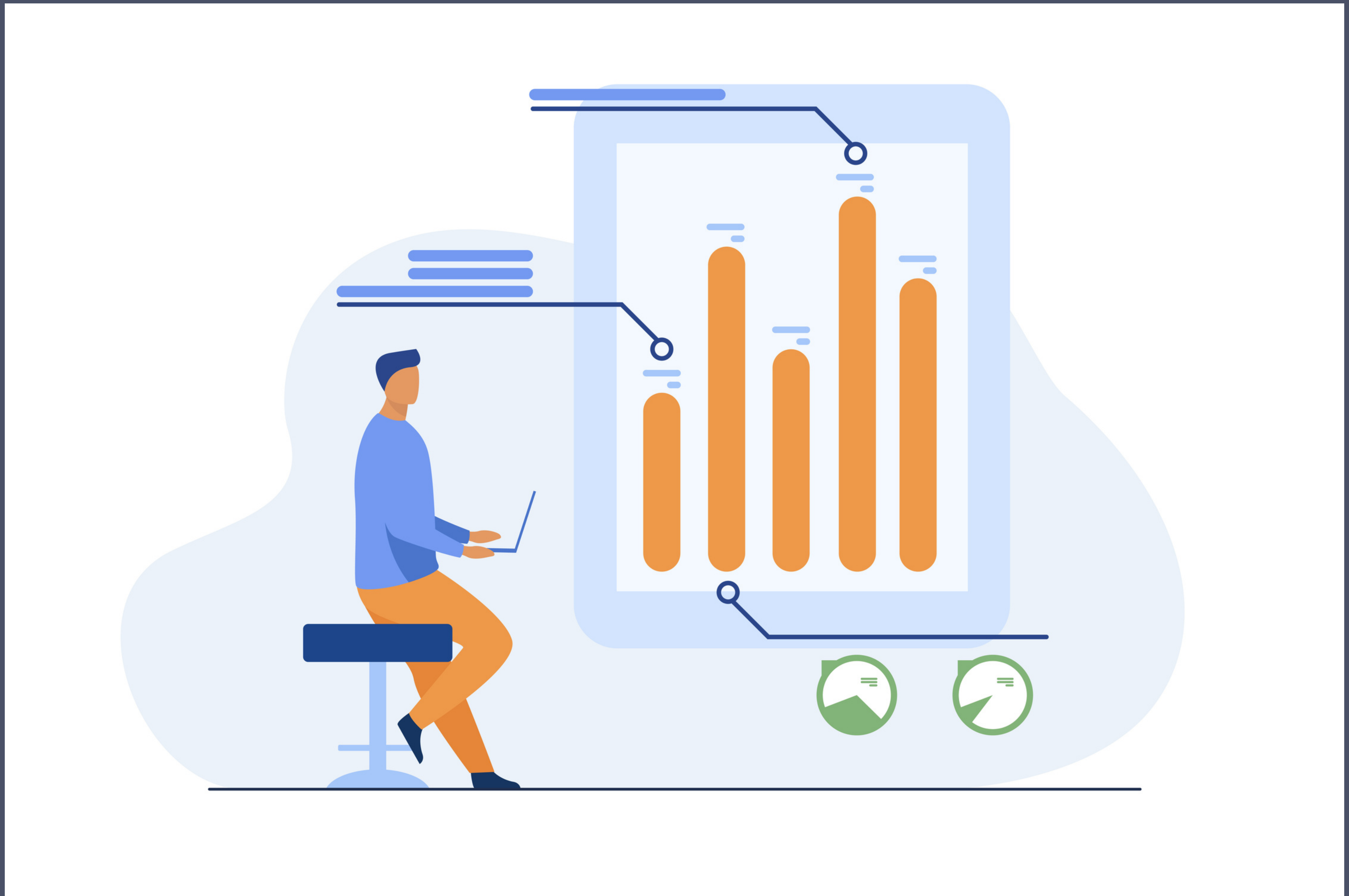
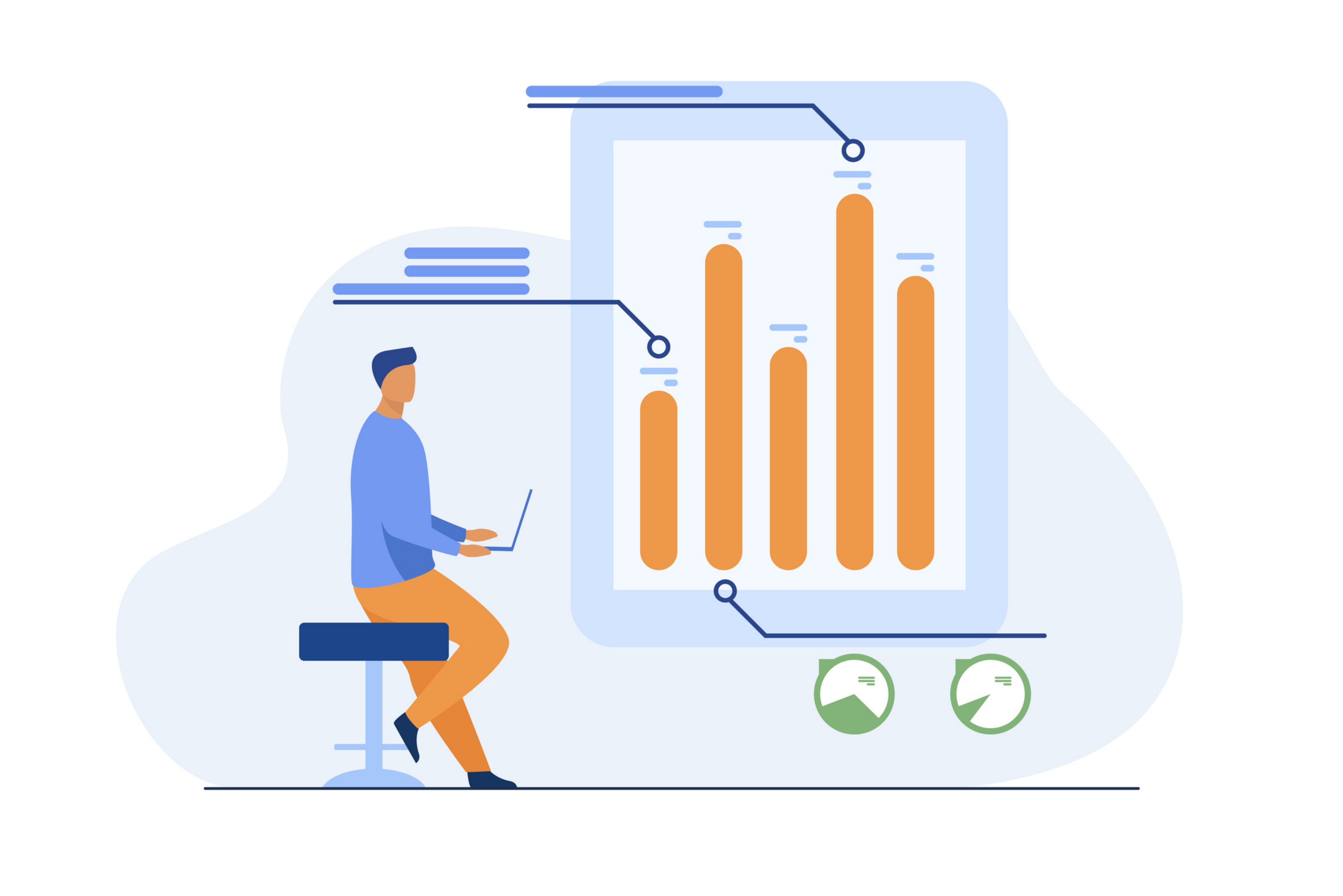
Predicting Diabetes Based on Diagnostic Measures

