FINAL REPORT

DIABETES PREDICTION

Problem statement

A hospital will conduct tests on various parameters of a person's health information which can then be used by an expert to make conclusion whether the person is having diabetes or not.

In some regions of the country, there is lack of specialised doctors present, and this might create difficulties for people of the region to avail the preferred service. While getting tests results of various parameters can be done through machines, can the process of concluding/predicting whether a person has diabetes or not can also be done through machines? Will it have high accuracy so that the predictions will be reliable and can be used without an expert?

Data Wrangling

The raw dataset consists of 2000 rows and 9 columns of data. The columns data consists of health-related parameters of a person like number of pregnancies occurred, glucose levels, blood pressure levels, skin thickness, BMI, DiabetesPedigreeFunction, age and outcome (whether the person has diabetes or not). The outcome column is the target function and it is of binary type. All other columns are of numerical type.

The dataset was fully filled with values and there were no null values present in the dataset. But as we explore the dataset, it has been found out that some columns have zero values filled for example the skin thickness value was zero, the BMI of a person was zero, the insulin level of a person was zero etc. Considering these columns, it is evident that such values could not be zero and there might be some data mishandling happened for such patients.

There were lot of such cases present in the dataset, so we cannot remove those rows as we don't have a large dataset. So, the values of such columns are replaced with the mean or median of the column.

Exploratory Data Analysis

Data distribution of columns present in the dataset are shown in below figures a, b and c.

Pregnancies column has a skewed dataset but there were not outliers present in the data.

Glucose dataset was approximately normally distributed and there was no outlier present in the data.

Blood pressure column has an approximately normal distribution graph with no outliers present in the data.

Skin thickness column is having an approximately normal distribution graph and has some outliers present in the data. But theses outlier values are possible in real cases and thus are not been removed.

BMI value column is having an approximately normal distribution and the data seems outlier is also possible in real scenario and thus it has also been not removed.

DPF or DiabetesPedigreeFunction is a skewed data distribution with no outliers in the data which needs to be removed.

Age column has a skewed data distribution with no outliers present in the data.

Outcome column in the target column which is of binary type data distribution.

