Diabetes Prediction using Machine Learning Algorithms

A PROJECT REPORT

Submitted by

NURUKURTHI VEERA VENKATA GANESH
[Reg No: RA1911027010109]

Under the Guidance of

Dr. T VEERAMAKALI
(Associate Professor, Department of Data Science and Business Systems)

In partial fulfillment of the Requirements for the Degree of

BACHELOR OF TECHNOLOGY
COMPUTER SCIENCE ENGINEERING
with specialization in Big Data Analytics

DEPARTMENT OF DATA SCIENCE AND BUSINESS SYSTEMS

FACULTY OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
KATTANKULATHUR – 603203
KATTANKULATHUR-603203

NOVEMBER2022

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

KATTANKULATHUR-603203

BONAFIDE CERTIFICATE

Certified that this project report titled "**Diabetes Prediction using Machine Learning Algorithms"** is the bonafide work of **Nurukurthi Veera Venkata Ganesh [Reg No: RA1911027010109]** who carried out the project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion for this or any other candidate.

<<Signature of the Supervisor>>	<<Signature>>
Dr. T VEERAMAKALI **GUIDE** Associate Professor Dept. of DSBS	**Dr. M.Lakshmi** **HEAD OF THE DEPARTMENT** Dept. of DSBS
Signature of Internal Examiner	Signature of External Examiner

ACKNOWLEDGEMENT

We express our humble gratitude to Dr C. Muthamizhchelvan, Vice-Chancellor, SRM Institute of Science and Technology, for the facilities extended for the project work and his continued support.

We extend our sincere thanks to Dean-CET, SRM Institute of Science and Technology, Dr T.V. Gopal, for his invaluable support.

We wish to thank Dr Revathi Venkataraman, Professor & Chairperson, School of Computing, SRM Institute of Science and Technology, for her support throughout the project work.

We are incredibly grateful to our Head of the Department, Dr M. Lakshmi, Professor, Department of Data Science and Business Systems, SRM Institute of Science and Technology, for her suggestions and encouragement at all the stages of the project work.

We want to convey our thanks to our program coordinators Dr.G.Vadivu, Professor, Department of Data Science and Business Systems, SRM Institute of Science anTechnology, for her inputs during the project reviews and support.

We register our immeasurable thanks to our Faculty Advisor, Dr. A. Shanthini, Associate Professor, Department of Data Science and Business Systems SRM Institute of Science and Technology, for leading and helping us to complete our course.

Our inexpressible respect and thanks to my guide, Dr. T. Veeramakali, Associate Professor, Department of Data Science and Business Systems, SRM Institute of Science and Technology, for providing me with an opportunity to pursue my project under her mentorship. She provided me with the freedom and support to explore the research topics of my interest. Her passion for solving problems and making a difference in the world has always been inspiring.

We sincerely thank the Data Science and Business Systems staff and students, SRM Institute of Science and Technology, for their help during our project. Finally, we would like to thank parents, family members, and friends for their unconditional love, constant support, and encouragement.

NURUKURTHI VEERA VENKATA GANESH

ABSTRACT

Diabetes Mellitus is a serious disease that affects a large number of people. Diabetes Mellitus can be caused by age, obesity, lack of exercise, hereditary diabetes, lifestyle, poor diet, high blood pressure, and other factors. Diabetes puts people at a higher risk of developing diseases such as heart disease, kidney disease, stroke, eye problems, nerve damage, and so on. Current hospital process is to gather the necessary information for diabetes diagnosis through various tests, and then provide appropriate treatment based on the diagnosis. Big Data Analytics is important in the healthcare industry. Data sets in the healthcare industry are massive. Using big data analytics, one can examine massive datasets for hidden information and patterns in order to discover knowledge from the data and predict outcomes accordingly. The accuracy of classification and prediction in the existing method is not very high. In this paper, we proposed a diabetes prediction model for better diabetes classification that includes a few external factors responsible for diabetes as well as regular factors such as glucose, BMI, age, insulin, and so on. When compared to existing datasets, the new dataset improves classification accuracy. Furthermore, a pipeline model for diabetes prediction was imposed with the goal of improving classification accuracy.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Diabetes is a disease that diminishes the body's potentiality to secrete insulin. The food we eat is converted into glucose in the blood. Insulin is used to regulate the glucose levels by pushing the glucose into the cells. These cells convert glucose into energy and perform their specialized function. Insufficient secretion of insulin leads to an increase in blood glucose levels. Diabetes can be majorly classified into 3 types: Type 1, Type 2 and Gestational. Type 1 diabetes is caused because of inadequate secretion of insulin. Immunity system destroys the beta cells within the body that produce insulin. The quantity of insulin produced is very little, which means that insulin should be furnished to the body through injections to maintain the blood glucose level. It is most predominant in children but also found in youngsters. Although the exact cause of type 1 diabetes is unknown, factors that may signal an increased risk include Family history, Environmental factors and presence of damaging immune system cells. Type 2 diabetes is also referred to as insulin resistance because it does not use the glucose produced in the body. Glucose is accrued in huge amounts within the blood, which causes diabetes. It is most widespread in adults. Researchers don't fully understand why some people develop type 2 diabetes and others don't. It's clear that certain factors increase the risk, however, including Weight, Family history, Age, High blood pressure, abnormal cholesterol and Race. Gestational diabetes is caused during the time of pregnancy. Change in the hormones produced leads to this scenario. It happens only during pregnancy. The conceived baby is likely to be infected by type 2 diabetes in the future. Complications in your baby includes Excess growth, Low blood sugar, Type 2 diabetes later in life and can also leads to Death. Risk factors for gestational diabetes include Age, Family history, Weight and Race.

1.1 Machine Learning:

Machine learning is the scientific field dealing with the ways in which machines learn from experience. For many scientists, the term "machine learning" is identical to the term "artificial intelligence", given that the possibility of learning is the main characteristic of an entity called intelligent in the broadest sense of the word. The purpose of machine learning is the construction of computer systems that can adapt and learn from their experience. A more detailed and formal definition of machine learning is given by Mitchel: A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. With the rise of Machine Learning approaches we have the ability to find a solution to this issue, we have developed a system using data mining which has the ability to predict whether the patient has diabetes or not. Furthermore, predicting the disease early leads to treating the patients before it becomes critical. Data mining has the ability to extract hidden knowledge from a huge amount of diabetes-related data. Because of that, it has a significant role in diabetes research, now more than ever. The aim of this research is to develop a system which can predict the diabetic risk level of a patient with a higher accuracy. This research has focused on developing a system based on three classification methods namely, Support Vector Machine, Logistic regression and Artificial Neural Network algorithms

1.2 Supervised Learning:

In supervised learning, the system must "learn" inductively a function called target function, which is an expression of a model describing the data. The objective function is used to predict the value of a variable, called dependent variable or output variable, from a set of variables, called independent variables or input variables or characteristics or features. The set of possible input values of the function, i.e. its domain, are called instances. Each case is described by a set of characteristics (attributes or features). A subset of all cases, for which the output variable value is known, is called training data or examples. In order to infer the best target function, the learning system, given a training set, takes into consideration alternative functions, called hypothesis and denoted by h. In supervised learning, there are two kinds of learning tasks: classification and regression.

