INTERNATIONAL JOURNAL OF ADVANCED INNOVATIVE TECHNOLOGY IN ENGINEERING

Published by Global Advanced Research Publication House Journal

Home page: www.ijaite.co.in

Diabetes Prediction using Machine Learning

[1]Daksh Ghatate, [2]Sanket Bhoyar, [3]Mohammed Rayyan, [4]Madhurmeet Jadhav, [5]Farhan Qureshi, [6]Ima Rahman, [7]Prof. Dr. M. S. Khatib

[1,2,3,4,5,6,7]Department of Computer Engineering, Anjuman College of Engineering and Technology, Nagpur, Maharashtra, India

[1]dakshghatate@gmail.com, [2]sanketbhoyar60@gmail.com, [3]rayyan.mohammed2000@gmail.com, [4]madurmeetj@gmail.com, [5]farhan09qureshi@gmail.com, [6]imarahman92@gmail.com, [7]mskhatib@anjumanengg.edu.inm

Article History

Received on: 10 April 2022

Revised on: 22 April 2022

Accepted on: 12 May 2022

Keywords: Type 2 Diabetes, Machine Learning, Random Forest, Prediction.

e-ISSN: 2455-6491

Production and hosted by
www.garph.org

ABSTRACT

Diabetes is a chronic disease that has the potential to trigger a global health disaster. According to the International Diabetes Federation, 382 million people worldwide have diabetes. By 2035, this figure will have more than quadrupled to 592 million. The primary goal of this study is to create prediction model using the supplied medical data of diabetic and non-diabetic patients. The goal of this research is to develop hybrid model that clinicians may utilize to treat diabetes patients. Key characteristics such as Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function and Age were initially picked using the PIMA Indian Diabetes Dataset to build the prediction model. The dataset was divided into two sections: training and testing. Based on these data, we next used a random forest machine learning algorithm to predict whether the patient will be normal (non-diabetic), or diabetic.

1. INTRODUCTION

Diabetes is a chronic condition that arises when the pancreas does not create enough insulin or when the body does not utilize the insulin that is produced adequately. Insulin is a hormone that controls blood sugar levels. Hyperglycaemia, or high blood sugar, is a typical side effect of untreated diabetes, and it may cause catastrophic harm to many of the body's systems, particularly the neurons and blood vessels, over time.

Diabetes type 1 - Your body does not produce insulin if you have type 1 diabetes. Your immune system targets and kills insulin-producing cells in your pancreas. Type 1 diabetes is most commonly diagnosed in children and young adults, but it can

occur at any age. To stay alive, people with type 1 diabetes must take insulin every day.
Diabetes type 2 - Your body does not produce or utilize insulin well if you have type 2 diabetes. Type 2 diabetes can strike at any age, including youth. However, this type of diabetes occurs most often in middle-aged and older people. The most frequent kind of diabetes is type 2.

People with diabetes commonly lack knowledge about the disease or are asymptomatic; diabetes frequently goes unreported; about one-third of diabetic people are unaware of their condition. Diabetes causes substantial long-term damage to multiple organs and bodily systems, including the kidneys, heart, nerves, blood vessels, and eyes, if it is not treated. Thus, early identification of the condition allows persons at risk to adopt preventative measures to slow disease development and increase the quality of life.

Diabetic symptoms include blurred vision, fatigue, weight loss, increased hunger and thirst, frequent urination, confusion, poor healing, frequent infections, and difficulty concentrating. Design principle, components and its merits over other conventional engines.

By exploiting the advantages of the advancement in machine learning techniques, we have proposed an approach for the classification, prediction of diabetes in this paper. We have employed the widely used classifier, i.e., random forest. To demonstrate the effectiveness of the proposed approach, PIMA Indian Diabetes Dataset is used for experimental evaluation. The accuracy results of our proposed approach demonstrate its adaptability in many healthcare applications.

Machine learning is a technique for directly training computers or machines. By constructing multiple classification and ensemble models from acquired datasets, various machine learning approaches give efficient results for collecting knowledge. This type of information can be used to predict diabetes. Various machine learning algorithms are capable of making predictions, but selecting the optimum methodology is difficult. Machine learning approaches are commonly employed in diabetes prediction, and they produce better outcomes. In the medical industry, decision trees are a prominent machine learning approach with good classification capability. The random forest creates a large number of decision trees. As a result, we random forests in our research (RF).

2. Related Work

A. Mujumdar, V. Vaidehi [2], in this research paper, the authors has shown the correlation/ comparative study between PIMA dataset and updated dataset. They have used PIMA dataset for their study in which they have used 800 records and 10 attributes in which after study they came to an conclusion that the parameter JOB TYPE is of no use, also they had proposed that glucose level has relation with age. So, that parameter is important for the dataset. Multiple machine learning algorithm is used like SVC, RFC, DTC, Extra tree classifier, Ada Boost Algorithm, Linear Regression. In which LR has accuracy of 96% with RFC having accuracy of 91% and Naive Bayes with accuracy of 93%.

