Diabetes detection and Analysis using Machine Learning

Marpu Jagadeesh[1], Bosubabu Sambana[2]
[1] *B.Tech, Department of Civil Engineering, Andhra University Visakhapatnam*
[2]*Associate Professor, Department of Computer Science and Engineering,*
Lendi Institute of Technology and Engineering (Autonomous) Vizianagaram
Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, India

Abstract- Diabetes is a disease that comes with an increase in blood glucose, also called blood sugar, which is too high. Glucose is our main source of energy and comes from the food we eat. Insulin is a hormone made by the pancreas it helps glucose from food to get into your cells to be used for energy. Sometimes your body doesn't enough or any insulin or doesn't insulin well. Glucose then stays in your blood and doesn't reach your cells. Over time, having too much glucose in your blood can cause health problems. So now it's very important to predict diabetes in an early stage via a simple blood test, our machine learning approach to detecting diabetes is what we have used in the process we answer the simple questions asked by the UI of our program, and boom you will get your result whether you suffer from diabetes or not.

Keywords: Diabetes, machine learning, glucose level, insulin stages of diabetes, symptoms, drugs.

I. INTRODUCTION

Diabetes is a disease that comes with an increase in blood glucose, also called blood sugar, which is too high. Glucose is our main source of energy and comes from the food we eat. Insulin is a hormone made by the pancreas it helps glucose from food to get into your cells to be used for energy. Sometimes your body doesn't make enough or any insulin or doesn't use insulin well. Glucose then stays in your blood and does n't reach your cells. Over time, having too much glucose in your blood can cause health problems. Although diabetes has no cure, you can take steps to manage your diabetes and stay healthy.

Diabetes affects just about everyone. The most common types of diabetes are type 1, type 2, and gestational diabetes. If you have type 1 diabetes, your body does not make insulin. Your immune system attacks and destroys the cells in your pancreas that make insulin. Type 1 diabetes is usually diagnosed in children and young adults, although it can appear at any age. People with type 1 diabetes need to take insulin every day to stay alive. If you have type 2 diabetes, your body does not make or use insulin well. You can develop type 2 diabetes at any age, even during childhood. However, this type of diabetes occurs most often in middle-aged and older people. Type 2 is the most common type of diabetes. Gestational diabetes develops in some women when they are pregnant. Most of the time, this type of diabetes goes away after the baby is born [1]. However, if you've had gestational diabetes, you have a greater chance of developing type 2 diabetes later in life. Over time, high blood glucose leads to problems such as heart stroke, kidney disease, eye problems dental disease, nerve damage, and foot problems.

The prediction of diabetes has become so fast and easy with the availability of our
Program, and also there is a provision for checking symptoms, drugs, side effects, statistics of diabetes and finally the prediction function used for prediction it asks several questions and those areas listed below

a). A number of pregnancies
b). Glucose level
c). Blood pressure
d). Skin thickness
e). Insulin level
f). Body mass index
g). Diabetes pedigree function
h). Age

Based on these outcomes the ML model predicts whether the diabetes is there or not.

Types Of Diabetes With Their Symptoms

Symptoms of Type 1 Diabetes:

People who have type 1 diabetes may also have nausea, vomiting, or stomach pains. Type 1 diabetes symptoms can develop in just a few weeks or months and can be severe. Type 1 diabetes usually starts when you're a child, teen, or young adult but can happen at any age [2].

Symptoms of Type 2 Diabetes:

Type 2 diabetes symptoms often take several years to develop. Some people don't notice any symptoms at all. Type 2 diabetes usually starts when you're an adult, though more and more children and teens are developing it. Because symptoms are hard to spot, it's important to know the risk factors for type 2 diabetes. Make sure to visit your doctor if you have any of them.

Symptoms of Gestational Diabetes

Gestational diabetes (diabetes during pregnancy) usually doesn't't have any symptoms., If you're pregnant, your doctor should test you for gestational diabetes between 24 and 28 weeks of pregnancy.

If needed, you can make changes to protect your health and your baby's health.

Risk Factors

Type 1 Diabetes

Type 1 diabetes is thought to be caused by an immune reaction (the body attacks itself by mistake).

Risk factors for type 1 diabetes are not as clear as for pre-diabetes and type 2 diabetes.

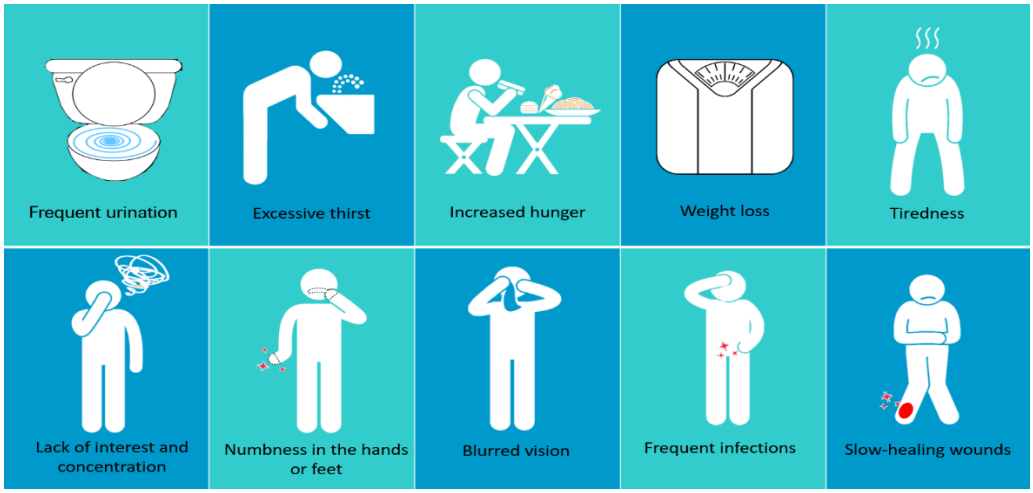


Figure.1: Symptoms of Diabetes diseases

Known risk factors include:

- Family history: Having a parent, brother, or sister with type 1 diabetes.
- Age: You can get type 1 diabetes at any age, but it's more likely to develop when you're a child, teen, or young adult.

In America, whites are more likely to develop type 1 diabetes than African Americans and Hispanic/Latino Americans and currently, no one knows how to prevent type 1 diabetes.

Type 2 Diabetes

You're at risk for developing type 2 diabetes if you:

a). Have pre-diabetes
b). Are overweight
c). Are 45 years or older
d). Have a parent, brother, or sister with type 2 diabetes
e). Are physically active less than 3 times a week
f). Have ever had gestational diabetes (diabetes during pregnancy) or given birth to a baby who weighed more than 9 pounds
g). Are African American, Hispanic/Latino American, American Indian, or Alaska Native (some Pacific Islanders and Asian Americans are also at higher risk)

If you have non-alcoholic fatty liver disease you may also be at risk for type 2 diabetes.

You can prevent or delay type 2 diabetes with simple, proven lifestyle changes such as losing weight if you're overweight, eating healthier, and getting regular physical activity.

Gestational Diabetes

You're at risk for developing gestational diabetes (diabetes while pregnant) if you:

a). Had gestational diabetes during a previous pregnancy
b). Have given birth to a baby who weighed more than 9 pounds
c). Are overweight
d). Are more than 25 years old
e). Have a family history of type 2 diabetes
f). Have a hormone disorder called polycystic ovary syndrome (PCOS)
g). Are African American, Hispanic/Latino American, American Indian, Alaska Native, Native Hawaiian, or Pacific Islander

Gestational diabetes usually goes away after your baby is born but increases your risk for type 2 diabetes later in life. Your baby is more likely to have obesity as a child or teen and is more likely to develop type 2 diabetes later in life too.

Before you get pregnant, you may be able to prevent gestational diabetes by losing weight if you're overweight, eating healthier, and getting regular physical activity.

II. LITERATURE SURVEY

Aishwarya Mujumdara, Dr. Vaidehi Vb," Diabetes Prediction using Machine Learning Algorithms ", has proposed a diabetes prediction model for better classification of diabetes which includes a few external factors responsible for diabetes along with regular factors like Glucose, BMI, Age, and insulin. Classification accuracy is boosted with a new dataset compared to the existing dataset. Further imposed a pipeline model for diabetes prediction intended towards improving the accuracy of classification

Machine Learning Algorithms that can be used for diabetes prediction. The task of choosing a machine learning algorithm includes feature matching of the data to be learned based on existing approaches. The taxonomy of machine learning algorithms is discussed below machine learning has numerous algorithms which are classified into three categories: Supervised learning, Unsupervised Learning, and Semi-supervised learning.

