Mini Project-2B Report on

Diabetes Prediction System

by

Rohan Hasabe(19CE1071)

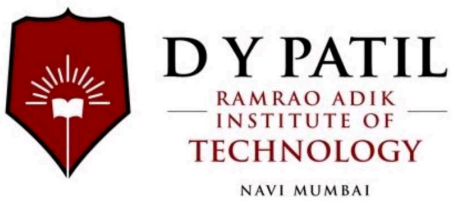
Rahul Jadhav(19CE1065)

Pradyumna Debadwar(19CE1103)

Jatin Dedhia(19CE1058)

Under the guidance of

Dr. Aditi Chhabria



Department of Computer Engineering

Ramrao Adik Institute of Technology

Dr. D. Y. Patil Vidyanagar, Nerul, Navi Mumbai

University of Mumbai

April 2022

Ramrao Adik Institute of Technology

Dr. D. Y. Patil Vidyanagar, Nerul, Navi Mumbai

CERTIFICATE

This is to certify that Mini Project-2B report entitled

Diabetes Prediction System

by

Rohan Hasabe(19CE1071)

Rahul Jadhav(19CE1065)

Pradyumna Debadwar(19CE1103)

Jatin Dedhia(19CE1058)

is successfully completed for Third Year Computer Engineering as prescribed by University of Mumbai.

Supervisor	Project Co-ordinator
(Dr. Aditi Chhabria)	**(Dr.Bharati Joshi)**
Head of Department	Principal
(Dr. Leena Ragha)	**(Dr. Mukesh D. Patil)**

Mini Project Report Approval

This is to certify that the Mini Project-2B entitled ***" Diabetes Prediction System "*** is a bonafide work done by ***Rohan Hasabe***, ***Rahul Jadhav***, ***Pradyumna Debadwar***, and ***Jatin Dedhia*** under the supervision of ***Dr. Aditi Chhabria***. This Mini Project has been approved for Third Year Computer Engineering.

Internal Examiner :

1. .....

2. .....

External Examiners :

1. .....

2. .....

Date : .../.../.....

Place : .....

DECLARATION

I declare that this written submission represents my ideas and does not invovle plagiarism. I have adequately cited and referenced the original sources wherever others' ideas or words have been included. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action against me by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Date: _____

Rohan Hasabe(19CE1071)

Rahul Jadhav(19CE1065)

Pradyumna Debadwar(19CE1103)

Jatin Dedhia(19CE1058)

Abstract

Nowadays, diabetes has become a common disease to the mankind from young to the old persons. Age, obesity, disease, kidney disease, stroke, eye problem, nerve damage, lack of exercise, hereditary diabetes, living style, bad diet, high blood pressure, etc. can cause diabetes mellitus. Current practice in hospital is to collect required information for diabetes diagnosis through various tests and appropriate treatment is provided based on diagnosis. Hence, an early prediction of diabetes can be very essential to save the human life.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

Diabetes prediction systems belong to a larger category of prediction systems that generate predictions based on user input. Our diabetes prediction system use deep learning algorithms to determine whether or not a person is diabetic. Many parameters, including as glucose and insulin levels, diabetes pedigree function, blood pressure, and age, can be used to develop a reliable diabetes prediction system.

1.2 Objectives

• Extracting the most useful features for the model building process from the raw data.
• Retrieving data for building the dataset and splitting the dataset into training set and testing set.
• Experiment on different Deep learning algorithms available for our processed data.
• Integrate the Model with the Web application for performing the prediction.

1.3 Motivation

According to a report of WHO, about 463 million people in the world were affected by diabetes in 2019 making it one of a most significant lethal chronic disease. Early diagnosed of diabetes plays a very important role in the treatment of the disease, however, the long asymptomatic phase makes it difficult to diagnose the said disease. According to International Diabetes Fed-

eration (IDF), about 232 million people (50% of the total population) remain undiagnosed of the disease.

There are three sorts of mistakes in contemporary medical diagnostic methods: The first type is the false-negative type, in which a patient is already diabetic but test results show that he or she does not have diabetes.The second type is the false-positive type. This occurs when inadequate knowledge is extracted from previous data, and a patient's type is projected to be unclassified. Such diagnostic mistakes may result in needless therapies or no treatments at all when they are needed. To minimise the severity of such an impact, a system based on machine learning algorithms and data mining techniques is required, which will offer correct findings while reducing human effort.

1.4 Organization of report

This report gives brief summary of project which includes details about every component of project. The standard components of this report are - Introduction, Literature survey and Proposed system. The purpose of this report is to give a brief overview about our mini project and provide specific information of the survey. The report is split into sections like introduction - which focuses on the importance of donation, motivation - which specifies motive of our project, problem statement - which simplifies the topic of project, objectives - that define measurable outcomes. The literature survey includes study of existing systems - which makes the analyzing simpler and efficient. Their features and limitations - which helps to explore more objectives in our project. The proposed system involves Introduction, Architecture/ Framework, Algorithm and Process Design, Details of Hardware & Software, Experiment and Results, Conclusion and Future Work. This gives us an idea about the working of system and process design. The hardware and software simplify the construction of project and provides functionality for smooth working of the system. The future work describes the area where the study and survey can be applied to build a strong understanding.

Chapter 2

Literature Survey

2.1 Existing Systems

1. **Title:** A novel classification method for diagnosis of diabetes mellitus using artificial neural networks.

Authors: T. Jayalakshmi, Dr. A. Santhakumaran

Year of Publication: 2010

Description: This paper approached the aim of diagnoses by using ANNs and demonstrated the need for preprocessing 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. Limitation: The research was conducted in respect to only artificial neural networks.

