International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 8 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT183892

Comparative Analysis of Gestational Diabetes using Data Mining Techniques

Geetha. V. R[1], Dr. Jayaveeran. N[2]

[1]Assistant Professor, Department of Computer Science, A.V.C College (Autonomous), Mannampandal, Tamil Nadu, India

[2]Associate Professor and HOD, Department of Computer Science, Khadir Mohideen College, Adhirampattinam, Tamil Nadu, India

ABSTRACT

Data mining is process of extracting hidden knowledge from large volumes of raw data. Data mining is used to discover knowledge out of data and presenting it in a form that is easily understand to humans Disease Prediction plays an important role in data mining. Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain. Data Mining is used intensively in the field of medicine to predict Gestational diabetics which is affected the pregnant women. *Gestational diabetes* mellitus (GDM) is defined as any degree of glucose intolerance with onset or first recognition during pregnancy. This paper analyzes the Gestational diabetics predictions using different classification algorithms. Medicinal data mining has high potential for exploring the unknown patterns in the data sets of medical domain. These patterns can be used for medical analysis in raw medical data using decision table, Multi-layer perceptron and Naives Bayes algorithm and number of experiment has been conducted in WEKA tool to compare the performance of predictive data mining technique on the same dataset and the outcome reveals that Naives Bayes algorithm outperforms than other predictive methods such as Decision table, Multi-layer perceptron.

Keywords : Data Mining, Clustering, Classification, Gestational Diabetics Prediction, Error Rate

I. INTRODUCTION

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. An interdisciplinary subfield of computer science, it is an essential process wherein intelligent methods are applied to extract data patterns the overall goal of which is to extract information from a data set, and transform it into an understandable structure for further use. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. Health organizations today are capable of generating and collecting a large amount of data. This increase in data volume automatically requires the data to be retrieved when needed. With the use of data mining techniques is possible to extract the knowledge and determine interesting and useful patterns. The knowledge gained in this way can be used in the proper order to improve work efficiency and enhance the quality of decision making. Above the foregoing is a great need for new generation of theories and computational tools to help people with extracting useful information from the growing volume of digital data. Information technologies are implemented increasingly often in healthcare organizations to meet the needs of physicians in their daily decision making. Computer

systems used in data mining can be very useful to control human limitations such as subjectivity and error due to fatigue and to provide guidance to decision-making processes. The medical data about blood sugar levels can be analyzed to predict the diabetics' diseases using clustering and classification algorithm.

Gestational diabetes mellitus (GDM) is defined as any degree of glucose intolerance with onset or first recognition during pregnancy. The definition applies whether insulin or only diet modification is used for treatment and whether or not the condition persists after pregnancy. It does not exclude the possibility that unrecognized glucose intolerance may have antedated or begun concomitantly with the pregnancy. Risk assessment for GDM should be undertaken at the first prenatal visit. Women with clinical characteristics consistent with a high risk of GDM (marked obesity, personal history of GDM, glycosuria, or a strong family history of diabetes) should undergo glucose testing (see below) as soon as feasible. If they are found not to have GDM at that initial screening, they should be retested between 24 and 28 weeks of gestation. Women of average risk should have testing undertaken at 24–28 weeks of gestation. Women with GDM are at increased risk for the development of diabetes, usually type 2, after pregnancy. Obesity and other factors that promote insulin resistance appear to enhance the risk of type 2 diabetes after GDM, while markers of islet cell-directed autoimmunity are associated with an increase in the risk of type 1 diabetes. Offspring of women with GDM are at increased risk of obesity, glucose intolerance, and diabetes in late adolescence and young adulthood. Fig 1 specifies insulin level variations in GDM

