Project Report

On

Diabetes Prediction System using Machine Learning

Submitted to

Sant Gadge Baba Amravati University, Amravati

In recognition to partial fulfillment of the requirement

For the award of degree in

Bachelor of Engineering

(Computer Science and Engineering)

By

Amol Sudokar **Vaishnavi Khare**

Trupti Ambuskar **Aniket Dudhe**

B.E. CSE (VIII Semester)

Under the Guidance of

Dr. D. G. Harkut
(Guide & Head of Department)

Department of Computer Science & Engineering

**Prof Ram Meghe College of Engineering and Management,
Badnera-Amravati**

2022-2023

CERTIFICATE

This is to certify that the Seminar Report on

Diabetes Prediction System using Machine Learning

is a bonafide work submitted by

Amol Sudokar **Vaishnavi Khare**

Trupti Ambuskar **Aniket Dudhe**

B.E. CSE (VIII Semester)

In recognition to partial fulfillment of the requirement for the award of degree in

Bachelor of Engineering in Computer Science & Engineering

To

Sant Gadge Baba Amravati University, Amravati

During the academic year 2022-2023 Under the guidance of Dr. D. G. Harkut

Dr. D. G. Harkut
(Guide & Head of Department)

Signature of Internal Examiner *Signature External Examiner*

Department of Computer Science & Engineering

Prof Ram Meghe College of Engineering and Management, Badnera-Amravati

2022-2023

ACKNOWLEDGEMENT

The acknowledgement of this project is given to all the people who have helped us in completing this Project Report. We would like to extend my sincere thanks to all of them for their help, support and encouragement which has enabled us to complete this Project report named **Diabetes Prediction System using Machine Learning** successfully.

We are highly indebted to **Dr. D. G. Harkut** Head of the Department, and Project Guide for his guidance, valuable suggestion and constant supervision as well as for providing necessary information regarding the Project report. We are thankful for his support in completing the report.

We would like to thank all those who have helped us in preparing this Project report. We would like to thank all staff members of Department of Computer Science & Engineering (CSE), without their kind co-operation and guidance, this report would not have been possible.

Lastly, we would like to thank the almighty and our parents for moral support and friends with whom we share our day-to-day experience and receive lots of suggestion that improve our quality of work.

Amol Sudokar (Roll. No.70 Sec-A)
Vaishnavi Khare (Roll. No. 39 Sec-A)
Trupti Ambuskar (Roll. No. 03 Sec-A)
Aniket Dudhe (Roll. No. 18 Sec-A)

B. E. (CSE) VIII Semester

CONTENTS

LIST OF FIGURES

ABSTRACT

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Diabetes is a disease caused due to the increase level of blood glucose.

This high blood glucose produces the symptoms of frequent urination, increased thirst, and increased hunger. Diabetes is a one of the leading cause of blindness, kidney failure, amputations, heart failure and stroke. When we eat, our body turns food into sugars, or glucose. At that point, our pancreas is supposed to release insulin. Insulin serves as a key to open our cells, to allow the glucose to enter and allow us to use the glucose for energy. But with diabetes, this system does not work.

Type 1 and type 2 diabetes are the most common forms of the disease, but there are also other kinds, such as gestational diabetes, which occurs during pregnancy, as well as other forms. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience.

The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques.

The algorithms like K nearest neighbor, Logistic Regression, Random forest, Support vector machine and Decision tree are used. The accuracy of the model using each of the algorithms is calculated. Then the one with a good accuracy is taken as the model for predicting the diabetes.

Big Data Analytics plays a significant role in healthcare industries. Healthcare industries have large volume databases. Using big data analytics one can study huge datasets and find hidden information, hidden patterns to discover knowledge from the data and predict outcomes accordingly.

In existing method, the classification and prediction accuracy is not so high. In this paper, we have proposed a diabetes prediction model for better classification of diabetes which includes few external factors responsible for diabetes along with regular factors like Glucose, BMI, Age, Insulin, etc.

Classification accuracy is boosted with new dataset compared to existing dataset. Further with imposed a pipeline model for diabetes prediction intended towards improving the accuracy of classification.

CHAPTER 1

INTRODUCTION

Diabetes is rapidly growing nowadays in individuals, particularly young peoples and become major challenge for the researcher, scientist and educationist. It is not only a disease but also a creator of different kind of disease like heart attack, blindness, kidney diseases, etc.

The main reason of diabetes is increase in the amount of sugar in the blood. The normal identifying process is that patients need to visit a diagnostic center, consult their doctor, and sit tight for a day or more to get their report.

The number of reports of diabetic patients is escalating day by day, due to innumerable reasons toxic or chemical contents mixed with the food, obesity, working culture and bad diet plan, unusual life style, eating food habits and environmental factors. Hence diagnosing of diabetes is essential to save the human lives.

Machine learning techniques can be used to develop an efficient healthcare system to predict a different type of diabetes diseases in advance.

1.1 Basic Definitions:

Diabetes is noxious diseases in the world. Diabetes caused because of obesity or high blood glucose level, and so forth. It affects the hormone insulin, resulting in abnormal metabolism of crabs and improves level of sugar in the blood. Diabetes occurs when body does not make enough insulin.

Population of India is now more than 100 million so the actual number of Diabetics in India is 40 million. Diabetes is major cause of death in the world. Early prediction of disease like diabetes can be controlled and save the human life. To accomplish this, this work explores prediction of diabetes by taking various attributes related to diabetes disease. For this purpose we use the Pima Indian Diabetes Dataset, we apply various Machine Learning classification and ensemble Techniques to predict diabetes. Machine Learning Is a method that is used to train computers or machines explicitly.

Various Machine Learning Techniques provide efficient result to collect Knowledge by building various classification and ensemble models from collected dataset. Such collected data can be useful to predict diabetes.

The WHO (World Health Organization) reported that around 1.6 million people die due to diabetes every year. Diabetes is one kind of disease that occurs when the blood glucose/blood sugar level in the human body is very high. According to health experts, diabetes occurs when the human body's gland called the pancreas cannot produce enough insulin (Type 1 diabetes), and the produced insulin cannot be used by the cell of the body (Type 2 diabetes).

When we eat food, after the digestion process, glucose gets released. Insulin is a blood hormone that moves from blood to cells and instructs cells to consume blood glucose and transform it into energy. When the pancreas cannot produce enough insulin, the cells cannot absorb glucose, and the glucose remains in the blood. Hence the blood glucose/blood sugar increases in the blood at a very unacceptable level . Due to high blood sugar, some symptom arises in the human body, such as extreme hunger, intense thirst, and frequent urination. The usual range of glucose levels in the human body is 70 to 99 mg per deciliter. If the glucose level is more than 126 mg/dl, it indicates diabetes. A person is considered to have prediabetes if body glucose concentration is 100 to 125 mg/dl.

1.2 Basic Concepts:

Diabetes can be divided into two classes: Type 1 and Type 2

Type 1 affects 10% of everyone with diabetes. While type 2 diabetes affects about 90%. The main thing to remember is that both are as serious as each other. Having high blood glucose (sugar) levels can leads to serious health complications, no matter whether you have type 1 or type 2 diabetes.

In Type 1, your body attacks the cells in your pancreas which means it cannot make any insulin. Type 1 diabetes can appear at any age but it's most common among children and adolescents.

In Type 2, the body still produce insulin, but it's unable to use it effectively. Type 2 diabetes can appear at over age of 45, have a lots of belly fats, are carrying access weight or have obesity. So, if you have either conditions you need to take the right steps to manage it.

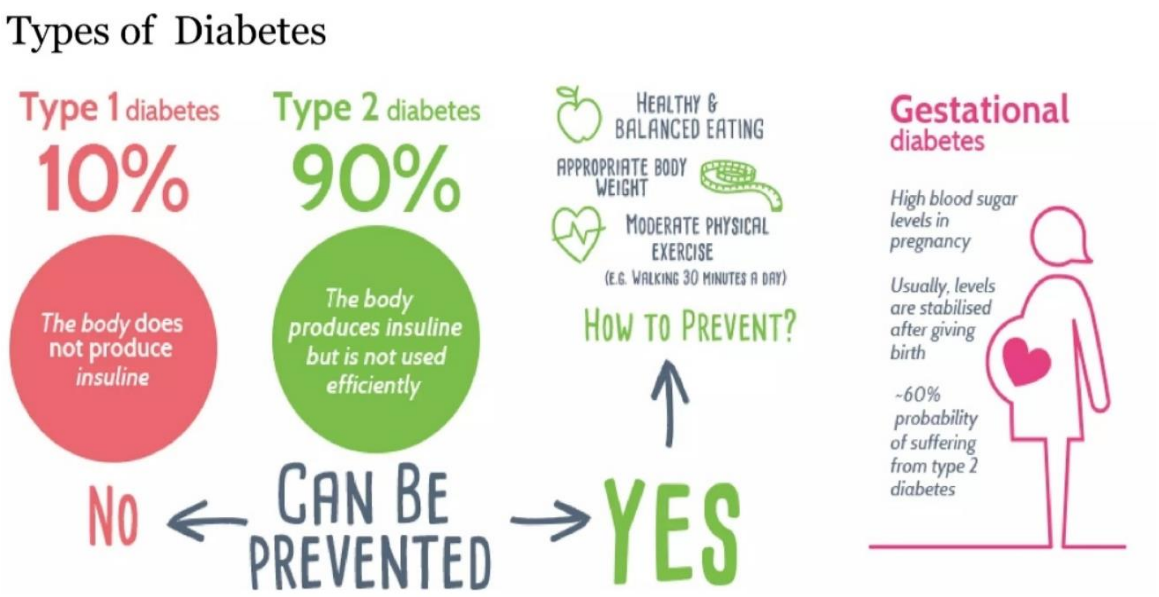

Types of Diabetes
Type 1 diabetes
10%
The body does not produce insuline
Type 2 diabetes
90%
The body produces insuline but is not used efficiently
HEALTHY & BALANCED EATING
APPROPRIATE BODY WEIGHT
MODERATE PHYSICAL EXERCISE
(E.G. WALKING 30 MINUTES A DAY)
HOW TO PREVENT?
NO
CAN BE PREVENTED
YES
Gestational diabetes
High blood sugar levels in pregnancy
Usually, levels are stabilised after giving birth
~60% probability of suffering from type 2 diabetes

Fig 1.1: Types of Diabetes

Are there different risk factors for type 1 and type 2?

We don’t know exactly what causes type 1 or type 2 diabetes, but we do know the different risk factors. So we know why you might be likely to get one type over the other. Even though we know this, it’s good to remember these aren’t set in stone.

Type 1

A big difference between the two is that type 1 isn’t affected by your lifestyle. Or your weight. That means you can’t affect your risk of developing type 1 by lifestyle changes.

People up to the age of 40 are more likely to be diagnosed with it, especially children. In fact, most children with diabetes have type 1. But, although it’s less common, people over 40 can also be diagnosed with it.

Type 2

It’s different for type 2 diabetes. We know some things put you at more risk:

- your family history
- ethnic background
- age
- If you’re overweight or obese.

We also know that there are things you can to reduce your risk of developing type 2 diabetes. Things like eating healthily, being active and maintaining a healthy weight can help you to prevent type 2.

You’re also more likely to get type 2 if you’re over 40. Or if you’re South Asian, if you’re over 25. But type 2 is also becoming more common in younger people. More and more children and young people get diagnosed with type 2 in the UK each year.

1.3 Purpose

The aim of the project is to determine the appropriate classification model or algorithm that gives the best accuracy results ever possible. So, that algorithm proven to be the best can then be used in the prediction of diabetes to figure out if a person is diabetic or non-diabetic so far. This is to avoid any kind of misconceptions due to the incompetent classification algorithm or model can cause if the best one is not chosen.

It is also one of the most chronic diseases in India or elsewhere, the prediction of this in the early stage or even before should be able to control and contain it more easily maybe with a proper diet or a less severe treatment. The Type 2 diabetes has a much stronger link to family history or lineage than the Type 1 diabetes. So, if a member of a family has Type 2 diabetes it is likely that any member of the family could possess the same, so it has to be eliminated before it gets too complicated.

