Diabetes Prediction Comparing Machine Learning Classifiers

Hasibur Rahman
Computer Science and Engineering
East West University
Dhaka, Bangladesh
shanto.ewu99@gmail.com

Arif Hasan Nayon
Computer Science and Engineering
East West University
Dhaka, Bangladesh
arifhasan588@gmail.com

Kazi Mushfiquer Rahman
Computer Science and Engineering
East West University
Dhaka, Bangladesh
mushfique1996@gmail.com

Sujit Kumar Roy
Computer Science and Engineering
East West University
Dhaka, Bangladesh
sujitkumarroyy@gmail.com

Abstract

Diabetes is one the most fatal and lifelong disease in most of the countries all over the world which need to be prevented. Most of the countries are involved in early stage diabetes diagnosis and prevention. Using of machine learning classifier algorithms are the best way to detect diabetes. The main purpose of the study is to compare the performance of the machine learning classification algorithms for diabetes prediction. In this study, we compare machine learning classifiers such as Naive Bayes, Decision Tree, Support Vector Machine, Random Forest and K- Nearest Neighbor in order to detect diabetes mellitus patients. The pima native American dataset was chosen for prediction of diabetes and find out the accuracy score, precision, true positive rate, true negative rate, false positive rate and false negative rate using the classification algorithms. After comparing the algorithms, we figure out that support vector machine gives the best performance as the accuracy and the precision is maximum and the true negative rate is minimum.

Keywords- Diabetes, prediction, machine learning, classification

I Introduction

There are many uncured diseases in the medical sector which should be predicted in early stage and the classification algorithms of data mining and machine learning can be applied here. Diabetes is also an uncured disease which need to be predict in early stage. Diabetes causes when the sugar substance of human's body cannot be controlled. Diabetes can be two types, 1) Low diabetes and 2) high diabetes. The cause of the diabetes can be genetic issues, family history, ethnicity and environmental setup. To support and recognize the diabetes, data mining and machine learning classification algorithms can play a vital role so that medical experts can be assisted. In this paper, we tried to have the maximum accurate prediction so that the patients can be well treated.

II Background

Diabetics cannot be cured if a human body is suffered by it ones. Diabetics is a serious disease. Because it can cause strokes, heart disease, blindness, kidney failure even death. Diabetics patients lose weight, power of eyes, kidney disease, infections and so on. So, diabetics is a disease which is the mother of a lot of diseases. It happens when the sugar level of the human body becomes high or

low. Diabetes can be categorized mainly into three types. They are Type 1, Type 2 and Gestational Diabetes Type 1 causes because of the lack of insulins. Type 2 causes the inside of an adult human body. In type 2, the human body resists insulin. It may attack the heart, lungs, stomach, kidney and also other parts of the body. Gestational Diabetes is affected to pregnant ladies. It can be harmful for the baby and the mother.

To identify and diagnose diabetes there are several tests in the field of medical. They include the following:

• Fasting Blood Glucose Test (FBS)

• Post Prandial Blood Sugar Test(PPBS)

• Random Blood Sugar Level (RBS)

• Oral Sugar Tolerance Test

• Glycosylated hemoglobin(HbAlc)

• Urine Test

But machine learning classification algorithms can predict better result from the dataset of previous prediction. One of the study, suggested that AdaBoost algorithm with decision stump as base classifier, they got good accuracy and can predict diabetes. They also use some other classification algorithms to get a better prediction [1]. In another study, their purpose of study was to compare the classification algorithms. [6]. It was globally researched that diabetics is a terrible disease all over the world. [8]. All these studies point that diabetes predictions are an important factor in medical research field