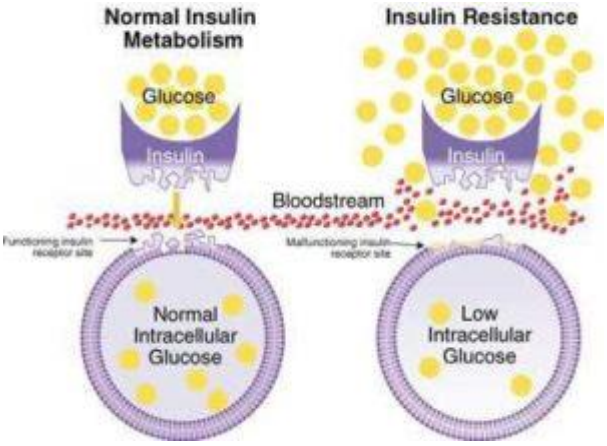


Figure 1 : Insulin level at GDM

Applications of data mining techniques in health care

The different classification algorithms mentioned below in figure 1 are used to predict or to analyses diabetic diseases. The fig 2 illustrated various clustering, association and classification algorithms in data mining. In classification field includes neural networks, decision tree, Bayesian networks, genetic algorithm, ANN, KNN, Naives Bayesian, Apirori algorithm, support vector machine, K-NN classification and so on.

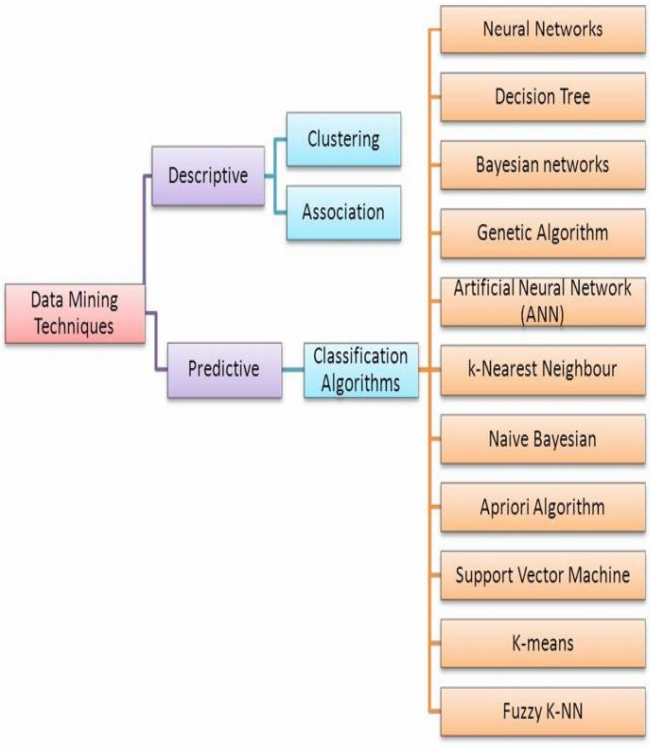


Figure 2 : Different techniques in Healthcare domain

In terms of prediction and decision making, Data mining techniques have substantial expansion in medical industry with respect to various diseases like diabetes, heart disease, liver diseases, cancer and others.

II. RELATED WORK

Lashari S, et.al,…(2013) [1] examines comparative analysis of applications of data mining techniques has been presented. Thus, the existing literature suggests that we do not lose sight of the current and future potential of applications of data mining techniques that can impact upon the successful classification of medical data into a thematic map. Thus, there is a great potential for the use of data mining techniques for medical data classification, which has not been fully investigated and would be one of the interesting directions for future research..

Kittipol Wisaeng, et.al,…(2013) [2] conducted a comparison between four data mining techniques namely, NBT, Ridor, NB and J48 relies on the careful KDD steps could be used to classification in medical databases. The performance of the techniques is validated by recall, precision and accuracy values. The NBT technique show better performance for medical databases (100%), but J48 and Ridor are also useful and may be better fit to deal with our case. In the process of KDD, choice of parameters and the construction of high quality training and test data sets are important steps. Overall, experimental results show that careful KDD steps and appropriate technique together provide best classification in medical databases This technique intends to help expert in risk factor analysis in diabetes mellitus to faster and more easily.

Anil Sharma, et.al,…(2017) [3] provide interface to get data and to retrieve some interesting patterns out of it which are further useful to attain new knowledge. There are varieties of parameters defined in the literature which provide base for a tool to perform analysis and different tools are available to perform these analysis. This is quite interesting to perform a comparative analysis of these tools and to observe their behavior based on some selected parameters which will further be helpful to find the most appropriate tool for the given data set and the parameters.