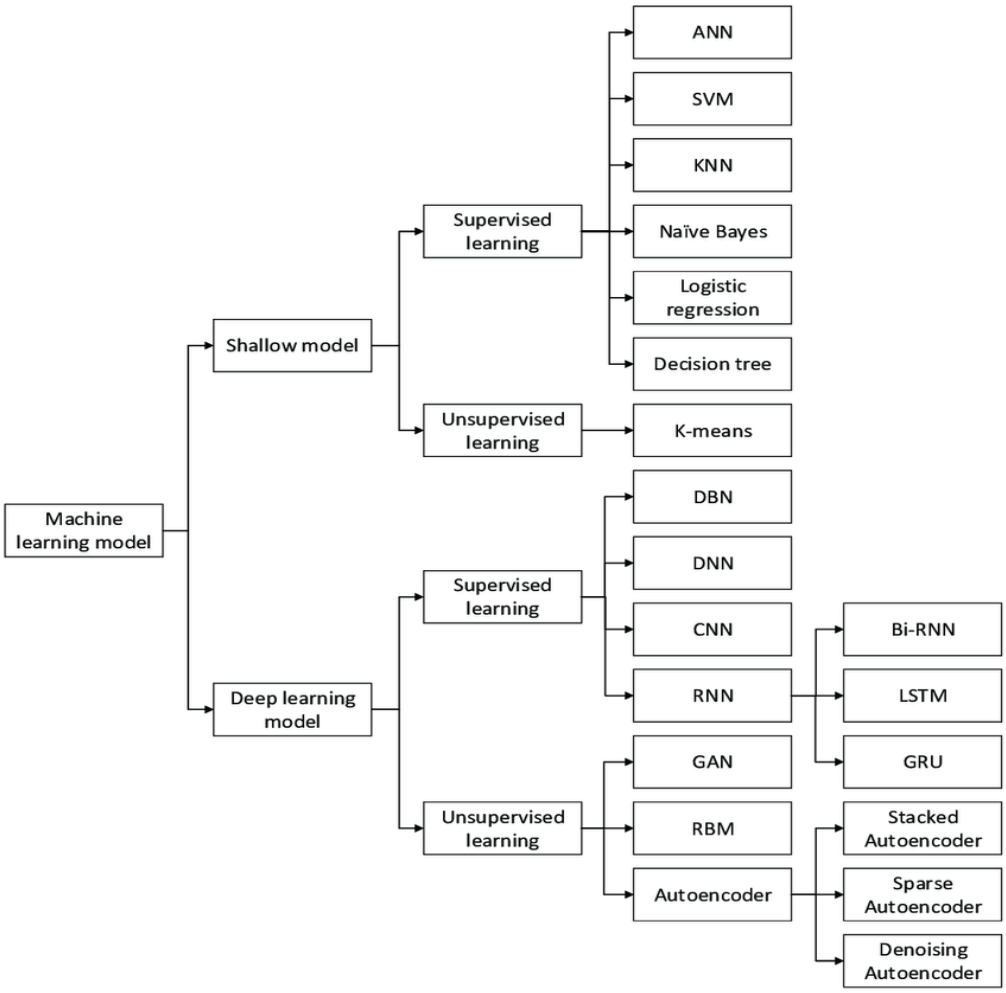


Figure.2: ML and Sub-Fields

The Supervised Learning/Predictive Models Supervised learning algorithms are used to construct predictive models. A predictive model predicts missing values using other values present in the dataset. A supervised learning algorithm has a set of input data and also a set of output, and builds a model to make realistic predictions for the response to the new dataset. Supervised learning includes Decision Tree, Bayesian Method, Artificial Neural Network, Instance-based learning, and Ensemble Method. These are booming techniques in Machine learning.

Unsupervised Learning / Descriptive Models Descriptive models are developed using the unsupervised learning method. In this model, we have a known set of inputs but the output is unknown. Unsupervised learning is mostly used on transactional data. This method includes clustering algorithms like k-Means clustering and k-Medians clustering.

Semi-supervised Learning Semi-Supervised learning method uses both labeled and unlabeled data on the training dataset. Classification, Regression techniques come under Semi-Supervised Learning. Logistic Regression and Linear Regression are examples of regression techniques.

III. EXISTING SYSTEM

However, the old system's categorization and accuracy were not as good. They presented a pipeline approach for diabetes prediction and classification accuracy. Haseen et al discussed how diabetes mellitus risk is classified. Four ML methods were studied: decision tree, ANN, logistic regression, and Naive Bayes. Later, the Bugging and Boosting procedures were used to improve the resilience of the models. Following examination, the random forest was determined to be the best disease model.

Decision Tree Algorithm:-

For predicting the class of a given dataset in a decision tree, the algorithm begins at the tree's root node. This algorithm compares the values of the root attribute with the values of the record (actual dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.

Here are the steps in which the algorithm works:-

Step-1: Begin the tree with the root node, which contains the entire dataset, says S.
Step-2: Using the Attribute Selection Measure, find the best attribute in the dataset (ASM).
Step-3: Subdivide the S into subsets containing potential values for the best qualities.
Step-4: Create the decision tree node with the best attribute.
Step-5: Create new decision trees recursively using the subsets of the dataset obtained in step 3. Continue this process until you reach a point where you can no longer categorize the nodes and refer to the final node as a leaf node.

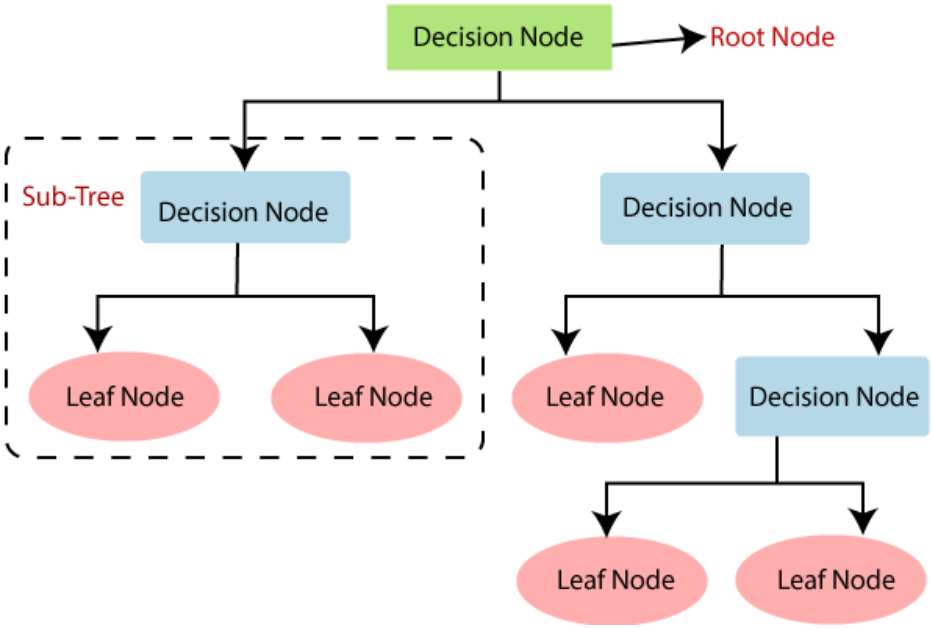


Figure.2: Decision-Tree approach

Artificial Neural Networks(ANN):

Artificial Neural networks, often known as artificial neural networks (ANN), are computational methods. It aimed to mimic the behavior of biological systems made up of "neurons." ANNs are computer models inspired by the central nervous systems of animals. It is capable of both machine learning and pattern recognition. These are depicted as interconnected "neurons" that can compute values from inputs.