2.**Title:** Performance Analysis of Classifier Models to Predict Diabetes Mellitus

Authors: J. Pradeep Kandhasamy, S. Balamurali

Year of Publication: 2015

Description: This research study compares the performance of algorithms those are used to predict diabetes using data mining techniques. Authors compared four prediction models for predicting diabetes mellitus 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. From this we can come to know that after removing the noisy data from our dataset it will provide good result for our problem.

Limitation: The research was conducted in respect to only artificial neural networks.

3. **Title:** Predictive analysis of diabetes using J48 algorithm of classification techniques.

Authors: K. Pradeep, N. Naveen

Year of Publication: 2016

Description: In this paper, the performance of machine learning techniques was compared and measured based on their accuracy. The accuracy of the technique is varied from before pre-processing and after pre-processing as they identified on this study. This indicates the in the prediction of diseases the pre-processing of data set has its own impact on the performance and accuracy of the prediction. The algorithms were used by this paper were five different algorithms GMM, ANN, SVM, EM, and Logistic regression. Finally, the researchers conclude that ANN (Artificial Neural Network) was providing High accuracy for prediction of Diabetes.

Limitation: The research was conducted using only small sample data for prediction of Diabetes.

2.2 Limitations of Existing System

In this section, we enlist a set of limitations of this research, which provides opportunities for future work.

• The research paper lacks predicting the type of diabetes a patient has.

• Diabetes prediction models usually are additive models and use linear terms and most do not account for interactions between risk factors.

• If nonlinear associations and interactions between variables are ignored, the accuracy of the models may be compromised.

Chapter 3

Proposed System

3.1 Problem Statement

Diabetes mellitus is a metabolic condition defined by an abnormal rise in blood sugar content caused by insulin insufficiency. According to the World Health Organization, 422 million people worldwide suffer with diabetes, with the majority living in low- or middle-income nations. Up to 2030, this figure might be boosted to 490 billion. As a result, data analytics might be used to make an early diabetes prediction. The act of analysing and detecting hidden patterns from massive amounts of data in order to make conclusions is known as data analytics. This analytical procedure is carried out in health care utilising machine or deep learning algorithms to analyse medical data and develop machine learning models to perform medical diagnosis.

3.2 Proposed Methodology/Techniques

Deep learning is a branch of machine learning that deals with artificial neural networks, which are algorithms inspired by the structure and function of the brain. Layers are used in deep learning models, and most models include at least three layers. Each layer receives data from the previous layer and passes it on to the next. Because the nervous system has neurons, each line may be thought of as a single neuron that is linked to the neurons in the next layer as well as neurons in the same layer. The good news is that the activation function, f(h), may be any function. This is what makes neural networks conceivable. When data enters a neuron, it is multiplied by the weight value assigned to that input.These weights begin as random numbers, but as the neural network learns more about the types of input data that contribute to correct

outcomes, they become more accurate. The network modifies the weights based on any classification mistakes caused by the prior weights.

The Dataset: Pima Indians Diabetes (PID) dataset of National Institute of Diabetes and Digestive and Kidney Diseases. PID is composed of 768 instances with following columns

No.	Attribute
1	Number of times pregnant
2	Plasma glucose concentration a two hours
3	Diastolic blood pressure
4	Triceps skin fold thickness
5	2-Hours Serum insulin
6	Body mass index
7	Diabetes pedigree function
8	Age

Table.3.2.1 Dataset Attributes

Data preparation:

The input data is processed through a series of procedures at this stage of the proposed system in order to increase the system's performance. To begin, data reduction is done to the input dataset in order to remove any noisy or inconsistent data. Following that, we examined the dataset for any null values and filled them with the medians of their respective characteristics. Then we looked for outliers or extreme values, which we replaced with median or extreme boundaries. The data is then normalised by using a scaler to put the entire dataset into a homogeneous range.

Algorithms:

1. Stochastic Gradient Descent Optimizer: This is a modified version of the GD technique that updates the model parameters on each iteration. It means that the loss function is checked and the model is modified after each training sample. These frequent updates allow the model to reach the minimum in less time, but at the cost of greater variance, which might cause the model to overshoot the desired location. However, this strategy has the benefit of using less memory

than the prior one since the past values of the loss functions are no longer need to be stored.
2. RMS Prop (Root Mean Square) Optimizer: RMS prop is a natural extension of RPPROP's work. The problem of varying gradients is solved by RPPROP. The issue with the gradients is that some were modest while others may be rather large. As a result, establishing a single learning rate may not be the best option. RPPROP adjusts the step size for each weight based on the sign of the gradient. The two gradients are initially compared for signs in this technique. If they both have the same sign, we're on the correct track and can reduce the step size by a modest amount. If they have the opposite indications, we must reduce the step size.Then we restrict the step size, and we're ready to update our weight.
3. Adam Optimizer: Adaptive Moment Estimation combines the capabilities of RMSProp (root-mean-square prop) and momentum-based GD into a single algorithm. Adam optimizers are a strong approach because of the capability of momentum GD to retain the history of updates and the adjustable learning rate offered by RMSProp. It also adds two additional hyper-parameters, beta1 and beta2, which are normally set to 0.9 and 0.99, respectively, but may be changed to suit your needs.

Deployment:
Cloud computing is the distribution of IT resources on-demand through the Internet with pay-as-you-go pricing. Instead of purchasing, operating, and maintaining physical data centres and servers, you may rent computing power, storage, and databases from a cloud provider like Amazon Web Services on an as-needed basis (AWS). The cloud provides you quick access to a wide range of technologies, allowing you to create more quickly and construct almost anything you can dream. You may instantly spin up resources as needed, including computation, storage, and databases, as well as Internet of Things, machine learning, data lakes and analytics, and much more.

