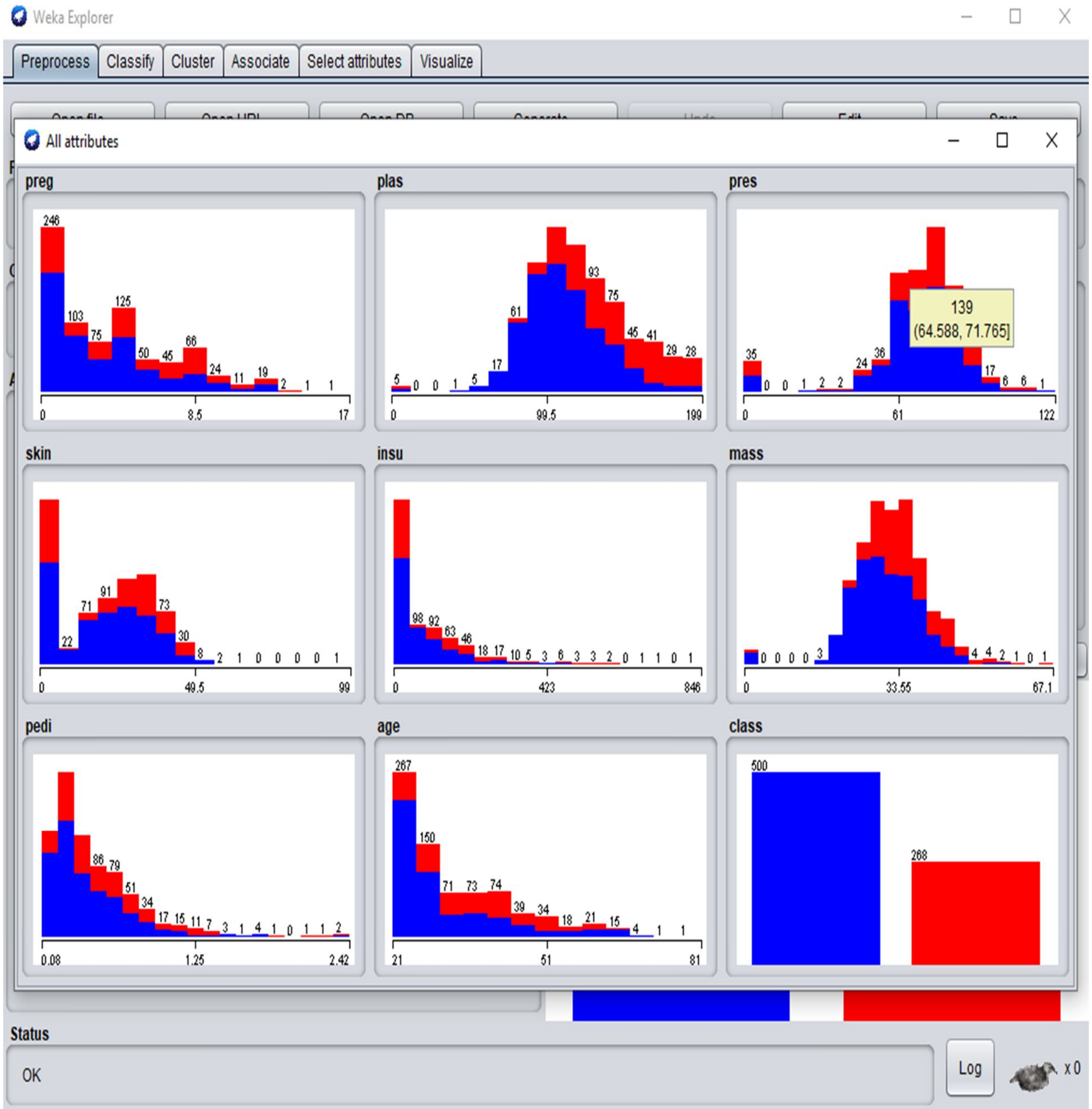


Figure 3.3: Attributes of Diabetes data set in WEKA.

