Supervised Learning Diabetes report

BERNARD ADEBOYE

Project Goals

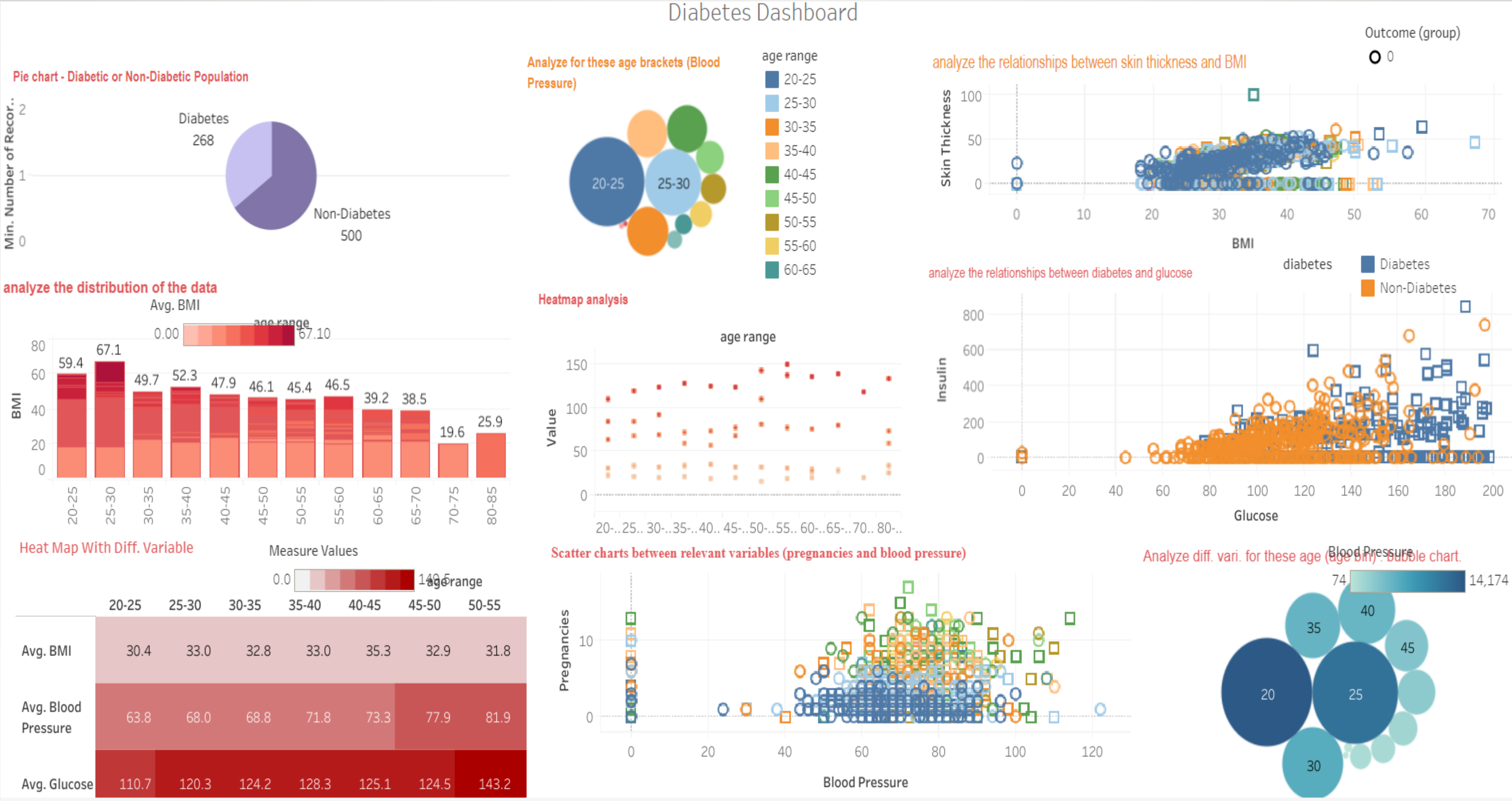
Use supervised learning techniques to build a machine learning model that can predict whether a patient has diabetes or not, based on certain diagnostic measurements.

- Perform exploratory data analysis,
- Preprocessing and feature engineering, and
- Training machine learning model.
- Testing machine learning model.
- Validating machine learning model.