3.3 Design of the System

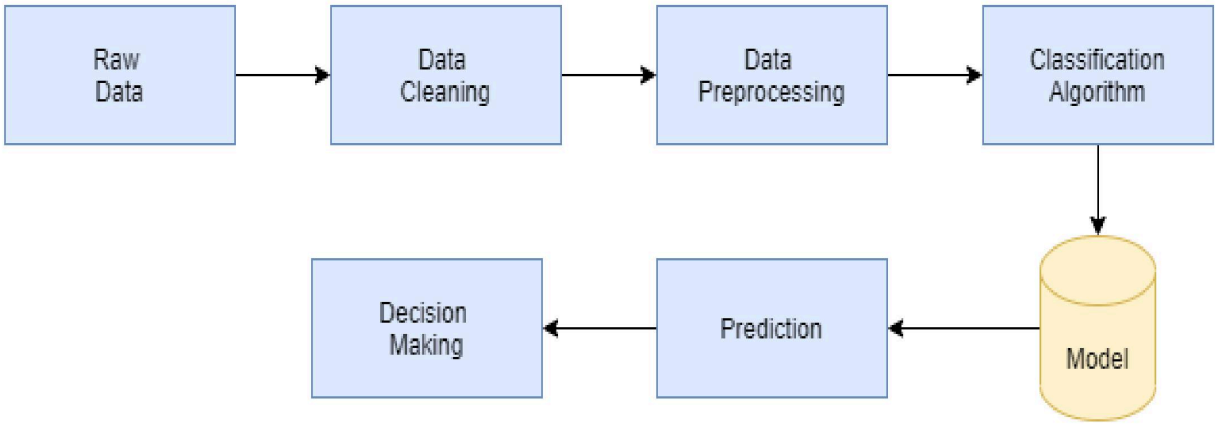


Fig.3.3.1 Flow Diagram for data processing and model creation

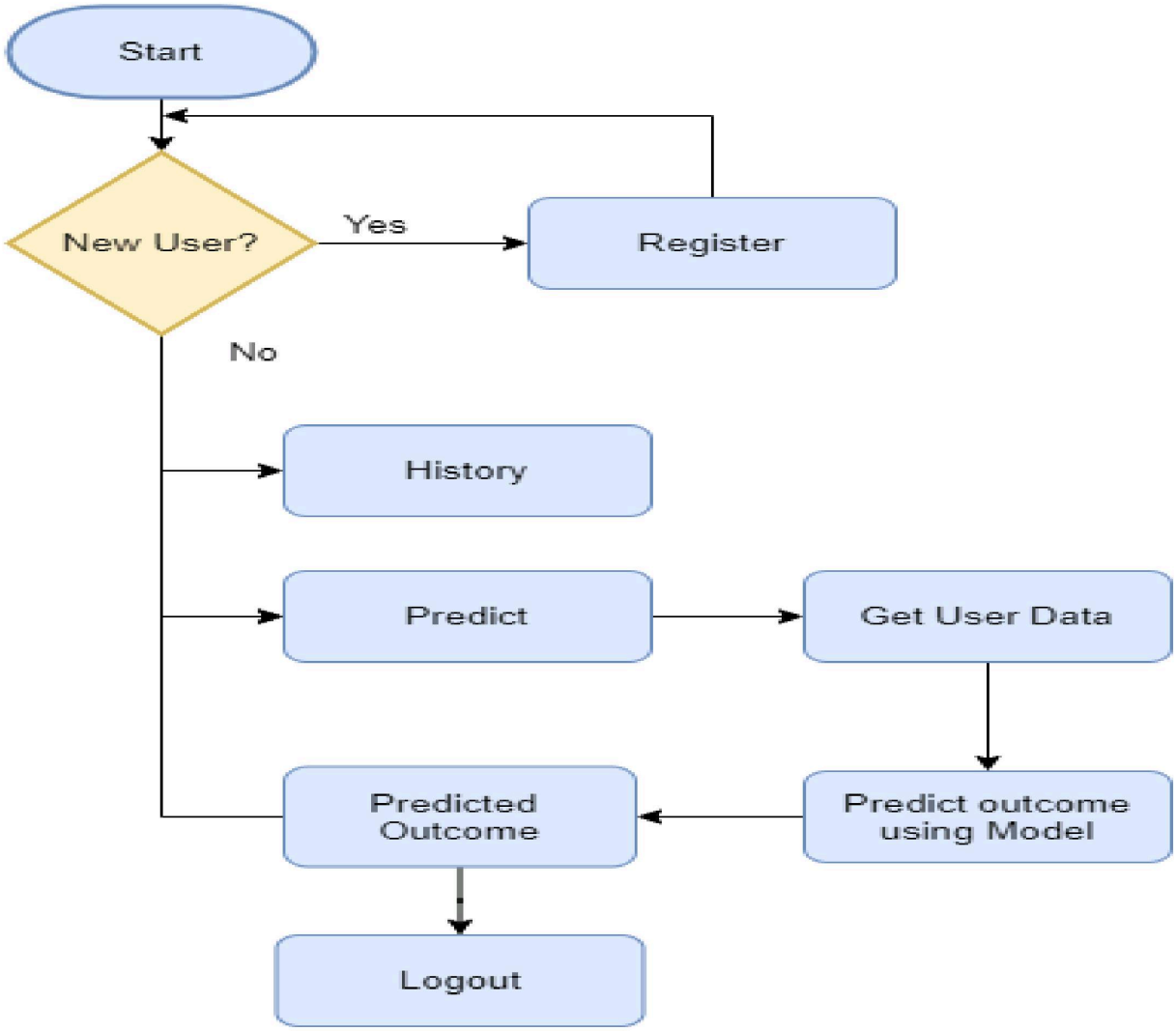


Fig.3.3.2 Flow diagram for our system when a user visits the webapp:

3.4 Hardware/Software Requirement

Hardware:

Operating System: Windows

RAM: 4GB Required

Software:

Browser: MS Edge, Google Chrome

Python Libraries: Pandas, Numpy, Matplotlib, Sklearn, Plotly

Language: Python

Framework: Django

Frontend: HTML, CSS, Javascript

Database: SQLite3

Cloud:AWS

3.5 Implementation Details

In the system firstly, we have started with data preprocessing in which we have eliminated the null values and outliers from the raw dataset.

