IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-6, June 2018 E-ISSN: 2347-2693

Proficiency Analysis of Various Data Mining Techniques for Diabetes Prognosis

Misba Reyaz[1*], Gagan Kumar[2]

[1] Dept. of CSE, Modern Institute of Engineering and Technology, Kurukshetra University, Kurukshetra, India
[2]Dept. of CSE, Modern Institute of Engineering and Technology, Kurukshetra University, Kurukshetra, India

Available online at: www.ijcseonline.org

Accepted: 12/Jun/2018, Published: 30/Jun/2018

ABSTRACT- In this paper, various data mining techniques are analyzed and their proficiencies have been evolved. Medical professionals need a reliable prediction methodology to diagnose factors influencing diabetes. There are large quantities of information about patients and their medical conditions. Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. The main aim of this thesis is to show the comparison of different classification algorithms such as Multilayer perceptron neural network (MLPNN), Zero R, K-Star based on computing time, Correctly Classified Instances, Incorrectly Classified Instances, Mean absolute error, Root mean squared error, Relative absolute error, Root relative squared error, precision value , Recall , F-measure, the data evaluated using 25 fold Cross Validation error rate, error rate focuses True Positive, True Negative, False Positive and False Negative and the clustering algorithm such as K-means algorithm based on varied number of clusters and Sum of squared error. Classification is an important data mining technique with broad applications. It classifies data of various kinds. Classification is used in every field of our life. Classification is used to classify each item in a set of data into one of predefined set of classes or groups. Clustering analysis method is one of the main analytical methods in data mining; in which k-means clustering algorithm is most popularly/widely used for many applications. K-means algorithm has higher efficiency and scalability and converges fast when dealing with large data sets. Clustering is an adaptive procedure in which objects are clustered or grouped together, based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Waikato Environment for Knowledge Analysis or in short, WEKA is used to obtain the accuracy of algorithms and find out which algorithm is most suitable for user working on data of diabetic patients. Weka is a data mining tool. It contains many machine leaning algorithms. It provides the facility to classify our data through various algorithms.

Keywords: Multilayer perceptron neural network (MLPNN), Zero R, K-Star, K-means, WEKA.

I. INTRODUCTION

Data mining, at its core, is the transformation of large amounts of data into meaningful patterns and rules. Further, it could be broken down into two types: directed/supervised and undirected/unsupervised. In directed data mining, we try to predict a particular data point for example - the sales price of a house given information about other houses for sale in the neighborhood. In undirected data mining, you are trying to create groups of data, or find patterns in existing data. Additionally, the term data mining is all-encompassing, referring to dozens of techniques and procedures used to examine and transform data. The ultimate goal of data mining is to create a model, a model that can improve the way you read and interpret your existing data and your future data. Since there are so many techniques with data mining, the major step to creating a good model is to determine what type of technique to use. Data mining isn't just about outputting a single number: It's about identifying patterns and rules. There are different types of diseases predicted in data mining namely Hepatitis, Lung Cancer, Liver disorder, Breast cancer, Thyroid disease, Diabetes etc… This paper analyzes the Diabetes predictions.

Diabetes mellitus is one of the most common diseases in developed countries, growing at a rapid rate in developing countries like India. Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood sugar. There are three types of diabetes. Type 1 diabetes, Type 2 diabetes and Gestational diabetes. Type 1 diabetes is characterized by deficient insulin production and requires daily administration of insulin. The cause of type 1 diabetes is not known and it is not preventable with current knowledge. Symptoms include excessive excretion of urine (polyuria), thirst (polydipsia), constant hunger, weight loss, vision changes and fatigue. These symptoms may occur suddenly. Type 2 diabetes (formerly called non-insulin-dependent or adult-onset) results from the body's ineffective use of insulin. Type 2 diabetes

comprises 90% of people with diabetes around the world, and is largely the result of excess body weight and physical inactivity. Symptoms may be similar to those of Type 1 diabetes, but are often less marked.

