# Prediction of Onset Diabetes using Machine Learning Techniques

Md. Aminul Islam
Assistant Professor
Bangladesh University of Health Sciences

Nusrat Jahan
Lecturer
Daffodil International University, Bangladesh

## ABSTRACT
Machine learning algorithms can help us to detect the onset diabetes. Early detection of diabetes can reduce patient's health risk. Physicians, patients, and patient's relatives can be benefited from the prediction's outcomes. In low resource clinical settings, it is necessary to predict the patient's condition after the admission to allocate resources appropriately. Several articles have been published analyzing Prima Indian data set applying on various machine learning algorithms. Shankar applied neural networks to predict the onset of diabetes mellitus on Prima Indian Diabetes dataset and showed that his approach for such classification is reliable [4, 5 and 6]. Machine learning techniques increase medical diagnosis accuracy and reduce medical cost [2, 3]. In this study, the main focus is to investigate different types of machine learning classification algorithms and show their comparative analysis. The purpose of this study is to detect the diabetic patient's onset from the outcomes generated by machine learning classification algorithms.

## Keywords
Machine Learning, SVM, Naive Bayes, Logistic Regression, J48, OneR.

## 1. INTRODUCTION
Diabetes is a chronic illness which can be caused by body's inability to produce, or when body cannot use the insulin that it produces [1]. The effects of diabetes mellitus include long–term damage, dysfunction and failure of various organs (WHO). As a result, it has significantly increased mortality in patients. There are mainly two types of diabetes: Type I (T1) and Type II (T2). T1 occurs when the body is no longer able to produce insulin whereas T1 is common in childhood and also known as juvenile diabetes. This form of diabetes is less common; only about 5-10% of people with diabetes have T1 (American Diabetes Association, 2010). T2 occurs when the body is unable to utilize the insulin produced or not enough insulin is produced [9, 10 and 11].

In addition, there is another type of diabetes named gestational diabetes which develops during pregnancy. Too much glucose in blood can damage eyes, kidneys, and nerves. It can also cause of heart disease, stroke, and insufficiency in blood flow to legs. Overweight, lack of exercise, family history and stress increased the possible risk of diabetes 14, 15].

In Bangladesh, people are not conscious about health. There are 7.1 million case of Diabetes in Bangladesh. The increasing level of Diabetes is up bound. People do not know about it and they do not go to check it.

## 2. REVIEW WORK
Diabetes has affected over 246 million people worldwide with a majority of them being women. According to the WHO report, by 2025 this number will expect to rise over 380 million [23].

In one of the research paper Mukesh kumari and et el. predicting diabetes using data mining techniques [21]. Veena Vijayan V. and Aswathy Ravikuma worked with data mining algorithms for prediction and diagnosis of diabetes mellitus [22]. This paper discussed about the data mining techniques to predict diabetes risk.

### 2.1 Naive Bayes (NB)
Naive Bayes classifiers assume attributes have independent distributions. It is considered to be fast and space efficient. It also provides simple approach, with clear semantics, representing and learning probabilistic knowledge [17]. It is known as Naive because it relies on two important simplifying assumptions. The predictive attributes are conditionally independent and secondly it assumes that no hidden attributes bias the prediction process. It is very fast to train and fast to classify.

### 2.2 Logistic Regression (LR)
Logistic regression is a type of probabilistic statistical classification model for analyzing a dataset in which there are one or more independent variables that determine an outcome. In logistic regression, the dependent variable is binary or dichotomous, that means it only contains data coded as 1 (TRUE, success, pregnant, etc.) or 0 (FALSE, failure, non-pregnant, etc.). Logistic regression generates the coefficients of a formula to predict a log it transformation of the probability of presence of the characteristic of interest.

### 2.3 Multilayer Perceptron (MlP)
MLP is one of the most widely used neural network classification algorithms [20]. This classifier uses back propagation to classify instances. The main problem with this algorithm is that prediction given by MLP is difficult to understand and explain by human being. MLP used in this experiment consisted of four layers: one input, 2 hidden layers, and one output layer.

### 2.4 Support Vector Machine
A support vector machine (SVM) is a machine learning algorithm that constructs a hyper plane or set of hyper planes in a high dimensional space, which can be used for classification. It is a group of supervised learning methods and good separation can be achieved by the hyper plane that has the largest distance to the nearest training data points of any class [18]. Sometimes it happen that sets are not linearly separable. Kernel function improving the SVM and solve dimensional and over fitting problem [7, 8].

