**DIABETIC PREDICTION**

**A PROJECT REPORT**

In partial fulfilment of the requirements for the award of the degree

**MASTER OF COMPUTER APPLICATION**

**IN**

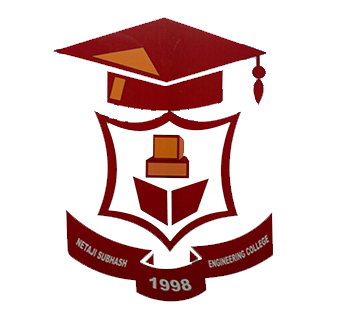
**COMPUTER SCIENCE**

Under the guidance of

**SOFIKUL MULLICK**

**BY**

**SWARNADEEP MONDAL**

****

**NETAJI SUBHASH ENGINEERING COLLEGE**

**In association with**

Engineers Study Center

**(Note: All entries of the pro forma of approval should be filled up with appropriate and complete information. Incomplete pro forma of approval in any respect will be summarily rejected.)**

|  |  |  |
| --- | --- | --- |
| 1. | Title of the Project: | **DIABETIC PREDICTION USING MACHINE LEARNING** |
| 2. | Project Members : | **SWARNADEEP MONDAL** |
| 3. | Name of the guide: | **Mr. SOFIKUL MULLICK** |
|  |  |  |

**Project Version Control History**

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Primary Author** | **Description of Version** | **Date Completed** |
| Final | Swarnadeep Mondal | Project Report | 3rd March,2021 |

Signature of Team Member Signature of Approver

Date: Date:

For Office Use Only **MR.SOFIKUL MULLICK**

|  |
| --- |
| **Approved** |

|  |
| --- |
| **Not Approved** |

Project Proposal Evaluator

**DECLARATION**

We hereby declare that the project work being presented in the project proposal entitled **“DIABETIC** **PREDICTION”** in partial fulfilment of the requirements for the award of the degree of **MASTER OF COMPUTER APPLICATION** at **ENGINEERS STUDY CENTER, SALTLAKE, KOLKATA, WEST BENGAL,** is an authentic work carried out under the guidance of **MR. SOFIKUL MULLICK**. The matter embodied in this project work has not been submitted elsewhere for the award of any degree of our knowledge and belief.

**Date:**

**Name of the Student:** Swarnadeep Mondal (ROLL:29101018018)

**Signature of the students:**

**CERTIFICATE**

This is to certify that this proposal of minor project entitled **“DIABETIC PREDICTION”** is a record of bona fide work, carried out by **Swarnadeep Mondal**  under my guidance at **Engineers Study Center**. In my opinion, the report in its present form is in partial fulfilment ofthe requirements for the award of the degree of **MASTER OF COMPUTER APPLICATION** and as per regulations of the **Engineers Study Center®.** To the best of my knowledge, the results embodied in this report, are original in nature and worthy of incorporation in the present version of the report.

**Guide / Supervisor**

------------------------------------------------

**MR. SOFIKUL MULLICK**

Project Engineer

Engineers Study Center

**ACKNOWLEDGEMENT**

Success of any project depends largely on the encouragement and guidelines of many others. I take this sincere opportunity to express my gratitude to the people who have been instrumental in the successful completion of this project work.

I would like to show our greatest appreciation to **Mr. Sofikul Mullick**, Project Engineer at Engineers Study Center, Kolkata. I always feel motivated and encouraged every time by his valuable advice and constant inspiration; without his encouragement and guidance this project would not have materialized.

Words are inadequate in offering our thanks to the other trainees, project assistants and other members at Engineers Study Center. for their encouragement and cooperation in carrying out this project work. The guidance and support received from all the members and who are contributing to this project, was vital for the success of this project.

**CONTENT**

|  |
| --- |
| * Abstract * Introduction * Problem Definition * Project Goal * Methodology * Project Objective * Project Workflow * Project Implementation * Step-by-Step Working * Project Limitations * Future Scope * Summary * Bibliography |

**ABSTRACT**

With the emerging increase of diabetes , there is a strong need for supporting the medical decision making process. In this project we aim to predict whether a person, having some certain situations, can have diabetes or not.

In this project, Logistic Regression is used to predict whether the person can be diabetic or not .

Logistic Regression is considered as one of the helpful methods for the diagnosis of many diseases.

With this, in fact , probable models are made which have been proved useful in displaying complex systems and showing the relationships between variables in a graphical way .

THIS MODEL THUS CAN PREDICT IF A PERSON CAN HAVE DIABETES OR NOT.

**INTRODUCTION**

**Python** is an interpreted, high-level and general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object oriented approach aim to help programmers write clear, logical code for small and large-scale projects

**Machine learning** (**ML**) is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence. Machine learning algorithms build a model based on sample data, known as "training data”, in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics

Machine learning involves computers discovering how they can perform tasks without being explicitly programmed to do so. It involves computers learning from data provided so that they carry out certain tasks. For simple tasks assigned to computers, it is possible to program algorithms telling the machine how to execute all steps required to solve the problem at hand; on the computer's part, no learning is needed.

**Logistic regression** is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, **logistic regression**(or **logit regression**) is estimating the parameters of a logistic model (a form of binary regression).

**PROBLEM DEFINITION**

The main objective of this project is to predict the diabetes.

The purposed work focuses on to predict diabetes using probabilistic model of Regression WITH A HANDSOME PERCENTAGE OF ACCURACY.

For detecting a disease a number of tests should be required from the patient. This tests play an important role in time and performance. Insulin is one of the most important hormones in the body. It aids the body in converting sugar ,starches and other food items into energy needed for daily life. This disease is refer to diabetes. The cause of diabetes is mystery, although obesity and lack of exercise, presence of pregnancy appear to possibly play significant roles. So, we trained our model with the blood glucose tolerance level, skin thickness, case of pregnancies, insulin levels, blood pressure ,age ,BMI and diabetes pedigree function so get more accuracy, as this is related to a serious disease prediction.