Section I contains the introduction of proficiency analysis of various data mining techniques for diabetes prognosis, Section II contains methodology of proficiency analysis of various data mining techniques for diabetes prognosis, section III explain the performance evaluation measures, Section IV describes results and discussions of proficiency analysis of various data mining techniques for diabetes prognosis and Section V contains conclusion.

II. METHODOLOGY

A. Data Source:

For this research Pima Indian Diabetes Dataset has been used that contains data of females aged at least 21 years.768 instances were considered with 9 attributes defining the features of diabetic patients.

The 9 attributes are as follows:
a. Number of times pregnant
b. Plasma Glucose concentration
c. Diastolic B.P
d. Triceps skin fold thickness
e. 2 hour serum insulin
f. BMI
g. Diabetes pedigree function
h. Age
i. class variable

Table 1: Datasets of Diabetic Patients

S. NO	Name	Description	Unit	Value range
01	Pregnancy	No of Times Pregnant	Numeric value	0-9
02	Plasma	Plasma Glucose Concentration	Numeric value	0-199
03	Press	Diastolic Blood Pressure	mmHg	0-122
04	Skin	Triceps skin folds thickness	Mm	0-99
05	Insulin	2-Hours Serum Insulin	mu/Uml	0-846
06	Mass	Body Mass Index	Weight in kg Height	0-67.1
07	Pedi	Diabetes Pedigree Function	Numeric value	0.08-2.42
08	Age	age	Numeric value	21-81
09	Class	Diabetes Mellitus Type II	Numeric value	Positive=1, Negative = 0

B. WEKA Tool:

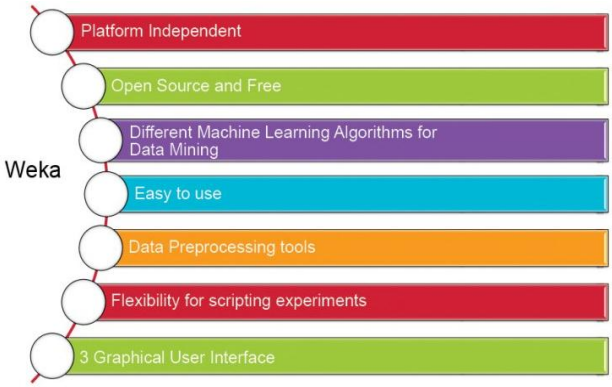

Figure 1: WEKA's features

WEKA stands for Waikato Environment for Knowledge Learning. It was developed by University of Waikato, New Zealand. The Weka or woodhen (Gallirallus australis) is an endemic bird of New Zealand. WEKA is open source software which consists of a collection of machine learning algorithms for data mining tasks. WEKA is freely available and is also platform independent. WEKA contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization. It is also well suited for developing new machine learning schemes. The data that is used for WEKA should be made into the ARFF (Attribute Relation file format) format and the file should have the extension dot ARFF (.arff). WEKA runs on almost any platform and is available on the web at www.cs.waikato.ac.nz/ml/weka.

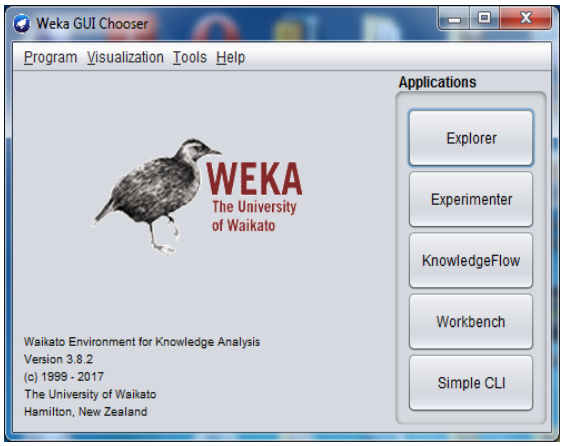

Figure 2: View of WEKA tool

The GUI Chooser consists of four buttons:

Explorer: An environment for exploring data with WEKA.

Experimenter: An environment for performing experiments and conducting statistical tests between learning schemes.

