A REPORT OF PROJECT
On
Diabetes Prediction

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE AWARD

OF THE DEGREE OF

BACHELOR OF ENGINEERING

(Computer Science & Engineering)

JAN-MAY, 2019

SUBMITTED BY:

ABHISHEK JOSHI
ISHAAN

UNIVERSITY UID:-

16BCS3171
16BCS1920

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CHANDIGARH UNIVERSITY GHARUAN, MOHALI

CONTENTS

CHAPTER 1 INTRODUCTION

CHAPTER 2 TRAINING WORK UNDERTAKEN

CHAPTER 3 RESULTS AND DISCUSSION

CHAPTER 4 CONCLUSION AND FUTURE SCOPE

CANDIDATE’S DECLARATION

I “ABHISHEK JOSHI” and “ISHAAN” hereby declare that we have worked for the whole semester on a project titles as “Diabetes Prediction” during a period from 2 JAN 2018 to 12 APRIL 2019 in partial fulfillment of requirements for the award of degree of B.E (COMPUTER SCIENCE & ENGINEERING) at CHANDIGARH UNIVERSITY GHARUAN, MOHALI. The work which is being presented in the report submitted to Department of Computer Science & Engineering at CHANDIGARH UNIVERSITY GHARUAN, MOHALI is an authentic record of training work.

Signature of the Student

ABSTRACT

A major challenge facing healthcare organizations (hospitals, medical centre) is the provision of quality services at affordable costs. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. Hospitals must also minimize the cost of clinical tests. They can achieve these results by employing appropriate computer-based information and/or decision support systems. Most hospitals today employ some sort of hospital information systems to manage their healthcare or patient data.

These systems typically generate huge amounts of data which take the form of numbers, text, charts and images. Unfortunately, these data are rarely used to support clinical decision making. There is a wealth of hidden information in these data that is largely untapped. This raises an important question: “How can we turn data into useful information that can enable healthcare practitioners to make intelligent clinical decisions?” Although data mining has been around for more than two decades, its potential is only being realized now. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. The two most common modelling objectives are classification and prediction. Classification models predict categorical labels (discrete, unordered) while prediction models predict continuous-valued functions. Decision Trees and Neural Networks use classification algorithms while Regression, Association Rules and Clustering use prediction algorithms.

ACKNOWLEDGEMENT

We would like to express my special thanks of gratitude to my mentor (Jyoti Mam) as well as our project teacher(Trapti Mam) who gave us the golden opportunity to do this wonderful project on the topic (Diabetes Prediction), which also helped me in doing a lot of Research and I came to know about so many new things I am really thankful to them. Secondly I would also like to thank my parents and friends who helped me a lot in finalizing this project within the limited time frame.

LIST OF FIGURES

1. INTRODUCTION

1.1 SOFTWARE/HARDWARE DETAIL

1.1.1 HARDWARE DETAILS

1. 4 GB RAM

2. Laptop / Desktop with supported internet

1.1.2 SOFTWARE REQUIRED

1. Python 3.5.1

2. Sqlite

3. Anaconda 3.7

4.Apache Server

1.2 BACKGROUND OF THE PROJECT

(Diabetes Prediction)

A major challenge facing healthcare organizations (hospitals, medical centers) is the provision of quality services at affordable costs. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. Hospitals must also minimize the cost of clinical tests. They can achieve these results by employing appropriate computer-based information and/or decision support systems. Most hospitals today employ some sort of hospital information systems to manage their healthcare or patient data. These systems typically generate huge amounts of data which take the form of numbers, text, charts and images. Unfortunately, these data are rarely used to support clinical decision making. There is a wealth of hidden information in these data that is largely untapped. This raises an important question: “How can we turn data into useful information that can enable healthcare practitioners to make intelligent clinical decisions?” Although data mining has been around for more than two decades, its potential is only being realized now. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. The two most common modeling objectives are classification and prediction. Classification models predict categorical labels (discrete, unordered) while prediction models predict continuous-valued functions. Decision Trees and Neural Networks use classification algorithms while Regression, Association Rules and Clustering use prediction algorithms.

2.2 CODING IN WEBSITE

(USING PYTHON)

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. It was created by Guido van Rossum during 1985- 1990. Like Perl, Python source code is also available under the GNU General Public License (GPL). This tutorial gives enough understanding on Python programming language

Python is derived from many other languages, including ABC, Modula-3, C, C++, Algol-68, SmallTalk, and Unix shell and other scripting languages.

The concepts and rules used in python programming provide these important benefits:

- Interactive
- Interpreted
- Modular
- Dynamic
- Object-oriented
- Portable
- High level
- Extensible in C++ & C

2.3 DATABASE (STRUCTURED QUERY LANGUAGE)

Structure Query Language (SQL) is a database query language used for storing and managing data in Relational DBMS. SQL was the first commercial language introduced for E.F Codd's **Relational** model of database. Today almost all RDBMS (MySQL, Oracle, Informix, Sybase, MS Access) use **SQL** as the standard database query language. SQL is used to perform all types of data operations in RDBMS.

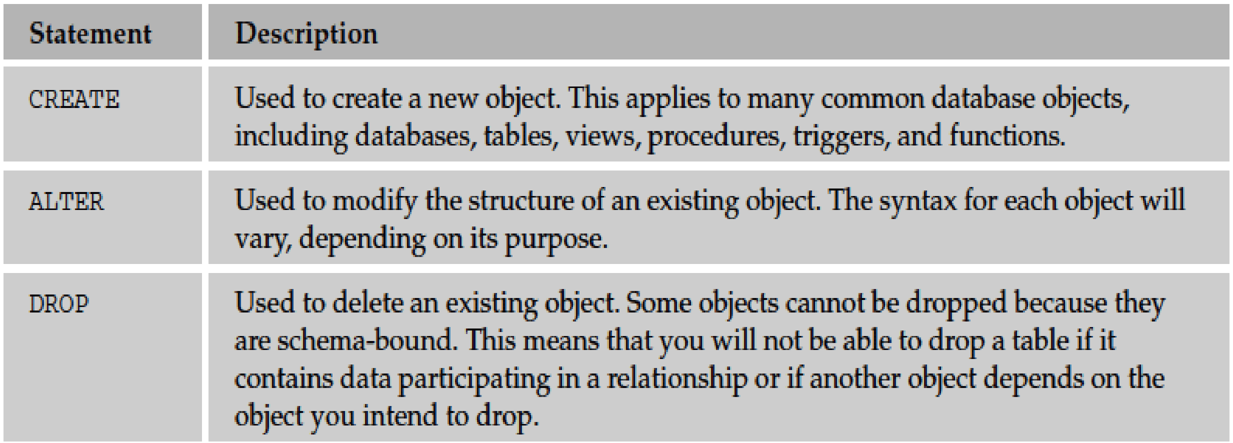
SQL Command

SQL defines following ways to manipulate data stored in an RDBMS.