Classification models try to predict distinct classes, such as e.g. blood groups, while regression models predict numerical values. Some of the most common techniques are Decision Trees (DT), Rule Learning, and Instance Based Learning (IBL), such as kNearest Neighbours (k-NN), Genetic Algorithms (GA), Artificial Neural Networks (ANN), and Support Vector Machines (SVM).

1.3 Unsupervised Learning:

In unsupervised learning, the system tries to discover the hidden structure of data or associations between variables. In that case, training data consists of instances without any corresponding labels. Association Rule Mining appeared much later than machine learning and is subject to greater influence from the research area of databases. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

CHAPTER 2

LITERATURE REVIEW

A. Sneha, N. and Gangil,T., Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big Data, 6(1), p.13.(2019):

Authors have focused on selecting the attributes in early detection of Diabetes Mellitus using predictive analysis and designing a prediction algorithm using Machine learning techniques. The data is collected from CImachine repository.15 attributes hasbeen used for the purpose of classification. Support Vector Machine, Random forest and Naïve Bayes are the classifiers used with an accuracy of 77.73 %, 75.39% and 73.48%.

B.K.VijayaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking,2019

The author proposes a random forest algorithm for diabetes prediction to develop a system that can perform early prediction of diabetes for patients with higher accuracy by using the random forest algorithms. The proposed model gives the best results to predict diabetes and the results show that the prediction system can predict diabetes effectively, efficiently, and most importantly instantaneously. Nanos Nnamoko et al presented Prediction of diabetes onset: a group-supervised learning approach, they used five widely used classifiers for groups and one used Meta classifier. Results are presented and compared with similar studies that have used the same data sets in the literature. It is shown that by using the proposed method, prediction of the onset of diabetes can be made with greater accuracy.

C. Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineering Research and Application, Vol. 8, Issue 1, (Part - II) January 2018, pp.-09-13

Diabetes prediction is presented by machine learning techniques to predict diabetes through three different supervised machine learning methods including SVM, logistic regression, ANN. This project proposes an effective technique for the early detection of diabetes. Deeraj Shetty et al. proposed diabetes disease prediction using data mining assemble Intelligent Diabetes Disease Prediction System that gives analysis of diabetes malady utilizing diabetes patients diagnoses information. In this system, they propose the use of algorithms like Bayesian and KNN (K-Nearest Neighbor) to apply on diabetes patient's databases and analyze them by taking various attributes of diabetes for prediction of diabetes disease. Muhammad Azeem Sarwar proposed a study of diabetes prediction using machine learning algorithms in healthcare. They applied six different algorithms of machine learning. The performance and accuracy of the applied algorithms will be discussed and compared. Comparing different machine learning techniques used in this study shows which algorithm is best suited to predict diabetes. Diabetes prediction has become an area of ??interest for researchers to train programs to identify patients with diabetes by applying the appropriate classifier on the data set. Based on previous research work, it has been observed that the classification process is not much improved. Hence a system is required as Diabetes Prediction is important, in computers, to handle the issues identified based on previous research.

D. Sisodia, D. and Sisodia, DS, 2018. "Prediction of diabetes using classification algorithms. Procedia computer science", 132, pp.1578-1585.(2018) .

The authors designed a support system for estimating disease, including diabetes, using the Pima Indian Selected Diabetes Database (PIDD). In this study, three machine learning recognition algorithms, including Bayes Naive, SVM, and Decision Tree, were used to diagnose diabetes at an earlier stage with an accuracy of 76.3%, 65.1. % and 73.82%.

E. Rahul Joshi and Minyechil Alehegn, "Analysis and prediction of diabetes diseases using machine learning algorithm": Ensemble approach, International Research Journal of Engineering and Technology Volume: 04 Issue: 10 | Oct - 2017.

The authors have proposed the ML techniques which are used to guess the data set at an initial phase to save the life. Using KNN and Naïve Bayes algorithm. In this study they proposed method provide high accuracy with accuracy value of 90.36% and decision Stump provided less accuracy than other by providing 83.72% accuracy.Random Forest, Naive Bayes, and KNN, are the most widely employed predictive algorithms here. The single algorithm offered less precision than ensemble one. The decision tree was highly accurate in most of the tests. Java and Weka are the tools in this hybrid study for predicting diabetes data. They proposed a theory based on Analysis and prediction of diabetes

diseases using machine learning algorithms: Ensemble approach. To make this system as an ensemble hybrid model, the following algorithms are used: KNN, Naive Bayes, Random forest and J48 which is used to increase the performance and accuracy. J48 is one of the most popular as well as better accuracy. All these algorithms are used to enhance the accuracy and all these are advanced when compared to others.

The random forest provides better accuracy than J48 as well as Naive Bayes in 10 cross-validation splitting methods. The fuzzy rule was developed to reduce the wrong treatment.

We can analyze the performance by using the result of this proposed theory

CHAPTER 3

METHODOLOGY

In this modern era, human beings encounter different health issues. Most of the health issues are due to the food habits of the individuals. In this project work, a predictive approach is proposed to pre-treat Diabetic Mellitus. The proposed approach has three phases namely data collection, data storage and analytics. This approach plays an important role in predicting diabetes and pre-treating diabetic patients. The phases in the proposed approach for diabetic prediction are presented in below figure