Process

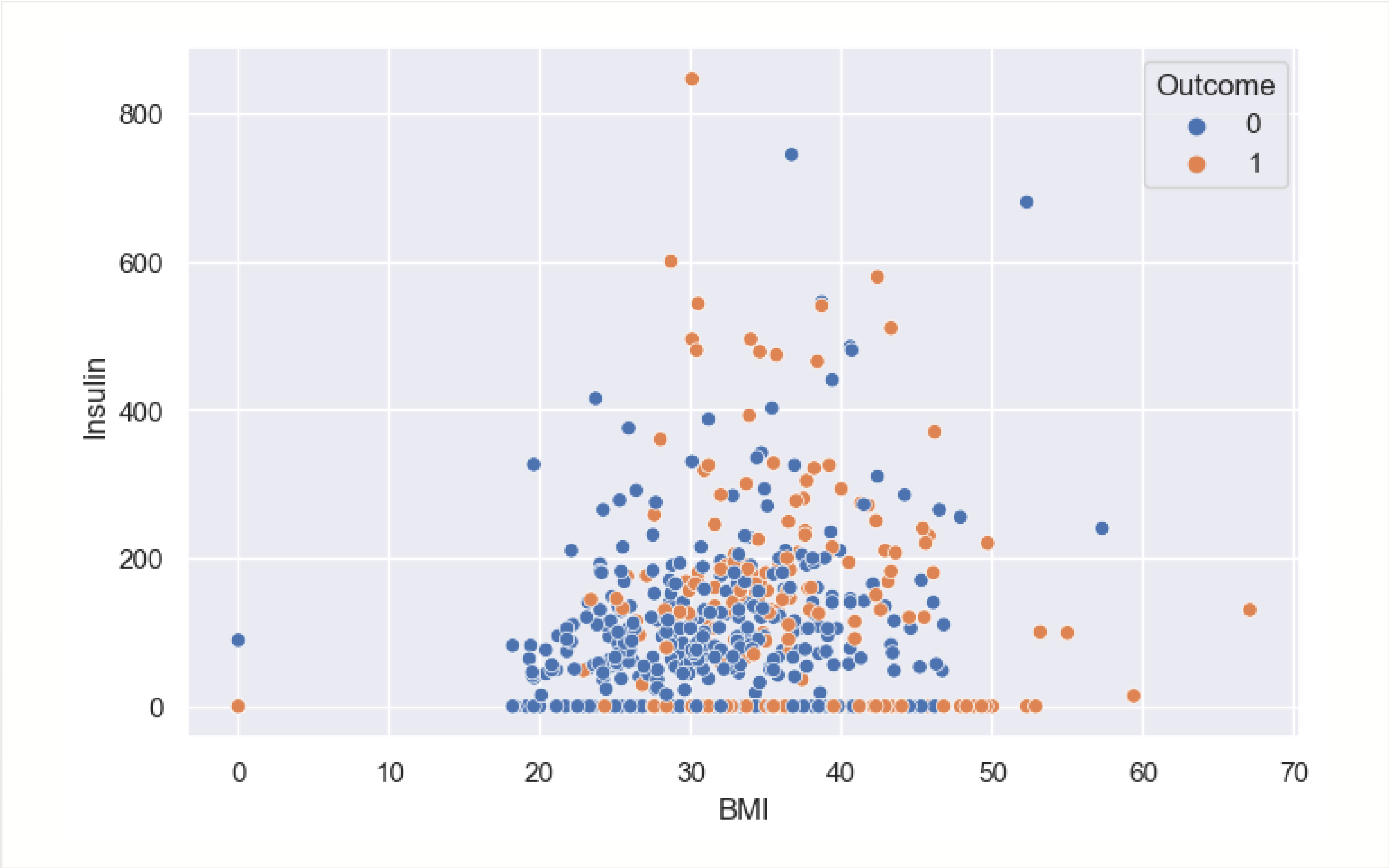
- **Step 1:** Import libraries and load dataset
- **Step 2:** Explorative Data Analysis to understand the dataset.
- **Step 3:** Check for missing information in the dataset.
- **Step 4:** Clean dataset by identifying and treating outliers, and filling missing values.
- **Step 5:** Apply various machine learning algorithms
- **Step 6:** Validate the ML algorithms to ascertain the best.
- Step 7: Interpret and summary findings