Knowledge Flow: This environment supports essentially the same functions as the Explorer but with a drag and-drop interface. One advantage is that it supports incremental learning.

Workbench: combines all GUI interfaces into one.

Simple CLI: Provides a simple command-line interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface.

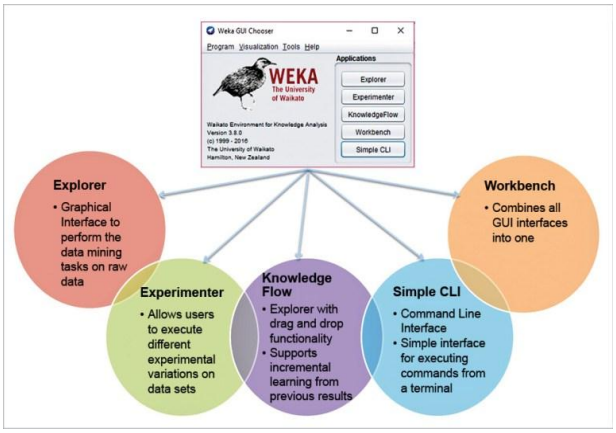

Figure 3: WEKA's application interfaces

III. PERFOMANCE EVALUATION MEASURES

In selecting the appropriate algorithms and parameters that best model the diabetes forecasting variable, the following performance metrics were used:

Time: This is referred to as the time required to complete training or modeling of a dataset. It is represented in seconds

Mean Absolute Error: Mean absolute error is the average of the difference between predicted and the actual value in all test cases; it is the average prediction error.

Mean Squared Error: Mean-squared error is one of the most commonly used measures of success for numeric prediction. This value is computed by taking the average of the squared differences between each computed value and its corresponding correct value. The mean-squared error is simply the square root of the mean-squared-error. The mean-squared error gives the error value the same dimensionality as the actual and predicted values.

Root relative squared error: Relative squared error is the total squared error made relative to what the error would have been if the prediction had been the average of the absolute value. As with the root mean squared error, the square root of the relative squared error is taken to give it the same dimensions as the predicted value.

Relative Absolute Error: Relative Absolute Error is the total absolute error made relative to what the error would have been if the prediction simply had been the average of the actual values.

Correctly Classified Instances: the percentage of instances whose predicted value agrees with actual value.

Analysis

• True positive rate .TP rate=TP/ (TP+TN)

Where, TP is True Positive, and FN is False Negative.

• True negative rate. TN rate=TN/ (TN+FP)

• F-measure: harmonic mean of precision and recall F measure= 2/ ((1/precision) + (1/recall))

• Recall: Fraction of relevant instances that are retrieved Recall= TP/ (TP+FN)

• Precision: Fraction of retrieved instances that are relevant Precision=TP/ (TP+FP)

To find the performance metrics such as sensitivity, specificity and accuracy, a distinguished confusion matrix is obtained based on the classification results from these algorithms. Confusion matrix is a matrix representation of the classification results as shown in Table 2.

Table.2

	Classified as Healthy	Classified as not Healthy
Actual Healthy	TP	FP
Actual not Healthy	FN	TN

Accuracy is the percentage of predictions that are correct. Sensitivity is the percentage of positive labeled instances that were predicted as positive.
These performance criterion for the classifiers in disease detection are evaluated as follows from the confusion matrix.

Accuracy = (TP+TN) / (TP+FP+TN+FN)
Sensitivity = TP / (TP+FN)
Specificity = TN / (FP+TN)

IV. RESULTS AND DISCUSSIONS

A. Diabetes Dataset:

The diabetes database consists of nine conditional attributes. The decisional attribute takes the values 0 or 1 (binary). As presented in the Figure 4 the attributes' values have different distributions. Distribution of the attributes *pregnant, pedigree* and *age* falls with the increase of the values of these attributes. The distribution of values of the attributes: *plasma*, *diastolic*, *triceps* and *mass* is of a bell-shape. The ranges with the highest cardinality are in the middle decreasing towards the ends of the graphs. All conditional attributes are multi-valued.