Satish Kumar David, et.al,…(2013) [4] compared algorithms based on their accuracy, learning time and error rate. The availability of huge amounts of data resulted in great need of data mining technique in order to generate useful knowledge. In the present study we provide detailed information about data mining techniques with more focus on classification techniques as one important supervised learning technique. And observed that there is a direct relationship between execution time in building the tree model and the volume of data records, while there is also an indirect relationship between execution time in building the model and the attribute size of the data sets.

Pon Periasamy, et.al,…(2015) [5] proposes a system to determine the presence of different dermatological diseases in Kottayam and Alappuzha. The system is built using Naïve Bayes classifier which is based on Bayesian theorem. Author collected data from various healthcare areas and used Naïve Bayes Algorithm as they produces higher predictive accuracies. The percentage of eight skin diseases is predicted effectively using the implementation in Java platform. On the basis of imported inputs, the prediction window gives the results indicating the chance of occurrence of diseases.

Disease prediction using data mining techniques

Data mining is the process of analyzing and summarizing data from different perspectives and converting it into useful information. It also helps the

healthcare researchers for making efficient healthcare policies, constructing drug recommendation systems, developing health profiles of individuals Medical databases are very bulky that need computerized programs to find latent trends that will help in medical diagnosis and treatment. In the wake of data mining techniques, especially medical data mining techniques, the health care domain has made significant progress in using the technologies in prevention and diagnosis of disease. Using diabetic's datasets to classify the diseases. Diabetic Data used in this research with 5 attributes. The dataset contains the attributes such as Maximum Glucose level, Chlorpropamide, Insulin level, Number of diagnosis, Diabetic_med. These attributes can perform classification and clustering using tool named as WEKA for WINDOWS OS with any configuration. The imported datasets and variables are shown in Table 1. The attributes that have considered in this proposed work are:

Table 1. Dataset Description

S.No	Attribute	Description
1	Maximum glucose	Glucose can be expressed as values
2	chlorpropamide	Representing as insulin class level as state Up, Down, Steady and No
3	Insulin level	Representing as blood sugar level as state Up, Down, Steady and No
4	Number of diagnosis	Represent as numeric values
5	Diabetic_med	Represent the as yes or no values

The process of proposed work is shown in fig 3.

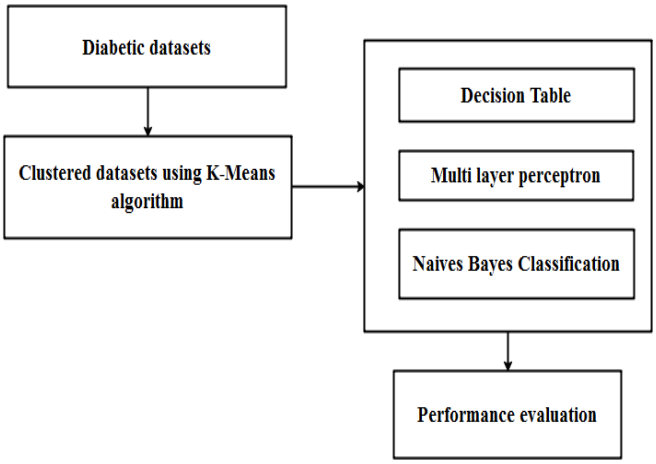


Figure 3: Proposed Framework

Fig 3 provide framework for the proposed system include steps such as clustering and classification. K-means clustering can be applied after that applies classification algorithms such as decision table, multi-layer perceptron and Naives Bayes algorithm. Finally compare the classification algorithm in terms of error rates in performance evaluation.

Simple K-means algorithm:

The diabetic datasets are clustered using simple K-means algorithm. They are provided with a hard and fast of data instances that have to be grouped in keeping with a few notion of correspondence. The algorithm devises access only to the set of features describing each object; it is not given any information as to where each of the instances should be placed within the partition. K-way clustering is a method generally used to mechanically partition a statistics set into okay organizations. It proceeds by selecting k initial cluster centers and then iteratively refining the results. The algorithm converges when there is no further change in assignment of instances to clusters. The diabetic datasets are grouped as two clusters named as normal and abnormal. K-means clustering can be applied after perform the preprocessing for uploaded diabetic datasets. In

WEKA tool, Choose cluster options and click drop down list to pick simple K means algorithm and then start cluster to group the classes as 0 and 1.

