Analyzing Diabetes Datasets using Data Mining

Saman Hina*, Anita Shaikh and Sohail Abul Sattar

Department of Computer Science & Software Engineering, NED University of Engineering & Technology, Karachi Pakistan

Abstract: Data mining techniques explore critical information in various domains (for example in CRM (customer relationship management), HR (Human Resource), GIS (Geographic Information System) etc.) but most importantly in medical domain. In medical domain, data mining can assist in minimizing the risk of developing some stereotyped diseases such as cancer, heart diseases, diabetes etc. In this paper, authors have focused data of Diabetic patients. Diabetic patient's body lacks ability to manage the glucose level in blood which can affect the other body mechanism. This can lead to the dysfunctioning of other physiological and psychological parameters such as reduced weight, skin folding. These parameters may be a valuable data source for the research. Diabetes mellitus placed 4th among Noncommunicable diseases-NCDs, caused 1.5 million global deaths each year worldwide [1]. The increase in digital information has elevated numerous challenges especially when it comes to automated content analysis and to make use of some machine learning techniques to aid mankind for predicting the non-communicable diseases like diabetics. . In this research different classifying algorithms such as Naïve bayes, MLP, J.48, ZeroR, Random Forest, and Regression were applied to depict the result. The conducted research aims to extract knowledge from the given set of data and to generate comprehensive and intelligent results.

Keywords: Data mining, Classification, Algorithm, Diabtes MelitusType II.

I. INTRODUCTION

Non-communicable diseases (NCDs) which include stroke, heart disease, cancer, chronic lung cancer and diabetes they together are responsible for almost 70% of the deaths worldwide in which Diabetes mellitus Type II is most common in all [1]. The number of patients suffered has quadrupled since 1980.it is estimated that 422 million people have diabetes all over the world and this figure may get doubles in the next 20 years [2]. The top 10 countries which are affected are India, China, USA, Indonesia, Japan, Pakistan, Russia, Brazil, Italy and Bangladesh [3].

About seven million Pakistanis had diagnosed Type II Diabetes mellitus it is estimated that in 2035, the figure will reach up to 12 million [4]. In this situation, it is necessary to look into the facts and the risk factors involved.

This paper meant to be written to give an idea of utilizing the information taken in different hospital as their procedures includes assessing the patient by taking some medical history before prescribing anything. This information may give some diagnostic details of the disease by comparing different data mining algorithm.

II. BACKGROUND

The process of data mining allows ascertaining the patterns in the provided datasets by simply applying combination of methods like artificial intelligence, machine learning, statistics and database system.

The objective of this research was to obtain information from the dataset and alter it to a more meaningful structure.

The data mining tool opt for this research is WEKA. WEKA is known for data mining and contains well-known algorithms for data pre-processing, classification, regression, clustering, association rules and visualization. It is also suited for developing new machine learning schemes [2].

In this particular example, different classifiers were used which include naïve bayes, decision tree and regression techniques and neural networks to get the best results out of it.

III. MATERIALS AND METHODS

The datasets had been taken from Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases these datasets includes records of 768 patients, out of which 500 tested negative while 268 of them were tested positive [5].

The description of the dataset with the nine attributes in Table **1**, help us to understand the possible prediction of this disease and which of the algorithm is more suitable for it.

In Table **1** the first eight attribute are the inputs set as input and the ninth attribute is the result which is used as a target which is either "Positive" or "Negative".

*Address correspondence to this author at the Department of Computer Science & Software Engineering, NED University of Engineering & Technology, Karachi Pakistan; Tel: 99261261, Ext: 2498;
E-mail: samhaque@neduet.edu.pk, saman.hina@gmail.com

ISSN: 1814-8085 / E-ISSN: 1927-5129/17

Table 1: Datasets of Diabetic Patients

S. NO	Name	Description	Unit	Value range
01	Preg	No of Times Pregnant	Numeric value	0-9
02	Plas	Plasma Gulocose Concentrartion	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 in ,m-2.	0-67.1
07	Pedi	Diabets Pedigree Function	Numeric value	0.08-2.42
08	Age	Age	Numeric value	21-81
09	Classs	Diabetes Melitis Type II	Numeric value	Postive =1 ,Negative = 0

IV. Graphical Representation of Attributes

Figure **1** is a graphical representation of the original test results shown as positive (blue) and negative (red) for different parameters (preg, plas, press, skin, insulin, mass, pedi, age, class).