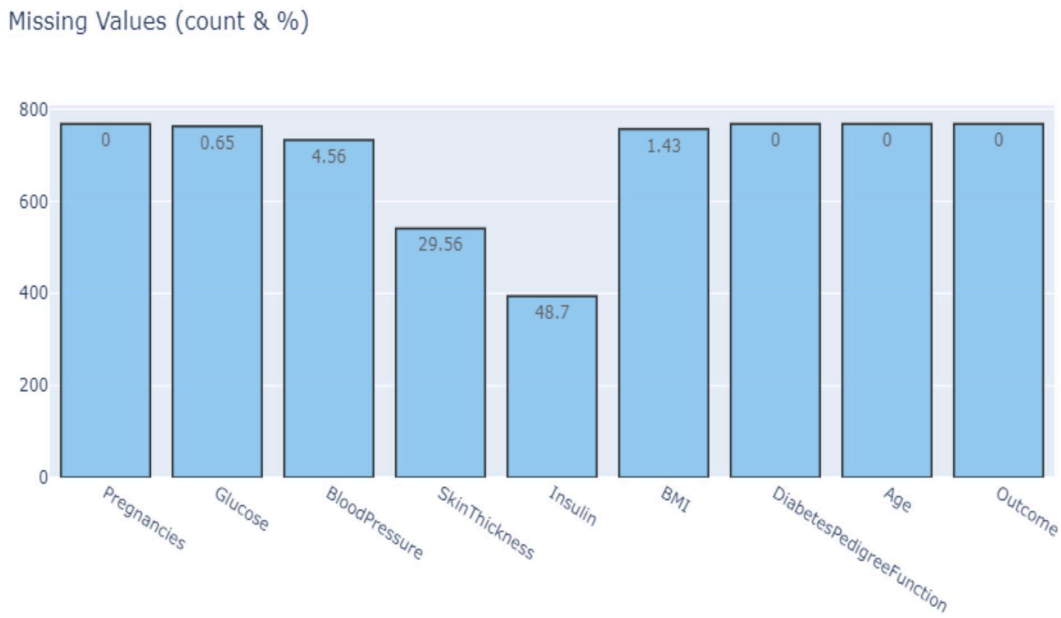


Fig.3.4.1 Missing values in the dataset

Outliers in some of the attributes:

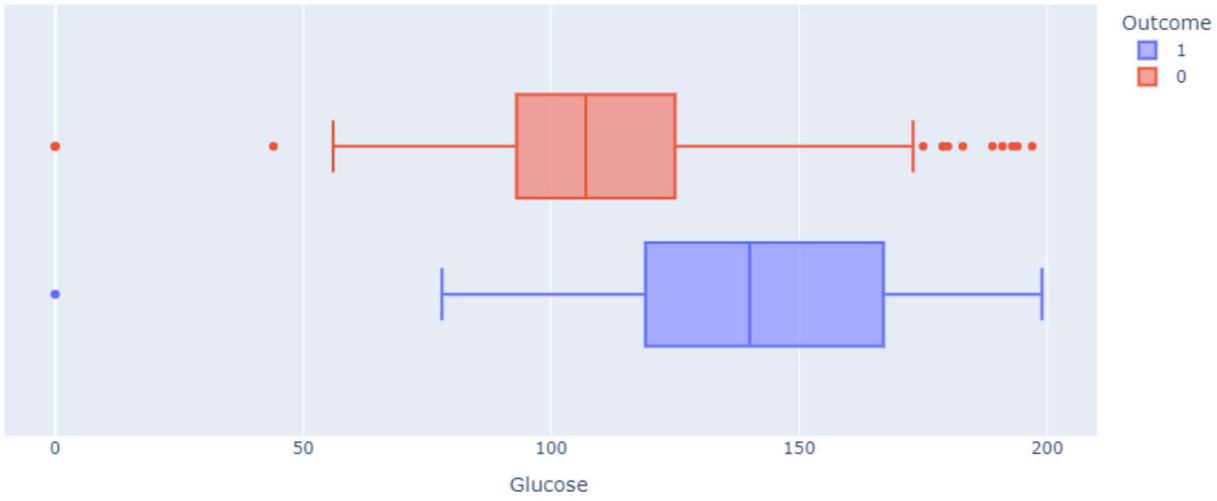


Fig.3.4.2

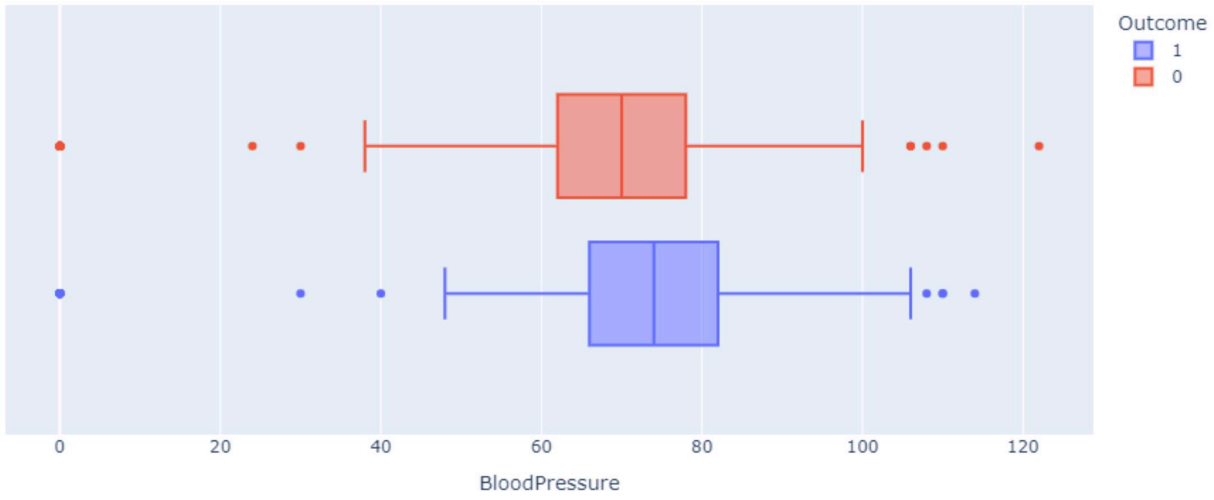


Fig.3.4.3

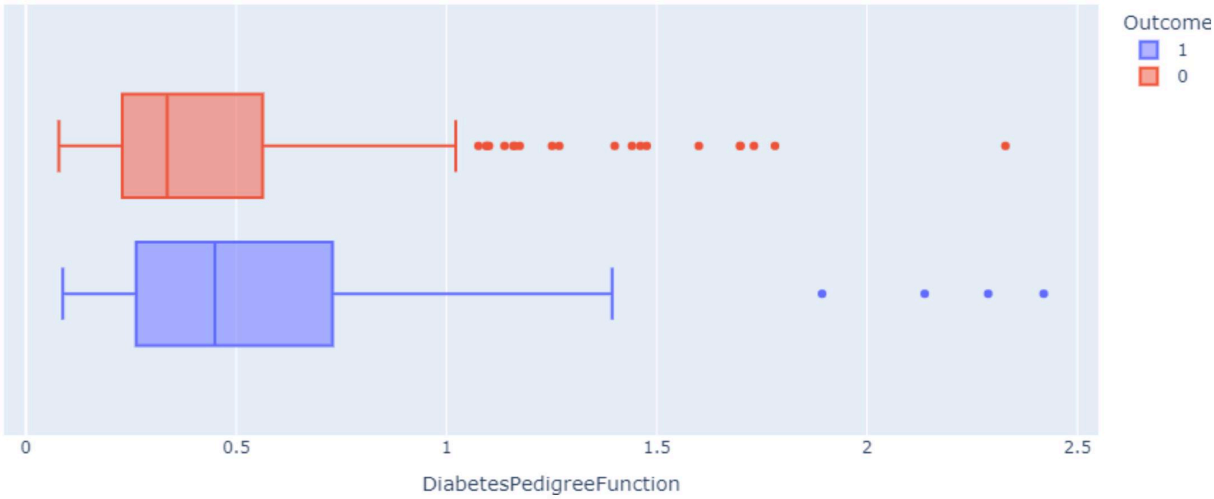


Fig.3.4.4

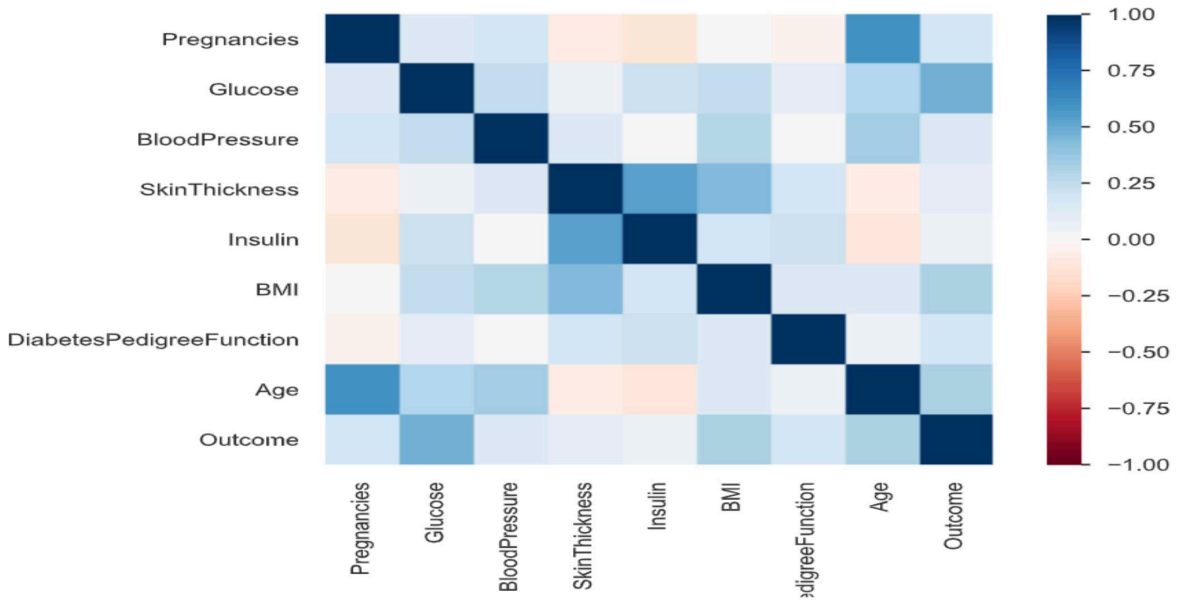


Fig.3.4.5 Correlation between different parameters

Later on, we have implemented different deep learning algorithms on the processed dataset for diabetes prediction and compared their accuracies. A good model with better prediction is then selected for deployment in the webapp. Finally, this webapp is hosted on cloud service like AWS.

