DOCUMENTATION ON gestational DIABETES PREDICTION

MODEL NAME:- GESTATIONAL DIABETES PREDICTION

AIM:-

The objective of the dataset is to diagnostically predict whether or not a patient has gestational diabetes, based on certain diagnostic measurements included in the dataset. 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.

Input to the model:-

The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

Output :-

whether a patient has gestational diabetes or not.

THIS MODEL HELPS TO WHICH DEPARMENT PEOPLE?

Gestational diabetes is a type of diabetes diagnosed for the first time during pregnancy.IF you have gestational diabetes during pregnancy ,generally your blood sugar returns to its usual level soon after the delivery,But if you’ve had gestational diabetes,you have a higher risk of getting type 2 diabetes.You’ll need to be tested for changes in blood sugar more often.

Diabetes can be controlled if it is predicted earlier. To achieve this goal this work we will do early prediction of Diabetes in a human body or a patient for a higher accuracy through applying, Various MACHINE LEARNING TECHNIQUES.

PROBLEM STATEMENT:-

Diabetes is major cause of death in the world mainly in the pregnancy woman. Early prediction of disease like diabetes can be controlled and save the human life. To accomplish this, this work explores prediction of diabetes by taking various attributes related to diabetes disease. For this purpose we use the Pima Indian Diabetes Dataset, we apply various Machine Learning classification and ensemble Techniques to predict diabetes. Machine Learning Is a method that is used to train computers or machines explicitly. Various Machine Learning Techniques provide efficient result to collect Knowledge by building various classification and ensemble models from collected dataset. Such collected data can be useful to predict diabetes. Various techniques of Machine Learning can capable to do prediction, however it’s tough to choose best technique. Thus for this purpose we apply popular classification and ensemble methods on dataset for prediction.

Dataset Description:-

The data is gathered from UCI repository which is named as Pima Indian Diabetes Dataset. The dataset have many attributes of 3142 patients

|  |  |
| --- | --- |
| S No. | ATTRIBUTES |
| 1 | Pregnancy |
| 2 | Glucose |
| 3 | Blood Pressure |
| 4 | Skin thickness |
| 5 | Insulin |
| 6 | BMI(Body Mass Index) |
| 7 | Diabetes Pedigree Function |
| 8 | Age |
| 9 | outcome |

Details about the dataset:

The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

* **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)

LIBRARIES:-

The libraries that I’ve used to build this model are

* Numpy-is used for working with arrays
* Pandas- is an open source library mainly made for working with relational or labeled data easily.
* Matplotlib-is used to create 2D graphs and plots by using module name pyplot which makes easy for plotting.
* Seaborn- is a data visualization library based on matplotlib.it provides high level interface for drawing attractive and informative graphics.
* SCIPY:-SciPy is a scientific computation library that uses NumPy underneath.Scipy stands for scientific Python.It provides more utility functions for optimization ,stats and signal processing.Like NumPy ,SciPy is also an open source,so we can use it freely.
* Sklearn (Scikit-learn)-is used to provide a selection of efficient tools for ML and statistical modeling including classification ,regression and clustering.

StandardScaler removes the mean and scales each feature/variable to unit variance. This operation is performed feature-wise in an independent way. StandardScaler can be influenced by outliers (if they exist in the dataset) since it involves the estimation of the empirical mean and standard deviation of each feature.

Train\_test\_split:- This procedure is used to estimate the performance of ML algorithms when they are used to make predictions on data.

XGBOOST ALGORITHM: XGBoost ,which stands for extreme gradient boosting,is a scalable ,Distributed Gradient Boosted Tree(GBDT)machine learning library.It provides parallel tree boosting and is the leading machine learning library for regression,classification,and ranking problems.

RANDOMFOREST CLASSIFIER: Random Forest is a classifier that contains a number of decision trees on various subsets of te given dataset and takes the average to improve the predictive accuracy of that dataset.

Logistic Regression:-It is one of the most popular ML algorithms,which comes under the SUPERVISED LEARNING technique .It is used for predicting the categorical dependent variable using a given set of independent variables.

MODEL BUILDING:-

This is most important phase which includes model building for prediction of diabetes. In this we have implemented various machine learning algorithms which are discussed above for diabetes prediction. Procedure of Proposed Methodology:

Step1: Import required libraries, Import diabetes dataset.

Step2: Pre-process data to remove missing data.

Step3: Perform percentage split of 75% to divide dataset as Training set and 25% to Test set.

Step4: Select the machine learning algorithm i.e. KNearest Neighbor, Support Vector Machine, Decision Tree, Logistic regression, Random Forest and Gradient boosting algorithm.

Step5: Build the classifier model for the mentioned machine learning algorithm based on training set.

Step6: Test the Classifier model for the mentioned machine learning algorithm based on test set.