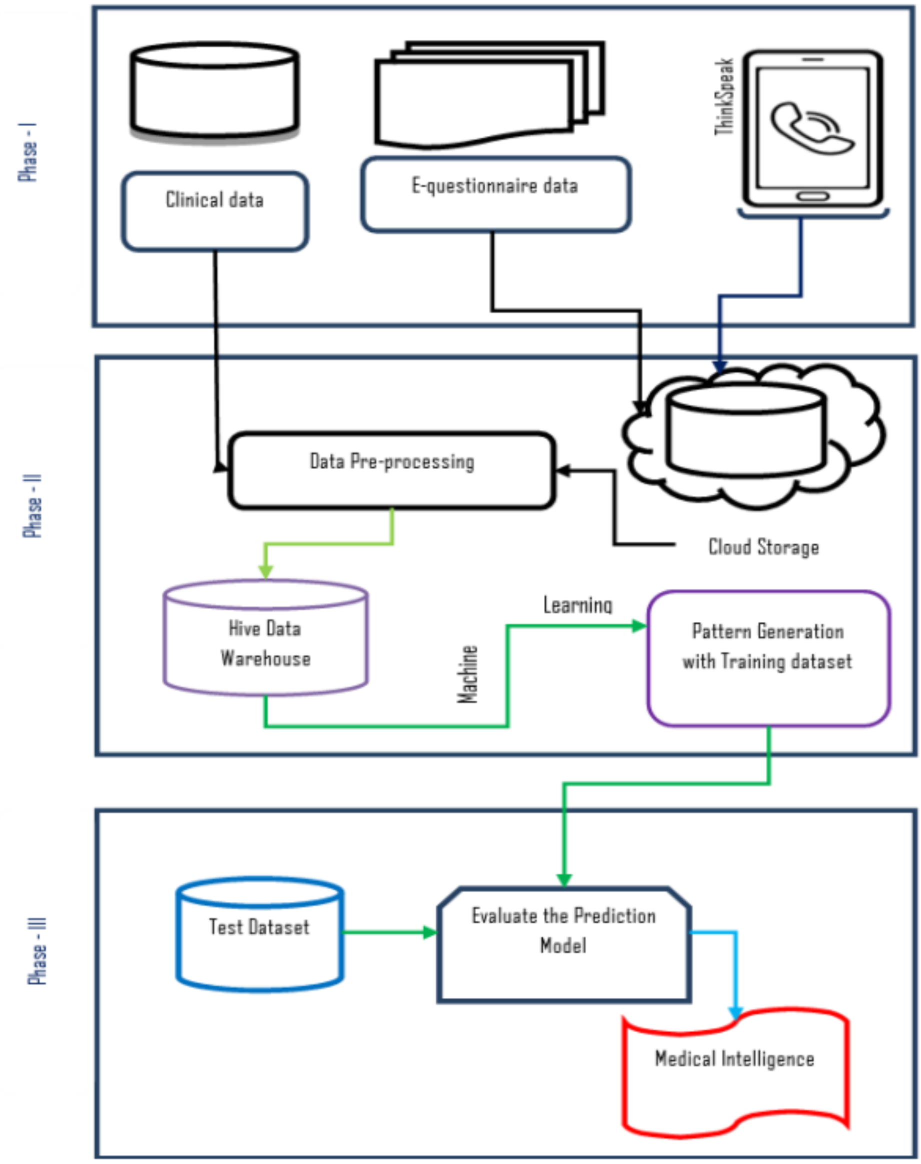


Fig 3.1 Methodological Diagram for Predicting Diabetes

In the first phase, data collection is done through IoT devices and other sources. The collected data are cleansed using pre-processing techniques. Phase two deals with data storage. The pre-processed data is stored in warehouses. To store massive amounts of data, cloud storage is used. The data stored in the cloud are analyzed to establish association between the various parameters such as BP, BMI, Air Pollution level etc., with Diabetic Mellitus. The third phase of the proposed approach deals with Predictive Analytics where the decisions are taken based on association rules with respect to diet pattern, physical fitness, current medicine intake etc.

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

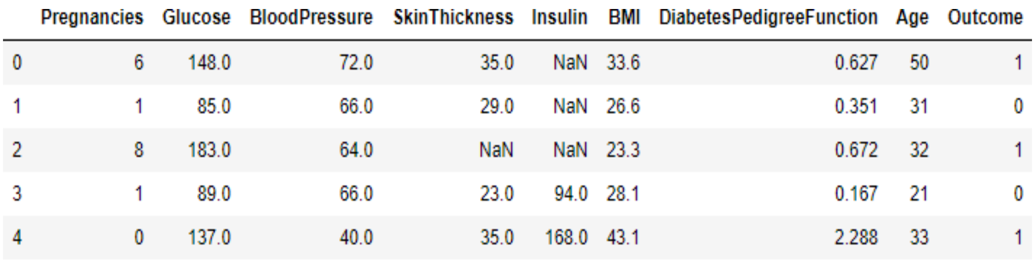
3.2 Data Pre-processing: This phase of model handles inconsistent data in order to get more accurate and precise results like in this dataset Id is inconsistent so we dropped the feature. This dataset doesn’t contain missing values. So, we imputed missing values for a few selected attributes like Glucose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

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

CHAPTER 4

DATA SET

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



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148.0	72.0	35.0	NaN	33.6	0.627	50	1
1	1	85.0	66.0	29.0	NaN	26.6	0.351	31	0
2	8	183.0	64.0	NaN	NaN	23.3	0.672	32	1
3	1	89.0	66.0	23.0	94.0	28.1	0.167	21	0
4	0	137.0	40.0	35.0	168.0	43.1	2.288	33	1

Fig 4.1 Sample dataset

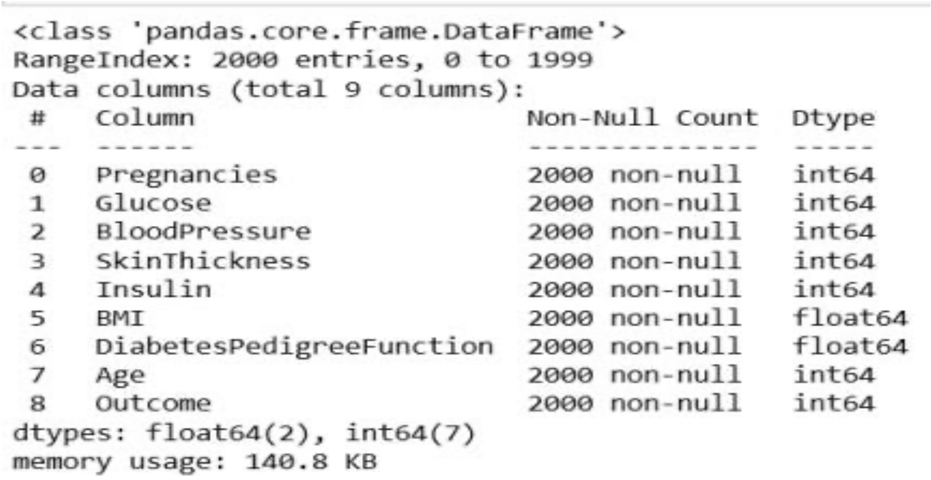
S.NO	Name	Description	Unit	Value Range
1	Pregnancy	No of times pregnant	Numeric value	0-9
2	Glucose	Glucose content	Numeric value	0-199
3	Blood Pressure	Diastole blood pressure	mmHg	0-122
4	Skin	Triceps skin fold thickness	Mm	0-99
5	Insulin	2-hours serum insulin	Mu/Uml	0-846
6	BMI	Body mass index	Weight in kg Height in m	0-67.1
7	Pedigree	Pedigree function	Numeric value	0.08-2.42
8	Age	Age	Numeric value	21-81

9	Class	Diabetes mellitus type 2	Numeric value	Positive=1
				Negative=0

Table 4.1 Dataset description, units and value range.

➔ The diabetes data set consists of 2000 data points, with 9 features each.

➔ “Outcome” is the feature we are going to predict, 0 means No diabetes, 1 means diabetes.



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

Fig 4.2 predictions

➔ There are no null values in the dataset.