DDL: Data Definition Language

This includes changes to the structure of the table like creation of table, altering table, deleting a table etc.

All DDL commands are auto-committed. That means it saves all the changes permanently in the database.



Statement	Description
CREATE	Used to create a new object. This applies to many common database objects, including databases, tables, views, procedures, triggers, and functions.
ALTER	Used to modify the structure of an existing object. The syntax for each object will vary, depending on its purpose.
DROP	Used to delete an existing object. Some objects cannot be dropped because they are schema-bound. This means that you will not be able to drop a table if it contains data participating in a relationship or if another object depends on the object you intend to drop.

DML: Data Manipulation Language

DML commands are used for manipulating the data stored in the table and not the table itself. DML commands are not auto-committed. It means changes are not permanent to database, they can be rolled back.

Command	Description
Insert	to insert a new row
Update	to update existing row
Delete	to delete a row
Merge	merging two rows or two tables

DQL: Data Query Language

Data query language is used to fetch data from tables based on conditions that we can easily apply.

Command	Description
Select	retrieve records from one or more table

CONNECTIVITY

1. Django determines the root URLconf module to use. Ordinarily, this is the value of the **ROOT_URLCONF** setting, but if the incoming **HttpRequest** object has a **urlconf** attribute (set by middleware), its value will be used in place of the**ROOT_URLCONF** setting.

2. Django loads that Python module and looks for the variable **urlpatterns**. This should be a Python list of **django.urls.path ()** and/or **django.urls.re_path ()** instances.

3. Django runs through each URL pattern, in order, and stops at the first one that matches the requested URL.

4. Once one of the URL patterns matches, Django imports and calls the given view, which is a simple Python function (or a class-based view). The view gets passed the following arguments:

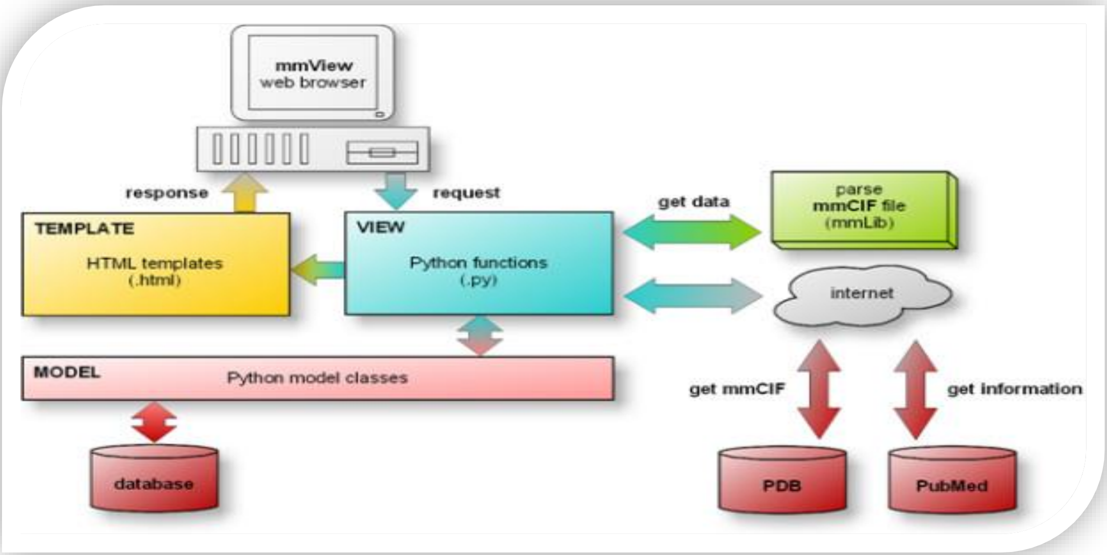
    - An instance of **HttpRequest**.

    - If the matched URL pattern returned no named groups, then the matches from the regular expression are provided as positional arguments.

    - The keyword arguments are made up of any named parts matched by the path expression, overridden by any arguments specified in the optional **kwargs** argument to **django.urls.path()** or **django.urls.re_path()**.

5. If no URL pattern matches, or if an exception is raised during any point in this process, Django invokes an appropriate error-handling view. See Error handling belo

- To capture a value from the URL, use angle brackets.
- Captured values can optionally include a converter type. For example, use **<int:name>** to capture an integer parameter. If a converter isn't included, any string, excluding a / character, is matched.



3. RESULTS AND DISCUSSIONS

3.1 RESULTS

3.1.1 BENEFIT OF PROJECT:-

This project results in enhancement of the DATA ANALYSIS in the can be viewed as a vast and way better application for users to save time in order to fetch the data from any website in minutes.

3.1.2 RESULT ON INDIVIDUAL DEVELOPMENT:-

On an individual basis this project helped me a lot in understanding concepts of **Python**. By this I was able to explore the use of skill i.e Data Analysis, Anaconda and in enhancement the concepts of hybrid programing.

At the due of all these things, I am able to create web applications using Django and Python.

3.2 DISCUSSIONS OF PROJECT

Exploring Project Problem

Diabetes and pre-diabetes are serious conditions in which people have high levels of sugar or glucose in their blood. The World Health Organization (WHO) reports that 422 million people worldwide had diabetes in 2014. In the US, according to the US Centers for Disease Control and Prevention (CDC), over 29 million people have diabetes and at least 86 million adults over 20 have pre-diabetes (blood sugar levels are higher than normal, but not high enough to be diagnosed with type 2 diabetes). Diabetes is a major cause of blindness, amputation, kidney failure, and cardiovascular disease.