### 2.5 IBK
In Weka, nearest neighbor classification algorithm is known as IBK (the IB stands for Instance-Based, and the K allows us

to specify the number of neighbors to examine). IBK is a useful data mining technique that allows us to use past data instances with known output values to predict an unknown output value of a new data instance. It predicts very accurate but often perform slow, generally perform well for large value ok K.

## 2.6 Adaboostm1 and Bagging

AdaBoost (Adaptive Boosting) is a Boosting machine learning meta-algorithm which theoretically can be used to significantly reduce the error of any learning algorithm that consistently generates classifiers whose performance is a little better than random guessing. It is a nominal class classifier using the Adaboost M1 ensemble method which means only nominal class problems can be solved. It is Often dramatically improves performance, but sometimes over fits. On the other hand, Bagging is an ensemble method that creates separate samples of the training dataset and creates a classifier for each sample. It reduces the variance [13].

## 2.7 OneR

OneR, short for "One Rule", is a simple classification algorithm that generates a one- level decision tree. OneR is able to infer typically simple, yet accurate, classification rules from a set of instances. This is for building and using a 1R classifier; in other words, uses the minimum-error attribute for prediction, discretizing numeric attributes. It is well known as a simple classification rules perform well on most commonly used datasets.

## 2.8 Decision Tree (J48) and Random Forest

Decision tree algorithm initially defined as C4.5 algorithm, Weka classifiers packages has its own version of it known as J48. J48 is an optimized implementation of C4.5. Random Forest consists of many decision trees and the method combines bagging and the random selection of features idea both together. It is one of the best learning algorithms available in machine learning and produces a highly accurate classifier. It can handle huge amount data and run efficiently without variable deletion. Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. Significant improvements in classification accuracy have resulted from growing an ensemble of trees and letting them vote for the most popular class. In order to grow these ensembles, often random vectors are generated that govern the growth of each tree in the ensemble [19]. After a large number of trees are generated, this algorithm has voting mechanism for selecting the most popular class after generating a large number of trees.

Above all these classification algorithms are considered to be among the popular machine learning algorithms. All classifiers were tested in this study. Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), Area Under Curve (AUC) and Accuracy were the determinant factors for the usefulness of the particular tested algorithms. Data has been analyzed and pre-processed to achieve the good accuracy.

# 3. MATERIALS AND METHODS

This datasets have been collected among the Pima Indian female population near Phoenix, Arizona. This particular dataset has been widely used in machine learning experiments and is currently available through the UCI repository of standard datasets. This population has been studied continuously by the National Institute of Diabetes, Digestive and Kidney. UCI repository contains 768 instances of observations and total 9 attributes with no missing values reported. Data sets contains 8 particular variables which were considered high risk factors for the occurrence of diabetes, like number of times pregnant, plasma glucose concentration at 2 hour in an oral glucose tolerance test (OGTT), diastolic blood pressure, 2 hour serum insulin, body mass index, diabetes pedigree. All the patients in this datasets are female at least 21 years old living near Phoenix, Arizona. All attributes are numeric values except class is nominal type. Attributes name and types are shown in table 1.

**Table 1. The Prima Indian datasets attributes**

| No | Name of attributes | Type |
|---|---|---|
| 1 | Number of times pregnant | Numeric |
| 2 | Plasma glucose concentration a 2 hours in an oral glucose tolerance test | Numeric |
| 3 | Diastolic blood pressure(mm Hg) | Numeric |
| 4 | Triceps skin fold thickness(mm) | Numeric |
| 5 | 2 hour serum insulin(mu U/ml) | Numeric |
| 6 | Body mass index(weight in $Kg/(height\ in\ m)^2$) | Numeric |
| 7 | Diabetes pedigree function | Numeric |
| 8 | Age(years) | Numeric |

For the prediction of outcome, the patient was classified into one of the two categories:

1. Tested Positive( class 1)
2. Tested Negative(class 2)

The patient who has been marked as tested positive means has higher risk of disease whereas patient who was tested negative have lower risk of diabetic. There are 268 instances of class 1 and 500 of class 2. Weka machine learning tool has been used to classify the data in this study. WEKA (Waikato Environment for Knowledge Analysis) is developed at University of Waikato, New Zealand. It is a collection of machine learning algorithms for data mining tasks written in Java and freely available under the GNU general public License [16]. It also contains tools for data pre- processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Weka Explorer was used in default configuration settings. Weka version 3.8 was used in this study.

Data has been analyzed carefully and found out that several attributes has impossible values, such as 0 body mass indexes, and 0 plasma glucose. There are 236 observations that have at least one impossible value for attribute of body mass index, blood pressure, glucose, and triceps skin thickness. There is no missing value attributes found from 768 observations [12]. To keep the datasets reasonably large, only observations are deleted those contain impossible value for multiple attributes. After cleaning the data we have 755 observations shown in table 2.

