# A

# Project Report

# on

**“ Disease Prediction ”**

Submitted To

D Y Patil International University

Akurdi, Pune

In the partial fulfillment for requirement of the degree

**Masters of Computer Application**

**(MCA – I, Sem – I, 2023-24)**

Submitted By

Mr. Vishal Uttam Katke – 20230804062

Mrs. Arya Madan Tingare - 20230804075

Under The Guidance of

Prof. DR. Kapil Sharma Sir

Mrs. Remya Praveen Ma’am

Through The

School of Computer Science,

Engineering and Applications Pune



**D Y Patil International University Akurdi, Pune**

School of Computer Science, Engineering and Applications

CERTIFICATE

This is certified that a Project Report titled **“Disease Prediction”** Submitted by **Mr. Vishal Uttam Katke, Mrs. Arya Madan Tingare** in partial fulfilment for requirement of the degree of ‘Masters of Computer Application’ submitted to D Y Patil International University Akurdi, Pune has been completed under my guidance.

To the best of my knowledge and belief matter presented by them is original in nature and has not been copied down from any sources.

**Place –** Pune

**Date –** 04th December 2023

|  |  |  |
| --- | --- | --- |
| Mrs. Remya Praveen Ma’am  (Technical Assistant) | Prof. DR. Kapil Sharma Sir.  (Subject Teacher) | Dr. Maheshwari Birdar  (Head of Department) |

Declaration

**To,**

**The Head of Department**

**Masters of Computer Application**

Respected Ma’am,

We the undersigned hereby declare that the Project report entitled, **“Disease Prediction”** written and submitted by **Mr. Vishal Uttam Katke, Mrs. Arya Madan Tingare** under the guidance of **Prof. DR. Kapil Sharma Sir, Mrs. Remya Praveen Ma’am** is our Original Work.

The Empirical findings in this report are based on information collected by us. We have not copied from any report submitted to D Y Patil International University Akurdi, Pune. We understand any such copying is such liable in a punishment in a way that the university authority demits this.

**Place –** Pune

**Date –** 04th December 2023

Mr. Vishal Uttam Katke – 20230804062

Mrs. Arya Madan Tingare – 20230804075

(MCA – 1, Sem - 1)

Acknowledgement

We take this opportunity to express our deep sense of obligation to the D Y Patil International University’s School of Computer Science, Engineering and Applications, Akurdi, Pune.

We express our sincere thanks to the **HOD** for giving us an opportunity to undertaking this Project work. We owe our deepest gratitude towards respective guide **Prof. DR. Kapil Sharma Sir** and all of our faculty members for their valuable guidance and motivation during the completion of the Project Report.

Finally, we are very much thankful to everyone who helped us a lot in the completion of this Report to the great extent.

Thanking you all.

**Place –** Pune

**Date –** 04th December 2023

Mr. Vishal Uttam Katke – 20230804062

Mrs. Arya Madan Tingare – 20230804075

(MCA – 1, Sem - 1)

## Index

|  |  |  |
| --- | --- | --- |
| *Sr. No.* | *Title* | *Page No.* |
| *1* | *Abstract* | *6* |
| *2* | *Introduction* | *7* |
| *3* | *Dataset Used* | *8* |
| *4* | *Algorithms Used* | *9* |
| *5* | *Flow Chart* | *10* |
| *6* | *Design of the system* | *11* |
| *7* | *Code & Output* | *12* |
| *8* | *Bibliography* | *21* |

## Abstract

Disease Prediction using Machine Learning is the system that is used to predict the diseases from the symptoms which are given by the patients or any user. The system processes the symptoms provided by the user as input and gives the output as the probability of the disease. Naïve Bayes classifier is used in the prediction of the disease which is a supervised machine learning algorithm.

The probability of the disease is calculated by the Naïve Bayes algorithm. With an increase in biomedical and healthcare data, accurate analysis of medical data benefits early disease detection and patient care. By using linear regression and decision tree we are predicting diseases like Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis.

People with diabetes for a long time can get several complications like heart disorder, kidney disease, nerve damage, diabetic retinopathy etc. But its risk can be reduced if it is predicted early. In this paper, an automatic diabetes prediction system has been developed using a private dataset of female patients in Bangladesh and various machine learning techniques.

The authors used the Pima Indian diabetes dataset and collected additional samples from 203 individuals from a local textile factory in Bangladesh. Feature selection algorithm mutual information has been applied in this work. A semi‐supervised model with extreme gradient boosting has been utilized to predict the insulin features of the private dataset. SMOTE and ADASYN approaches have been employed to manage the class imbalance problem. The authors used machine learning classification methods, that is, decision tree, SVM, Random Forest, Logistic Regression, KNN, and various ensemble techniques, to determine which algorithm produces the best prediction results.

The explainable AI approach with LIME and SHAP frameworks is implemented to understand how the model predicts the final results. Finally, a website framework and an Android smartphone application have been developed to input various features and predict diabetes instantaneously.

## Introduction

Machine Learning is the domain that uses past data for predicting. Machine Learning is the understanding of

computer system under which the Machine Learning model learn from data and experience. The

machinelearning algorithm has two phases: 1) Training & 2) Testing. To predict the disease from a patient’s

symptoms and from the history of the patient, machine learning technology is struggling from past decades.

Healthcare issues can be solved efficiently by using Machine Learning Technology. We are applying

complete machine learning concepts to keep the track of patient’s health.

ML model allows us to build models to get quickly cleaned and processed data and deliver results faster. By

using this system doctors will make good decisions related to patient diagnoses and according to that, good

treatment will be given to the patient, which increases improvement in patient healthcare services. To

introduce machine learning in the medical field, healthcare is the prime example. To improve the accuracy of

large data, the existing work will be done on unstructured or textual data. For the prediction of diseases, the

existing will be done on linear, KNN, Decision Tree algorithm.

Machine Learning is the domain that uses past data for predicting. Machine Learning is the understanding of

computer system under which the Machine Learning model learn from data and experience. The

machinelearning algorithm has two phases: 1) Training & 2) Testing. To predict the disease from a patient’s

symptoms and from the history of the patient, machine learning technology is struggling from past decades.

Healthcare issues can be solved efficiently by using Machine Learning Technology. We are applying

complete machine learning concepts to keep the track of patient’s health.

ML model allows us to build models to get quickly cleaned and processed data and deliver results faster. By

using this system doctors will make good decisions related to patient diagnoses and according to that, good

treatment will be given to the patient, which increases improvement in patient healthcare services. To

introduce machine learning in the medical field, healthcare is the prime example. To improve the accuracy of

large data, the existing work will be done on unstructured or textual data. For the prediction of diseases, the

existing will be done on linear, KNN, Decision Tree algorithm.