Analyzing the Problem

The diabetes is classified in mainly two types and these are described below:

Type 1 Diabetes (Insulin Dependent)

- The pancreas does not make any insulin
- Person must take insulin every day to survive
- Usually begins in childhood or adolescence

Type 2 Diabetes (Non-insulin Dependent)

- The pancreas makes some insulin, and the body does not respond normally to the insulin your body does make
- Some people with type 2 diabetes are able to control it with diet and exercise; many others need diabetes medication, and some need insulin
- Most common form of diabetes

Defining The Problem

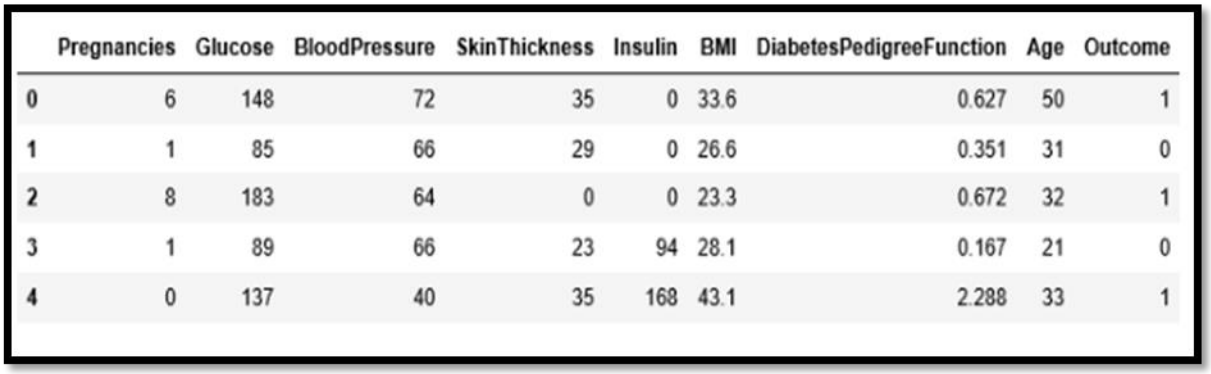
The diabetes is a severe problem that can be detected by finding a specific pattern in the glucose level and blood sugar level. The main is to use a dataset of the UCI Machine Learning Repository to analyze the visual patterns in the dataset to define the particular range that describe the possibility of presence of diabetes. This aim was achieved using Data mining and Machine learning algorithm to make a efficient model that best classify a sample for diabetes positive and diabetes negative

NEED OF PROJECT

Diabetes and pre-diabetes are serious conditions in which people have high levels of sugar or glucose in their blood. The World Health Organization (WHO) reports that 422 million people worldwide had diabetes in 2014. In the US, according to the US Centres for Disease Control and Prevention (CDC), over 29 million people have diabetes and at least 86 million adults over 20 have pre-diabetes (blood sugar levels are higher than normal, but not high enough to be diagnosed with type 2 diabetes). Diabetes is a major cause of blindness, amputation, kidney failure, and cardiovascular disease

DATASET USED

The diabetes is a severe problem that can be detected by finding a specific pattern in the glucose level and blood sugar level. The main is to use a dataset of the UCI Machine Learning Repository to analyze the visual patterns in the dataset to define the particular range that describe the possibility of presence of diabetes. This aim was achieved using Data mining and Machine learning algorithm to make a efficient model that best classify a sample for diabetes positive and diabetes negative



	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

METHODOLOGY

A. Data cleaning

The collected data is messy and need to be pre-processed before using it as for modelling purpose. The data consist of the various anomalies that needed to be resolved before data modelling. Most of the researchers spend two third of their time in data cleaning. It is mandatory step to pre-process and normalize the data before data analysis. The main aim of this phase is to convert the unstandardized data into standardized data.

B. Class Balancing

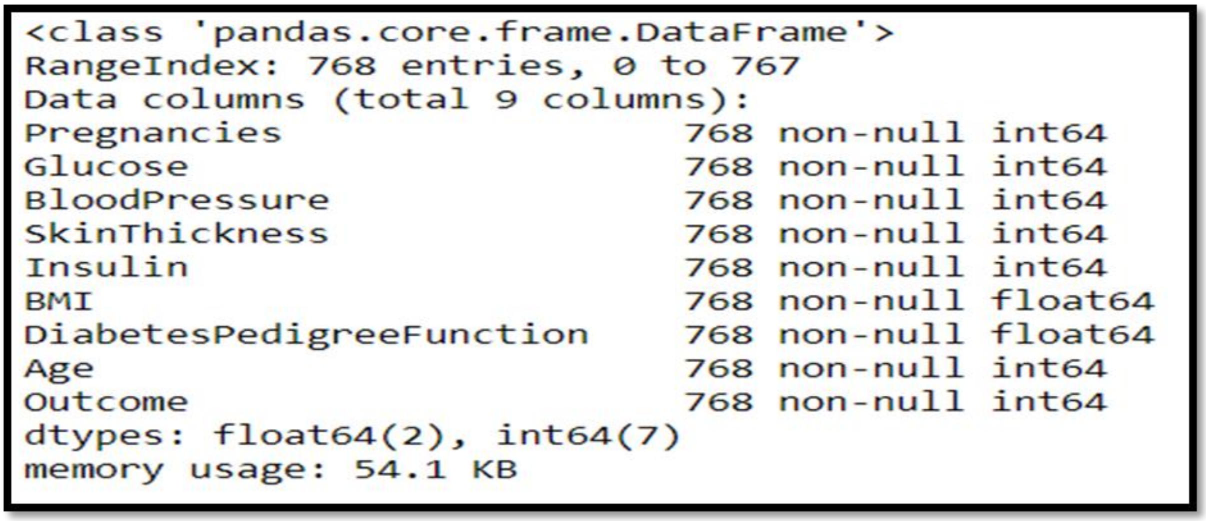
Class balancing is a required feature to avoid the biasing of the data. The data with low bias and high variance is considered good for analytics purpose. The balanced data can provide very efficient and useful insights. SMOTE is a Synthetic Minority Over Sampling Technique that is used to generate a synthetic sample by the fusion of two randomly fetched nearest neighbour samples from a minority class samples.

C. Data Preprocessing

The dataset is normalized to attain a persistent accuracy. The dataset is reshuffled and resampled over a count of 10 to avoid under fitting or overfitting of the data. The throughput and data flow rate are most contributing over the protocol. There is a consistent data flow rate whereas there are sharp up and down in the throughput.