**Table 2. Number of observations after cleaning attributes for multiple impossible values 0**

| No | Label | Count |
|---|---|---|
| 1 | tested_negative | 491 |
| 2 | tested_positive | 264 |

10 folds cross validation has been applied as a test option because this option perform better for over-fitting problem and small datasets. In 10 folds cross validation, each data point used 1 time for testing and 9 times for training, repeat it 10 times. Weka does stratified cross validation by default. With 10 fold cross validation Weka invokes the learning algorithm 11 times. There is a rule of thumb, if we have lots of data normally percentage splits is good but for small and medium data size cross validation performs good results. It reduces the variance of estimate. Different classification techniques were applied in this study using Weka. 10 classification algorithms from 6 different categories were selected to compare the prediction outcomes. The selected machine learning techniques are described below:

1. Bayes: Naïve Bayes (NB)
2. Function: Logistic Regression(LR), Multilayer perception(ML), Support Vector Machine(SVM)
3. Lazy: IBK
4. Meta: AdaBoostM1, Bagging.
5. Rules: OneR
6. Trees: J48, Random Forrest

## 4. CLASSIFICATION OUTCOMES

In this study, the performance of different classification methods tested by Weka is shown in Table 3. With the default configuration, logistic regression had the highest accuracy (78.01%) and AUC (0.833) whereas IBK obtained lowest accuracy (70.99%) and simple classification algorithm OneR obtained the lowest AUC (0.642). SVM, MLP, NaiveBayes and Bagging classifiers performed good from the rest of the classifiers. Total accuracy is above 70% in all cases. Baseline accuracy found 65% using rules classifiers ZeroR applying test option use training set. Generally, any result above baseline is considered to be good.

The maximum AUC obtained from LR classifiers but overall AUC is quite high and satisfactory. Only IBK and OneR obtained low AUC compare to others. All the ROC curves are shown in Figure 1.

The second highest accuracy (77.08%) obtained from SVM classifiers but obtained comparatively lower AUC (0.716). Sensitivity and specificity are quite satisfactory. Thus, the SVM classifier appears to perform second best among all 10 classifiers. On the other hand, MLP obtained third highest accuracy (76.82%), sensitivity (0.81), PPV (0.85), and AUC (0.817) respectively.

**Table 3. Different machine learning classifiers outcomes**

| Category | Classifier Name | Sensitivity | Specificity | PPV | NPV | AUC | Total Accuracy |
|---|---|---|---|---|---|---|---|
| Bayes | NaiveBayes | 0.80 | 0.67 | 0.84 | 0.61 | 0.815 | 75.76 |
| Function | LR | 0.80 | 0.74 | 0.89 | 0.58 | 0.833 | 78.01 |
| | MLP | 0.81 | 0.68 | 0.85 | 0.63 | 0.817 | 76.82 |
| | SVM | 0.78 | 0.74 | 0.90 | 0.53 | 0.716 | 77.08 |
| Lazy | IBK | 0.77 | 0.59 | 0.79 | 0.55 | 0.668 | 70.99 |
| Meta | AdaBoostM1 | 0.78 | 0.66 | 0.85 | 0.56 | 0.805 | 74.70 |
| | Bagging | 0.79 | 0.66 | 0.84 | 0.59 | 0.822 | 75.09 |
| Rules | OneR | 0.73 | 0.64 | 0.88 | 0.40 | 0.642 | 71.25 |
| Trees | J48 | 0.81 | 0.63 | 0.79 | 0.66 | 0.763 | 74.30 |
| | Random Forest | 0.78 | 0.67 | 0.86 | 0.54 | 0.808 | 74.83 |

## 5. RESULT DISCUSSION

In medical context, sensitivity is the ability to identify the proportions of actual positives correctly identified, also known as true positive rate whereas specificity is the ability to identify the proportions of the actual negatives are correctly identified, also known as true negative rate. 100% sensitivity and 100% specificity is the perfect predictor but it is almost impossible to achieve such performance in any datasets in reality. Logistic regression along with SMO and MLP could be used to predict the onset diabetes. Other methods like NB, Bagging also performed well but not up to the mark considering all the parameters. 6 out of 10 classifiers AUC outcomes are above than 0.80 which is considered as good result. AUC equals to 1 is the perfect and AUC equals to 0.5 considered random guessing. So, overall AUC is satisfactory in this study.