Deberneh, H.M.; Kim, I. [3], in the following research, the researchers had used electronic dataset from Korean hospital with 10000 records from 2011 till 2016. In which the data contained of the patients which were being getting diagnosed from multiple years in the hospital. Dataset contains 12 features. According to study glucose is the prime parameter according to American Diabetes Association (ADA). Feature selection techniques were used in this study so initially they had 18 parameters and by feature selection they sorted 12 important features for the prediction.
Algorithm used are RFA, XGBoost, SVM, LR. As per studies they proposed that RFA has the highest accuracy and it will improve as per addition of new records.

KM Jyoti Rani. [4], the researchers have proposed the hybrid model using RFA and NB. They had used the PIMA Indian Dataset with 768 records
and 9 features. In which RSA was used for feature selection. As per their study they have proposed that skin thickness is the weak feature, and shown the training accuracy of RFA as 98%.

Priyanka Indoria, Yogesh Kumar Rathore [5], this research focuses on using machine learning methods for improving ailment perception and diagnostic accuracy. The numerous machine learning approaches utilised to categorise the data sets include supervised, unsupervised, reinforcement, semi-supervised, deep learning, and evolutionary learning algorithms. It also compares the two approaches, namely Nave Bayes and Artificial Neural Networks (ANN). The Nave Bayes theorem is used by the Bayesian Network, which implies that the existence of any characteristic in a class is unrelated to the presence of any other attribute, making it far more beneficial, efficient, and independent.

3. Proposed Methodology

Diabetes prediction is used in this study to detect the onset of diabetes. Diabetes prediction system is built in Python using machine learning to improve the medical profession. Diabetes is recognised in backend processing using sklearn's Random Forest algorithm. Using our suggested technique, the system will identify the accuracy. If the system detects certain Diabetes traces based on the input values in the supplied dataset, our system will detect them using our suggested technique.

The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome.

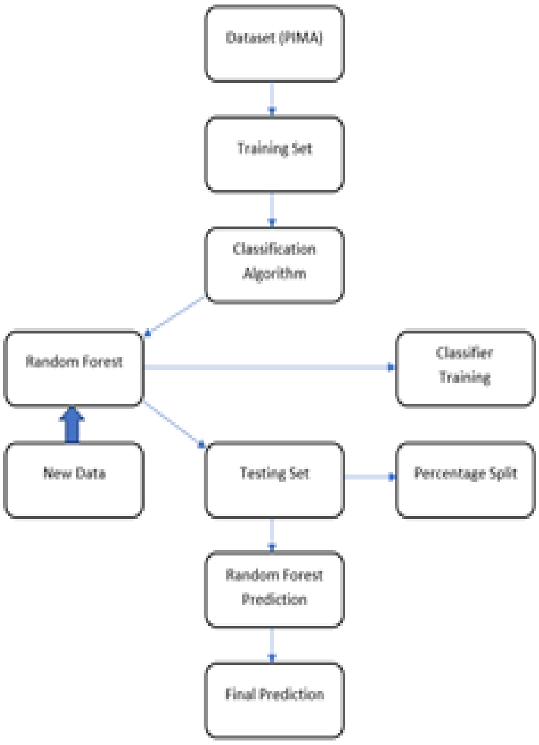


Figure 1: Steps of Proposed System

A. *Dataset Collection*

The first step is dataset preparation. We used the Pima Indians Diabetes Dataset from the Kaggle site. The dataset accommodates 2000 rows with 9 columns. These values are congregated from the diabetes.csv (Diabetes dataset).

B. *Splitting the Dataset*

Dividing the dataset into training and test data is one of the crucial steps in analysis. This process is basically carried out to ensure test data is different from the training data because we need to test the model followed by the training process. First, the training data undergoes through learning and then, the data which is trained is generalized on the other data, based on which the prediction is made. The dataset in our case is split into multiple variants and prediction is performed accordingly. The dataset has multiple columns that are medical predictors and one target column, that of the diabetic's outcome. The medical predictors are given as inputs to a variable and the target variable is given as input to another variable.

Using the inbuilt function, train_test_split, the dataset is split into arrays and is mapped to training and test subsets. In our case, we are performing splits of 80/20,70/30,75/25,60/40 and the accuracy of each is recorded. It was noticed that the dataset contains some null values, to streamline the analysis and the prediction, the null values were filled with the mean values of the respective columns.

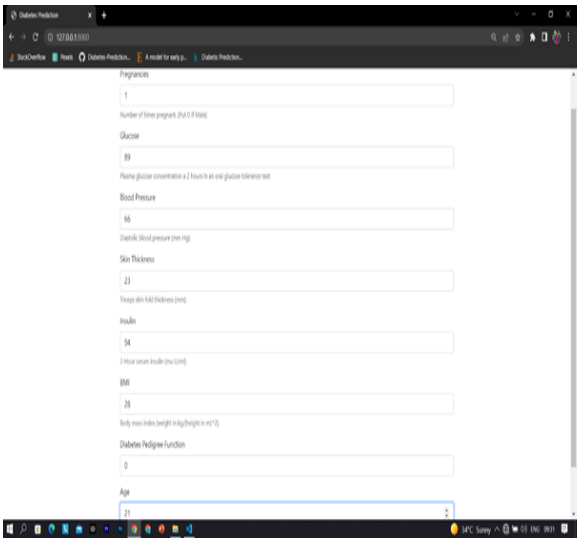
C. *Random Forest*

A random forest [6] is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems.