The basic algorithm pseudo code as follows:

Input: X = {x1, x2, x3,…..,xn} be the set of data points , Y= {y1,y2,y3…yn} be the set of data points and V = {v1,v2,v3,….,vn} be the set of centers
Step 1: Select 'c' cluster centers arbitrarily
Step 2: Calculate the distance between each pixels and cluster centers using the Euclidean Distance metric as follows

$$Dist(X,Y) = \sqrt{\sum_{j=1}^{n}(X_{ij} - Y_{ij})^2}$$

X, Y are the set of data points
Step 3: Pixel is assigned to the cluster center whose distance from the cluster center is minimum of all cluster centers
Step 4: New cluster center is calculated using

$$V_i = \frac{1}{C_i}\sum_{1}^{ci} x_i$$

Where Vi denotes the cluster center, ci denotes the number of pixels in the cluster
Step 5: The distance among every pixel and new obtained cluster facilities is recalculated
Step 6: If no pixels were reassigned then stop. Otherwise repeat steps from 3 to 5

Decision Table:

Decision tables are a concise visual representation for specifying which actions to perform depending on given conditions. They are algorithms whose output is a set of actions. The information expressed in decision tables could also be represented as decision trees or in a programming approach in terms of if then else rules if-then-else rules. Each decision corresponds to a variable, relation or predicate whose possible values are listed among the condition alternatives. In WEKA tool, choose classify option and pick the rules options to perform Decision table classifier based on class attribute. To build a decision tree, we need to calculate two types of entropy using frequency tables as follows:

a) Entropy using the frequency table of one attribute

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

b) Entropy using the frequency table of two attributes

$$E(T,X) = \sum_{c\in X} P(c)E(c)$$

Multi-layer Perceptrons:

A multilayer perceptron (MLP) is a class of feed forward artificial neural network. An MLP consists of at least three layers of nodes. Except for the input nodes, each node is a neuron that uses a nonlinear activation function.MLP utilizes a learning technique called back propagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable. Multilayer perceptrons are sometimes colloquially referred to as "vanilla" neural networks, especially when they have a single hidden layer. A perceptron is a linear classifier; that is, it is an algorithm that classifies input by separating two categories with a straight line. In WEKA tool, choose classify option and pick the functions options to perform Multilayer perceptron based on class attribute. Input is typically a feature vector x multiplied by weights wand added to a bias

$$b: y = w * x + b.$$

The basic layout is shown in fig 4.

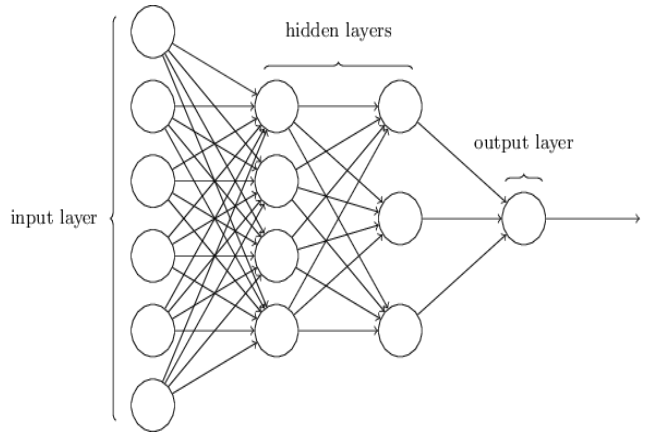


Figure 4 : Multi-layer Perceptron

Naives Bayes Algorithm:

In machine learning, naïve bayes classifiers are a family of simple probabilistic classifiers with strong independence assumptions between the features. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. In WEKA tool, choose classify option and pick the bayes options to perform Naives Bayes classifier based on class attribute. Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:

$$P(c|x) = \frac{P(X|C)P(c)}{P(x)} \text{ ---------- Eqn(1)}$$

$P(c|X) = P(x_1|c) * P(x_2|c)*....*P(x_n|c)*P(c)$

Where

- P(c|x) is the posterior probability of class (c, target) given predictor(x, attributes).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

The diabetic dataset is applied to above classification algorithms to analyze the performance in terms of error rates. Based on these results proposed work provide naives bayes algorithm has reduced error rates than the decision table and multi-perceptron approaches.