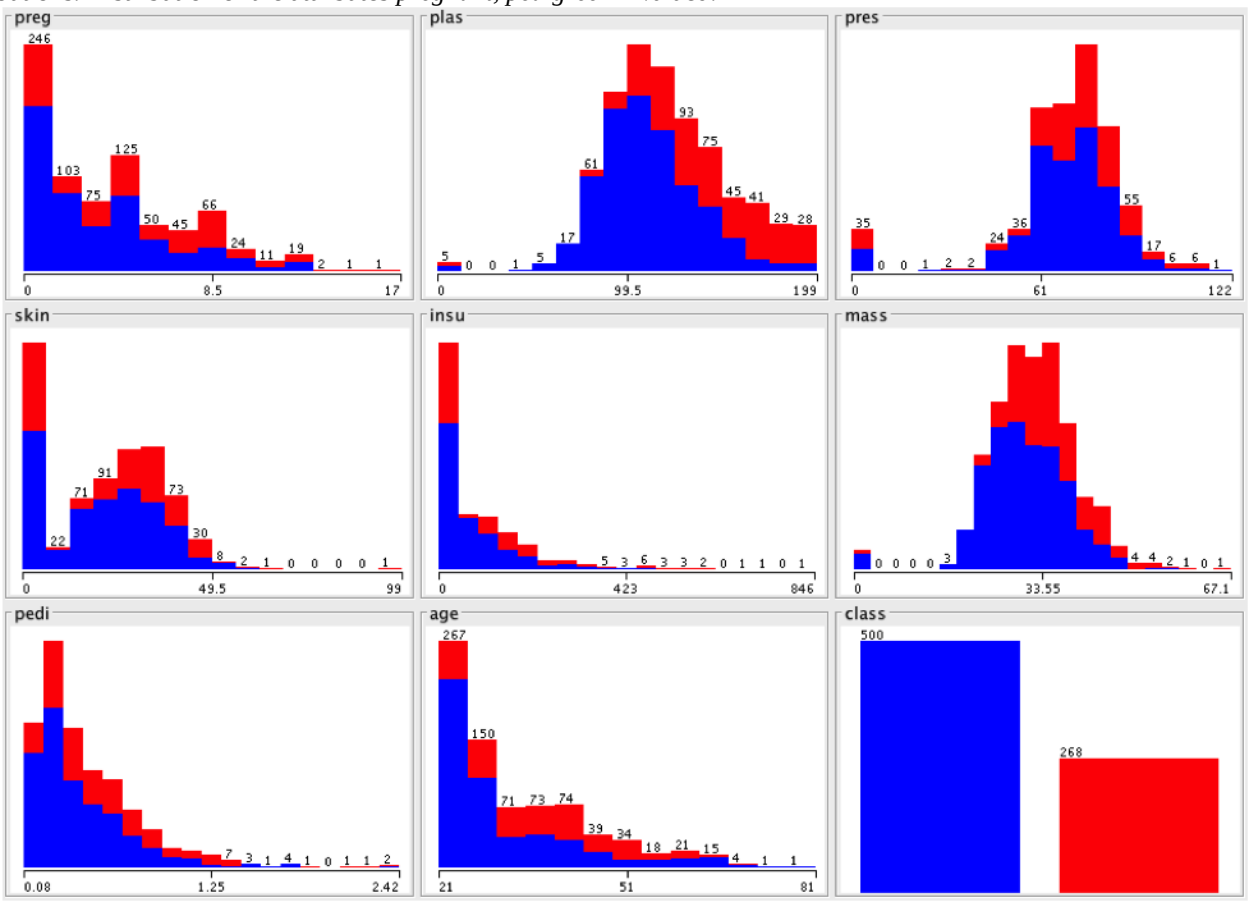

Figure 4: Distributions of the attributes of the diabetes data

Three different supervised classification algorithms i.e. Multilayer perceptron neural network (MLPNN), Zero R, K-Star have been used to analyze dataset in WEKA tool.