1.4 Scope

Type 2 diabetes is very different from Type 1 diabetes, which was previously called as insulin dependent diabetes mellitus (IDDM). Before the 2000s Type 2 diabetes was considered a disease of elderly and middle-aged individuals (hence it was also called adult onset diabetes). But once it started to show up on teenagers the name faced away as it no longer was confined to middle-aged or adults.

According to U.S. NIH, Type 2 diabetes contributes to the respective conditions directly:

1. Stroke and Heart diseases. Adults who are subjected to diabetes die due to heart diseases 2x to 4x times than those of adults who are not subjected to diabetes. The risk of the stroke the take place is also 2x to 4x times than those adults who are not subjected to diabetes.

2. Nervous system disease. Half of the population with diabetes feel impaired sensation, pain in the hands or feet, carpal tunnel syndrome, slower digestion, and many other nervous problems.

3. Possess High blood pressure. Most of the adults have blood pressure that goes higher than 130/80 mmHg.

4. Blindness. It is one of the new causes of being diabetic for the ones who are between the ages of 20 to 74.

5. Amputation. 60% of this occurs among the people with diabetes, non-traumatic lower limb amputation.

6. Kidney disease. One of the leading cause of kidney failure. About 150,000 individuals having diabetes survive on chronic dialysis or due to a kidney transplant.

7. Immune system disorder. Individuals with diabetes have fewer abilities to fight or reject viral and bacterial infections. They have more chances of dying of influenza or pneumonia from the individuals who do have diabetes.

8. Pregnancy complication. Mothers having diabetes having a greater number of abortions, and their babies tend to have a greater risk of major born defects and of being susceptible to diabetes later in their life.

Diabetes is a disease that is considered to be as one of the leading causes of death In India, 72 Million diabetes cases were recorded in 2017 and this is expected to double by 2025. This poses a serious public Health Issue in a country where population keeps increasing

Exponentially every year. Among the Indian states, Tamil Nadu has been having the highest Death Rates from Diabetes. Diabetes often leads to long term disabilities and complications. It leads to Heart Attacks, Kidney Failure, Blindness and Gestational Diabetes causes birth defects to the new born babies.

Around 1.95 Lakh Crores will be needed as the Annual Cost to treat Diabetes. Urban Poor in India spend 34% of their income on Diabetes Treatment.

These trends indicate that there is a rise in premature death and this is a major threat to global development. Technological advancements have been useful in reducing hyperglycaemia. But irrespective of all of these Technological Advancements, Diabetes still poses a serious threat to life.

We aim to perform Prediction and analysis on a PIMA Dataset that can be used to find the efficiency and accuracy. This can be used to find the most suitable algorithm and the one that has the highest accuracy. We split the dataset into 4 different splits namely: 60 / 40, 70 / 30, 80 / 20, 65 / 35. The Input Training and Test Data is fitted to the model and we then classify the Training data into different arrays for the purpose of prediction. We then find the accuracy by comparing the predicted values with the original set of values that we have.

1.5 Motivation

Diabetes is indeed one of the chronic health problems that are devastating and with preventable consequences. The driving agents would be high blood glucose levels due to low insulin production. Type 2 diabetes affects men and women proportionately, around 12 million men and 11.5 million women have diabetes. To improve the quality of life means to take one’s own diabetes into control, for which additional support and education need to be provided to the patients.

Though the technology has evolved and new treatments are found in controlling diabetes, the challenges of self-comprehension are the most overwhelming for most of the individuals. It demands individual patient self-management that includes monitoring the blood glucose levels, maintaining a healthy diet, taking medication and regularly exercising. There are high non-compliance patterns in self-management behaviour’s, this could be usually due to the changes that are encountered in the patient's routine life. To adapt and inherit to such changes the patients are usually motivated to achieve their goals and them their new way of living to create a long living life which allows them to manage diabetes. The support, assistance, and feedback play a very important role in achieving self-management goals. Organizations, where there are peer diabetes people support groups, are a valuable source for the patients as they can cling on to the mutual changes and awareness of themselves.

CHAPTER 2

LITERATURE REVIEW

2.1 Classification of Studied Literature:

Raja Krishnamurthy proposed a diabetes healthcare disease prediction framework using machine learning techniques. The dataset contains 768 rows and 9 columns and 90% of the data is used for training and 10% used for the testing purpose and they performed hyper-parameter tuning to evaluate the Machine Learning models and used to increase the accuracy. Out of 5 algorithms best one is identified and hyper parameter tuning has been applied to provide better accuracy as a result of 86%.

Sonu Kumari and Archana Singh proposed an intelligent and effective methodology for the automated detection of Diabetes Mellitus using Neural Network. The paper approached the aim of diagnoses by using ANNs and demonstrated the need for pre-processing and replacing missing values in the dataset being considered. Through the Modified training set, a better accuracy was achieved with lesser time required for training the set. Sajida by using CPCSSN (Canadian primary care sentinel surveillance Network) dataset and three machine learning methods to predict the diabetes Disses (DD) in early stage to safe human life at from early death .On this study Bagging, Adaboost, and decision tree (J48) were used to predict the diabetes and the researcher was compare the result of those methods and concluded that Adaboost method was provide effective and better accuracy than the other methods in weka data mining tools. Sadri used Naive Bayes, RBF Network and J48 datamining algorithms for diagnosing type II diabetes. They used WEKA tool. Finally they found Naive Bayes, having the accuracy rate of 76.96% than other algorithms.

In this paper, Prediction of diabetes is done using ensemble voting classifiers for pima Indian diabetes dataset, in comparison with different classification algorithms, the highest accuracy of 80% and 81% is achieved for data set by using 10-fold cross validation and by spitting data into 30% testing and 70% training.

J. Pradeep Kandhasamy, S. Balamurali’s research study compare the performance of algorithms those are used to predict diabetes using data mining techniques. Also authors classifiers J48 Decision Tree, K Nearest Neighbors, and Random Forest, Support Vector Machines to classify patients with diabetes mellitus. Authors compared four prediction models for predicting diabetes mellitus using 8 important attributes under two different situations. One is before pre-processing the dataset. Here the studies conclude that the decision tree J48 classifier achieves higher accuracy of 73.82 % than other three classifiers.

After pre-processing, the dataset given more accurate result when compared to the previous studies. In this case, both KNN (k=1) and Random Forest performance much better than the other three classifiers and they provide 100% accuracy. From this we can come to know that after removing the noisy data from our dataset it will provide good result for our problem.

Preeti Verma, Inderpreet Kaur, Jaspreet Kaur paper work, the performance of this method is evaluated using 10-fold cross validation accuracy, confusion matrix. The obtained classification accuracy using 10-fold cross validation is 96.58% in comparison with other spline SSVM technique. The results of this study showed that the modified spline SSVM was effective to detect diabetes disease diagnosis and this is very promising result compared to the previously reported results.

Dr. B .L. Shivkumar and S Thiyagarajan work, an effective machine learning algorithm is proposed for the classification of type dm patients. This machine learning algorithm used for classification will find the optimal hyper-plane which divides the various classes. Sneha and Tarun proposed a method that aims to focus on selecting the attributes that ail in early detection of Diabetes Miletus using Predictive analysis. The result shows the decision tree algorithm and the Random forest has the highest specificity of 98.20% and 98.00%, respectively holds best for the analysis of diabetic data. Naïve Bayesian outcome states the best accuracy of 82.30%. The research also generalizes the selection of optimal features from dataset to improve the classification accuracy.

This paper focuses that the use of data mining algorithms can be very helpful in early prediction and in consequence early precautions before the diagnosis of disease. The main goal of this paper is to provide a comparison and suggest best algorithm which can be used for the pattern recognition or prediction in healthcare fields. After the implementations of these algorithms it can be said that for PID dataset Decision Tree gives best accuracy 75.65%. The tool used for testing and validation is Rapid Miner while all algorithms worked with 70:30 ratio for training and testing.

2.2 Analysis of Studied Literature:

Sr No.	Originator With Title	Dataset	Algorithm	Tool	Outcome & Accuracy
1.	Aishwarya Iyar, S. jeyalatha and RonakSumbaly, "Diagnosis Of Diabetes Using Classification Mining Techniques"	Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases	Decision tree and Naïve Bayes algorithm	WEKA	The Naïve Bayes algorithm is obtained 79.5652% of accuracy
2.	AkankashaRathore, Simran Chauhan, SakshiGujral, "Detecting and Predicting Diabetes Using Supervised Learning: An Approach towards Better Healthcare for Women"	Pima	SVM and Decision Tree	R-Studio	SVM gives the 82% of accuracy
3.	Salim Amour Diwani studied all patients data are trained and tested using 10 cross validation, then performance was evaluated, investigated and compared with other classifications algorithms	Pima	Naïve Bayes and Decision tree algorithm	WEKA	The result predicted that the best algorithm is Naïve Bayes with an accuracy of 76.3021%
4.	N. P. Tigga applied logistic regression on PIDD for diabetes Prediction	Pima Indians Diabetes dataset	Logistic Regression algorithm	R-Studio	Good prediction with an accuracy of 75.32%

Table 2.1: Literature Review

CHAPTER 3

PROBLEM DEFINATION AND REQUIREMENT ANALYSIS

3.1 Problem Domain and Definition:

Diabetes is a most common disease caused by a group of metabolic disorders. It is also diabetic mellitus. It affects the organs of human body. It can be controlled by predicting this disease earlier.

Now a days, healthcare industries generating large volume of data. Machine learning algorithms and statistics are used to predict the diseases with the help of current and past data.

Machine learning techniques helps the doctors to predict early stage for diabetics. Diabetic’s patient’s medical record and different types of algorithms are added in datasets for experiential analysis.

We can used Random forest, Decision tree, Support vector machine (SVM) and Logistic regression to predict whether a patient has diabetes based on diagnostic measurements.

Performance and accuracy of the applied algorithms is discussed and compare.

Statement of the problem

Patients with the potential of diabetes have to go through a series of tests and exams to diagnose the disease properly. These tests might embody redundant or inessential medical procedures that cause complications and wastage of time and resources. The burden of this sickness on the economy way exceeds the direct medical prices within the care sector because diabetes reduces the standard of life and hinders labour productivity. The absence of a correct diagnosis scheme, deficiency of economic means, and a general lack of awareness represent the main reasons for these negative effects. Hence, preventing the sickness altogether through early detection may doubtless cut back a considerable burden on the economy and aid the patient in diabetes management.

3.2 Requirement Analysis

We will be predicting that whether the **patient has diabetes or not** on the basis of the features we will provide to our machine learning model, and for that, we will be using the famous Pima Indians Diabetes Database.

1. **Data analysis:** Here one will get to know about how the data analysis part is done in a data science life cycle.

2. **Exploratory data analysis:** EDA is one of the most important steps in the data science project life cycle and here one will need to know that how to make inferences from the visualizations and data analysis
3. **Model building:** Here we will be using 4 ML models and then we will choose the best performing model.
4. **Saving model:** Saving the best model using pickle to make the prediction from real data.

3.2.1 Functional Requirements

In order for every software application to run properly, it needs to satisfy a lot of functions that are to be deployed in it. These functions are nothing but various operations that are performed in each step while developing the application. This step comes under the best practices of developing an application. Functional and Non-Functional Requirements together set a list of rules that govern the smooth running of an application and it also helps the developer and the user to determine the software and hardware requirements that are needed to run the application. Functional Requirements that are required are:

Python:

Python programming language was developed in the year 1991 by Guido Van Rossum. The syntaxes used in the language makes it very comfortable and easier for developers to work with. Because of this very reason, this programming language can be used both in small and large scale. They are dynamic and garbage collected.

Numpy:

Numpy is a universally useful array processing package. It gives an elite multidimensional cluster object, and devices for working with these arrays. It is the principal package for logical processing with Python.

Matplotlib:

Matplotlib is a stunning perception library in Python for 2D plots of arrays. Matplotlib is a multi-stage information perception library based on NumPy arrays and intended to work with the more extensive SciPy stack. It was presented by John Hunter in the year 2002.