Data visualization dashboard

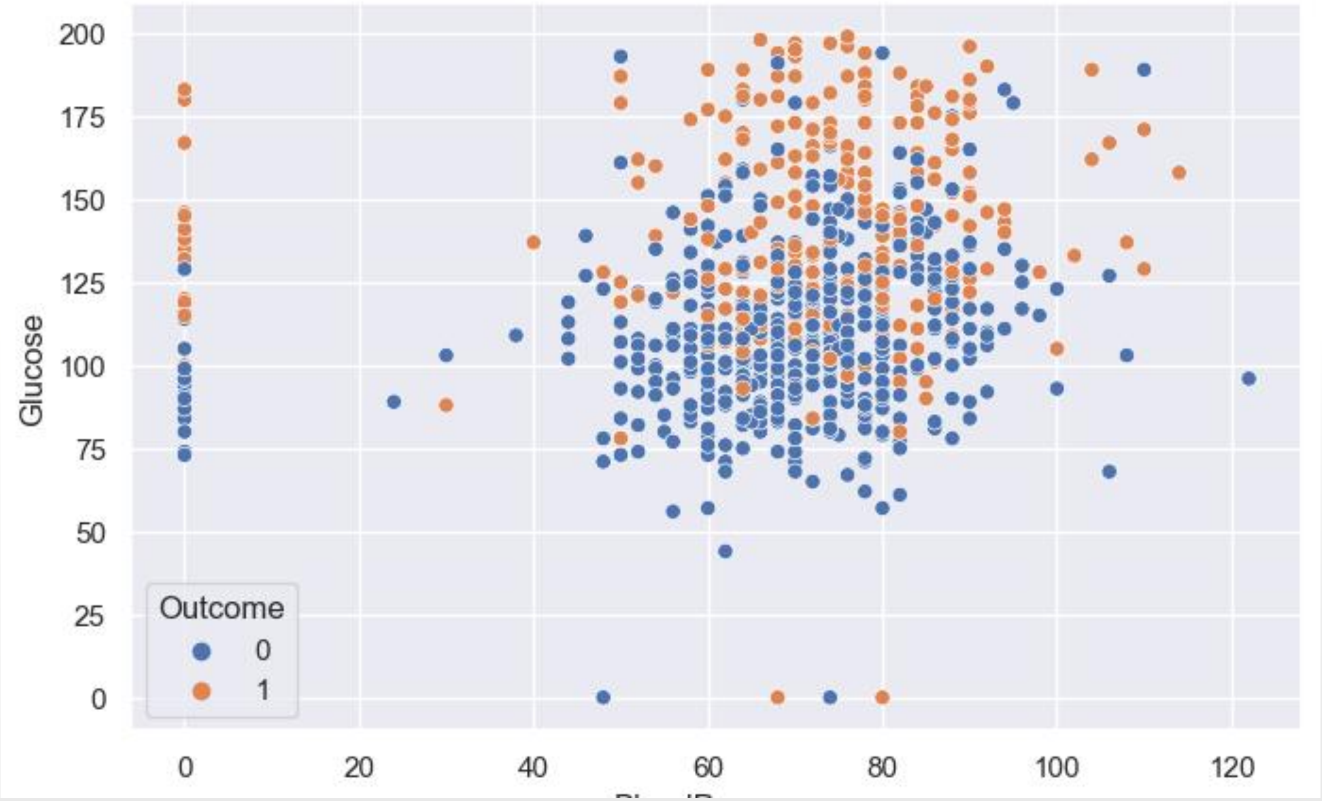


	20-25	25-30	30-35	35-40	40-45	45-50	50-55
Avg. BMI	30.4	33.0	32.8	33.0	35.3	32.9	31.8
Avg. Blood Pressure	63.8	68.0	68.8	71.8	73.3	77.9	81.9
Avg. Glucose	110.7	120.3	124.2	128.3	125.1	124.5	143.2

