



# Diabetes Prediction using Machine Learning

Daksh Ghatate<sup>1</sup>, Sanket Bhoyar<sup>2</sup>, Farhan Qureshi<sup>3</sup>

Madhurmeet Jadhav<sup>4</sup>, Ima Rahman<sup>2</sup>, Mohammed Rayyan<sup>6</sup>

Student, Computer Science and Engineering, Anjuman College of Engineering and Technology, Nagpur, India<sup>1-6</sup>

**Abstract:** Diabetes is a chronic illness with the potential to cause a worldwide health catastrophe. Diabetes affects 382 million people globally, according to the International Diabetes Federation. This headcount will have more than tripled to 592 million by 2035. The fundamental purpose of this study is to develop a prediction model based on the medical data provided by diabetic and non-diabetic individuals. The purpose of this study is to create a hybrid model that physicians may use to manage diabetic patients. To begin building the prediction model, key parameters such as Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age were selected from the PIMA Indian Diabetes Dataset. The dataset was separated into two parts: training and testing. We then proceeded based on these findings. Following that, we utilised a random forest machine learning system to predict whether the patient will be normal (non-diabetic) or diabetic.

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

## I. INTRODUCTION

Diabetes is a chronic disease that occurs when the pancreas does not make enough insulin or when the body does not properly use the insulin that is produced. Insulin is a hormone that regulates blood glucose levels. Hyperglycaemia, or high blood sugar, is a common side effect of uncontrolled diabetes, and it can cause long-term damage to many of the body's systems, including the neurons and blood vessels.

Diabetes type 1 - If you have type 1 diabetes, your body does not make insulin. Insulin-producing cells in your pancreas are targeted and killed by your immune system. Type 1 diabetes is more frequent in children and young adults, although it can strike anybody at any age. People with type 1 diabetes must take insulin every day to stay alive.

Type 2 diabetes means that your body does not manufacture or use insulin properly. Diabetes type 2 can arise at any age, including childhood. This kind of diabetes, however, is more common in middle-aged and older adults. Kind 2 diabetes is the most common type.

People with diabetes usually lack knowledge about the disease or are asymptomatic; diabetes is routinely underreported; around one-third of diabetics are ignorant of their condition. Diabetes, if left untreated, causes significant long-term damage to various organs and physiological systems, including the kidneys, heart, nerves, blood vessels, and eyes. Thus, early detection of the condition allows those at risk to take preventative actions to delay disease progression and improve quality of life.

Blurred vision, weariness, weight loss, increased appetite and thirst, frequent urination, disorientation, poor healing, recurrent infections, and difficulties focusing are all signs of diabetes. The design philosophy, components, and advantages over other conventional engines are all discussed.

In this study, we offer a strategy for diabetes categorization and prediction that takes use of advances in machine learning techniques. We utilised a commonly popular classifier called random forest. PIMA Indian Diabetes Dataset is utilised for experimental assessment to demonstrate the efficacy of the proposed strategy. Our suggested approach's accuracy results illustrate its flexibility in a wide range of healthcare applications.

Machine learning is a method of teaching computers or machines directly. Various machine learning algorithms provide efficient outcomes for knowledge collection by generating numerous classification and ensemble models from collected datasets. This sort of data can be used to forecast diabetes. Various machine learning algorithms can provide predictions, but choosing the best approach is tricky. Machine learning algorithms are widely used in diabetes prediction and offer improved results. Decision trees are a popular machine learning technique with high categorization power in the medical field. A vast number of decision trees are generated by the random forest. As a consequence, in our investigation, we used random forests.



## II. RELATED WORK

A. Mujumdar and V. Vaidehi [2] demonstrated a correlation/comparison analysis between the PIMA dataset and the revised dataset in their research work. They used the PIMA dataset for their study, which included 800 records and 10 characteristics. After doing their research, they concluded that the parameter JOB TYPE is useless, and they also claimed that glucose levels are related to age. As a result, that parameter is critical for the dataset. SVC, RFC, DTC, Extra tree classifier, Ada Boost Algorithm, and Linear Regression are among the machine learning algorithms employed. In which LR has a 96 percent accuracy, RFC has a 91 percent accuracy, and Naive Bayes has a 93 percent accuracy.

Deberneh, H.M., and I. Kim [3], in the following study, the researchers used an electronic dataset from a Korean hospital including 10,000 records from 2011 to 2016. At which the data includes patients who had been diagnosed for several years in the hospital. The dataset has 12 characteristics. According to the American Diabetes Association, the primary measure is glucose (ADA). In this study, feature selection techniques were utilised, therefore they started with 18 parameters and used feature selection to rank 12 relevant characteristics for prediction.