4.1 Correlation matrix

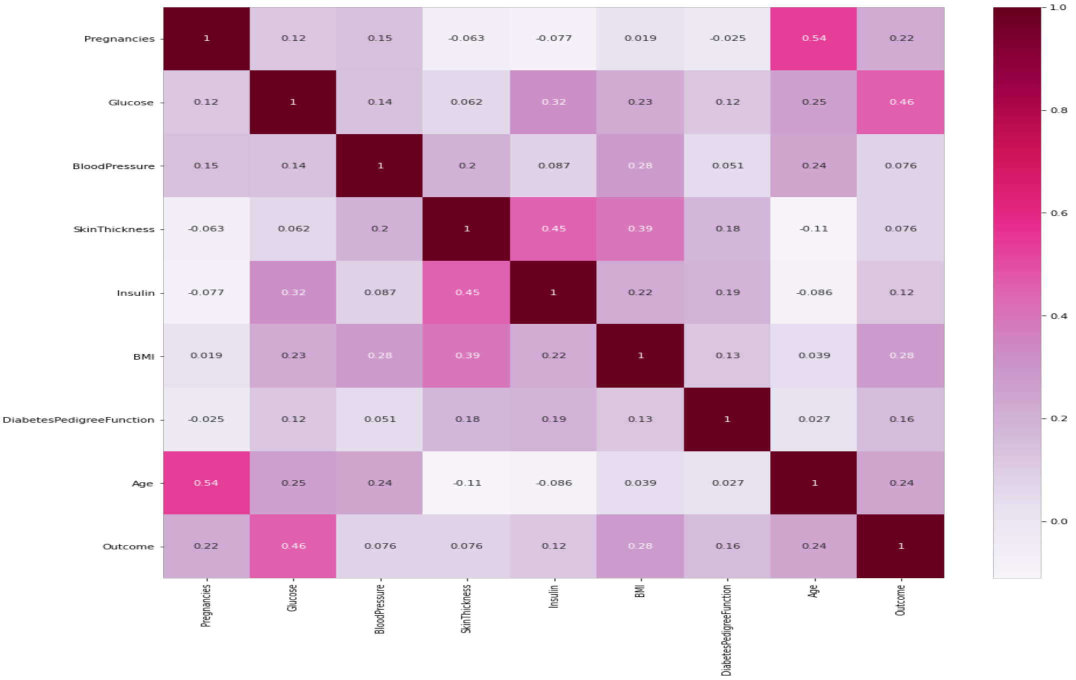


Fig 4.3 correlation matrix

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

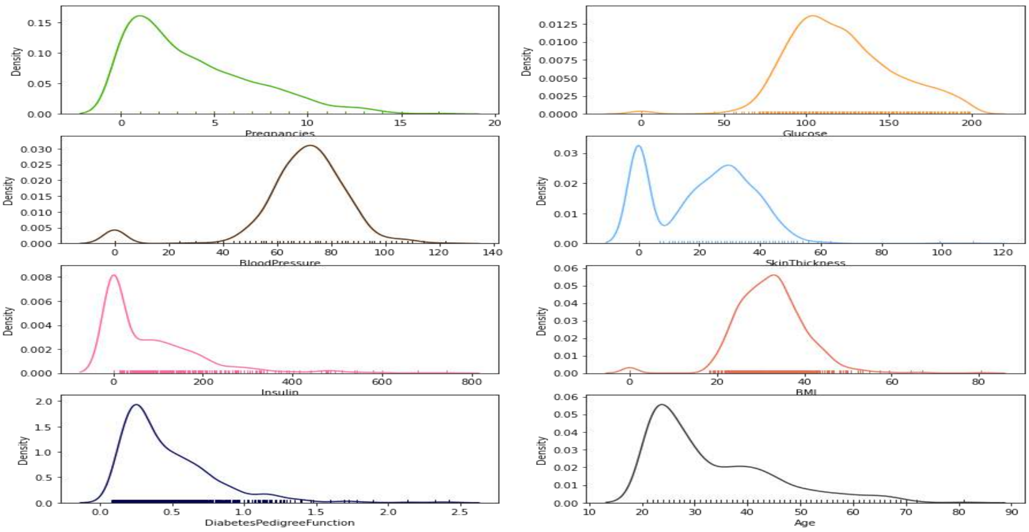
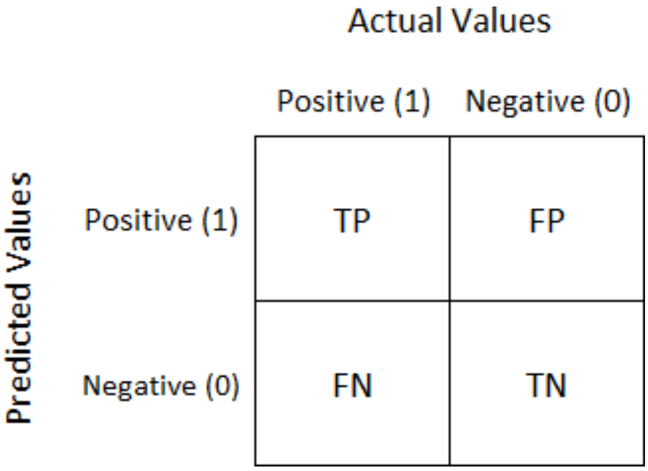


Fig 4.4 skew of data

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

4.2 Confusion matrix

Which provides an output matrix with complete description performance of the model?