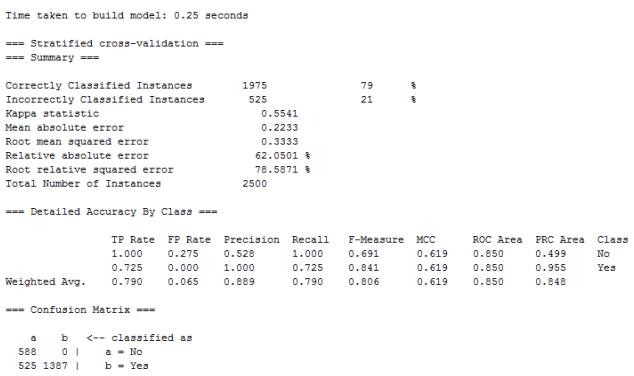
III. EXPERIMENTAL RESULTS

We can import the datasets from UCI repository and then perform the K-means clustering as illustrated in fig 4.

```
Cluster 0: None,6,No,No,Yes
Cluster 1: None,9,No,No,Yes

Missing values globally replaced with mean/mode

Final cluster centroids:
                                Cluster#
Attribute          Full Data          0          1
                  (101766.0) (41561.0) (60205.0)
=================================================
max_glu_serum           None       None       None
number_diagnoses      7.4226       5.38     8.8326
chlorpropamide            No         No         No
insulin                   No         No         No
diabetesMed              Yes        Yes        Yes


Time taken to build model (full training data) : 0.58 seconds

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

Clustered Instances

0       41561 ( 41%)
1       60205 ( 59%)
```

Figure 5 : K- Means cluster result

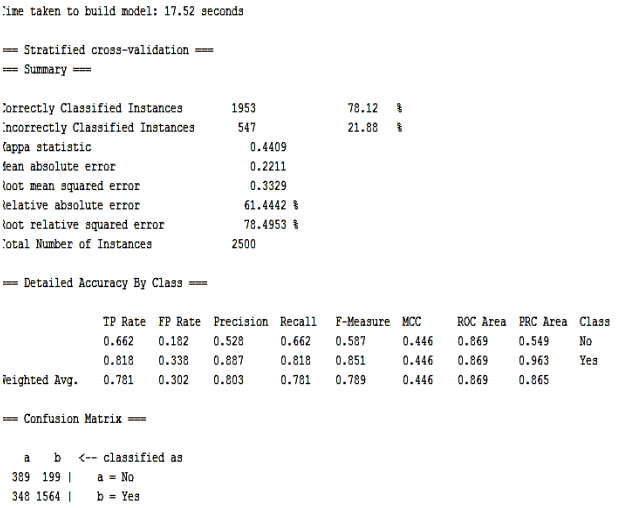
The five attributes are clustered as cluster 0 and cluster 1. The k value is 2. Each attribute is read and calculate the mean value. The cluster matrix is constructed as cluster 0 with 41% and cluster 1 with 59%. The classification results with various algorithms illustrated in fig 6, 7 and fig 8.



```
Time taken to build model: 0.25 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1975               79      %
Incorrectly Classified Instances       525               21      %
Kappa statistic                          0.5541
Mean absolute error                      0.2233
Root mean squared error                  0.3333
Relative absolute error                 62.0501 %
Root relative squared error             78.5871 %
Total Number of Instances             2500

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               1.000    0.275    0.528      1.000    0.691      0.619    0.850     0.499     No
               0.725    0.000    1.000      0.725    0.841      0.619    0.850     0.955     Yes
Weighted Avg.  0.790    0.065    0.889      0.790    0.806      0.619    0.850     0.848

=== Confusion Matrix ===

   a    b   <-- classified as
 588    0 |    a = No
 525 1387 |    b = Yes
```