Machine Learning is the domain that uses past data for predicting. Machine Learning is the understanding of

computer system under which the Machine Learning model learn from data and experience. The

machinelearning algorithm has two phases: 1) Training & 2) Testing. To predict the disease from a patient’s

symptoms and from the history of the patient, machine learning technology is struggling from past decades.

Healthcare issues can be solved efficiently by using Machine Learning Technology. We are applying

complete machine learning concepts to keep the track of patient’s health.

ML model allows us to build models to get quickly cleaned and processed data and deliver results faster. By

using this system doctors will make good decisions related to patient diagnoses and according to that, good

treatment will be given to the patient, which increases improvement in patient healthcare services. To

introduce machine learning in the medical field, healthcare is the prime example. To improve the accuracy of

large data, the existing work will be done on unstructured or textual data. For the prediction of diseases, the

existing will be done on linear, KNN, Decision Tree algorithm.

Machine Learning is the domain that uses past data for predicting. Machine Learning is the understanding of computer system under which the Machine Learning model learn from data and experience.

The machine learning algorithm has two phases:

1. Training
2. Testing.

To predict the disease from a patient’s symptoms and from the history of the patient, machine learning technology is struggling from past decades. Healthcare issues can be solved efficiently by using Machine Learning Technology.

We are applying complete machine learning concepts to keep the track of patient’s health. ML model allows us to build models to get quickly cleaned and processed data and deliver results faster. By using this system doctors will make good decisions related to patient diagnoses and according to that, good treatment will be given to the patient, which increases improvement in patient healthcare services.

To introduce machine learning in the medical field, healthcare is the prime example. To improve the accuracy of large data, the existing work will be done on unstructured or textual data. For the prediction of diseases, the existing will be done on linear, KNN, Decision Tree algorithm.

Any technology user today has benefitted from machine learning. Machine learning is a subfield of artificial Intelligence. The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people.

Machine learning is used across many spheres around the world. The health care industry is not an exception.

Machine learning is used in Medical diagnosis, Stock market trading, Image recognition, Self-driving cars and more.

Such information, if predicted well in advance, can provide important insights to doctors who can then adapt their diagnosis and treatment per patient basis.

## Dataset Used

*Context :*

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict based on diagnostic measurements whether a patient has diabetes.

*Content :*

Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

*Columns of Dataset :*

* Pregnancies: Number of times pregnant
* Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
* BloodPressure: Diastolic blood pressure (mm Hg)
* SkinThickness: Triceps skin fold thickness (mm)
* Insulin: 2-Hour serum insulin (mu U/ml)
* BMI: Body mass index (weight in kg/(height in m)^2)
* DiabetesPedigreeFunction: Diabetes pedigree function
* Age: Age (years)
* Outcome: Class variable (0 or 1)

## Algorithms Used

***The Supervised Learning/Predictive Models***

Supervised learning algorithms are used to construct predictive models. A predictive model predicts missing valueusing other values present in the dataset. Supervised learning algorithm has a set of input data and also a set of output, and builds a model to make realistic predictions for the response to new dataset. Supervised learning includes Decision Tree, Bayesian Method, Artificial Neural Network, Instance based learning, Ensemble Method. These are booming techniques in Machine learning.

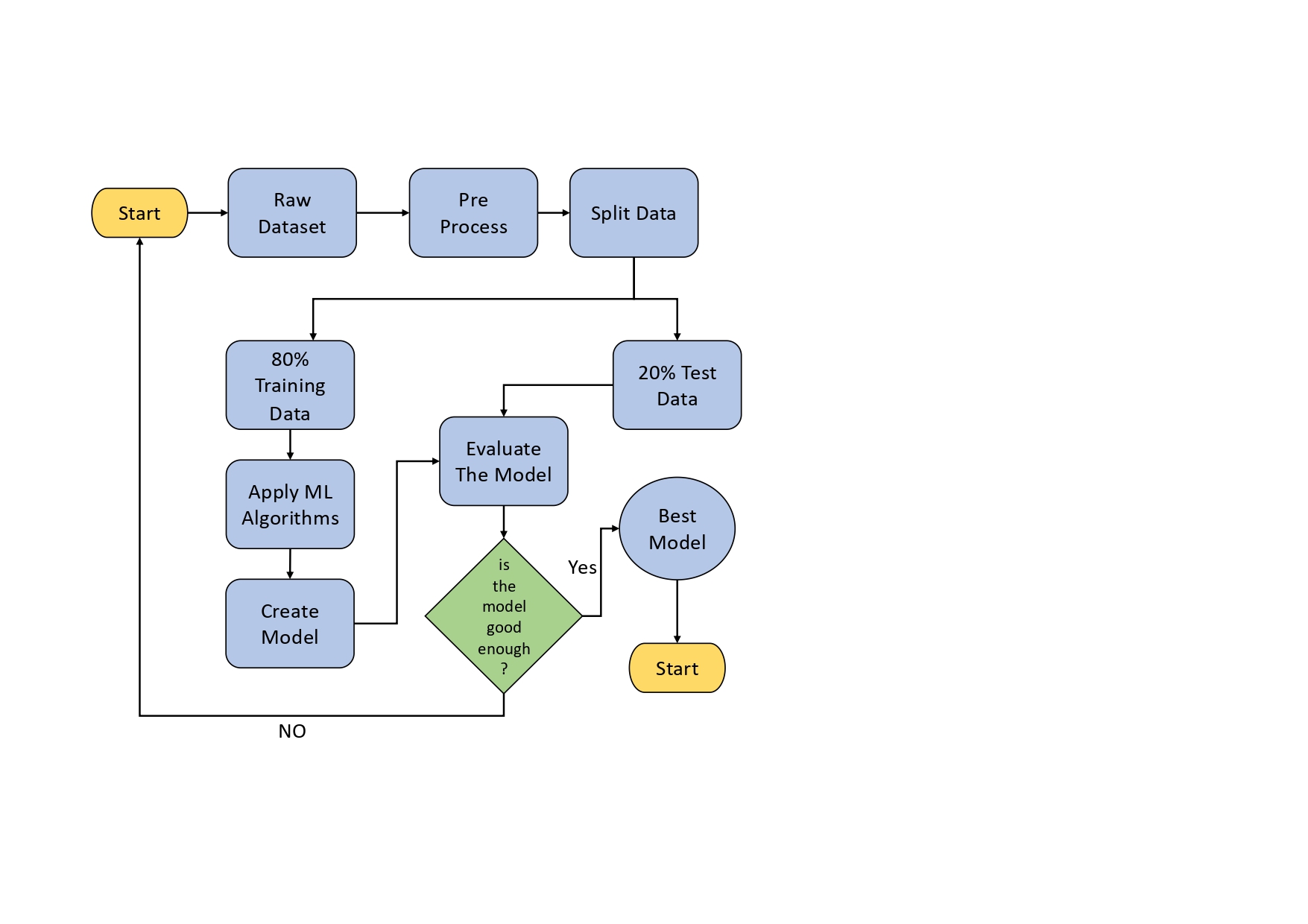
***The Unsupervised Learning / Descriptive Models***