Analysis Report

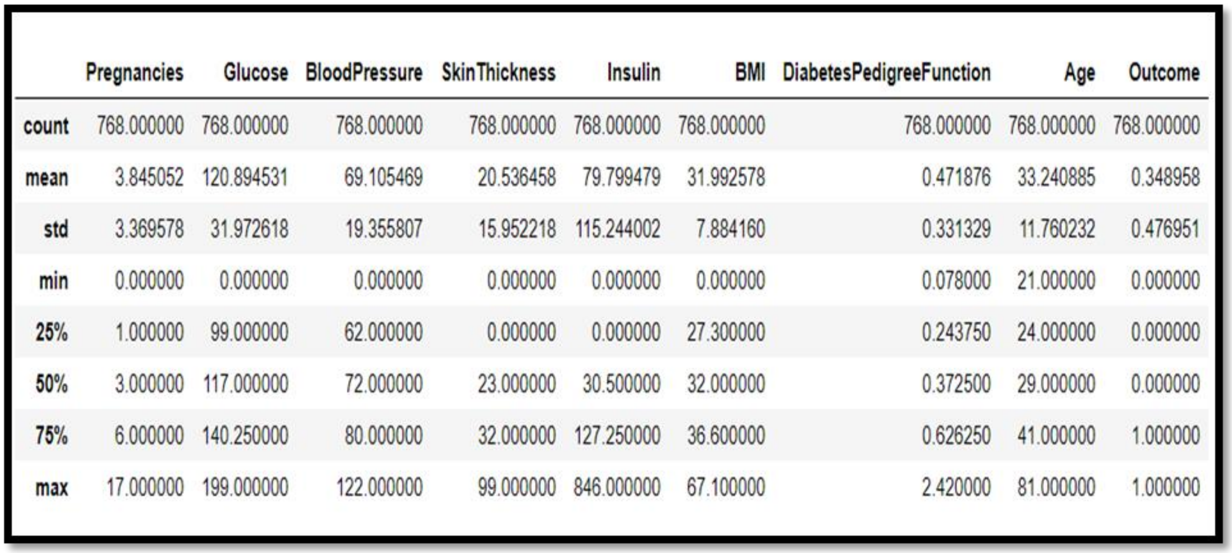
Overview of Dataset: - Here is the description about the dataset that are used for data mining. It consists of the 9 attributes that are recorded for the analysis of the diabetes.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
Pregnancies                 768 non-null int64
Glucose                     768 non-null int64
BloodPressure               768 non-null int64
SkinThickness               768 non-null int64
Insulin                     768 non-null int64
BMI                         768 non-null float64
DiabetesPedigreeFunction    768 non-null float64
Age                         768 non-null int64
Outcome                     768 non-null int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

As we can clearly see that all the attributes are in the numerical form and BMI and Function are float and rest are the integer datatype.

Statistical View of Dataset: - Here is the description of the statistical parameters such as the mean median mode and standard deviation.



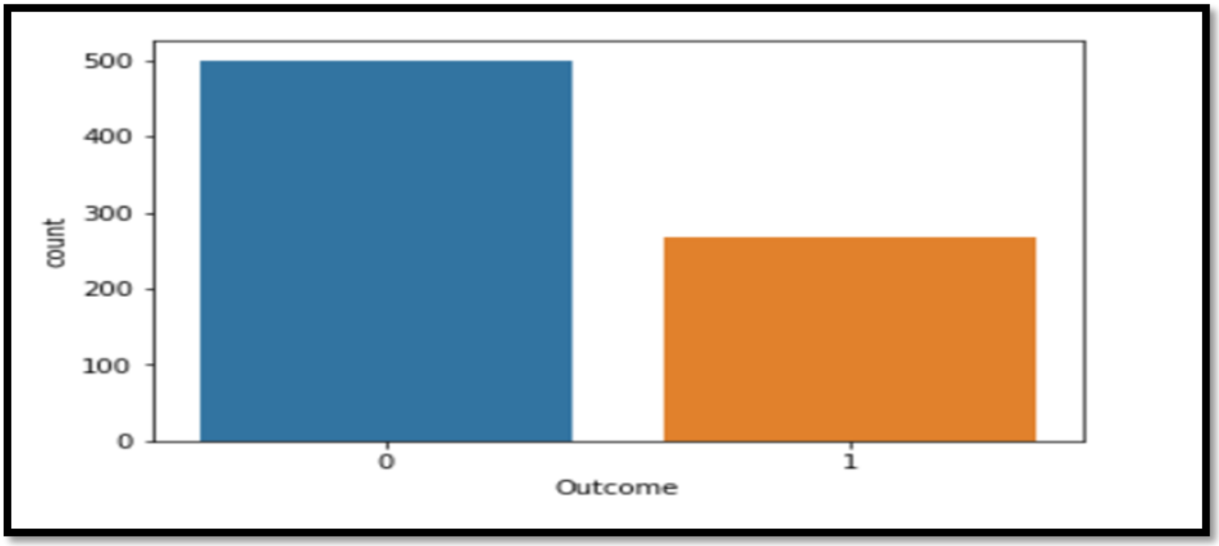
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

As we can see that standard deviation in Glucose, BMI, Insulin is very large. It shows that there are possible attribute that needed to be investigated for the outliers.

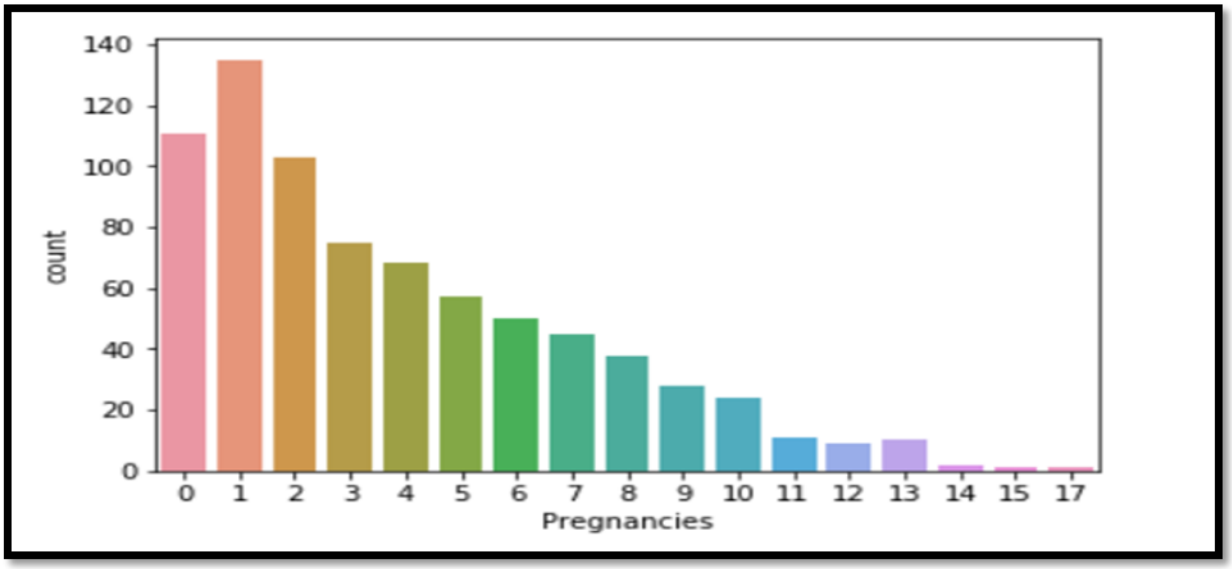
As we can clearly see that count is 768 for each attribute so means this dataset does not have a missing data.