Figure 6 : Decision table

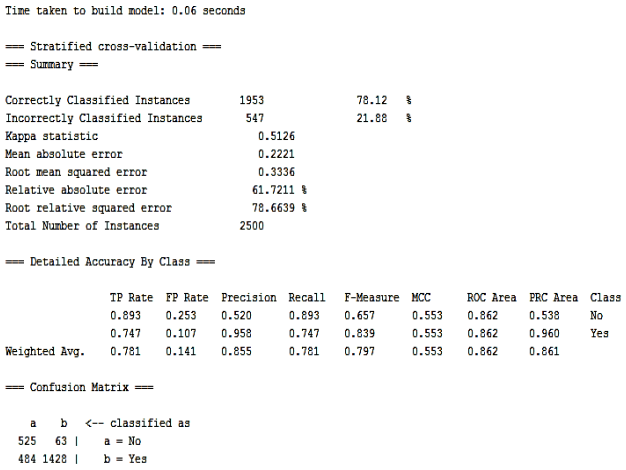
The decision table is constructed for the clustered datasets with 10 cross validation results. The summary of the result listed as error rate as 0.23 and with true positive rate is 1.00.



```
Time taken to build model: 17.52 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1953               78.12   %
Incorrectly Classified Instances       547               21.88   %
Kappa statistic                          0.4409
Mean absolute error                      0.2211
Root mean squared error                  0.3329
Relative absolute error                 61.4442 %
Root relative squared error             78.4953 %
Total Number of Instances             2500

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.662    0.182    0.528      0.662    0.587      0.446    0.869     0.549     No
               0.818    0.338    0.887      0.818    0.851      0.446    0.869     0.963     Yes
Weighted Avg.  0.781    0.302    0.803      0.781    0.789      0.446    0.869     0.865

=== Confusion Matrix ===

   a    b   <-- classified as
 389  199 |    a = No
 348 1564 |    b = Yes
```

Figure 7 : Multi-layer perceptron

The Multi-layer is constructed for the clustered datasets with 10 cross validation results. The summary of the result listed as error rate as 0.22 and with true positive rate is 1.00.



```
Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1953               78.12   %
Incorrectly Classified Instances       547               21.88   %
Kappa statistic                          0.5126
Mean absolute error                      0.2221
Root mean squared error                  0.3336
Relative absolute error                 61.7211 %
Root relative squared error             78.6639 %
Total Number of Instances             2500

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.893    0.253    0.520      0.893    0.657      0.553    0.862     0.538     No
               0.747    0.107    0.958      0.747    0.839      0.553    0.862     0.960     Yes
Weighted Avg.  0.781    0.141    0.855      0.781    0.797      0.553    0.862     0.861

=== Confusion Matrix ===

   a    b   <-- classified as
 525   63 |    a = No
 484 1428 |    b = Yes
```

Figure 8 : Naives Bayes algorithm

The probability values are constructed for the clustered datasets with 10 cross validation results. The summary of the result listed as error rate as 0.21 and with true positive rate is 1.00. From above model construction we can evaluate the performance of each algorithm and compare the performance based on MSE, RMSE, RAE, RRSE and shown in table and performance graph. The performance is listed in table 2 and performance graph in fig 9.

Table 2. Performance of various algorithms

Algorithms	MSE	RMSE	RAE	RRSE
Decision table	0.23	0.33	0.62	0.78
Multi-layer Perceptron	0.22	0.32	0.61	0.77
Naives Bayes	0.21	0.31	0.60	0.76

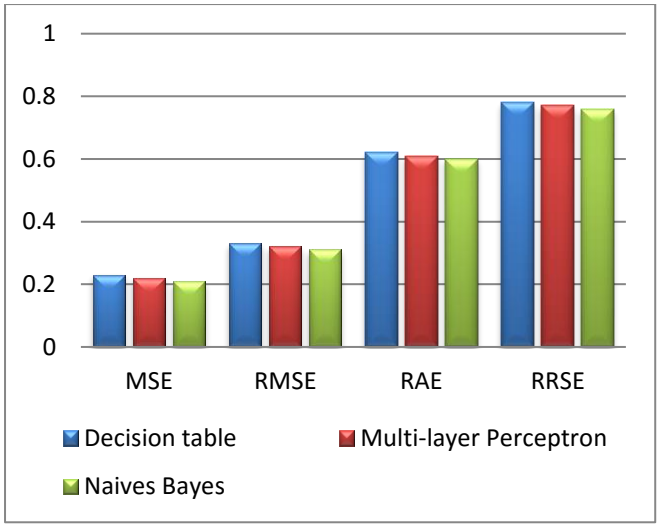


Figure 9 : Performance chart

From the above comparison, can be illustrated in fig 8, Naives Bayes outperforms than the existing algorithms and provides reduce number of MSE, RMSE, RAE and RRSE values.

