Volume 9, Issue 2
April 2020

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X



# Data Mining Approaches to Diagnose Diabetes Using Clinical Dataset

**[1]Md. Toukir Ahmed, [2]Md. Niaz Imtiaz and [3]Eshita Nahar Lahm**

[1,2,3] Department of Computer Science and Engineering,
Pabna University of Science and Technology, Pabna, Bangladesh

E-mail: [1]toukirahmedreal@gmail.com, [2]imtiaz.cse.buet@gmail.com, [3]nahar.eshita@gmail.com

## ABSTRACT

Nowadays, diabetes have become a common disease which is affecting many people all over the world. When human body fails to produce enough insulin to meet their body's need, blood sugar level increases from normal to higher and causes diabetes. Diabetes causes blindness, kidney failure, heart failure, brain strokes and so on. Data Mining and machine learning plays an important role in the field of healthcare, as disease diagnosis and analysis have huge size of data. In this study various data mining techniques are used on the dataset including Bayesian, Logistic Regression, Neural network, Naïve Bayes, CART and AdaBoost. A comparative analysis is done among the above mentioned classifiers. Assessing the usual performance metrices on the dataset it was noticeable that, Neural Network gives the best accuracy.

**Keywords:** *Data mining, Diabetes, Machine Learning, Prediction Models, Performance metrices.*

## 1. INTRODUCTION

Machine learning is an use of artificial intelligence. The goal of machine learning is to make the computers learn automatically without human intervention and it adjust it's actions according to it. This analytical process is being completed by using machine learning algorithm for analyzing medical data in many health care. Machine learning algorithms are mainly classified as being supervised or unsupervised (www). Diabetes is a fast growing diseases among the people even among the youngster in the whole world. Diabetes is caused by increasing of blood sugar level(high blood glucose). There are 3 major diabetes types can develop: type 1, type 2, and gestational diabetes (Chandan Kumar, 2019). Diabetes usually mean Diabetes Mellitus (DM). People with DM are called "diabetics" (htt). Mainly four major types of diabetes- Type 1, Type 2, Gestational diabetes. Type 1 diabetes is an autoimmune disease. In this case, the body destroys the cells that are essential to produce insulin to absorb the sugar to produce energy. Type 2 diabetes usually affects the adults who are obese. In this type, the body resist observing insulin or fails to produce insulin. Type 2 generally occurs in the middle or aged group (Chandan Kumar, 2019). One of the main reasons for type-2 diabetes is Obesity. The type-2 diabetes can be controlled by doing proper exercise and taking appropriate diet. If the glucose level is not reduced by the above methods then medicines can be prescribed. National Diabetes Statistics Report 2014 says that 29.1 million people or 9.3% of the U.S. population have diabetes (Umatejaswi, 2017).

Data mining is the process of extracting valuable knowledge from huge dataset and it has played an important role in health care domain. Data mining techniques would be a valuable asset for diabetes researchers because it can expose hidden knowledge from a huge amount of diabetes related data (B. Senthil Kumar 1, 2016). The algorithms have been applied to the PIMA Indians Diabetes Dataset of National Institute of Diabetes and Digestive and Kidney Diseases that contains the various kind of data (S.Subashree, 2019).

## 2. LITERATURE SURVEY

The objective of the research is "Prediction of Diabetes by consequence the various Data Mining Classification Techniques" describes the various Data Mining Classification Techniques. There are many classification techniques used in this paper for predicting diabetes (A. Ayer, 2015). Research paper, "Disease Prediction in Data Mining Technique"– A Survey. The disease prediction plays an important role in data mining. This paper analyses about various diseases like Heart disease prediction, Breast cancer prediction, Diabetes by using many techniques like Classification, Decision Tree (CART), Naive Bayes, SVM etc. methods in order to predict the diabetes disease. This paper also tells about predictive and descriptive type about the data. Prediction involves some fields in the data set to predict the values of other variables. The different algorithm of data mining is used in the field of medical prediction are discussed in this paper (Kannan, 2011).

The diabetes prediction system is being presented in this section for diabetes diagnosis. There are five steps. Initially we need diabetes data-set/clinical data. The diabetes data-set is given to the data pre-processing module. It removes the unrelated feature from the dataset, then it goes to the machine learning and data mining algorithm with related feature. After that, machine learning algorithm develops a prediction model from the pre-processed data-set. It is also known as knowledge model. Furthermore, the diabetes is predicted for a person's medical report or data using the knowledge model (Chandan Kumar, 2019).

