SMARTBRIDGE
Let's Bridge the Gap

INTERNSHIP REPORT APPROVAL FORM

July 1, 2019

With immense pleasure, this is to approve that the students of Sona College of Technology i.e.,

Harikarasuriya Devarajan(1516106022),

Dhivya Murali(1516106015),

Madhumitha M(1516106044) and

Kiruthika R(1516106036)

successfully completed their Project and Project Report on **"Diabetes Mellitus Prediction"** under our guidance.

We are highly impressed with the work that they have done and commend them on their quick grasping skills. They have shown good intent to learn and have put the knowledge gained into application in the from of this project. We appreciate the hard work and commitment shown by them.

We, hereby approve that this document is completely checked and accepted by SmartBridge Technical Team. Its been an absolute pleasure to educate and mentor these students. We hope that this document will also serve as a Letter of Recommendation, to whomsover applied.

We wish them success in all future endeavors and a great career ahead.

GD Abhishek

AI Developer

1. DIABETES MELLITUS PREDICTION

1.1 INTRODUCTION:

AI:

Artificial Intelligence is a simulation of human intelligence processed by machines. It is an area of computer science that emphasizes the creation of intelligent machines that work and react like humans. Research associated with artificial intelligence is highly technical and specialized. Modern machine capabilities generally classified as AI include successfully understanding human speech, competing at the highest level in strategic game systems, in content delivery networks and military simulations.

Artificial intelligence is a science and technology based on disciplines such as Computer Science, Biology, Psychology, Linguistics, Mathematics, and Engineering. The domain of artificial intelligence is huge in breadth and width. Machine learning is a part of Artificial Intelligence. Machine Learning is a system that can learn from example through self-improvement and without being explicitly coded by programmer. Machine learning combines data with statistical tools to predict an output.

Python:

Machine learning involves computer to get trained using a given dataset and use this training to predict the properties of a given new data. Process of training and prediction involves use of specialized algorithms. Python community has developed many modules to help programmers implement machine learning. Numpy is a very popular python library for large multi-dimensional array and matrix processing, with the help of a large collection of high-level mathematical functions.

Scikit-learn another popular library of python supports most of the supervised and unsupervised learning algorithms. Python also has libraries for data visualizations such as matplotlib and seaborn. Python makes it easier to learn and implement machine learning.

1.2 OBJECTIVES OF RESEARCH:

Diabetes has reached an epidemic magnitude in many countries and more so in developing countries, causing a great burden from life threatening complications of varied nature. The rising tide of diabetes and its complications will place an increasingly heavy burden of morbidity and mortality on patients and their families for decades to come.

In developing countries more than 50% of diagnosis remain undiagnosed. Many do not seek medical help until debilitating complications force them to do so. The pandemic of diabetes, along with its high human and economic costs, is showing no signs of reduction and therefore, new approaches are urgently needed to prevent or slow the progression and limit the consequence of the disease.

Evidences suggest that early detection of diabetes by appropriate methods, especially in subjects with high risk of diabetes will help to prevent or delay complications and thus reduce the clinical, social and economic burden of the disease.

The main objective is to identify individuals who are at increased risk of diabetes and delay or prevent the progression. Therefore, the present study was aimed to predict whether a person has diabetes or not by using Machine Learning Model.

Machine Learning develops a programmed model using data, algorithms and computing power. This process requires more computing power as the number of data variables increases.

In statistics, logistic regression is a regression model where the dependent variable is categorical namely binary dependent variable where it can take only two values 0 and 1 which represent outcomes such as diabetes and no diabetes.

1.3 PROBLEM STATEMENT:

Diabetes mellitus is a creator of different kinds of diseases like heart attack, blindness, kidney diseases, etc. Insulin deficiency results in elevated blood glucose levels and impaired metabolism of carbohydrates, fat and proteins. Various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes.

In initial days patient need to visit a diagnostic center, consult their doctor, and sit tight for a day or more to get their reports. Moreover, every time they want to get their diagnosis report, they must waste their money in vain. By using machine learning concepts, we can predict accurately whether the patient is a sufferer of diabetic or not.

The aim of this project is to develop a model which can perform prediction of diabetes for a patient with a higher accuracy by using the results of different machine learning techniques. This project has focused on developing a model based on Logistic regression. Classification is one of the most important decision-making techniques in many real-world problems.

In this work, the main objective is to classify the data as diabetic or non-diabetic and improve the classification accuracy. In many cases, the performance of algorithm is high in the context of speed, but the accuracy of data classification is low. The main objective of our model is to achieve high accuracy.