**PROJECT GOAL**

The goal of the project to build a model with can predict if a person, having some certain conditions, can have diabetes or not.

**METHODOLOGY**

* **Data Selection :**

Data is the foundation for any machine learning project. The job is to find ways and sources of collecting relevant and comprehensive data, interpreting it, and analysing results with the help of statistical techniques.

* **Data Visualisation:**

A large amount of information represented in graphic form is easier to understand and analyse. Some companies specify that a data analyst must know how to create slides, diagrams, charts, and templates.

* **Data cleaning:**

This set of procedures allows for removing noise and fixing inconsistencies in data. A data scientist can fill in missing data using imputation techniques. A specialist also detects outliers observations that deviate significantly from the rest of distribution.

* **Data Splitting:**

A dataset used for machine learning should be partitioned into two subsets training and test sets.

* **Model Selection:**

After a data scientist has pre-processed the collected data and split it into two subsets, he or she can proceed with a model training. This process entails “feeding” the algorithm with training data. An algorithm will process data and output a model that is able to find a target value in new data .The purpose of model training is to develop a model.

* **Model Evaluation:**

The goal of this step is to develop the simplest model able to formulate a target value fast and well enough and check the accuracy.

* **Model Testing:**

Final step is testing the model, that what will it actually show when the user gives the certain set of conditions to know if he can have diabetes or not.

**PROJECT OBJECTIVE**

Our objective of the project is to predict if a person can have diabetes or not using Machine learning with the implementation of logistic regression.

**PROJECT WORKFLOW**

This is the detailed work architecture where we are showing the process of Diabetes Prediction using machine learning.

Data Visualization

Data Selection

Data Cleaning

Data Splitting

Model Training

Model Evaluation

Model Testing

**PROJECT IMPLEMENTATION**

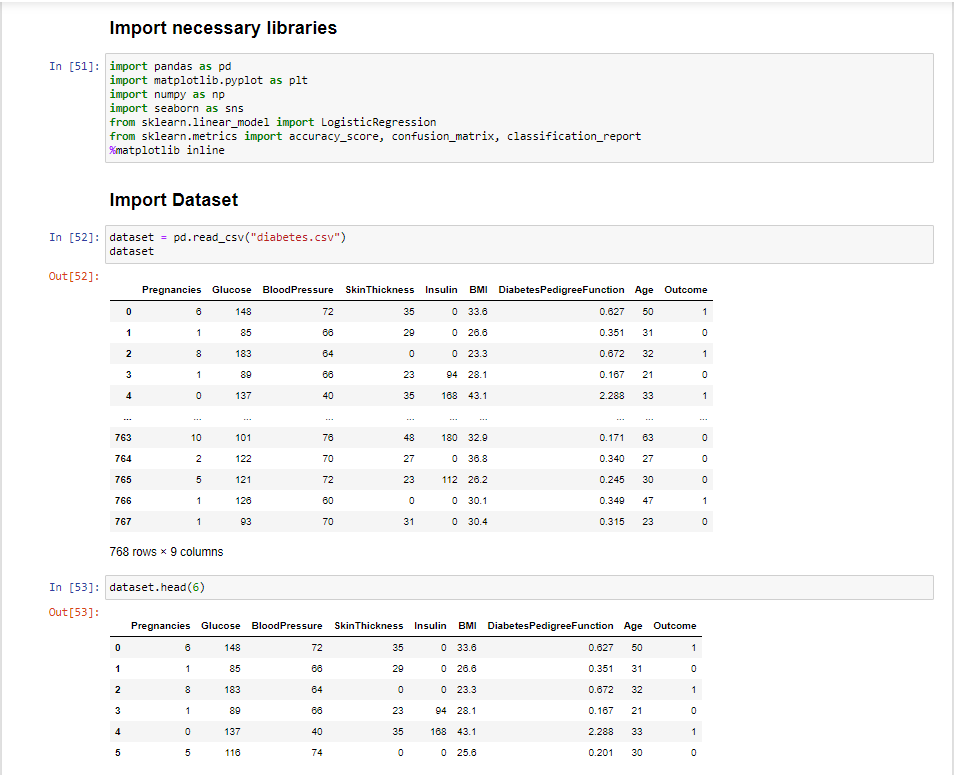
* **SELECTION OF DATA:** The process of selecting data depends on the type of project we desire to do. The data set can be collected from various sources such as a file, database, sensor and many other such sources.
* **VISUALIZATION OF DATA:** Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in a visual manner.
* **DATA PRE-PROCESSING :** As we know that data pre-processing is a process of cleaning the raw data into clean data, so that can be used to train the model. So, we definitely need data pre-processing to achieve good results from the applied model in machine learning and deep learning projects.

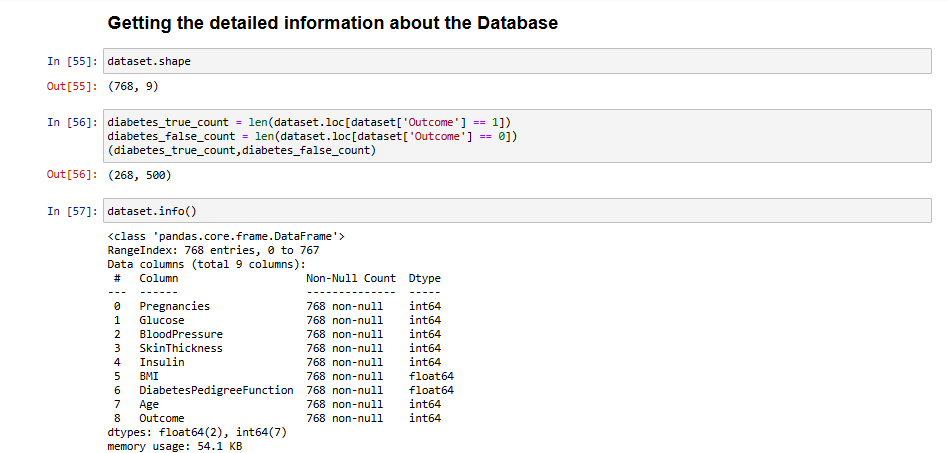
Here we checked that the data we selected was completely clean, so we did not have to clean it any further.