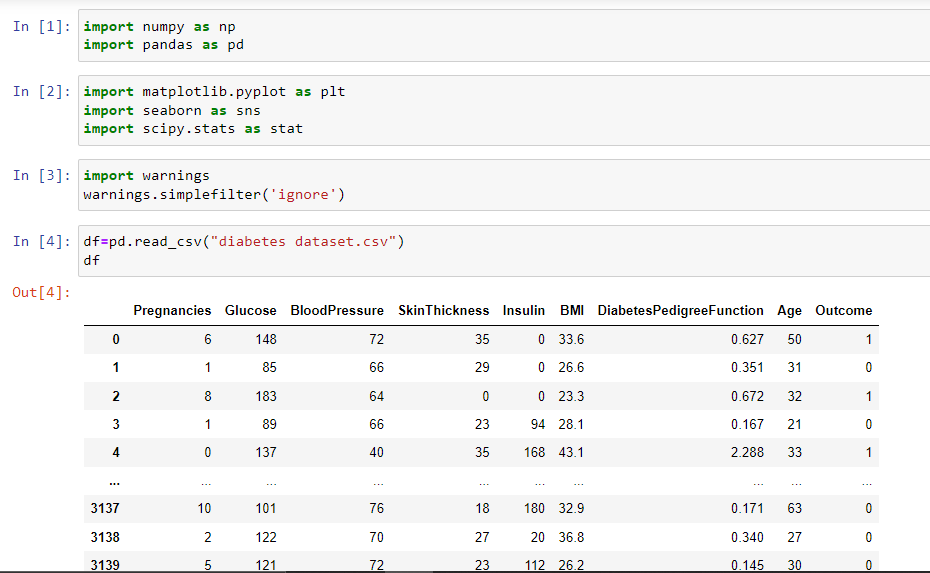
Step7: Perform Comparison Evaluation of the experimental performance results obtained for each classifier.

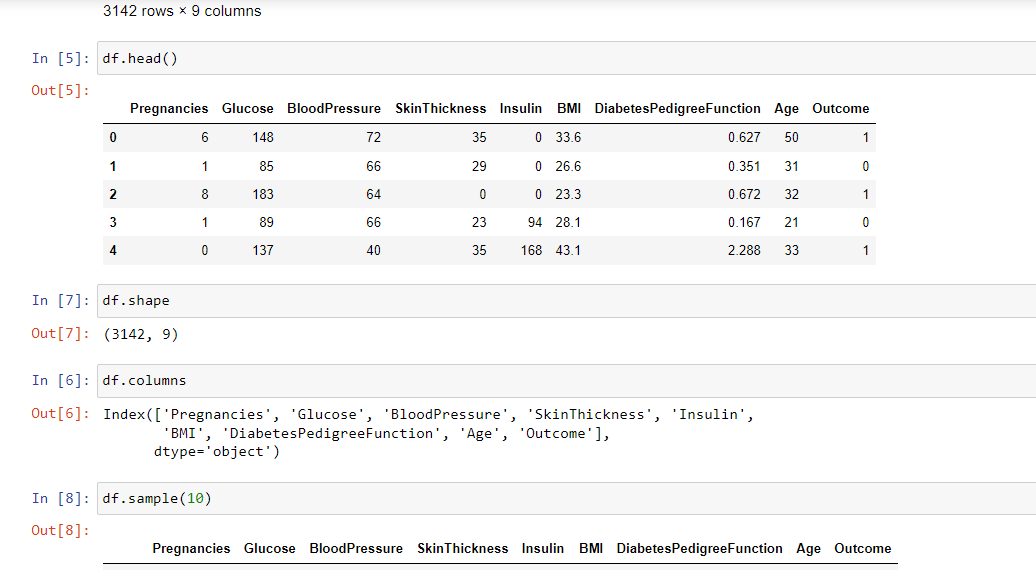
Step8: After analyzing based on various measures conclude the best performing algorithm.