3.2.2 Non-Functional Requirements

Non-functional requirements are used to set conditions to monitor the performance characteristic of the application. It describes how a specific function in the application works. They also determine the overall quality of the project and hence it is a very important aspect in any software development process. The Non-Functional Requirements include

1. **Usability:** It refers to the easiness of the application of models and determines the ease with which it can be used by the user. Usability can be said to be high when the knowledge required to use the models is less and the efficiency of its functionality is high. It is also a main criterion which can determine the accuracy of the results.
2. **Accuracy:** Accuracy determines the relative closeness of the value produced by the system to that of the ideal value. It is also one way to determine how the classification models works better compared to the other similar models.
3. **Responsiveness:** Responsiveness is determined by completing the software operations with minimal errors or no errors. It is directly proportional to the stability and the performance of the application. The Robustness and Recoverability can also be determined by this criterion.
4. **Scalability:** Scalability is used to determine the growth of the project. It determines how much room the application can have in order to include more features in the future. It determines the sustainability of the project.

Steps for develop project:

1. Installing the Libraries
2. Importing the Dataset
3. Filling the Missing Values
4. Exploratory Data Analysis
5. Feature Engineering
6. Implementing Machine Learning Models
7. Predicting Unseen Data
8. Concluding the Report

3.2.3 Statement of Scope

Healthcare professions found it hard to find healthcare data and perform analysis on them due to lack of tools, resources. But using ML, we can overcome this and can perform analysis on real-time data leading to better modeling, predictions. This enhances and improves overall healthcare services. Now, IoTs being integrated with ML in order to make smart healthcare devices that sense if there is any change in the person's body, health data when he uses the device, and this will notify the person regarding this through an app. This helps in easy monitoring, advanced prediction and analysis thereby reducing errors, saving time and life of people.

3.2.4 Aim of the Project

The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques. The algorithms like K nearest neighbor, Logistic Regression, Random forest, Support vector machine and Decision tree are used. The accuracy of the model using each of the algorithms is calculated. Then the one with a good accuracy is taken as the model for predicting the diabetes.

This research work aims to analyse the Diabetes dataset, design, and implement a Diabetes prediction and recommendation system utilizing machine learning classification techniques. The specific objectives of this project work are:
(i) To review existing literature along the area of diabetes diagnosis and prediction.
(ii) Design and develop a model using machine learning techniques.
(iii) To analyse the Diabetes dataset and use Support Vector Machine and Random forest algorithms or others to develop a prediction engine.
(iv) To identify and discuss the benefits of the designed system along with effective applications.

3.2.5 Objective to be achieved

The primary aim of this project is to analyses the Diabetes Dataset and use the Random Forest, Logistic Regression, Support Vector Machine, Naïve Bayes, K-Nearest Neighbors algorithms for prediction and to develop a prediction engine. The secondary aim is to develop a web application with following feature.

• Allow users to predict diabetes utilizing the prediction engine.

The objective is set to achieve the aims of the project through a Research on statistical models in machine learning and to understand how the algorithms works.

CHAPTER 4

PROPOSED APPROACH AND DESIGN

4.1 Proposed Approach

Classification is one of the most important decision-making techniques in many real-world problems. In this work, the main objective is to classify the data as diabetic or non-diabetic and improve the classification accuracy. For many classifications problem, the higher number of samples chosen but it doesn't lead to higher classification accuracy. In many cases, the performance of algorithm is high in the context of speed but the accuracy of data classification is low.

The main objective of our model is to achieve high accuracy. Classification accuracy can be increase if we use much of the data set for training and few data sets for testing. This survey has analyzed various classification techniques for classification of diabetic and non-diabetic data. Thus, it is observed that techniques like Support Vector Machine, Logistic Regression, and Artificial Neural Network are most suitable for implementing the Diabetes prediction system.

Diabetes is a most common disease caused by a group of metabolic disorders. It is also diabetic mellitus. It affects the organs of human body. It can be controlled by predicting this disease earlier. Now a days, healthcare industries generating large volume of data. Machine learning algorithms and statistics are used to predict the diseases with the help of current and past data. Machine learning techniques helps the doctors to predict early stage for diabetics. Diabetic's patient's medical record and different types of algorithms are added in datasets for experiential analysis.

We can used Random Forest, Decision tree, Support vector machine (SVM) and Logistic regression to predict whether a patient has diabetes based on diagnostic measurements. Performance and accuracy level of the applied algorithms is discussed and compare.

In this project we used Random Forest algorithm to predict the diabetes. Random forest algorithm used for both classification and regression problems in ML. It predicts output with high accuracy, even for the large dataset it runs efficiently. Random forest is a classifier that contains number of decision trees on various subsets of the given dataset and takes average to improve the predictive accuracy of that datasets. The greater number of trees in the forest leads to higher accuracy and prevents the problem of over fitting.

4.1.1 Block Schematic of the Approach

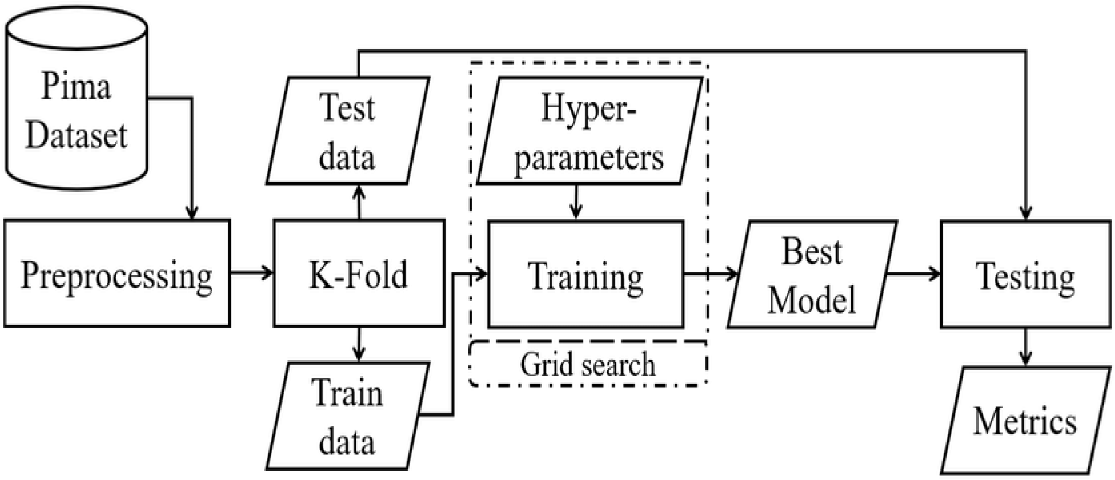


Figure 4.1: Workflow

The algorithms like K nearest neighbor, Logistic Regression, Random Forest, Support vector machine and Decision tree are used. The accuracy of the model using each of the algorithms is calculated. Then the one with a good accuracy is taken as the model for predicting the diabetes.

4.1.2 Algorithms

Now a days, healthcare industries generating large volume of data. Machine learning algorithms and statistics are used to predict the diseases with the help of current and past data.

We can used Random Forest, Decision tree, Support vector machine (SVM) and Logistic regression to predict whether a patient has diabetes based on diagnostic measurements.

1. **Logistic Regression: -** It is machine learning technic used when dependent variables are able to categorize. It used probabilistic estimations which helps in understanding the relationship between the dependent and one or more independent variable.
2. **K-nearest neighbors (KNN): -** It uses feature similarity to predict the values of new data points which further means that the new data point will be assigned a value based on how closely it matches the points in the training sets.
3. **Support Vector Machine (SVM): -** SVM are set of supervised learning methods used for classification, regression and outlier's detection. In SVM we have to identify the right hyper plane to classify the data correctly.

4. **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.
5. **Random forest: -** Random Forest is the popular supervised learning technique. It is a process of combing multiple classifiers to solve complex problem and to improve performance of the model.

❖ Random Forest Algorithm

In this project we used the Random Forest Algorithm to predict the diabetics of the patient.

- Random forest algorithm used for both classification and regression problems in ML.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- Random forest is a classifier that contains number of decision trees on various subsets of the given dataset and takes average to improve the predictive accuracy of that datasets.
- The greater number of trees in the forest leads to higher accuracy and prevents the problem of over fitting.

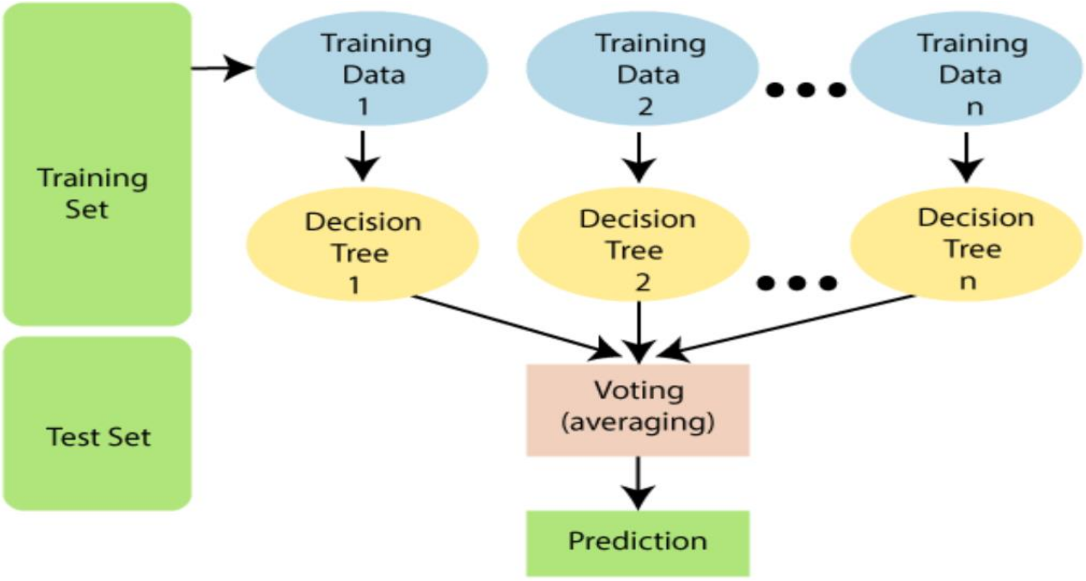


Figure 4.2: Workflow of Random Forest

Why we used Random Forest Algorithm?

- ✓ This algorithm takes less training time as compared to other algorithms.
- ✓ This algorithm predicts output with high accuracy, even for the large dataset it runs efficiently.
- ✓ In random forest algorithm combines a multiple trees to predict the class of dataset, it is possible that some decision trees may predict the correct output, while others may not.
- ✓ But together, all the trees predict the correct output. Therefore, below are two assumptions for better random forest classifier:
- ✓ There should be some actual values in the feature variables of the dataset so that classifier can predict accurate results rather than a guessed result.
- ✓ The predictions from each tree must have very low correlations.

Advantages of Random Forest

1. Random forest can balance the errors in unbalanced datasets. This helps in the reduction of errors while performing analysis.
2. It computes prototypes that provides information between the classification of the variables and the variables itself. It classifies the variables that are important for the classification.
3. Larger datasets can be analysed with the help of Random Forest as it tends to give better results.
4. Missing data in a dataset can be estimated with the help of Random Forest and also helps in maintaining the accuracy of the result while large sets of data are missing.
5. Variable interactions can be detected using experimental methods of Random Forest.
6. The learning of the algorithm is very fast and it trains very fast. If we split the data into training and testing data, then the data is learned by the algorithm very quickly to the point where the accuracy of the results are efficient compared to any other machine learning algorithm.
7. It does not delete any variables while handling thousands of variable inputs. This helps in better accuracy of the prediction results.

