

# **PROJECT REPORT**

(Project Term August-December 2022)

## **Diabetes Prediction using Machine Learning**

Submitted By

**Manoj Varma Lakkammraju**  
**11907874**

**Suresh Kumar Mettela**  
**11907286**

**INT 248 (Advance Machine Learning)**

**(B. Tech CSE)**

Under the Guidance of

**Niharika Thakur**

**School of Computer Science and Engineering**

**Lovely Professional University**  
**Phagwara, Punjab.**



**L**OVELY  
**P**ROFESSIONAL  
**U**NIVERSITY

## DECLARATION

We here by declare that the project work entitled Diabetes prediction using Machine learning is an authentic record of our own work carried out as requirements of project for the award of B. Tech degree in computer science and Engineering from Lovely Professional University ,Punjab. under the guidance of Niharika Thakur, during August to November 2022.All the information finished in this project report is based on our own intensive work and is genuine.

**Name of student 1: L.Manoj Varma.**  
**Reg.No:11907874.**

**Name of student 2: M.Suresh Kumar.**  
**Reg.No:11907286.**

# **CERTIFICATE**

This is to certify that the declaration statement made by this group of students is correct to the best of my knowledge and belief. They have completed this project under my guidance and supervision. The present work is the result of their original investigation effort and study. No part of the work has ever been submitted for any other degree at any university. The Project is fit for the submission and partial fulfillment of the conditions for the award of B. Tech degree in Computer Science and Engineering from Lovely Professional University, Punjab.

**Name of Mentor: Niharika Thakur**

**School of Computer Science & Engineering**

**Lovely Professional University, Punjab.**

# **ACKNOWLEDGEMENT**

We take this opportunity to express our deep gratitude and most sincere thanks to our teachers, parents, and friends for giving most valuable suggestion, helpful guidance and encouragement in the execution of this project work.

We would like to thank our course mentor for guiding us in making the Project work successful.

# CONTENTS

1. **INTRODUCTION**
2. **RELATED WORKS**
3. **DATASET**
4. **PROPOSED METHODS**
  - I) **DATASET COLLECTION**
  - II) **DATA PRE PROCESSING**
  - III) **MISSING VALUE IDENTIFICATION**
  - IV) **FEATURE SELECTION**
  - V) **SCALING AND NORMALISATION**
  - VI) **SPLITTING OF DATA**
  - VII) **DESIGN AND IMPLEMENTATION OF CLASSIFICATION MODEL**
  - VIII) **MACHINE LEARNING CLASSIFIER**
5. **MODELING AND ANALYSIS:**
  - a) **LOGISTIC REGRESSION**
  - b) **K-NEAREST NEIGHBORS**
  - c) **SVM**
  - d) **NAIVE BAYES**
  - e) **DECISION TREE**
  - f) **RANDOM FOREST**
  - g) **ADABOOST CLASSIFIER**
6. **MEASUREMENT**
7. **RESULTS AND DISCUSSION**
8. **RESULTS AND ANALYSIS**
9. **CONCLUSION**

# 1. Introduction

All around there are numerous ceaseless infections that are boundless in evolved and developing nations. One of such sickness is diabetes. Diabetes is a metabolic issue that causes blood sugar by creating a significant measure of insulin in the human body or by producing a little measure of insulin. Diabetes is perhaps the deadliest sickness on the planet. It is not just a malady yet, also a maker of different sorts of sicknesses like a coronary failure, visual deficiency, kidney ailments and nerve harm, and so on.

Subsequently, the identification of such chronic metabolic ailment at a beginning period could help specialists around the globe in forestalling loss of human life. Presently, with the ascent of machine learning, AI, and neural systems, and their application in various domains [1, 2] we may have the option to find an answer for this issue. ML strategies and neural systems help scientists to find new realities from existing well-being-related informational indexes, which may help in ailment supervision and detection. The current work is completed utilizing the Pima Indians Diabetes Database. The point of this framework is to make an ML model, which can anticipate with precision the likelihood or the odds of a patient being diabetic. The ordinary distinguishing process for the location of diabetes is that the patient needs to visit a symptomatic focus. One of the key issues of bio-informatics examination is to achieve precise outcomes from the information. Human mistakes or various laboratory tests can entangle the procedure of identification of the disease. This model can foresee whether the patient has diabetes or not, aiding specialists to ensure that the patient in need of clinical consideration can get it on schedule and also help anticipate the loss of human lives.

DNA makes neural networks the apparent choice. Neural networks use neurons to transmit data across various layers, with each node working on a different weighted parameter to help predict diabetes.

Presently, with the ascent of machine learning, AI, and neural systems, and their application in various domains [1, 2] we may have the option to find an answer for this issue. ML strategies and neural systems help scientists to find new realities from existing well-being-related informational indexes, which may help in ailment supervision and detection. The current work is completed utilizing the Pima Indians Diabetes Database.

## **Causes of Diabetes**

Genetic factors are the main cause of diabetes. It is caused by at least two mutant genes in the chromosome 6, the chromosome that affects the response of the body to various antigens. Viral infection may also influence the occurrence of type 1 and type 2 diabetes. Studies have shown that infection with viruses such as rubella, Cocksackievirus, mumps, hepatitis B virus, and cytomegalovirus increase the risk of developing diabetes.