Descriptive models are developed using unsupervised learning method. In this model we have known set of inputs but output is unknown. Unsupervised learning is mostly used on transactional data. This method includes clustering algorithms like k-Means clustering and k-Medians clustering.[3]

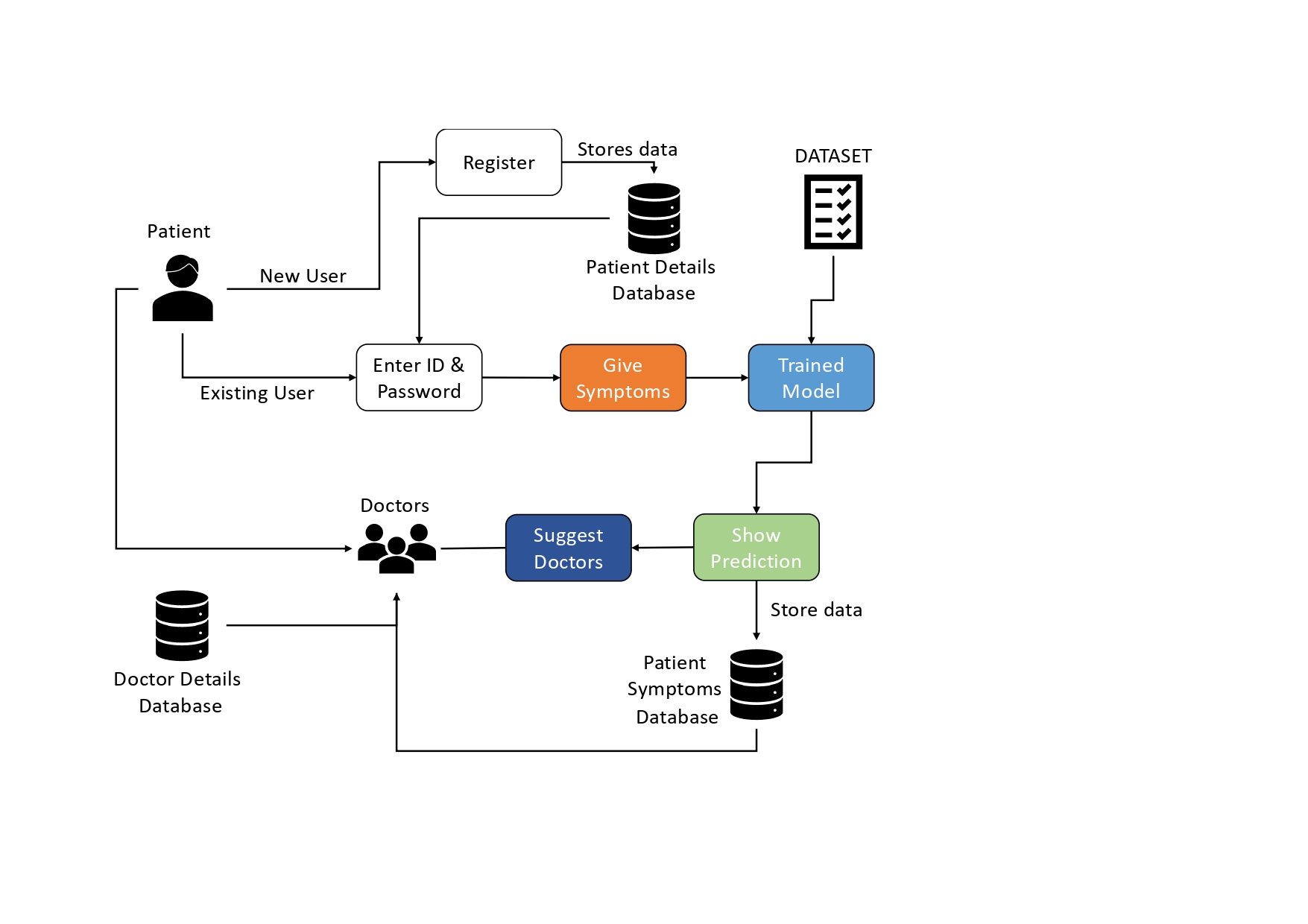
***Semi-supervised Learning***

Semi Supervised learning method uses both labelled and unlabelled data on training dataset. Classification, Regression techniques come under Semi Supervised Learning. Logistic Regression, Linear Regression are examples of regression techniques.