4.2 Data Flow Diagram:

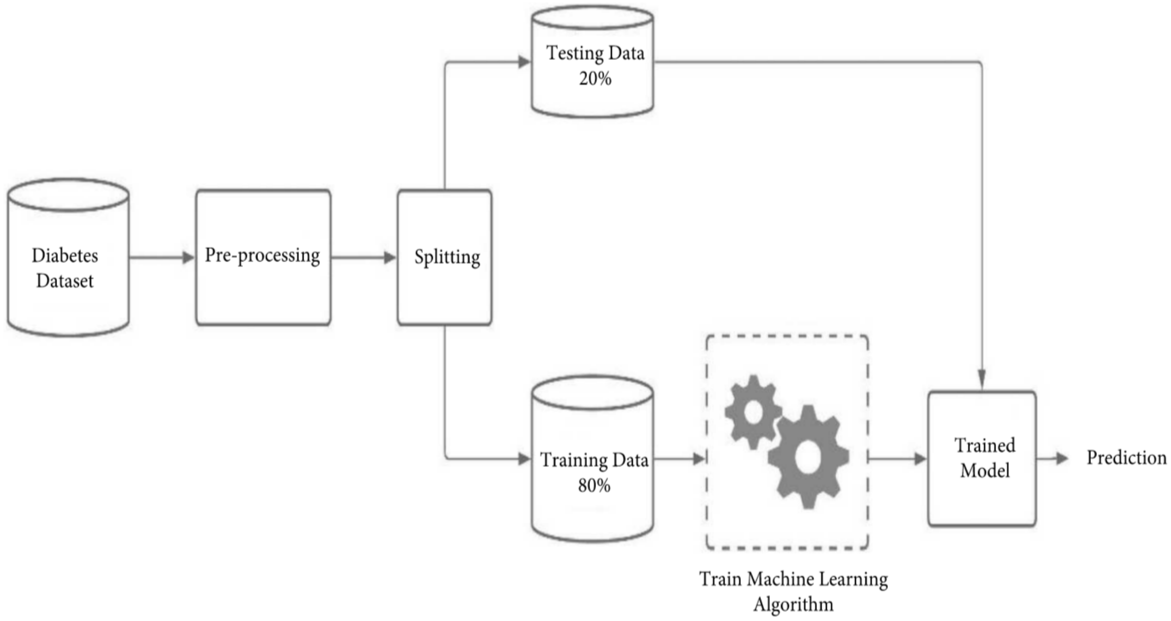


Fig 4.3: Data Flow Diagram

- **Dataset collection: -** It includes collection of understanding data to study the hidden patterns and trends which helps to predict and evaluating the results. Dataset carries number of rows (i.e., total number of data) and number of columns (i.e., total number of attributes).
- **Data Pre-processing: -** This phase of model handles inconsistent data in order to get a more accurate and precise result like in this dataset is inconsistent so we dropped the feature. This dataset doesn't content missing values. So, we imputed missing value for few selected attributes.
- **Missing Value Identification: -** Using the panda's library and SK-learn, we got the missing values in datasets. We replace this missing value with the corresponding mean value i.e., zero.
- **Scaling and normalization: -** We performed features scaling by normalizing the data from 0 to 1 range, which boosted the algorithms calculations speed.
- **Feature Selection: -** Pearson's correlation method is popular method to find most relevant attributes/features. The coefficient value remains in a range between -1 and 1. The value above 0.5 and below -0.5 indicates a notable correlation, and the zero value means no correlation.

4.3 Architectural Flow Diagram:

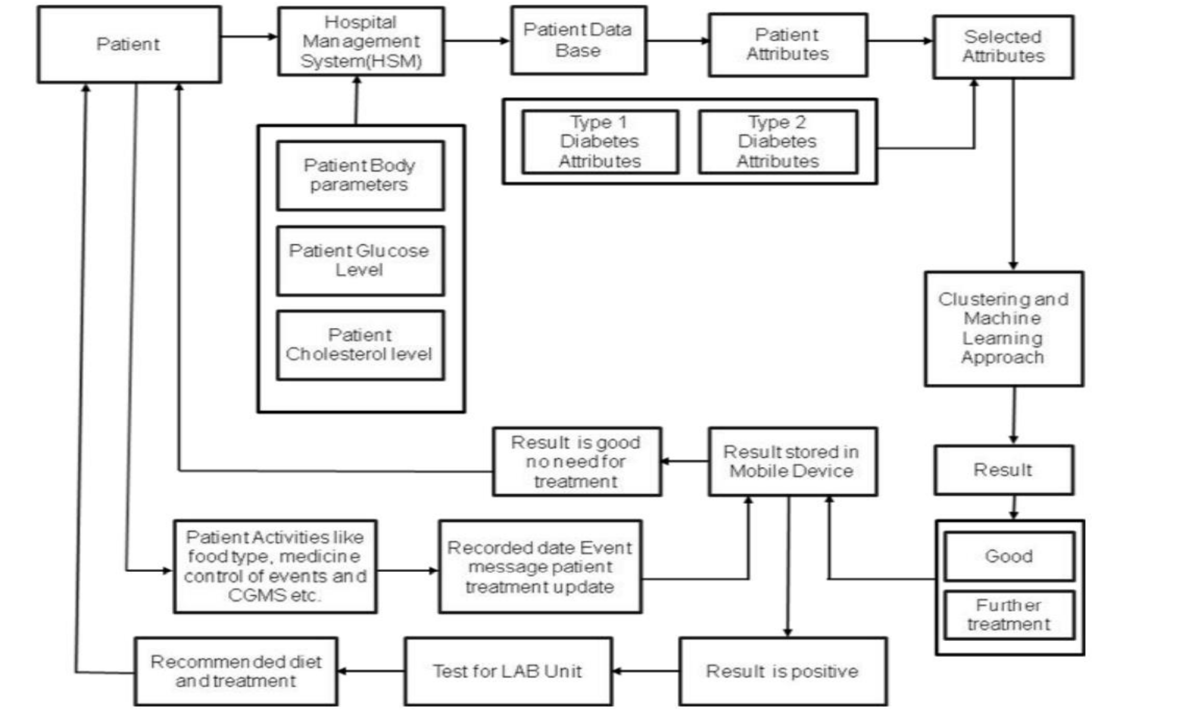

Patient
Hospital Management System(HSM)
Patient Data Base
Patient Attributes
Selected Attributes
Patient Body parameters
Patient Glucose Level
Patient Cholesterol level
Type 1 Diabetes Attributes
Type 2 Diabetes Attributes
Clustering and Machine Learning Approach
Result is good no need for treatment
Result stored in Mobile Device
Result
Good
Further treatment
Patient Activities like food type, medicine control of events and CGMS etc.
Recorded date Event message patient treatment update
Recommended diet and treatment
Test for LAB Unit
Result is positive

Fig 4.4: Architectural Flow Diagram

The algorithms like K nearest neighbor, Logistic Regression, Random Forest, Support vector machine and Decision tree are used. The accuracy of the model using each of the algorithms is calculated. Then the one with a good accuracy is taken as the model for predicting the diabetes.

The prediction diabetes is plays very important role for the human life because it leads to death. The off erred system is used to initial detection of diabetes and time prediction whereas time prediction means when the patients the diabetes it will be help to improve the habit of the patients. The proposed system is mainly concentred on development of machine learning model and also it helpful in the medical sector to identify the diseases. This offer system is an automation to predict the diabetes using old patient’s data.

Designing of system is the process in which it is used to define the interface, modules and data for a system to specify the demand to satisfy. System design is seen as the application of the system theory. The main thing of the design a system is to develop the system architecture by giving the data and information that is necessary for the implementation of a system.

4.4 Control Flow Diagram:

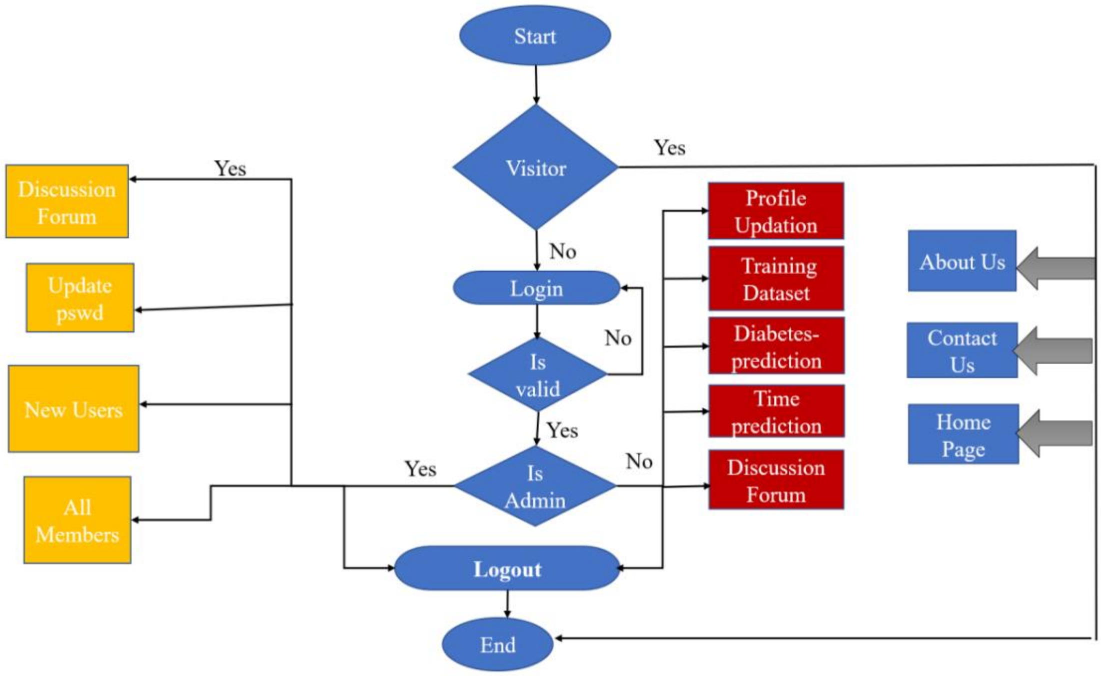


Fig 4.5: Control Flow Diagram

Implementation can be described as the realization of an application, or execution of the plans, ideas, models, design and system development, specification of the model, standard, algorithms used in the system, or authority. In computer science, an implement is explained as the realization of technically specified or algorithms as a programmed, a software component, or any others computer systems through computer programming and deployment. Many of the implementations may existed for a given specification or standard.

CHAPTER 5

PROPOSED METHODOLOGY

5.1 Procuring the Dataset

The dataset used here is the PIMA Indian Dataset. It is the data obtained from the National Institute for Diabetes. It contains of several medical predictor variables and one target variable. The various medical variables are BMI, Glucose levels, Blood Pressure etc. It contains 768 rows and 9 columns. The columns that are present in the dataset are as follows:

5.1.1 Skin Thickness

Skin thickness is a column in the dataset which denotes the thickness of an individual’s skin. Skin thickness varies from person to person depending upon their health and various other factors which can affect the skin. A person’s skin thickness can play a factor in denoting whether the person has diabetes or not, but in the dataset, there are a few rows where the skin thickness is set to 0. Skin thickness cannot be 0 for a person, so we try to avoid this column mainly to get the accurate results while performing prediction. While performing analysis, the skin thickness column is removed from the code we write so as to get a more accurate prediction result using Naïve Bayes and Random Forest.

3.1.2 Number of Pregnancies

When a woman gets pregnant, they may or may not go through gestational pregnancy. Gestational pregnancy is a common form of pregnancy where the woman develops diabetes. After the birth, the diabetes usually goes away. The diabetes is caused due to the high levels of sugar in the body which does not happen when the woman is not pregnant. This is due to the making of hormones by the placenta. The number of pregnancies plays a key factor when it comes to diabetes in women. So we record the number of pregnancies and if it is a male, the pregnancy is set to 0 in the dataset. It can also denote that a woman has not been pregnant during her life.

5.1.3 Glucose Concentration

The glucose concentration is the level of glucose that is present in a person’s blood. A teaspoon of glucose is required for a human body to function normally per day. The glucose present in the body travels through the bloodstream to other parts of the body. The glucose level is required to determine the amount of insulin present in the body. If the insulin is not able to handle the amount of glucose in the body, then this causes diabetes. The glucose levels in a person’s body is an important factor in determining if the person has diabetes or not. In the dataset, we have a column to represent the glucose level of each person.