Classification accuracy can be increase if we use much of the data set for training and few data sets for testing. This survey has analyzed various classification techniques for classification of diabetic and non-diabetic data. Thus, it is observed that techniques like Logistic Regression, is most suitable for implementing the Diabetes prediction model.

2. REVIEW OF LITERATURE

The history of diabetes mellitus begins with the mention of polyuria in Ebers papyrus in 1550 BC. The earliest mention of honey urine (Madhumeha) was made by Sushrutha in 400 BC. Though Celsus (30 BC-50 BC) recognized the disease, Aretaeus or Cappadocia gave the name “Diabetes” (a siphon). He made a complete description of diabetes mellitus describing it as “melting down of the flesh and limbs into urine” (George F and Cahill JR, 2005). In the 3rd to 5th centuries AD scholars in China, Japan and India wrote of a condition with polyuria in which the urine was sweet and sticky.

In 2008, Abbas Ali Mansour et al., conducted a cross-sectional population-based study to screen for diabetes in al-Madina a rural area located in the north of Basrah, Iraq. A total of 3176 subjects were screened and overall prevalence of undiagnosed diabetes was 2.14%, known diabetics constituted 5.29%, IFG was seen in 2.02%, subjects with abnormal glycemia (diabetes and IFG) constitute 9.45%. Previously undiagnosed diabetics constitute of 28.81% of all diabetics in this study.

In 2009, Shaoyan Zhang et al, have done a study to assess the prevalence of diabetes and IFG and to compare the risk factors between diabetes and IFG in the Mongolian population, China. A total of 2589 Mongolians aged 20 years or more were screened and the overall prevalence of diabetes and IFG was 3.7% (males 3.9%; females 3.5%) and 18.5% (males 17.7%; females 19.0%) respectively.

In 2010, Qiang Lu et al., conducted a cross-sectional study in 3937 Han adolescents aged 13-18 years, to evaluate the prevalence of impaired fasting glucose (IFG) and its relationship with cardiovascular risk factors. The prevalence of IFG was found to be 3.5% similar in both genders.

In 2011, Amina Khambalia et al., conducted a nationwide survey of people aged 15-64 years (n=1592) for the prevalence and risk factors of diabetes and impaired fasting glucose in Nauru. The sex standardized prevalence of diabetes and prediabetes was found to be 13.7% and 6.0%.

3.DATA COLLECTION

Data Collection is the process of gathering and measuring information from countless different sources. Collecting data allows you to capture a record of past events so that we can use data analysis to find recurring patterns. The data can be numeric, categorical and free text. The dataset we used for prediction is PIMA Indian Diabetics Dataset from UCI Machine Learning Repository. The dataset was originally from the national institute of diabetes and digestive and kidney disease.

The dataset includes data from 768 women with 8 characteristics. The dataset consists of several medical predictor (independent) variables and one target (dependent) variable. The dataset includes attributes such as pregnancies, glucose level, age, BMI, insulin, diabetes pedigree function, skin thickness and sex.

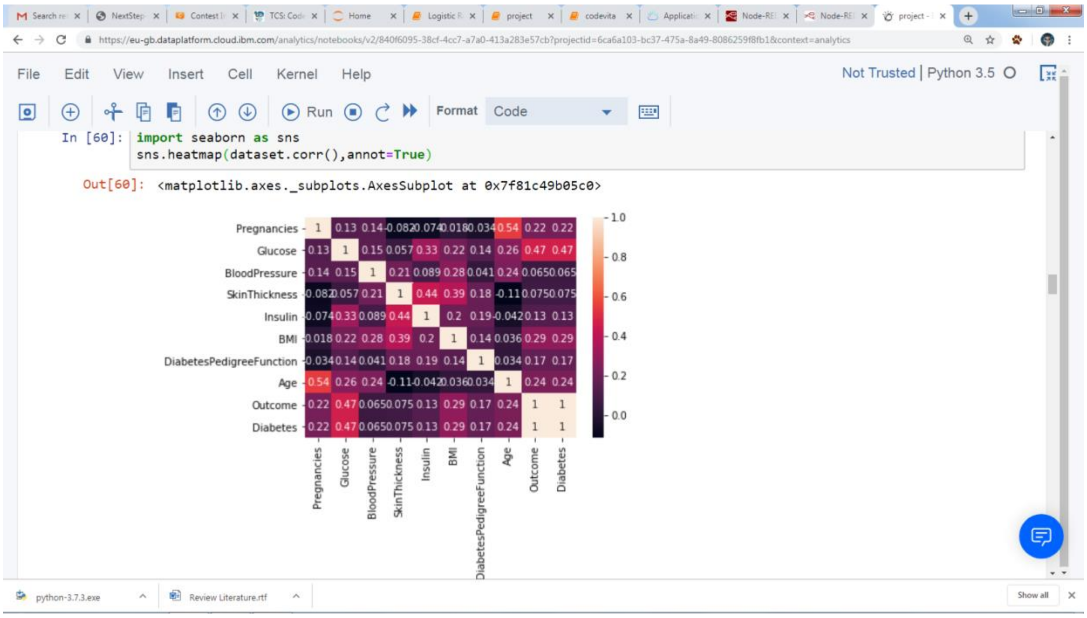
The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community. UCI Machine Learning repository currently maintains 378 datasets as a service to the machine learning community. Each dataset gets its own webpage that lists all the details known about it including any relevant publications that investigate it. The datasets themselves can be downloaded as ASCII files, often useful CSV format.