## **Types of Diabetes**

### **Type 1**

Type 1 diabetes means that the immune system is compromised and the cells fail to produce insulin in sufficient amounts. There are no eloquent studies that prove the causes of type 1 diabetes and there are currently no known methods of prevention.

### **Type 2**

Type 2 diabetes means that the cells produce a low quantity of insulin or the body can't use the insulin correctly. This is the most common type of diabetes, thus affecting 90% of persons diagnosed with diabetes. It is caused by both genetic factors and the manner of living.

Data mining and machine learning have been developing, reliable, and supporting tools in the medical domain in recent years. The data mining method is used to pre-process and select the relevant features from the healthcare data, and the machine learning method helps automate diabetes prediction [14]. Data mining and machine learning algorithms can help identify the hidden pattern of data using the cutting-edge method; hence, a reliable accuracy decision is possible. Data Mining is a process where several techniques are involved, including machine learning, statistics, and database system to discover a pattern from the massive amount of dataset [15]. According to Nvidia: Machine learning uses various algorithms to learn from the parsed data and make predictions.

## 2. Related Works

Diabetes prediction is a classification technique with two mutually exclusive possible outcomes, either the person is diabetic or not diabetic. After extensive research, we came to conclusion that although numerous classification techniques can be used for the purpose of prediction, the observed accuracy varied. On careful examination of the performance of techniques used in prevalent works, logistic regression, KNN, Naive Bayes [3], random forest, decision tree, and neural network [4], we found them at par when applied to our dataset. KNN and logistic regression techniques were able to achieve 80% accuracy.

The primary factor which influenced our algorithm selection was its adaptability and compatibility with future applications. The inevitable shift of data storage toward DNA makes neural networks the apparent choice. Neural networks use neurons to transmit data across various layers, with each node working on a different weighted parameter to help predict diabetes.

The point of this framework is to make an ML model, which can anticipate with precision the likelihood or the odds of a patient being diabetic. The ordinary distinguishing process for the location of diabetes is that the patient needs to visit a symptomatic focus. One of the key issues of bio-informatics examination is to achieve precise outcomes from the information. Human mistakes or various laboratory tests can entangle the procedure of identification of the disease. This model can foresee whether the patient has diabetes or not, aiding specialists to ensure that the patient in need of clinical consideration can get it on schedule and also help anticipate the loss of human lives.

### 3. Data Set

The dataset collected is originally from the Pima Indians Diabetes Database is available on Kaggle. It consists of several medical analyst variables and one target variable. The objective of the dataset is to predict whether the patient has diabetes or not. The dataset consists of several independent variables and one dependent variable, i.e., the outcome. Independent variables include the number of pregnancies the patient has had their BMI, insulin level, age, and so on as Shown in Following Table 1:

**Table 1** Dataset description

Serial no	Attribute Names	Description
1	Pregnancies	Number of times pregnant
2	Glucose	Plasma glucose concentration
3	Blood Pressure	Diastolic blood pressure
4	Skin Thickness	Triceps skin fold thickness (mm)
5	Insulin	2-h serum insulin
6	BMI	Body mass index
7	Diabetes pedigree function	Diabetes pedigree function
8	Outcome	Class variable (0 or 1)
9	Age	Age of patient

Fig 3.1 dataset description.