Step 1: The three algorithms can be filtered by using lowest computing time.

K-star had lowest computing time with highest computing time for Multilayer perceptron neural network.

Step 2: The above algorithms can filtered by using True Positive rate which is highest in zeroR and least for k-star.

Step 3: The above algorithms can filter by using Cross Validation Error rate i.e. Root Relative Squared Error (lowest error rate).which is lowest in Multilayer perceptron neural network and highest in zeroR.

Step 4: The above algorithms can filter by using highest accuracy and lowest computing time. The above three algorithms can be reduced to one best algorithm.

Step 5: Stop the process. We get the best one.

The step5 consists of values of different classification. According to these values the accuracy was calculated and it shows the highest accuracy which is shown in table 4. The table (3) represents the resultant values of above classified dataset using data mining supervised classification algorithms and lowest computing among the three. Logical chart (figure 5) is shown that compared algorithms on the basis of performance and computing time, precision value, Error rate

(25 fold Cross Validation, Bootstrap Validation) and finally the highest accuracy and again lowest computing time.

Table 3: Performance evaluation of different algorithm using WEKA tool

	KSTAR (25 fold cross validation)	ZEROR (25 fold cross validation)	Multilayer perceptron (25 fold cross validation)
Time taken to build model(sec)	0	0.01	0.25
Correctly Classified Instances	78%	66%	84%
Incorrectly Classified Instances	22%	34%	16%
Mean absolute error	27.9%	45.1%	16.65%
Root mean squared error	38.58%	47.5%	35.79%
Relative absolute error	61.85%	100%	36.92%
Root relative squared error	81.22%	100%	75.34%
precision	77.7%	66%	83.8%
Recall	78%	66%	84%
F-measure	77.8%	79.5%	83.9%
True positive	56%	66%	59%
True negative	22%	0%	25%
False positive	12%	34%	9%
False negative	10%	0%	7%