V. CLASSIFICATION ALGORITHM AND THEIR EVALUATION

Output Prediction

The results were based on 90% percentage split. The comparison of the two initial results of different algorithms can be seen in Table **2**. "Actual " and "predicted" represents the original results and the predicted results respectively.

However in Tables **3-8** the column "error" represents the prediction error.

A. Naïve Bayes

This algorithm is named after Thomas Bayes who proved the bayes theorem. Naive Bayes is suitable in our situation in it solve the problem of identifying the possibilities of how many people are more prone towards diabetes.

Test Positive= Red Test Negative =Blue

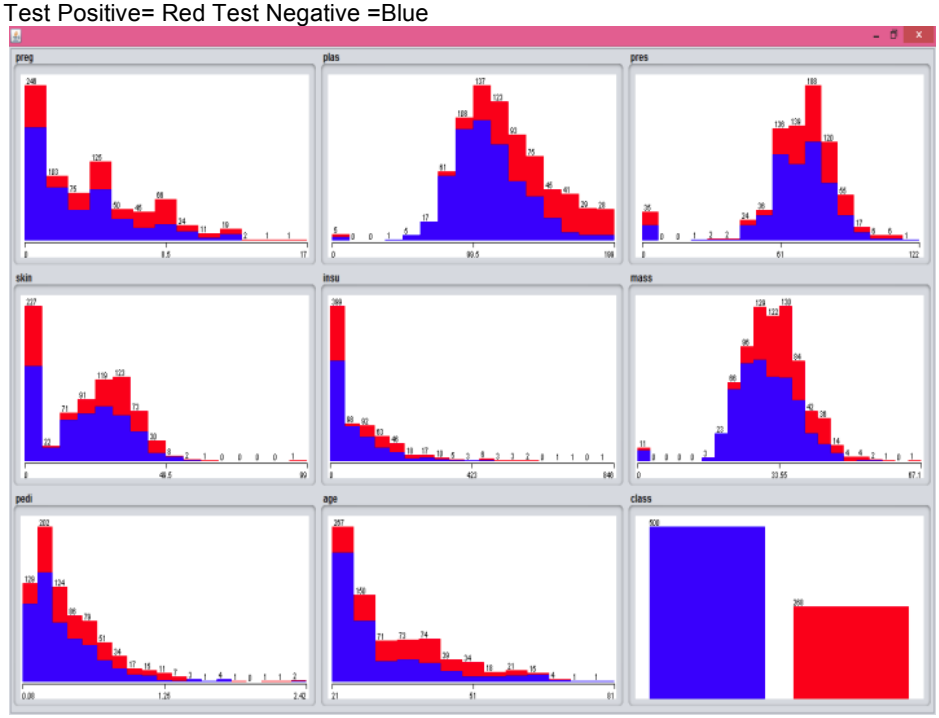

Figure 1: Weka Output (Negative and Positive outcomes with respect to different classes).

Table 2: Comparison Of Prediction Of First Two Instances By Using Different Algorithm

Decision Attributes	Logistic Regression	Naïve Bayes	ZeroR	J.48	MLP	Random Foest
Instance 1						
Actual	N	N	N	N	N	N
Predicted	N	N	N	N	N	N
Prediction (True/False)	True	True	True	True	True	True
Instance 2						
Actual	P	P	P	P	P	P
Predicted	P	N	N	P	N	N
Prediction (True/False)	True	False	False	True	False	False

N= tested_negative.
P= tested_positive.

Table 3: Prediction using Naive Bayes

inst#	actual	predicted	error
1	1:tested_negative	1:tested_negative	0.99
2	2:tested_positive	1:tested_negative	+0.67
3	1:tested_negative	1:tested_negative	0.501
4	1:tested_negative	1:tested_negative	0.825

This algorithm works on probability distribution function.

In Table **3** Error column 0.99 means there is 99% chance of that instance to test negative which is true and 1% possibility that the instances could test positive.

"+ "means prediction came out untrue. However, in the second instance 67% chance for the instance to test negative as compared to the instance in which it can have 99% surety that it proved wrong.

0.67 is not to close to 0.99 which gives the algorithm a benefit of doubt as to predict positive or negative.

B. Zero R

ZeroR is the simplest classification method. It is that type of classification method which would lean on the target and ignore other attributes invasion.

In Table **4**, it always generates the same result for every instance either 65% (0.352 test negative) or 35% (0.352 test positive) means there is no other possibility of changing the output either it is Yes or No.

This algorithm is very useful when the involvement of every other parameter is less significant.