The advantage of using UCI Machine Learning Repository is that most of the datasets are already preprocessed and cleaned. Most of the datasets are small meaning you can easily load the dataset into the MS excel file and review them. Datasets are limited to tabular column mainly classification datasets, this limiting for working on natural language processing, computer vision and recommender systems.

4.METHODOLOGY

4.1 EXPLORATORY DATA ANALYSIS:

4.1.1 FIGURES AND TABLES:

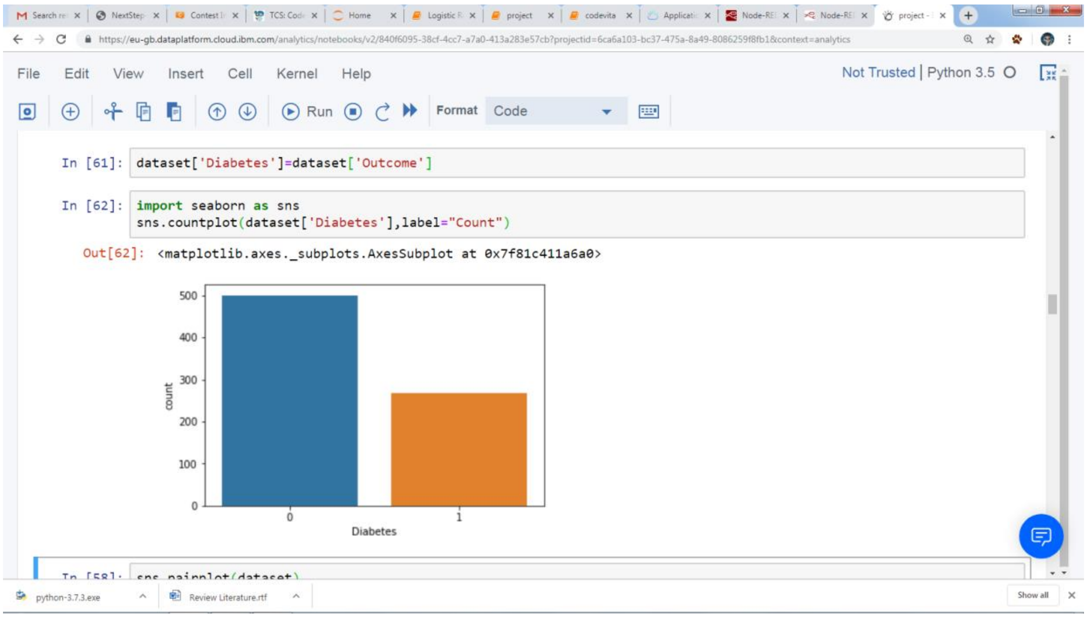


https://eu-gb.dataplatform.cloud.ibm.com/analytics/notebooks/v2/840f6095-38cf-4cc7-a7a0-413a283e57cb?projectid=6ca6a103-bc37-475a-8a49-8086259f8fb1&context=analytics
File Edit View Insert Cell Kernel Help
Not Trusted | Python 3.5
Run
Format Code
Pregnancies - 1 0.13 0.14 -0.082 0.074 0.018 0.034 0.54 0.22 0.22
Glucose - 0.13 1 0.15 0.057 0.33 0.22 0.14 0.26 0.47 0.47
BloodPressure - 0.14 0.15 1 0.21 0.089 0.28 0.041 0.24 0.065 0.065
SkinThickness - 0.082 0.057 0.21 1 0.44 0.39 0.18 -0.11 0.075 0.075
Insulin - 0.074 0.33 0.089 0.44 1 0.2 0.19 -0.042 0.13 0.13
BMI - 0.018 0.22 0.28 0.39 0.2 1 0.14 0.036 0.29 0.29
DiabetesPedigreeFunction - 0.034 0.14 0.041 0.18 0.19 0.14 1 0.034 0.17 0.17
Age - 0.54 0.26 0.24 -0.11 -0.042 0.036 0.034 1 0.24 0.24
Outcome - 0.22 0.47 0.065 0.075 0.13 0.29 0.17 0.24 1 1
Diabetes - 0.22 0.47 0.065 0.075 0.13 0.29 0.17 0.24 1 1
Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age
Outcome
Diabetes
1.0
0.8
0.6
0.4
0.2
0.0
python-3.7.3.exe
Review Literature.rtf
Show all

```
In [60]: import seaborn as sns
         sns.heatmap(dataset.corr(),annot=True)

Out[60]: <matplotlib.axes._subplots.AxesSubplot at 0x7f81c49b05c0>
```