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

Fig 4.5 Actual values

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

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

4.3 Data Visualization

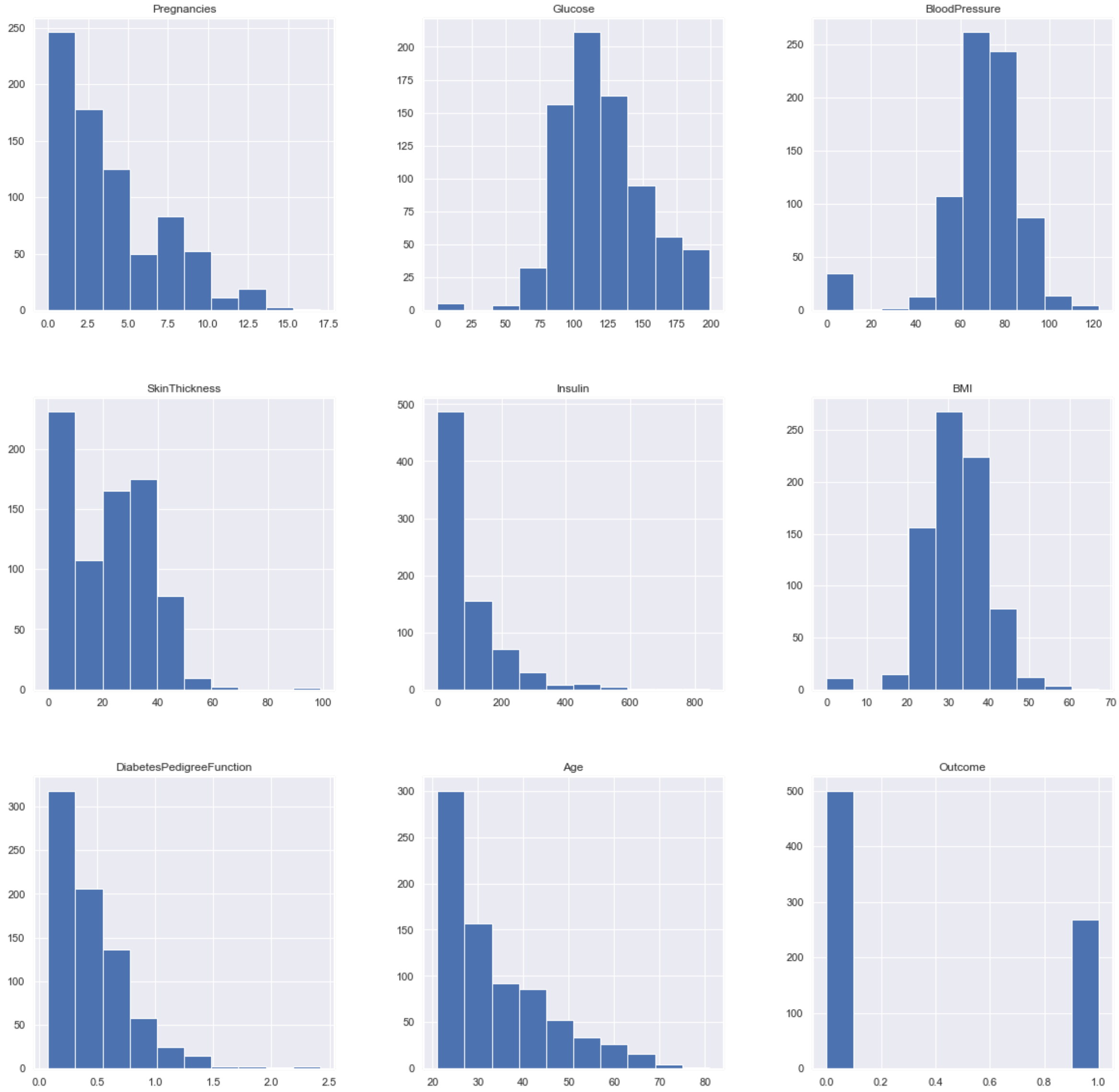
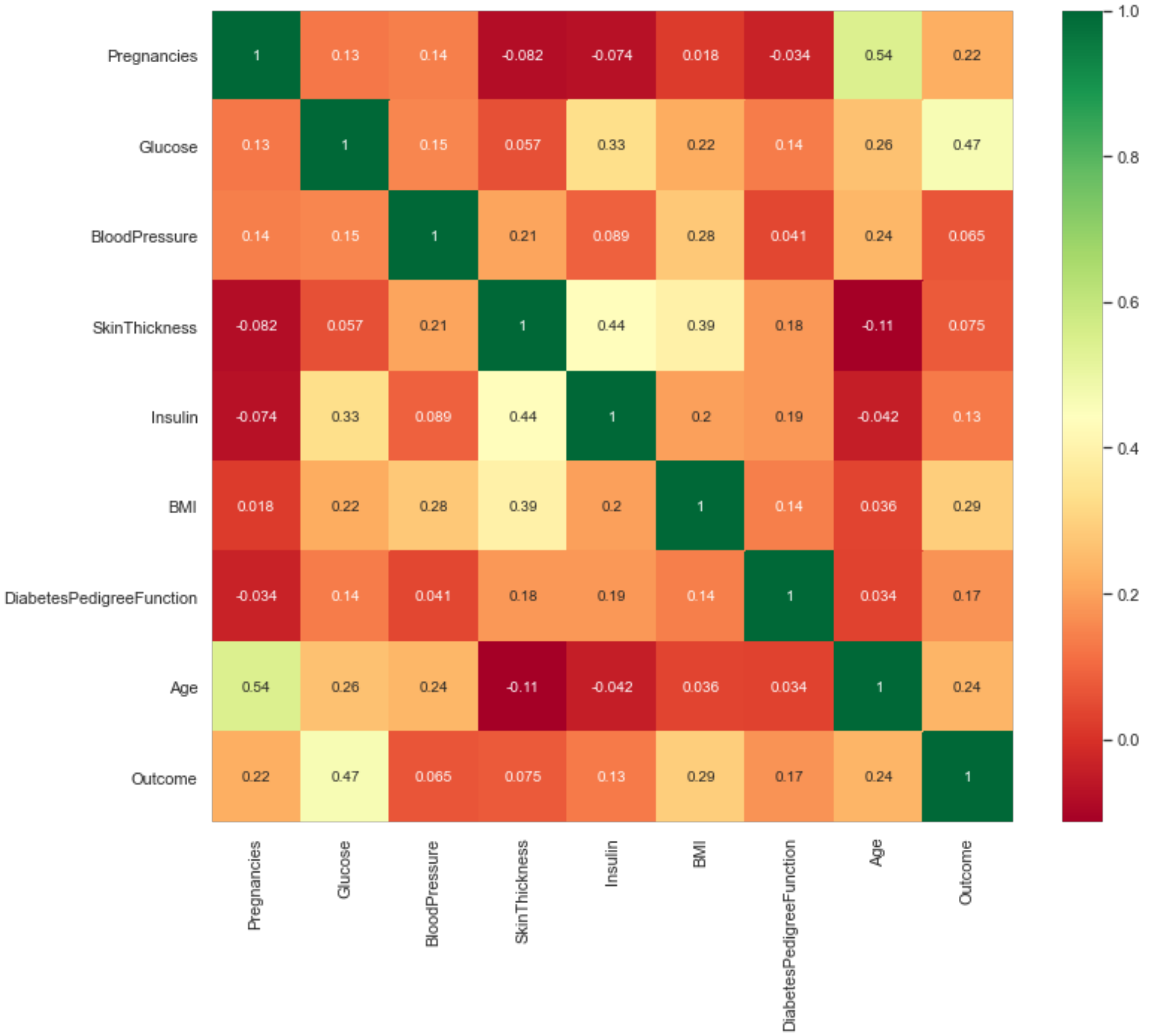


Fig 4.6 Data set graphs

4.4 Correlation between all the features



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1	0.13	0.14	-0.082	-0.074	0.018	-0.034	0.54	0.22
Glucose	0.13	1	0.15	0.057	0.33	0.22	0.14	0.26	0.47
BloodPressure	0.14	0.15	1	0.21	0.089	0.28	0.041	0.24	0.065
SkinThickness	-0.082	0.057	0.21	1	0.44	0.39	0.18	-0.11	0.075
Insulin	-0.074	0.33	0.089	0.44	1	0.2	0.19	-0.042	0.13
BMI	0.018	0.22	0.28	0.39	0.2	1	0.14	0.036	0.29
DiabetesPedigreeFunction	-0.034	0.14	0.041	0.18	0.19	0.14	1	0.034	0.17
Age	0.54	0.26	0.24	-0.11	-0.042	0.036	0.034	1	0.24
Outcome	0.22	0.47	0.065	0.075	0.13	0.29	0.17	0.24	1