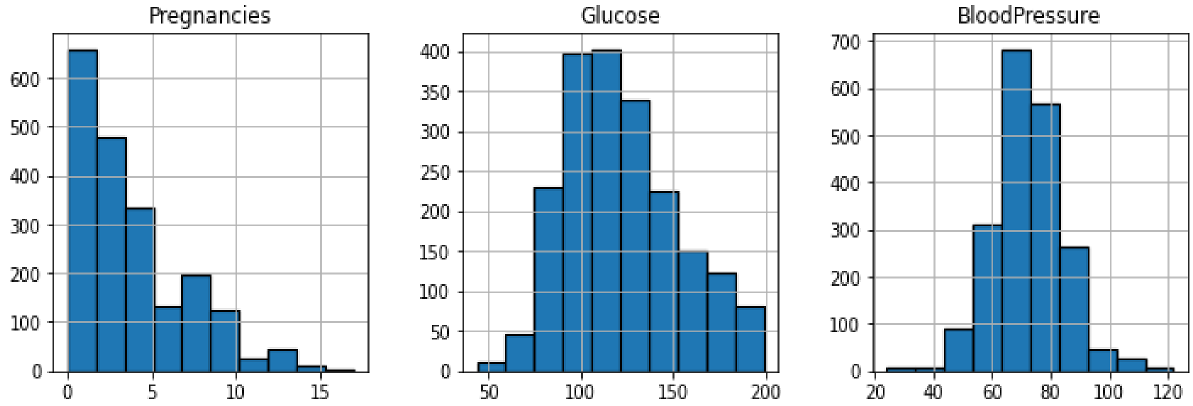


fig a

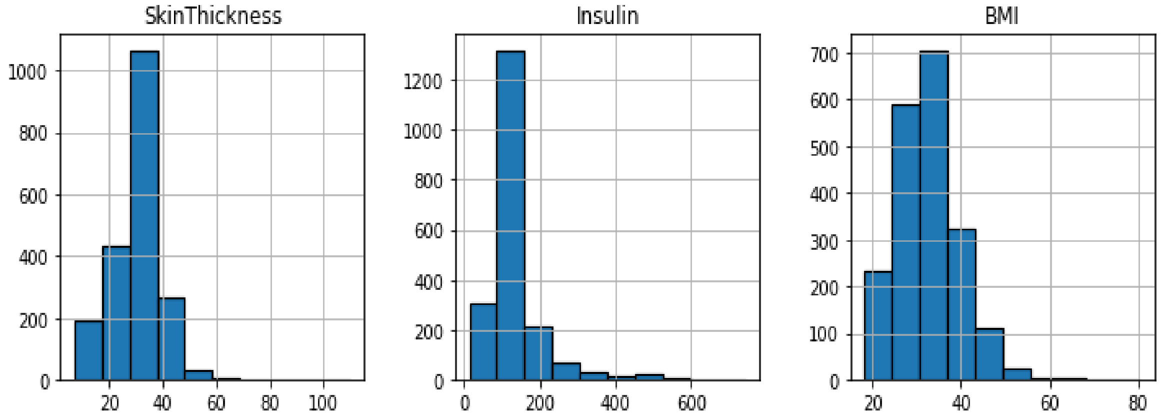


fig b

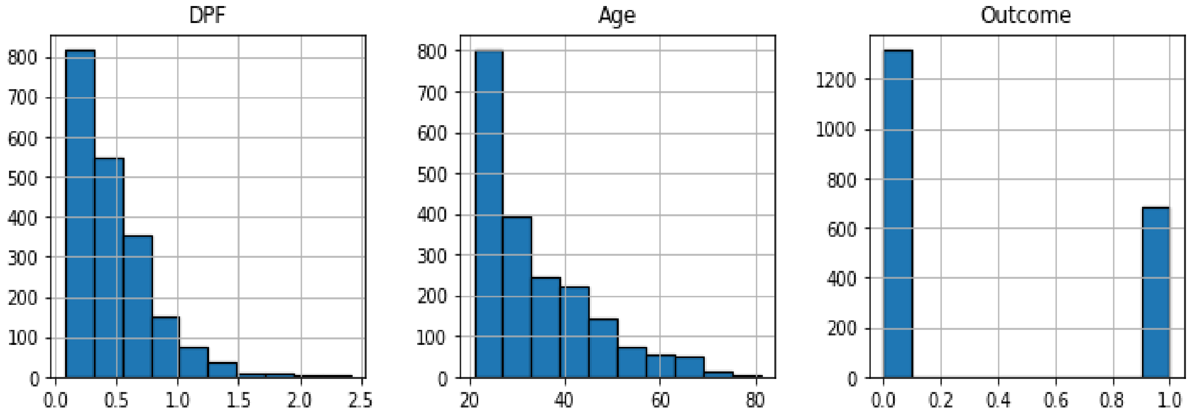
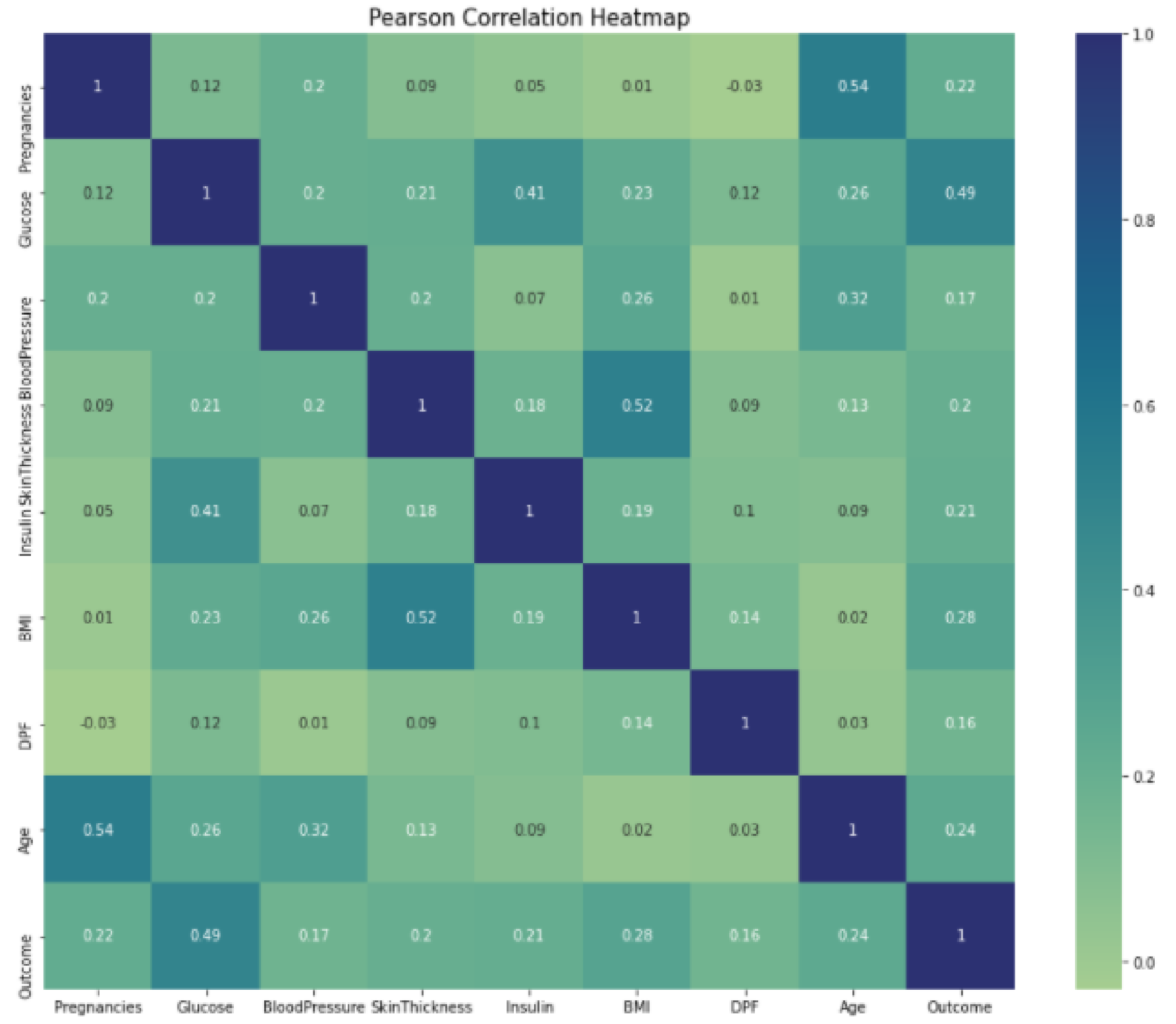


fig c

Pearson correlation heatmap between all the columns of the dataset is as shown below.

There is a good correlation between the glucose and outcome columns. Also, there is a fairly good correlation between the BMI and outcome columns.



Pre-processing and Training Data Development

All the independent columns of the dataset were of numerical type and thus were good to go.

The dataset was divided into training and test dataset in the ratio 75 is to 25.

Following are the split made in the dataset:

a) Training dataset with independent columns (75%)
b) Training dataset with dependent column (75%)
c) Test dataset with independent variables (25%)
d) Test dataset with dependent column (25%)

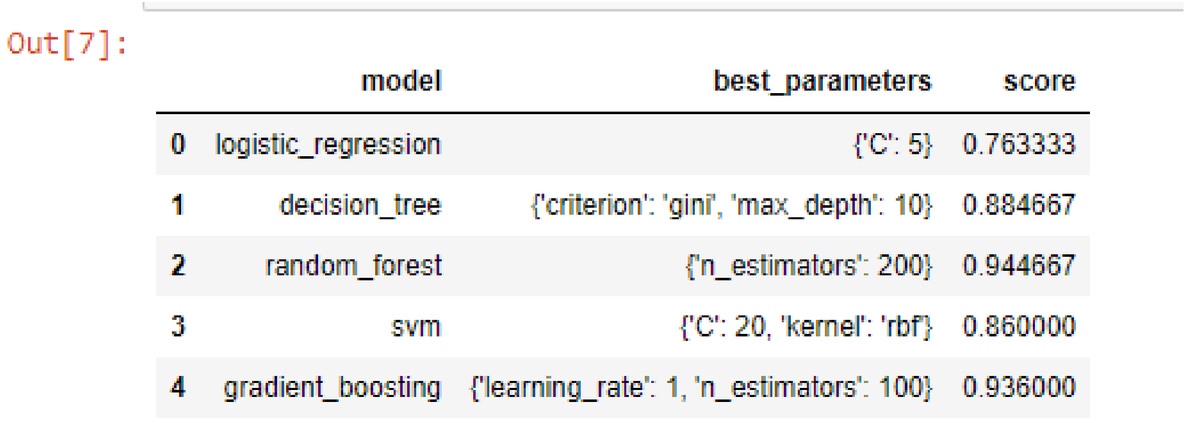
The training and test dataset of independent columns were standardized to remove any outlier cause and also to scale each feature/ variance to unit variance.

Model Evaluation and Optimization

In this unit the dataset has been trained on different models.

Here we have worked on applying five different models to train the model and test the accuracy of the models.

Below figure shows the models been used to train the model and the accuracy they have achieved. Few parameters tuning is done while training the models.

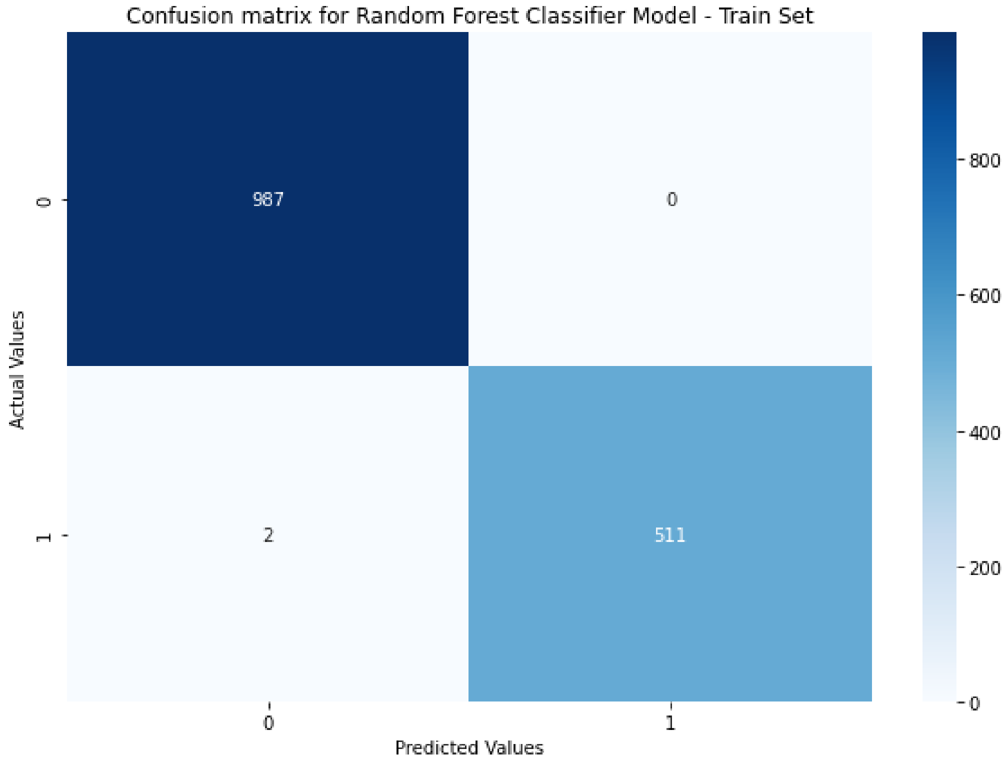


Out[7]:

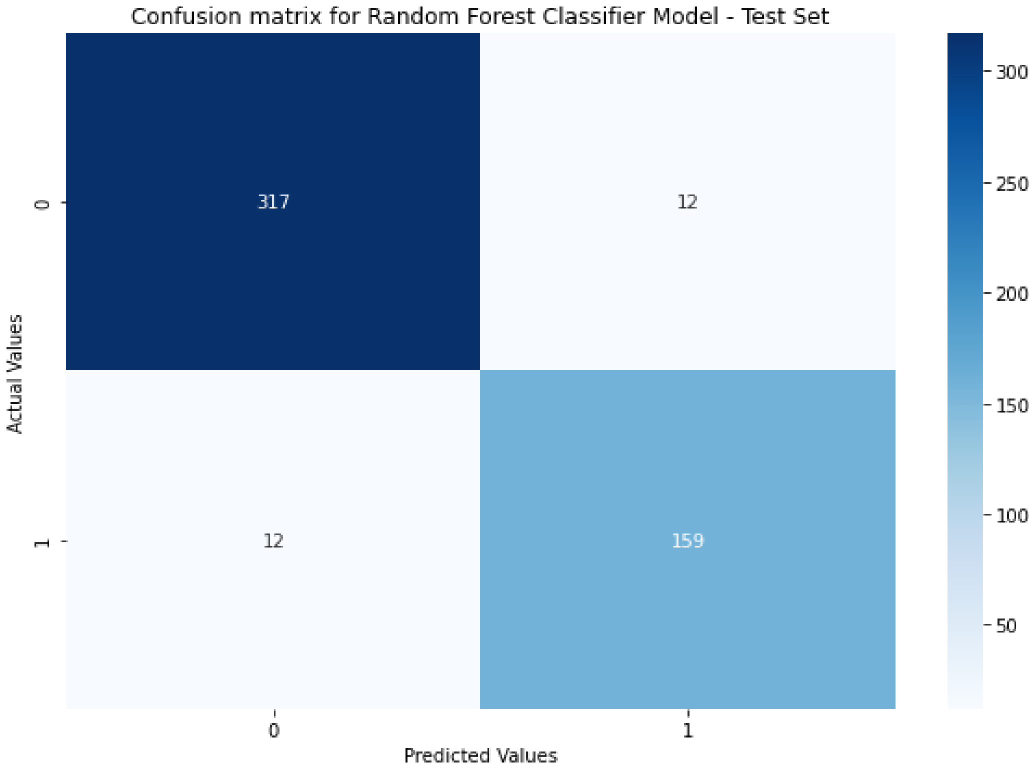
	model	best_parameters	score
0	logistic_regression	{'C': 5}	0.763333
1	decision_tree	{'criterion': 'gini', 'max_depth': 10}	0.884667
2	random_forest	{'n_estimators': 200}	0.944667
3	svm	{'C': 20, 'kernel': 'rbf'}	0.860000
4	gradient_boosting	{'learning_rate': 1, 'n_estimators': 100}	0.936000

Here it is seen that random forest has achieved the highest accuracy among all the models. So, we will perform more hyperparameter tuning in the random forest model for getting better accuracy results.

After tuning more through hyperparameter optimization, we get 99.8% accuracy on training dataset with the best parameters. Below is the confusion matrix for training dataset.



For test dataset we get an accuracy of 96% while fitting the model with the best parameters. Below shown the confusion matrix for the test dataset.



Further Research on the Problem

Here we have used the data to predict whether a person has diabetes or not.

Diabetes is also divided into two types. This project can be further expanded into the following:

1) Predicting Type I diabetes
2) Predicting Type II diabetes