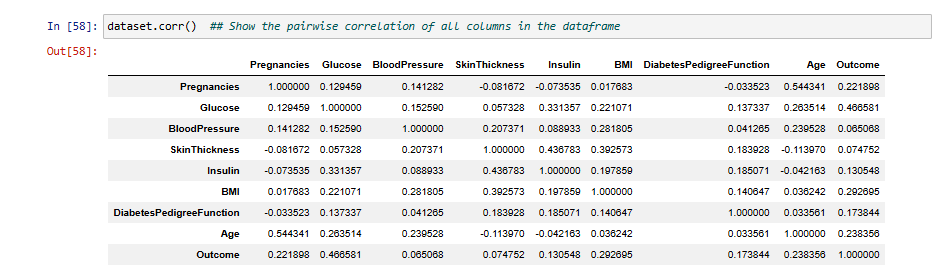
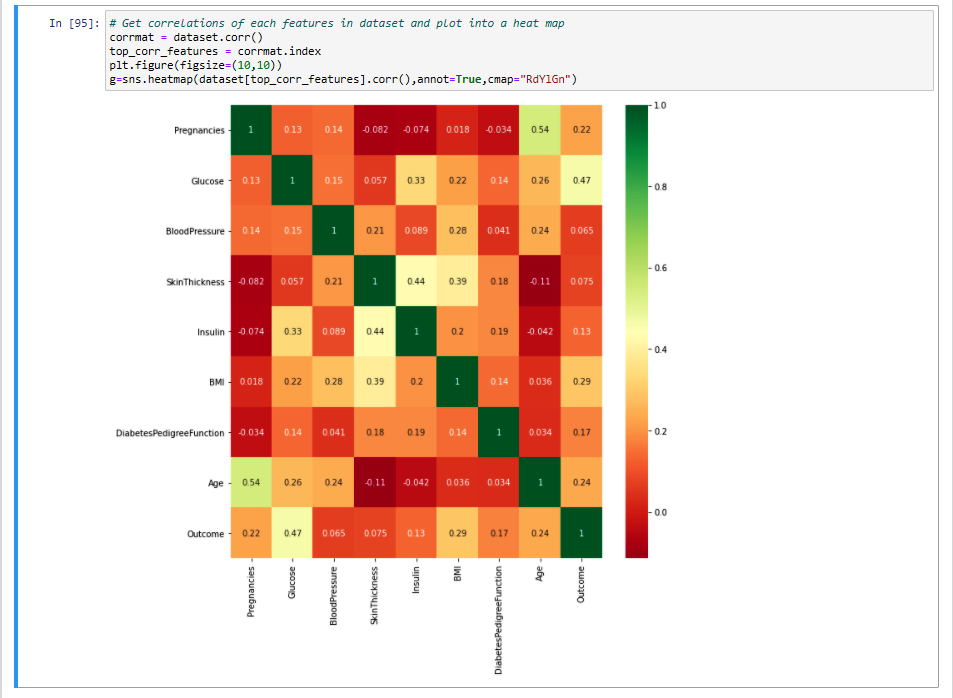
* **SELECTION OF DEPENDENT AND INDEPENDENT DATA:** We need to select the dependent and independent data and store them in y and x.
* **SPLITTING OF THE DATA:** We train the classifier using ‘**training data set**’, then test the performance of your classifier on unseen ‘**test data set**’.We split the data for training and testing by using the ‘train\_test\_split’ .
* **FITTING THE MODEL:** In a data set, a training set is implemented to build up a model. Once the model is trained we can use the same trained model to predict using the testing data i.e. the unseen data. Once this is done we can develop a confusion matrix, this tells us how well our model is trained.
* **MODEL EVALUATION :** It is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future.
* **MODEL TESTING:** Final step is testing the model, that what will it actually show when the user gives the certain set of conditions to know if he can have diabetes or not.