5.1.4 Blood Pressure

The blood in our body moves through our body by the means of blood pressure. It helps in the movement of oxygen and nutrients throughout our body through the blood. The white blood cells in our body are also delivered by the means of blood pressure. The normal blood pressure for a person is usually below 120 mm Hg systolic and 80 mm Hg diastolic. Variations in blood pressure can be a major cause of diabetes mellitus. So we take this factor in our dataset for the prediction of diabetes.

5.1.5 Insulin

To control the blood sugar in our body, insulin is required. Insulin is a hormone created by the pancreas to balance all the sugars in our body. The insulin controls the glucose concentration, which is a major factor in development of diabetes. If the insulin is not able to keep up with the levels of glucose in our body, it causes diabetes. Insulin also helps in the breaking down of fats and proteins in our body to form energy. Insulin resistance is the inability of insulin to exert its effects on the tissues in our body. In the dataset, the insulin level plays a key role in the prediction of diabetes.

5.1.6 Body Mass Index (BMI)

The Body Mass Index of a person can be defined as the person’s weight divided by the square of the height. The weight is defined in kgs and the height is defined in meters. The BMI of a person varies according to the weight and height and it calculates whether the person is normal or obese. The following table denotes the various BMIs that distinguish a person into four categories:

BMI	**Category**
Under 18.5 kg/m^2	Underweight
18.5 to 25	Normal Weight
25 to 30	Overweight
Over 30	Obese

Table 5.1: BMI & its Category

5.1.7 Diabetes Pedigree Function

A pedigree shows relationships between family members and indicates which individuals have certain genetic pathogenic variants, traits, and diseases within a family as well as vital status. A pedigree can be used to determine disease inheritance patterns within a family. Enlarge.

5.1.8 Age

Age is a common factor for diabetes. When it comes to age, usually, people above the age of 40 are diagnosed with diabetes. But, sometimes, even people who are younger are diagnosed with diabetes. Type 1 diabetes usually occurs in people above the age of 40 but sometimes, people at ages as young as 15 – 16 can also be diagnosed. This all depends on factors such as family history, diet etc.

5.1.9 Value of Diabetes Diseases

In the dataset, this column is used to define if the person has diabetes or not. We define it using True or False. The dataset has predefined values for each person whether the person has diabetes or not and our project is to find whether the given values are accurate or not.

The following features are the key to finding whether a person has diabetes or not. There are various other factors as to determining diabetes, but in our project, we are mainly focusing on these features for the prediction.

The dataset file is in a .csv(Comma Separated Values) format. Using the help of Python’s inbuilt library Pandas, which is a dataframe library, we import the file into our Python environment. The other libraries that are imported into the environment are:

Numpy – a library that is used mainly to operate with large dimensional arrays and matrices, providing high level mathematical functionalities to work on data.

Matplotlib – the library that provides Python with the functionality of plotting graphs and plots. It works in tandem with NumPy. Pandas has a function named read_csv(), which essentially reads a file of the format (.csv).

Once the dataset is loaded into the environment, we can check the dimensions of the dataset by the function .shape () which returns the number of rows and columns. Basic lookup of the data is done, by using the inbuilt commands .head () and .tail () which print the number of rows from the start of the dataset and the bottom of the dataset respectively.

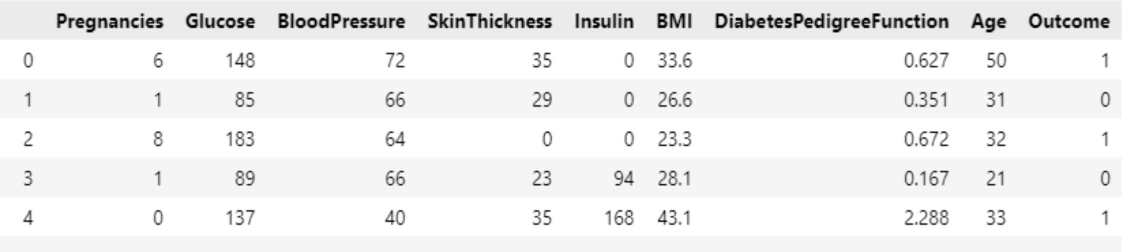
After procuring the dataset, we see if we can make any changes to the dataset. Operations such as initialization of the variables, cleansing the data, making appropriate labels for the data takes place. In our case, the dataset contains a parameter skin thickness, which is column that has a weak correlation to the contribution of a person being diabetic. Hence, we remove the column for our analysis. In this stage, we can calculate the numeric aspects of the data, such as the average of a particular column, number of cases of the column based on conditions etc.

The dataset contains the values for the people having diabetes and people who don't. Hence, we calculated the count for each case. The result turned out like this:

People with Diabetes: 268

People without Diabetes: 500

In the given data, around 35% of the people have been diagnosed with diabetes.



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Fig 5.1: Dataset Samples

5.2 Splitting the Data

Splitting the data into training and test data, is one of the most crucial steps in the analysis. The split of the training data is more than the training data. The training data undergoes through learning. This data which is trained is later 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 column that are medical predictors and one target column, that of the diabetes outcome. The medical predictors are given as inputs to a variable and the target variable is 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 contained values that were null, hence in order to streamline the analysis and the prediction, the null values were filled with the mean values of the respective columns.

5.3 Random Forest

In our case, we have split the labels into two variables, which is input to the classifier. One of the greatest strengths of Random Forest classifiers is its ability to be used with practically in any kind of data, especially with feature selection.

In our dataset we have a feature selection process, hence the use of the model. The model is available in the sklearn library, hence we utilize that to begin our prediction. In our case, we use the RandomForestClassifier() function to perform prediction on the data. The input training and test data, is fitted to the model using fit (). The training data is then classified into arrays during prediction. The accuracy of the model is obtained by comparing the predicted values against the original set of values.

The model.fit() function trains the model for a given number of epochs (iterations on a dataset).

Arguments

x: Numpy array of training data (if the model has a single input), or list of Numpy arrays (if the model has multiple inputs). If input layers in the model are named, you can also pass a dictionary mapping input names to Numpy arrays. X can be None (default) if feeding from framework-native tensors (e.g. TensorFlow data tensors).

y: Numpy array of target (label) data (if the model has a single output), or list of Numpy arrays (if the model has multiple outputs). If output layers in the model are named, you can also pass a dictionary mapping output names to Numpy arrays. Y can be None (default) if feeding from framework-native tensors (e.g. TensorFlow data tensors).

Y_train.ravel() returns contiguous flattened array(1D array with all the input-array elements and with the same type as it). A copy is made only if needed.

Prediction

To predict values using the training data, we use the predict function. A class prediction is: given the finalized model and one or more data instances, predict the class for the data instances.

We do not know the outcome classes for the new data. That is why we need the model in the first place. We can predict the class for new data instances using our finalized classification model in scikit-learn using the predict() function.

To perform prediction, we make use of the tool sklearn. Scikit-learn is probably the most useful library for machine learning in Python. It is on NumPy, SciPy and matplotlib, this library contains a lot of effiecient tools for machine learning and statistical modeling including classification, regression, and clustering and dimensionality reduction.

The goal of ensemble methods is to combine the predictions of several base estimators built with a given learning algorithm in order to improve generalizability / robustness over a single estimator. We use the sklearn.ensemble tool to import RandomForestClassifier.

Train_test_split splits arrays or matrices into random train and test subsets. That means that everytime we run it without specifying random_state, we will get a different result, and this is expected behavior.

If use random_state = some number, then we can guarantee that the outputs will be equal i.e. The split will be always the same. It doesn't matter what the actual random_state number is 42, 0, 21, the important thing is that everytime we use 42, we will always get the same output the first time we make the split. This is useful if we want reproducible results, for example in the documentation, so that everybody can consistently see the same numbers when they run the examples. In practice I would say, you should set the random_state to some fixed number while we test stuff, but then remove it in production if we really need a random (and not a fixed) split.

Finding the Accuracy

The next step is to find the accuracy of the training and testing data. To find the accuracy, we use a function called metrics.accuracy_score. In multilabel classification, this function computes subset accuracy: the set of labels predicted for a sample must *exactly* match the corresponding set of labels in y_true.

First, we check the accuracy of the training data by passing the arguments for the training data split. After that, we check the accuracy of the testing data by doing the same with the testing data as the parameters. By comparing both, we print a confusion matrix.

A confusion matrix is used to evaluate the accuracy of a classification. By definition a confusion matrix C is such that $C_{i,i}$ is equal to the number of observations known to be in group i but predicted to be in group j. Thus in binary classification, the count of true negatives is $C_{0,0}$, false negatives is $C_{1,0}$, true positives is $C_{1,1}$ and false positives is $C_{0,1}$.

Parameters:

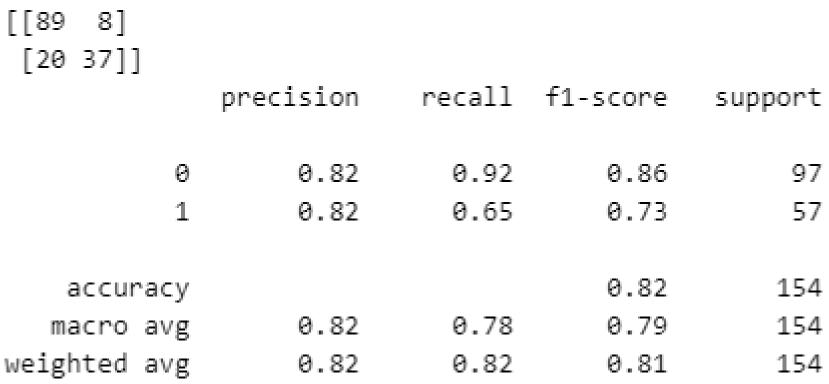
y_true : array, shape = [n_samples]. Ground truth (correct) target values. **y_pred :** array, shape = [n_samples]. Estimated targets as returned by a classifier.

labels : array, shape = [n_classes], optional. List of labels to index the matrix. This may be used to reorder or select a subset of labels. If none is given, those that appear at least once in y_true or y_pred are used in sorted order.

sample_weight : array-like of shape = [n_samples], optional. Sample weights. **Returns:** C : array, shape = [n_classes, n_classes].Confusion matrix

5.4 Classification Report

The classification report visualizer shows the exactness, review, F1, and bolster scores for the model. So as to help simpler elucidation and issue recognition, the report coordinates numerical scores with a shading coded heatmap. All heatmaps are in the range (0.0, 1.0) to encourage simple examination of classification models crosswise over various classification reports.



```
[[89  8]
 [20 37]]
              precision    recall  f1-score   support

           0       0.82      0.92      0.86        97
           1       0.82      0.65      0.73        57

    accuracy                           0.82       154
   macro avg       0.82      0.78      0.79       154
weighted avg       0.82      0.82      0.81       154
```

Fig 5.2: Classification Report of Random Forest

CHAPTER 6

EXPERIMENTAL SETUP AND RESULTS

6.1 Experimental Setup

1. Software Required:

- ✓ OS: Windows 7 and above /LINUX
- ✓ Software: Jupyter Notebook
- ✓ Additional requirements: Numpy, Matplotlib

2. Hardware Required:

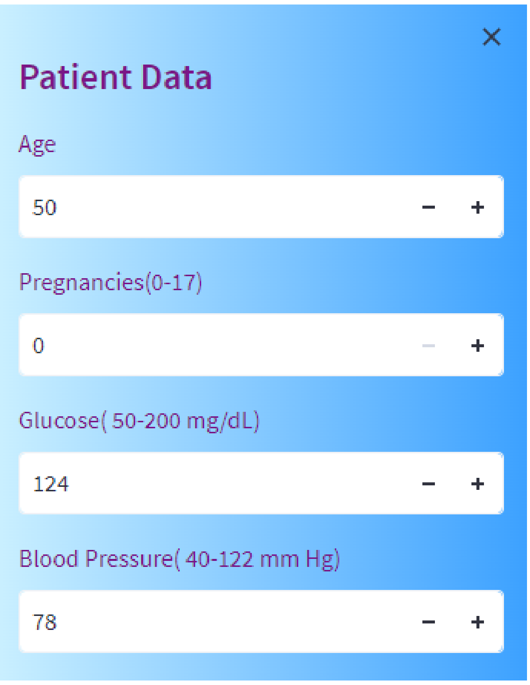
- ✓ Processor: Intel I5 processor
- ✓ Storage Space: 500 GB.
- ✓ Screen size: 15” LED
- ✓ Devices Required: Monitor, Mouse and a Keyboard
- ✓ Minimum Ram: 8GB