The mean is value that can be used to provide a central median limit for the clustering part.

Closure look at Outcome: -Here is the visualisation describing the outcome classes binary count and we can clearly see that it has more biasing over the cases that are negative of diabetes prediction.

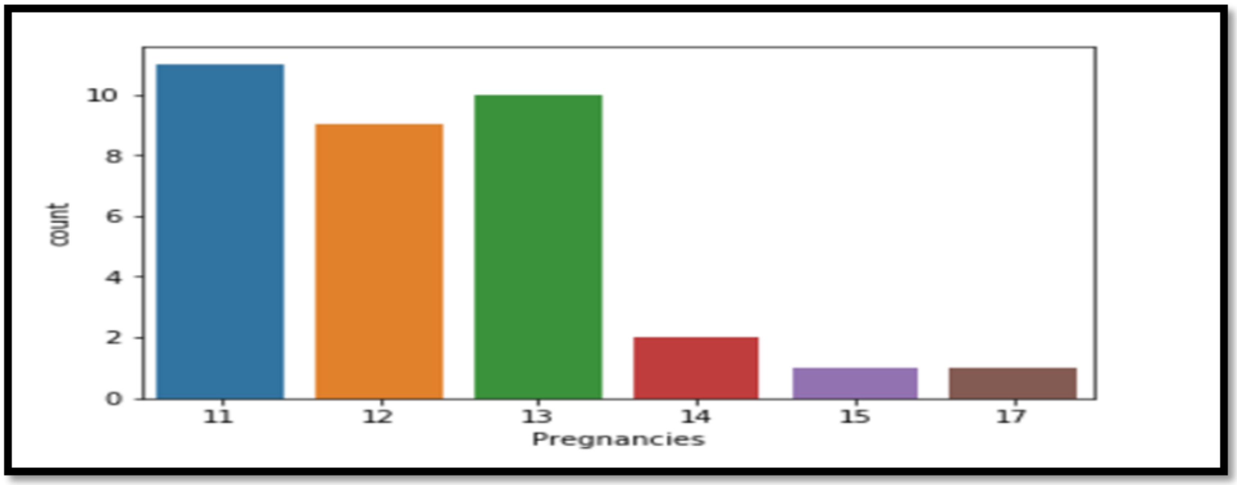


Closure look at Pregnancies: - Here is what we found during a closure look at attribute showing the pregnancies.



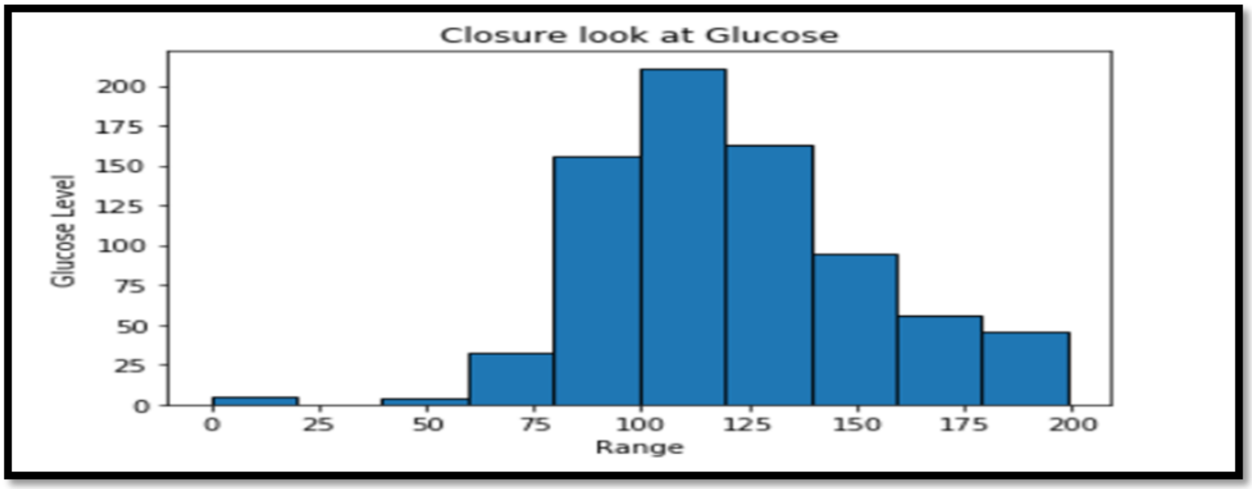
It can be clearly seen that this attribute contains the outlier that provide an extreme failure over diabetes prediction.

We assumed that the at most pregnancies can be 10 and rest others are considered as faulty outliers in the data.