A neural network may have the following three layers:

a). **Input Layer** — The raw information that can be fed into the network is represented by the activity of the input units.
b). **Hidden Layer:** To determine the activity of each hidden unit, use the hidden layer. The input units' actions and the weights on the links between the input and the hidden units. One or more hidden layers are possible.

c). **Output Layer:** The activity of the hidden units and the weights between the hidden and output units determine the behavior of the output units.

The below framework shows how the artificial neural network works

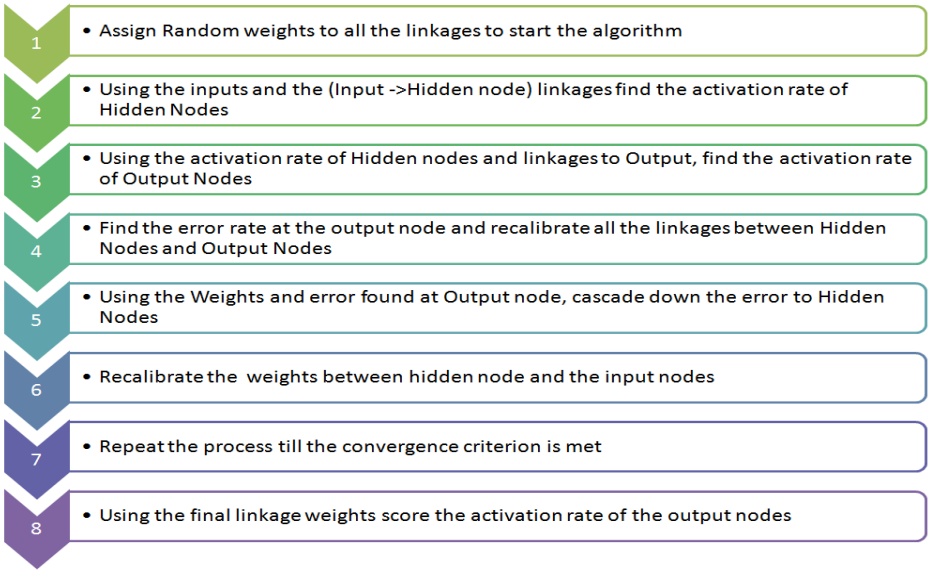


Figure.3: Process Model Mechanism

Logistic Regression: Logistic regression is a common Machine Learning method that belongs to the Supervised Learning technique. It is used to forecast the categorical dependent variable from a group of independent variables. A categorical dependent variable's output is predicted using logistic regression. As a result, the outcome must be categorical or discrete. It can be Yes or No, 0 or 1, true or False, and so on, but instead of presenting the exact values like 0 and 1, it presents the probability values that fall between 0 and 1. Except for how they are employed, Logistic Regression and Linear Regression are very similar. Logistic regression is used to solve classification difficulties, whereas linear regression is used to solve regression problems.

Logistic Regression can be divided into three types based on the categories:

Binomial: In binomial Logistic regression, the dependent variables can only be of two sorts, such as 0 or 1, Pass or Fail, and so on.

Multinomial: In multinomial Logistic regression, the dependent variable might be one of three or more unordered kinds, such as "cats," "dogs," or "sheep."

Ordinal: Ordinal Logistic regression allows for three or more ordered sorts of dependent variables, such as "low," "medium," or "high."

Procedures in Logistic Regression: To develop Logistic Regression in Python, we will follow the same steps that we did in earlier Regression subjects. The steps are as follows:

(a). Pre-processing of data
(b). Logistic Regression Fitting to the Training Set
(c). Predicting the outcome of a test
(d). The result's accuracy was tested (Creation of Confusion matrix)
(e). Visualizing the outcome of the test set.

Naive Bayes Algorithm:

The Naive Bayes method is a supervised learning technique that uses the Bayes theorem to solve classification issues.
It is mostly utilized in text classification with a large training dataset.
The Naive Bayes Classifier is a simple and effective Classification method that aids in the development of fast machine learning models capable of making quick predictions.
It is a probabilistic classifier, which means it predicts based on an object's likelihood.
The Naïve Bayes algorithm is made up of the phrases Naïve and Bayes, which can be translated as:

Naïve: It is dubbed Naïve because it assumes that the occurrence of one trait is unrelated to the occurrence of others. For example, if the fruit is classified based on color, shape, and taste, then a red, spherical, and delicious fruit is identified as an apple. As a result, each feature contributes to identifying it as an apple independently of the others.

Bayes: It is so named because it is based on the principle of Bayes' Theorem.

Working of Nave Bayes' Classifier: The following example will help you understand the working of Nave Bayes' Classifier:

Assume we have a dataset of weather conditions and a target variable called "Play." So, given this dataset, we must select whether or not to play on a certain day based on the weather circumstances. So, in order to overcome this problem, we must take the following steps:

1. Create frequency tables from the given dataset.
2. Create a Likelihood table by calculating the probability of the provided features.
3. Now, apply the Bayes theorem to determine the posterior probability.