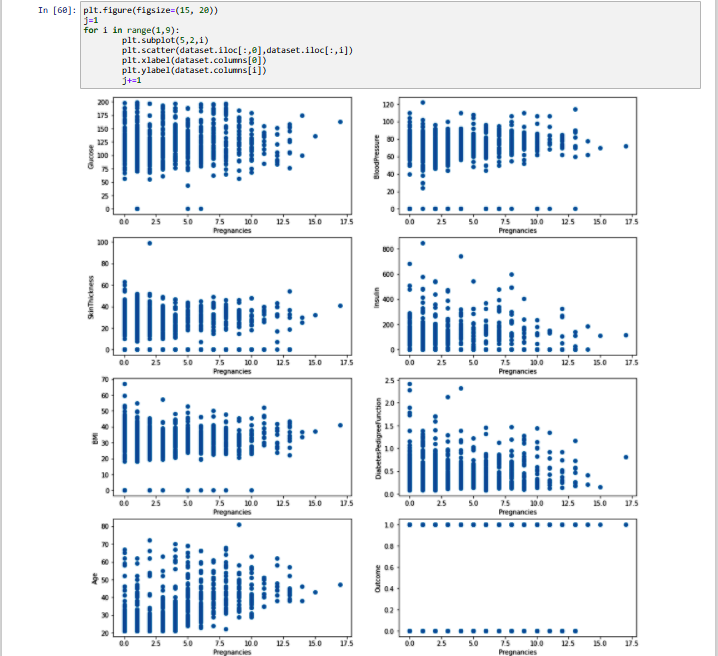
**STEP-BY-STEP WORKING**

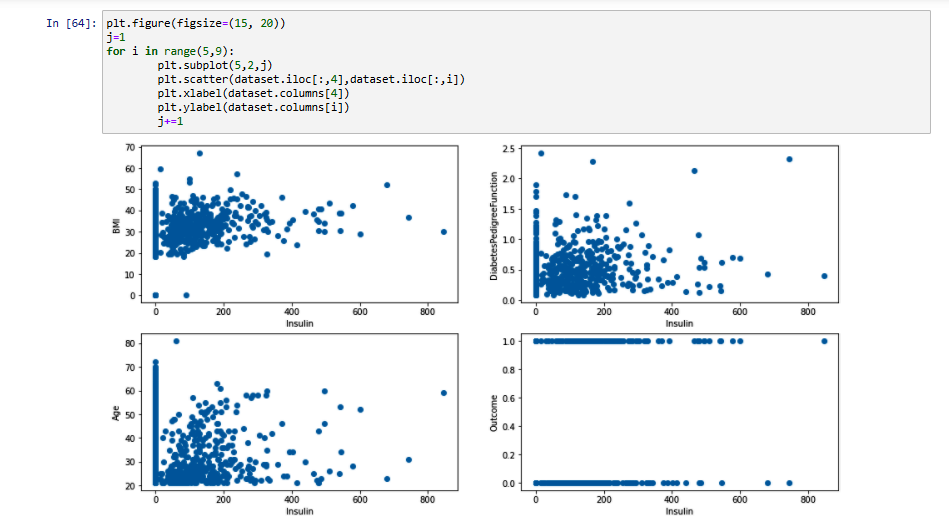
* **DATA SELECTION**

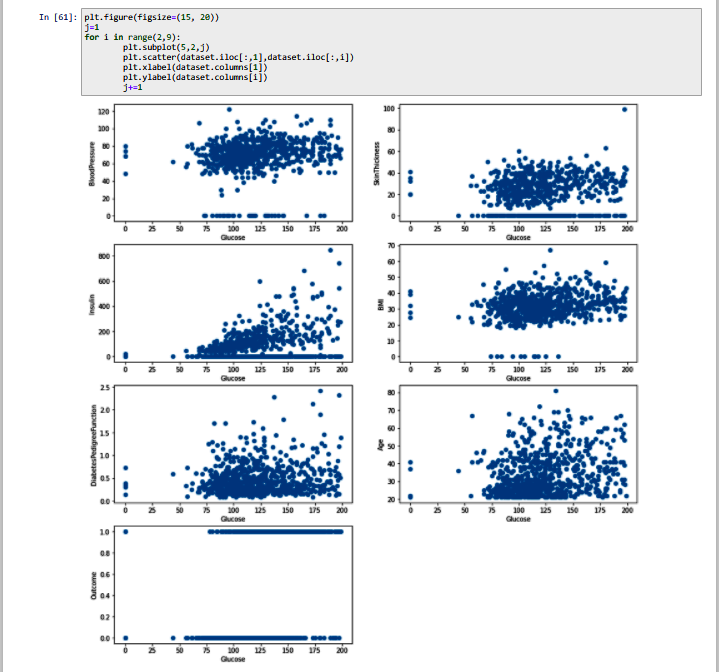