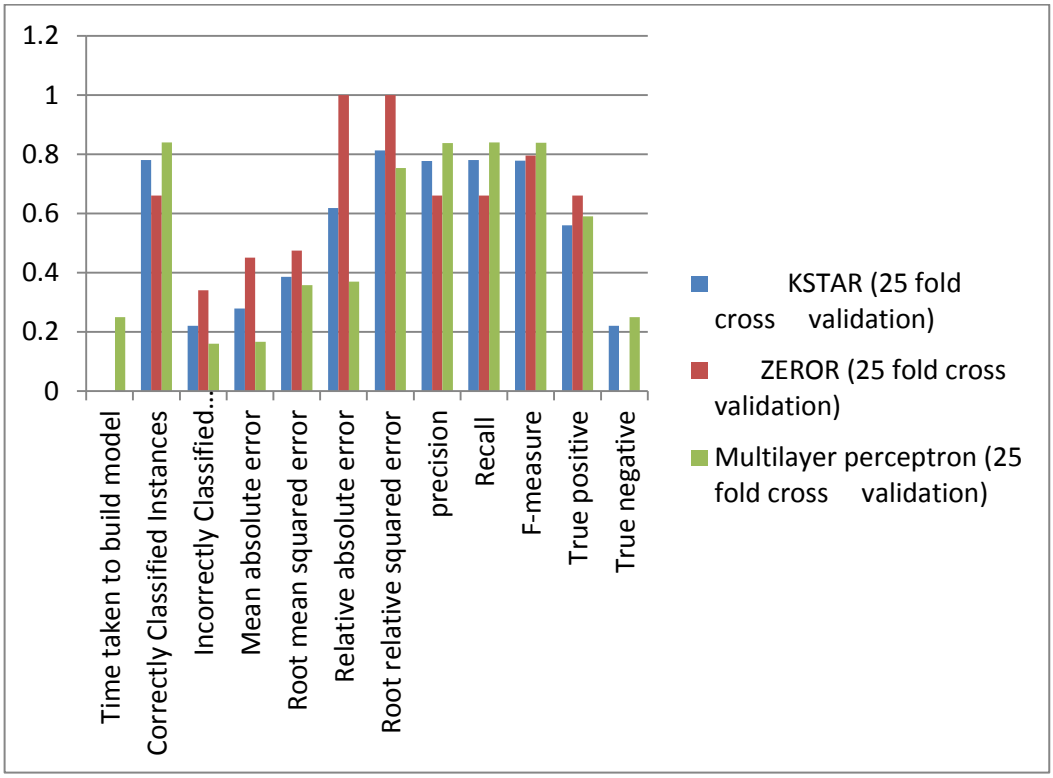

Figure 5: Graphical representation of comparison of the above data mining techniques for diabetic prognosis.

Table 4: Comparison of Accuracy between Different Algorithms

Accuracy (%)	
MLPNN	84%
ZeroR	66%
K-star	78%

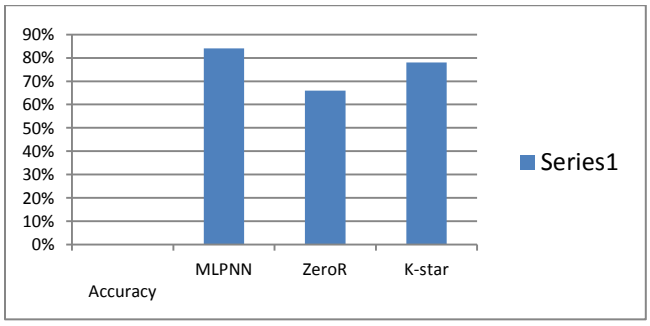

Figure 6: Graphical representation of Accuracy over different algorithm

Based on the above Fig. and Table, we can clearly see that the highest accuracy is 84% and the lowest accuracy is 66%. We can say that Multilayer perceptron neural network (MLPNN) is better.

B. K-MEANS CLUSTERING:

Clustering refers to the natural grouping of the data objects in such a way that the objects in the same group are similar with respect to the objects present in the other groups. There are broadly three types of clustering, namely, Hierarchal clustering, Density based clustering, and Partition based clustering. K-means clustering algorithm is simplest algorithm as compared to other algorithms and its performance is better than Hierarchical Clustering algorithm. Density based clustering algorithm is not suitable for data having very huge variations in density and hierarchical clustering algorithm is more susceptible to noisy data. Density based algorithm takes relatively less time to build a cluster but it's not better than the k-mean algorithm since density based algorithm has high log likelihood value, if the value of log likelihood is high then it makes bad cluster. Hence k-mean is best algorithm because it takes very less time to build a model. Hierarchal algorithm take more time than k-mean algorithm and cluster instances are also not good in hierarchal algorithm. Clustering is a vivid method. The solution is not exclusive and it firmly depends upon the analysts' choices [11].

K-means clustering is one of the simplest unsupervised classification techniques **.**The basic step of k-means clustering is simple. In the beginning, we determine number of cluster K and we assume the centroid or center of these clusters. We can take any random objects as the initial centroids or the first K objects can also serve as the initial centroids. Then the K means algorithm will do the three steps below until convergence. Iterate until *stable* (= no object move group):