- The diabetes data set consists of 2000 data points, with 9 features each.
- “Outcome” is the feature we are going to predict, 0 means No diabetes, 1 means diabetes.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 9 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Pregnancies            2000 non-null   int64
1   Glucose                2000 non-null   int64
2   BloodPressure          2000 non-null   int64
3   SkinThickness          2000 non-null   int64
4   Insulin                2000 non-null   int64
5   BMI                    2000 non-null   float64
6   DiabetesPedigreeFunction 2000 non-null   float64
7   Age                   2000 non-null   int64
8   Outcome                2000 non-null   int64
dtypes: float64(2), int64(7)
memory usage: 140.8 KB
```

Fig 3.2 predictions

- There is no null values in dataset.



## Correlation Matrix:

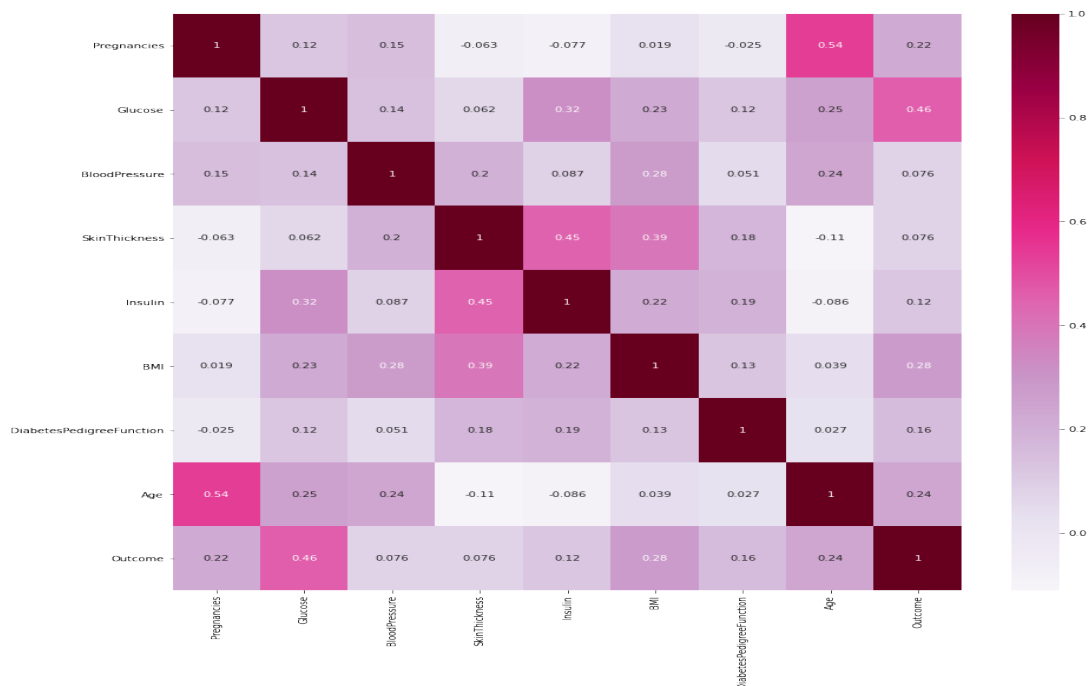


Fig 3.3 correlation matrix

It is easy to see that there is no single feature that has a very high correlation with our outcome value. Some of the features have a negative correlation with the outcome value and some have positive.

## Skew Of Data:

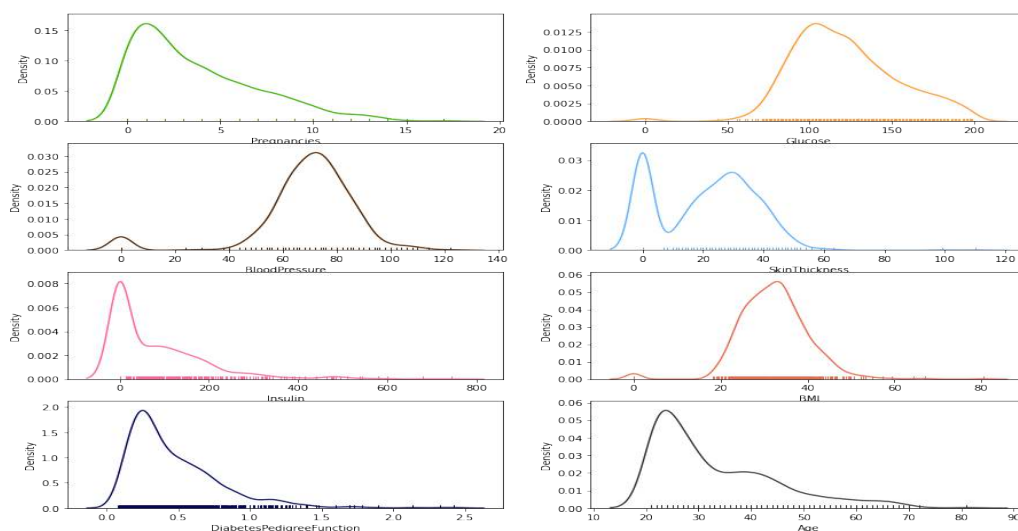


Fig 3.4 skew of

data

It shows how each feature and label is distributed along different ranges, which further confirms the need for scaling. It basically means that each of these is actually a categorical variable. We will need to handle these categorical variables before applying Machine Learning. Our outcome labels have two classes, 0 for no disease and 1 for disease.

## Bar Plot for Outcome Class

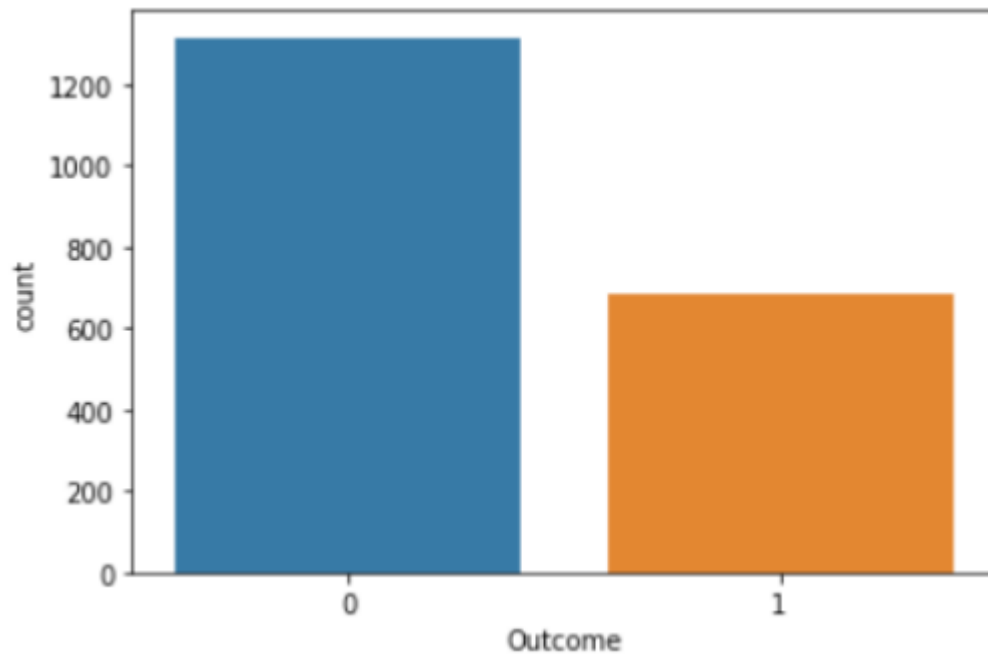


fig 3.5 Bar plot for outcomes class

The above graph shows that the data is biased towards datapoints having outcome value as 0 where it means that diabetes was not present actually. The number of non-diabetics is almost twice the number of diabetic patients.

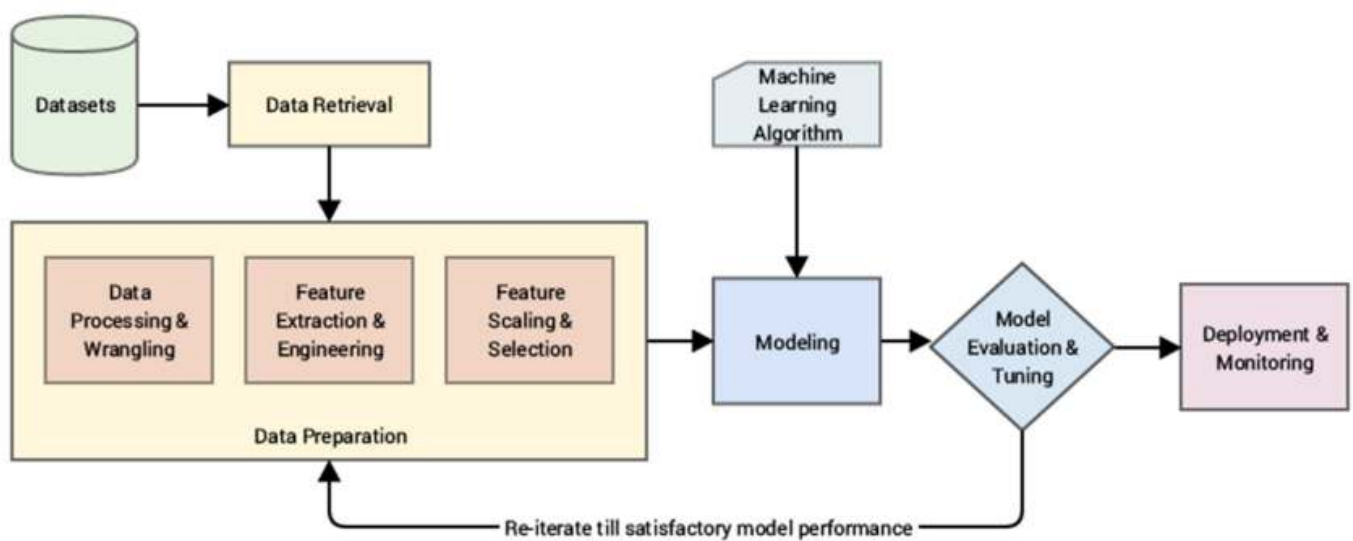


fig 3.6 Graph

## 4. PROPOSED METHODS

**I| Dataset collection** – It includes data collection and understanding the data to study the hidden patterns and trends which helps to predict and evaluating the results. Dataset carries 1405 rows i.e., total number of data and 10 columns i.e., total number of features. Features include Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, DiabetesPedigreeFunction, Age

### II| Data Pre-processing:

This phase of model handles inconsistent