## Flow Chart

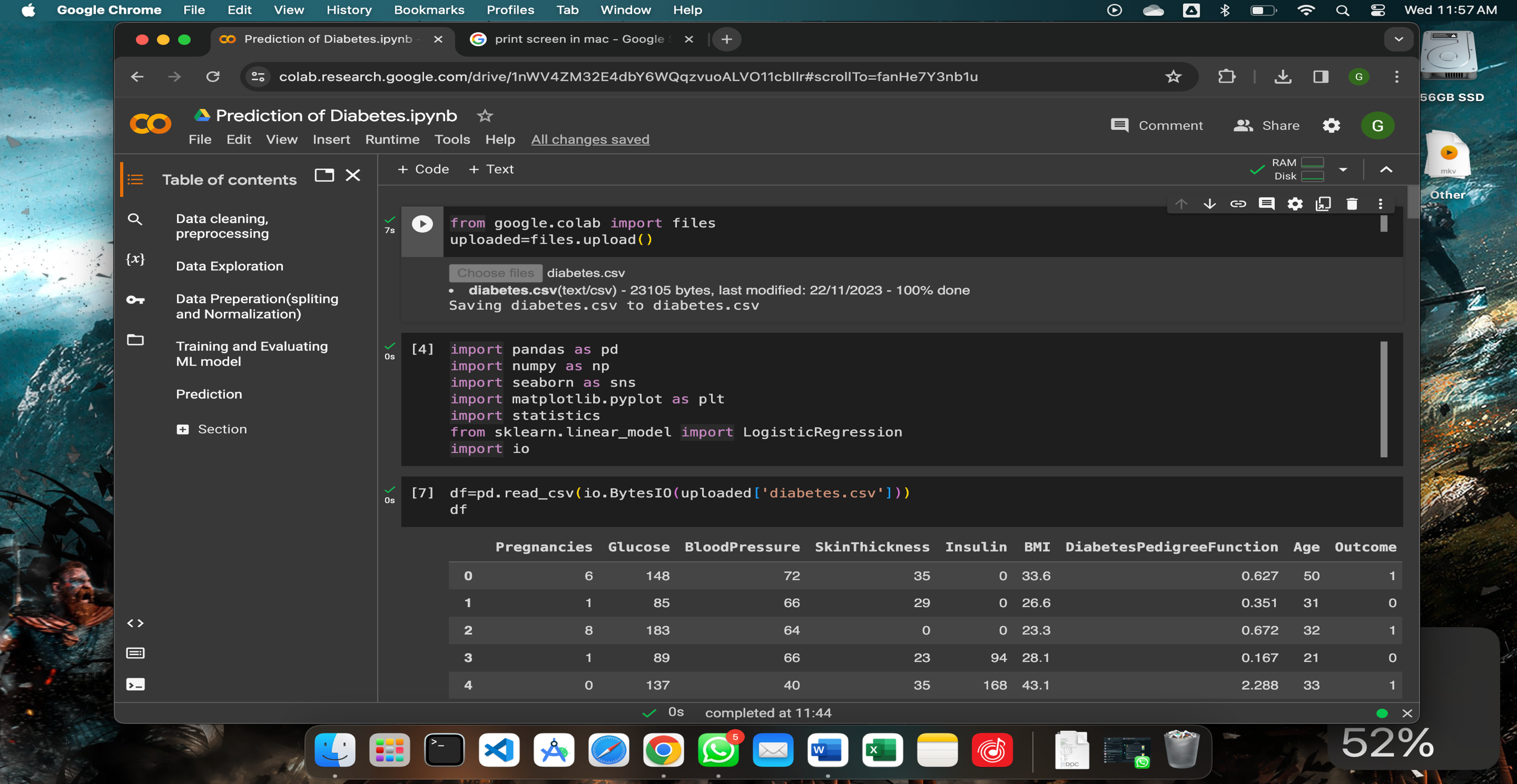


## Design of the system

****

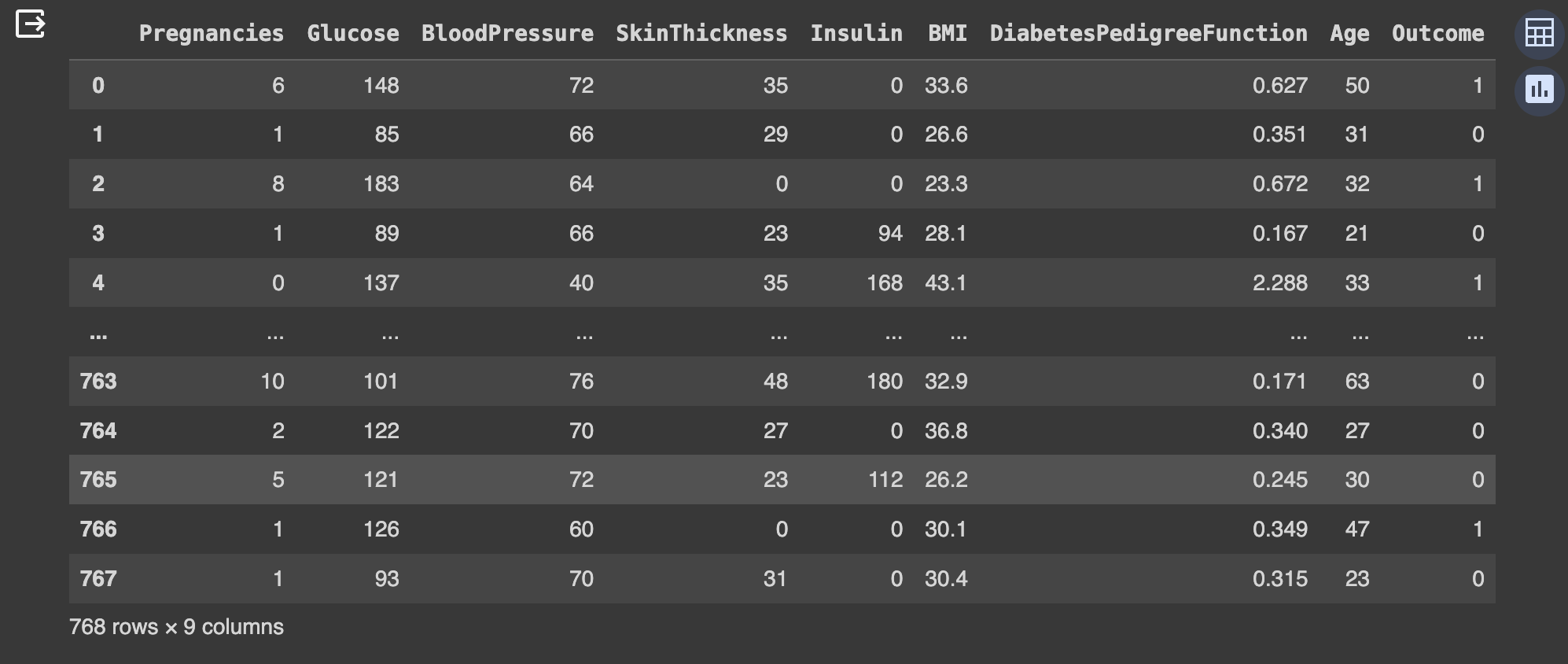
## Code & Output

from google.colab import files  
uploaded=files.upload()



import pandas as pd  
import numpy as np  
import seaborn as sns  
import matplotlib.pyplot as plt  
import statistics  
from sklearn.linear\_model import LogisticRegression  
import io