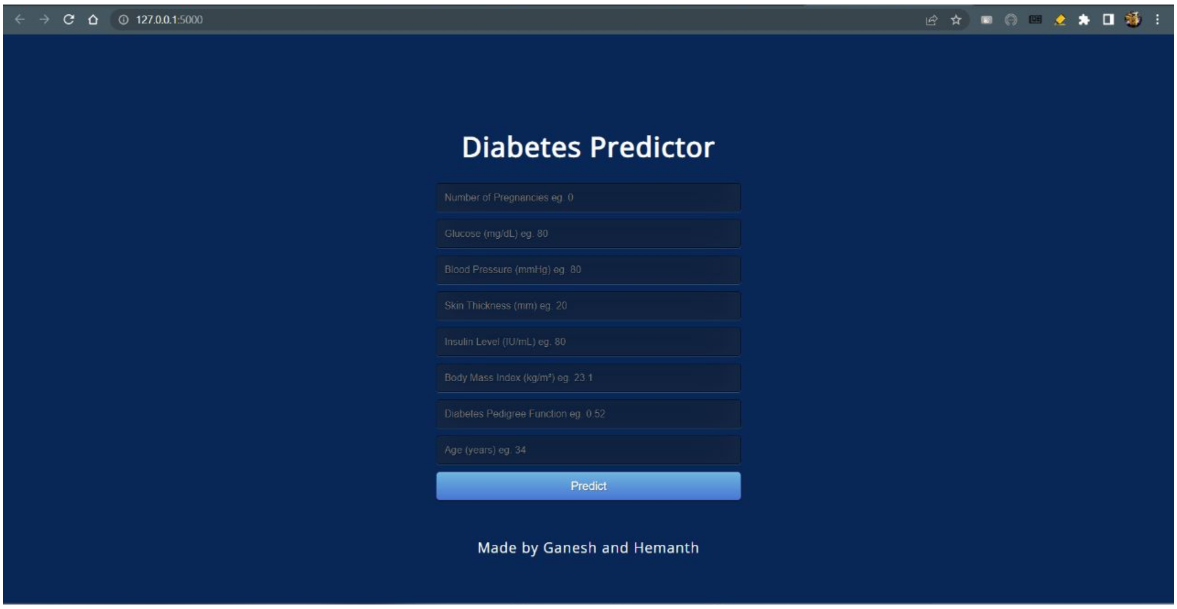
Fig 4.7 Correlation between all the features before cleaning