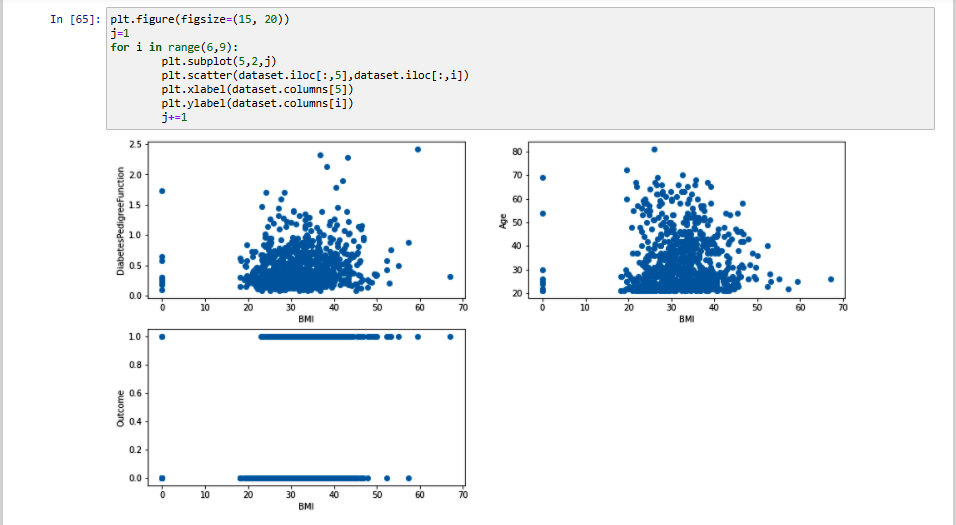
****

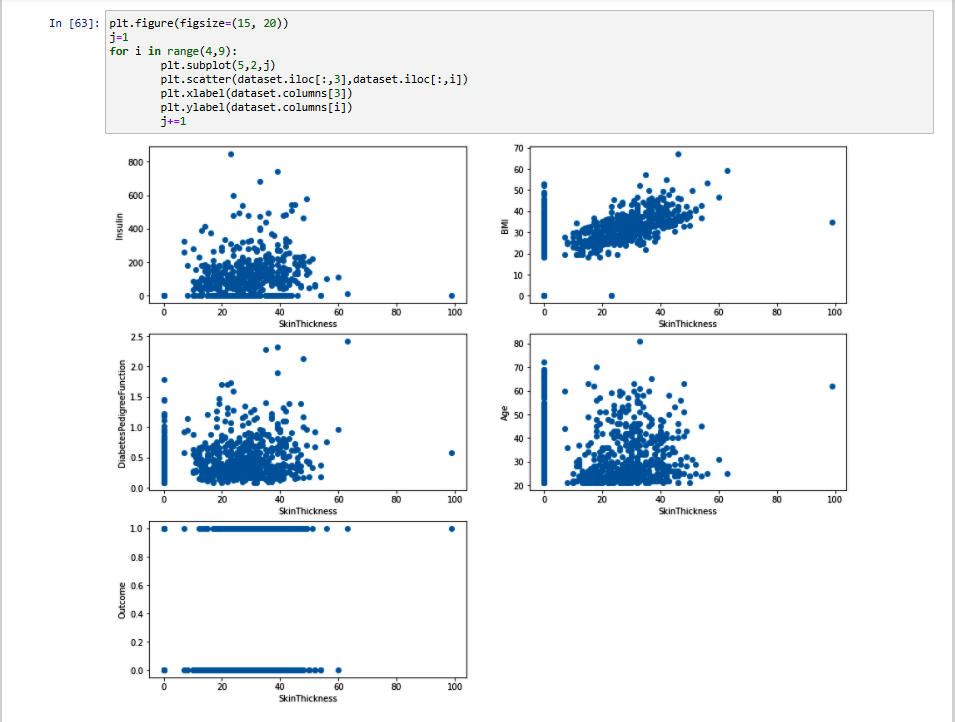
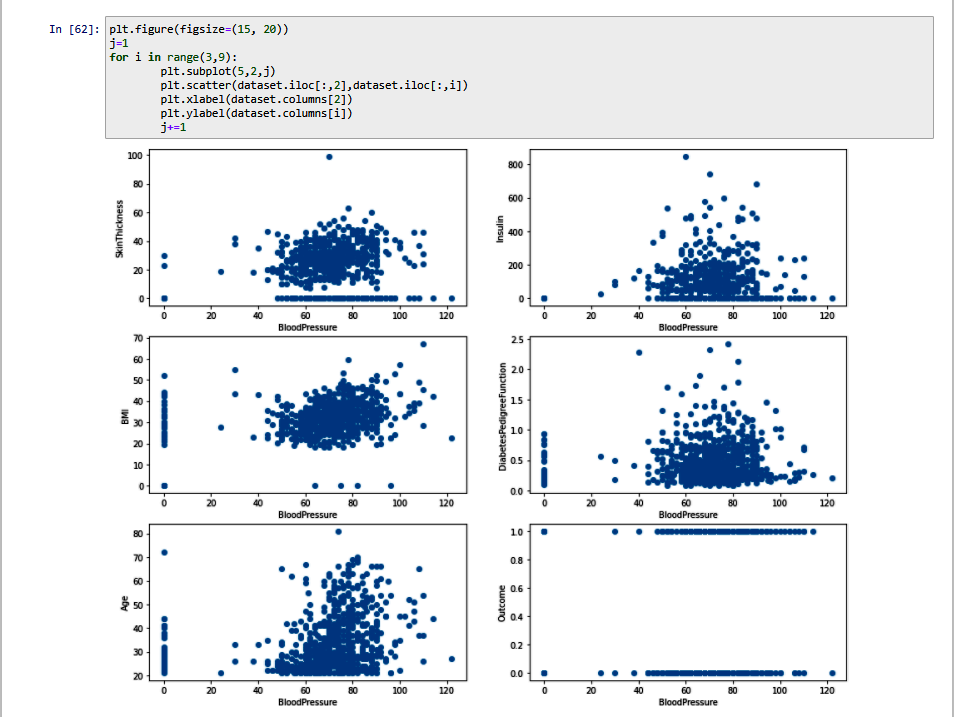
****