data in order to get more accurate and precise results like in this dataset Id is inconsistent so we dropped the feature. This dataset doesn't contain missing values. So, we imputed missing values for few selected attributes like Glucose level, Blood Pressure, Skin Thickness, BMI and Age because these attributes cannot have values zero. Then data was scaled using StandardScaler. Since there were a smaller number of features and important for prediction so no feature selection was done.

### III|Missing value identification:

Using the Panda library and SK-learn , we got the missing values in the datasets, shown in Table 2. We replaced the missing value with the corresponding mean value.

<i>Pregnancies</i>	<i>0</i>
<i>Glucose</i>	<i>13</i>
<i>Blood Pressure</i>	<i>90</i>
<i>Skin Thickness</i>	<i>573</i>
<i>Insulin</i>	<i>956</i>
<i>BMI</i>	<i>28</i>
<i>DPF</i>	<i>0</i>
<i>Age</i>	<i>0</i>
<i>Outcome</i>	<i>0</i>

Fig 4.1 Missing value identified

#### IV] Feature selection:

Pearson's correlation method is a popular method to find the most relevant attributes/features. The correlation coefficient is calculated in this method, which correlates with the output and input attributes. The coefficient value remains in the range by between  $-1$  and  $1$ . The value above  $0.5$  and below  $-0.5$  indicates a notable correlation, and the zero value means no correlation

Attributes	Correlation coefficient
Glucose	0.484
BMI	0.316
Insulin	0.261
Preg	0.226
Age	0.224
Skin Thickness	0.193
BP	0.183
DPF	0.178

Fig 4.2 *Correlation*

Table:

#### V] Scaling and Normalization:

We performed feature scaling by normalizing the data from 0 to 1 range, which boosted the algorithm's calculation speed.

scaling means that you're transforming your data so that it fits within a specific scale, like 0-100 or 0-1. You want to scale data when you're using methods based on measures of how far apart data points are, like support vector machines (SVM) or k-nearest neighbours (KNN). With these algorithms, a change of "1" in any numeric feature is given the same importance.

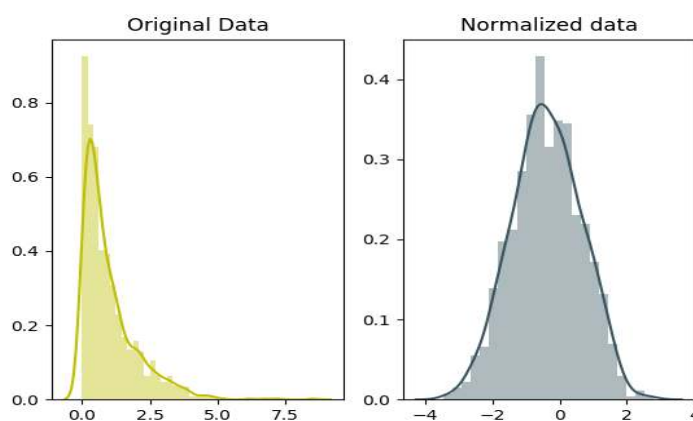


Fig 4.3 Scaling and normalization

#### VI] Splitting of data:

After data cleaning and pre-processing, the dataset becomes ready to train and test. In the train/split method, we split the dataset randomly into the training and testing set. For Training we took 1600 sample and for testing we took 400 sample.