PPV used to indicate the probability of patients really has specified disease. Almost all cases PPV values are quite high. That means the detection of diabetes patient is quite high from this prediction. On the other hand, almost all cases the NPV value obtained comparatively low. The highest accuracy is still unable to exceed the accuracy of 80% with the best performing algorithm, Logistic Regression. This may be due to limitations of the data. In some cases, such as with IBK and OneR, accuracy drops drastically. ROC curves were generated

based on the predicted outcomes using Weka's "visualize threshold curve" option and is shown in Figure 1. All 10 classifier's ROC curves are shown in Figure 1 to visualize the machine learning prediction performance and compare the outcome one another.

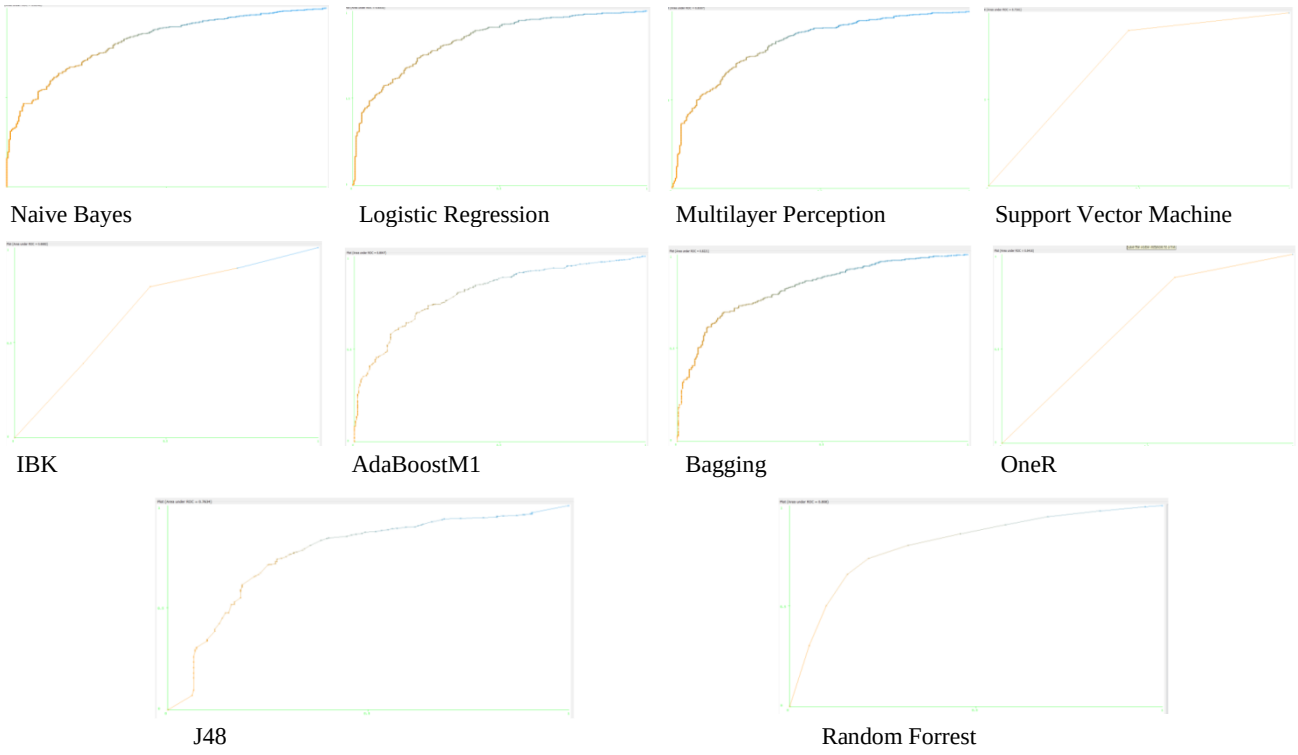


**Fig. 1. ROC curves for classifiers outcomes using Weka**

In this study, authors compared the performance of various algorithms and found that Logistic Regression performed well on the standard. They tried to understand how different classifiers affected outcomes. Moreover, authors paid close attention to data sets and analyzed carefully to understand the medical value. The work here gave me a better understanding of machine learning applications in medical diagnosis. This was also an important lesson on data transforms and algorithm analysis. It is somewhat unfortunate that many medical datasets are small (this may be due to patient confidentiality), since a larger dataset would give us more flexibility and robustness in analysis. However, authors strongly believe that this study is a good start for building methods that help diagnose patients, and bridge the gap between doctors and large datasets.