Deployment on AWS:

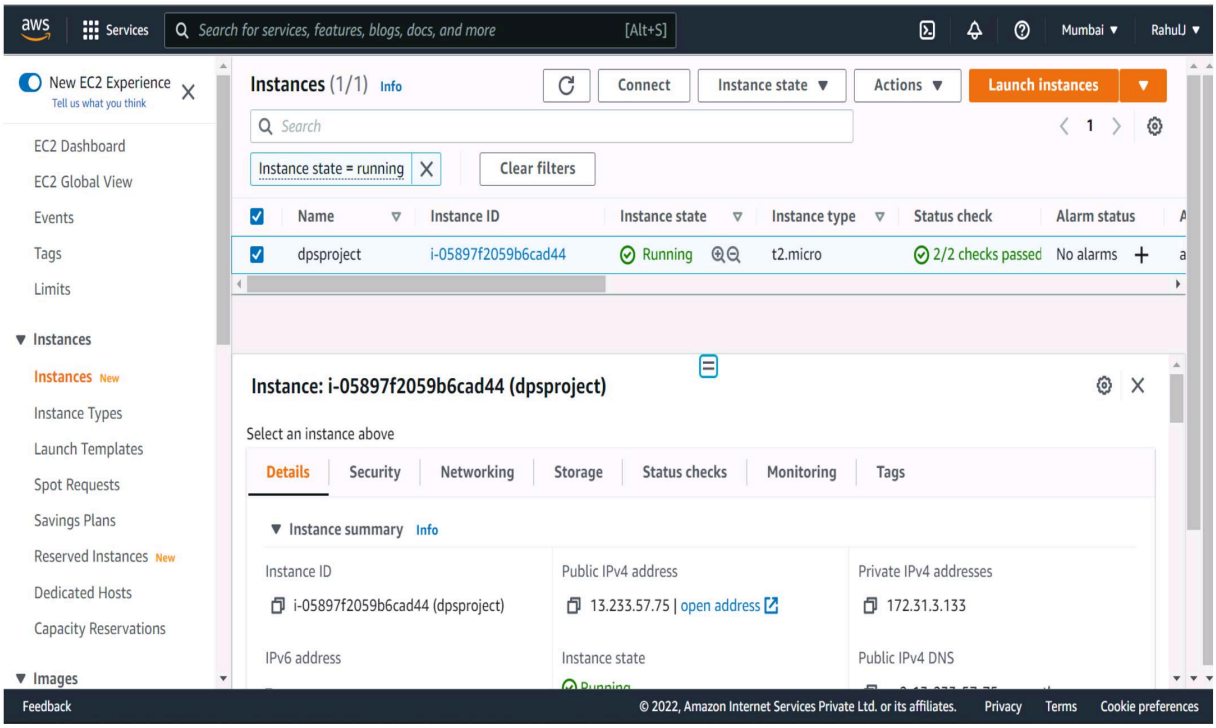


Fig.3.4.6 AWS EC2

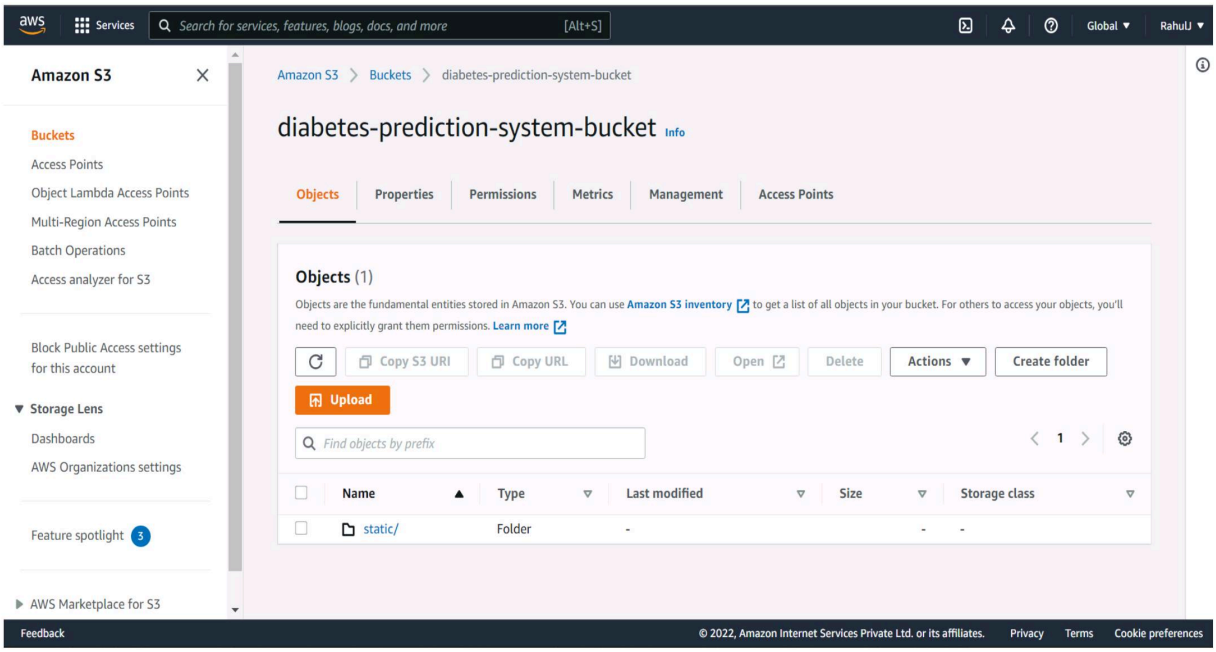


Fig.3.4.7 AWS S3

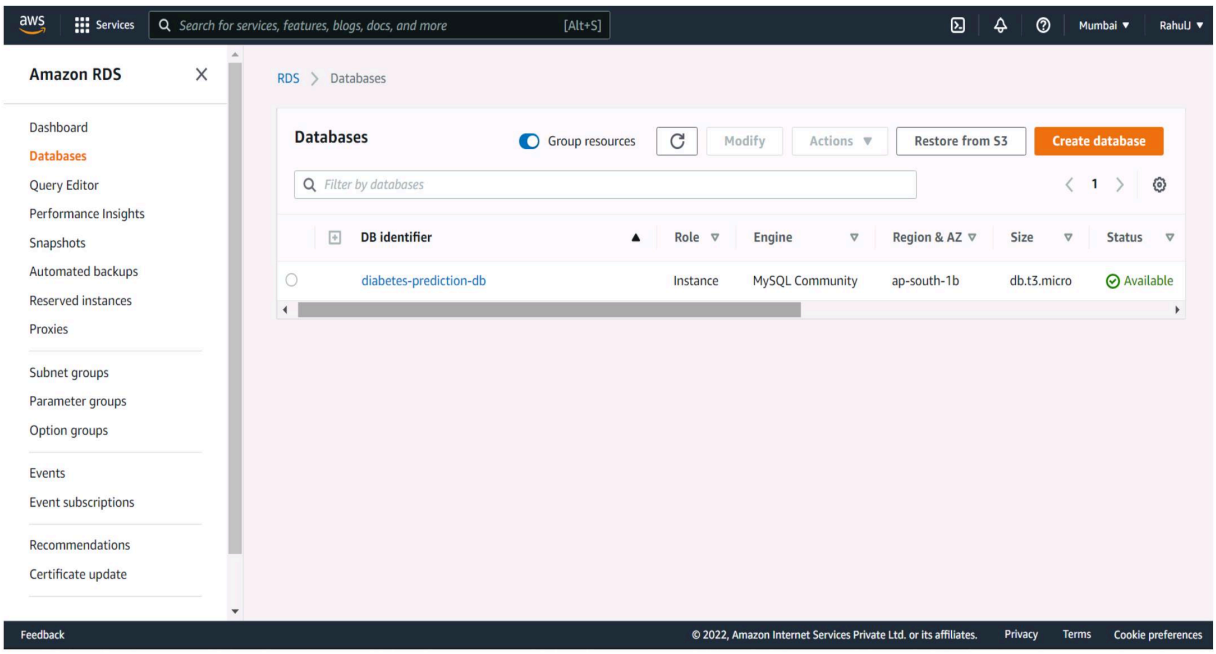


Fig.3.4.8 AWS RDS

Chapter 4

Results and Discussion

4.1 Result and Analysis

In our data, there are 8 independent variables and 1 dependent variable. There are total 768 observations with 500 as healthy and 268 as diabetic. From the data representation graph is it clear that the dataset has lots of zeros as missing values. To counter this, we can replace these values with the medians of their respecting columns. Secondly, the correlation matrix states that skin thickness and blood pressure contribute the least to the outcome. So, we can get rid of them to improve the accuracy of the model. Through Box Plots we can get to see a lot of mild and extreme outliers present in the dataset. This can be handled by eliminating these extreme outliers and replacing them