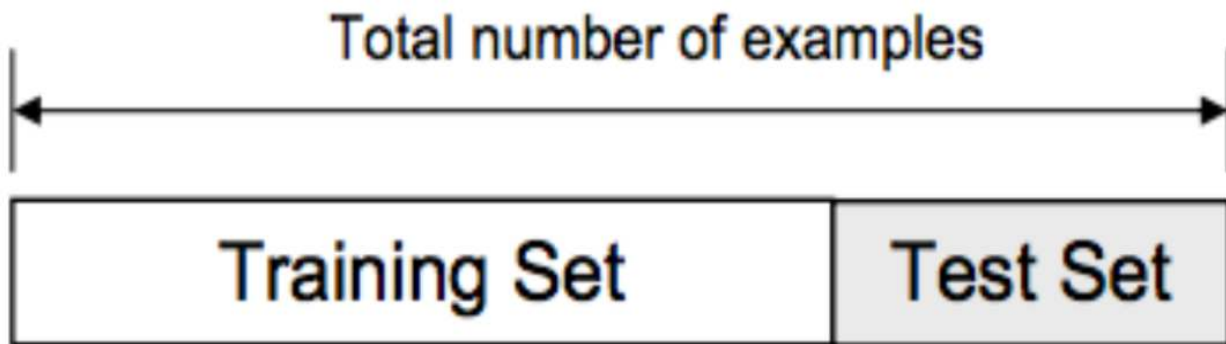


Fig 4.4 Spitting data

#### **VII] Design and implementation of classification model:**

In this research work, comprehensive studies are done by applying different ML classification techniques like DT, KNN, RF, NB, LR, SVM.

#### **VIII] Machine learning classifier:**

We have developed a model using Machine learning Technique. Used different classifier and ensemble techniques to predict diabetes dataset. We have applied SVM, LR, DT and RF Machine learning classifier to analyse the performance by finding accuracy of each classifier All the classifiers are implemented using scikit learn libraries in python. The implemented classification algorithms are described in next section.

## 5. MODELING AND ANALYSIS:

### A] Logistic Regression:

Logistic regression is a machine learning technique used when dependent variables are able to categorize. The outputs obtained by using the logistic regression is based on the available features. Here sigmoidal function is used to categorize the output.

### B] K-Nearest Neighbors:

K-nearest neighbors (KNN) algorithm uses 'feature similarity' to predict the values of new datapoints which further means that the new data point will be assigned a value based on how closely it matches the points in the training set.

Predictions are made for a new instance (x) by searching through the entire training set for the K most similar instances (the neighbors) and summarizing the output variable for those K instances.

### C] SVM:

SVM is supervised learning algorithm used for classification. In SVM we have to identify the right hyper plane to classify the data correctly. In this we have to set correct parameters values. To find the right hyper plane we have to find right margin for this we have choose the gamma value as 0.0001 and rbf kernel. If we select the hyper plane with low margin leads to miss classification.

### D] Naive Bayes:

Naive Bayes classifiers are a collection of classification algorithms based on **Bayes' Theorem**. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

### E] Decision Tree:

Decision tree is non parametric classifier in supervised learning. In this method all the details are represented in the form of tree, where leaves are corresponds to the class labels and attributes are corresponds to internal node of the tree. We have used Gini Index for splitting the nodes.