A Supervised Learning Approach

Lara Onipede



Project Goals



To perform Exploratory Data Analysis on dataset

To apply Supervised learning techniques to gain insights from the dataset.

To communicate insights using visualizations

To build Appropriate Machine Learning Projects for prediction

Dataset

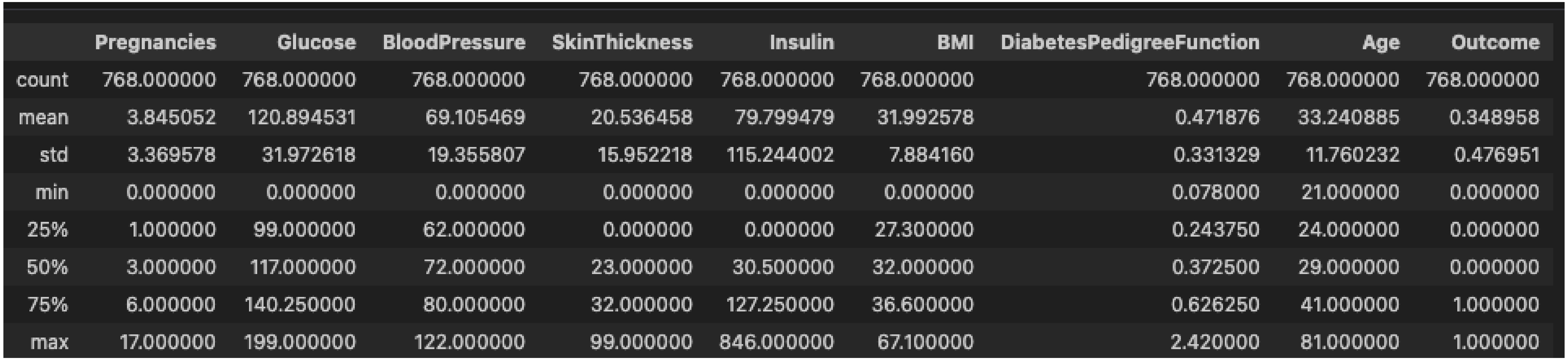
Diabetes dataset from the National Institute of Diabetes and Digestive and Kidney Diseases.

768 rows, 9 Columns

Exploratory Data Analysis

EDA involves examination of the Dataset to gain insights, understand patterns, and identify potential capabilities as well as challenges that lies ahead in the course of the project