## 6. CONCLUSION

There is no cure for diabetics but early detection can reduce the long term complications and reduce the cost. Millions of people in the world have diabetes. Many of these people do not even know whether they have it or not. The ability to predict diabetes early plays an important role for the patient's appropriate treatment strategy. However, the correct prediction accuracy of current machine learning algorithms is often low. LR performed the best among all 10 classifiers. It tried to predict whether an individual was diabetes positive or not. Thus, this article applied several machine learning algorithm and analyzed the data for enhancing the diabetes prediction accuracy. Further analysis of attributes and different combination of feature selection is required to achieve better accuracy. The outcomes might help the care process in the low resource settings. It also helps for preventive care of diabetes patients.

## 7. ACKNOWLEDGMENTS

The authors would like to convey their earnest and heartiest thanks to all the teacher of Faculty of Health Engineering & Technology and Allied Health Sciences, Bangladesh University of Health Sciences, for their valuable suggestions and co-operation.

## 8. REFERENCES

[1] World Health Organisation [Internet]. 2013, Available from : http://www.who.int/diabetes/en/

[2] Kayaer K, Yildirim T. Medical Diagnosis on Prima Indian Diabetes Using General Regression Neural Networks. [Internet]. Available from: http://www.yildiz.edu.tr/~tulay/publications/Icann-Iconip2003-2.pdf.

[3] A comparative study on diabetes disease diagnosis using neural networks. Volume 36, Issue 4, May 2009, Pages 8610–8615. ELSEVIER.

[4] Shibendra Pobi and Lawrence O. Hall. Predicting Juvenile Diabetes from Clinical Test Results. 2006 International Joint Conference on Neural Networks, Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada, July 16-21, 2006.

[5] Pradhan M, Sahu RK. Predict the onset of diabetes disease using Artificial Neural Network (ANN). International Journal of Computer Science & Emerging Technologies. 2011; Volume 2. Issue 2.

[6] Shanker MS. Using Neural Networks to Predict the Onset of Diabetes Mellitus. American Chemical Society. 1996 Jan 1; 36: 35-41.

[7] Wu J, Diao YB, Li ML, Fang YP, Ma DC. A Semi-supervised Learning Based Method: Laplacian Support Vector Machine Used in Diabetes Disease Diagnosis. Interdiscip Sci Comput Life Sci. 2009; 1:151-155.

[8] Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case study of diabetes and pre- diabetes. BMC Medical Informatics and Decision Making. 2010; 10:16.

[9] Selvakuberan K, Kayathiri D, Harini B, Devi MI. An efficient feature selection method for classification in Health care system using machine learning Techniques. IEEE. 2011; 223-226.

[10] Shankaracharya, Odedra D, Mallick M, Shukla P, Samanta S, et al. Java-Based Diabetes Type 2 Prediction Tool for Better Diagnosis. Diabetes Technology & Therapeutics. 2012; 14: 251-256.

[11] Temurtas H, Yumusak N, Temustas F.A comparative study on diabetes disease diagnosis using neural networks.2009;36:8610-8615.

[12] Bellazi R, Abu-Hanna A. Data Mining Technologies for Blood Glucose and Diabetes Management. Journal of Diabetes Science and Technology.2009; 3(3): 603-612.

[13] Smith JW, Everhart JE, Dickson WC, Knowler WC, Johannes RS. Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. SCAMC. 1988; 261- 265.

[14] Pobi S. A study of machine learning performance in the prediction of juvenile diabetes from clinical test results[Graduate Thesis].South Florida, University of South Florida;2006[cited 2013 March 21]. Available from: http://scholarcommons.usf.edu/etd/2661/.

[15] Davidson M, Schriger DL, Peters AL. An alternative Approach to the Diagnosis of Diabetes with a Review of the Literature. Diabetes Care. 1995; 18(7): 1065-1071.

[16] Cs.waikato.ac.nz. (2014). Weka 3 - data mining with open source machine learning software in java. [online] Retrieved from: http://www.cs.waikato.ac.nz/ml/weka/ [Accessed: 6 Feb 2016].

[17] G.H John and P. Langley, "Estimating Continuous Distributions in Bayesian Classifiers," Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, San Francisco, 1995,pp.338-345.

[18] Ivanciuc, O.Support Vector Machine [internet].2005. Available from: http://www.support-vector-machines.org/SVM_review.html.

[19] Breiman L. Machine Learning. Editor. Robert E. Schapir. Netherlands: Kluwer Academic Publishers; 2011. P. 5-32. (Random Forests; vol 45).

[20] P. J. Werbos. Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. PhD thesis, Harvard University, 1974.

[21] Mukesh kumari, Dr. Rajan Vohra and Anshul arora, "Prediction of Diabetes Using Bayesian Network" International Journal of Computer Science and Information Technologies(IJCSIT), Vol. 5 (4) , 2014, 5174-5178.

[22] Veena Vijayan V. and 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.

[23] 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.