### F] Random Forest:

Random forest is an ensemble learning method for classification. This algorithm consists of trees and the number of tree structures present in the data is used to predict the accuracy. Where leaves are corresponds to the class labels and attributes are corresponds to internal node of the tree. Here number of trees in forest used is 100 in number and Gini index is used for splitting the nodes.

## G] AdaBoost Classifier:

**Boosting** is an ensemble modeling technique that attempts to build a strong classifier from the number of weak classifiers. It is done by building a model by using weak models in series. Firstly, a model is built from the training data. Then the second model is built which tries to correct the errors present in the first model. This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models are added.

**AdaBoost** was the first really successful boosting algorithm developed for the purpose of binary classification. *AdaBoost* is short for *Adaptive Boosting* and is a very popular boosting technique that combines multiple “weak classifiers” into a single “strong classifier”. It was formulated by Yoav Freund and Robert Schapire. They also won the 2003 Gödel Prize for their work.

## 6. Measurements

To find the efficient classifier for diabetes prediction we have applied performance matrices. Confusion matrix and accuracy are discussed as follows:

Confusion matrix: - which provides output matrix with complete description of the model's performance.

Here, TP: True positive

FP: False positive

TN: True negative

FN: False negative

**Fig 6.1 Actual values**

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN



The following performance metrics are used to calculate the presentation of various algorithms.

- True positive (TP) – person has disease, and the prediction also has a positive
- True negative (TN) – person not having disease and the prediction also has a negative
- False positive (FP) – person not having disease but the prediction has a positive
- False negative (FN) – person having disease and the prediction also has a positive
- TP and TN can be used to calculate accuracy rate and the error rates can be computed using FP and FN values.
- True positive rate can be calculated as TP by a total number of persons have disease in reality.
- False positive rate can be calculated as FP by a total number of persons do not have disease in reality.