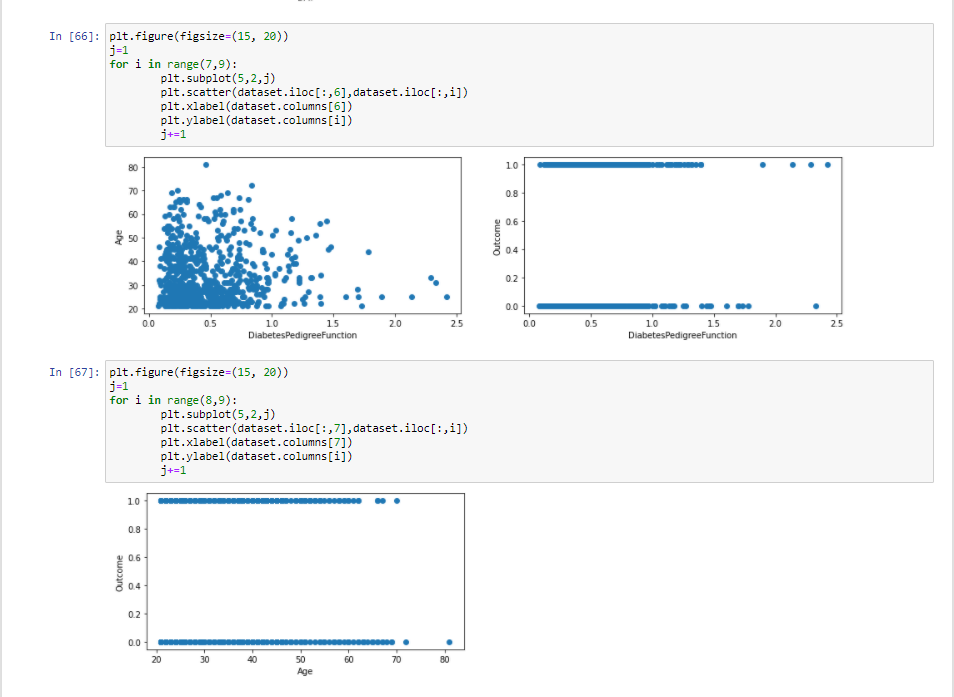
* **DATA VISUALIZATION**

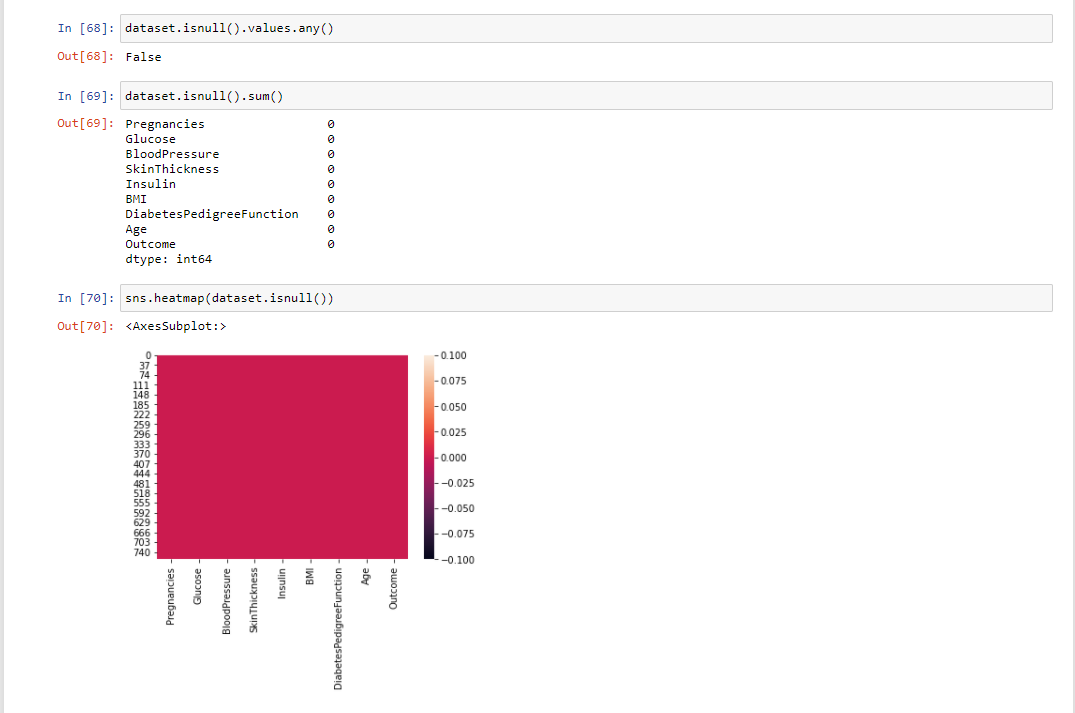
****

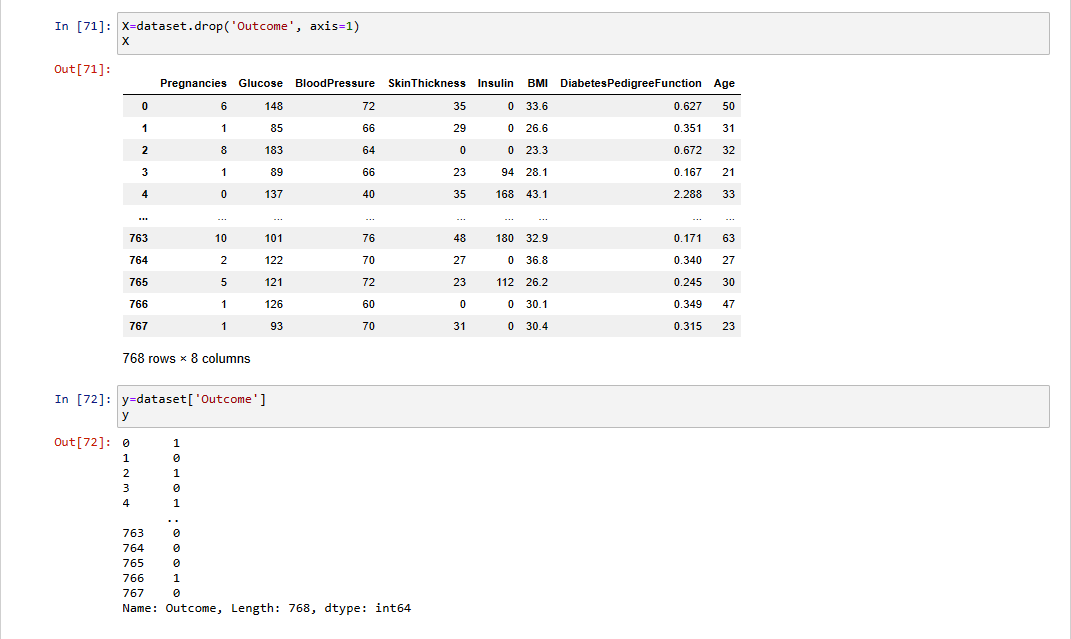
****

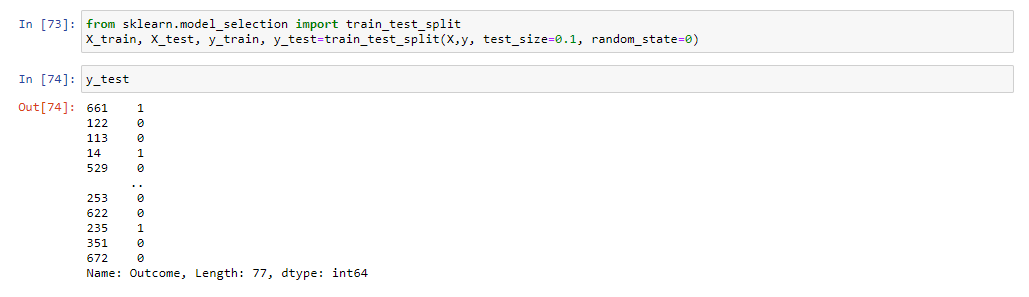
****

****

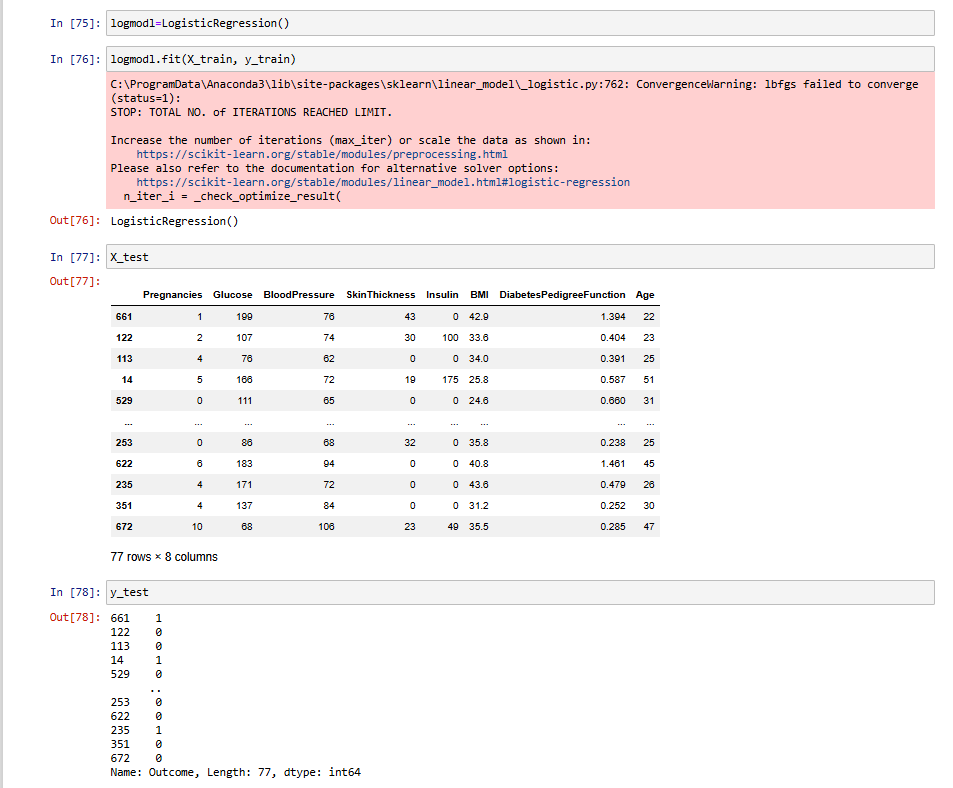
****

* **DATA CLEANING**
* **DEPENDENT AND INDEPENDENT DATA**

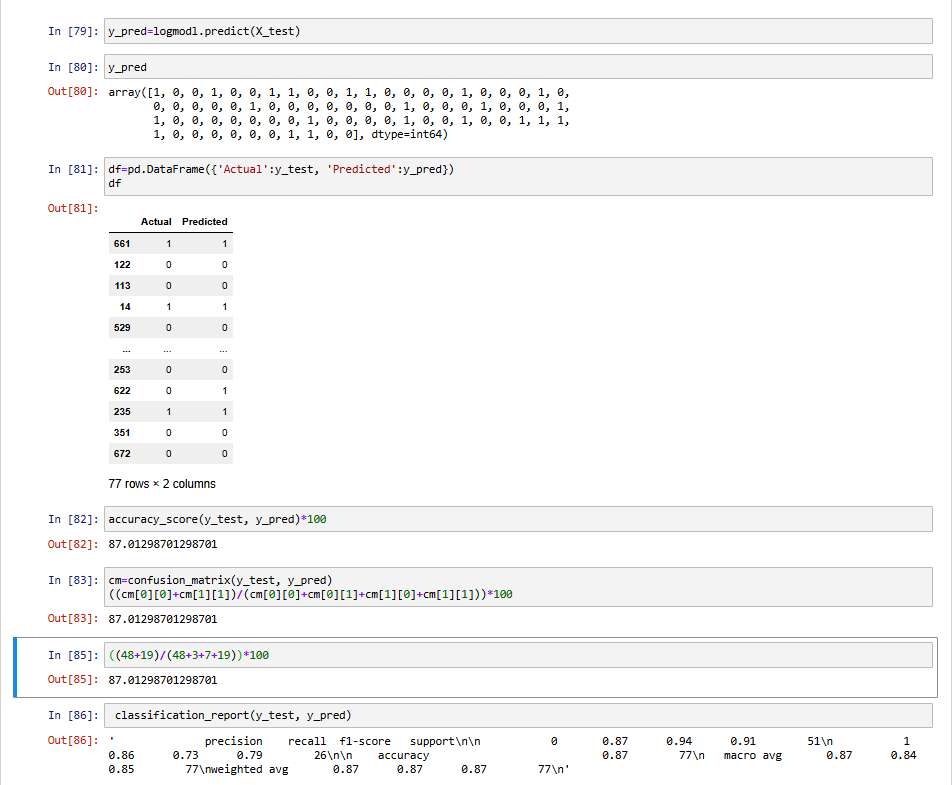