FIGURE 4.1.1 HEAT MAP



https://eu-gb.dataplatform.cloud.ibm.com/analytics/notebooks/v2/840f6095-38cf-4cc7-a7a0-413a283e57cb?projectid=6ca6a103-bc37-475a-8a49-8086259f8fb1&context=analytics
File Edit View Insert Cell Kernel Help
Not Trusted | Python 3.5
Run
Format Code
500
400
300
200
100
0
count
0
1
Diabetes
python-3.7.3.exe
Review Literature.rtf
Show all

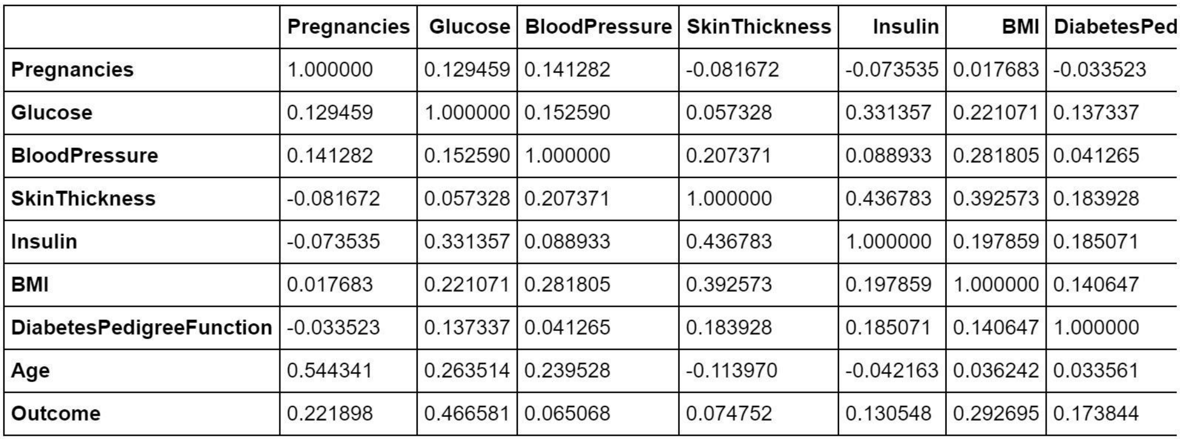
```
In [61]: dataset['Diabetes']=dataset['Outcome']

In [62]: import seaborn as sns
         sns.countplot(dataset['Diabetes'],label="Count")

Out[62]: <matplotlib.axes._subplots.AxesSubplot at 0x7f81c411a6a0>

In [58]: sns.pairplot(dataset)
```

FIGURE 4.1.2 COUNT PLOT



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPed
Pregnancies	1.000000	0.129459	0.141282	-0.081672	-0.073535	0.017683	-0.033523
Glucose	0.129459	1.000000	0.152590	0.057328	0.331357	0.221071	0.137337
BloodPressure	0.141282	0.152590	1.000000	0.207371	0.088933	0.281805	0.041265
SkinThickness	-0.081672	0.057328	0.207371	1.000000	0.436783	0.392573	0.183928
Insulin	-0.073535	0.331357	0.088933	0.436783	1.000000	0.197859	0.185071
BMI	0.017683	0.221071	0.281805	0.392573	0.197859	1.000000	0.140647
DiabetesPedigreeFunction	-0.033523	0.137337	0.041265	0.183928	0.185071	0.140647	1.000000
Age	0.544341	0.263514	0.239528	-0.113970	-0.042163	0.036242	0.033561
Outcome	0.221898	0.466581	0.065068	0.074752	0.130548	0.292695	0.173844