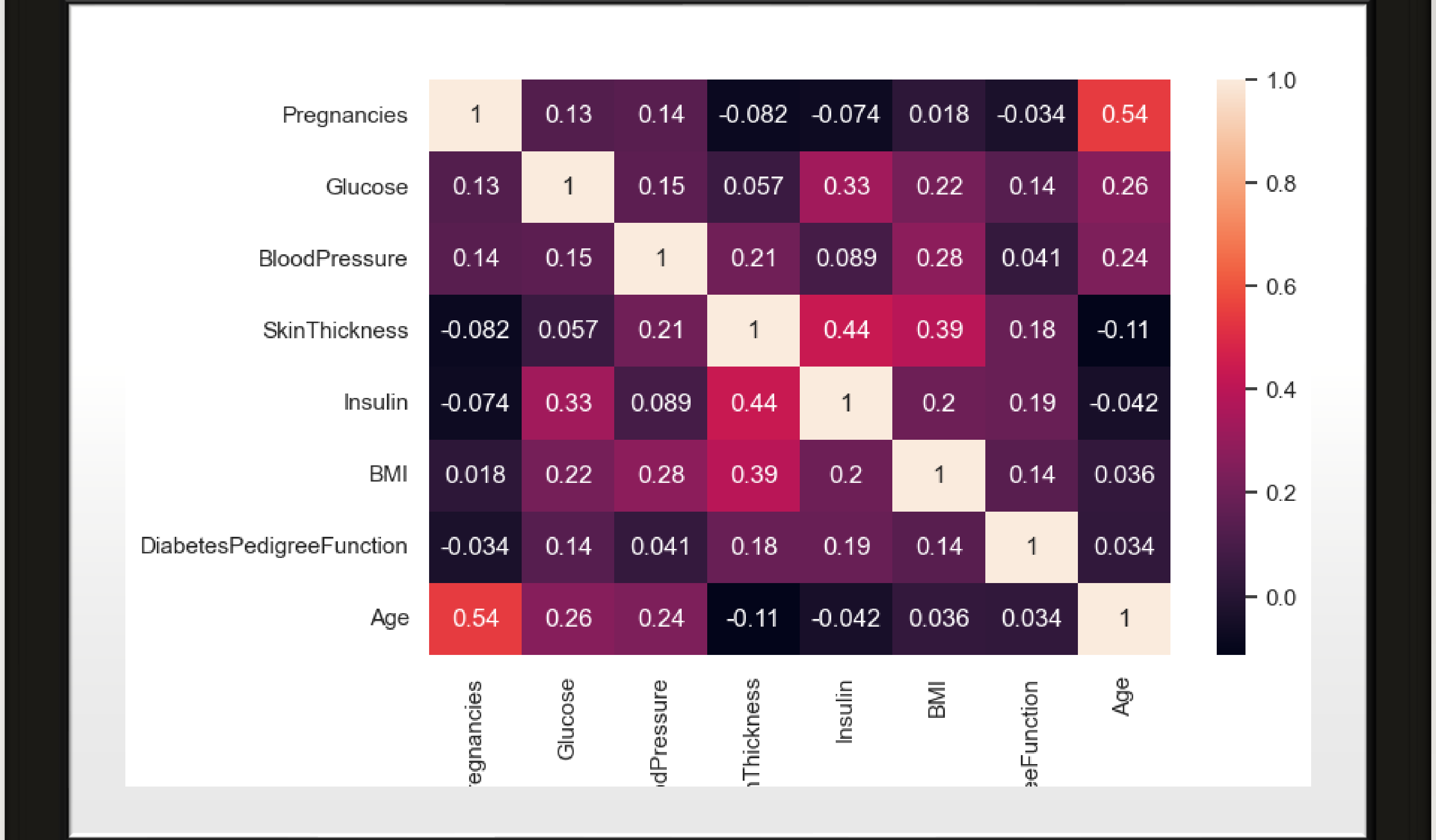
Insulin and BMI chart



Blood Pressure and glucose



Correlation Matrix



	egnancies	Glucose	dPressure	nThickness	Insulin	BMI	eeFunction	Age
Pregnancies	1	0.13	0.14	-0.082	-0.074	0.018	-0.034	0.54
Glucose	0.13	1	0.15	0.057	0.33	0.22	0.14	0.26
BloodPressure	0.14	0.15	1	0.21	0.089	0.28	0.041	0.24
SkinThickness	-0.082	0.057	0.21	1	0.44	0.39	0.18	-0.11
Insulin	-0.074	0.33	0.089	0.44	1	0.2	0.19	-0.042
BMI	0.018	0.22	0.28	0.39	0.2	1	0.14	0.036
DiabetesPedigreeFunction	-0.034	0.14	0.041	0.18	0.19	0.14	1	0.034
Age	0.54	0.26	0.24	-0.11	-0.042	0.036	0.034	1

Dataset findings

After analyzing the histogram we can identify that there are some outliers in some columns.

For Example:-

Blood Pressure - A living person cannot have a diastolic blood pressure of zero.

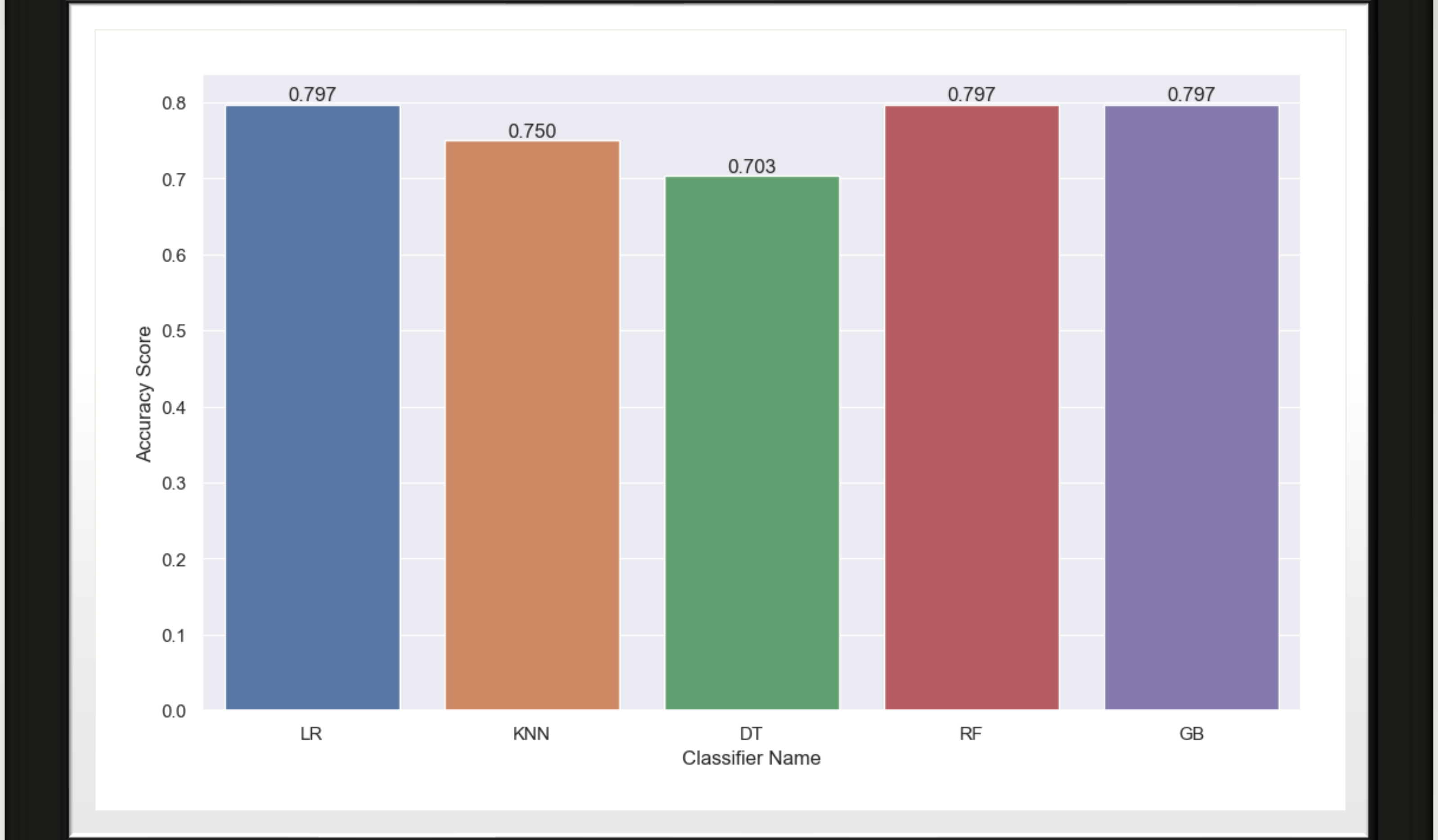
Plasma glucose levels - Zero is invalid number as fasting glucose level would never be as low as zero.

Skin Fold Thickness - For normal people, skin fold thickness can’t be less than 10 mm better yet zero.