C. Logistic Regression

Logistic regression was developed by statistician David Cox in 1958. Logistic regression measures the

Table 4: Prediction using Zero R

inst#	actual	predicted	Error prediction
1	1:tested_negative	1:tested_negative	0.648
2	2:tested_positive	1:tested_negative	+0.648
3	1:tested_negative	1:tested_negative	0.648
4	1:tested_negative	1:tested_negative	0.648
5	2:tested_positive	1:tested_negative	+0.648

Table 5: Prediction using Logistic Regression

inst#	actual	predicted	error prediction
1	1:tested_negative	1:tested_negative	0.981
2	1:tested_ positive	1:tested_ positive	0.517
3	1:tested_negative	1:tested_ positive	+0.5
4	1:tested_negative	1:tested_negative	0.721
5	1:tested_ positive	2:tested_positive	0.582
6	1:tested_negative	1:tested_negative	0.841
7	1:tested_ positive	2:tested_positive	0.921
8	1:tested_negative	2:tested_ negative	0.927

relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function [6].

Some of the initial prediction bases on test split data can be seen in Table **5**.

D. Random Forest

Random forest generates many single classification trees. To classify a new object from an input, put the input vector down each of the trees in the forest. Each tree generates their own results and then they select one set of a class as shown in Figure **2** [7].

```
plas < 111.5
|  preg < 7.5
|  |  skin < 29.5
|  |  |  age < 30.5
|  |  |  |  skin < 19.5 : tested_negative (122/0)
|  |  |  |  skin >= 19.5
|  |  |  |  |  plas < 94.5 : tested_negative (40/0)
|  |  |  |  |  plas >= 94.5
|  |  |  |  |  |  mass < 32.7 : tested_negative (18/0)
|  |  |  |  |  |  mass >= 32.7
|  |  |  |  |  |  |  preg < 0.5 : tested_positive (2/0)
|  |  |  |  |  |  |  preg >= 0.5
|  |  |  |  |  |  |  |  skin < 23.5 : tested_positive (1/0)
|  |  |  |  |  |  |  |  skin >= 23.5 : tested_negative (10/0)
|  |  |  |  |  pedi >= 0.22
|  |  |  |  |  |  mass < 37 : tested_positive (15/0)
|  |  |  |  |  |  mass >= 37
|  |  |  |  |  |  |  pres < 89
|  |  |  |  |  |  |  |  skin < 36.5 : tested_negative (5/0)
|  |  |  |  |  |  |  |  skin >= 36.5 : tested_positive (2/0)
|  |  |  |  |  |  |  pres >= 89 : tested_positive (3/0)
|  |  |  plas >= 146.5 : tested_positive (48/0)

Size of the tree : 189
```

Figure 2: Random Forest (Tree).

E. Multilayer Perception

It works on how different attributes results process and interact with one another and alter their results in such a way that the final outcome is the filtered version of each node (neuron).

Table 6: Prediction using Random Forest

inst#	actual	predicted	Error prediction
1	1:tested_negative	1:tested_negative	0.58
2	1:tested_negative	1:tested_negative	0.55
3	1:tested_negative	1:tested_negative	0.95
4	1:tested_negative	1:tested_negative	1
5	1:tested_negative	1:tested_negative	0.6
6	1:tested_negative	1:tested_negative	0.81
7	1:tested_negative	2:tested_positive	+0.83
8	1:tested_negative	2:tested_positive	+0.65

Multi-Layer perception bestows great advantages as it is used for pattern classification, recognition, prediction and approximation. In Table **7**.

Table 7: Prediction using Multilayer Perception

inst#	actual	predicted	error prediction
1	1:tested_negative	1:tested_negative	0.692
2	1:tested_negative	1:tested_negative	0.583
3	1:tested_negative	1:tested_negative	0.942
4	1:tested_negative	1:tested_negative	0.954
5	1:tested_negative	1:tested_negative	0.945
6	1:tested_negative	1:tested_negative	0.894
7	1:tested_negative	2:tested_positive	+80.85
8	1:tested_negative	2:tested_positive	+0.55

In Figure **3**, a network of different layers namely input layer, hidden layer and output layer consisting of

input nodes (green) or “neurons”, output nodes (yellow) and some hidden nodes (red) some of them are visible.