****

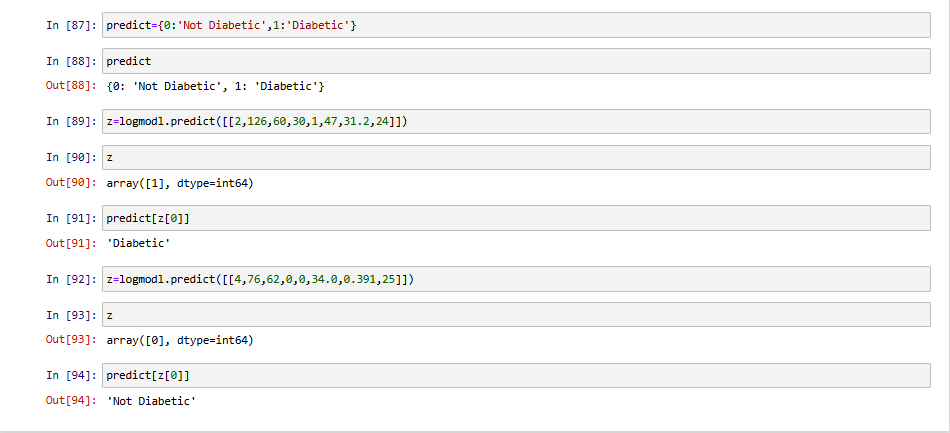
****

* **DATA SPLITTING**
* **MODEL TRAINING**

****

* **MODEL EVALUATION**

****

* **MODEL TESTING**

**PROJECT LIMITATIONS**

We worked on the backend part of the system thus there is no frontend work associated which can result in a more realistic look and focus on user experience.