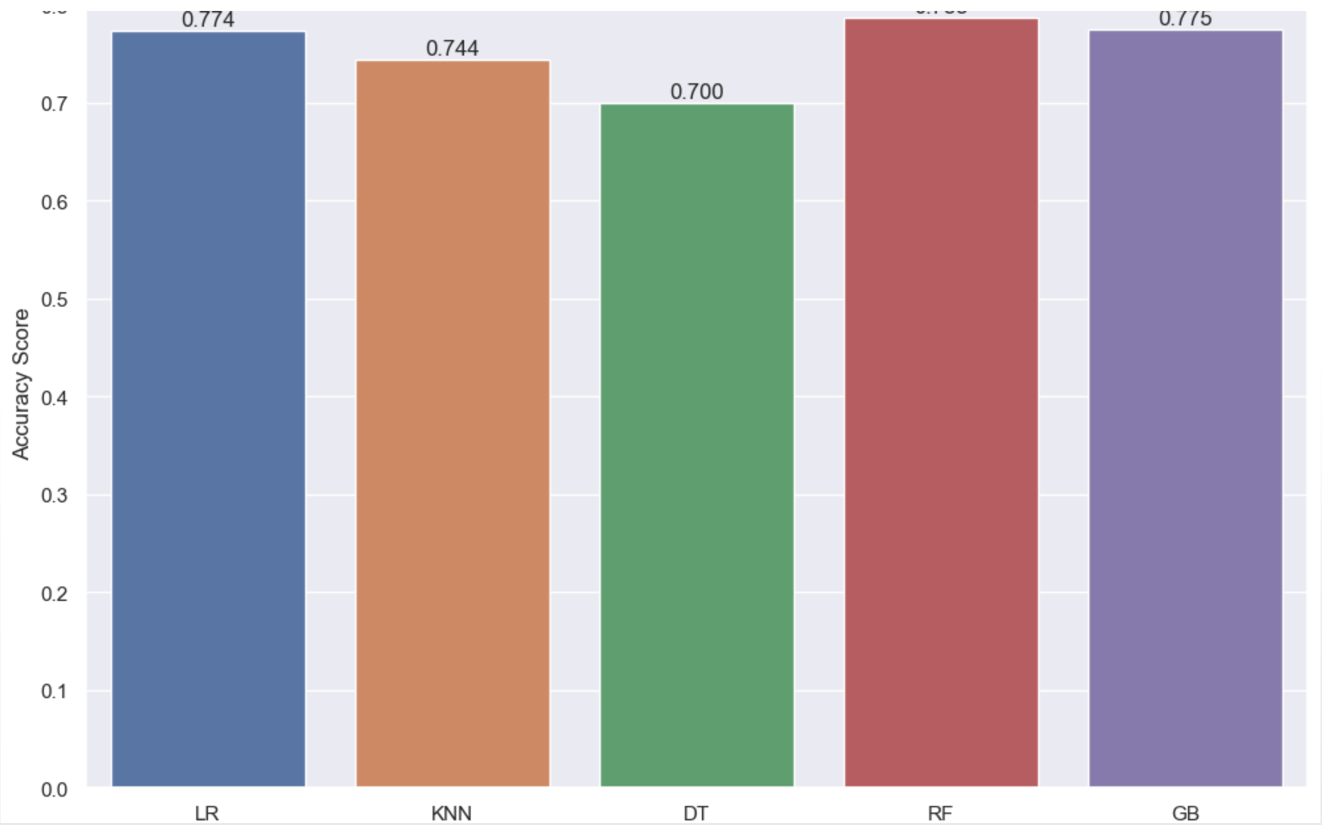
BMI: Should not be 0 or close to zero unless the person is really underweight which could be life-threatening.