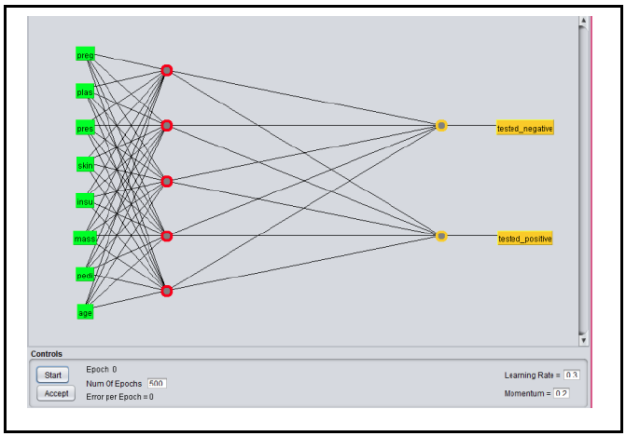

Figure 3: Neural network (MLP).

The nodes in the network are all sigmoid. Each connected network has some value in it which will be pass on to other nodes and each nodes perform a weighted sum of its input and pass it on until it generate some results. Hidden layer depends upon the complexity of the data [8].

MLP does show result with minimum error rate but it processes slow as compared to others.

J.48

Jr8 is basically an implementation of C4.5 algorithm [9]. J48 decision tree decides which attributes is the most decisive one and which one is least and over and then these attributes further divided into sub tree. It generates a binary tree, unlike Random Forest decision tree it use the concept of entropy, difference in entropy gives us the attribute which is free to make decisions.

Table 8: Prediction using J.48

inst#	actual	predicted	error prediction
1	1:tested_negative	1:tested_negative	0.982
2	2:tested_positive	2:tested_positive	0.635
3	1:tested_negative	2:tested_positive	+0.635
4	1:tested_negative	1:tested_negative	0.867
5	2:tested_positive	1:tested_negative	+0.9
6	1:tested_negative	1:tested_negative	0.867

VI. CLASSIFICATION RESULTS

Positive = identified

Negative = rejected.

Therefore:

TP=True positive = correctly identified.

FP=False positive = incorrectly identified.

TN=True negative = correctly rejected.

FN=False negative = incorrectly rejected [6].

Accuracy= (TP + TN) / (TP + FP + TN + FN) [8].

Table 9: Comparison of Accuracy between Different Algorithms

S. No	Classification Type	Accuracy (%)
1	Naïve Bayes	76.3 %
2	MLP	81.8182%
3	J.48	75.3%
4	ZeroR	67.5%
5.	Random Forest	79.2%
6.	Regression	76.8%
7.	Logistic Regression	79.2%

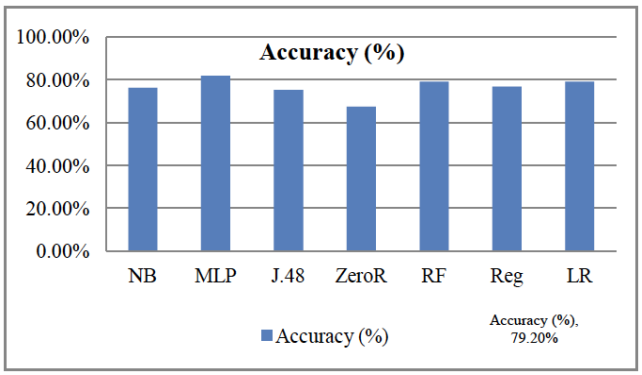

Figure 4: Graphical representation of Accuracy over different algorithm.

VII. CONFUSION MATRIX

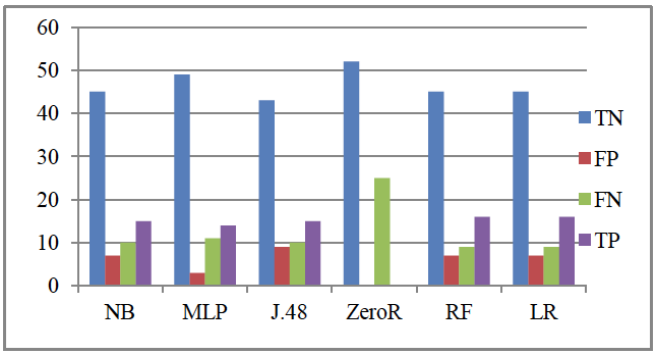

Figure 5: Graphical representation of Confusion Matrix over different algorithm.

VIII. ERROR

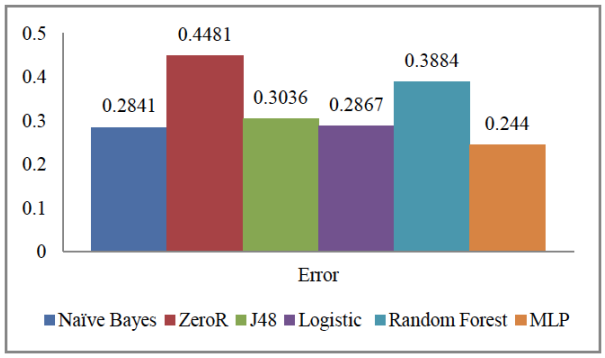

Figure 6: Graphical representation of absolute error over different algorithm.