with an upper bound or 0.75 quantile for that column. Algorithms like in Deep Learning use distance for their prediction. So the models require a uniform data which is inside a normalized range. An uniformed dataset may led to poor predictions. So we can use minmaxscaler or standardscaler to normalize our dataset. To prevent overfitting of the model, we will split our data into training and testing sets. This lowers the chances of overfitting in our model. We can also use cross validation or remove unwanted features to avoid overfitting. For selecting a suitable model for the webapp, we would compare the accuracies of the model against the same input data.

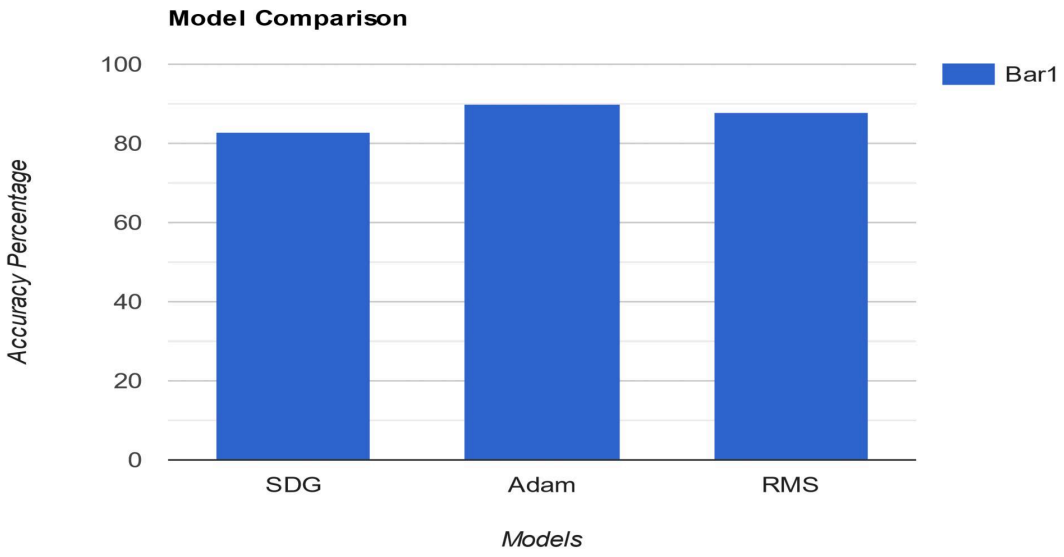


Fig.4.1 Models

From the above bar chart, it is clear that Model developed using Adam optimiser gives the best predictions for the processed data. Considering this chart, we are going to use Deep Learning model with Adam optimizer for our project.