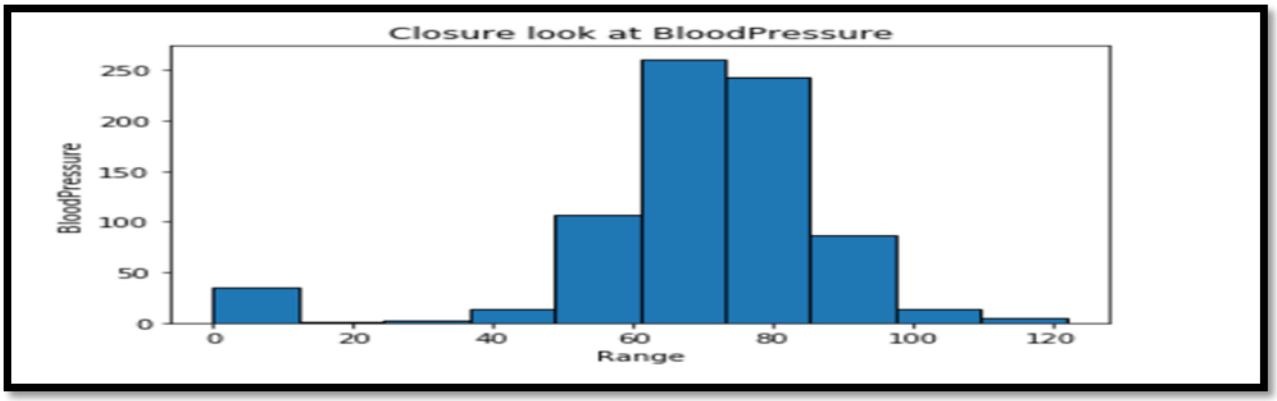
Here is the picture showing the outliers in the dataset in pregnancies.

Closure look at Glucose: - Glucose is the attribute that has a standard deviation of around 30 so below is the visualisation presenting glucose.



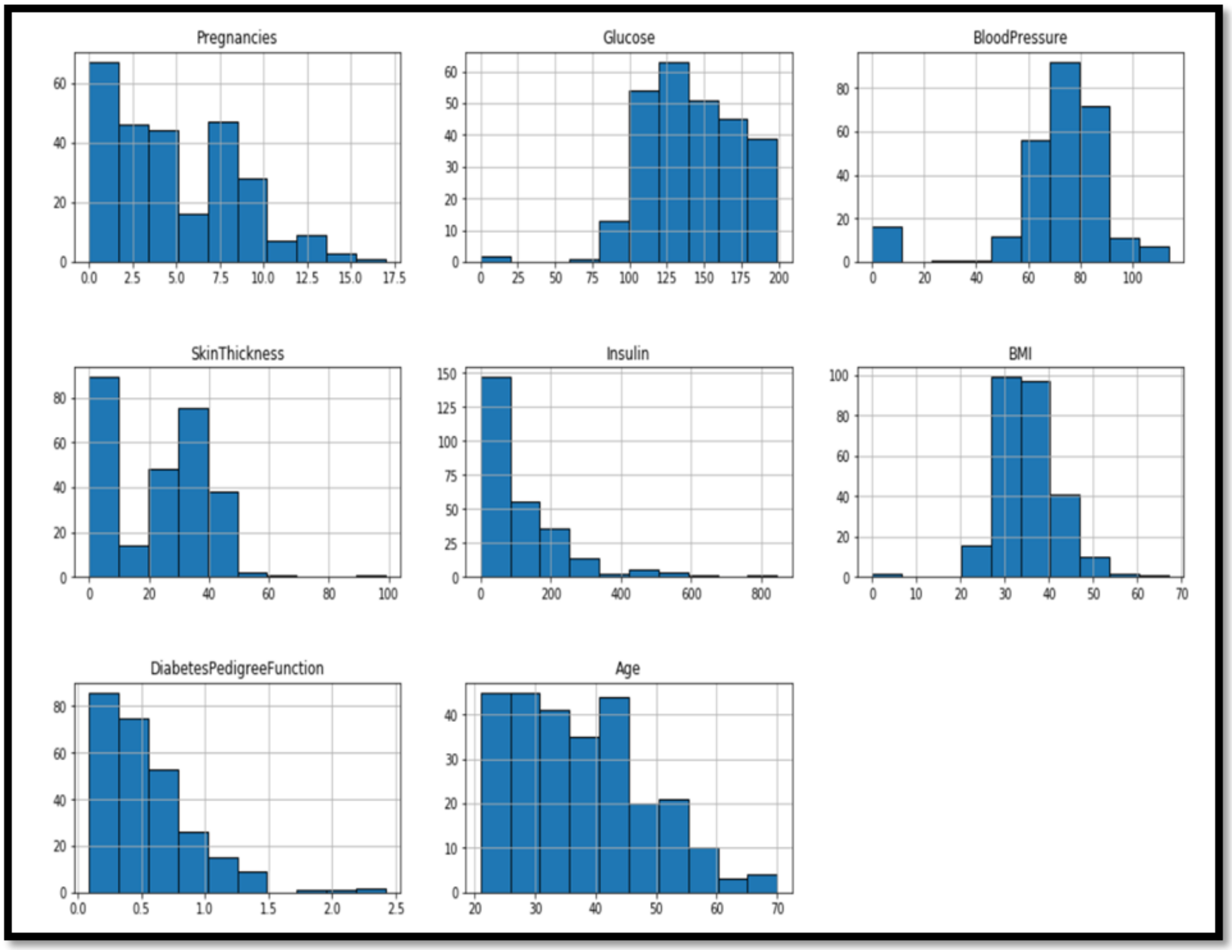
We can clearly see that there are outliers that are below 25 because such amount of glucose may result in death. Therefor Glucose below 25 is considered as outlier.

Closure Look at Blood Pressure: -Let’s take a look over the Blood Pressure level described in the dataset.



It can be clearly seen that the data is divided into two clusters therefor one cluster that is from 0-20 is an outlier so not to be considered.