CHAPTER 5

RESULTS AND ANALYSIS

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

5.1 Index page of Diabetes Predictor



127.0.0.1:5000

Diabetes Predictor

Number of Pregnancies eg. 0

Glucose (mg/dL) eg. 80

Blood Pressure (mmHg) eg. 80

Skin Thickness (mm) eg. 20

Insulin Level (IU/mL) eg. 80

Body Mass Index (kg/m²) eg. 23.1

Diabetes Pedigree Function eg. 0.52

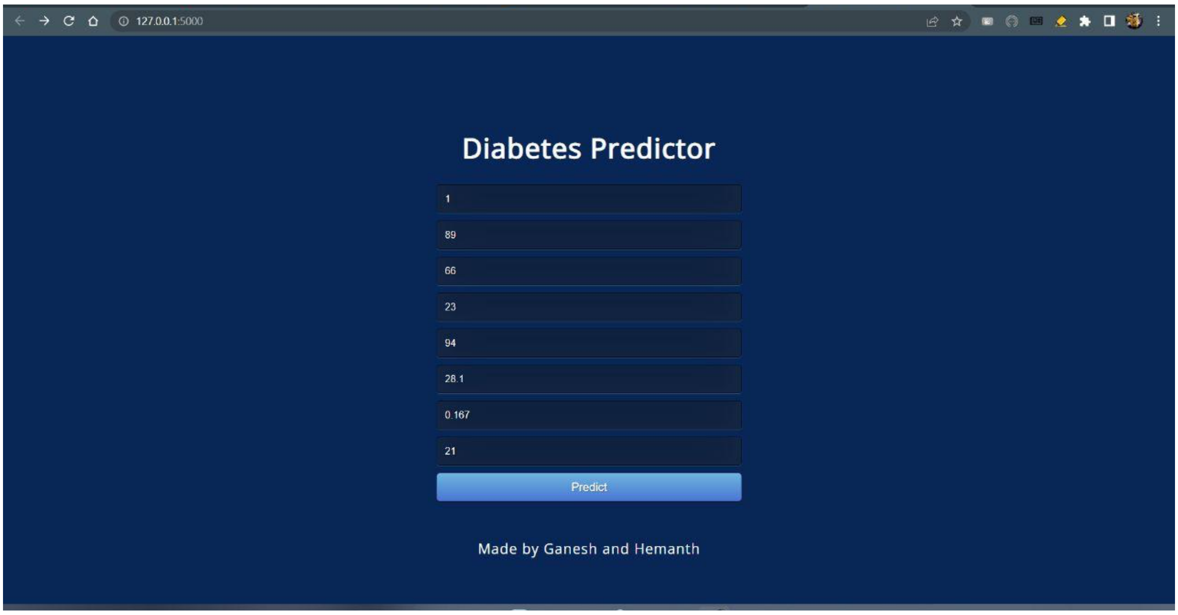
Age (years) eg. 34

Predict

Made by Ganesh and Hemanth

Fig 5.1: Basic Design of UI

5.2 Prediction input for non-diabetic person



127.0.0.1:5000

Diabetes Predictor

1

89

66

23

94

28.1

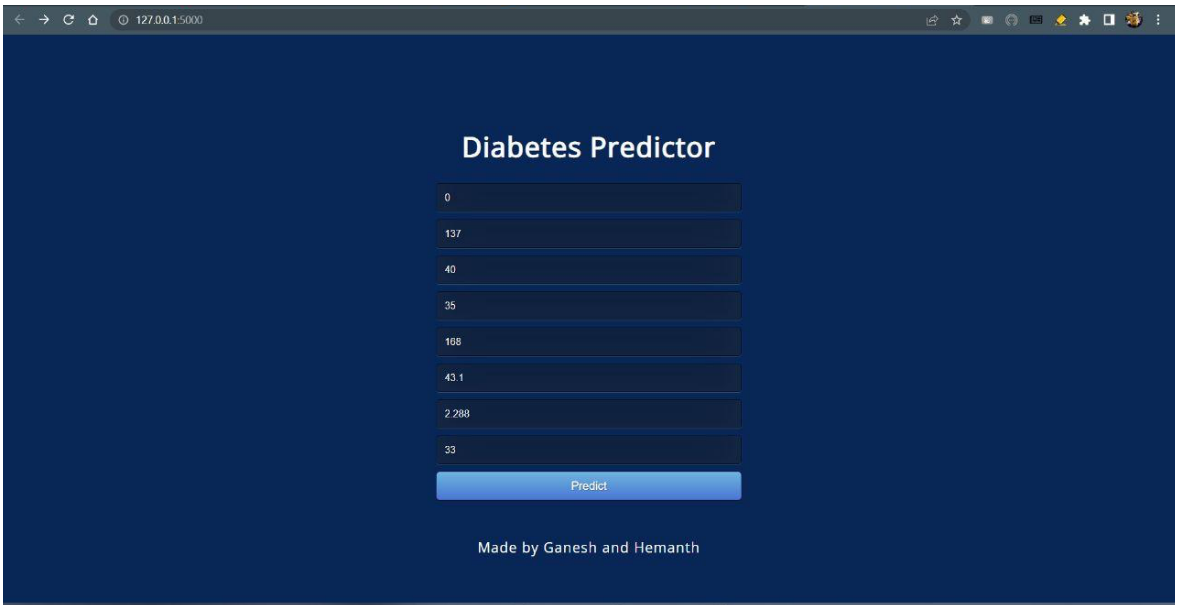
0.167

21

Made by Ganesh and Hemanth

Fig 5.2: Prediction input for non-diabetic person

5.3 Prediction input for diabetic person



127.0.0.1:5000

Diabetes Predictor

0

137

40

35

168

43.1

2.288

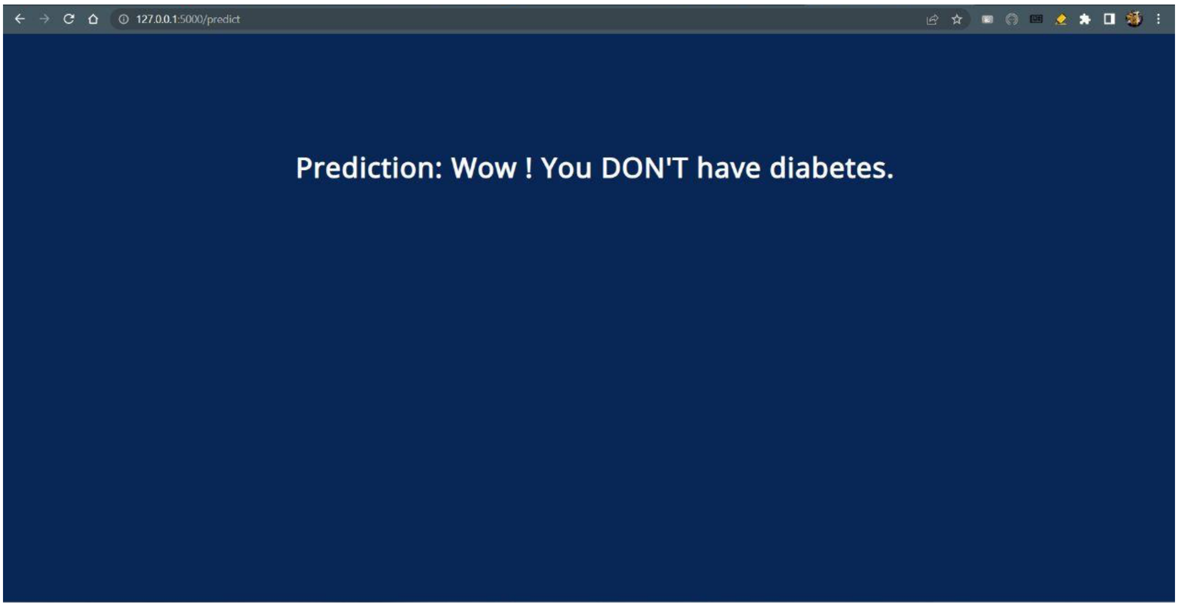
33

Made by Ganesh and Hemanth

Fig 5.3: Prediction input for diabetic person

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

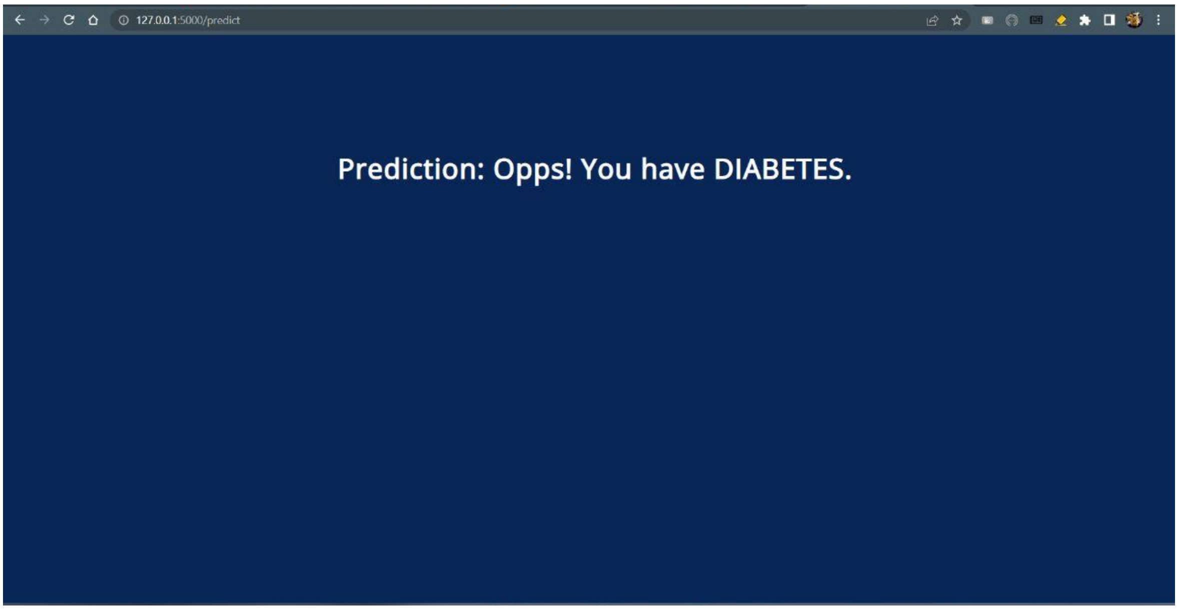
5.4 Prediction output for non-diabetic person



Prediction: Wow ! You DON'T have diabetes.

Fig 5.4: Prediction output for non-diabetic person

5.5 Prediction output for diabetic person



Prediction: Opps! You have DIABETES.

Fig 5.5: Prediction output for diabetic person

CHAPTER 6

CONCLUSION

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

REFERENCES

[1] Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar,” Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop”, International Conference On I-SMAC,978-1-5090-3243-3,2017.

[2] Ayush Anand and Divya Shakti,” Prediction of Diabetes Based on Personal Lifestyle Indicators”, 1st International Conference on Next Generation Computing Technologies, 978-1-4673-6809-4, September 2015.

[3] B. Nithya and Dr. V. Ilango,” Predictive Analytics in Healthcare Using Machine Learning Tools and Techniques”, International Conference on Intelligent Computing and Control Systems, 978-1-5386-2745-7,2017.

[4] Dr Saravana kumar N M, Eswari T, Sampath P and Lavanya S,” Predictive Methodology for Diabetic Data Analysis in Big Data”, 2nd International Symposium on Big Data and Cloud Computing,2015.

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

[6] P. Suresh Kumar and S. Pranavi “Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics”, International Conference on Infocom Technologies and Unmanned Systems, 978-1-5386-0514-1, Dec. 18-20, 2017.

[7] Mani Butwall and Shraddha Kumar,” A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier”, International Journal of Computer Applications, Volume 120 - Number 8,2015.

[8] K. Rajesh and V. Sangeetha, “Application of Data Mining Methods and Techniques for Diabetes Diagnosis”, International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.

[9]Humar Kahramanli and Novruz Allahverdi,”Design of a Hybrid System for Diabetes and Heart Disease”, Expert Systems with Applications: An International Journal, Volume 35 Issue 1-2, July, 2008.

[10] B.M. Patil, R.C. Joshi and Durga Toshniwal,”Association Rule for Classification of Type-2 Diabetic Patients”, ICMLC '10 Proceedings of the 2010 Second International Conference on Machine Learning and Computing, February 09 - 11, 2010.