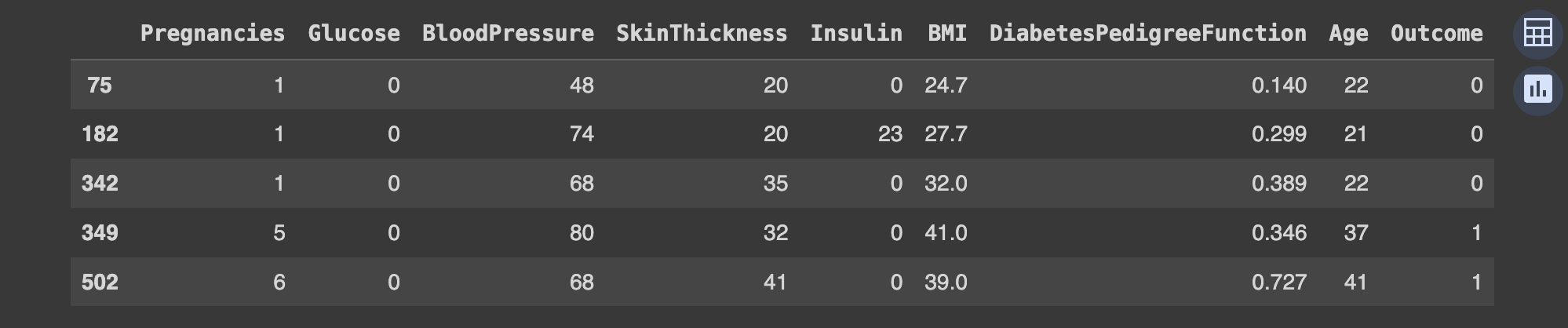
df=pd.read\_csv(io.BytesIO(uploaded['diabetes.csv']))  
df



# Data cleaning, pre-processing

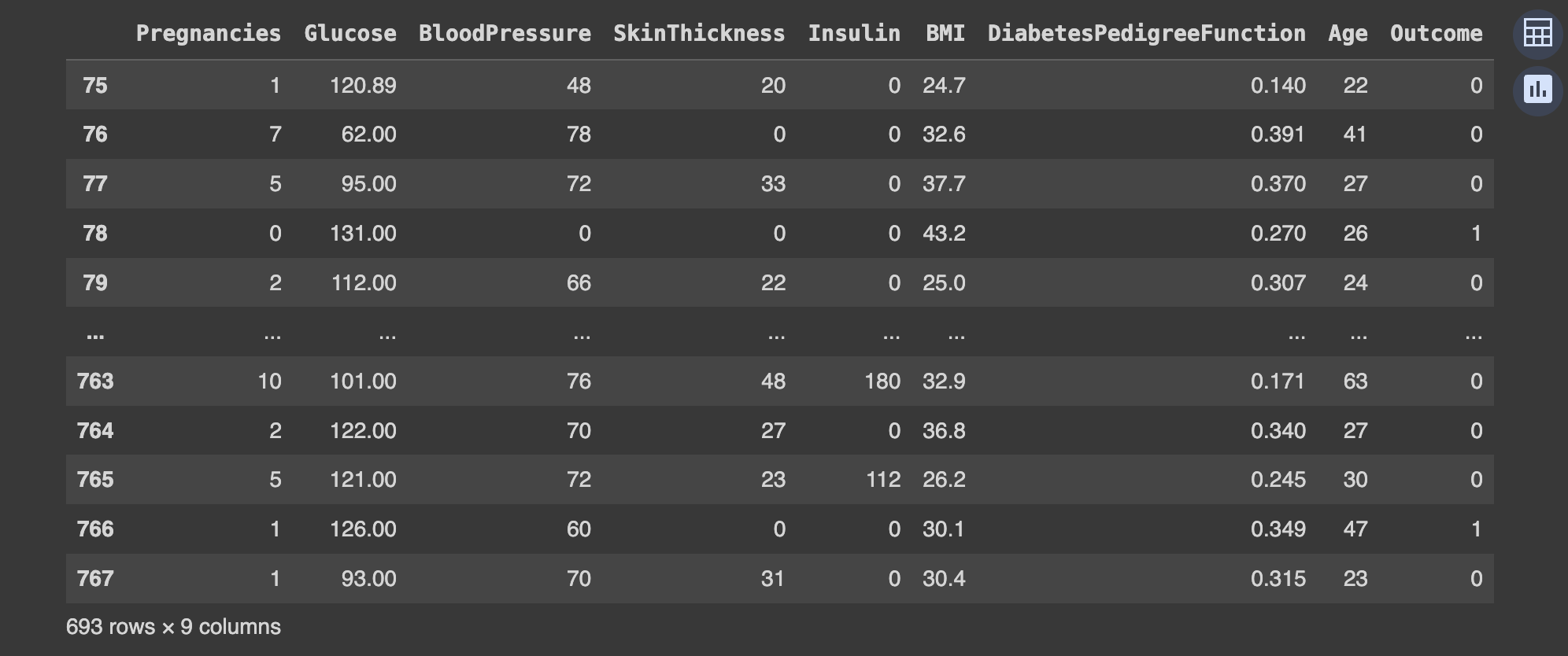
***Removing noisy data***

df[df['Glucose']==0]



df['Glucose'].replace({0:round(df['Glucose'].mean(),2)},inplace=True)

df.loc[75:]



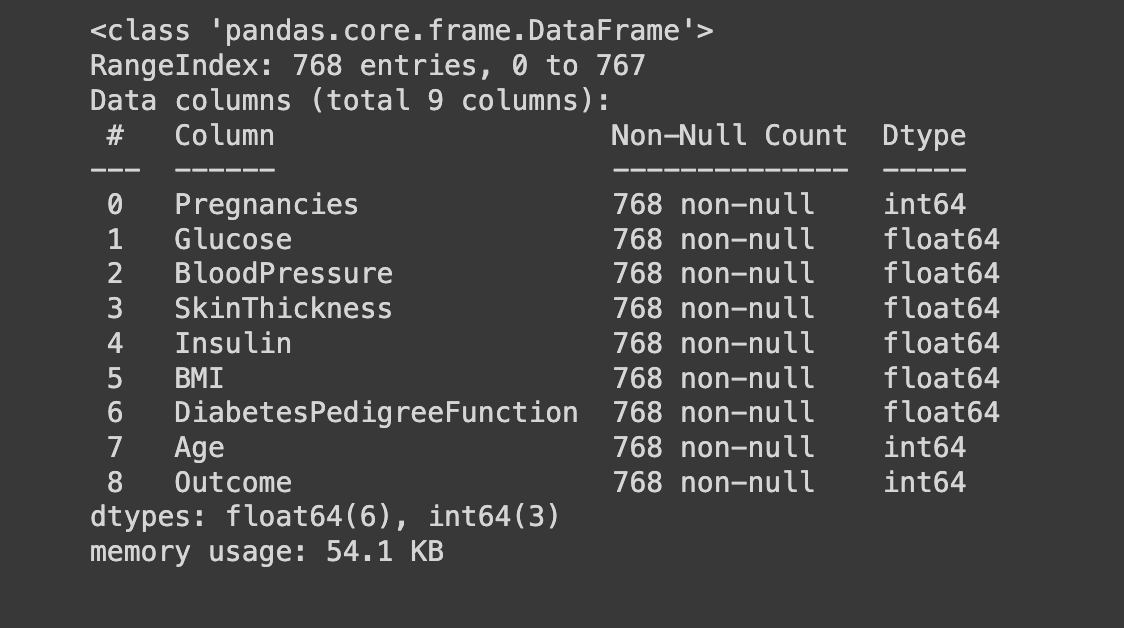
df['BloodPressure'].replace({0:round(df['BloodPressure'].mean(),2)},inplace=True)

df['SkinThickness'].replace({0:round(df['SkinThickness'].mean(),2)},inplace=True)

df['Insulin'].replace({0:round(df['Insulin'].mean(),2)},inplace=True)

df['BMI'].replace({0:round(df['BMI'].mean(),2)},inplace=True)

df.info()

