Title:

PIMA-INDIANS

Name-Ganesh N Anil   
[Email-ganesh.n.anil007@gmail.com](mailto:Email-ganesh.n.anil007@gmail.com)

Ph-7034007173

DIABETES PREDICTION MODEL

Predicting the onset of diabetes**.**

***Abstract***-The diabetes dataset is a binary classification problem where it needs to be analysed whether a patient is suffering from the disease or not on the basis of many available features in the dataset. Different methods and procedures of cleaning the data, feature extraction, feature engineering and algorithms to predict the onset of diabetes are used based for diagnostic measure on Pima Indians Diabetes Dataset.

# **Dataset details:**

The data sets comprise several variables of the medical predictor, and one objective variable. The forecasting variables include the patient’s number of pregnancies, BMI levels, insulin levels, age, etc.

1: Pregnancies: Number of times pregnant

2: Glucose: Plasma glucose concentration 2 hours in an oral glucose tolerance test.

3: Blood Pressure: Diastolic blood pressure (mm Hg)

4: Skin Thickness: Triceps skinfold thickness (mm)

5: Insulin: 2-Hour serum insulin (mu U/ml)

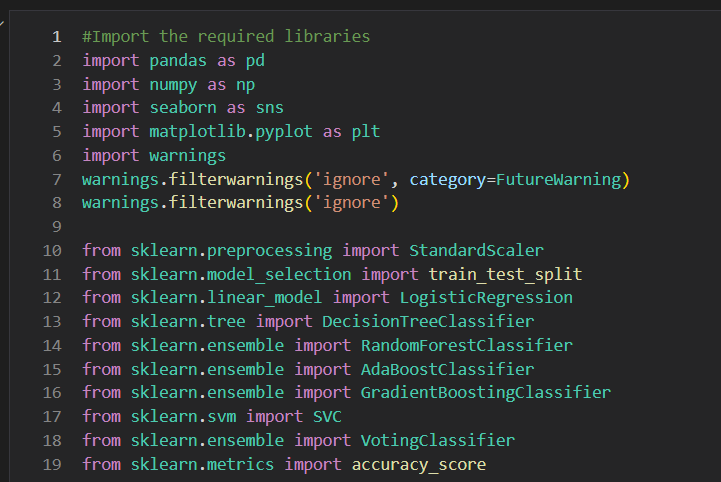
6: BMI: Body mass index (weight in kg/ (height in m²))

7: Diabetes Pedigree Function: Diabetes pedigree function

8: Age: Age (years)

9: Outcome: Class variable (0 or 1) 268 of 768 are 1, the others are 0

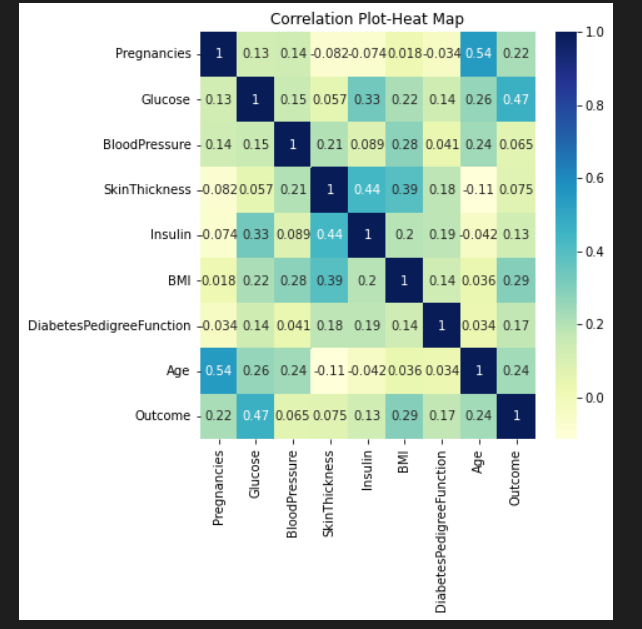
# **Importing the required libraries:**



**EDA (Exploratory Data Analysis):**

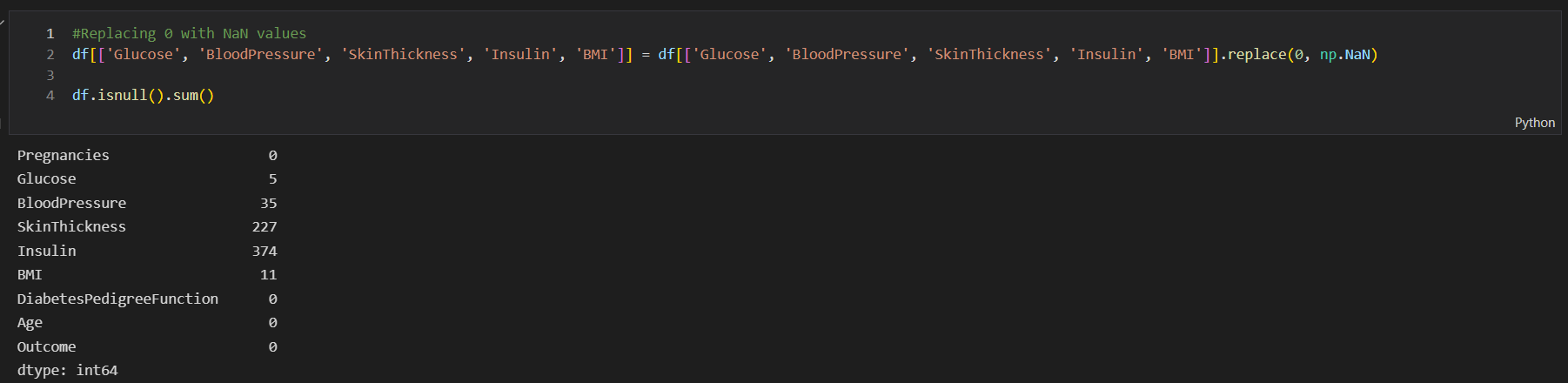
* Checking the unique values in target variable and plotting the target variable.
* Value count of output variable was found out and outcome is dataset is imbalance.
* Checking the basic info of the dataset such as shape, size, info etc.
* Finding out the missing values in each of the features.
* Plotting the heatmap and pairplot.

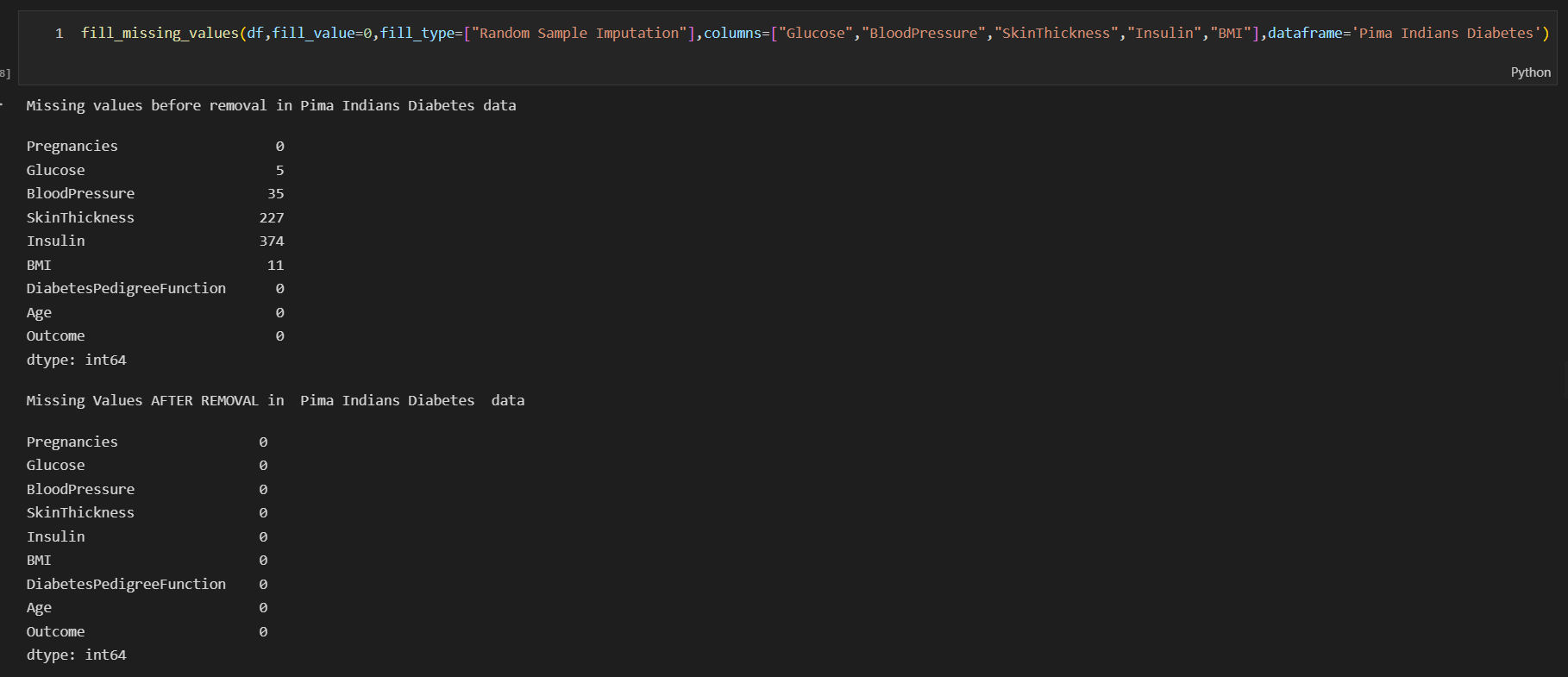




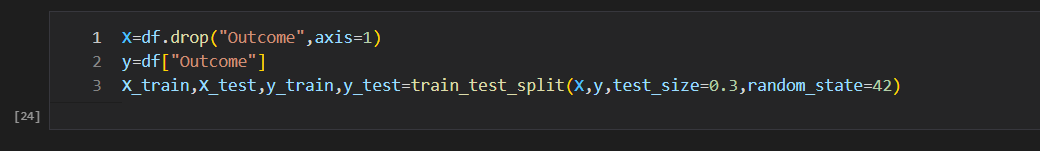
**Feature Engineering:**

* Handling the missing values using the function.
* Handling the outlier treatment using standard deviation method.





**Train Test Split:**

* Splitting into X\_train, X\_test, y\_train, y\_test.

**Scaling:**

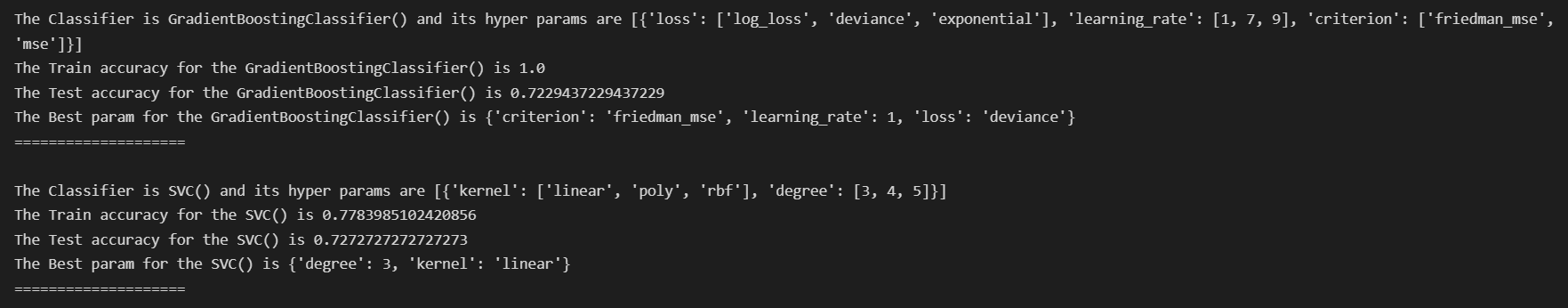
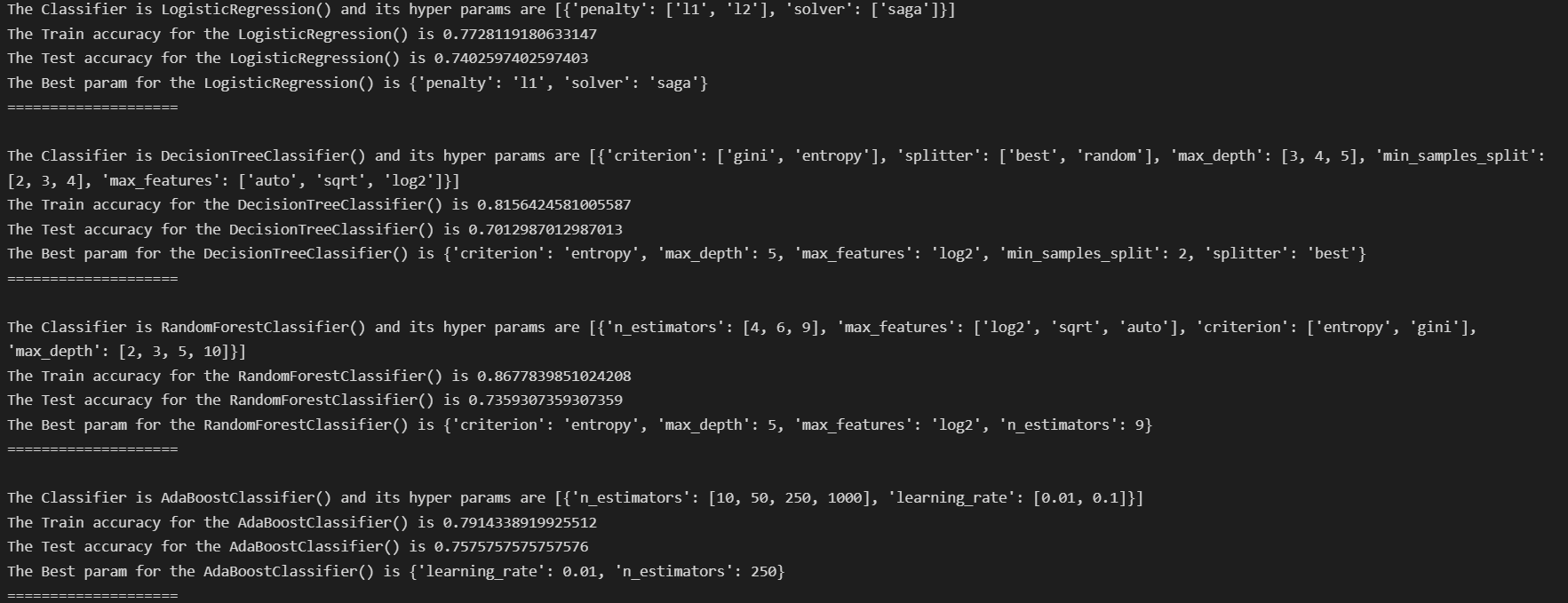
* Scale down using Standardscaler function.
* Train using fit\_transform and test using transform.
* Created pickle file using joblib.

**Data Modelling.**

**Applying All Algorithms:**

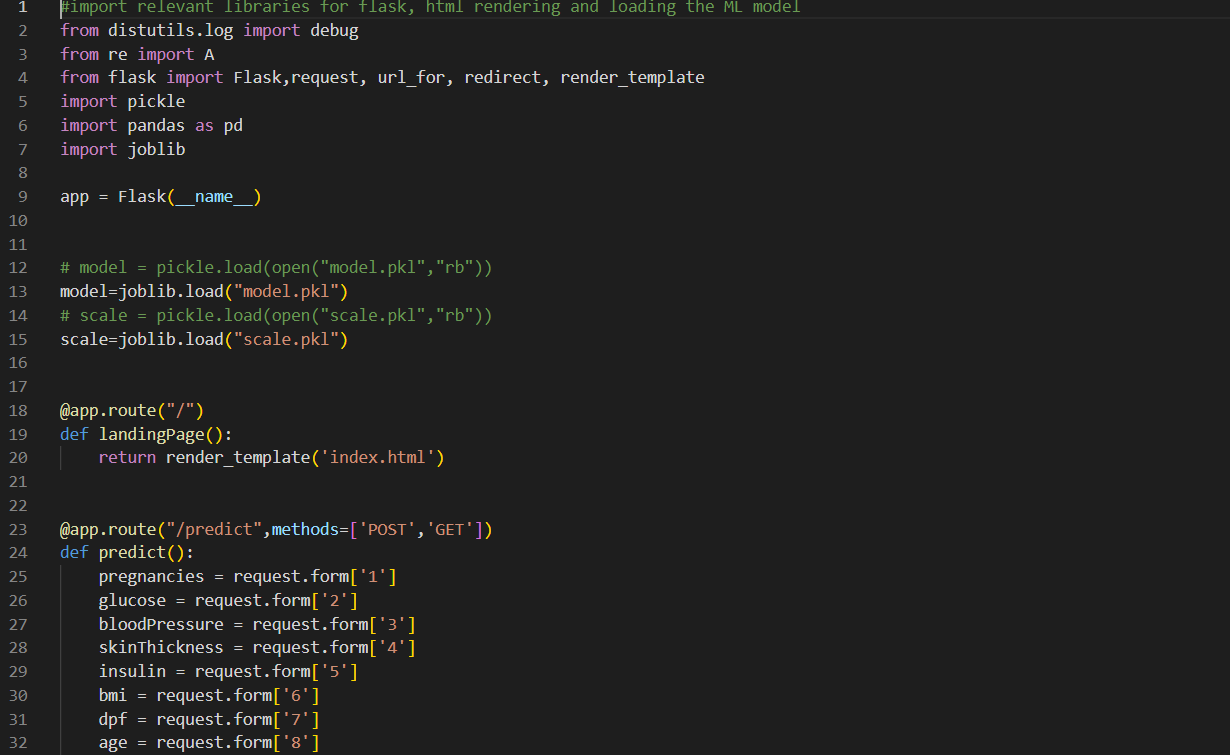
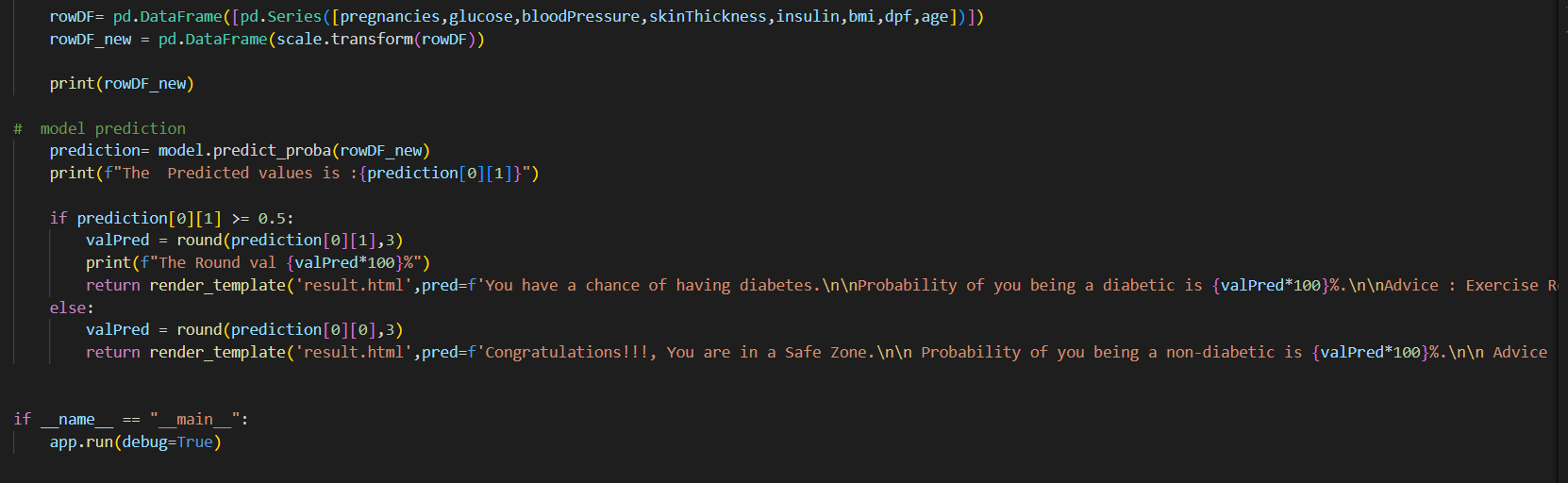
* Applying all classification algorithms such as logistic regression, decision tree classifier, random forest classifier, adaboost classifier, gradient boosting classifier and svc classifier.

**Hyperparameter Tuning:**

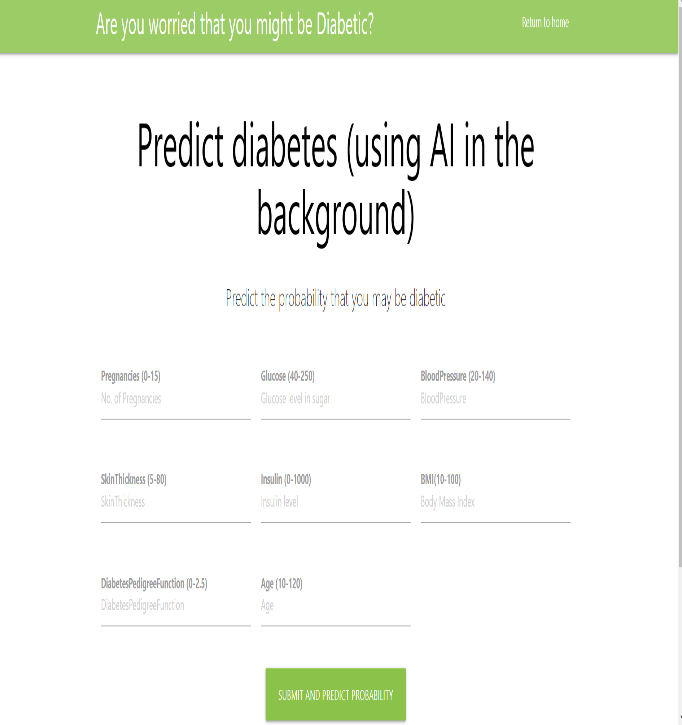
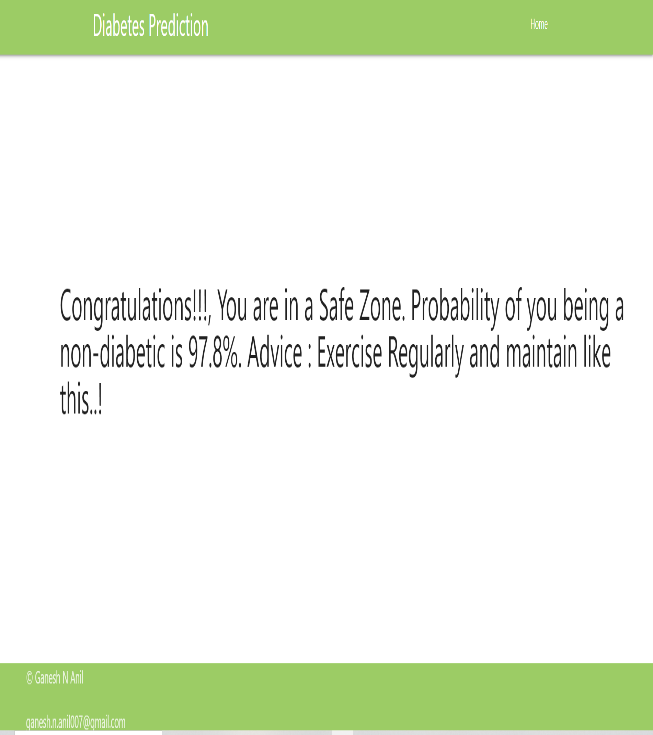
* Train accuracy and test accuracy for each classifier are found out using hyperparameter tuning.
* Best parameter for each classifier is found out.
* The best model was selected by comparing the train and test accuracy.
* ****In my case, I selected Random Forest Classifier as it is the best generalised model.
* Created pickle file using joblib.

**Web Framework.**

* Using the flask backend api was set
* UI (User Interface) was set using html code.



**User Interface.**

****

**Model Deployment.**

* I Have done the project in Google Collab and later pushed the file into my GitHub repository ( <https://github.com/GaneshNAnil/Pima_Indians_Diabetes.git>).
* I Have deployed the project through Heroku ([https://diabetispred.herokuapp.com/)](https://diabetispred.herokuapp.com/)%20)