Volume 9, Issue 2
April 2020

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X



## 3. METHODOLOGY

### 3.1 Data-Set

There are 9 attributes and 768 number of instances in our data-set. The data-set is based on Pima Indian Diabetic set from University of California, Irvine Repository of machine learning databases (Lichman, 2013).

**Table-1: Attributes of the dataset**

| No | Attribute Name |
|---|---|
| 1 | Number of times pregnant |
| 2 | Glucose tolerance test |
| 3 | Blood pressure |
| 4 | Triceps skin fold thickness(mm) |
| 5 | 2 hour serum insulin in mu/ml |
| 6 | Body mass index |
| 7 | Diabetes pedigree function |
| 8 | Age in years |
| 9 | Class/Outcome(1-positive,0-negative) |

Class variable takes the values 0 or 1, where 1 means tested positive for diabetes and 0 means tested negative for diabetes. The Pima Indian diabetes dataset is widely used for testing classification algorithm. There are total 768 samples are there in Pima Indian diabetes dataset, 268 samples are diabetes positive and 500 samples are diabetes negative.

### 3.2 Prediction Model

There are various methods through which make prediction about future or unknown events by analyzing current or historical facts. There are various data mining and machine learning techniques through which one can make a prediction model. We have use the models are - Logistic Regression, Naïve Bayes, CART, Neural Network, AdaBoost. Using these methods we predict the diabetes and accuracy comparison.

#### 3.2.1 Naïve Bayes

Naïve bayes are a collection of classification algorithm based on Bayes theorem. Naïve bayes is not a single algorithm. It is a collection of algorithm where all of them share a common principle, i.e. every pair of features is independent classified of each other. Naïve bayes is a powerful algorithm for predictive modeling (Chandan Kumar, 2019). 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:

$$P(c|x)=\frac{P(X|C)P(c)}{P(x)} \quad (1)$$

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \ldots\ldots \times P(x_n|c) \times P(c) \quad (2)$$

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 (Chandan Kumar, 2019).

#### 3.2.2 Logistic regression

Logistic Regression algorithm can be used as a linear equation with independent predictors to predict a value. The predicted value can be anywhere between negative infinity to positive infinity. We need the output of the algorithm to be class variable, i.e 0-no, 1-yes. Therefore, we are squashing the output of the linear equation into a range of [0,1]. To squash the predicted value between 0 and 1, we use the sigmoid function (Def).

$$Z=\theta_0+\theta_1 x_1 + \theta_2 x_2 + \cdots \quad (3)$$

$$g(x)=\frac{1}{1+e^{-x}} \quad (4)$$

$$h=g(z)=\frac{1}{1+e^{-z}} \quad (5)$$

These are the equations of sigmoid Function.

#### 3.2.3 CART Algorithm

Classification and regression trees (CART (D.Senthil Kumar1, May 2011 )) is a nonparametric technique that produces either classification or regression trees, depending on whether the dependent variable is categorical or numeric, respectively. Trees are formed by a collection of rules based on values of certain variables in the modelling data set. Rules are selected based on how well splits based on variables' values can differentiate observations based on the dependent variable Once a rule is selected and splits a node into two, the same logic is applied to each "child" node (i.e. it is a recursive procedure). Splitting stops when CART detects no further gain can be made, or some pre-set stopping rules are met. The basic idea of tree growing is to choose a split among all the possible splits at each node so that the resulting child nodes are the "purest". In this algorithm, only univariate splits are considered. That is, each split depends on the value of only one predictor variable. All possible splits consist of possible splits of each predictor. CART innovations include:

a. solving the "how big to grow the tree"- problem;
b. using strictly two-way (binary) splitting;
c. incorporating automatic testing and tree validation, and;
d. Providing a completely new method for handling missing values.

#### 3.2.4 Neural Network

Volume 9, Issue 2
April 2020

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X



Neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns. They interpret sensory data through a kind of machine perception, labeling or clustering raw input. The patterns they recognize are numerical, contained in vectors, into which all real-world data, be it images, sound, text or time series, must be translated (A.I. Wiki, A Beginner's Guide to Important Topics in AI, Machine Learning, and Deep Learning, n.d.).