RFA, XGBoost, SVM, and LR are the algorithms employed. According to research, RFA has the highest accuracy and will improve with the addition of additional records.

KM Jyoti Rani and colleagues [4] presented a hybrid model based on RFA and NB. They utilised the PIMA Indian Dataset, which had 768 entries and 9 characteristics. In which RSA was used to select features. According to their findings, skin thickness is the weakest aspect, and RFA training accuracy is 98 percent.

Yogesh Kumar Rathore and Priyanka Indoria [5], this study focuses on the application of machine learning approaches to improve disease perception and diagnosis accuracy. To categorise the data sets, several machine learning methodologies such as supervised, unsupervised, reinforcement, semi-supervised, deep learning, and evolutionary learning algorithms were used. It also contrasts the two methods, Nave Bayes and Artificial Neural Networks (ANN). The Bayesian Network employs the Nave Bayes theorem, which states that the presence of any attribute in a class is unrelated to the presence of any other attribute, making it significantly more advantageous, efficient, and independent.

## III. PROPOSED METHODOLOGY

In this work, diabetes prediction is utilised to identify the start of diabetes. To better the medical industry, a diabetes prediction system is constructed in Python utilising machine learning. Diabetes is detected in backend processing using the Random Forest method from Sklearn. The system will determine the accuracy using our proposed approach. If the system recognises particular Diabetes traces based on the input values in the given dataset, it will do so using the way we recommend.

The outcome is determined by the (random forest) algorithm based on the predictions of the decision trees. It forecasts by averaging or averaging the output of several trees. The precision of the output improves as the number of trees grows.

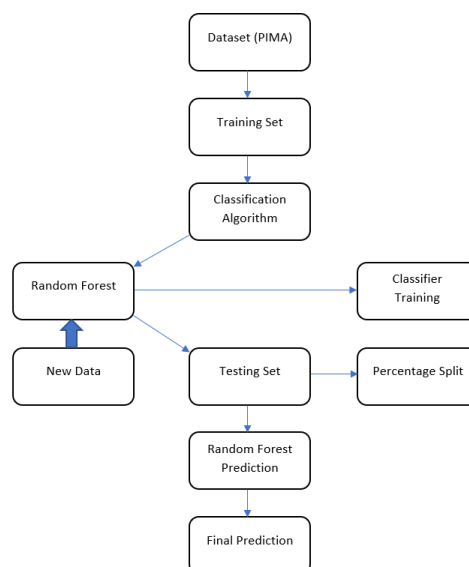


Figure 1: Steps of Proposed System

### A. Dataset Gathering

The first step is to prepare the dataset. We utilised the Kaggle site's Pima Indians Diabetes Dataset. The dataset has 2000 rows and 9 columns. These values are derived from diabetes.csv (Diabetes dataset).

### B. Dataset Segmentation

One of the most important processes in analysis is dividing the dataset into training and test data. This procedure is used to verify that test data differs from training data since we need to test the model after the training phase. First, the training data is learned, and then the trained data is generalised to the other data, on which the prediction is formed. In our situation, the dataset is divided into several versions, and prediction is conducted appropriately. The dataset contains many columns of medical predictors and one goal column, which represents the diabetic's result. The medical predictors are fed into one variable, while the goal variable is fed into another.

The dataset is split into arrays and assigned to training and test subsets using the built-in function train test split. In our example, we execute splits of 80/20, 70/30, 75/25, and 60/40 and report the accuracy of each. The dataset had some null values, which were filled with the mean values of the appropriate columns to simplify the analysis and prediction.

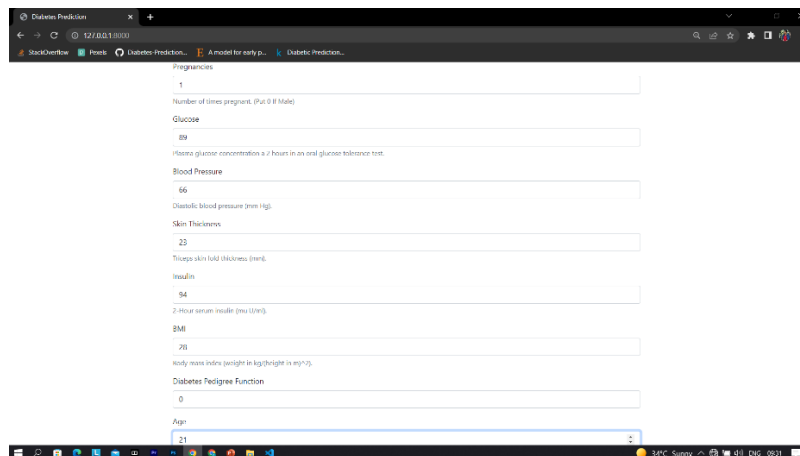
### C. Random Forest

A random forest [6] is a machine learning approach used to tackle regression and classification issues. It employs ensemble learning, a technique that combines several classifiers to find answers to complicated problems.