Insulin: In a rare situation a person can have zero insulin but by observing

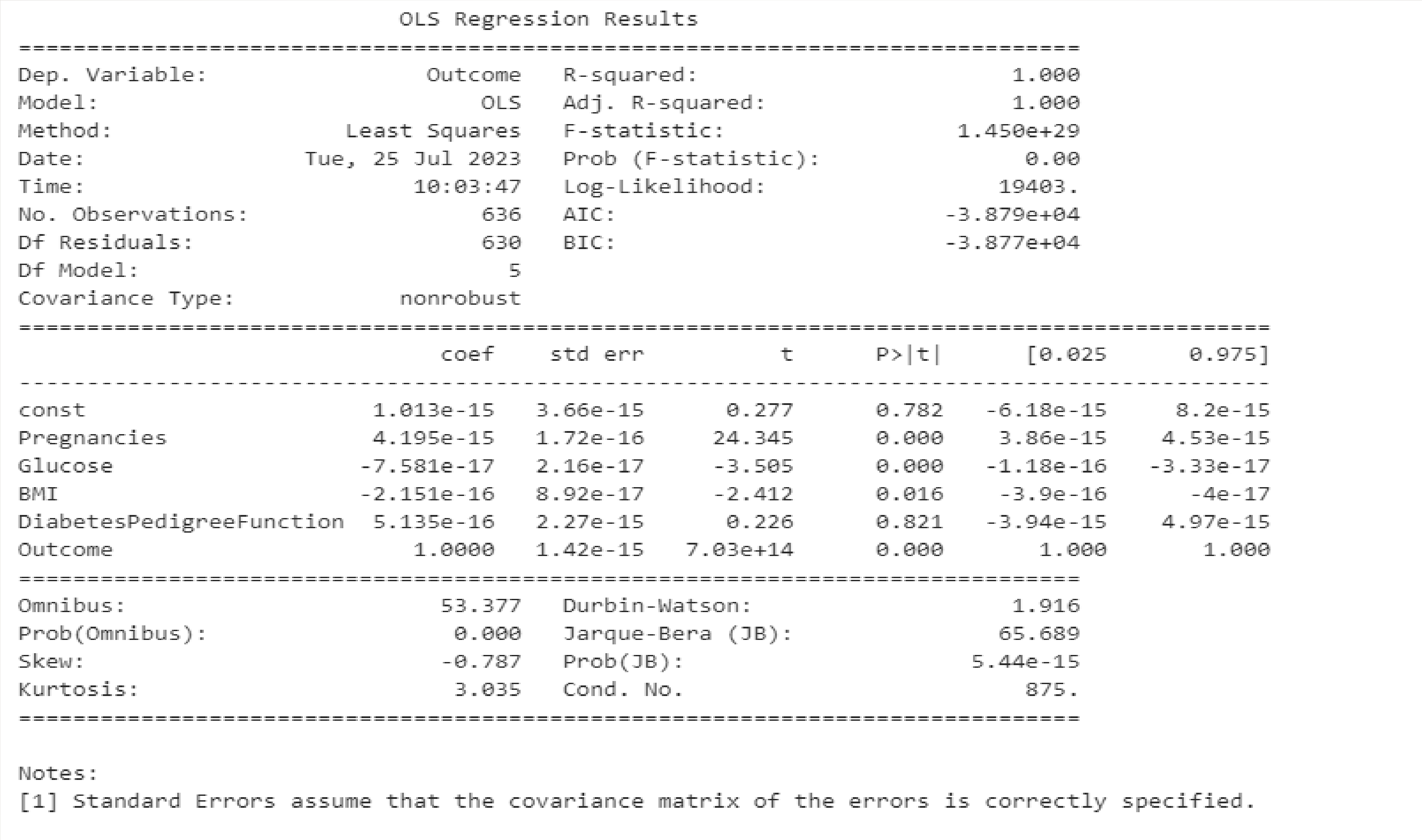
Models performance- Accuracy score metrics