**FUTURE SCOPE**

1. Since here only the backend part of the system is built, we can create a custom frontend which can result in a more realistic look and focus on user experience.
2. We can work on the project and use it for greater size dataset.

**SUMMARY**

We load previous datasets to the system. Visualization of data is done to properly know about the data . Data pre-processing is not required as there was no Nan. Following operations are performed on the dataset after that. User input data to the system in order to diagnose whether he has the disease or not. Building model using Logistic Regression Algorithm and train the data set. Test the dataset using model. Get the evaluation result. Get the predicted voting from all classifiers and gives the diagnostic result.

**BIBLIOGRAPHY**

[**https://www.wikipedia.org/**](https://www.wikipedia.org/)

[**https://www.altexsoft.com/blog/datascience/machine-learning-project-structure-stages-roles-and-tools/**](https://www.altexsoft.com/blog/datascience/machine-learning-project-structure-stages-roles-and-tools/)

[**https://stackify.com/aws-lambda-with-python-a-complete-getting-started-guide/**](https://stackify.com/aws-lambda-with-python-a-complete-getting-started-guide/)

[**https://towardsdatascience.com/workflow-of-a-machine-learning-project-ec1dba419b9**](https://towardsdatascience.com/workflow-of-a-machine-learning-project-ec1dba419b9)

**SOURCE CODE**

import pandas as pd

import matplotlib.pyplot as plt

import numpy as np

import seaborn as sns

from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import accuracy\_score, confusion\_matrix, classification\_report

%matplotlib inline

dataset = pd.read\_csv("diabetes.csv")

dataset

dataset.head(6)

dataset.tail(6)

dataset.shape

diabetes\_true\_count = len(dataset.loc[dataset['Outcome'] == 1])

diabetes\_false\_count = len(dataset.loc[dataset['Outcome'] == 0])

(diabetes\_true\_count,diabetes\_false\_count)

dataset.info()

dataset.corr()

corrmat = dataset.corr()

top\_corr\_features = corrmat.index

plt.figure(figsize=(10,10))

g=sns.heatmap(dataset[top\_corr\_features].corr(),annot=True,cmap="RdYlGn")

plt.figure(figsize=(15, 20))

j=1

for i in range(1,9):

plt.subplot(5,2,i)

plt.scatter(dataset.iloc[:,0],dataset.iloc[:,i])

plt.xlabel(dataset.columns[0])

plt.ylabel(dataset.columns[i])

j+=1

plt.figure(figsize=(15, 20))

j=1

for i in range(2,9):

plt.subplot(5,2,j)

plt.scatter(dataset.iloc[:,1],dataset.iloc[:,i])

plt.xlabel(dataset.columns[1])

plt.ylabel(dataset.columns[i])

j+=1

plt.figure(figsize=(15, 20))

j=1

for i in range(3,9):

plt.subplot(5,2,j)

plt.scatter(dataset.iloc[:,2],dataset.iloc[:,i])

plt.xlabel(dataset.columns[2])

plt.ylabel(dataset.columns[i])

j+=1

plt.figure(figsize=(15, 20))

j=1

for i in range(4,9):

plt.subplot(5,2,j)

plt.scatter(dataset.iloc[:,3],dataset.iloc[:,i])

plt.xlabel(dataset.columns[3])

plt.ylabel(dataset.columns[i])

j+=1

plt.figure(figsize=(15, 20))

j=1

for i in range(5,9):

plt.subplot(5,2,j)

plt.scatter(dataset.iloc[:,4],dataset.iloc[:,i])

plt.xlabel(dataset.columns[4])

plt.ylabel(dataset.columns[i])

j+=1

plt.figure(figsize=(15, 20))

j=1

for i in range(6,9):

plt.subplot(5,2,j)

plt.scatter(dataset.iloc[:,5],dataset.iloc[:,i])

plt.xlabel(dataset.columns[5])

plt.ylabel(dataset.columns[i])

j+=1

plt.figure(figsize=(15, 20))

j=1

for i in range(7,9):

plt.subplot(5,2,j)

plt.scatter(dataset.iloc[:,6],dataset.iloc[:,i])

plt.xlabel(dataset.columns[6])

plt.ylabel(dataset.columns[i])

j+=1

plt.figure(figsize=(15, 20))

j=1

for i in range(8,9):

plt.subplot(5,2,j)

plt.scatter(dataset.iloc[:,7],dataset.iloc[:,i])

plt.xlabel(dataset.columns[7])

plt.ylabel(dataset.columns[i])

j+=1

dataset.isnull().values.any()

dataset.isnull().sum()

sns.heatmap(dataset.isnull())

X=dataset.drop('Outcome', axis=1)

X

y=dataset['Outcome']

y

from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test=train\_test\_split(X,y, test\_size=0.1, random\_state=0)

y\_test

logmodl=LogisticRegression()

logmodl.fit(X\_train, y\_train)

X\_test

y\_test

y\_pred=logmodl.predict(X\_test)

y\_pred

df=pd.DataFrame({'Actual':y\_test, 'Predicted':y\_pred})

df

accuracy\_score(y\_test, y\_pred)\*100

confusion\_matrix(y\_test, y\_pred)

((48+19)/(48+3+7+19))\*100

classification\_report(y\_test, y\_pred)

predict={0:'Not Diabetic',1:'Diabetic'}

predict

z=logmodl.predict([[2,126,60,30,1,47,31.2,24]])

z

predict[z[0]]

z=logmodl.predict([[4,76,62,0,0,34.0,0.391,25]])

z

predict[z[0]]