IV. PROPOSED SYSTEM

We examined four Machine Learning algorithms in the suggested system.

This model is made up of four distinct components. Among these modules are:

a). Dataset Collection
b). Data Pre-processing
c). Exploratory analysis
d). Build Model
e). Evaluation

Dataset Collection: This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Column descriptions

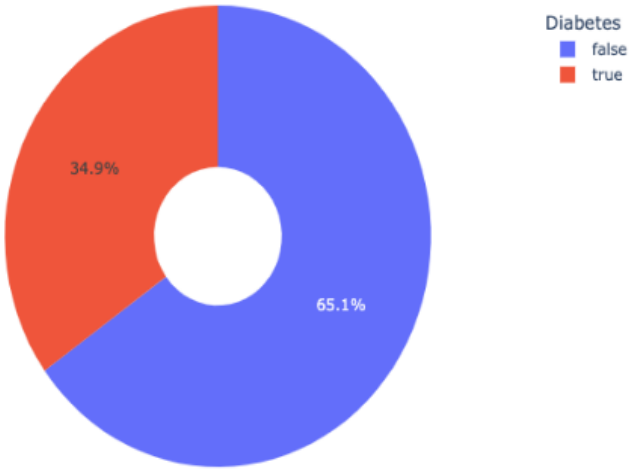
- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- BloodPressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in kg/(height in m)^2)
- DiabetesPedigreeFunction: Diabetes pedigree function
- Age: Age (years)
- Outcome: Class variable (0 or 1) 268 of 768 are 1, the others are 0

We have changed the column name Outcome to Diabetes, and replaced all the 1 and 0 in the Diabetes column with True and False, for visualization purposes

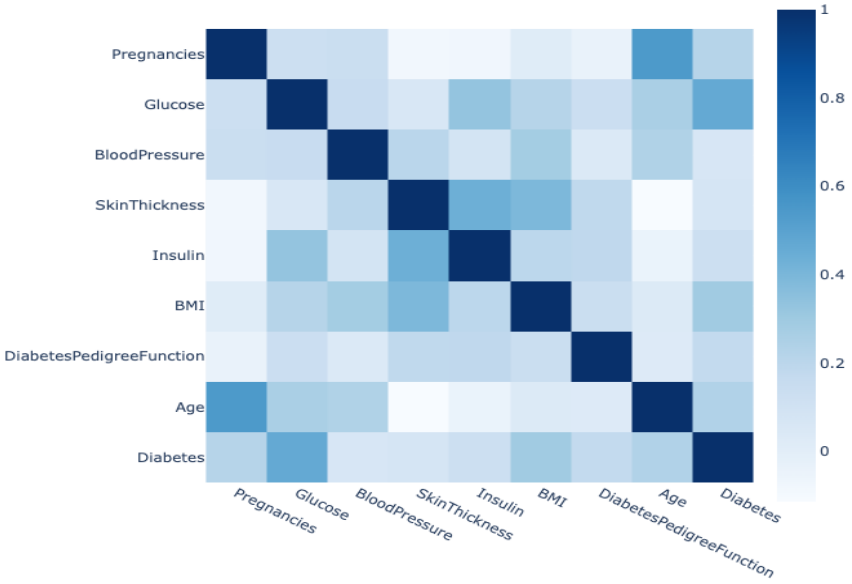
Column Statistics

	mean	std	Min	25%	50%	75%	max
Pregnancies	3.845052	3.369578	0.000000	1.000000	3.000000	6.000000	17.000000
Glucose	120.894531	31.972618	0.000000	99.000000	117.000000	140.250000	199.000000
BloodPressure	69.105469	19.355807	0.000000	62.000000	72.000000	80.000000	122.000000
SkinThickness	20.536458	15.952218	0.000000	0.000000	23.000000	32.000000	99.000000
Insulin	79.799479	115.244002	0.000000	0.000000	30.500000	127.250000	846.000000
BMI	31.992578	7.884160	0.000000	27.300000	32.000000	36.600000	67.100000
DiabetesPedigreeFunction	0.471876	0.331329	0.078000	0.243750	0.372500	0.626250	2.420000
Age	33.240885	11.760232	21.000000	24.000000	29.000000	41.000000	81.000000