3. Technology And Language:

- ✓ Language: - Python programming language (for backend)
- ✓ HTML/CSS, Python (for frontend)
- ✓ IDE Platform-Visual Studio, Python IDLE
- ✓ Libraries- NumPy, pandas, Matplotlib, sklearn, seaborn, **Streamlit**

6.2 Results

Input Data:



Patient Data

Age

50

Pregnancies(0-17)

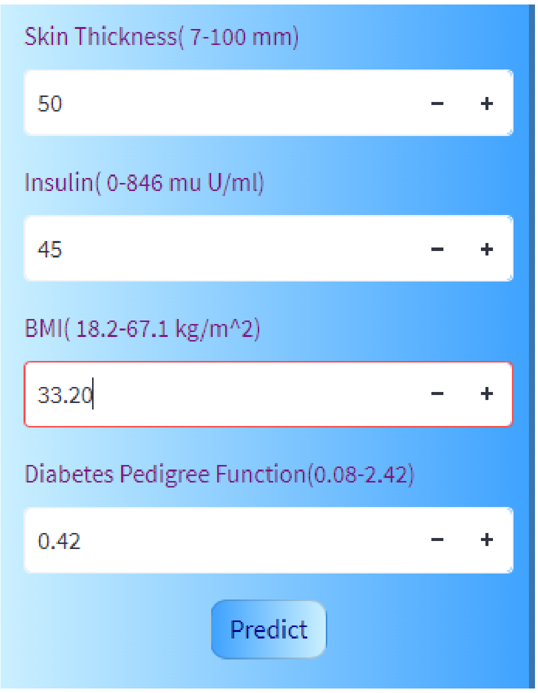
0

Glucose(50-200 mg/dL)

124

Blood Pressure(40-122 mm Hg)

78



Skin Thickness(7-100 mm)

50

Insulin(0-846 mu U/ml)

45

BMI(18.2-67.1 kg/m^2)

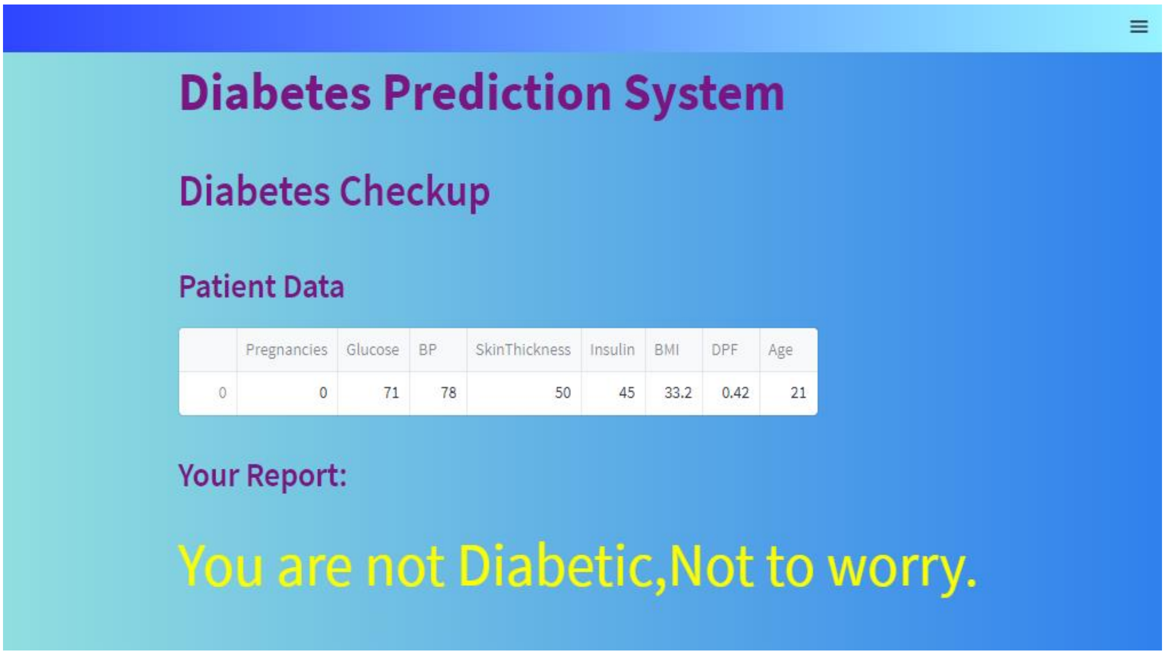
33.20

Diabetes Pedigree Function(0.08-2.42)

0.42

Output:

1) Non Diabetic



Diabetes Prediction System

Diabetes Checkup

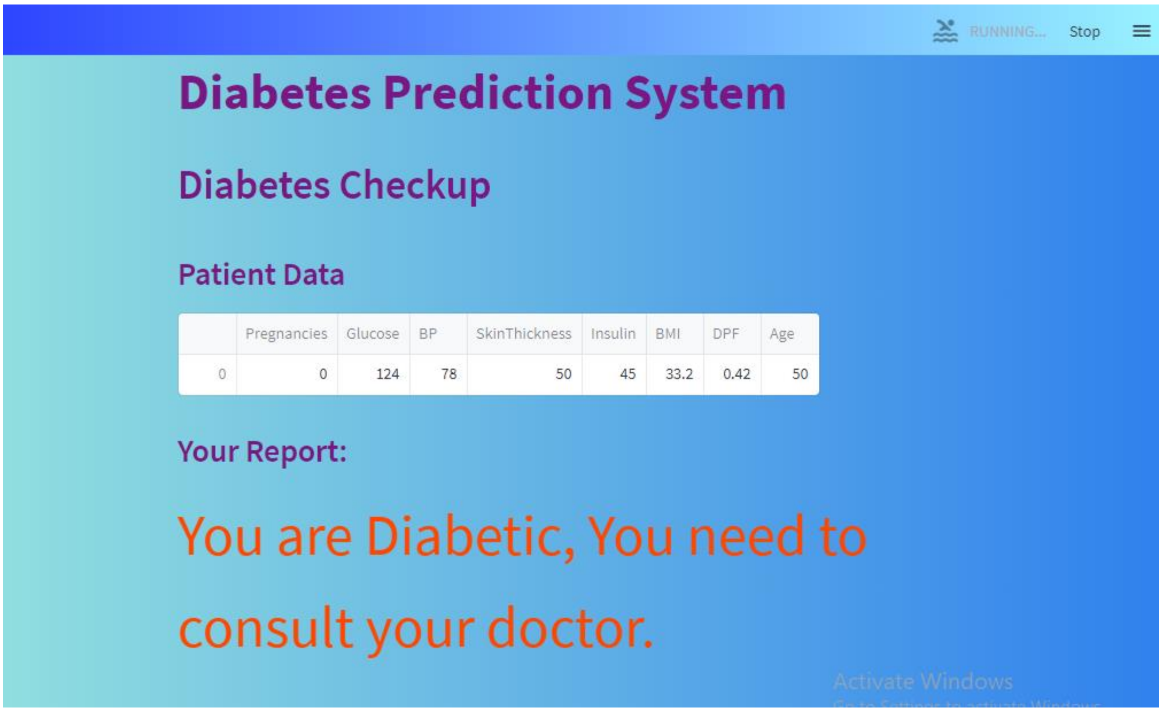
Patient Data

	Pregnancies	Glucose	BP	SkinThickness	Insulin	BMI	DPF	Age
0	0	71	78	50	45	33.2	0.42	21

Your Report:

You are not Diabetic,Not to worry.

2) Diabetic



Diabetes Prediction System

Diabetes Checkup

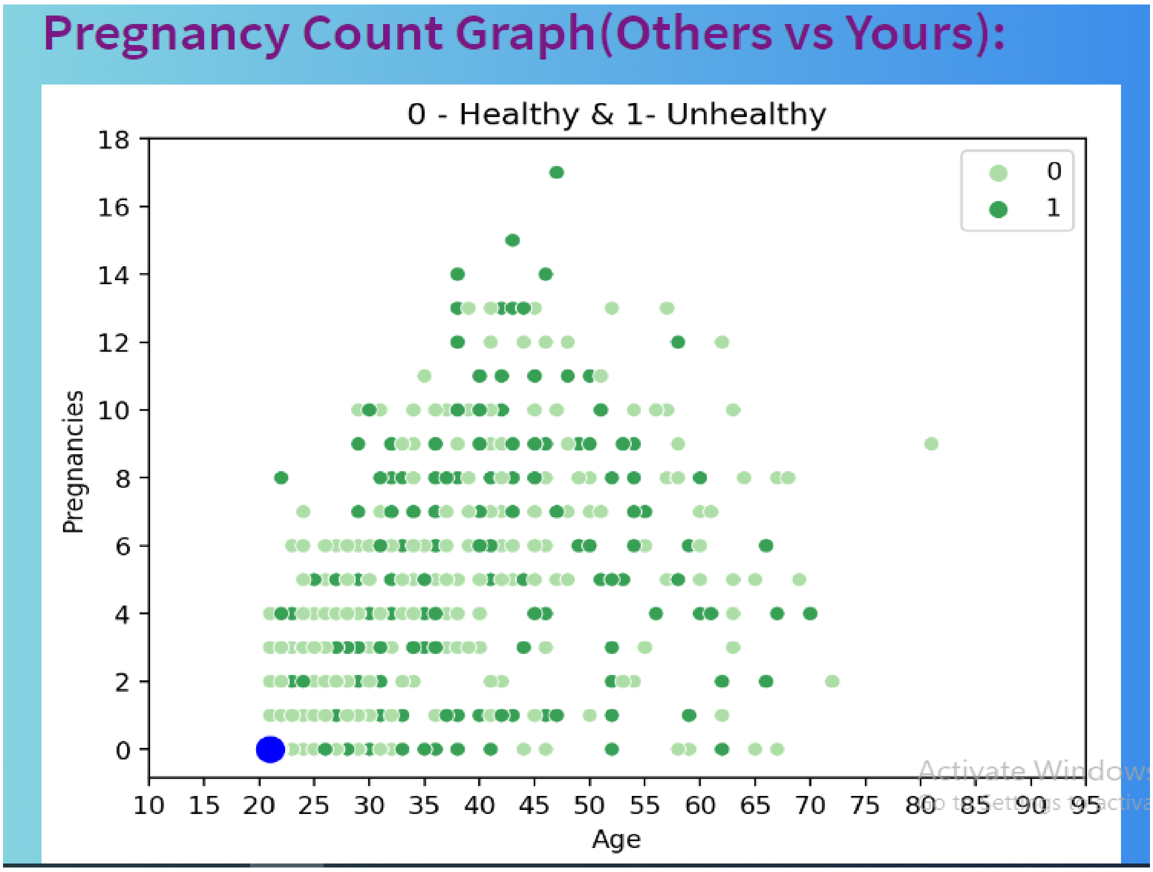
Patient Data

	Pregnancies	Glucose	BP	SkinThickness	Insulin	BMI	DPF	Age
0	0	124	78	50	45	33.2	0.42	50