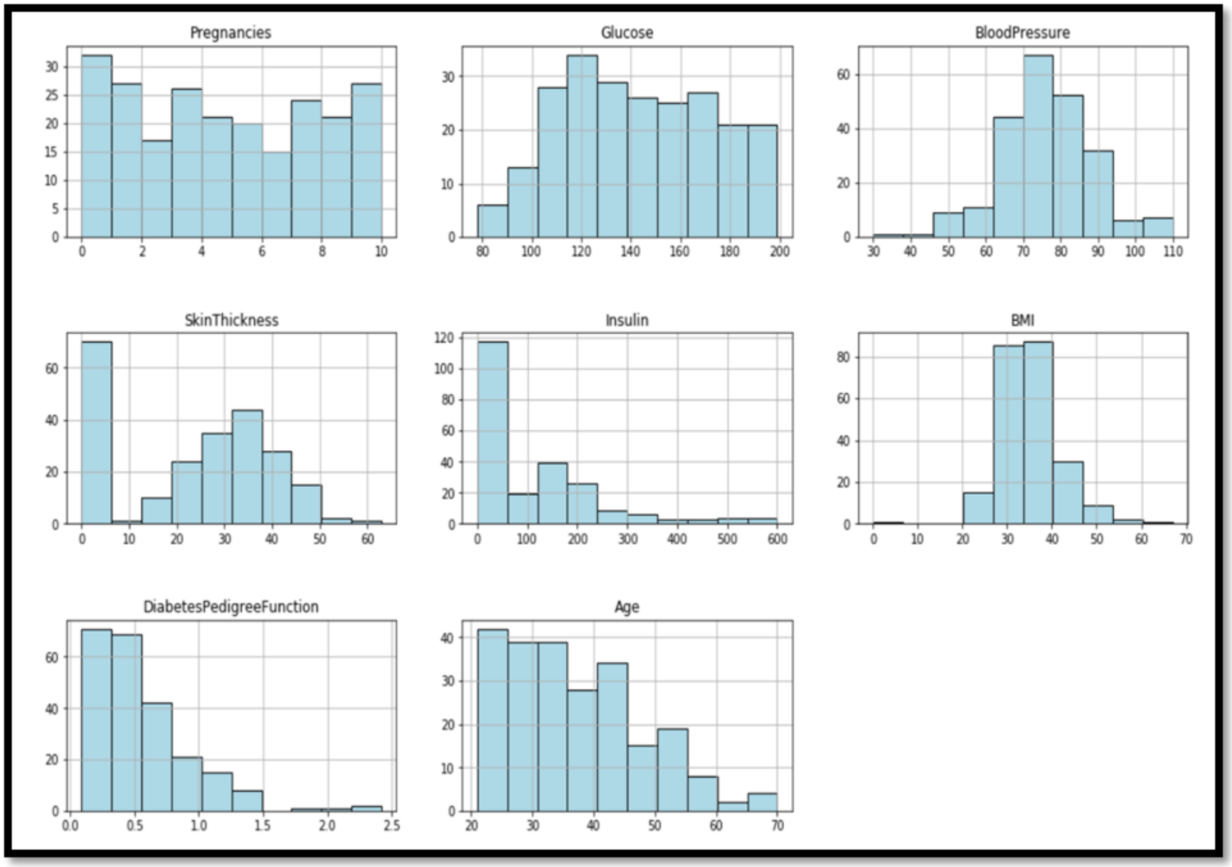
Target attributes without Pre-processing: - The data used to present the visualisation belong to positive cases of diabetes. The data is not pre-processed therefore we can differentiate between the pre-processed and non-pre-processed data.



The outcome of the visualisation is expected that data is not biased and not ready for the fitting inside a model. Therefor we conclude that this data require pre-processing over the major issues. The attributes that requires pre-processing are: -

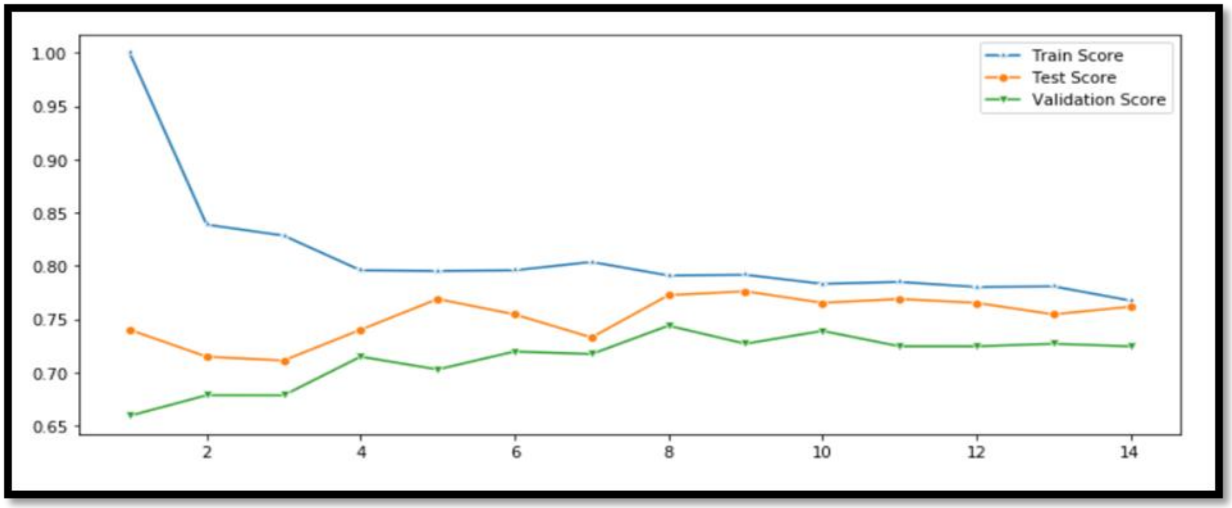
Pregnancies
Glucose
Blood Pressure
Skin Thickness
Diabetes Pedigree Function
BMI

Target with Pre-Processing: - As seen earlier we had pre-processed the data over a critical limits and the visualisation of the new dataset is shown below: -