A. Result Analysis

We compare the clustering algorithm of K-Means, DBSCAN and Filtered cluster terms of sum of squared error rate and cluster quality. We can analysis the error rate different clustering algorithms on diabetes dataset.

Table 4.2: Error rate of clustering algorithm

ALGORITHMS	DIABETES
K-Means	14951
DBSCAN	10197
Filtered Cluster	26729

We recorded sum of squared error rate for K-Means is higher than DBSCAN and filtered clustering algorithm in dataset of diabetes.

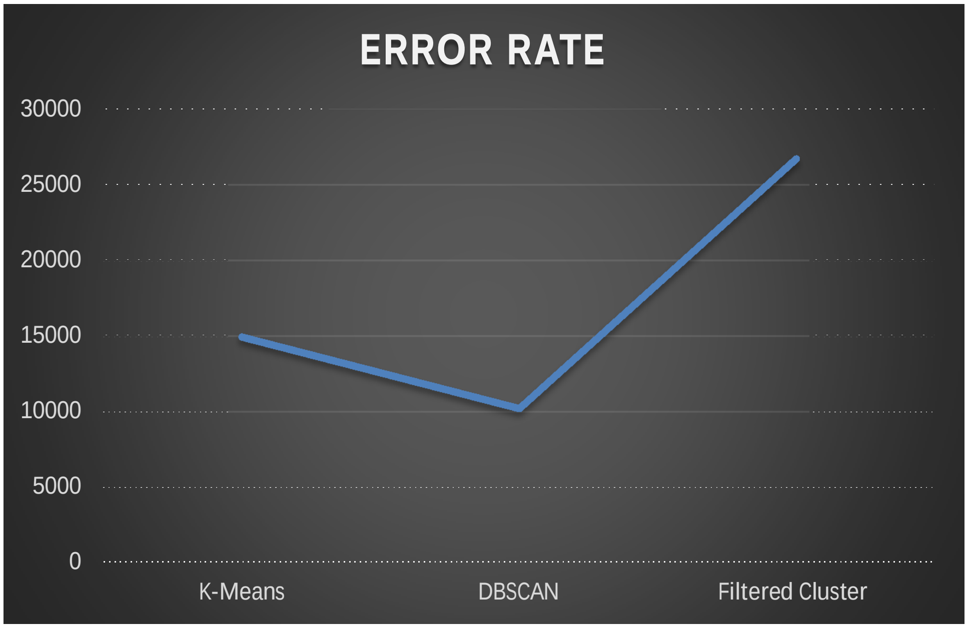


Figure 3.4: shows comparison of error rate between K-means, DBSCAN and filtered clustering algorithms on Diabetes datasets.

We compare the clustering algorithm of K-Means, DBSCAN and Filtered cluster terms of execution time and cluster quality. We can analysis the execution time different clustering algorithms on diabetes dataset.

Table 4.3: Execution time of clustering algorithm in Seconds.

ALGORITHMS	DIABETES
K-Means	0.05
DBSCAN	0.01
Filtered Cluster	0.02

We recorded execution time for K-Means is higher than DBSCAN and filtered clustering algorithm in diabetes dataset.