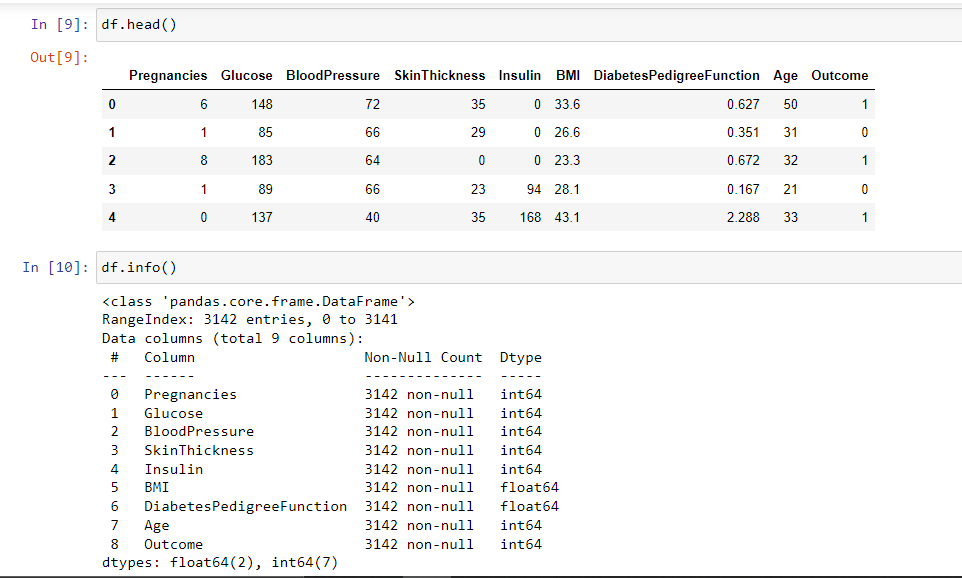
CONCLUSION:-

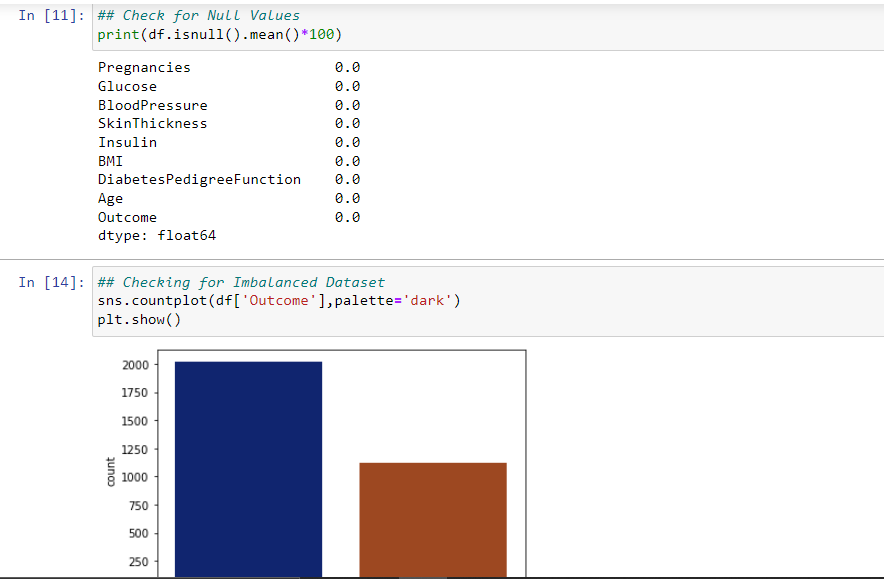
The main aim of this project was to design and implement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of that methods and it has been achieved successfully. The proposed approach uses XGBOOST classification ,92% classification accuracy has been achieved. The Experimental results can be assist health care to take early prediction and make early decision to cure diabetes and save humans life.