Model performance – Kfold Cross validation

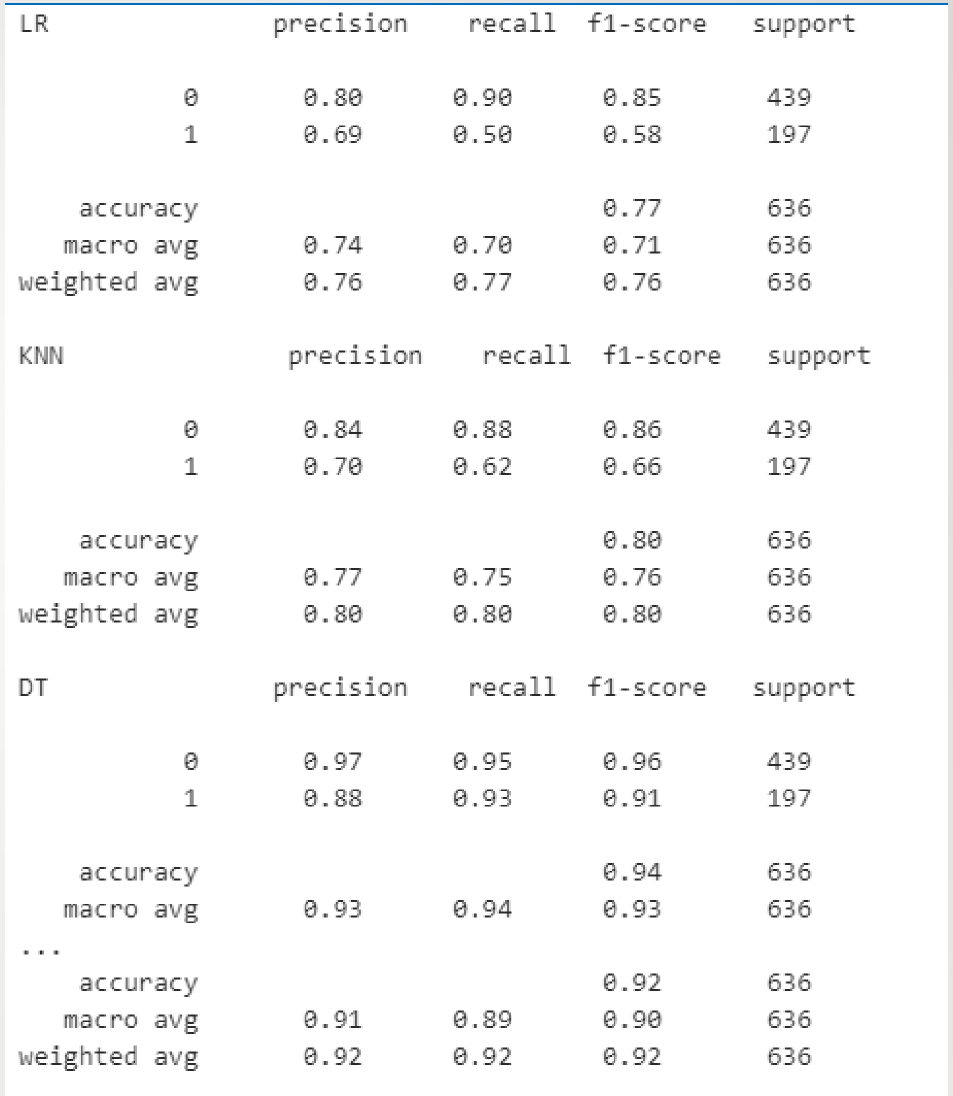


Model summary – Logistic Regression



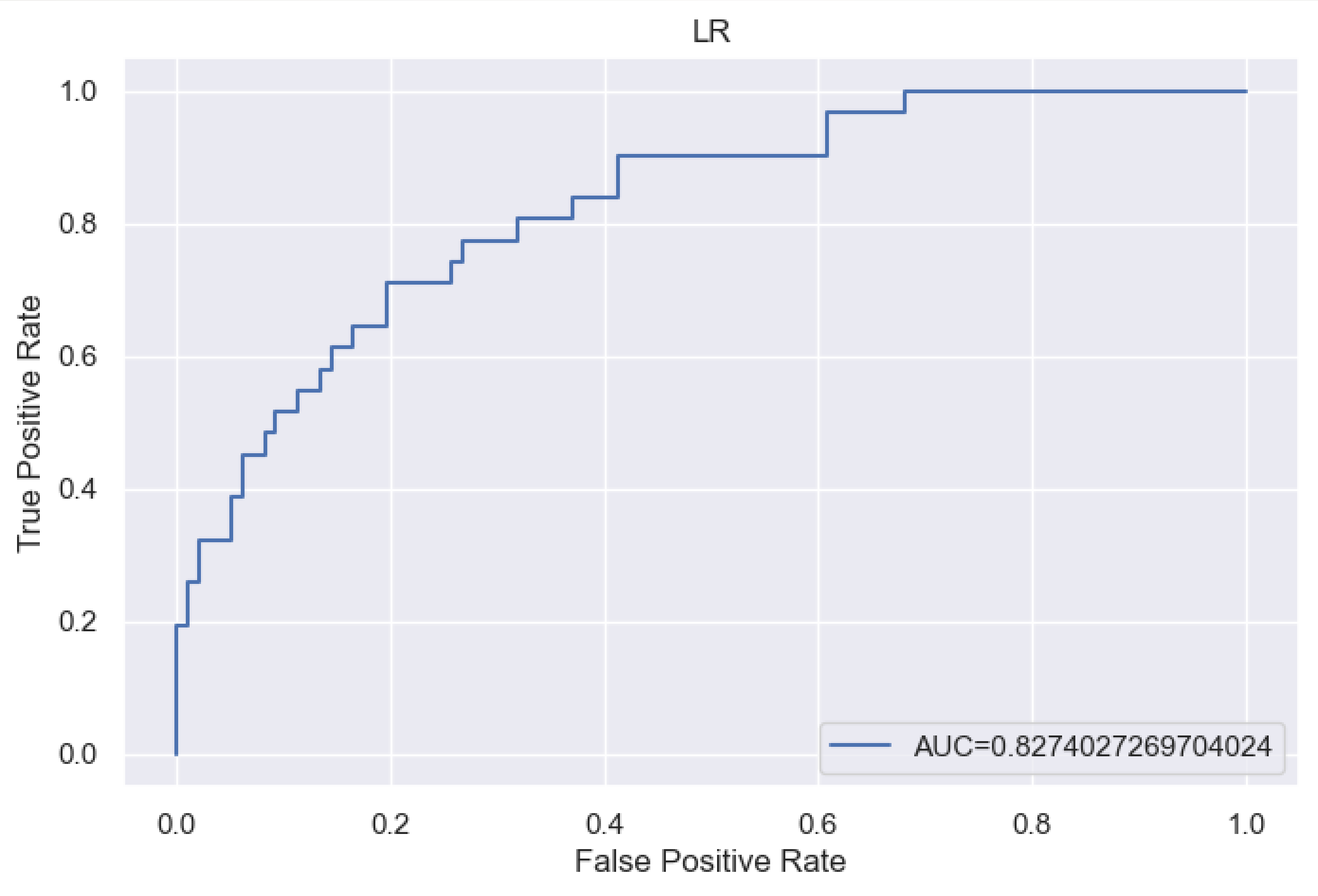
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                Outcome   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 1.450e+29
Date:                Tue, 25 Jul 2023   Prob (F-statistic):               0.00
Time:                        10:03:47   Log-Likelihood:                 19403.
No. Observations:                 636   AIC:                        -3.879e+04
Df Residuals:                     630   BIC:                        -3.877e+04
Df Model:                           5
Covariance Type:            nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                     1.013e-15   3.66e-15      0.277      0.782   -6.18e-15     8.2e-15
Pregnancies               4.195e-15   1.72e-16     24.345      0.000    3.86e-15    4.53e-15
Glucose                  -7.581e-17   2.16e-17     -3.505      0.000   -1.18e-16   -3.33e-17
BMI                      -2.151e-16   8.92e-17     -2.412      0.016    -3.9e-16      -4e-17
DiabetesPedigreeFunction  5.135e-16   2.27e-15      0.226      0.821   -3.94e-15    4.97e-15
Outcome                      1.0000   1.42e-15   7.03e+14      0.000       1.000       1.000
==============================================================================
Omnibus:                       53.377   Durbin-Watson:                   1.916
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               65.689
Skew:                          -0.787   Prob(JB):                     5.44e-15
Kurtosis:                       3.035   Cond. No.                         875.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Model classification report-recall, precision, f1-score

