Diabetes Prediction Comparing Machine Learning Classifiers

Hasibur Rahman Shanto, Arif Hasan Nayon, Kazi Mushfiquer Rahman, Sujit Kumar Roy

I Introduction

There are many uncured diseases in the medical sector which should be predicted in early stages and the classification algorithms of data mining and machine learning can be applied here. Diabetes is also an uncured disease which needs to be predicted in early stages. Diabetes occurs 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 diabetes can be genetic issues, family history, ethnicity and environmental setup. To support and recognize 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 and 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 by 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 results from the dataset of previous predictions. 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 the medical research field.

III Related work

Machine learning is one of the key ways 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 predict 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 an 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 the value of K is higher [5]. Another study suggests a method for the classification of diabetes patients 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 classifiers likeJ48 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 is used to detect diabetes and indulges about the diabetes complications [7].

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 attributes, values 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.

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.

$$df = df / df.\max()$$

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. ……..

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.