### 3.2.5 AdaBoost Algorithm

Boosting is an ensemble technique that attempts to create a strong classifier from a number of weak classifiers. AdaBoost was the first really successful boosting algorithm developed for binary classification. It is the best starting point for understanding boosting. AdaBoost is best used to boost the performance of decision trees on binary classification problems. AdaBoost was originally called AdaBoost.M1 by the authors of the technique Freund and Schapire. More recently it may be referred to as discrete AdaBoost because it is used for classification rather than regression. AdaBoost can be used to boost the performance of any machine learning algorithm. It is best used with weak learners. These are models that achieve accuracy just above random chance on a classification problem. The most suited and therefore most common algorithm used with AdaBoost are decision trees with one level. Because these trees are so short and only contain one decision for classification, they are often called decision stumps. Each instance in the training dataset is weighted. The initial weight is set to:

Weight (xi) = 1/n

Where xi is the i'th training instance and n is the number of training instances.

## 4. RESULTS AND DISCUSSIONS

The proposed model is validated using four parameters namely the Accuracy of the classifier, Area Under ROC Curve, Precision and Sensitivity(Recall) (Dr. B. Sarojini, 2011).

Precision = TP/ (TP + FP).

Recall = TP/ (TP + FN).

Accuracy = (TP + TN)/ (TP + TN + FP + FN).

TP - Positive tuples.

TN - Negative tuples.

FP - Incorrectly classified positive tuples.

**Table-2: Accuracy, Precision, Recall and ROC values of different Models**

The accuracy of these 5 different models can also be very clear by giving a bar chart of accuracy score. It gives a clear vision of difference and comparison among these algorithms. In bar diagram X-axis holds names of different

classifiers and Y-axis represents the accuracy score. So that we can see a clear view of accuracy comparison bar diagram. By this type of figure we can say which classifier shows the best result and which classifier is representing the lowest accuracy score. It is the graphical representation we can see neural network gives the highest accuracy score 81% than other used algorithm. It accuracy level is acquired by using 768 numbers of instances. Graphically we can see that comparison.

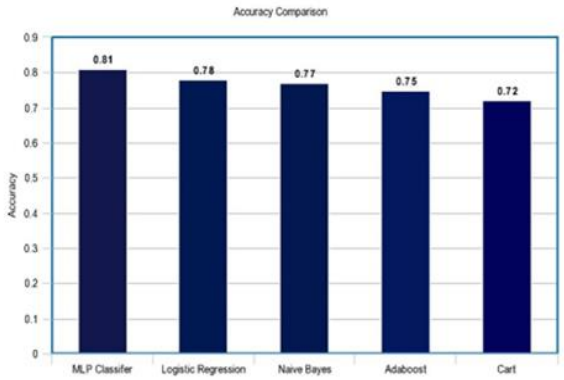


**Fig 1: Accuracy Score graph of different models**

Here, both Fig 2 shows different ROC curve. Y-axis represent the true positive rate. It is the ability which is used to find the high true-positive rate. The true-positive rate is also called as sensitivity. And X-axis represents False positive rate. the false negative rate is the proportion of positives which yield negative test outcomes with the test, i.e., the conditional probability of a negative test result given that the condition being looked for is present. ROC shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).

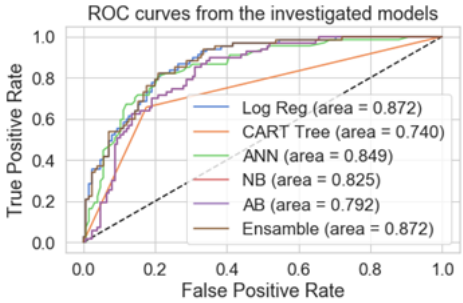


**Fig 2: ROC Area Rate for different models**

The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. The area under the ROC curve ( AUC ) is a measure of how well a parameter can distinguish between two diagnostic groups. The AUC for the ROC can be calculated using the 'roc_auc_score()' function. Like the 'roc_curve()' function, the AUC function takes both the true outcomes (0,1) from the test set and the predicted probabilities for the 1 class. It returns the AUC score between 0.0 and 1.0 for no skill and perfect skill respectively (Ambika Rani Subhash, May 2019).

## 5. CONCLUSION AND FUTURE WORK

The most significant aspect of this study is to apply data mining technology for predicting diabetes. We performed a pre-processing step to deal with dataset like feature selection

method, normalization and assessed machine-learning technique such as neural network. Feature selection reduces the no of dimensions by selecting most informative features based on some statistical score. F-score gives better performance of classification. Then performance of different classifiers are evaluated in term of accuracy, precision, recall and AUC (Khyati K. Gandhi, 2014). In future, We can extend our work by developing accuracy applying meta heuristics and performing principal component analysis.