III Related work

Machine learning is one of the key way to predict or identify diseases. Diabetes is a long life disease, so it should be predicted anyhow. A recent paper discusses designing a model that can predicted the chance of among the patients with high accuracy. This study uses three classification algorithms of machine learning, like- Decision Tree, SVM and Naïve Bayes, to detect diabetes at an early stage on Pima Indians Diabetes Database (PIDD). Different measures like accuracy, precision, F-measure and recall evaluate the performance of these algorithms. The obtained results show that Naïve Bayes performs more than other algorithms with the greatest accuracy of 76.30 percent [2]. Based on the Pima Indians Diabetes Database (PIDD) another study develops and amalgam model which combines K- means with KNN with multistep preprocessing. In general, higher values of K reduce the effect of noise. The cascaded K-mean and KNN model achieved 97.4% accuracy while value of K is higher [5]. Another study suggests a method for the classification of diabetes patient using a set of characteristics in accordance with World Health Organization criteria in order to help and boost diabetes diagnosis. Here a precision value of 0.070 and a recall value of 0.775 are obtained by the algorithm Hoeffding Tree [3]. Another research compares machine learning classifier like- J48 Decision Tree, KNN, support vector machine and random forest to classify patients with diabetes mellitus. In terms of sensitivity, accuracy and specificity, the performance of the algorithms has both been measured with noisy (before preprocessing) and without noisy (after preprocessing) dataset [4]. Support Vector Machine is one the promising methods of machine learning. SVM used to detect diabetes and indulges about the diabetes complications [7].

IV Proposed Method

As this study aims to predict diabetes using machine learning classification algorithm, so we need clear view of data.

A. Dataset collection

The dataset was collected from mldata.io that speaks about some features to detect diabetes. This dataset is about "Pima Native American Diabetes". The dataset consists of 768 instances and 9 attributes. All attributes are numerical. The 'class' attribute can be used as the class label. For class attribute value of 0 or 1 correspond to no diabetes and diabetes. Those nine attributes from the dataset are used as an input in the proposed classification model.

B. Implementation Details

The classifiers were built using python. In python, scikit learn library was used to build the classifier. A few other additional libraries were used and the code was written in Spyder IDE. Numpy, pandas, sklearn.model, sklearn.metrics, matplotlib.pyplot, sklearn.naive_bayes python packages were used to implement the code.

There were no missing values in the dataset, but there were some zero values which were replaced by the mean value of that particular column. After that, every column was normalized to convert every column between 0 and 1.

$$df = \frac{df}{df.\max()}$$

Then dataset was split into two parts which are input and target. To train the model, we used 75% of the dataset, and the rest 25% was used for prediction. After doing all that, we import different kinds of model such as GaussianNB for Naïve Bayes algorithm,

RandomForestClassifier for Random Forest algorithm, KNeighborsClassifier for KNN algorithm, SVC Support vector machine algorithm, DecisionTreeClassifier for decision tree algorithm to compare those classification machine learning algorithms, we trained the model with the training dataset and then predicted for test dataset. We find out the accuracy by comparing test dataset and predicted result. We used the confusion matrix to determine the performance of classification algorithms.

V Results and Discussion

For classification problems machine learning algorithms are the best choice. For predicting purposes, calculation of precision, accuracy score had done and printed as the confusion matrix. True positive rate, true negative rate, false positive rate and false negative rate from the dataset are also calculated. For calculating accuracy and precision there are a popular formula,

Accuracy = TP+TN/TP+FP+FN+TN

Precision = TP/TP+FP

Algorithms	Accuracy score	Precision	True positive rate	True negative rate	False positive rate	False negative rate
Naïve Bayes	0.72	0.62	0.65	0.77	0.23	0.35
Decision tree	0.72	0.60	0.68	0.75	0.25	0.32
Support vector machine	0.76	0.70	0.61	0.85	0.15	0.39
KNN	0.73	0.64	0.60	0.81	0.19	0.41
Random forest	0.75	0.65	0.71	0.83	0.20	0.37

Fig1: accuracy, precision, TPR, TNR, FPR, FNR for different classifier algorithms.

Accuracy score shows that best performer was Random Forest classifier and Support Vector Machine. Random Forest is best performer in case of TPR. That means Random forest can detect diabetes most efficiently when a person has actually diabetes. In terms of False positive rate

Decision tree gives the worst performance. And SVM gives best performance. These means that decision tree often detects diabetes when person have not actually had it. Random forest is like an extended version of decision tree algorithm. It implements a lot of tree rather than just one tree. Each tree gives a prediction and the class that receives the most prediction is declared as the winner. This probably gives it some edge over other algorithms. When evaluating accuracy value, it shows that all algorithms have over 70% accuracy. In this dataset accuracy is not a good measure since this dataset has a class named "1" which happens rarely. Any model that always returns false as class will automatically score a decent accuracy rate. So, accuracy can't be a good measure here. So, we need other measures. That's where True positive rate, false positive rate come into play. TPR means how many of the actual positive data were correctly classified as true. FPR means how many of the negative data were incorrectly classified as true. Random Forest excelled at TPR while decision tree didn't do too bad itself with a rate of 68%. FPR was topped by SVM followed by K-nearest Neighbors with the rate of 19%. Precision is the ratio of true positives and the all positive predicted values. It basically indicates how much of the positively predicted values are actually positive. SVM tops this with a rate of 70%. So basically, what it means that if the SVM classifier predicts the class to be positive then the chances of it really being positive is 70% which is a fairly good score. Using the true positive, true negative, false positive and false negative rate confusion matrix was implemented.