A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms.

The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome.

4. RESULTS



Pregnancies

1

Number of times pregnant. (Put 0 if Male)

Glucose

89

Plasma glucose concentration a 2 hours in an oral glucose tolerance test.

Blood Pressure

66

Diastolic blood pressure (mm Hg).

Skin Thickness

23

Triceps skin fold thickness (mm).

Insulin

94

2-Hour serum insulin (mu U/ml).

BMI

28

Body mass index (weight in kg/(height in m)^2).

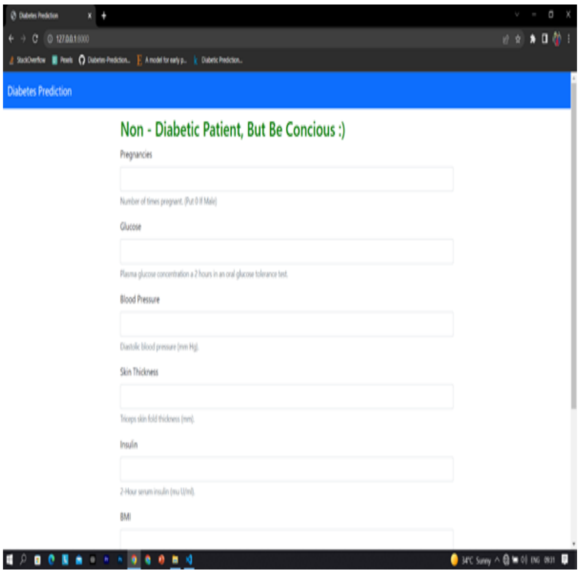
Diabetes Pedigree Function

0

Age

21

Figure 2: User giving values through UI



Diabetes Prediction

Non - Diabetic Patient, But Be Concious :)

Pregnancies

Number of times pregnant. (Put 0 If Male)

Glucose

Plasma glucose concentration a 2 hours in an oral glucose tolerance test.

Blood Pressure

Diastolic blood pressure (mm Hg).

Skin Thickness

Triceps skin fold thickness (mm).

Insulin

2-Hour serum insulin (mu U/ml).

BMI

Figure 3: Outcome (Patient is non-diabetic)

On training and testing, in our Diabetes Prediction System, we got 80.2% accuracy using Random Forest Algorithm.

5. Advantages

Our system is fast, accurate and fully automatic method for diabetes detection.

Random forest has the benefit that it can handle large datasets easily and also provide the accurate outcome.

6. Application

- The apps' primary goal is diabetes detection.
- The goal of developing this application is to deliver adequate treatment as quickly as feasible and to safeguard human lives that are in risk.
- We can detect diabetes with the assistance of our study. This approach can assist physicians in making early decisions, allowing treatment to begin at an earlier stage.
- This application is advantageous towards both physicians and patients.
- Manual identification is slower, but more accurate and efficient for the user. This application was created to address those issues.
- It is an easy-to-use application.

7 Conclusion

Diabetes is a condition that can lead to a variety of problems. It's important looking at how machine learning may be used to accurately forecast and diagnose this condition. The primary goal of this project was to develop and implement Diabetes Prediction using Machine Learning Approach, as well as to analyse the performance of such methods, which was accomplished effectively. The suggested method employs Random Forest. Based on the information in the database, we created a machine learning-based classifier that predicts whether a patient is diabetic or not. We conclude from this article that Random Forest is the best strategy for predicting diabetes. After separating and analysing the training and testing data, this approach yields an estimated result.

However, not every task in this development sector is stated to be ideal, and more enhancement in this application may be conceivable. We've learnt a lot and gained a lot of information about the development sector.

Conflict Of Interest

The authors declare that they have no conflict of interest.

Funding Support

The author declares that they have no funding support for this study.

References

[1] Pima Indians Diabetes Database https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[2] A. Mujumdar, V. Vaidehi, Diabetes Prediction using Machine Learning Algorithms DOI: 10.1016/j.procs.2020.01.047Corpus ID: 212837137 (2019)

[3] Deberneh, H.M.Kim, I. Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. Int. J. Environ. Res. Public Health 2021, 18,3317. https://doi.org/10.3390/ijerph18063317

[4] KM Jyoti Rani. Diabetes Prediction Using Machine Learning doi: https://doi.org/10.32628/CSEIT206463(IJSRCSEIT-2020)

[5] Priyanka Indoria, Yogesh Kumar Rathore. A survey: Detection and Prediction of diabetics using machine learning techniques. IJERT, 2018.

[6] 'Onesmus Mbaabu' Introduction to Random Forest in Machine Learning.

[7] https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/