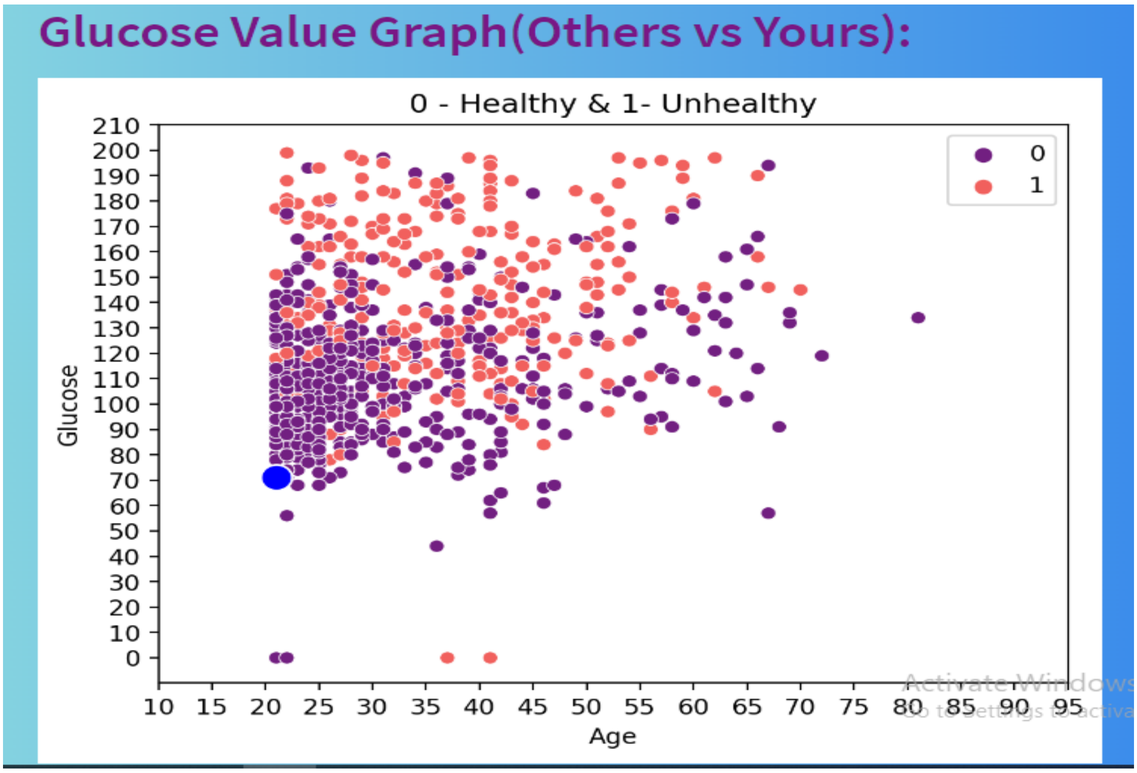
Your Report:

You are Diabetic, You need to consult your doctor.

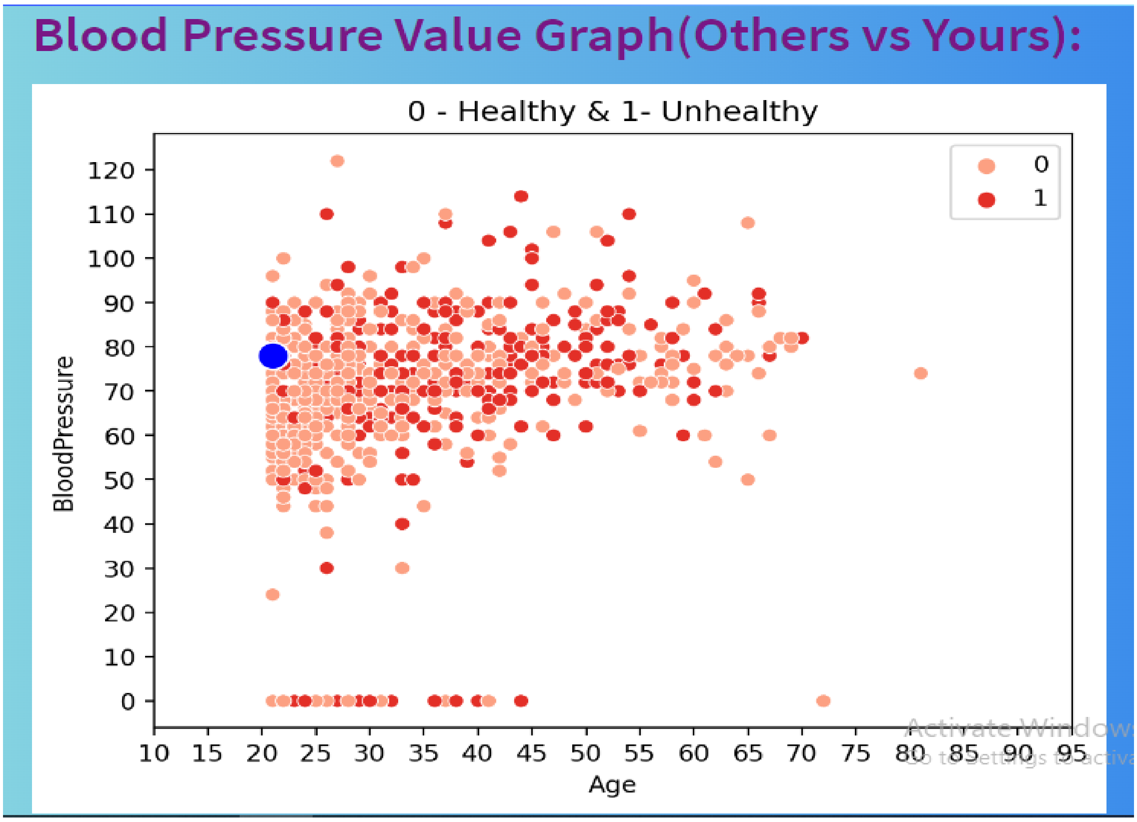
3) Relation Between Age and Pregnancy


Pregnancy Count Graph(Others vs Yours):