1. Determine the centroid coordinate

2. Determine the distance of each object to the centroids

3. Group the object based on minimum distance (find the closest centroid). This is showed in figure 7 in steps

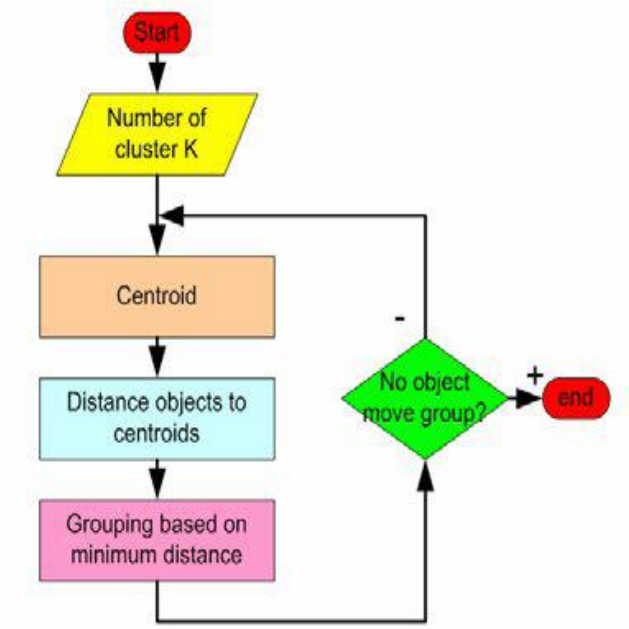

Figure 7: k-means clustering process

In this study, the standard Euclidean distance is used for the distance measure. The k-means clustering algorithm groups the data into k clusters. As shown

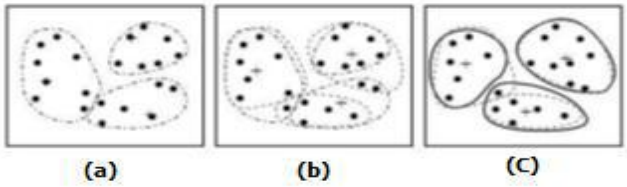

Fig. 8: K-means Algorithm

Table 5: K-Means Clustering Performance

No. of clusters	Cluster distribution	No. of iterations	Sum of squared error	Time taken to build model(sec)
5	80(10%) 112(15%) 144(19%) 164(21%) 268(35%)	16	115.4	0.1
6	61(8%) 106(14%) 178(23%) 135(18%) 268(35%) 20(3%)	24	109.7	0.18
7	48(6%) 105(14%) 140(18%) 132(17%) 268(35%) 20(3%) 55(7%)	27	107.1	0.13

. Figure 9: The results for no. of clusters, sum of squared error and no. of iterations are shown

Based on the above Fig. and Table, we can clearly see that increasing the no. of clusters reduced the rate of squared error which has to be achieved. We can say that for large datasets the higher no. of clusters is better as they provide lower rate of errors.
When a set of k=5 clusters are used the sum of squared error is 115.4, when set of k=6 clusters then the sum of squared error is 109.7 and when we used set of k= 7 clusters we got the sum of squared error least as 107.1.

V. CONCLUSION

The main goal of the research was to identify the most common data mining algorithms, implemented in modern Medical Decision Support Systems, and evaluate their performance on diabetic dataset. four algorithms were chosen: Multilayer Perceptron, , ZeroR, k-star and k-means. For evaluation, UCI diabetes database was used. Several performance metrics were utilized: percent of correct classifications, True/False Positive rates, AUC, Precision, Recall, F-measure and a set of errors. Weka tool was used to analyze these performances. This framework could be

stretched out further to discover conceivable outcomes of different diseases. In spite of the fact that the planned framework is exceedingly productive and matches the doctor's finding, different strategies like data mining algorithm could likewise be gone for. These aspects could be left for further examinations.

The results showed that for the present diabetes dataset, the percentage of correctly classified cases were highest in Multilayer perceptron followed by k-star, and zeroR respectively. The root mean squared error was found minimum in multilayer perceptron. Precision and recall was highest in multilayer perceptron. The underlying reason for such a research was the fact that no work was found which would analyze these three classification algorithms under identical conditions along with the clustering k-means algorithm. In this paper one of the approaches is Normalization which can affect the performance of a clustering algorithm, since we know that the normalized data would produce different result in comparison to the data which is not normalized and here we used a no. of clusters to normalize the data. Though the time taken to cluster set k=6 is minimum but the cluster set k=7 shows the less squared error (as shown in above figure). A good clustering method produces high-quality clusters to ensure that the inter-cluster similarity is low and the intra-cluster similarity is high, in other words, members of a cluster are more like each other than they are like members of a different cluster. As we have discussed in this paper k-means algorithm is good for large datasets.

REFERENCES

[1] P.Yasodha, M.Kannan, "Analysis of a Population of Diabetic Patients Databases in Weka Tool", International Journal of Scientific & Engineering Research, Volume 2, Issue 5, May-2011.