Confusion matrix shows the following results for the classification algorithms.

Naïve Bayes Algorithm:

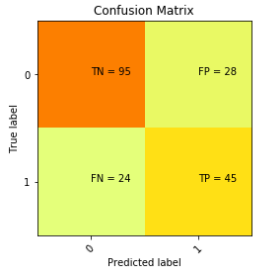

Fig2: confusion matrix for Naïve Bayes algorithm.

Figure 2 shows that for Naïve Bayes algorithm we get 45 true positive rate, 28 false positive rate, 95 true negative rate and 24 false negative rates for the test dataset.

Decision Tree Algorithm:

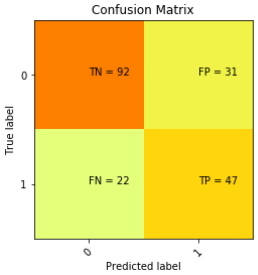

Fig3: confusion matrix for Decision tree algorithm.

Figure 3 shows that for decision tree algorithm we get 47 true positive rate, 31 false positive rate, 92 true negative rate and 22 false negative rates for the test dataset.

Support Vector Machine Algorithm:

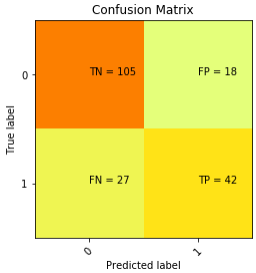

Fig4: confusion matrix for Support vector machine algorithm.

Figure 4 shows that for Support vector machine algorithm, we get 42 true positive rate, 18 false positive rate, 105 true negative rate and 27 false negative rates for the test dataset.

Random forest algorithm:

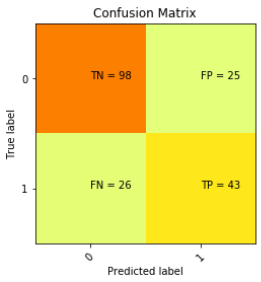

Fig5: confusion matrix for Random forest algorithm.

Figure 5 shows that for Random forest algorithm, we get 43 true positive rate, 25 false positive rate, 98 true negative rate and 26 false negative rates for the test dataset.

KNN algorithm:

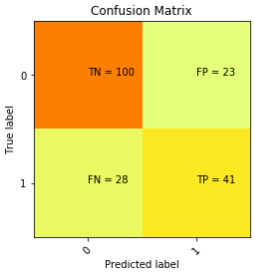

Fig6: confusion matrix for KNN algorithm.

Figure 6 shows that for KNN algorithm we get 41 true positive rate, 23 false positive rate, 100 true negative rate and 28 false negative rates for the test dataset.

After comparing the machine learning classifiers, we get the following results,

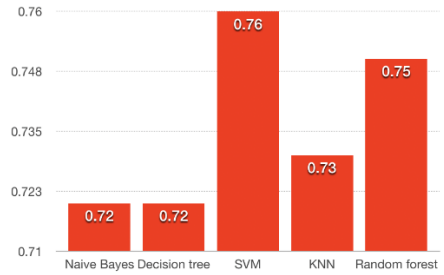

Fig7: Accuracy for Different classifiers.

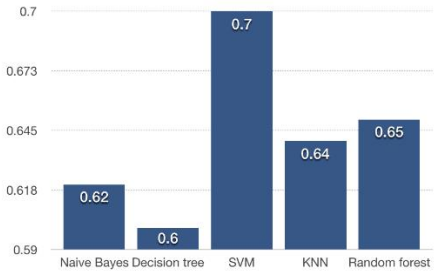

Fig8: Precision for Different classifiers.