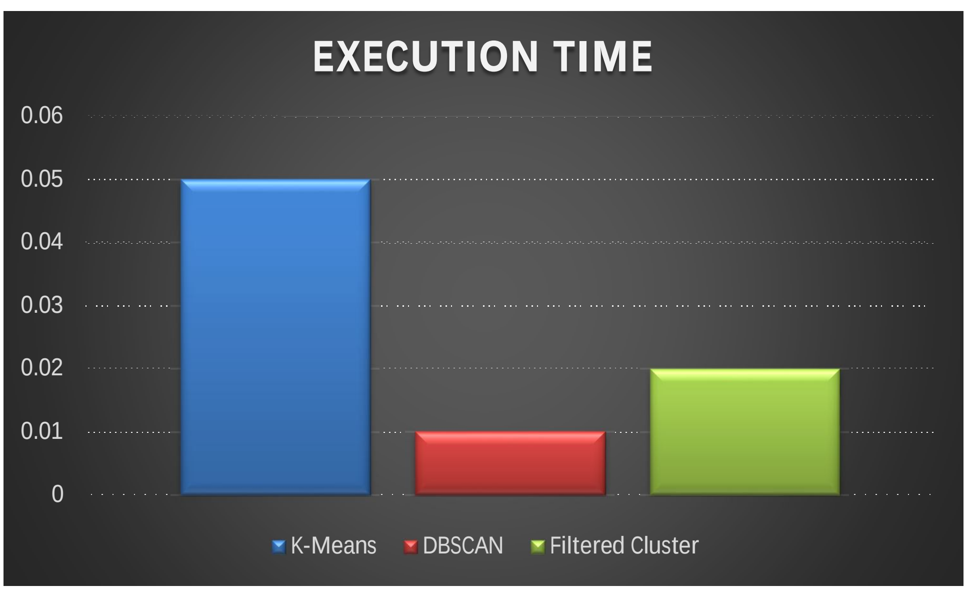


Figure 3.5: shows comparison of execution time between K-means, DBSCAN and filtered clustering algorithms on Diabetes datasets.

IV. CONCLUSION

A chronic disease that causes major causalities in the worldwide that is Diabetes. As per International Diabetes Federation (IDF) around the world estimated 285 million people are suffering from diabetes. This range and data will increase in nearby future as there is no appropriate method till date that minimize the effects and prevent it completely. Type 2 diabetes (TTD) is the most common type of diabetes. The major issue was the detection of TTD as it was not easy to predict all the effects. Therefore, data mining was used as it provides the optimal results and help in knowledge discovery from data. In this work, it is concluded that diabetes prediction is the complex problem due to high complexity of the dataset. In this research shows comparative study has been performed on the K- means, filtered cluster and Density based clustering algorithms. Comparison is performed on Bank and segmentation dataset using WEKA tool and the comparative results are presented in the form of table and graph. The comparative study is performed on the basis of accuracy and efficiency parameters. Every algorithm has their own significance and we use them on the nature of the data, but on the basis of this research we concluded that k-means clustering algorithm is simplest algorithm as compared to other algorithms. Filtered cluster is more accurate than other clustering algorithms in case of error rate but in case of execution time the DBSCAN is better than other clustering algorithms.

Clustering is a vivid method. The solution is not exclusive and it firmly depends upon the analysts' choices. Clustering always provides groups or clusters, even if there is no predefined structure. While applying cluster analysis we are contemplating that the groups exist. But this speculation may be false. The outcome of clustering should never be generalized.

A. Future Work

The motive of this research was to compare some of the clustering algorithms in terms of the execution time, number of iterations and sum of squared error. As a future work, we would attempt to compare all of the algorithms above in terms of different factors other than those mentioned in this research.