[11] Dost Muhammad Khan1, Nawaz Mohamudally2, “An Integration of K-means and Decision Tree (ID3) towards a more Efficient Data Mining Algorithm ”, Journal Of Computing, Volume 3, Issue 12, December 2011.

APPENDIX

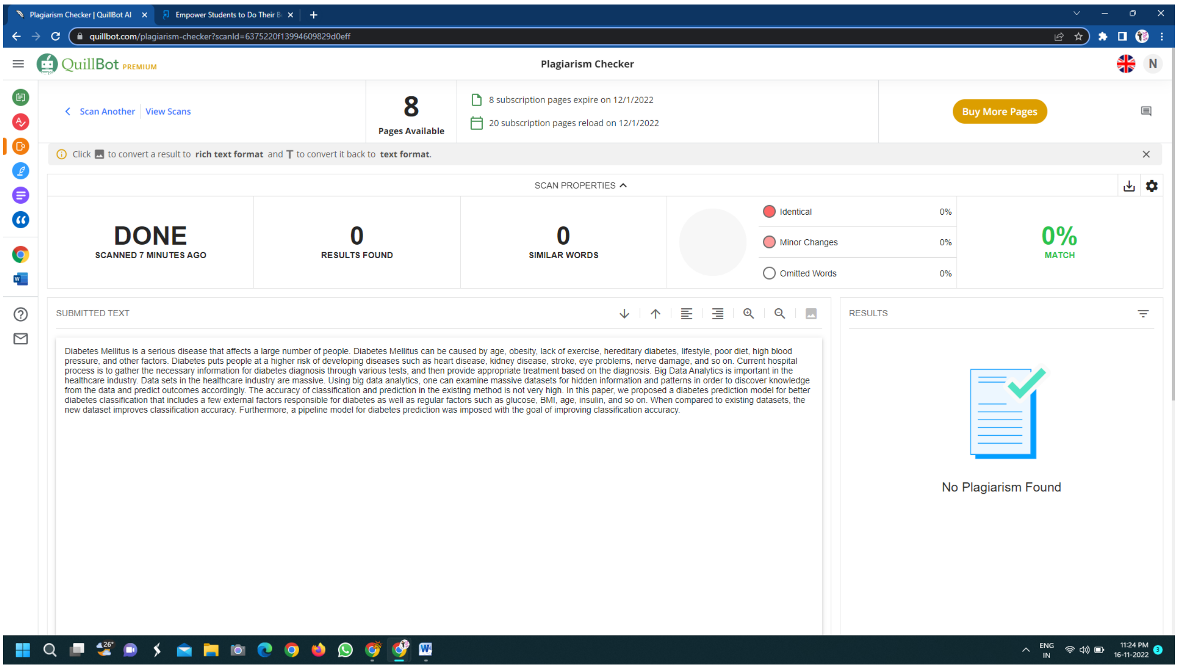
```
# Importing essential libraries
import numpy as np
import pandas as pd
import pickle

# Loading the dataset
df = pd.read_csv('kaggle_diabetes.csv')

# Renaming DiabetesPedigreeFunction as DPF
df = df.rename(columns={'DiabetesPedigreeFunction':'DPF'})

# Replacing the 0 values from ['Glucose','BloodPressure','SkinThickness','Insulin','BMI']
by NaN
df_copy = df.copy(deep=True)
df_copy[['Glucose','BloodPressure','SkinThickness','Insulin','BMI']] =
df_copy[['Glucose','BloodPressure','SkinThickness','Insulin','BMI']].replace(0,np.NaN)

# Replacing NaN value by mean, median depending upon distribution
df_copy['Glucose'].fillna(df_copy['Glucose'].mean(), inplace=True)
df_copy['BloodPressure'].fillna(df_copy['BloodPressure'].mean(), inplace=True)
df_copy['SkinThickness'].fillna(df_copy['SkinThickness'].median(), inplace=True)
df_copy['Insulin'].fillna(df_copy['Insulin'].median(), inplace=True)
df_copy['BMI'].fillna(df_copy['BMI'].median(), inplace=True)

# 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)

# Creating a pickle file for the classifier
filename = 'diabetes-prediction-rfc-model.pkl'
pickle.dump(classifier, open(filename, 'wb'))




# Importing essential libraries
from flask import Flask, render_template, request
import pickle
import numpy as np


# Load the Random Forest CLassifier model
filename = 'diabetes-prediction-rfc-model.pkl'
classifier = pickle.load(open(filename, 'rb'))

app = Flask(__name__)

@app.route('/')
def home():
        return render_template('index.html')

@app.route('/predict', methods=['POST'])
def predict():
    if request.method == 'POST':
        preg = int(request.form['pregnancies'])
        glucose = int(request.form['glucose'])
        bp = int(request.form['bloodpressure'])
        st = int(request.form['skinthickness'])
        insulin = int(request.form['insulin'])
        bmi = float(request.form['bmi'])
        dpf = float(request.form['dpf'])
        age = int(request.form['age'])

        data = np.array([[preg, glucose, bp, st, insulin, bmi, dpf, age]])
        my_prediction = classifier.predict(data)

        return render_template('result.html', prediction=my_prediction)

if __name__ == '__main__':
    app.run(debug=True, use_reloader=False)
```

PAPER PUBLICATION STATUS

Paper not yet published in any Conference. We will publish the paper this week at the IEEE International conference.

PLAGIARISM REPORT



Plagiarism Checker | QuillBot AI
Empower Students to Do Their B
quillbot.com/plagiarism-checker?scanId=6375220f13994609829d0eff
QuillBot PREMIUM
N
Scan Another
View Scans
Buy More Pages
SCAN PROPERTIES
SUBMITTED TEXT
RESULTS
ENG IN
11:24 PM
16-11-2022

Plagiarism Checker

8 Pages Available

8 subscription pages expire on 12/1/2022

20 subscription pages reload on 12/1/2022

Click to convert a result to **rich text format** and T to convert it back to **text format**.

DONE SCANNED 7 MINUTES AGO

0 RESULTS FOUND

0 SIMILAR WORDS

Identical	0%
Minor Changes	0%
Omitted Words	0%

0% MATCH

Diabetes Mellitus is a serious disease that affects a large number of people. Diabetes Mellitus can be caused by age, obesity, lack of exercise, hereditary diabetes, lifestyle, poor diet, high blood pressure, and other factors. Diabetes puts people at a higher risk of developing diseases such as heart disease, kidney disease, stroke, eye problems, nerve damage, and so on. Current hospital process is to gather the necessary information for diabetes diagnosis through various tests, and then provide appropriate treatment based on the diagnosis. Big Data Analytics is important in the healthcare industry. Data sets in the healthcare industry are massive. Using big data analytics, one can examine massive datasets for hidden information and patterns in order to discover knowledge from the data and predict outcomes accordingly. The accuracy of classification and prediction in the existing method is not very high. In this paper, we proposed a diabetes prediction model for better diabetes classification that includes a few external factors responsible for diabetes as well as regular factors such as glucose, BMI, age, insulin, and so on. When compared to existing datasets, the new dataset improves classification accuracy. Furthermore, a pipeline model for diabetes prediction was imposed with the goal of improving classification accuracy.

No Plagiarism Found