IX. CONCLUSION AND FUTURE SCOPE

In order to make effective and efficient results, the requirement is to work on different algorithm and to make sure which suits best. Diagnosing diabetes through data mining tool over medical records of patients though it has been done by a majority of the researchers [10-15] but the research demands more deep digging in terms of domain knowledge to get more operative medical diagnosis.

In terms of performance, it was found that multi layer perception function is most effective hence it shows fewer errors however it takes too much processing time because it requires calculation of weights of each node. ZeroR is useful to determine baseline performance for others classification method. Naïve Bayes is also very efficient as it gives a predominant result after each validation but its performance is not quit impressive. J4.8 gives a graphical image of the precedence of the attribute as it calculates the priority of each attribute with other and yet it also predicts accurate results with least error hence it requires time.

The objective of comparing the algorithm on the same dataset, analyzing and predicting the results out of it has been achieved. In future, authors are interested in gathering information among our own neighborhood and authors were keen to get new results which lead them toward more precise and accurate divination. Also more parameters can be added (such as thirst, fatigue, frequency of urination etc) for improvement.

REFERENCES

[1] World Health Organization, Diabetes Programm. http://www.who.int/diabetes/en/

[2] Machine Learning Group at the University of Waikato. Weka 3: Data Mining Software in Java. Retrieved September 4, 2016, from http://www.cs.waikato.ac.nz/ml/weka/

[3] Sanofi, Diabetes Pakistan, Statistics. http://www.sanofidiabetes.com.pk/web/about_diabetes/statistics

[4] The News International. https://www.thenews.com.pk/print/73051-seven-million-pakistanis-suffering-from-type-2-diabetes

[5] Pima Indians Diabetes Data Set. https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

[6] Logistic Regression. https://en.wikipedia.org/wiki/Logistic_regression

[7] Singh S, Kaur K. A Review on Diagnosis of Diabetes in Data Mining. International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value 2013; 6.14 | Impact Factor (2013): 4.438.

[8] Witten IH. Department of Computer Science University of Waikato New Zealand, Simple neural networks", "More Data Mining with Weka. More Data Mining with Weka, Simple Neuarl Network, https://drive.google.com/file/d/0B-f7ZbfsS9-xcEFUZ095UUpnVlU/edit 5

[9] Sathees Kumar B, Gayathri P. Department of Computer Science, Bishop Heber College, Analysis of Adult-Onset Diabetes Using Data Mining Classification Algorithms, International Journal of Modern Computer Science (IJMCS) ISSN: 2320-7868 (Online) Volume No.-2, Issue No.-3, June, 2014 Conference proceeding.

[10] Radha P, Srinivasan B. Predicting Diabetes by cosequencing the various Data Mining Classification Techniques. IJISET - International Journal of Innovative Science, Engineering & Technology 2014; 1(6).

[11] Iyer A, Jeyalatha S, Sumbaly R. Diagnosis of Diabetes Using Classification Mining Techniques. International Journal of Data Mining & Knowledge Management Process (IJDKP) 2015; 5(1):

[12] Satyanandam N, Satyanarayana Ch, Riyazuddin Md, Amjan S. Data Mining Machine Learning Approaches and Medical Diagnose Systems. International Journal of Computer & Organization Trends 2012; 2(3).

[13] Sa'di S, Maleki A, Hashemi R, Panbechi Z, Chalabi K. Comparison Of Data Mining Algorithms In The Diagnosis Of Type II diabetes. International Journal on Computational Science & Applications (IJCSA) 2015; 5(5).

[14] Ezaz Ahmed D, Mathur YK, Kumar V. Knowledge Discovery in Health Care Datasets Using Data Mining Tools. (IJACSA) International Journal of Advanced Computer Science and Applications 2012; 3(4): 117.

[15] Daghistani T, Alshammari R. Diagnosis of Diabetes by Applying Data Mining Classification TechniquesComparison ofThree Data Mining Algorithms. IJACSA) International Journal of Advanced Computer Science and Applications 2016; 7(7): 16.

Received on 03-08-2017 Accepted on 11-08-2017 Published on 29-08-2017

https://doi.org/10.6000/1927-5129.2017.13.77