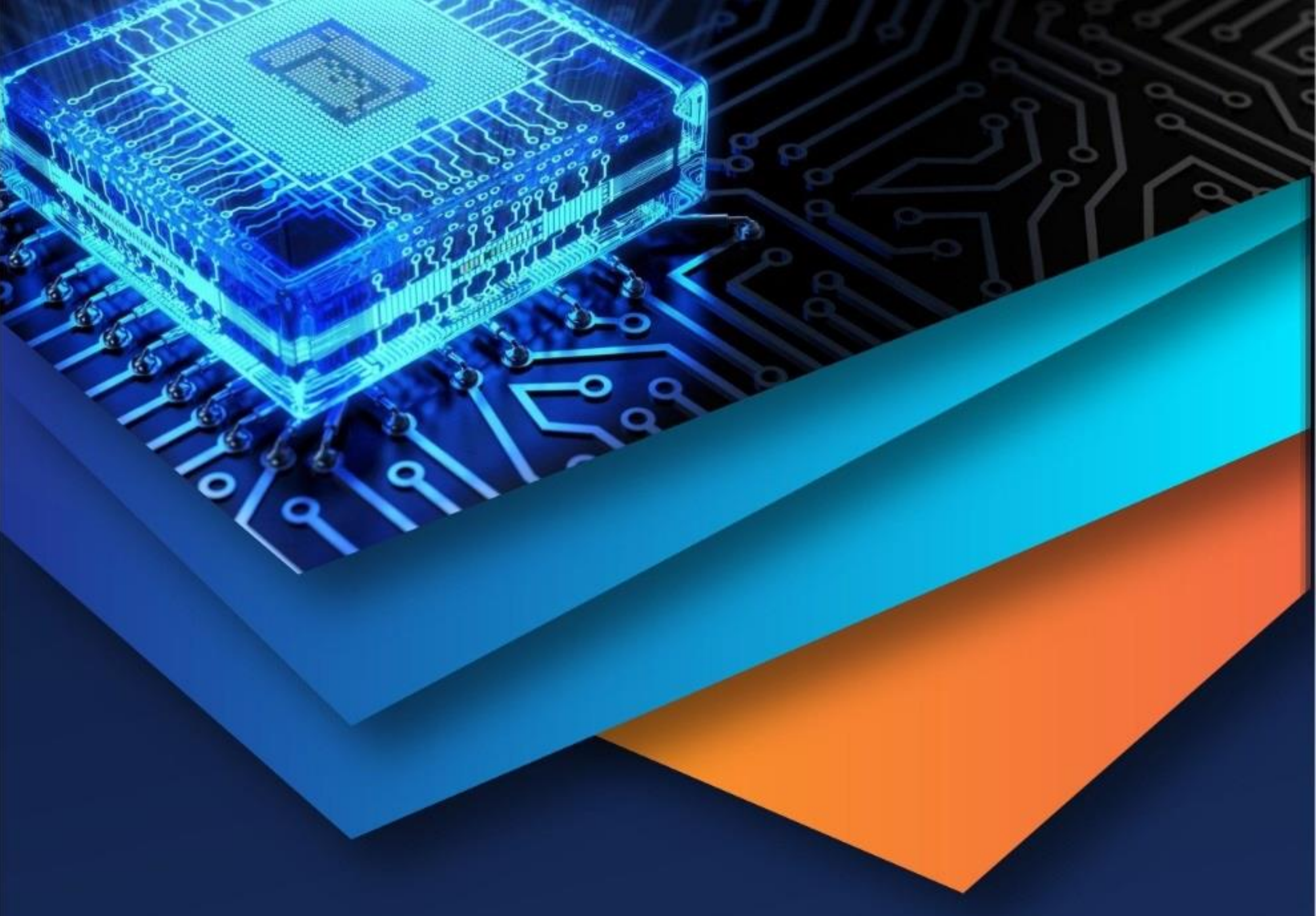
Following are the various futures prospective of this work:-

1) The proposed algorithm can be further improved using the method of clustering. The clustering method can differ the data which can be easily classified.
2) The proposed algorithm can be compared with other algorithms to validate the results.