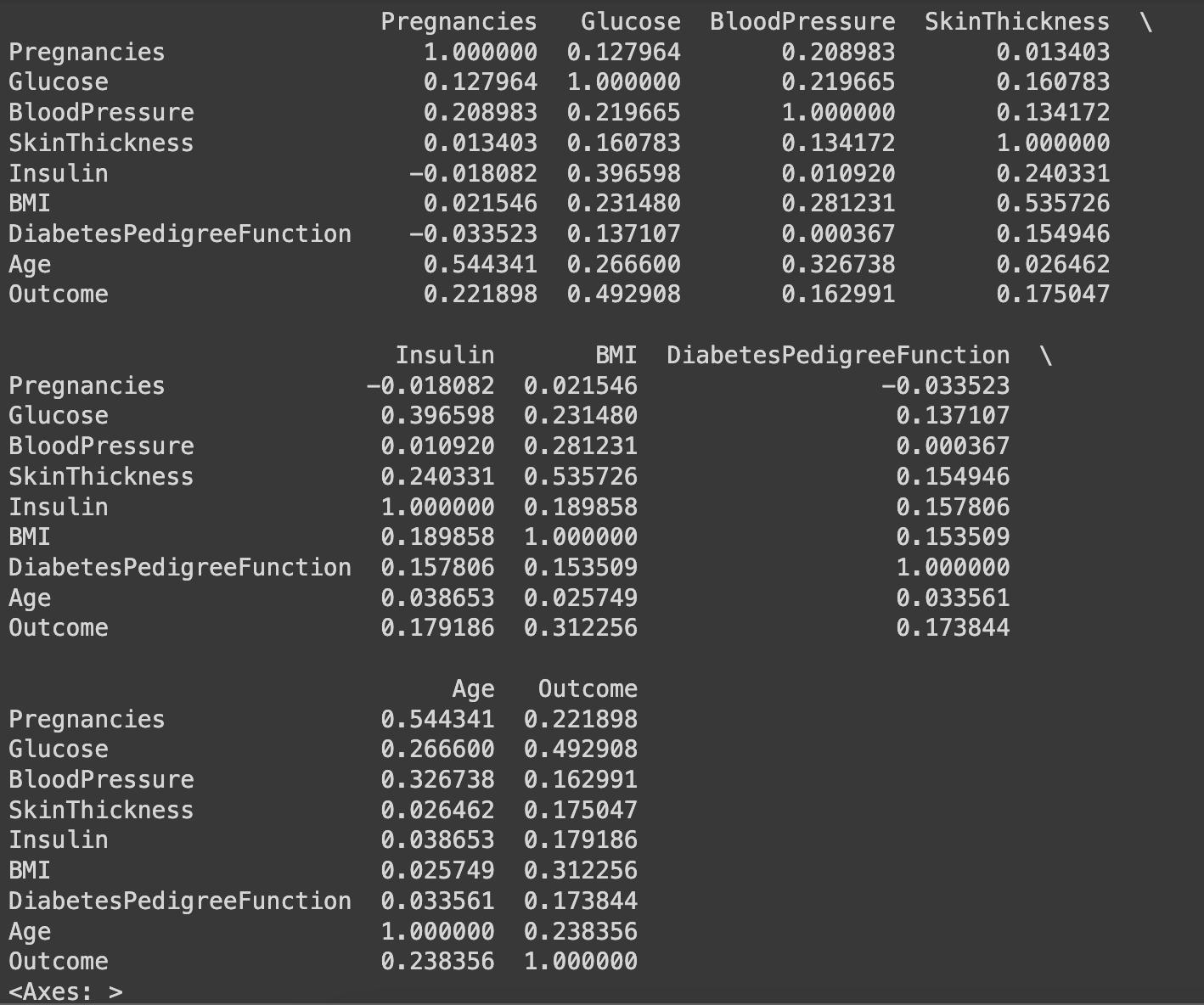


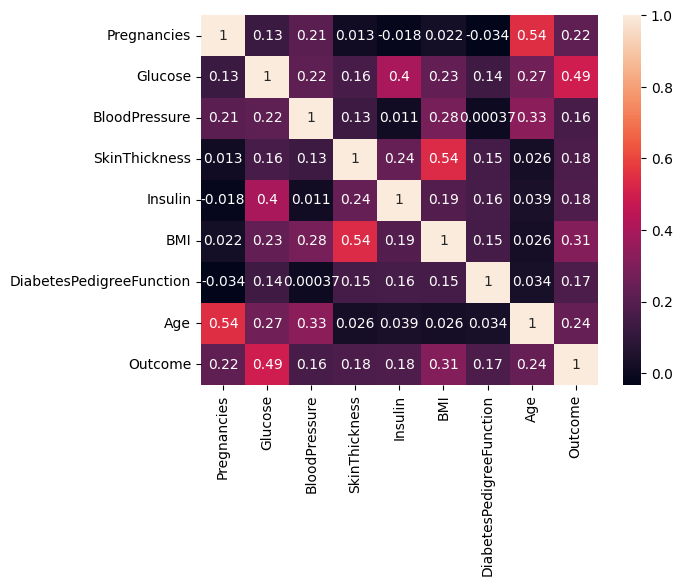
# 

# Data Exploration

***Finding correlation of every pair of features (and the outcome variable), and visualize the correlations using a heatmap.***

corr = df.corr()  
print(corr)  
sns.heatmap(corr,  
 xticklabels=corr.columns,  
 yticklabels=corr.columns,  
 annot=True)





In the above heatmap, brighter colors indicate more correlation. Glucose levels, age, BMI and number of pregnancies all have significant correlation with the outcome variable. Also notice the correlation between pairs of features, like age and pregnancies, or BMI and skin thickness.