- Precision is TP/ total number of person have prediction result is yes.
- Accuracy is the total number of correctly classified records

Accuracy- We have chooses accuracy matrix to measure the performance of all the models. The ratio of number of correct predictions to the total number of predictions Made.

$$\text{Accuracy} = \frac{\text{Number of correct Prediction}}{\text{Total numbers of predictions made.}}$$

## 7. RESULTS AND DISCUSSION

Machine learning classification algorithms developed for prediction of diabetes in earlier stage. We used 70% of data for training and 30% of data for testing. In this ratio of data splitting Here we found that Random Forest Classifier predicted with 99% of accuracy as highest accuracy for the dataset. Comparison of results of all the implemented classifiers are listed in below.

Machine Learning Algorithms	Result
<b>Logistic Regression</b>	79.0
<b>K-Nearest Neighbors</b>	80.5
<b>SVM</b>	84.5
<b>Naive Bayes</b>	76.83
<b>Decision Tree</b>	96.0
<b>Random Forest</b>	<b>98.0</b>
<b>AdaBoost Classifier</b>	81.16

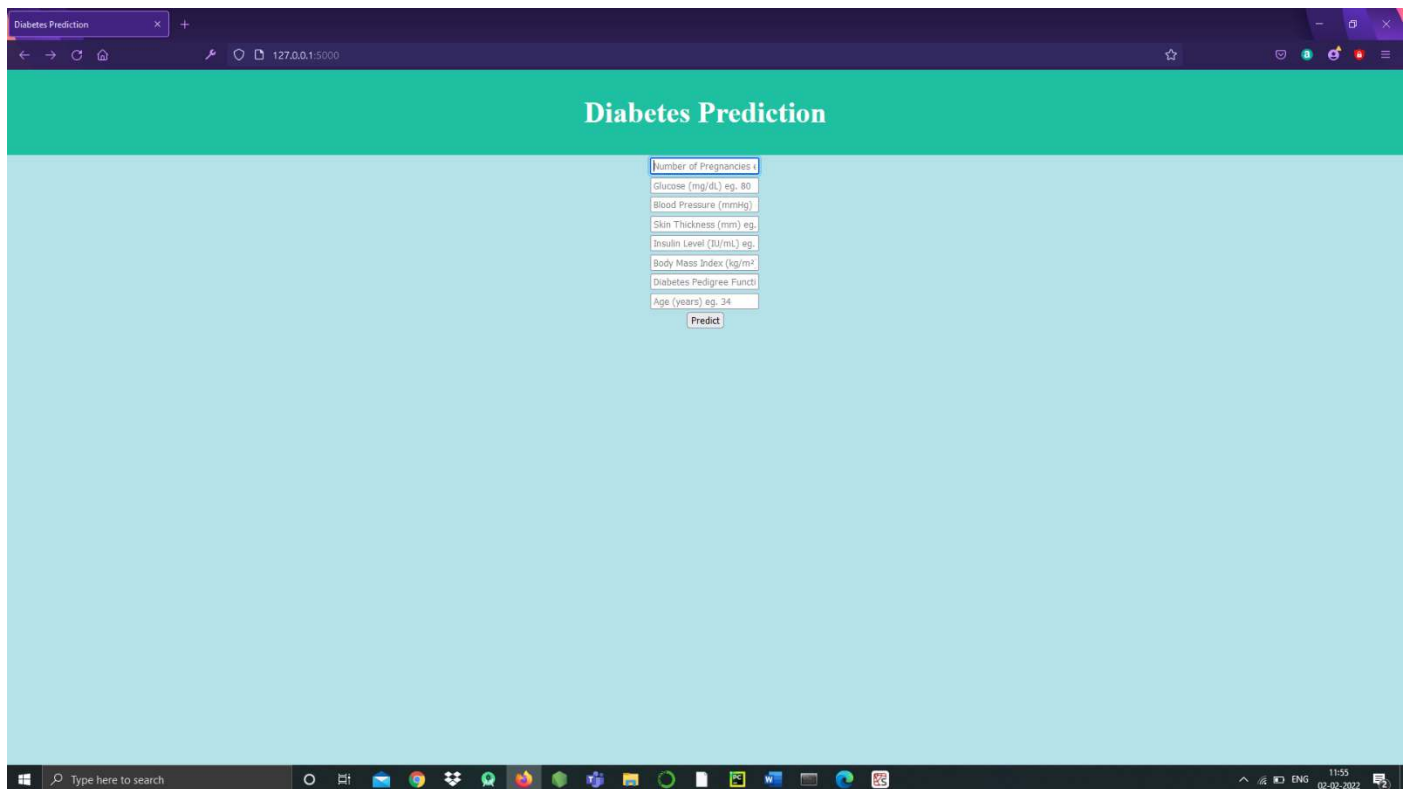
Fig 7.1 Results

### Creating a User Interface for Accessibility:

The last part of the project is the creation of a user interface for the model. This user interface is used to enter unseen data for the model to read and then make a prediction. The user interface is created using “Flask” Web app, Hyper Text Markup Language, and Cascading Style Sheets.

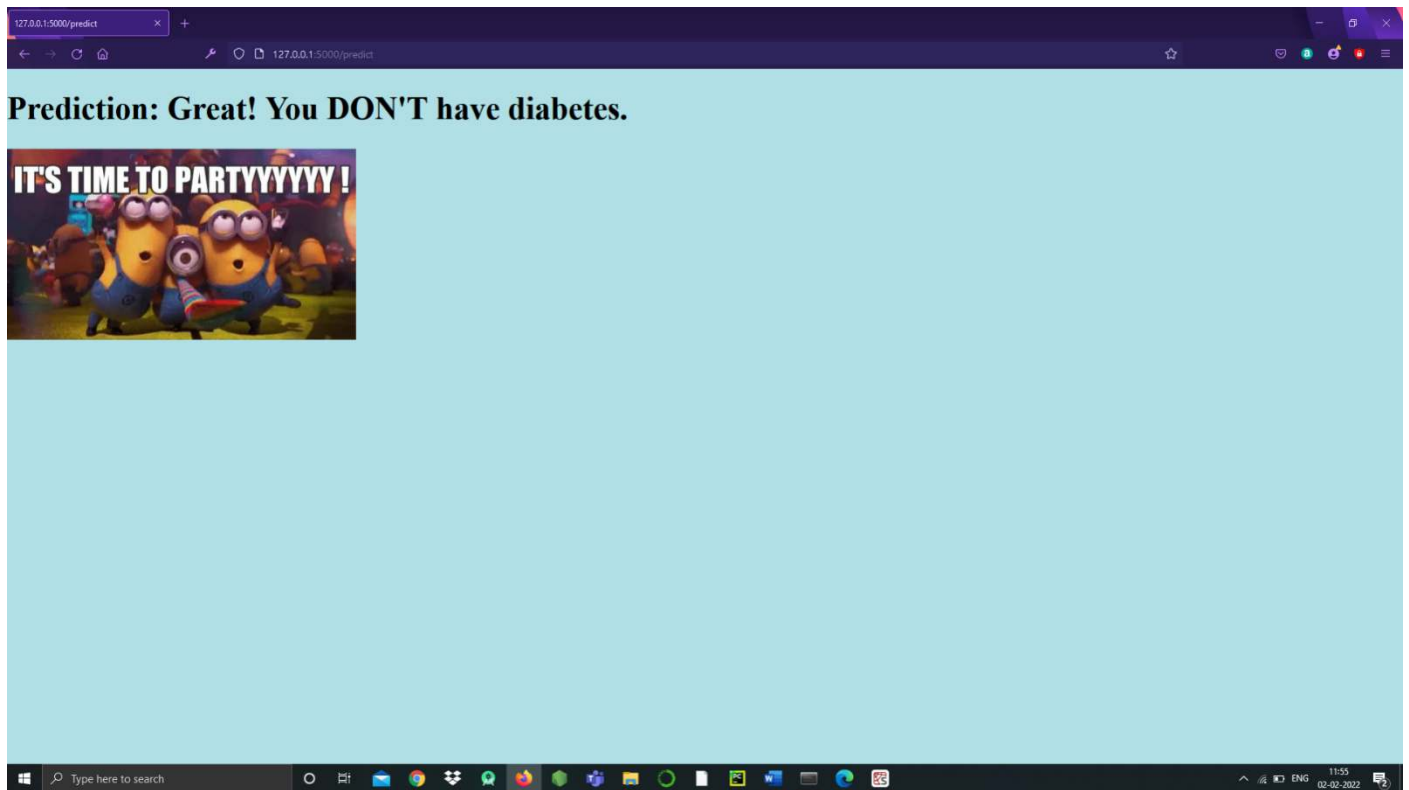
## 8. Results and Analysis

The project predicts the onset of diabetes in a person based on the relevant medical details collected. When the person enters all the relevant medical data required in the online Web portal, this data is then passed on to the trained model for it to make predictions whether the person is diabetic or non-diabetic the model then makes the prediction with an accuracy of 98%, which is fairly good and reliable. Following figure shows the basic UI form which requires the user to enter the specific medical data fields. These parameters help determine if the person is prone to develop diabetes Our research has the added benefit of an associated Web app, which makes the model more user friendly and easily understandable for a novice

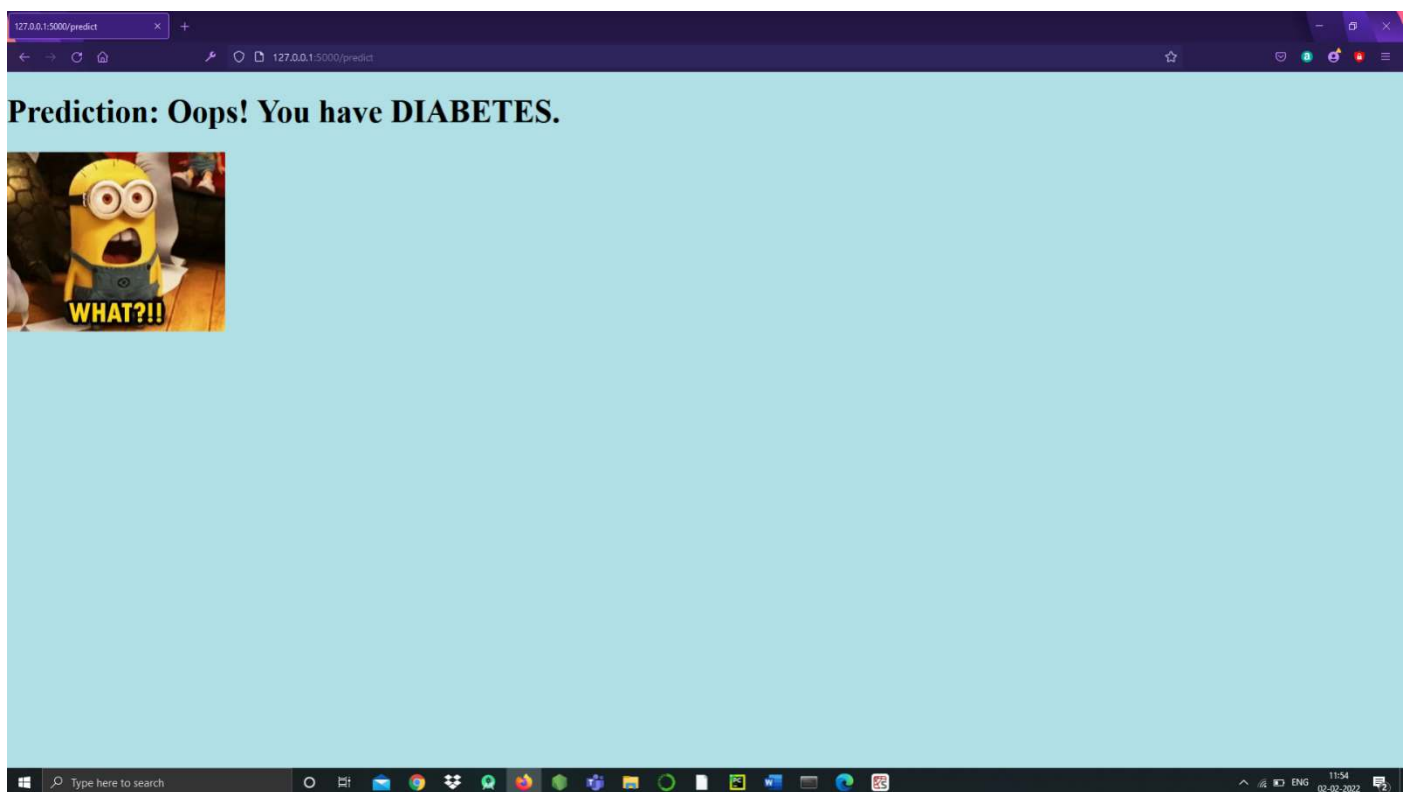