[2] K. Rajesh, V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis", International Journal of Engineering and Innovative Technology (IJEIT), Volume 2, Issue 3, September 2012.

[3] Sukhjinder Singh, Kamaljit Kaur,"A Review on Diagnosis of Diabetes in Data Mining", International Journal of Science and Research (IJSR), 2013.

[4] Veena Vijayan V, Aswathy Ravikumar, "Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus", International Journal of Computer Applications (0975 – 8887) Volume 95– No.17, June 2014.

[5] P.Yasodha, N.R. Ananthanarayanan, "Comparative Study of Diabetic Patient Data's Using Classification Algorithm in WEKA Tool", International Journal of Computer Applications Technology and Research, Volume 3– Issue 9, 554 - 558, 2014 .

[6] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly,"DIAGNOSIS OF DIABETES USING CLASSIFICATION MINING TECHNIQUES", International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol.5, No.1, January 2015.

[7] Sabreena Jan, Vinod Sharma," A Study of various data mining techniques for diabetic prognosis", International Journal of Modern Computer Science (IJMCS), Volume 4, Issue 3, June, 2016.

[8] 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 .

[9] Saman Hina*, Anita Shaikh and Sohail Abul Sattar, "Analyzing Diabetes Datasets using Data Mining", Journal of Basic & Applied Sciences, 2017, 13.

[10] S.Selvakumar, K.Senthamarai Kannan and S.GothaiNachiyar,"Prediction of Diabetes Diagnosis, Using Classification Based Data Mining Techniques", International Journal of Statistics and Systems, Volume 12, Number 2 (2017).

[11] Prakash Singh, Aarohi Surya, "PERFORMANCE ANALYSIS OF CLUSTERING ALGORITHMS IN DATA MINING IN WEKA", International Journal of Advances in Engineering & Technology, Jan., 2015**.**

[12] M.Mounika, S.D.Suganya, B.Vijayashanthi, S.KrishnaAnand," Predictive Analysis of Diabetic Treatment Using Classification Algorithm", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3), 2015, 2502-2505.

[13] Bharat Chaudhari, Manan Parikh,"A Comparative Study of clustering algorithms Using weka tools", International Journal of Application or Innovation in Engineering & Management (IJA*IEM)*, Volume 1, Issue 2, October 2012.

[14] Pallavi , Sunila Godara," A Comparative Performance Analysis of Clustering Algorithms", International Journal of Engineering Research and Applications (IJERA), Vol. 1, Issue 3, pp.441-445.

[15] Y. S. Thakare, S. B. Bagal," Performance Evaluation of K-means Clustering Algorithm with Various Distance Metrics, International Journal of Computer Applications (0975 – 8887), Volume 110 – No. 11, January 2015.

[16] Nidhi Singh, Divakar Singh,"Performance Evaluation of K-Means and Hierarchal Clustering in Terms of Accuracy and Running Time", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (3), 2012, 4119-4121.

[17] Arka Haldar, G.Prudhvi Raj, S.V.S.S Lakshmi, "Comparison of Different Classification Techniques Using WEKA for Diabetic Diagnosis", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 6, Issue 1, January 2018.

[18] Santosh Rani, Dr. Sandeep Kautish," Application of Data Mining Techniques for Prediction of Diabetes - A Review, International Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307.

Authors Profile

Ms. Misba Reyaz has completed her B.Tech in Information Technology at Model Institute of Engineering and Technology affiliated to Jammu University, Jammu, India, in 2015 and currently she is pursuing her M.Tech. In Computer Science and Engineering at Modern Institute of Engineering & Technology,, affiliated to Kurukshetra University, Kurukshetra, India. She is an IBM certified network associate. His areas of interest include Computer Networks, Network Security and Data Mining.

ER. Gagan Kumar is currently an Assistant Professor in the Department of Computer Science and Engineering at Modern Institute of Engineering & Technology, affiliated to Kurukshetra University, Kurukshetra, India. His research areas include Database Systems, Data Mining and Software Engineering.