# Data Preperation(spliting and Normalization)

***The data set consists of record of 767 patients in total. To train - 667records. 100 records for testing.***

dfTrain = df[:667]  
dfTest = df[667:]

***Separate the label and features (for both training and test dataset). Convert them into NumPy arrays as the machine learning algorithm process data in NumPy array format.***

trainLabel = np.asarray(dfTrain['Outcome'])  
trainData = np.asarray(dfTrain.drop('Outcome',1))  
testLabel = np.asarray(dfTest['Outcome'])  
testData = np.asarray(dfTest.drop('Outcome',1))

******

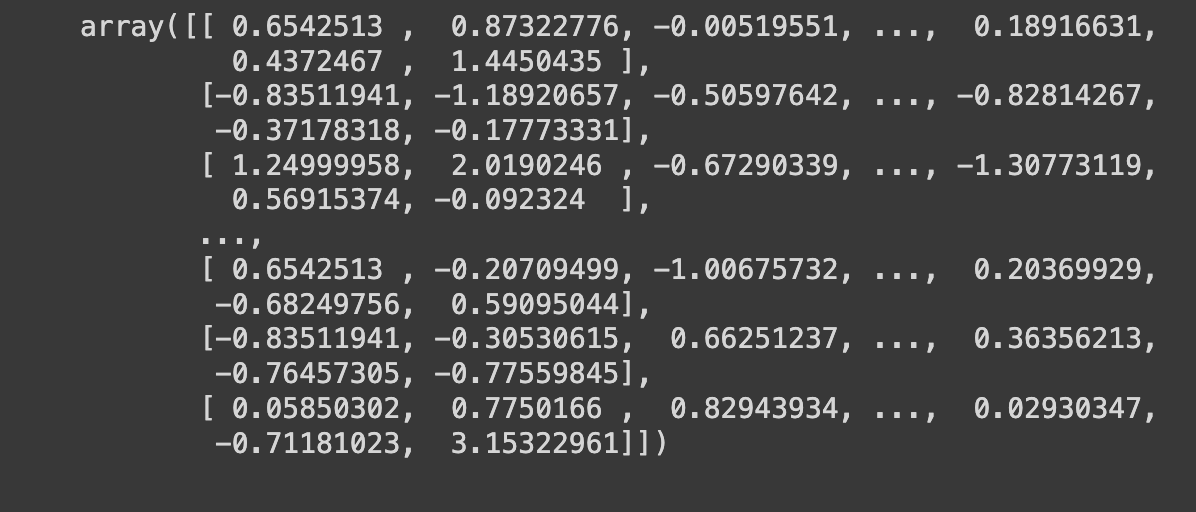
***Normalize the data such that each variable has 0 mean and standard deviation of 1. Standardizing the features around the center and 0 with a standard deviation of 1 is important when we compare measurements that have different units. Variables that are measured at different scales do not contribute equally to the analysis and might end up creating a bias.***

#splitting  
dfTrain = df[:667]  
dfTest = df[667:]  
#seperate label features  
trainLabel = np.asarray(dfTrain['Outcome'])  
trainData = np.asarray(dfTrain.drop('Outcome',1))  
testLabel = np.asarray(dfTest['Outcome'])  
testData = np.asarray(dfTest.drop('Outcome',1))  
#Normalization  
means = np.mean(trainData, axis=0)  
stds = np.std(trainData, axis=0)  
trainData = (trainData - means)/stds  
testData = (testData - means)/stds

******

means = np.mean(trainData, axis=0)  
stds = np.std(trainData, axis=0)  
trainData = (trainData - means)/stds  
testData = (testData - means)/stds

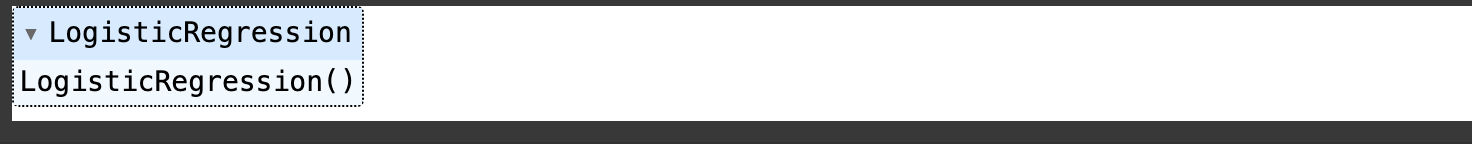
trainData



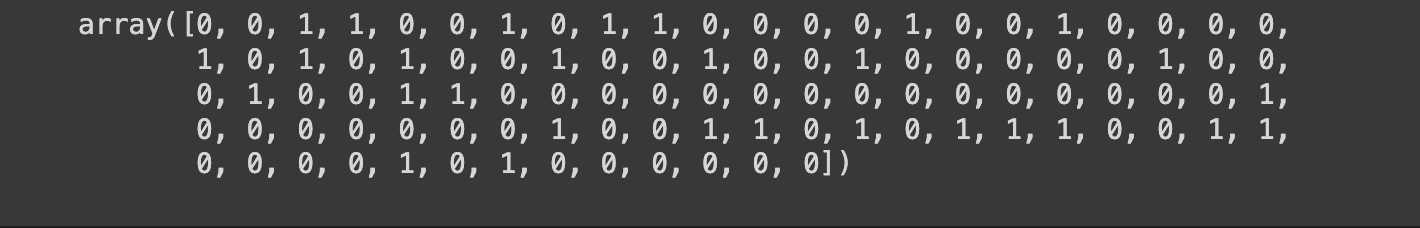
# 

# Training and Evaluating ML model

model = LogisticRegression()  
model.fit(trainData, trainLabel)



model.predict(testData)