Distribution of Diabetes

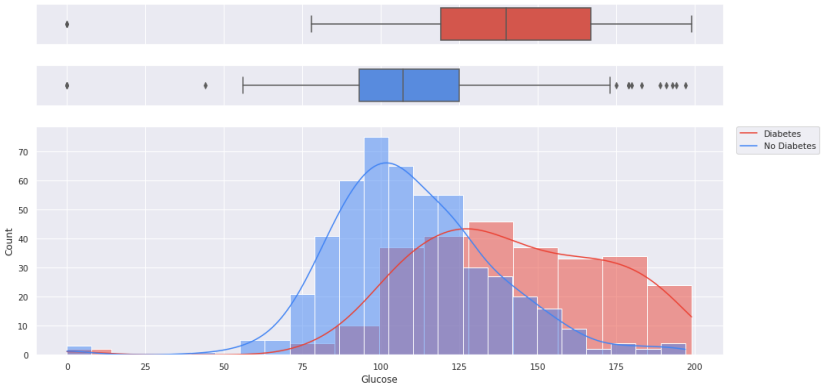


Correlation Matrix



Exploratory Analysis

- Glucose and diabetes



Insights

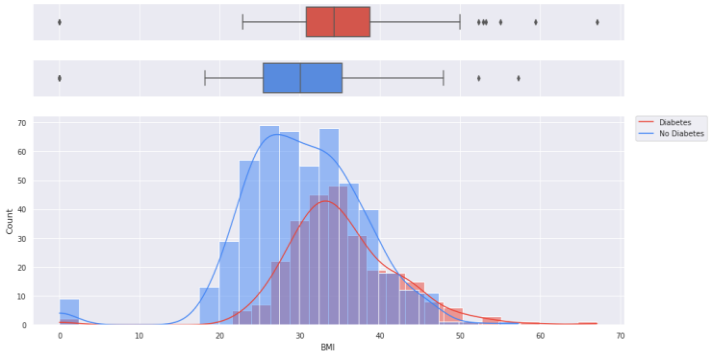
- Glucose has a correlation value of 0.467
- A higher glucose level usually mean a higher chance of diabetes

Skin Thickness and BMI to diabetes

Insights

- SkinThickness and BMI have a correlation value of 0.393
- A high SkinThickness usually means a higher BMI
- A high BMI means a higher chance of Diabetes

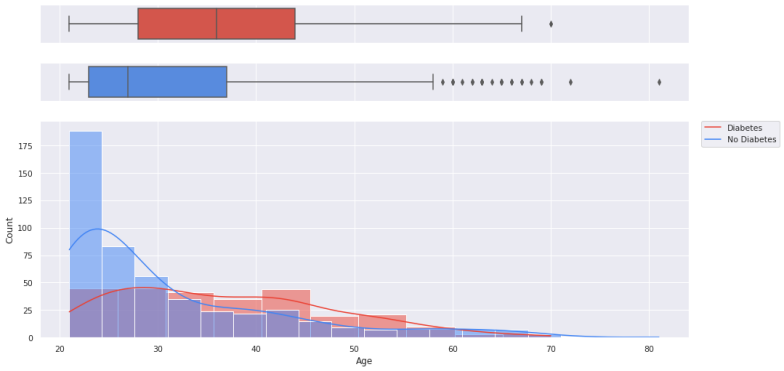
BMI and Diabetes



Insights

- BMI and Diabetes have a correlation value of 0.293
- A higher BMI usually means a higher chance of Diabetes

Age and diabetes



Insights

- Age and Diabetes have a correlation value of 0.238
- A higher age usually means a higher chance of diabetes

Data Preprocessing

- Normalizing continuous features

	min	mean	Max
Pregnancies	0.000000	3.845052	17.000000
Glucose	0.000000	120.894531	199.000000
BloodPressure	0.000000	69.105469	122.000000
SkinThickness	0.000000	20.536458	99.000000
Insulin	0.000000	79.799479	846.000000
BMI	0.000000	31.992578	67.100000
DiabetesPedigreeFunction	0.078000	0.471876	2.420000

All features are continuous, but they all have different ranges, so I am normalizing them to be between 0 and 1

```
for col in ['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness',
        'Insulin', 'BMI', 'DiabetesPedigreeFunction']:
   df[col] = df[col]/df[col].max()
```

Preparing Training and Validation arrays

Here I am creating arrays for features and labels
And splitting the dataset:

- 20% for validation
- 80% for training

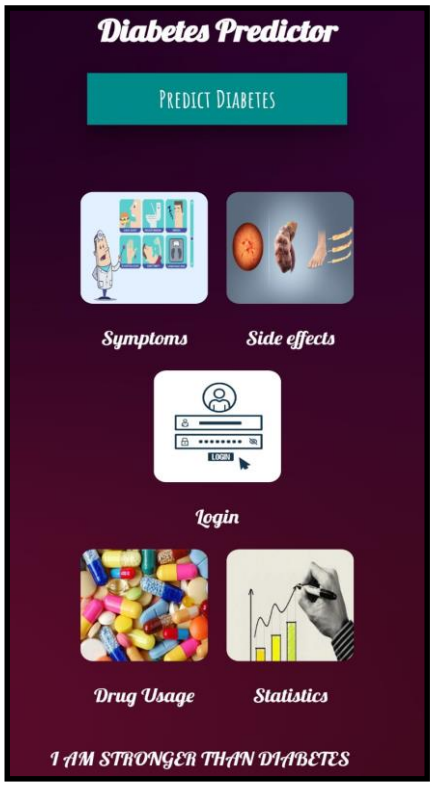
Coding of random forest model

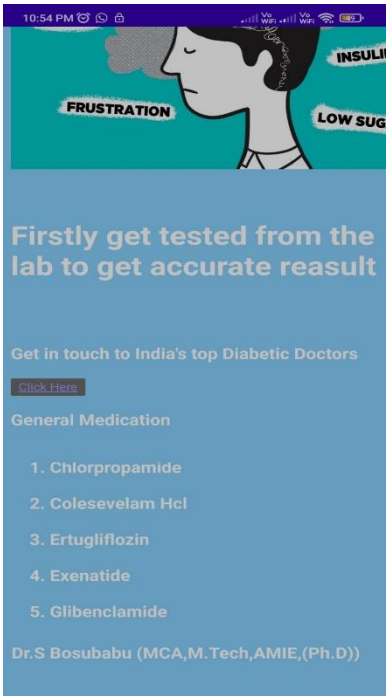
Model Building

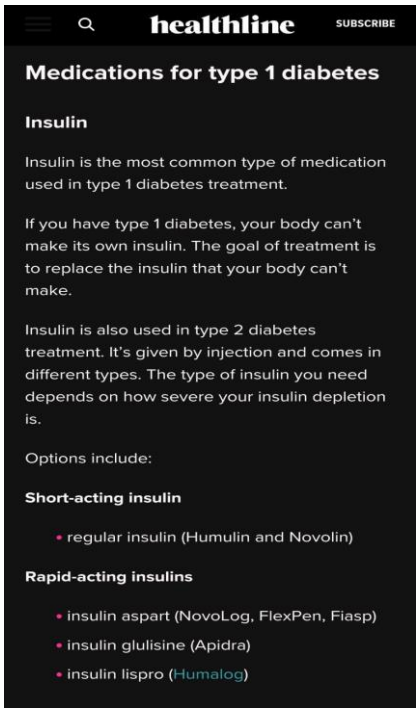
```
from sklearn.model_selection import train_test_split
X = df.drop(columns='Outcome')
y = df['Outcome']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20, random_state=0

# Creating Random Forest Model
from sklearn.ensemble import RandomForestClassifier
classifier = RandomForestClassifier(n_estimators=20)
classifier.fit(X_train, y_train)
```

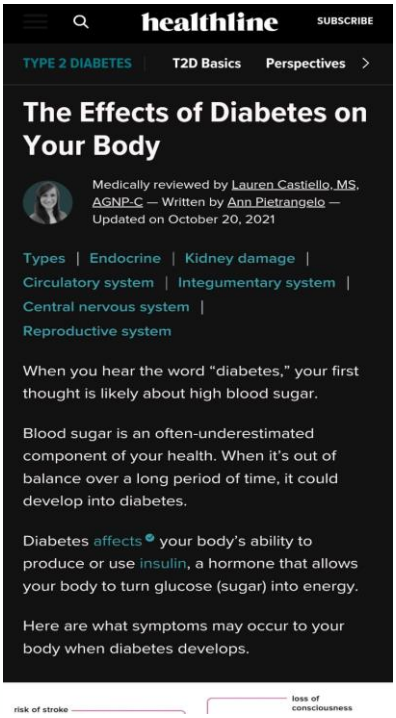
V. RESULTS

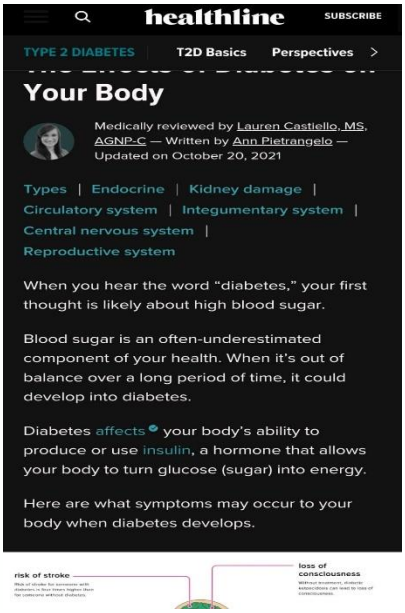
Application View on Mobile APP



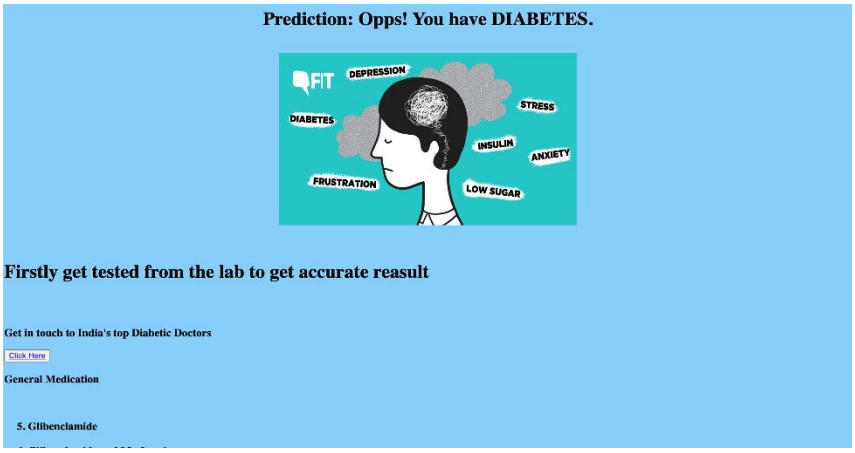








Application View on Website



Prediction: Wow ! You DON'T have diabetes.