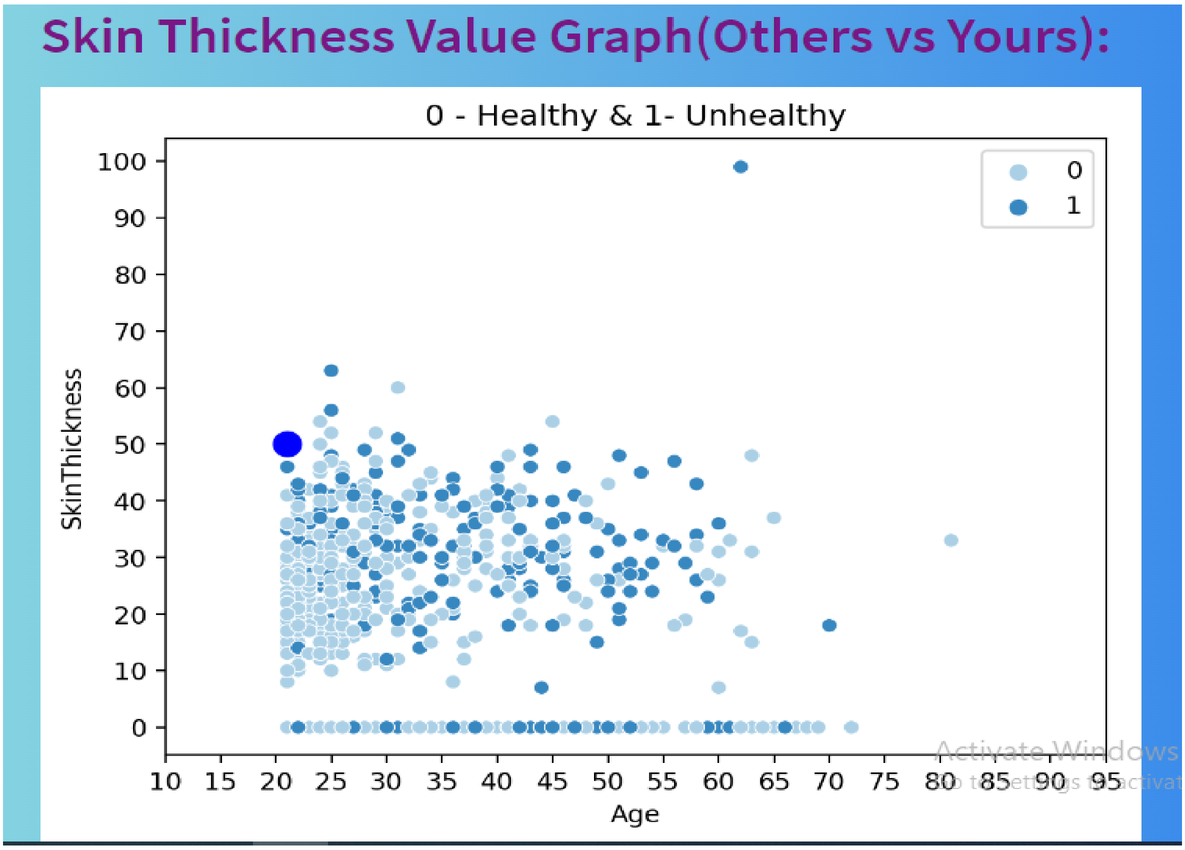
4) Relation Between Age and Glucose



5) Relation Between Age and Blood Pressure

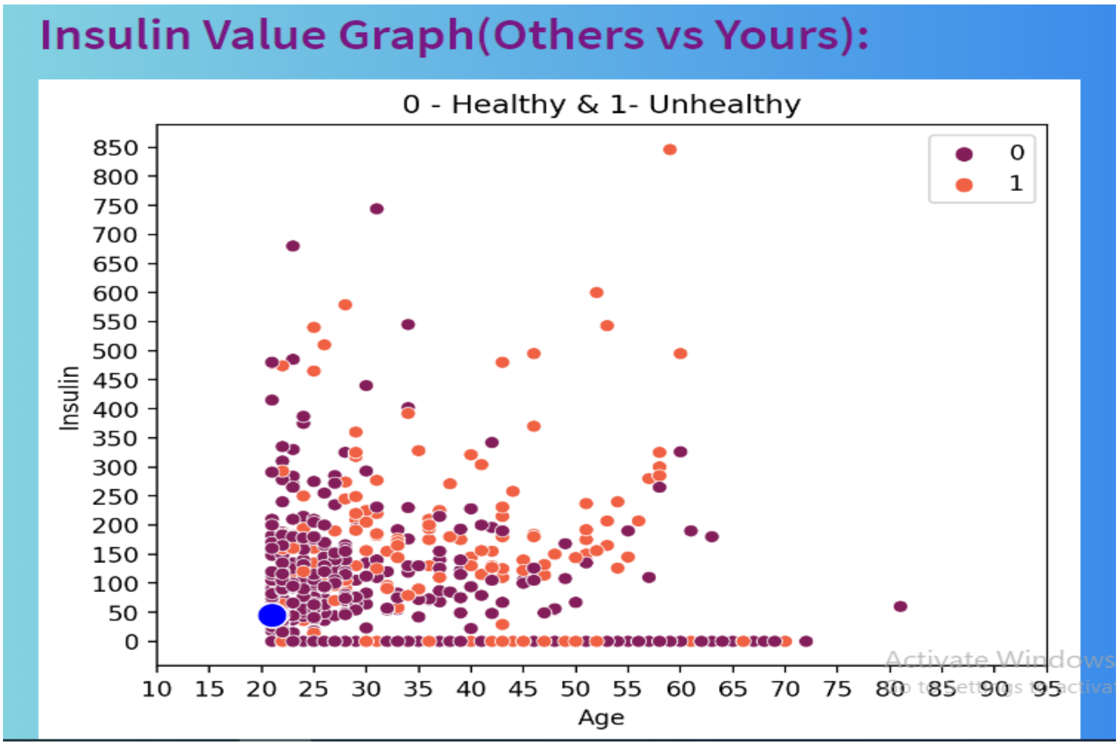


6) Relation Between Age and Skin Thicknes



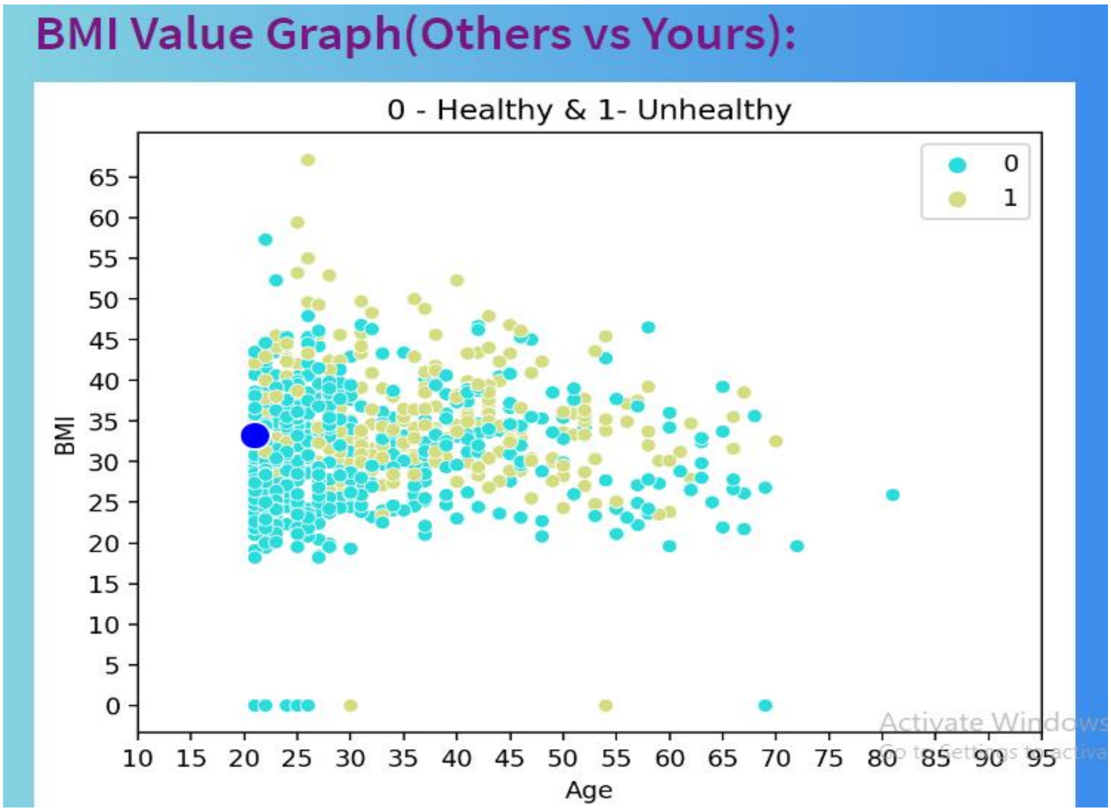
Skin Thickness Value Graph(Others vs Yours):

7) Relation Between Age and Insulin

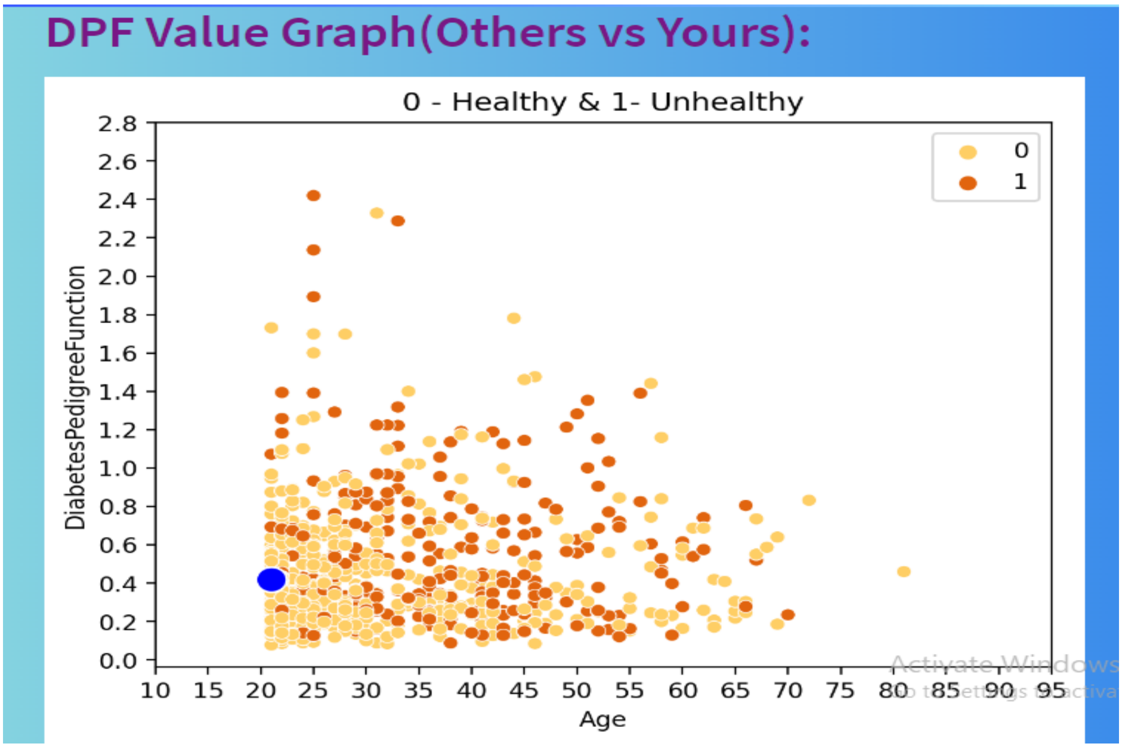


Insulin Value Graph(Others vs Yours):

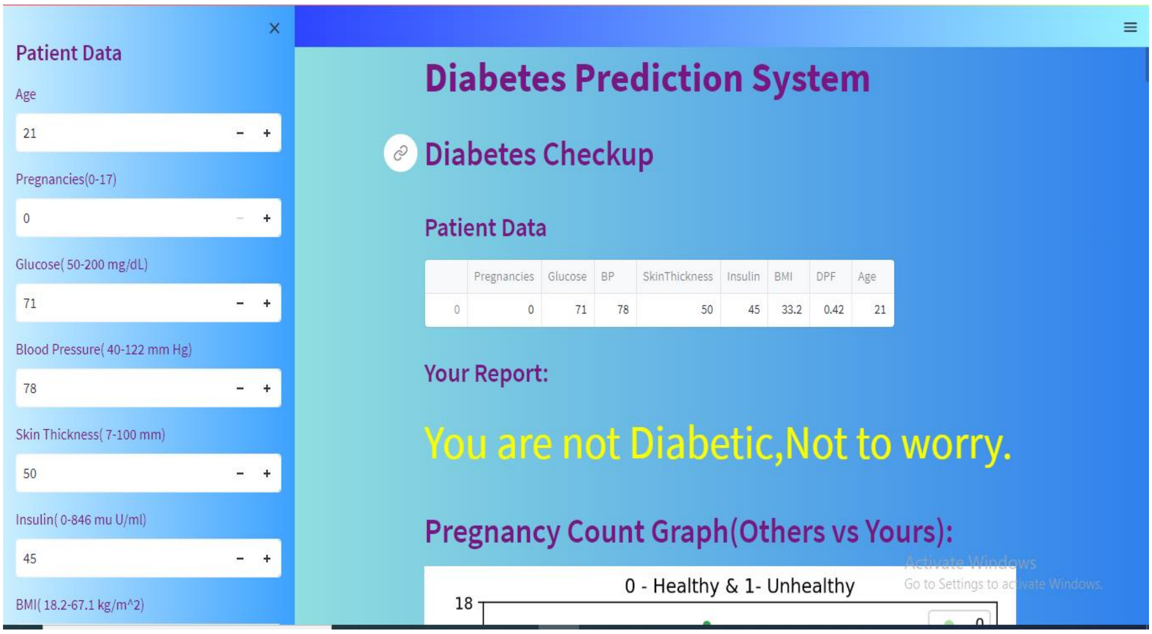
8) Relation Between Age and Pregnancy



9) Relation between Age and Diabetes Pedigree Function



10) Webpage View



Patient Data

Age

21

Pregnancies(0-17)

0

Glucose(50-200 mg/dL)

71

Blood Pressure(40-122 mm Hg)

78

Skin Thickness(7-100 mm)

50

Insulin(0-846 mu U/ml)

45

BMI(18.2-67.1 kg/m^2)

Diabetes Prediction System

Diabetes Checkup

Patient Data

	Pregnancies	Glucose	BP	SkinThickness	Insulin	BMI	DPF	Age
0	0	71	78	50	45	33.2	0.42	21

Your Report:

You are not Diabetic,Not to worry.

Pregnancy Count Graph(Others vs Yours):

×
−
+
≡
0 - Healthy & 1- Unhealthy
18
0
Activate Windows
Go to Settings to activate Windows.

CHAPTER 7

CONCLSION AND FUTURE SCOPE

7.1 Discussion and Conclusion

The main aim of this project is to design and implement Diabetes Prediction using Machine Learning Method and Performance Analysis of that methods and it has been achieved successfully. After using all the patient records we are able to build a machine learning model (Random forest algorithm is the best one which have highest accuracy) to accurately predict whether or not the patients in the datasets have diabetes or not along with that we were able to draw some insights from the data via data analysis and visualization.

The prediction of diabetes is one of the great importance in today scenario, and concerning with its severe complications. Due to the biggest reason for the death in worldwide is diabetes. The System model is mainly focus to identification of diabetes using some of the parameters. System is useful to physicians to predict the diabetes in initial dais. So, that conventional treatments and solutions may be given to the patients. System used some of the techniques like ML for the prediction, so that to get the more precise results. There has been fortune of investigation on the diabetes imprint. Building diabetes disease prediction system is useful for hospitals and doctors. System predicts disease at early stages, so doctors can treat patients in a better way. Proposed model is the real time application in which is meant for multiple hospitals and predicts disease in less time. As we use machine learning algorithms for disease prediction, we will get more accurate and efficient results.

7.2 Future Scope

The above model is used to predict whether a person has diabetes or not using their health records and in future we can build a perfect model using deep learning techniques and providing best accuracy and further we can also build a Web application using flask so that users can give the parameters and based on those attributes the model will predict.

Random forest algorithms are used to predict whether a person having diabetics or not, by keeping his health conditions in mind. Thus, this process enables doctors to easily group, classify and categorize the disease type accordingly treatment can be given to them.

The long-term effects of diabetes include damage to large and small blood vessels, which can lead to heart attack and stroke, and problems with the kidneys, eyes, feet and nerves. The good news is that the risk of long-term effects of diabetes can be reduced.

References

[1] Berger, Ashton C., et al. "A comprehensive pan-cancer molecular study of gynecologic and breast cancers." Cancer cell 33.4 (2018): 690-705.

[2] Dean, Laura, and Jo Mc Entyre. "Introduction to Diabetes" The Genetic Landscape of Diabetes [Internet].National Centre for Biotechnology Information (US), 2004

[3] Aiswarya Iyar, S. Jeyalatha and RonakSumbaly, “Diagnosis of Diabetes Using Classification Mining Techniques”, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.

[4] Zhao, Chunhui, and Chengxia Yu. "Rapid model identification for online subcutaneous glucose concentration prediction for new subjects with type I diabetes." IEEE Transactions on Biomedical Engineering 62.5 (2015): 1333-1344

[5] AakanshaRathore, Simran Chauhan, SakshiGujral, “Detecting and Predicting Diabetes Using Supervised Learning: An Approach towards Better Healthcare for Women”,Volume 8, No. 5, May-June 2017, ISSN No. 0976-5697, Available Online at www.ijarcs.info.

[6] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, H. Tang, Predicting Diabetes Mellitus with Machine Learning Techniques, Vol. 9, Frontiers in genetics, 2018, p. 515, http://dx.doi.org/10.3389/fgene.2018.00515.

[7] N.P. Tigga, S. Garg, Predicting type 2 Diabetes using Logistic Regression accepted to publish in: Lecture Notes of Electrical Engineering, Springer.

[8] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, H. Tang, Predicting Diabetes Mellitus with Machine Learning Techniques, Vol. 9, Frontiers in genetics, 2018, p. 515, http://dx.doi.org/10.3389/fgene.2018.00515.

[10] S. Perveen, M. Shahbaz, A. Guergachi, K. Keshavjee, Performance analysis of data mining classification techniques to predict diabetes, Procedia Comput. Sci. 82 (2016) 115–121.

[11] kaggle.com/datasets/mathchi/diabetes-data-set?select=diabetes.csv