EXECUTING THE MODEL BY USING JUPITER NOTEBOOK:-

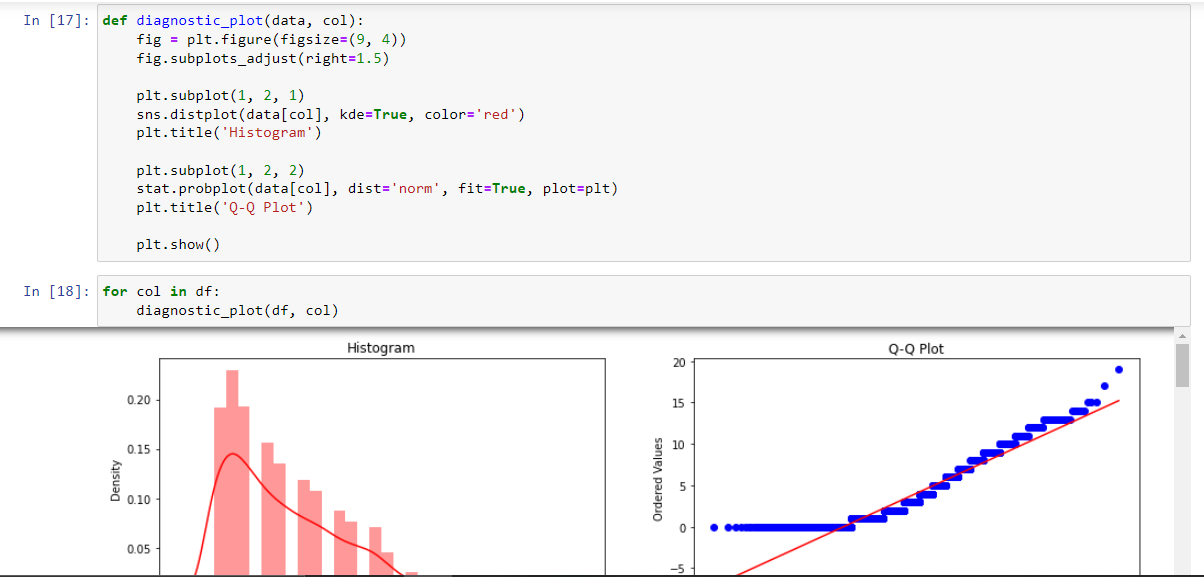


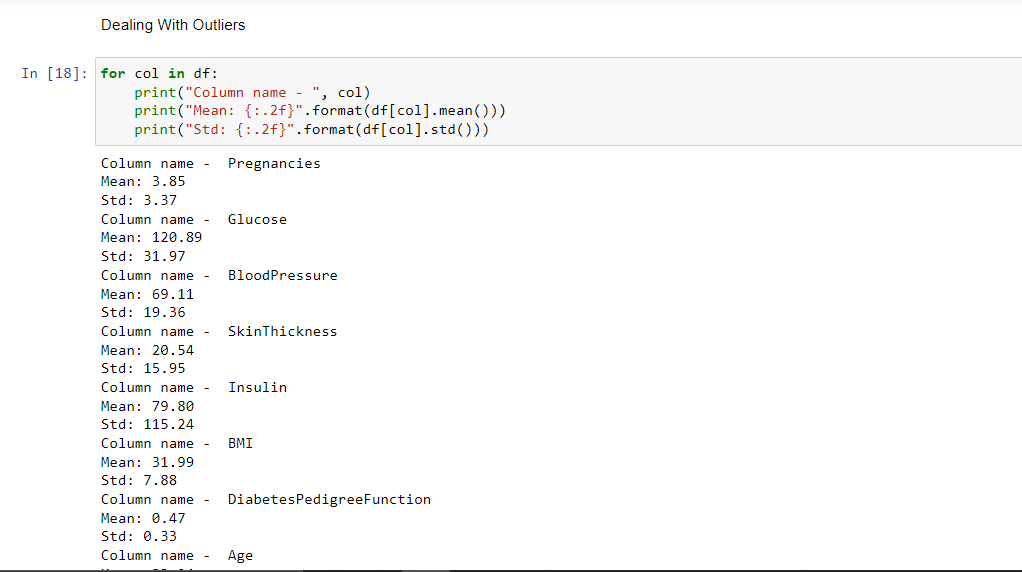




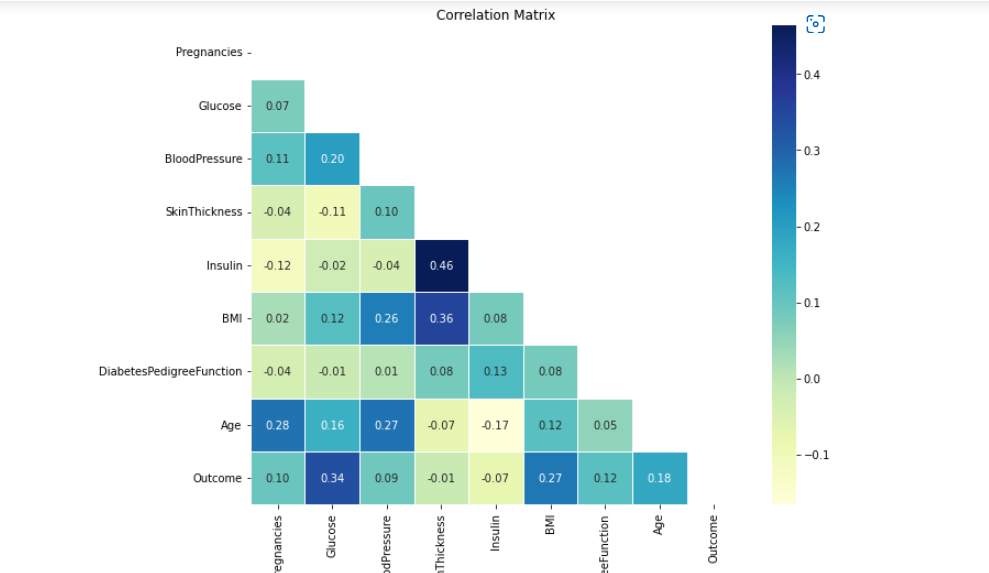




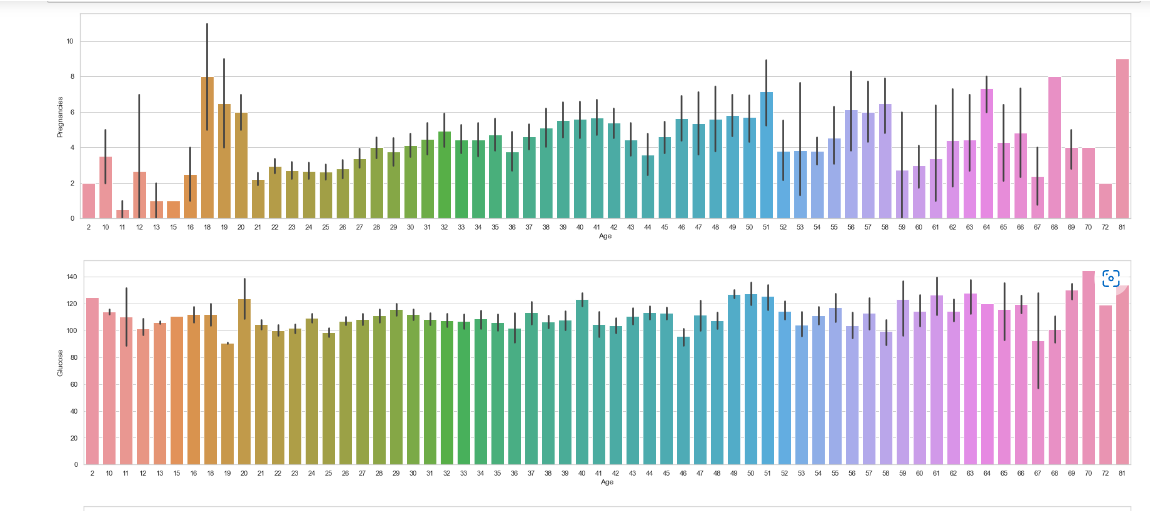


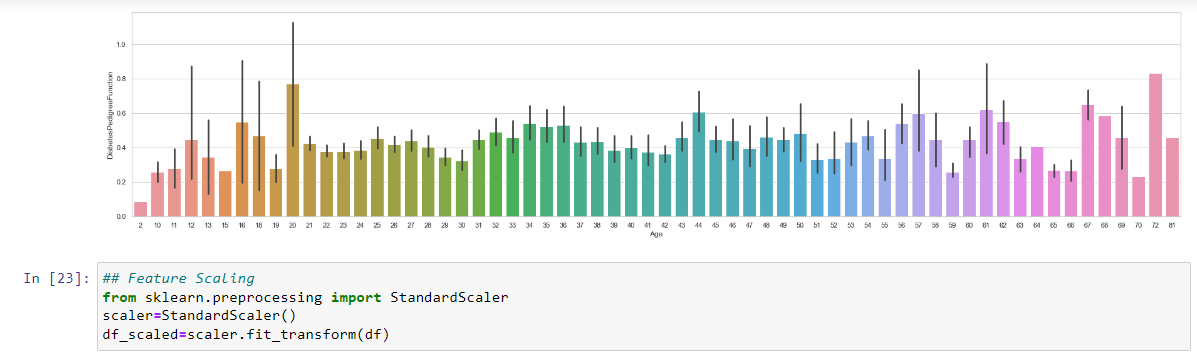
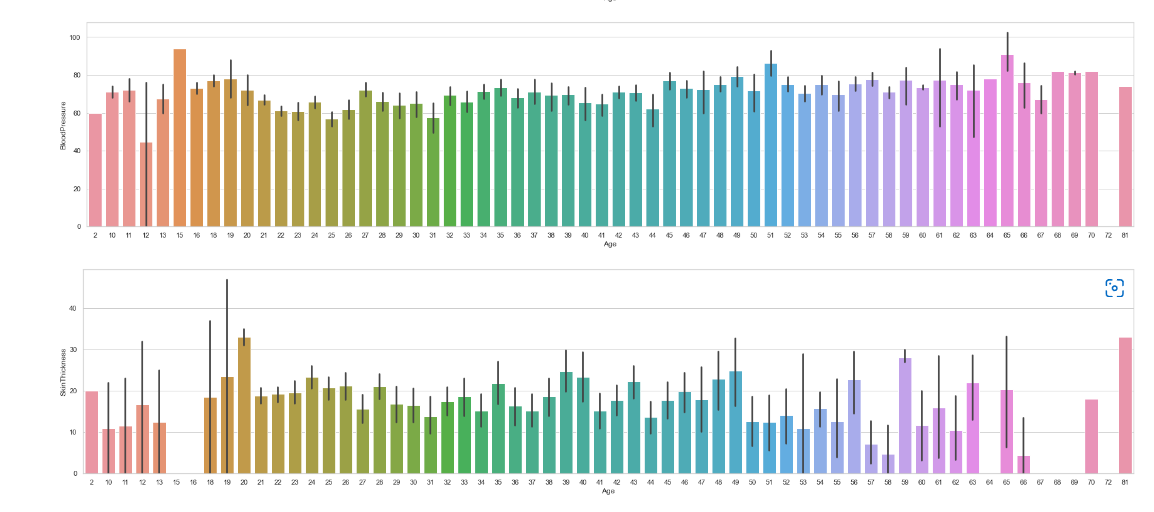
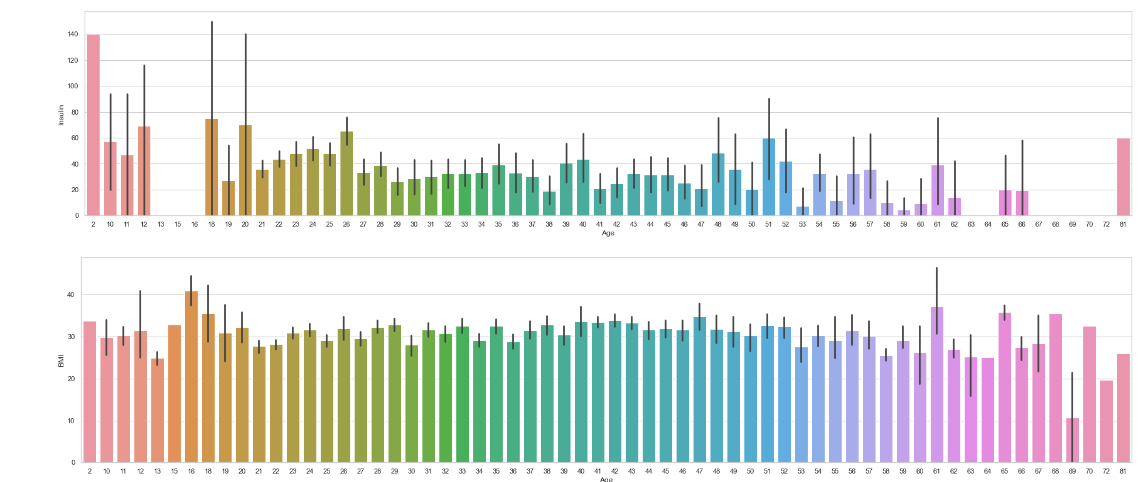


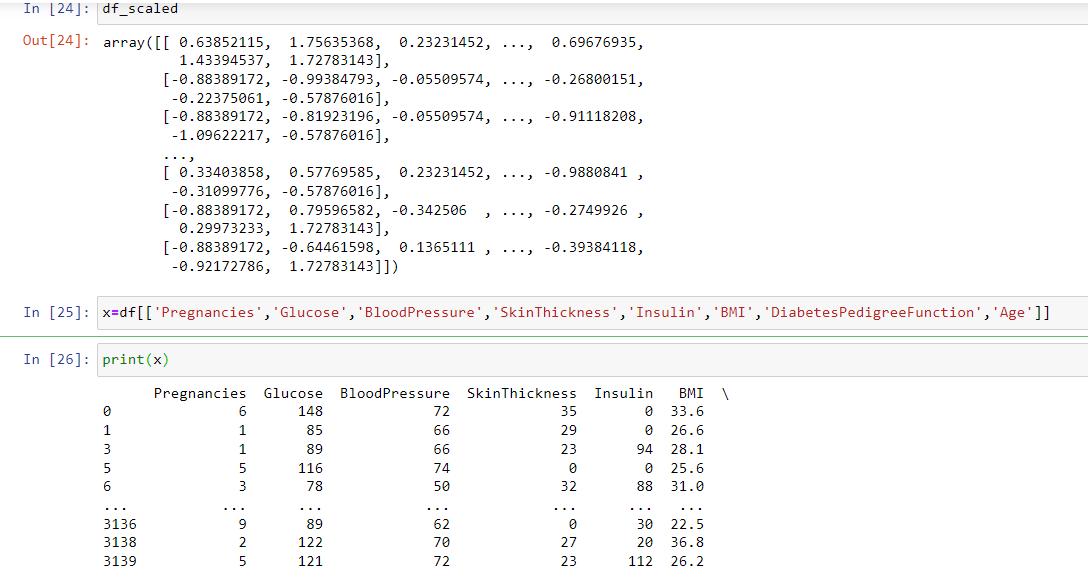


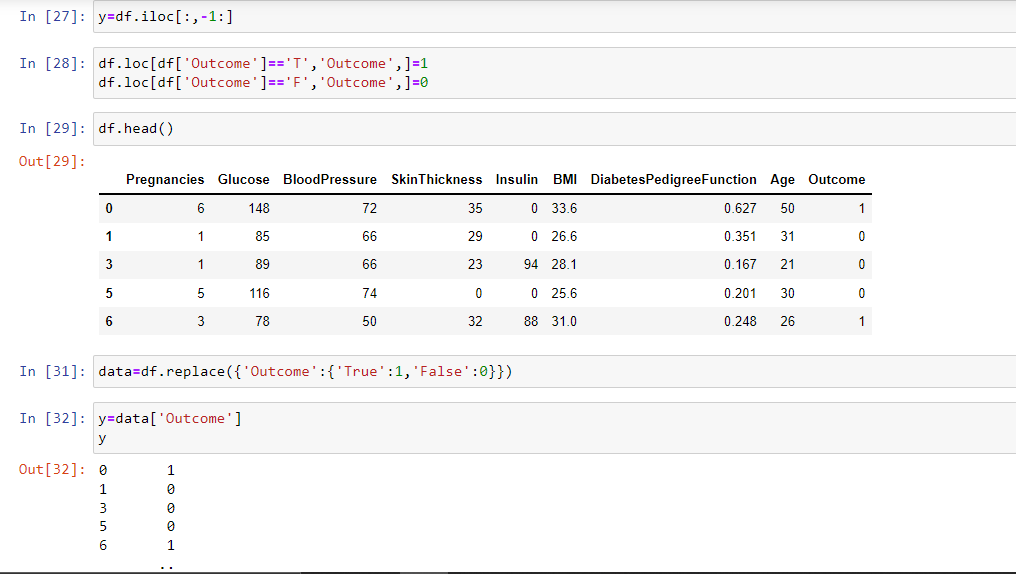


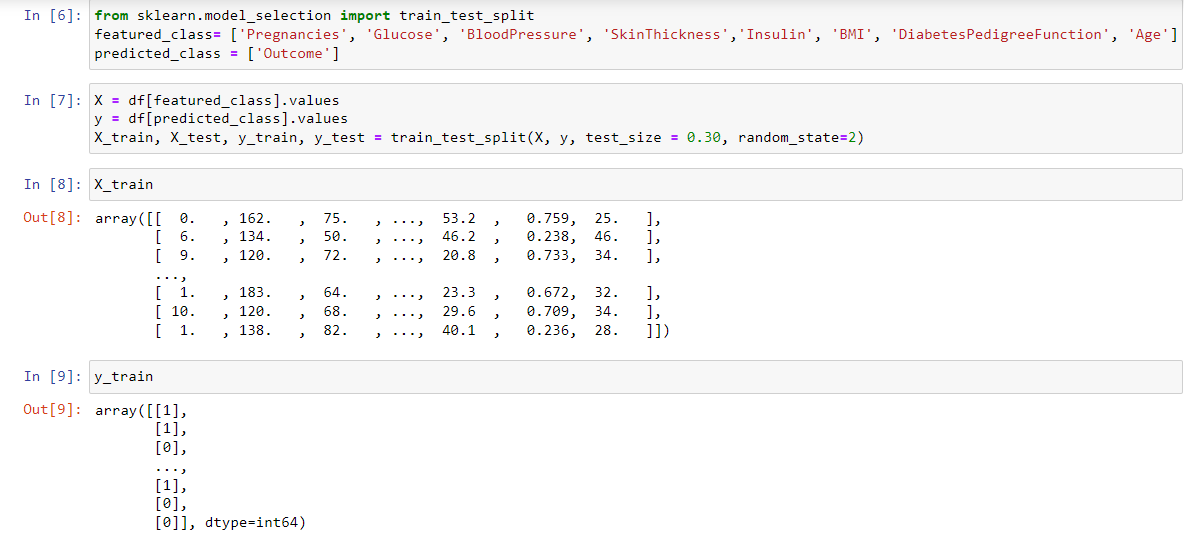


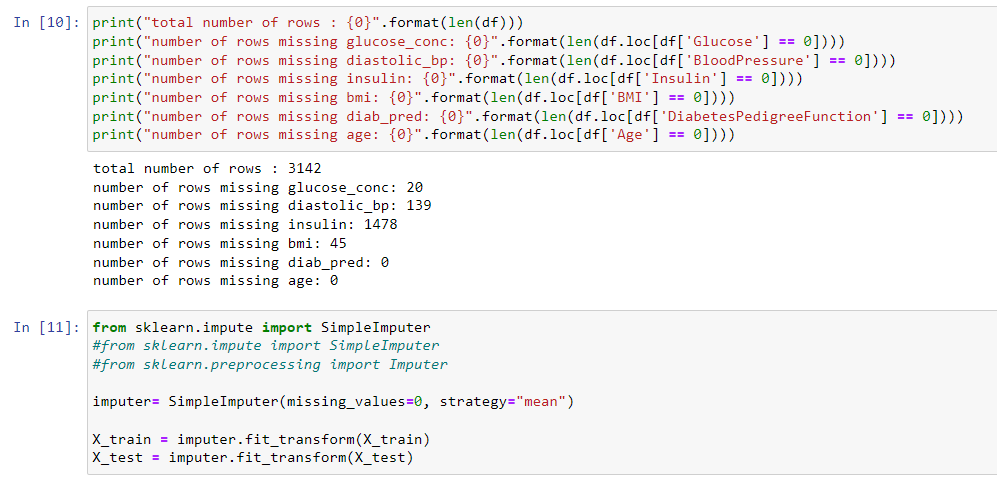


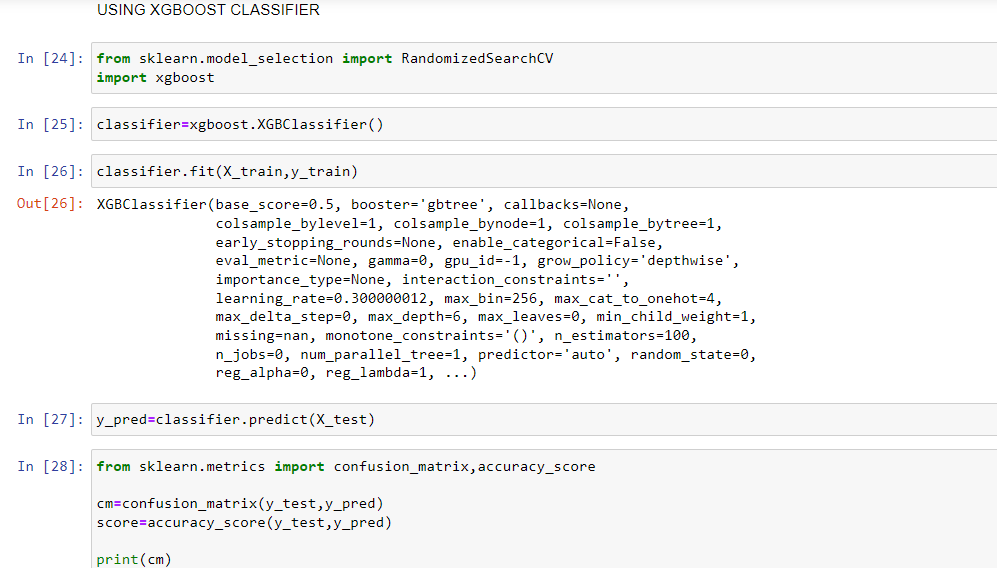










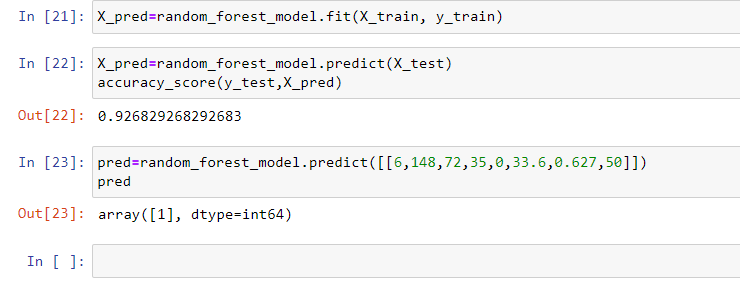


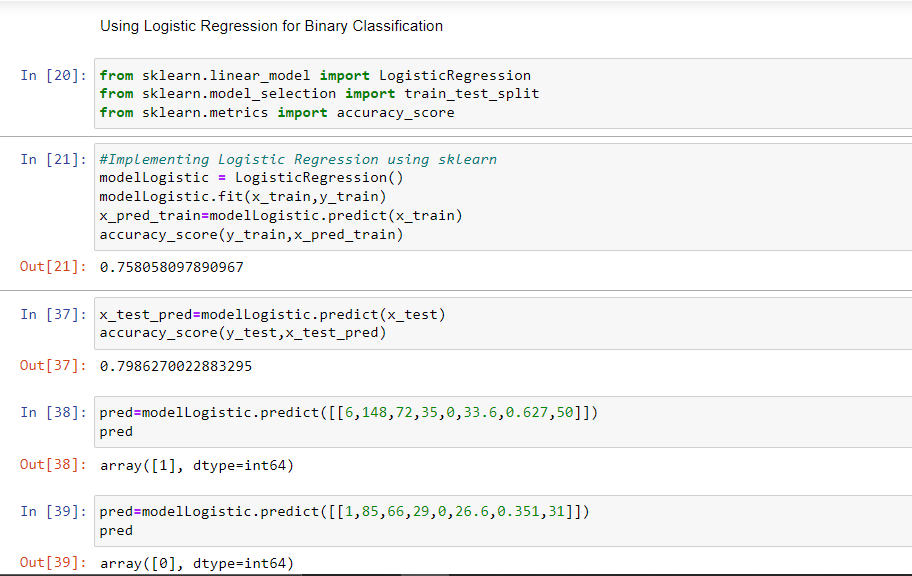


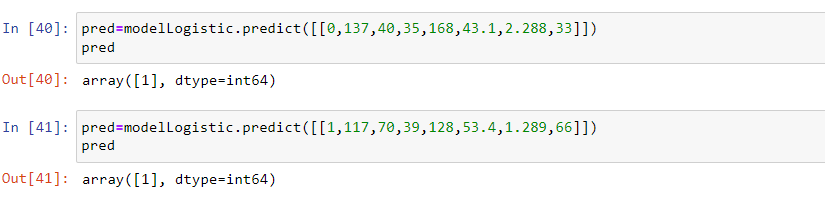
Task:-

BY USING OTHER CLASSIFICATIONS:-







 **B.YASASWINI**

**S180811**

**2C**

**S180811@rguktsklm.ac.in**