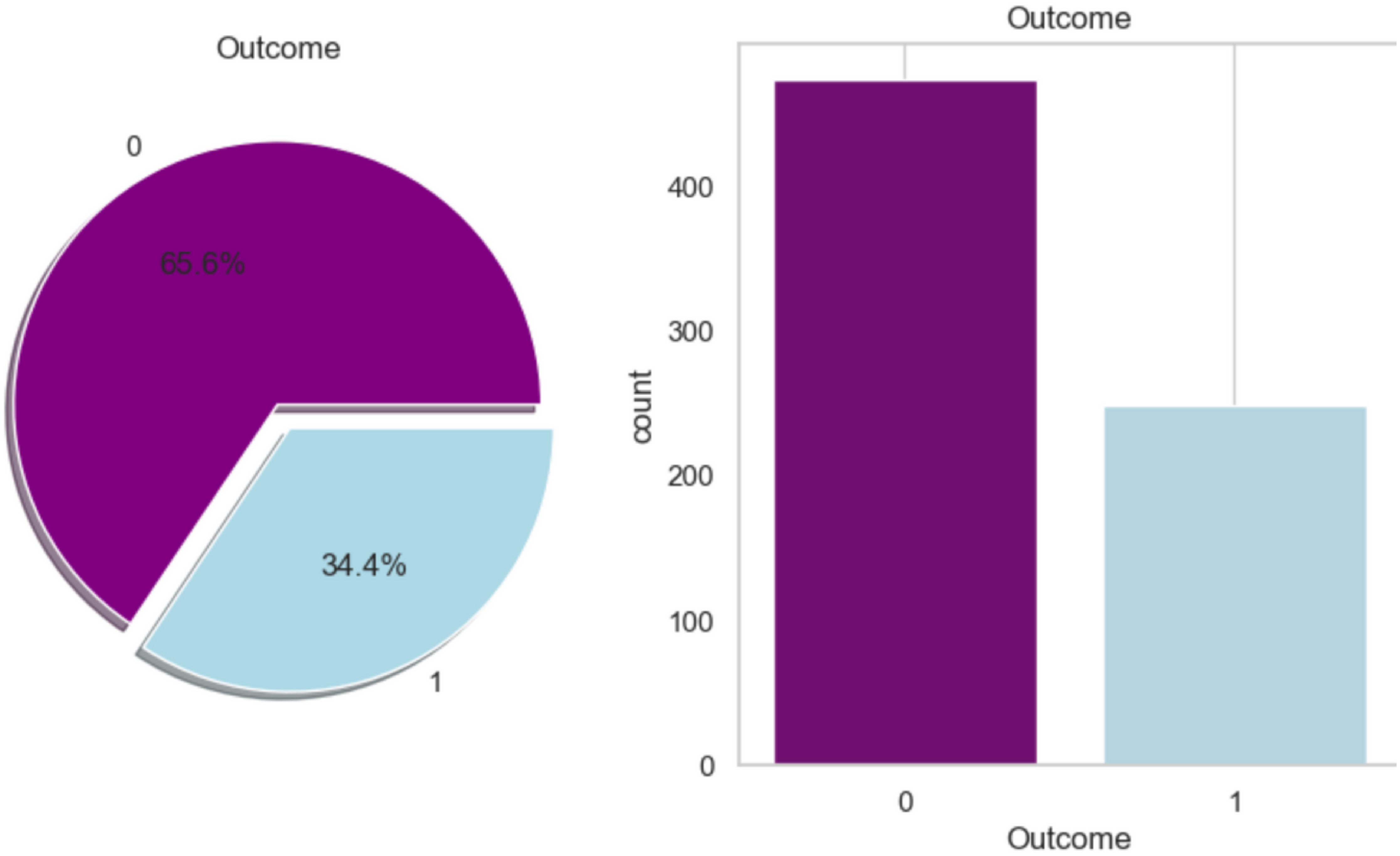
Summary Statistics of Diabetes Dataset



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

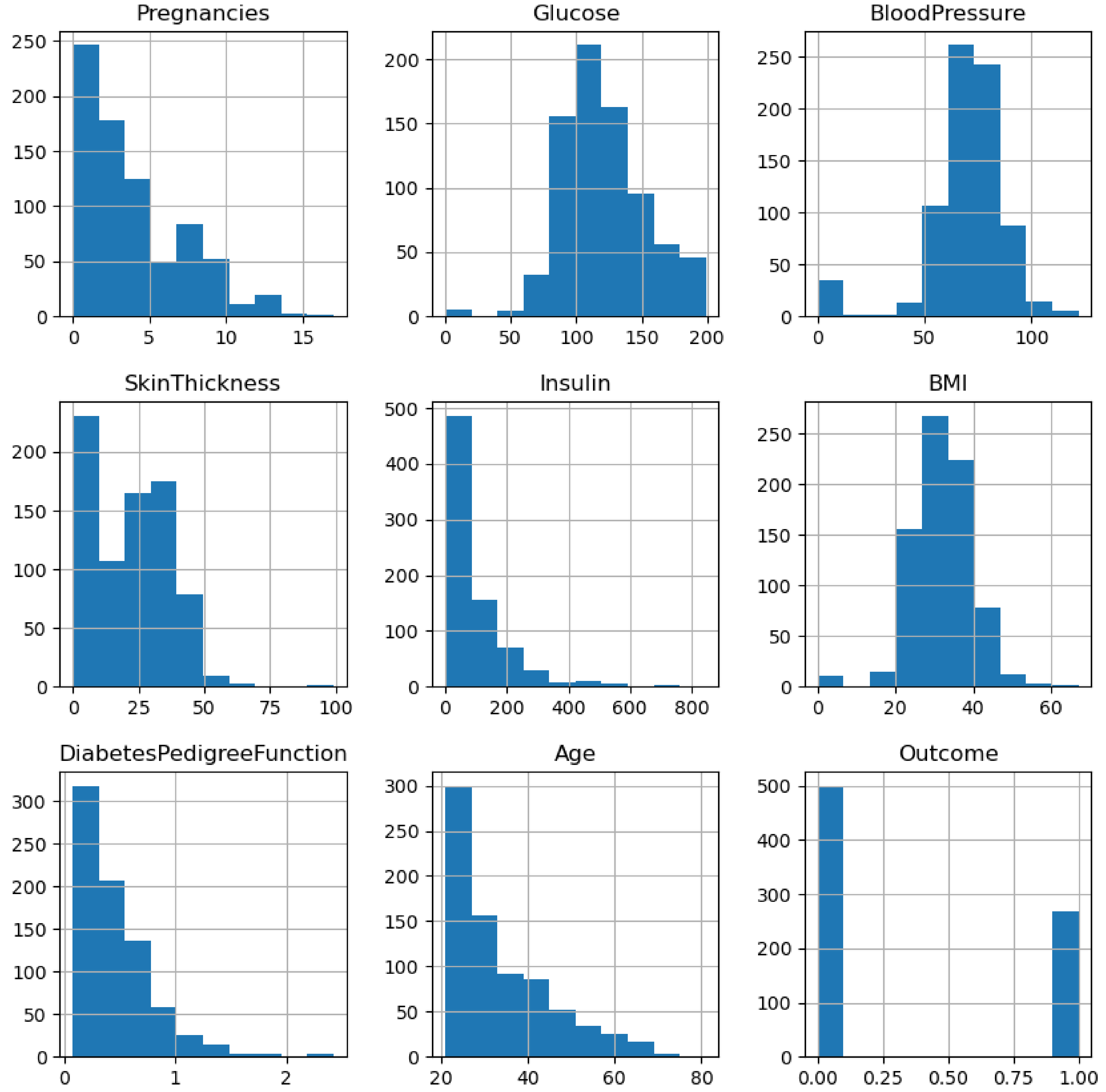
Data Visualizations

Diabetic: 500
Non-diabetic: 268

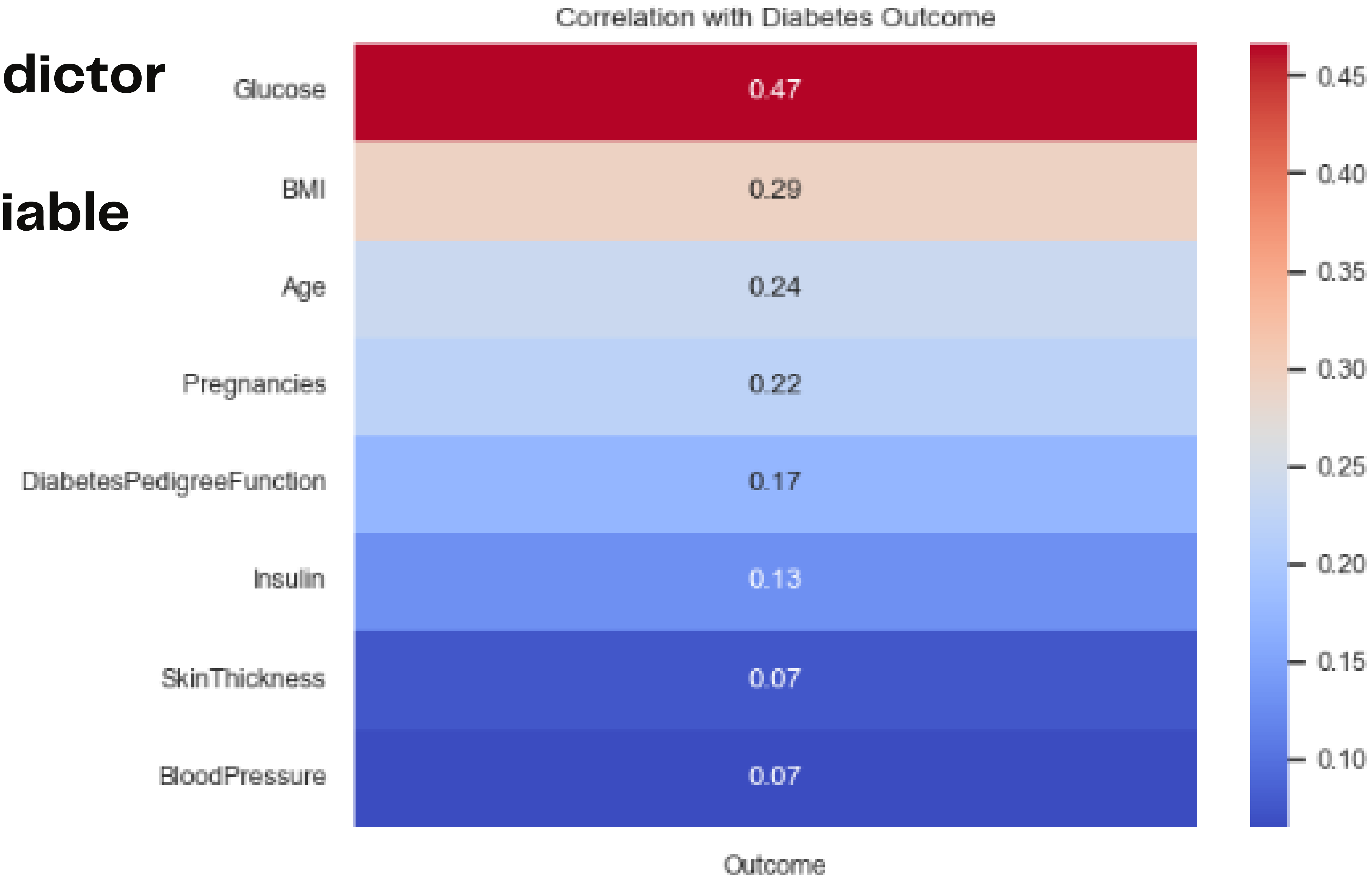


Proportion and Count of each category of the target Variable

Distribution of features



Correlation between predictor variable and outcome variable



Training Machine Learning Models

Data was split into 80% training set and 20% testing set.