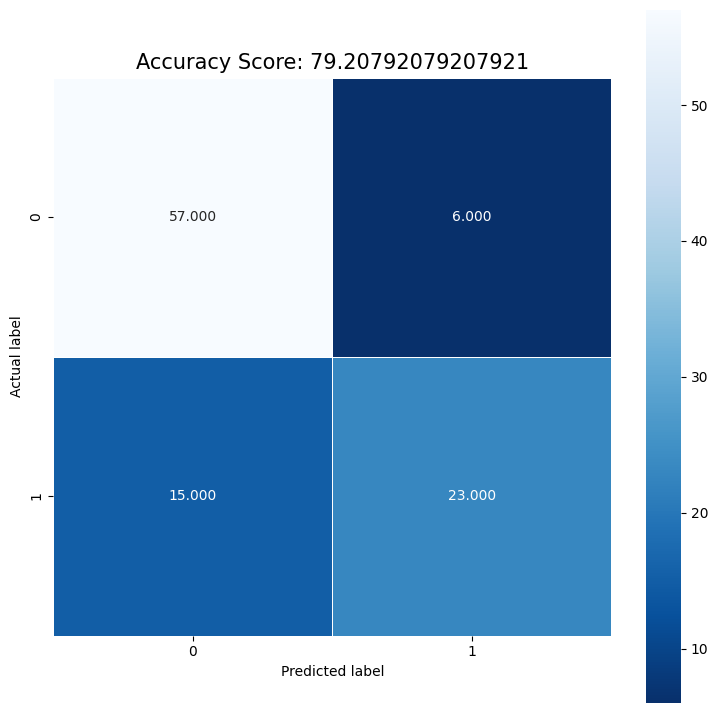


predictions = model.predict(testData)

from sklearn import metrics  
cm = metrics.confusion\_matrix(testLabel, predictions)  
print(cm)



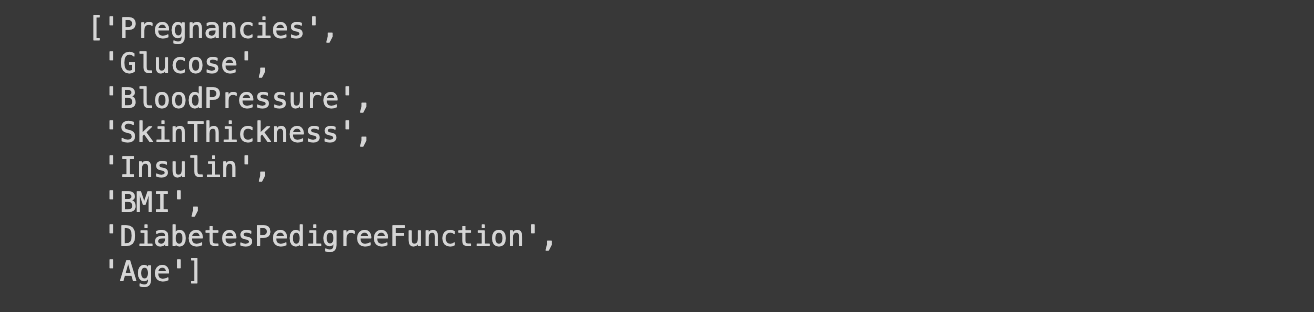
model = LogisticRegression()  
model.fit(trainData, trainLabel)  
  
from sklearn import metrics  
cm = metrics.confusion\_matrix(testLabel, predictions)  
  
plt.figure(figsize=(9,9))  
sns.heatmap(cm, annot=True, fmt=".3f", linewidths=.5, square = True, cmap = 'Blues\_r');  
plt.ylabel('Actual label');  
plt.xlabel('Predicted label');  
all\_sample\_title = 'Accuracy Score: {0}'.format(model.score(testData, testLabel)\*100)  
plt.title(all\_sample\_title, size = 15);



accuracy = model.score(testData, testLabel)  
print("accuracy = ", accuracy \* 100, "%")



lab=df.drop(columns=['Outcome'])  
lab.columns.to\_list()



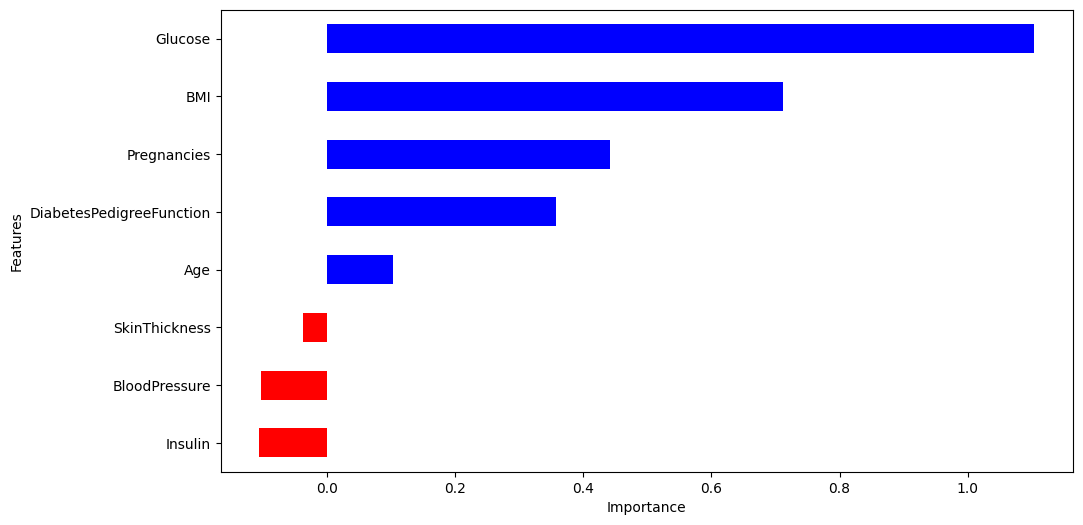
***Visualization of the weights in the Logistic Regression model corresponding to each of the feature variables***

df



coeff = list(model.coef\_[0])  
labels = list(lab.columns)  
features = pd.DataFrame()  
features['Features'] = labels  
features['importance'] = coeff  
features.sort\_values(by=['importance'], ascending=True, inplace=True)  
features['positive'] = features['importance'] > 0  
features.set\_index('Features', inplace=True)  
features.importance.plot(kind='barh', figsize=(11, 6),color = features.positive.map({True: 'blue', False: 'red'}))  
plt.xlabel('Importance')

Text(0.5, 0, 'Importance')



# Prediction

dfCheck = pd.DataFrame(df)

sampleData = dfCheck[:1]  
print(sampleData)  
# prepare sample  
sampleDataFeatures = np.asarray(sampleData.drop('Outcome',1))  
sampleDataFeatures = (sampleDataFeatures - means)/stds



# predict  
predictionProbability = model.predict\_proba(sampleDataFeatures)  
prediction = model.predict(sampleDataFeatures)  
print('Probability:', predictionProbability)  
print('prediction:', prediction)

Probability: [[0. 1.]]  
prediction: [1]



## Bibliography

Book References :-

1) Data Science And Machine Learning

2) Data Science And Machine Learning With R

3) Big Data Science Analytics & Machine Learning

4) Data Science: a beginner’s guide

5) Capitalizing Data Science

Web-Site References :-

1) https://www.kaggle.com/

2) https://www.w3schools.com/

3) https://websitesetup.org/

4) https://www.youtube.com/

5) https://stackoverflow.com/