FIGURE 4.1.3 CORRELATION

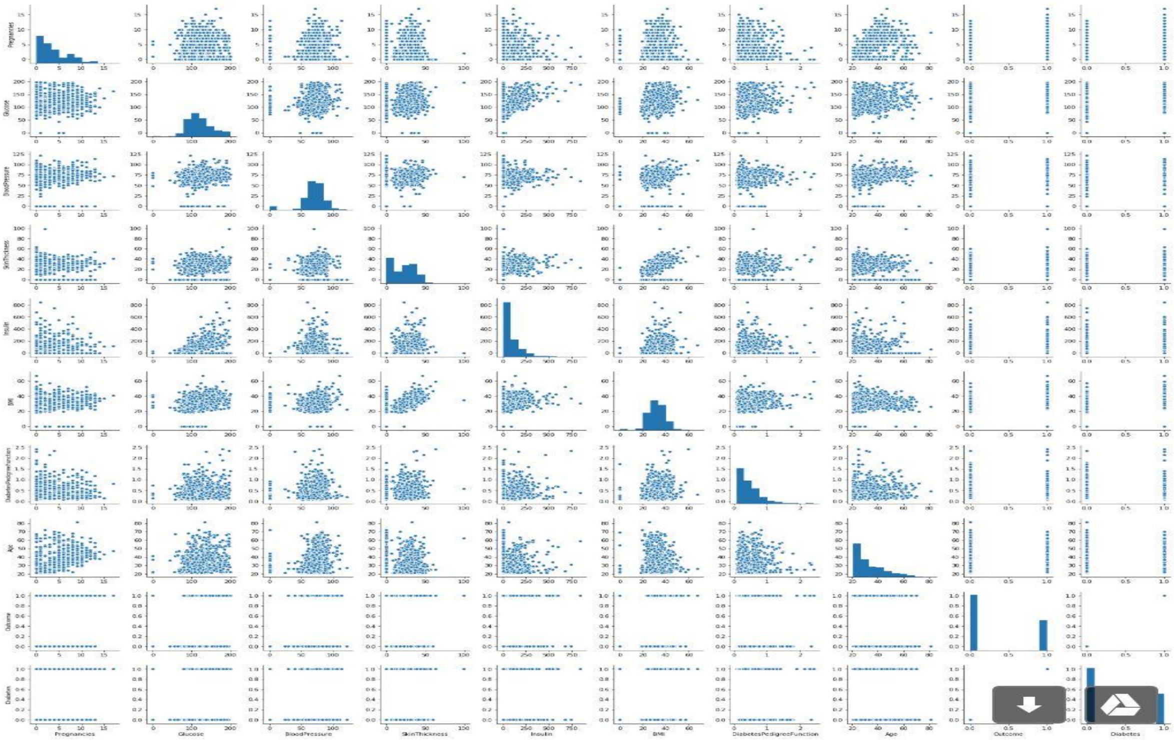


FIGURE 4.1.4 PAIR PLOT

4.2 DATA MODELLING:

By using Classification technique, Logistic Regression algorithm we predict whether a person has diabetes or not. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. A binary logistic model has a dependent variable with two possible values, such as whether the person is the victim of diabetes or not which is represented by an indicator variable, where the two values are labeled"0"and“1”.

Logistic Regression classifier is used in our model, this is because of its easy implementation and its better accuracy comparatively with other classifier algorithms. It is a linear kind of algorithm; it has a special function called sigmoid function which is applied to achieve a logistic curve. When linear regression is applied to sigmoid function it becomes logistic regression algorithm .

This regression class is present in sklearn.linear_model. The range of sigmoid function is between 0 and 1 . This prediction is done based on the information provided by the patient such as blood pressure, body mass index (BMI), age, glucose level , insulin , skin thickness , diabetes pedigree function and sex . Thus, the logistic regression algorithm which we used mainly focused to find the chances of occurrence of diabetes or not

5.REFERENCES:

https://www.kaggle.com/uciml/pima-indians-diabetes-database

https://datahub.io/machine-learning/diabetes#resource-diabetes_arff

https://catalog.data.gov/dataset?tags=diabetes

https://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/diabetes.arff

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6232260/

https://care.diabetesjournals.org/content/31/10/2056

https://towardsdatascience.com/logistic-regression-classifier-8583e0c3cf9

https://nodered.org/docs/getting-started/ibmcloud

6.CONCLUSION:

The objective of this project is to build a predictive model which identifies the patients who has the likelihood of Diabetes Mellitus. To classify the patients into diabetic and non-diabetic we have developed a model and analysed it by using the predictive model Logistic Regression. Thus, the experimental results suggested that the model has achieved good results, which is evaluated based on the parameters ROC and by accuracy metrics.