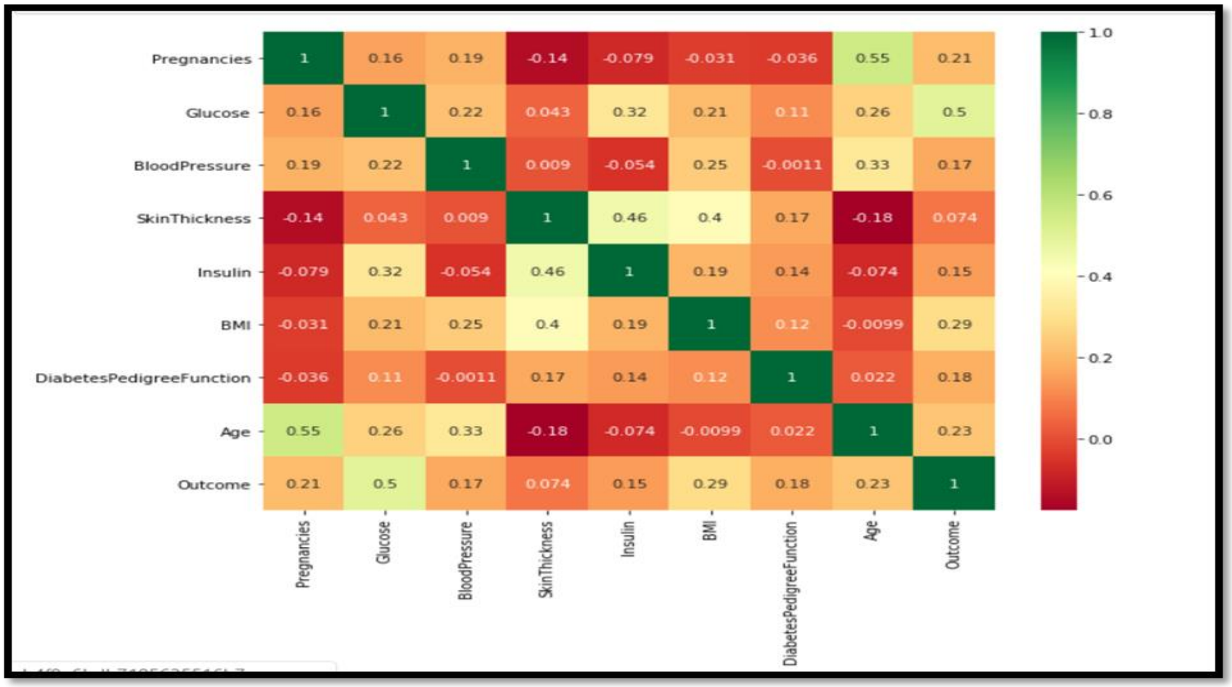


As seen above now each attribute seems to be biased properly therefore we can conclude that now the data is ready to use for modelling purpose.

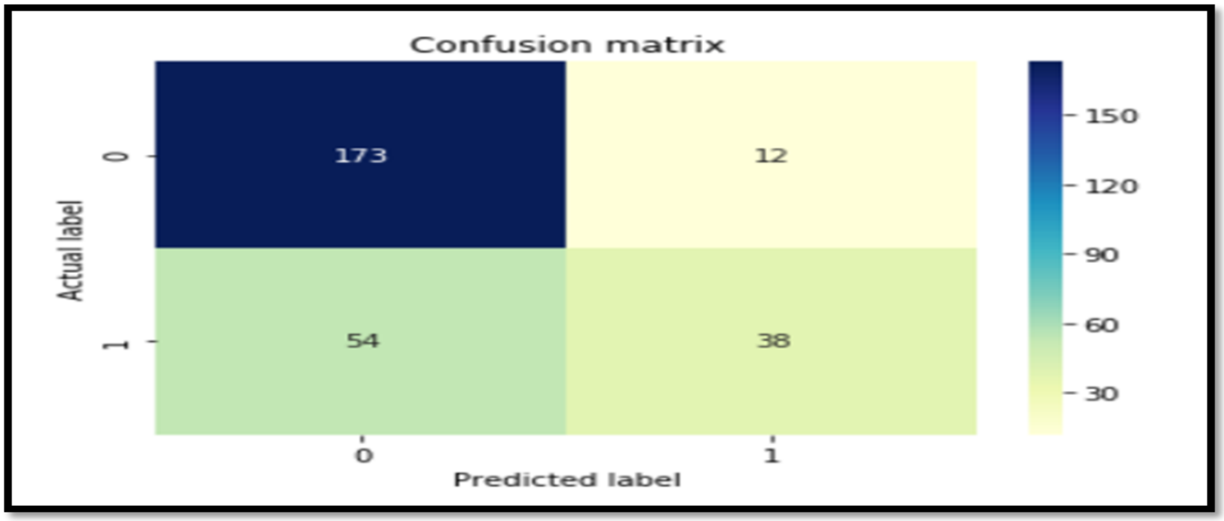
Test-Train Split: - We divided the data in the ration of 70:30. The data used for training is 70% and for testing we used 30% of the data. In addition to it we use a cross validation with fold=10 for split purpose.



Correlation of Data: -The correlation of the data defines the interdependence between the attributes. Below is the correlation of the data that describe the dependency of the data on each attribute.



Confusion Metrix: -It is the best way to validate the model using precision and recall. The model used for training is K-nearest neighbour. Here is the pictorial representation of the confusion metric describing the true positive rate, false positive rate, true negative rate and false negative rate.



4. CONCLUSION AND FUTURE SCOPE

4.1 CONCLUSION

This project results in enhancement of the DATA ANALYSIS in the can be viewed as a vast and way better application for users to identify the diabetes in their body and also help them to cure th disease before it gets too late.

On an individual basis this project helped me a lot in understanding concepts of **Python**. By this I was able to explore the use of Django, Apache Server and in enhancement the concepts of hybrid programing, JavaScript, html, css, python and enhancing the concepts of HTML and CSS and Django.

At the due of all these things, I am able to create web applications using Django and Python.

4.2 FUTURE SCOPE

Scrapping methods can be changed in future but it will be always in demand. Because

- Project is very Helpful In Medical Sciences
- There is a need to constantly monitor the health for such diseases and these diseases if caught can be cured before it gets too late.
- Provide a way to cluster the outlier and the valid data point to perform the analysis.

REFRENCES

1. UKPDS Study Group. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33).

2.Looker HC, Nyangoma SO, Cromie D, Olson JA, Leese GP, Black M, et al. Diabetic retinopathy at diagnosis of type 2 diabetes in Scotland.

3.Pierce MB, Zaninotto P, Steel N, Mindell J. Undiagnosed diabetes: data from the English longitudinal study of ageing.

4.Goyder E, Wild S, Fischbacher C, Carlisle J, Peters J. Evaluating the impact of a national pilot screening programme for type 2 diabetes in deprived areas of England.

5.Rahman M, Simmons RK, Hennings SH, Wareham NJ, Griffin SJ. Effect of screening for type 2 diabetes on population-level self-rate health outcomes and measures of cardiovascular risk: 13 year follow-up of the Ely cohort.

6.Mozaffarian D, Kamineni A, Carnethon M, Djousse L, Mukamal KJ, Siscovick D. Lifestyle risk factors and new-onset diabetes mellitus in older adults: the cardiovascular health study. Arch Intern Med 2009;169:798–807.

7.Tuomilehto J, Lindstrom J, Eriksson JG, Valle TT, Hamalainen H, Ilanne-Parikka P, et al. Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. N Engl J

8.Lindstrom J, Ilanne-Parikka P, Peltonen M, Aunola S, Eriksson JG, Hemio K, et al. Sustained reduction in the incidence of type 2 diabetes by lifestyle intervention: follow-up of the Finnish Diabetes Prevention Study.