## REFERENCES

[1] R. R. J. MinyechilAlehegn, "Type II Diabetes Prediction Using Combo of SVM," International Journal of Engineering and Advanced Technology (IJEAT), vol. 8, no. 6, p. ISSN: 2249 – 8958, August 2019

[2] N. S. J. S. Chandan Kumar, "Prediction of Diabetes using Data Mining Algorithm," International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 7, no. 11, pp. 2321-9653, 2019.

[3] N. J. Md. Aminul Islam, "Prediction of Onset Diabetes using Machine Learning Techniques," International Journal of Computer Applications, December, 2017.

[4] P. S. K. a. V. Umatejaswi, "Diagnosing Diabetes using Data Mining Techniques," International Journal of Scientific and Research Publications, vol. 7, no. 6, pp. ISSN 2250-3153, 2017.

[5] D. R. G. B. Senthil Kumar 1, " A Survey on Data Mining Approaches to Diabetes Disease Diagnosis and Prognosis," International Journal of Advanced Research in Computer and Communication Engineering , vol. 5, no. 12, p. 3297:2007, 2016.

[6] S. S.Subashree, "ANALYSIS AND PREDICTION OF DIABETES USING MACHINE LEARNING," International Journal of Emerging Technology and Innovative Engineering , vol. 5, no. 4, p. ISSN: 2394 – 6598, 2019.

[7] J. S. a. R. S. A. Ayer, "Diagnosis of Diabetes Using Classification Mining Techniques," IJDKP, vol. 5, no. 1, pp. 1-14, 2015.

[8] Y. a. M. Kannan, "Analysis of a Population of Diabetic Patients Databases in Waikato," International Journal of Scientific & Engineering Research, vol. 5, no. 2, 2011.

[9] M. Lichman, "UCI Machine Learning Repository," Irvine CA: University of California, School of Information and Computer Science , 2013.

[10] D. K. A. K. A. B. Thiyagarajan C, "A Survey on Diabetes Mellitus Prediction," International Journal of

Applied Engineering Research, vol. 11, no. 3, pp. ISSN 0973-4562, 2016.

[11] G. a. S. D.Senthil Kumar1, "Decision Support System for Medical Diagnosis Using Data Mining," IJCSI International Journal of Computer Science Issues, vol. 8, no. 3, pp. ISSN (Online): 1694-0814 , May 2011 .

[12] C. G. S. J. S. Jatin N Bagrecha, "Diabetes Disease Prediction using Neural Network," International Journal for Research in Applied Science & Engineering Technology, vol. 7, no. 4, pp. ISSN: 2321-9653, April-2019.

[13] P. N. B. Khyati K. Gandhi, "Diabetes prediction using feature selection and classification," International Journal of Advance Engineering and Research Development (IJAERD) , vol. 1, no. 5, pp. e-ISSN: 2348 - 4470 , 2014.

[14] D. N. R. Dr. B. Sarojini, "Enhancing Medical Prediction using Feature Selection (IJAE)," vol. 1, no. 3, 2011.

[15] A. k. U. Ambika Rani Subhash, "Accuracy of Classification Algorithms for Diabetes prediction," International Journal of Engineering and Advanced Technology (IJEAT), vol. 8, no. 5s, p. ISSN: 2249 – 8958, May 2019.

## Authors Biography

**Md. Toukir Ahmed** received his B.Sc. Engineering in Computer Science and Engineering from Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh in 2017, Bangladesh. He is currently working as a Lecturer in the Department of Computer Science and Engineering, Pabna University of Science and Technology, Pabna, Bangladesh. His research interest includes Data Mining, HCI, Big Data Analysis and Machine Learning.

**Md. Niaz Imtiaz** received his B.Sc. Engineering in Computer Science and Engineering from Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh in 2011, Bangladesh. He is currently working as Assistant Professor in the Department of Computer Science and Engineering, Pabna University of Science and Technology, Pabna, Bangladesh. His research interest includes Data Mining, Network Security, Big Data Analysis and Machine Learning.

**Eshita Nahar Lahm** is pursuing her B.Sc. Engineering in Computer Science and Engineering from Pabna University of Science and Technology, Pabna, Bangladesh. Her research interest includes Data Mining, HCI, Big Data Analysis and Machine Learning.