IV. Conclusion

Using data mining technology for disease prediction and diagnosis has become the focus of attention. Data mining technology provides an important means for extracting valuable medical rules hidden in medical data and acts as an important role in disease prediction and clinical diagnosis. In the current study, have demonstrated, using a large sample of patients hospitalized with classification. In this research work, the classification rule algorithms namely Decision table, Multi-layer perceptron and Naives bayes are used for classifying datasets which are uploaded by user. By analyzing the experimental results it is observed that the Naïve Bayes classification technique has yields better result than other techniques. In future we tend to improve efficiency of performance by applying other data mining techniques and algorithms.

V. REFERENCES

[1]. Lashari, Saima Anwar, and Rosziati Ibrahim. "Comparative analysis of data mining techniques for medical data classification." 4th International Conference on Computing and Information. 2013.

[2]. Wisaeng, Kittipol. "An empirical comparison of data mining techniques in medical databases." International Journal of Computer Applications 77.7 (2013).

[3]. Sharma, Anil, and Balrajpreet Kaur. "A Research Review On Comparative Analysis Of Data Mining Tools, Techniques And Parameters." International Journal of Advanced Research in Computer Science 8.7 (2017).

[4]. David, Satish Kumar, A. T. Saeb, and Khalid Al Rubeaan. "Comparative analysis of data mining tools and classification techniques using weka in medical bioinformatics." Computer Engineering and Intelligent Systems 4.13 (2013): 28-38.

[5]. Pon Periasamy. "A Review on Health Data Using Data Mining Techniques." International Research Journal of Engineering and Technology (IRJET) 2.07 (2015): 2395-0056.

[6]. Lashari, Saima Anwar, and Rosziati Ibrahim. "Comparative analysis of data mining techniques for medical data classification." 4th International Conference on Computing and Information. 2013.

[7]. Wisaeng, Kittipol. "An empirical comparison of data mining techniques in medical databases." International Journal of Computer Applications 77.7 (2013).

[8]. Sharma, Anil, and Balrajpreet Kaur. "A Research Review On Comparative Analysis Of Data Mining Tools, Techniques And Parameters." International Journal of Advanced Research in Computer Science 8.7 (2017).

[9]. David, Satish Kumar, A. T. Saeb, and Khalid Al Rubeaan. "Comparative analysis of data mining

tools and classification techniques using weka in medical bioinformatics." Computer Engineering and Intelligent Systems 4.13 (2013): 28-38.

[10]. Pon Periasamy. "A Review on Health Data Using Data Mining Techniques." International Research Journal of Engineering and Technology (IRJET) 2.07 (2015): 2395-0056.

[11]. Payal Dhakate , Suvarna Patil , K. Rajeswari , Dr. V.Vaithiyananthan , Deepa Abin, ?Preprocessing and Classification in WEKA using different classifiers?, Journal of Engineering Research and Applications www.ijera.com ISSN : 2248-9622, Vol. 4, Issue 8(Version 1), August 2014, pp.

[12]. Remco R. Bouckaert, Eibe Frank, Mark A. HallGeoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten, ?WEKA—Experiences with a Java Open-Source Project?, Journal of Machine Learning Research, November 2010.

[13]. Trilok Chand Sharma, Manoj Jain, ?WEKA Approach for Comparative Study of Classification Algorithm?, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 4, April 2013.

[14]. P.Yasodha, M. Kannan, ?Analysis of a Population of Diabetic PatientsDatabases in Weka Tool?, Research Vol 2, Issue 5, May-2011.

[15]. Vikas Chaurasia, Saurabh Pal, ?Data Mining Approach to Detect Heart Dieses?, International Journal of Advanced Computer Science and Information Technology Vol. 2.

Cite this article as :

Geetha. V. R, Dr. Jayaveeran. N, "Comparative Analysis of Gestational Diabetes using Data Mining Techniques", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 3 Issue 8, pp. 326-334, November-December 2018. Available at
doi : https://doi.org/10.32628/CSEIT183892
Journal URL : http://ijsrcseit.com/CSEIT183892