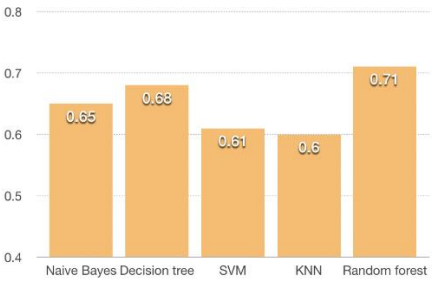

Fig9: True Positive Rate for Different classifiers.

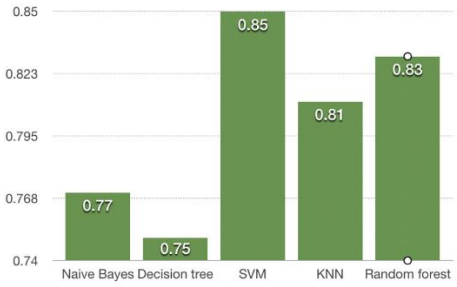

Fig10: True Negative Rate for Different Classifier

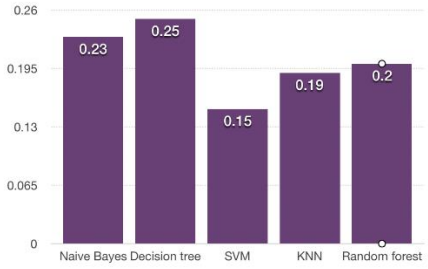

Fig11: False Positive Rate for Different Classifier

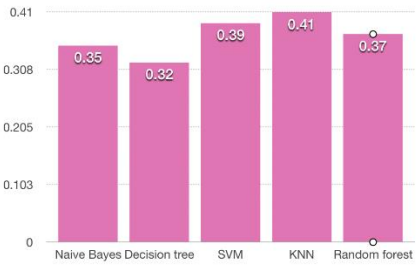

Fig12: False Negative Rate for Different Classifier

VI Conclusion and Future work

Now-a-days, the most important thing is to find diabetics as soon as possible. After using machine learning classification algorithms such as- naive Bayes, decision tree, support vector machine, K-nearest neighbor, random forest we predicted a patient have diabetes or not. Among these algorithms, Support Vector Machine was the most effective and we got the better performance when we used it. Support Vector Machine gives better performance because we have more '0' class then 1 in our dataset. So, accuracy rate is not a good measurement to get a better prediction here, but true positive rate and true negative rate can give a better result. Support vector machine have a better rate in true positive rate and true negative rate than other algorithms. In future, researchers can use these algorithms on different clinical dataset so that prediction of other diseases and can assist the medical researcher. Future researcher can also use other machine learning algorithms on different dataset to get better prediction.

Reference

[1] V. V. Vijayan and C. Anjali, "Prediction and diagnosis of diabetes mellitus - A machine learning approach," *2015 IEEE Recent Adv. Intell. Comput. Syst. RAICS 2015*, no. December, pp. 122–127, 2016.

[2] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 1578–1585, 2018.

[3] F. Mercaldo, V. Nardone, and A. Santone, "Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques," *Procedia Comput. Sci.*, vol. 112, pp. 2519–2528, 2017.

[4] J. P. Kandhasamy and S. Balamurali, "Performance analysis of classifier models to predict diabetes mellitus," *Procedia Comput. Sci.*, vol. 47, no. C, pp. 45–51, 2015.

[5] M. Nirmaladevi, S. A. Alias Balamurugan, and U. V. Swathi, "An amalgam KNN to predict diabetes mellitus," *2013 IEEE Int. Conf. Emerg. Trends Comput. Commun. Nanotechnology, ICE-CCN 2013*, no. Iceccn, pp. 691–695, 2013.

[6] 2000 Gabir M M et al, "The 1997 American Diabetes Association and 1999 World Health Organization Criteria for Hyperglycemia," vol. 23, no. 8, 2000.

[7] Aishwarya, R., Gayathri, P., Jaisankar, N., "A Method for Classification Using Machine Learning Technique for Diabetes. International Journal of Engineering and Technology (IJET) 5", 2903–2908, 2013.

[8] Curt L Rohlfing, Hsiao-Mei Wiedmeyer, Randie R Little, Jack D England, Alethea Tennill, and David E Goldstein., "Defining the relationship between plasma glucose and HbA1c", Diabetes care 25, 2 (2002), 275–278,2002.