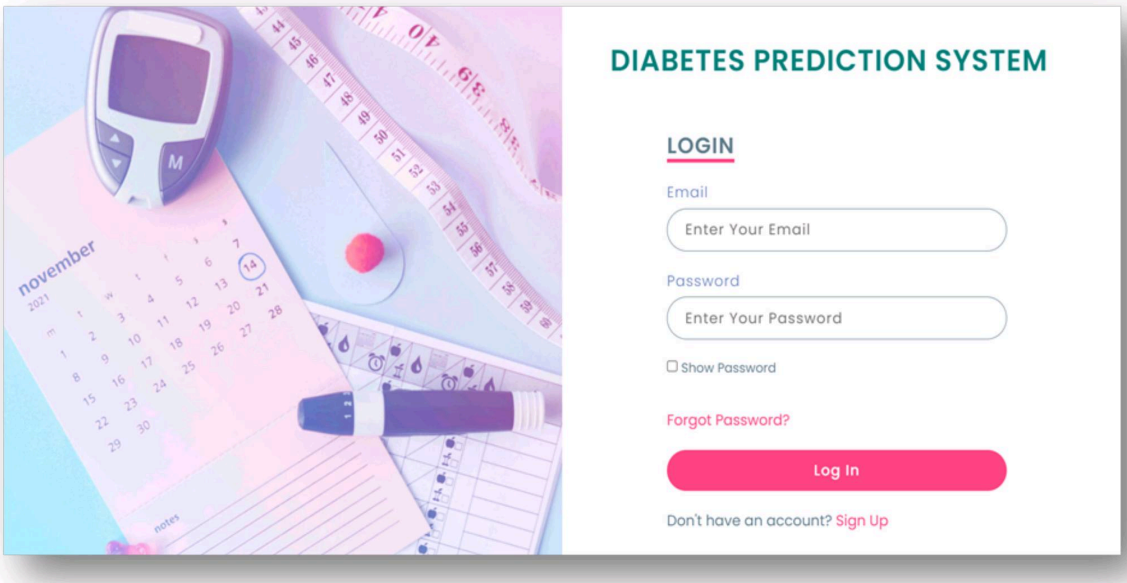


Fig.4.2 Login Page

Fig.4.3 Home Page

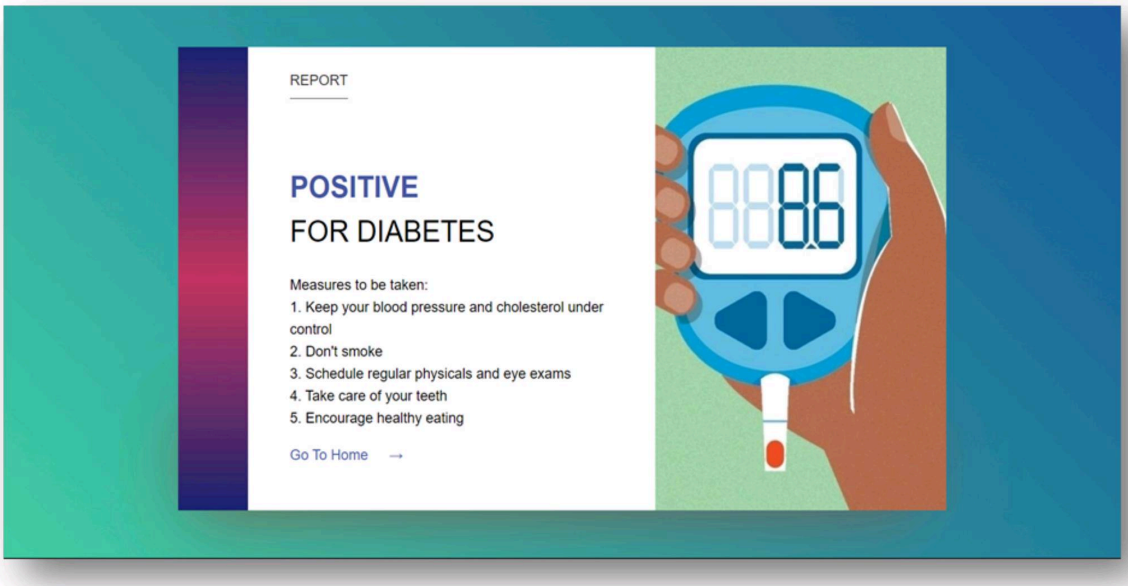


Fig.4.4 Prediction Page

Chapter 5

Conclusion and Further Work

5.1 Conclusion

Detection of diabetes in its early stages is the key for treatment. Machine learning has the great ability to revolutionize the diabetes risk prediction with the help of advanced computational methods and availability of large amount of epidemiological and genetic diabetes risk dataset. The focus of this project has been on implementing a software system that will take into consideration the various factors that affect diabetics.The limitation of this study is that a structured dataset has been selected but in the future we can also apply the models on unstructured data. Due to the data, we also cannot predict the type of diabetes, so in future we aim to predicting type of diabetes, which will led to a more better prediction.

5.2 Further Work

The focus for future work will be to get better predictions by enchancing the model training with inclusion of more sophisticated algorithms. Also, a detailed prediction for diabetes i.e predicting if the patient has Type 1 or Type 2 based on his symptoms. In future we will also consider the add more healthcare functionalities allowing more disease prediction with acceptable accuracy.

Acknowledgments

We would like to express our heartiest gratitude to our project supervisor Mrs. Aditi Chhabria for her guidance and support. We would like to express our thanks to our project coordinator Dr. Bharti Joshi, who gave us a golden opportunity to do this wonderful project on the Diabetes Prediction System. We are also thankful to our Head of Department Dr. Leena Ragha for giving us the chance to conduct this presentation. Our special gratitude goes to our Principal Dr. Mukesh D. Patil for letting us use the resources required. Our sincere thanks to all those who helped us in finalizing this topic within a limited time frame

Date: _____