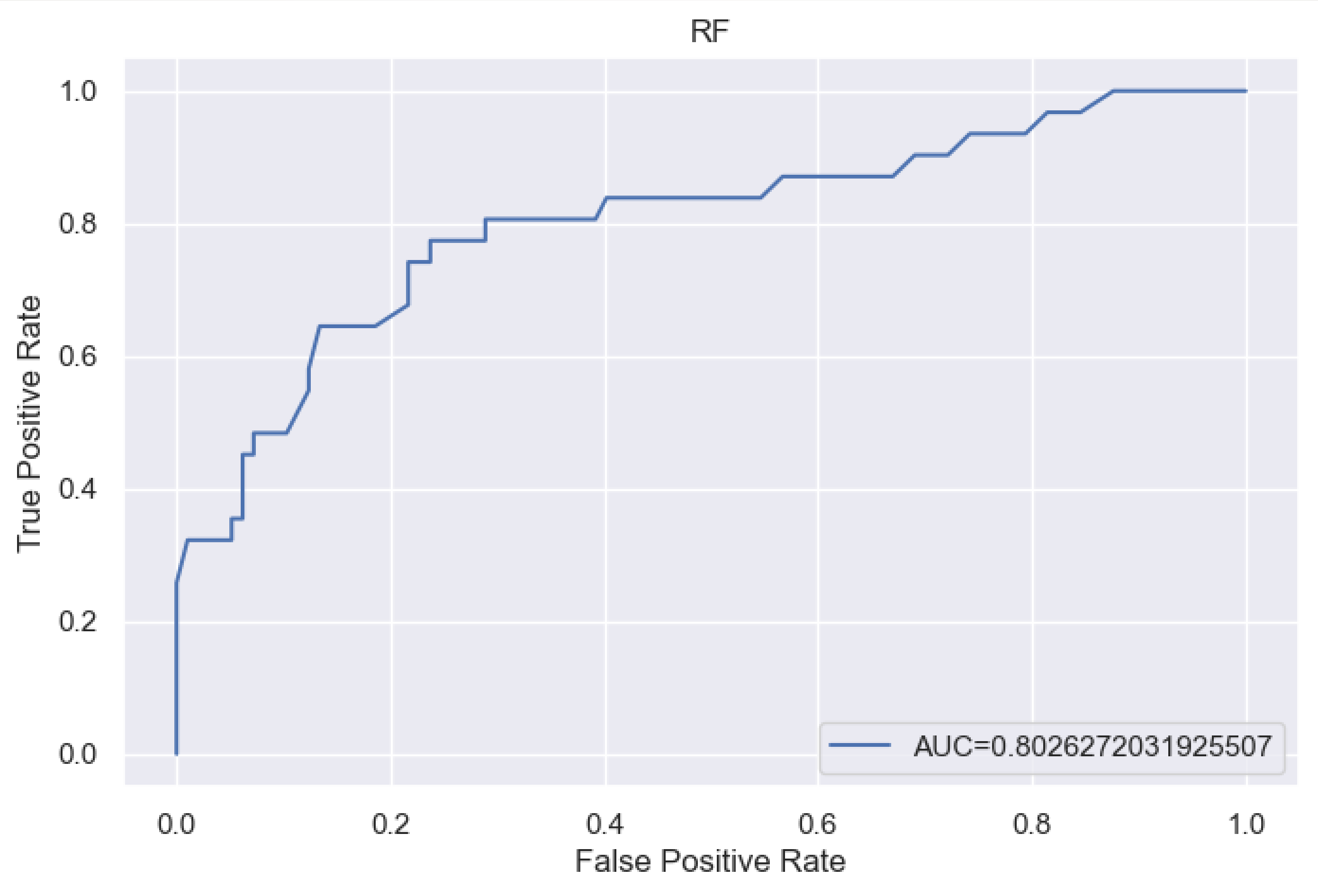


```
LR                precision    recall  f1-score   support

           0       0.80      0.90      0.85       439
           1       0.69      0.50      0.58       197

    accuracy                           0.77       636
   macro avg       0.74      0.70      0.71       636
weighted avg       0.76      0.77      0.76       636

KNN                precision    recall  f1-score   support

           0       0.84      0.88      0.86       439
           1       0.70      0.62      0.66       197

    accuracy                           0.80       636
   macro avg       0.77      0.75      0.76       636
weighted avg       0.80      0.80      0.80       636

DT                precision    recall  f1-score   support

           0       0.97      0.95      0.96       439
           1       0.88      0.93      0.91       197

    accuracy                           0.94       636
   macro avg       0.93      0.94      0.93       636
...
    accuracy                           0.92       636
   macro avg       0.91      0.89      0.90       636
weighted avg       0.92      0.92      0.92       636
```

Roc-AUC curve for Logistic regression



Roc-AUC curve for Random Forest



Conclusion

- We can see the Logistic Regression, Random Forest and Gradient Boosting have performed better than the rest.
- Diabetic and non-diabetic groups shows similar distribution pattern.
- Most variables shows relative positive relationship between themselves.
- Skin thickness and Insulin shows a lot of outliers due to the numbers of zeros.
- Pregnancies, glucose and BMI variables help to explain the outcome variables better.