The image shows a web browser window with the title "Diabetes Prediction". The page has a green header bar with the text "Diabetes Prediction" in white. Below the header, the background is light blue. In the center, there is a vertical stack of input fields, each with a label and a placeholder value: "Number of Pregnancies" (placeholder: 4), "Glucose (mg/dL) eg. 80", "Blood Pressure (mmHg)", "Skin Thickness (mm) eg.", "Insulin Level (IU/mL) eg.", "Body Mass Index (kg/m²)", "Diabetes Pedigree Functi", and "Age (years) eg. 34". Below these fields is a "Predict" button. The browser's address bar shows "127.0.0.1:3000". The Windows taskbar is visible at the bottom with various application icons and a system clock showing 11:55 on 02-02-2022.

**Fig8.1: Basic Design of UI.**

On submission of this form, data the model gives the result in the form of Text; as shown in following figures;



**Fig 8.2: Prediction for non-diabetic person**



**Fig 8.3: Prediction for diabetic person**

## 9. Conclusion

The objective of the project was to develop a model which could identify patients with diabetes who are at high risk of hospital admission. Prediction of risk of hospital admission is a fairly complex task. Many factors influence this process and the outcome. There is presently a serious need for methods that can increase healthcare institution's understanding of what is important in predicting the hospital admission risk. This project is a small contribution to the present existing methods of diabetes detection by proposing a system that can be used as an assistive tool in identifying the patients at greater risk of being diabetic. This project achieves this by analyzing many key factors like the patient's blood glucose level, body mass index, etc., using various machine learning models and through retrospective analysis of patients' medical records. The project predicts the onset of diabetes in a person based on the relevant medical details that are collected using a Web application. When the user enters all the relevant medical data required in the online Web application, this data is then passed on to the trained model for it to make predictions whether the person is diabetic or nondiabetic. The model is developed using artificial neural network consists of total of six dense layers. Each of these layers is responsible for the efficient working of the model. The model makes the prediction with an accuracy of 98%, which is fairly good and reliable.

**PROJECT URL:**

**<https://github.com/Manojvarma2207/Machine-Learning-Projects>**

# THANK YOU