A random forest algorithm is made up of numerous decision trees. The 'forest' formed by the random forest technique is trained using bagging or bootstrap aggregation. Bagging is an ensemble meta-algorithm that increases the accuracy of machine learning systems.

The (random forest) algorithm determines the outcome based on the predictions of the decision trees. It forecasts by taking the average or mean of the output from different trees. Increasing the number of trees improves the outcome's accuracy.

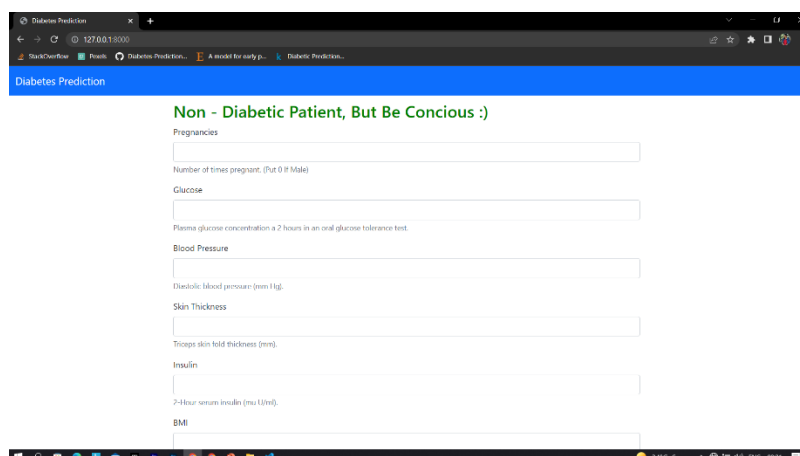
## IV.RESULT



The screenshot shows a web browser window titled "Diabetes Prediction". The page contains a form with the following input fields and values:

- Pregnancies: 1
- Number of times pregnant (Put 0 if Male): 0
- Glucose: 89
- Blood Pressure: 66
- Skin Thickness: 23
- Insulin: 94
- BMI: 21
- Age: 21

Figure 2: User giving Input using UI



The screenshot shows the same web browser window as Figure 2, but now displaying the prediction result. The result is "Non - Diabetic Patient, But Be Concious :)" in green text. Below the result, the input form is visible again.

Figure 3: Result (Patient is non-diabetic)



Using the Random Forest Algorithm, we achieved 80.2 percent accuracy in our Diabetes Prediction System during training and testing.

## V. ADVANTAGES

Our approach is a quick, accurate, and entirely automated way for detecting diabetes.

Random forest has the advantage of being able to manage enormous datasets while still producing reliable results.

## VI. APPLICATIONS

- The major purpose of the app is diabetes identification; the goal of designing this software is to offer proper treatment as soon as possible and to protect human lives that are at risk.
- With the help of our research, we can identify diabetes. This strategy can help clinicians make early judgments, allowing therapy to begin sooner.
- This application is beneficial to both clinicians and patients.
- Manual identification is more time-consuming, but it is more accurate and efficient for the user. This application was designed to solve such concerns.
- It is a simple application.

## VII. CONCLUSION

Diabetes is a disease that can cause a range of complications. It is critical to consider how machine learning may be utilized to effectively predict and diagnose this illness. The fundamental purpose of this research was to create and deploy diabetes prediction techniques using a machine learning approach, as well as to analyse the performance of such approaches, which was successfully completed. Random Forest is used in the proposed technique. We developed a machine learning-based classifier that predicts whether a patient is diabetic or not based on the information in the database. This paper concludes that Random Forest is the best technique for predicting diabetes. This method produces an approximated result after separating and analyzing the training and testing data.

However, not every activity in this development sector is said to be optimal, and more advancement in this application is possible. We've learned a lot about the development industry and gained a lot of knowledge.

## 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.2020.01.047 Corpus 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\(IJSRCSEIT-2020\)](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/>