11 Ways to Prevent from Diabetes

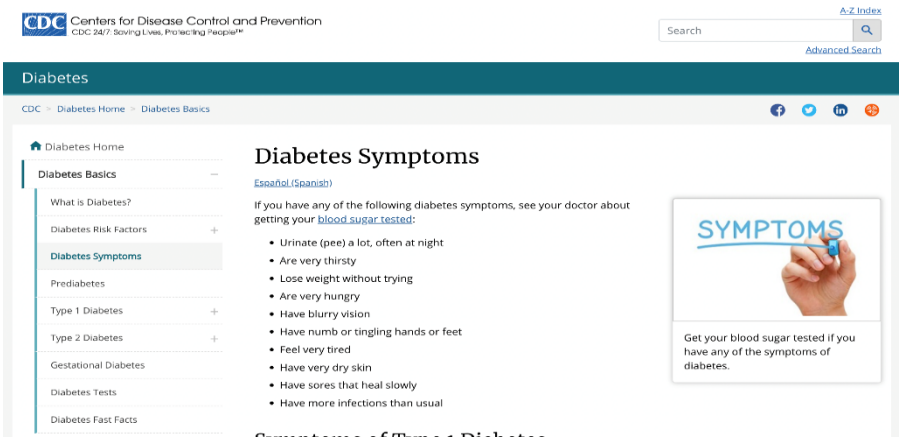
Diabetes is a chronic medical condition that affects millions of people worldwide. Unmanaged diabetes may lead to bli ndness, kidney failure, heart disease, and other serious conditions.

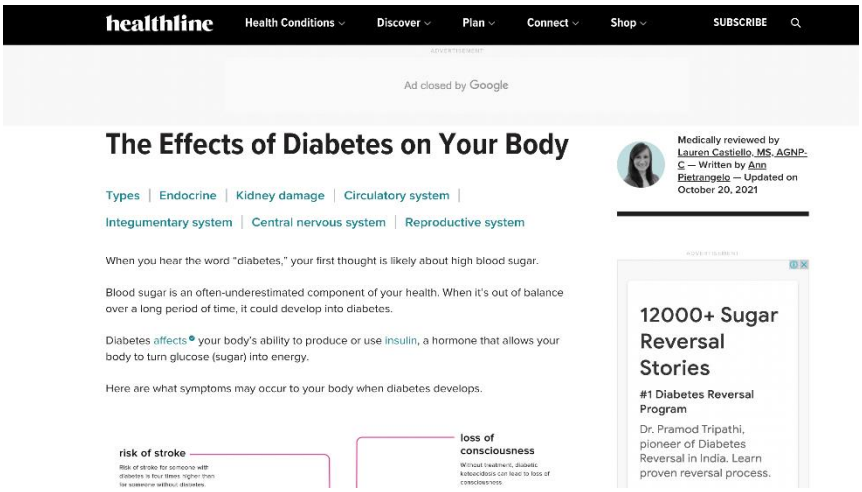
Before diagnosis, your blood sugar levels may be high — but not high enough to indicate diabetes. This is known as prediabetes. Taking a test like this one T rusted Source can help you figure out your risk factors for this condition.

It's estimated that up to 37% of people with untreated prediabetes develop t ype 2 diabetes within 4 years

Progressing from prediabetes to diabetes isn't inevitable. Although you can' t change certain factors like your genes or age, several lifestyle and dietary modifications may reduce your risk.

Here are 11 ways to lower your risk of getting diabetes.





International Diabetes Federation

ABOUT DIABETES WHO WE ARE NETWORK ACTIVITIES E-LIBRARY NEWS EVENTS

HOME > ABOUT DIABETES > WHAT IS DIABETES > FACTS & FIGURES

SHARE THIS PAGE

EN ES FR

Diabetes facts & figures

Diabetes facts & figures

LAST UPDATE: 09/12/2021

The IDF Diabetes Atlas Tenth edition 2021 provides the latest figures, information and projections on diabetes worldwide.

In 2021,

- Approximately **537 million adults** (20-79 years) are living with diabetes.
- The total number of people living with diabetes is projected to rise to **643 million by 2030** and **783 million by 2045**.
- **3 in 4** adults with diabetes **live in low- and middle-income countries**
- **Almost 1 in 2 (240 million)** adults living with diabetes are undiagnosed
- Diabetes caused **6.7 million deaths**
- Diabetes caused at least **USD 966 billion dollars** in health expenditure – 9% of total spending on adults
- **More than 1.2 million children and adolescents** (0-19 years) are living with type 1 diabetes
- **1 in 6 live births** (21 million) are affected by diabetes during pregnancy
- **541 million** adults are at increased risk of developing type 2 diabetes

Download the IDF Diabetes Atlas 10th Edition 2021 and other resources at www.diabetesatlas.org.

The result is we have developed an app and a website that can be used to predict diabetes using the random forest algorithm with an accuracy of 76 percent

VI. CONCLUSION

The prediction of diabetes become so fast and easy with the availability of our program and also there is a provision for checking symptoms, drugs, side effects, and statistics of diabetes finally the prediction function used for prediction asks several questions and those are Number of pregnancies, Glucose level, Blood pressure, Skin thickness, Insulin level, Body mass index, Diabetes pedigree function, Age. Based on these outcomes our ML model predicts whether the diabetes is there or not With an accuracy of 76%

VII. REFERENCES

[1]. " Diabetes Prediction using Machine Learning Algorithms" Aishwarya Mujumdar, Dr. Vaidehi V, International Conference On Recent Trends In Advanced Computing 2019. 10.1016/J.Procs.2020.01.047
[2]. M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of
[3]. different machine learning classifiers," IEEE Access, vol. 8, 2020.
[4]. analyticsvidhya.com/blog/2022/01/diabetes-prediction-using-machine-learning/
[5]. https://www.kaggle.com/code/mushfirat/diabetes-eda-model-comparison
[6]. https://www.healthline.com/health/diabetes
[7]. https://idf.org/