REFERENCES

[1] Ioannis Kavakiotis, Olga Tsave, AthanasiosSalifoglou, NicosMaglaveras, IoannisVlahavas, IoannaChouvarda, "Machine Learning and Data Mining Methods in Diabetes Research", Computational and Structural Biotechnology Journal 15 (2017) 104–116
[2] BayuAdhi Tama,1 Afriyan Firdaus,2 Rodiyatul FS, "Detection of Type 2 Diabetes Mellitus with Data Mining Approach Using Support Vector Machine", Vol. 11, issue 3, pp. 12-23, 2008.
[3] Yu-Xuan Wang, QiHui Sun, Ting-Ying Chien, Po-Chun Huang, "Using Data Mining and Machine Learning Techniques for System Design Space Exploration and Automatized Optimization", Proceedings of the 2017 IEEE International Conference on Applied System Innovation, vol. 15, pp. 1079-1082, 2017.
[4] ZhiqiangGe, Zhihuan Song, Steven X. Ding, Biao Huang, "Data Mining and Analytics in the Process Industry: The Role of Machine Learning", 2017 IEEE. Translations and content mining are permitted for academic research only, vol. 5, pp. 20590-20616, 2017.
[5] JahinMajumdar, Anwesha Mal, Shruti Gupta, "Heuristic Model to Improve Feature Selection Based on Machine Learning in Data Mining", 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), vol. 3, pp. 73-77, 2016.
[6] MS.Tejashri n. Giri, prof. S.r.todamal, "data mining approach for diagnosing type 2 diabetes", international journal of science, engineering and technology, vol. 2 issue 8, 2014.
[7] P. Suresh Kumar and V. Umatejaswi, " Diagnosing Diabetes using Data Mining Techniques", International Journal of Scientific and Research Publications, Volume 7, Issue 6, June 2017.
[8] M. Sharma, G. Singh, R. Singh, "Stark Assessment of Lifestyle Based Human Disorders Using Data Mining Based Learning Techniques", Elsevier, vol. 5, pp. 202-222, 2017.
[9] Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, Xiaoyi Wang, "Type 2 diabetes mellitus prediction model based on data mining", ScienceDirect, Vol. 11, issue 3, pp. 12-23, 2018.
[10] Yan Luo, Charles Ling, Ph.D., Jody Schuurman, Robert Petrella, MD, "GlucoGuide: An Intelligent Type-2 Diabetes Solution Using Data Mining and Mobile Computing", 2014 IEEE International Conference on Data Mining, Vol. 9, issue 8, pp. 12-23, 2014.
[11] AiswaryaIyer, S. Jeyalatha and RonakSumbaly, "DIAGNOSIS OF DIABETES USING CLASSIFICATION MINING TECHNIQUES", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, 2015.
[12] Alexis Marcano-Cede˜no, Diego Andina, "Data mining for the diagnosis of type 2 diabetes", IEEE, Vol. 11, issue 3, pp. 9-19, 2016.
[13] B. M. Patil, R. C. Joshi, DurgaToshniwal, "Association rule for classification of type -2 diabetic patients", 2010 Second International Conference on Machine Learning and Computing, Vol. 8, issue 3, pp. 7-23, 2010.
[14] Prova Biswas1,2, Ashoke Sutradhar3, PallabDatta, "Estimation of parameters for plasma glucose regulation in type-2 diabetics in presence of meal", IET Syst. Biol., 2018, Vol. 12 Iss. 1, pp. 18-25, 2018.
[15] Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analysing Feature Importances for Diabetes Prediction using Machine Learning", 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)

[16] AyushAnand, Divya Shakti, "Prediction of diabetes based on personal lifestyle indicators", 2015 1st International Conference on Next Generation Computing Technologies (NGCT), Conference Paper
[17] SunghwanBae, Taesung Park, "Risk prediction using common and rare genetic variants: Application to Type 2 diabetes", 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)
[18] Sneha Joshi, MeghaBorse, "Detection and Prediction of Diabetes Mellitus Using Back-Propagation Neural Network", 2016 International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE)
[19] MessanKomi, Jun Li, YongxinZhai, Xianguo Zhang, "Application of data mining methods in diabetes prediction", 2017 2nd International Conference on Image, Vision and Computing (ICIVC)
[20] SalimChemlal, Sheri Colberg, Marta Satin-Smith, Eric Gyuricsko, Tom Hubbard, Mark W. Scerbo, Frederic D. McKenzie, "Blood glucose individualized prediction for type 2 diabetes using iPhone application", 2011 IEEE 37th Annual Northeast Bioengineering Conference (NEBEC)



10.22214/IJRASET

INDEX COPERNICUS
45.98

IMPACT FACTOR:
7.129

TOGETHER WE REACH THE GOAL
IMPACT FACTOR:
7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY