KISHAN KUMAR

DIABETES PREDICTION

USING MACHINE LEARNING ALGORITHMS

USECASE

Diabetes is a chronic disease with the potential to cause a worldwide healthcare crisis. According to International Diabetes Federation, 382 million people are living with diabetes worldwide. By 2035, this will be doubled to 592 million. Diabetes mellitus or imply diabetes is a disease caused due to an increased level of blood glucose. Various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes. However, early prediction of diabetes is a quite challenging task for medical practitioners due to complex interdependence on various factors as diabetes affects human organs such as kidneys, eyes, heart, nerves, feet, etc. Data science methods have the potential to benefit other scientific fields by shedding new light on common questions. One such task is to help make predictions on medical data. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. The aim of this project is to develop a system that can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques. This project aims to predict diabetes via three supervised machine learning methods including SVM, Logistic regression, and KNN. This project also aims to propose an effective technique for earlier detection of diabetes disease using Machine learning.

DATASET

The dataset collected is originally from the Pima Indians Diabetes Database and is available on Kaggle. It consists of several medical analyst variables and one target variable. The objective of the dataset is to predict whether the patient has diabetes or not. The dataset consists of several independent variables and one dependent variable, i.e., the outcome. Independent variables include the number of pregnancies the patient has had their BMI, insulin level, age, and so on as Shown in the Following Table :

Serial no	Attribute Names	Description
1	Pregnancies	Number of times pregnant
2	Glucose	Plasma glucose concentration
3	Blood Pressure	Diastolic blood pressure
4	Skin Thickness	Triceps skin fold thickness (mm)
5	Insulin	2-h serum insulin
6	BMI	Body mass index
7	Diabetes pedigree function	Diabetes pedigree function
8	Outcome	Class variable (0 or 1)
9	Age	Age of patient

EDA

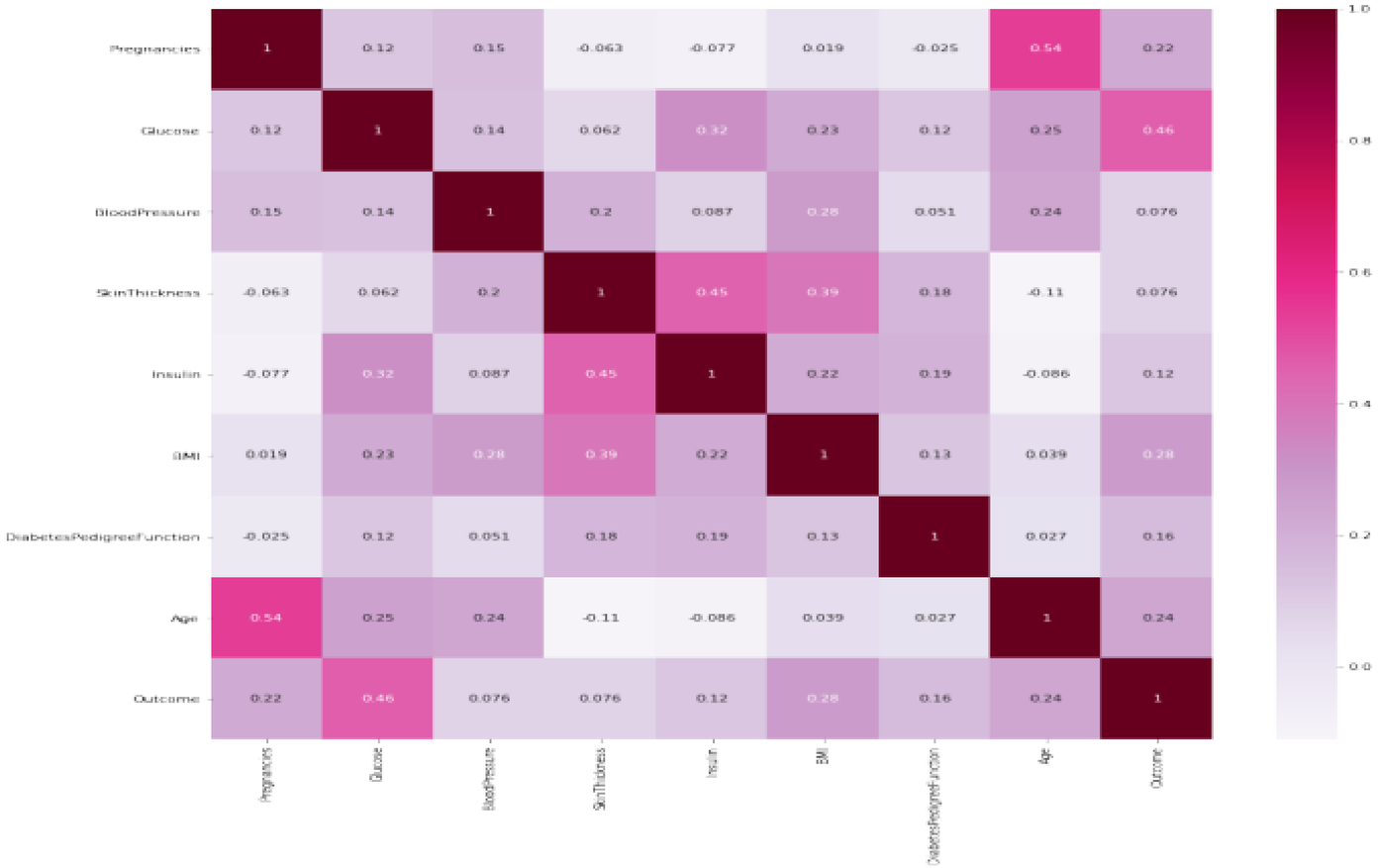
- The diabetes data set consists of 2000 data points, with 9 features each.
- “Outcome” is the feature we are going to predict, 0 means No diabetes, and 1 means diabetes.
- There are no null values in the dataset.

```
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               2000 non-null   int64  
 1   Glucose                   2000 non-null   int64  
 2   BloodPressure             2000 non-null   int64  
 3   SkinThickness             2000 non-null   int64  
 4   Insulin                   2000 non-null   int64  
 5   BMI                       2000 non-null   float64
 6   DiabetesPedigreeFunction  2000 non-null   float64
 7   Age                       2000 non-null   int64  
 8   Outcome                   2000 non-null   int64  
dtypes: float64(2), int64(7)
memory usage: 140.8 KB
```

EDA

- **CORRELATION MATRIX :**

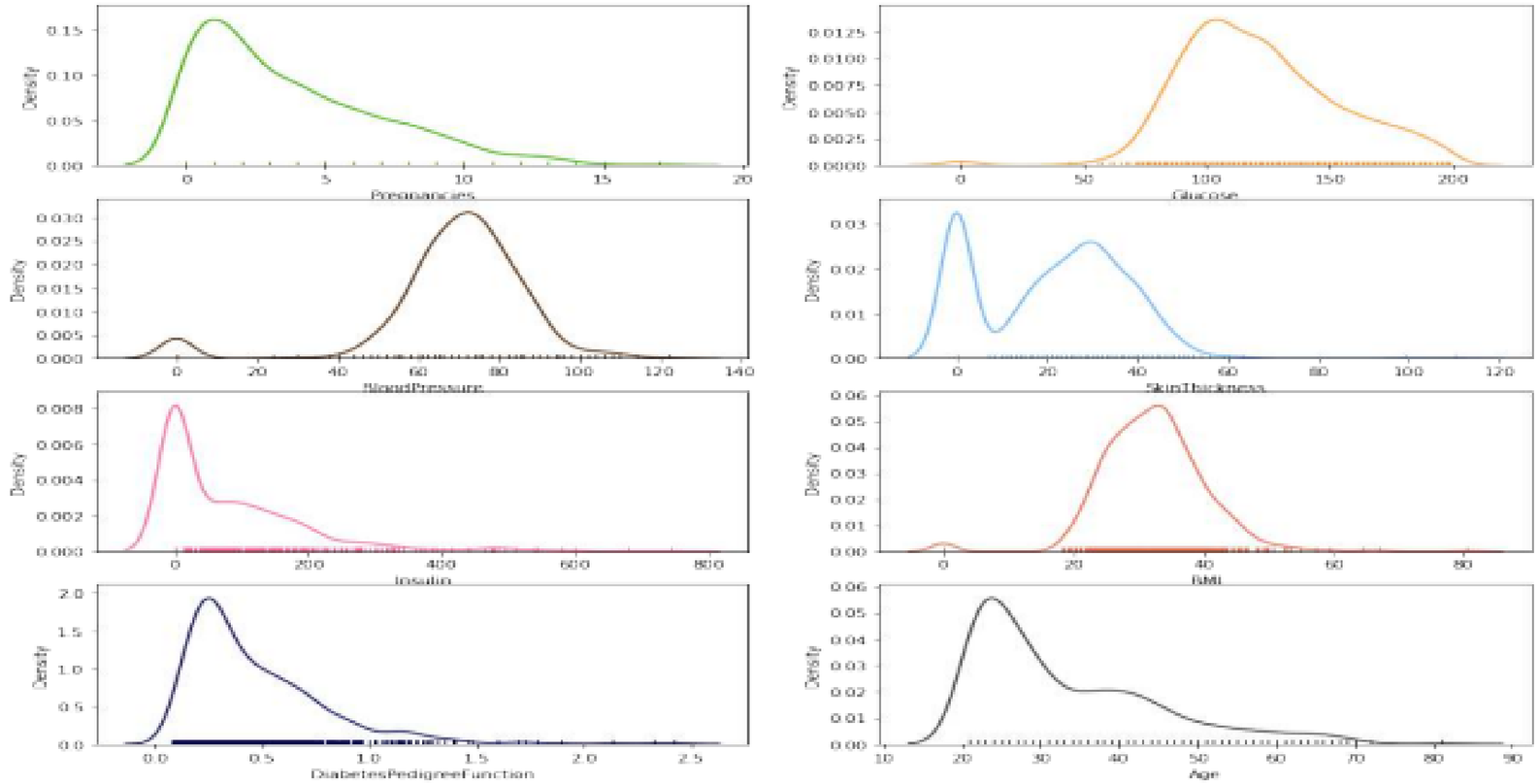
It is easy to see that there is no single feature that has a very high correlation with our outcome value. Some of the features have a negative correlation with the outcome value and some have positive.



EDA

- **SKEW OF DATA :**

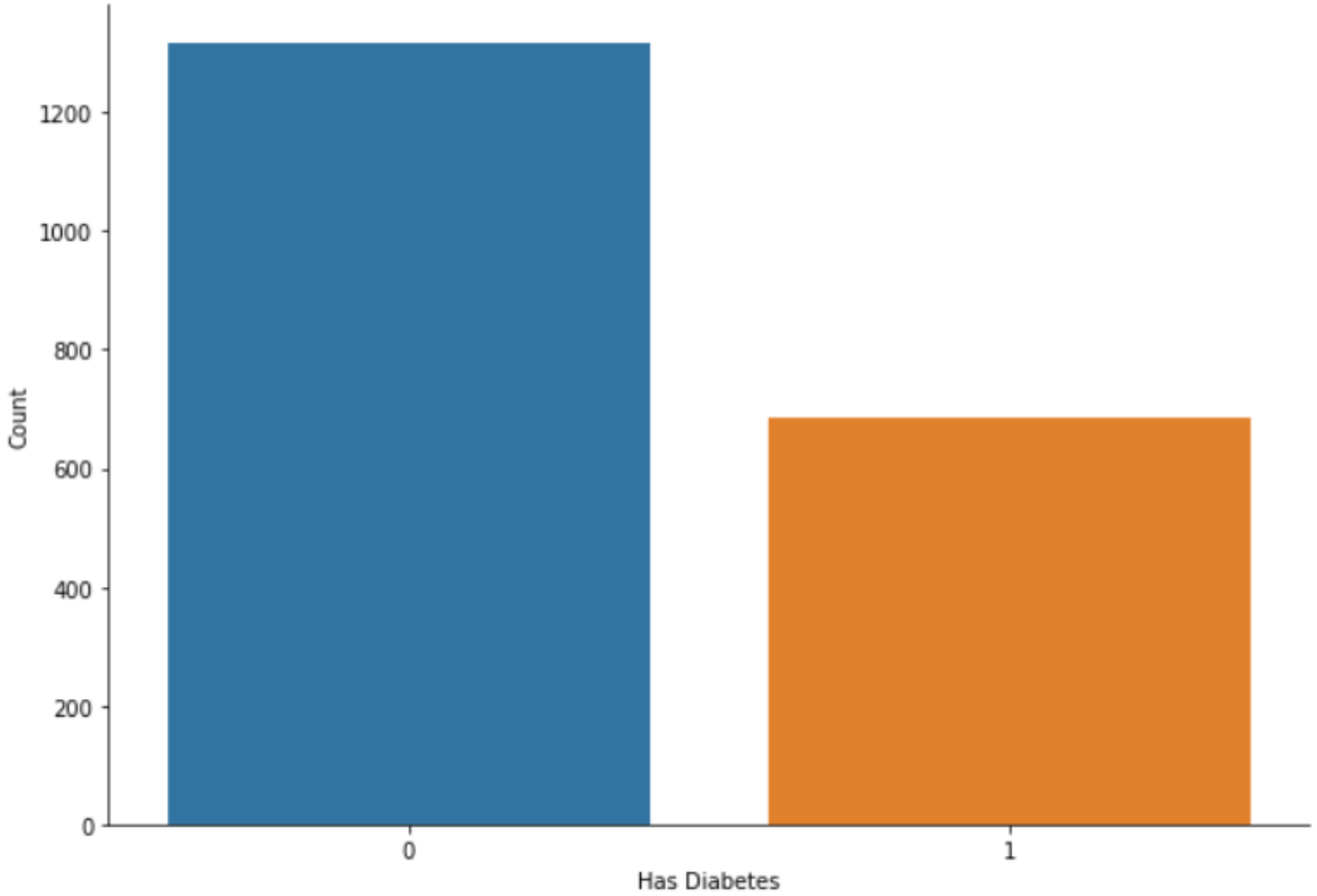
It shows how each feature and label is distributed along different ranges, which further confirms the need for scaling. It basically means that each of these is actually a categorical variable. We will need to handle these categorical variables before applying Machine Learning. Our outcome labels have two classes, 0 for no disease and 1 for disease.



EDA

- **BAR PLOT FOR OUTCOME CLASS :**

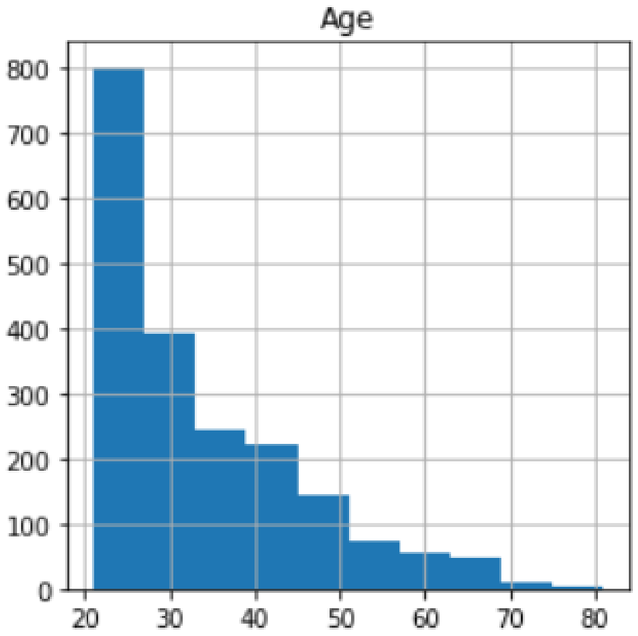
The below graph shows that the data is biased toward data points having outcome values of 0 which means that diabetes was not present actually. The number of non-diabetics is almost twice the number of diabetic patients.

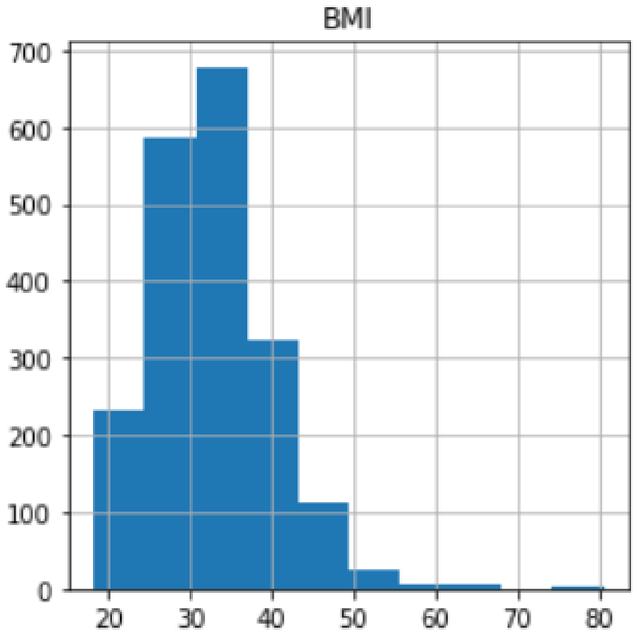


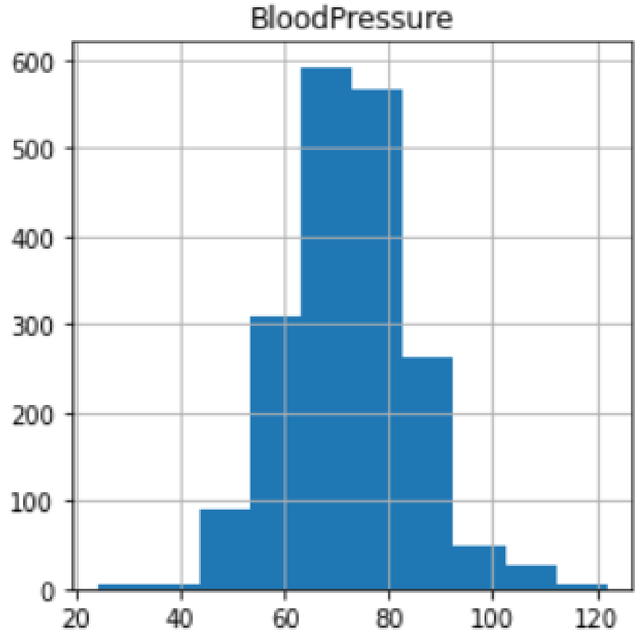
EDA

- **DATA CLEANING :**

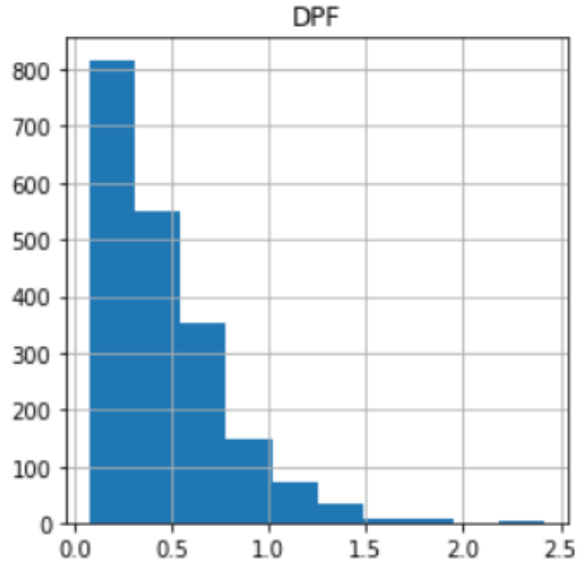
Replacing the 0 values from['Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI'] by NaN. To fill these Nan values the data distribution needs to be understood.

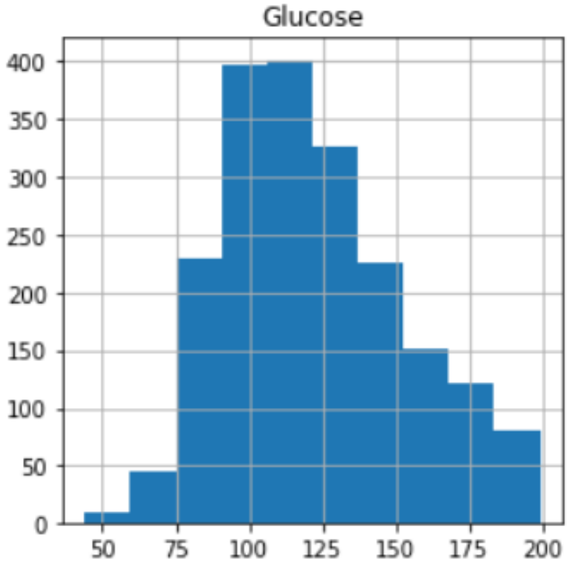


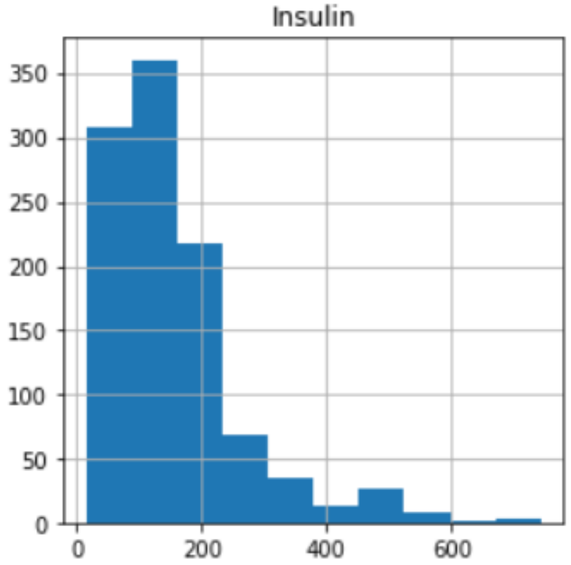


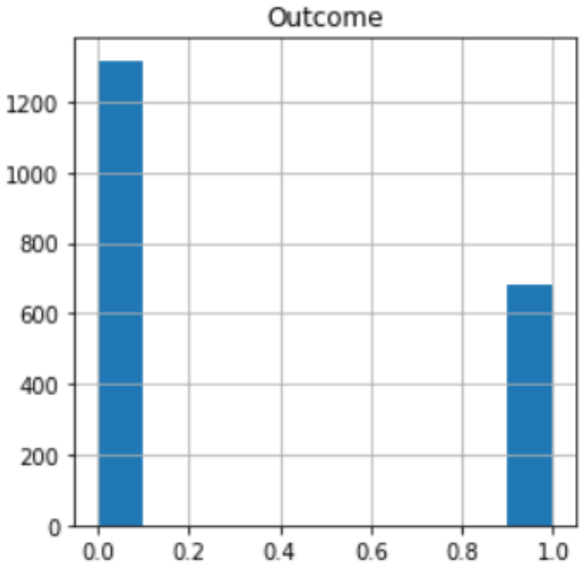


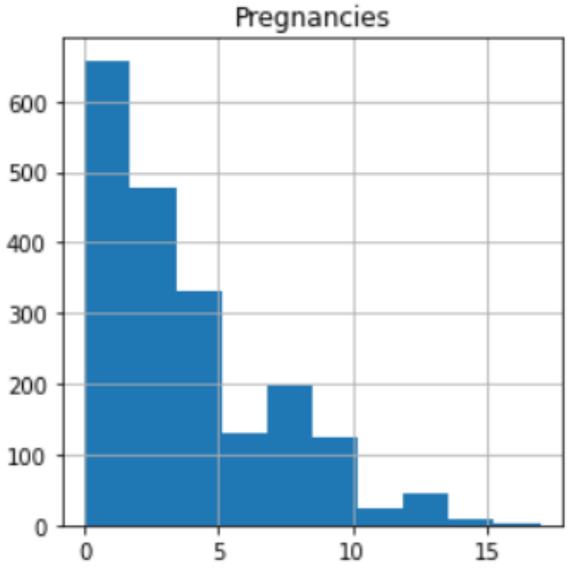
EDA

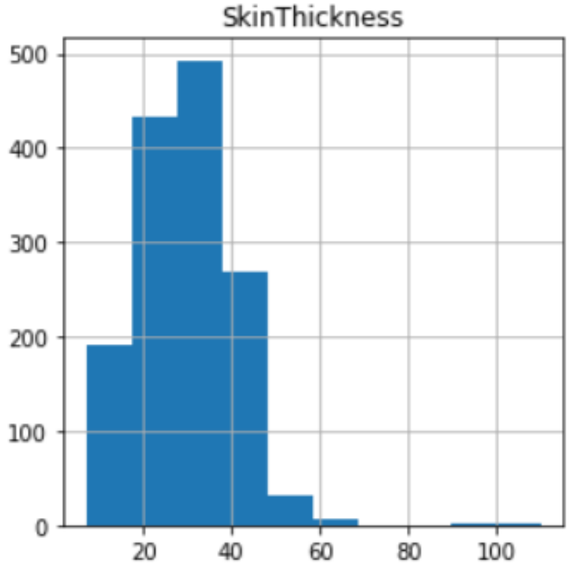

DPF
800
700
600
500
400
300
200
100
0
0.0
0.5
1.0
1.5
2.0
2.5


Glucose
400
350
300
250
200
150
100
50
0
50
75
100
125
150
175
200


Insulin
350
300
250
200
150
100
50
0
0
200
400
600


Outcome
1200
1000
800
600
400
200
0
0.0
0.2
0.4
0.6
0.8
1.0

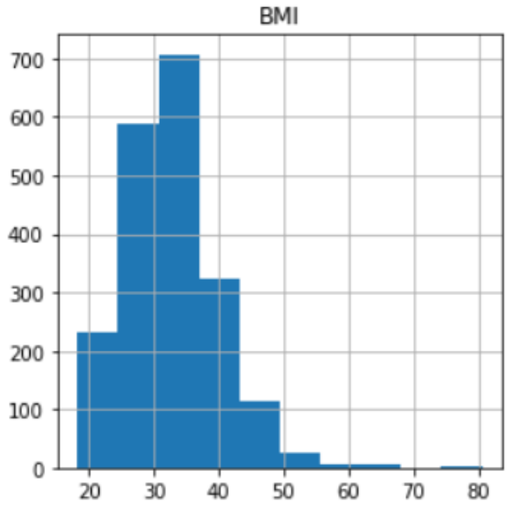

Pregnancies
600
500
400
300
200
100
0
0
5
10
15

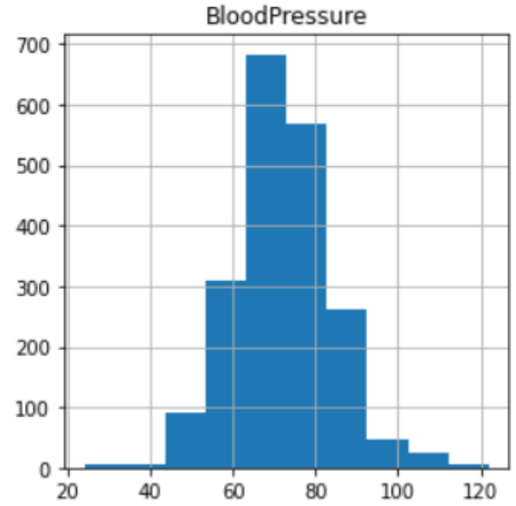

SkinThickness
500
400
300
200
100
0
20
40
60
80
100

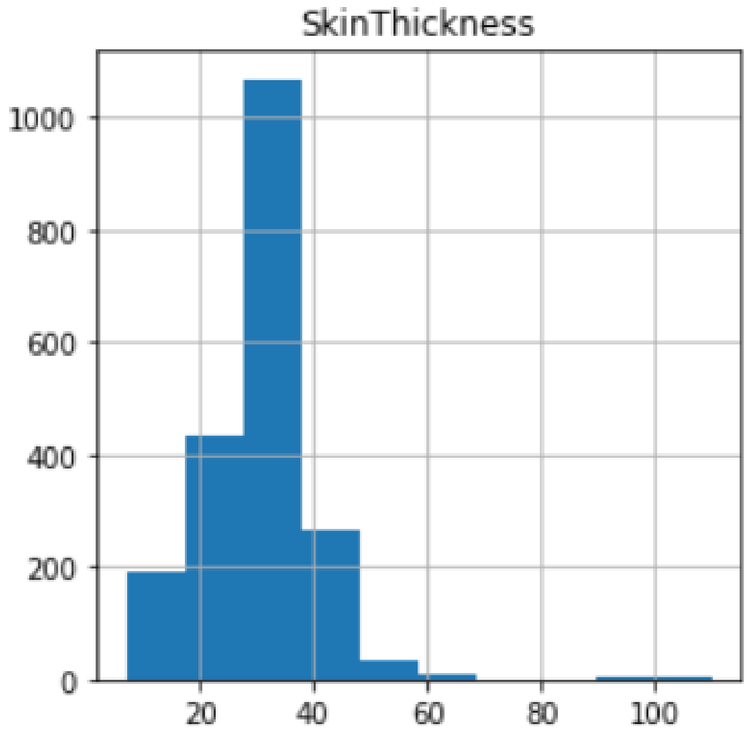
EDA

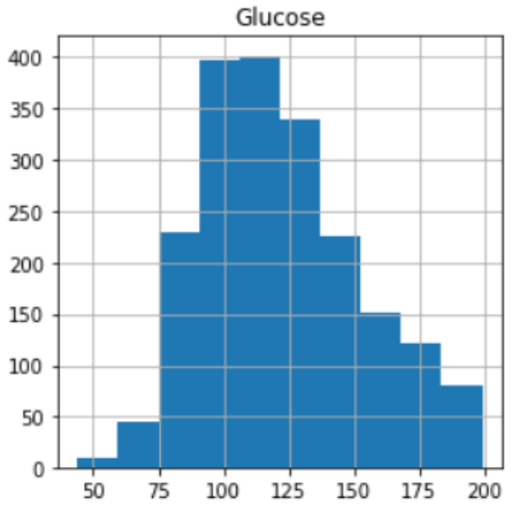
- **DATA CLEANING :**

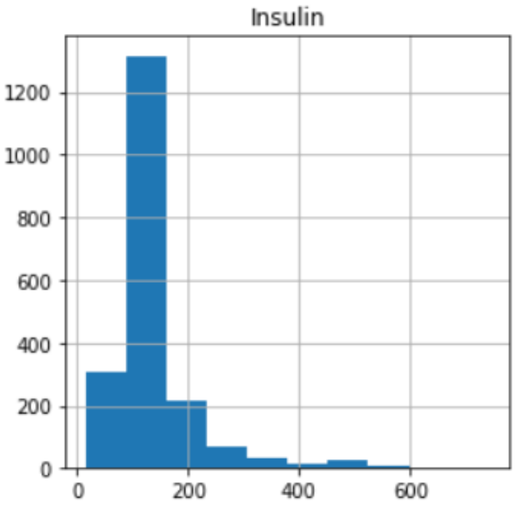
  Replacing NaN value by mean, and median depending upon the distribution











MODELLING

I've considered the following algorithms for building the model:

- **LOGISTIC REGRESSION :**

Logistic regression is a machine learning technique used when dependent variables are able to categorize. The outputs obtained by using the logistic regression is based on the available features. Here the sigmoidal function is used to categorize the output.

- **DECISION TREE :**

Decision tree is a nonparametric classifier in supervised learning. In this method, all the details are represented in the form of a tree, where leaves correspond to the class labels and attributes are corresponds to the internal node of the tree. We have used Gini Index for splitting the nodes.

MODELLING

- **RANDOM FOREST :**

it is an ensemble learning method for classification. This algorithm consists of trees and the number of tree structures present in the data is used to predict the accuracy. Where leaves correspond to the class labels and attributes are corresponds to the internal node of the tree. Here the number of trees in the forest used is 100 in number and the Gini index is used for splitting the nodes.

- **SVM :**

SVM is a supervised learning algorithm used for classification. In SVM we have to identify the right hyperplane to classify the data correctly. In this, we have to set correct parameter values. To find the right hyperplane we have to find the right margin for this we have to choose the gamma value as 0.0001 and rbf kernel. If we select the hyperplane with a low margin leads to miss classification.

MODELLING

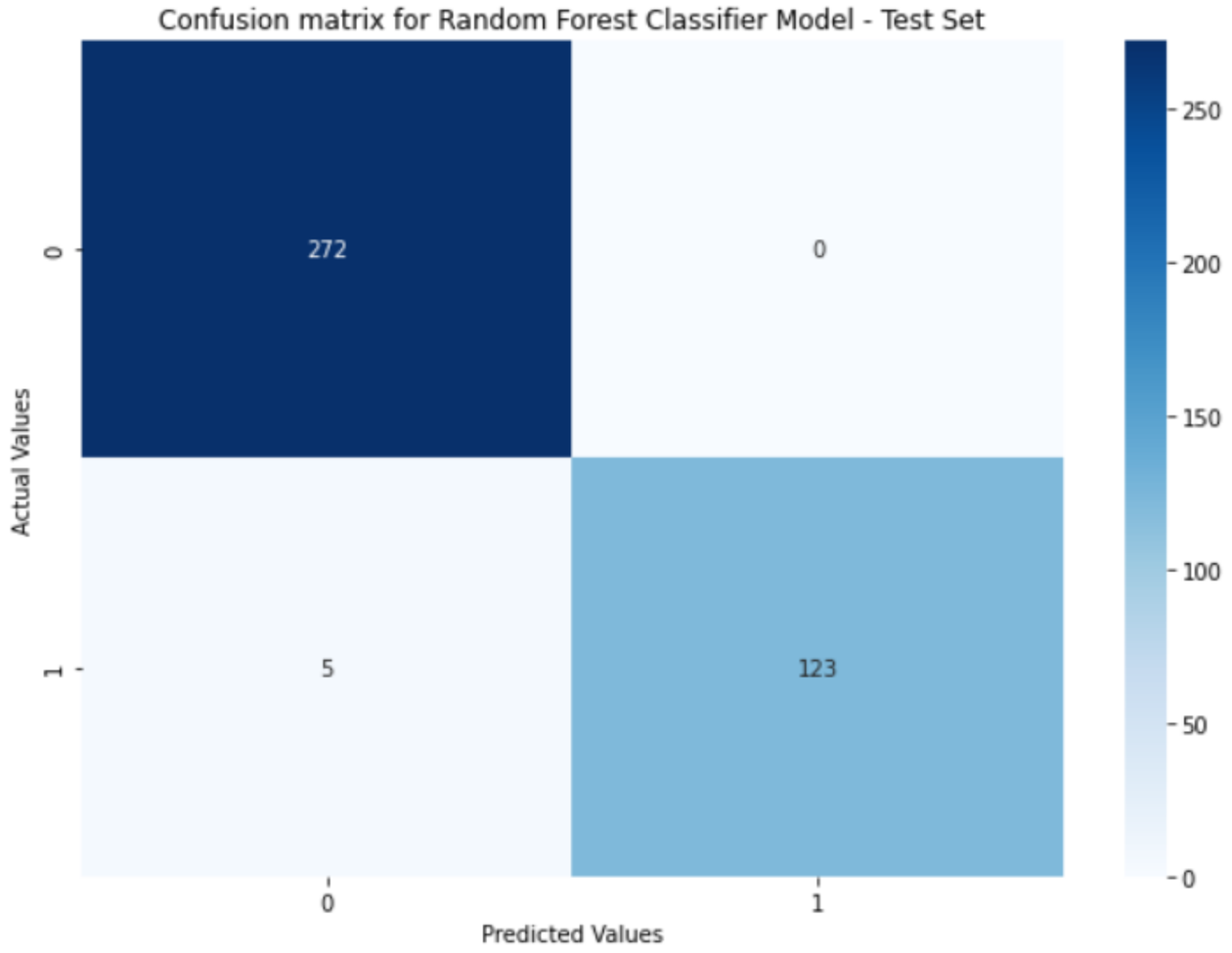
I am using **GridSearchCV** to find the best algorithm for this problem.

	model	best_parameters	score
0	logistic_regression	{'C': 10}	0.763125
1	decision_tree	{'criterion': 'entropy', 'max_depth': 10}	0.896250
2	random_forest	{'n_estimators': 100}	0.948125
3	svm	{'C': 20, 'kernel': 'rbf'}	0.869375

Since the Random Forest algorithm has the highest accuracy, we further fine-tune the model using hyperparameter optimization.

MODELLING

- **CONFUSION MATRIX :**



- **Confusion matrix** provides an output matrix with a complete description performance of the model.
- **Accuracy** is the ratio of the number of correct predictions to the total number of predictions Made.

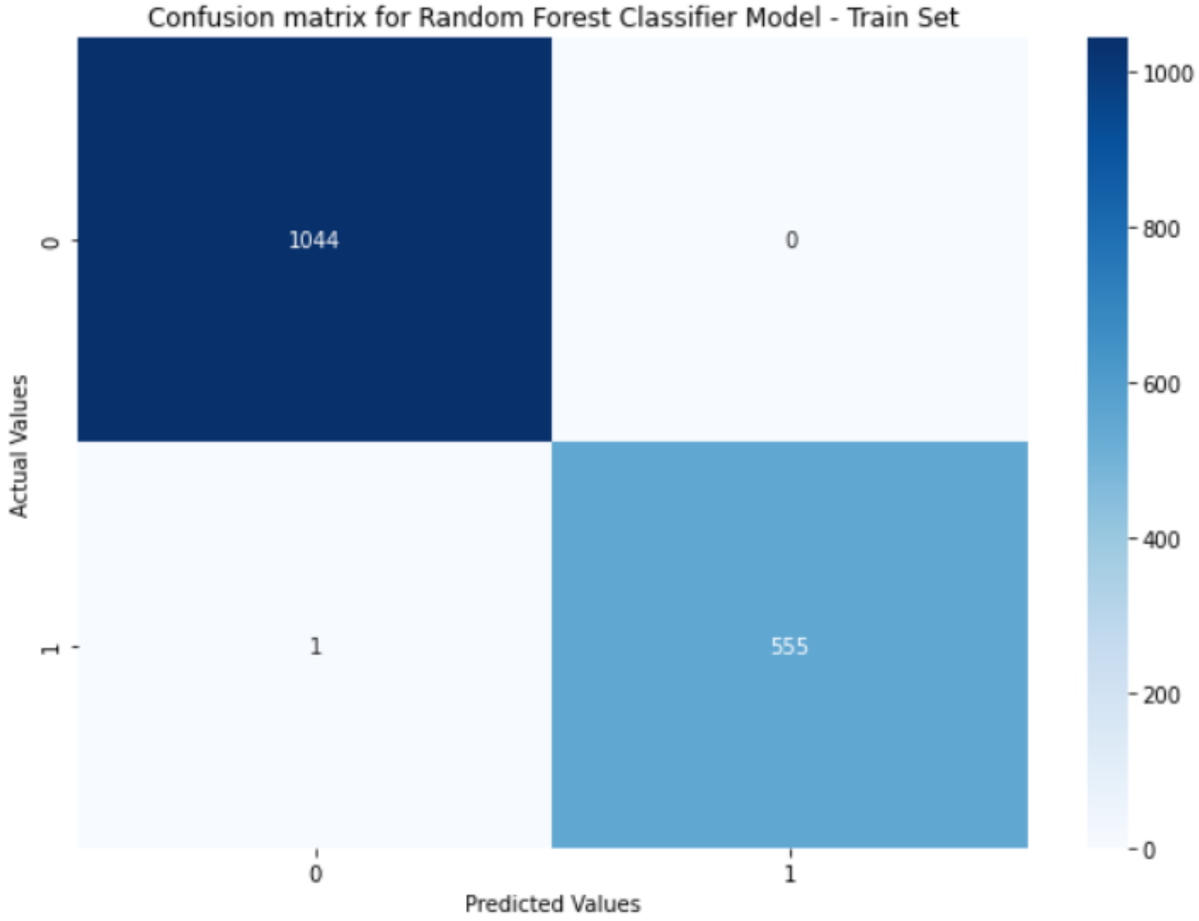
```
Accuracy on test set: 98.75%
```

MODELLING

- **CLASSIFICATION REPORT :**

	precision	recall	f1-score	support
0	0.98	1.00	0.99	272
1	1.00	0.96	0.98	128
accuracy			0.99	400
macro avg	0.99	0.98	0.99	400
weighted avg	0.99	0.99	0.99	400

MODELLING



Accuracy on trainning set: 99.94%

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1044
1	1.00	1.00	1.00	556
accuracy			1.00	1600
macro avg	1.00	1.00	1.00	1600
weighted avg	1.00	1.00	1.00	1600

RESULT

The project predicts the onset of diabetes in a person based on the relevant medical details collected. The data is passed on to the model for training it to make predictions about whether the person is diabetic or non-diabetic the model then makes the prediction with an **accuracy of 98%**, which is fairly good and reliable.

PREDICTION FOR NON-DIABETIC PERSON :

```
Great! You don't have diabetes.
```

PREDICTION FOR DIABETIC PERSON :

```
Oops! You have diabetes.
```

CONCLUSION

After using all these patient records, we are able to build a machine learning model (random forest – best one) to accurately predict whether or not the patients in the dataset have diabetes or not along with that we were able to draw some insights from the data via data analysis and visualization.