Data Was trained on the following Models

1 Logistic Regression

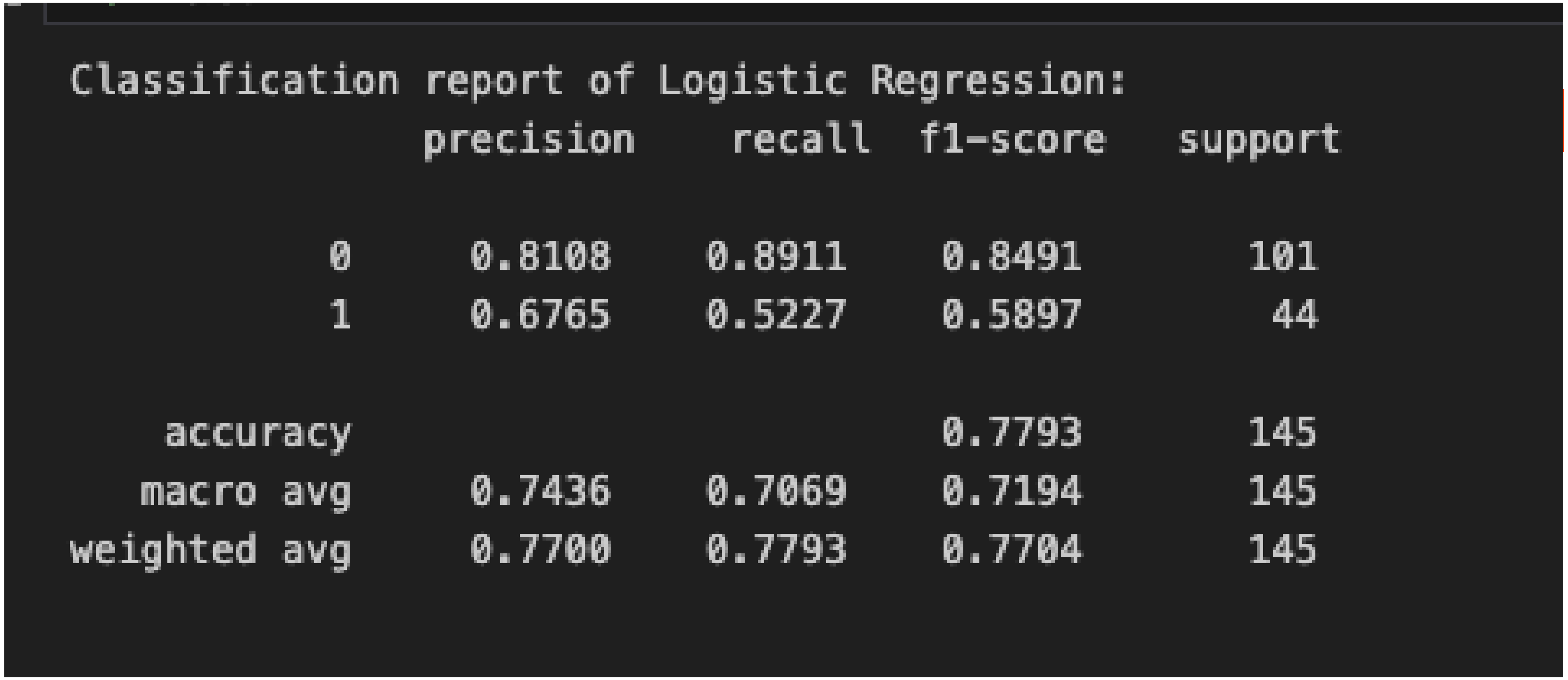
An interpretable Classification Algorithm for Binary and multiclass Classification tasks. I considered it for it's simplicity and ease of interpretaton

2 Random Forest

This is an ensemble learning technique that is renowned for handling categorical features for classification and regression tasks

3 K-Nearest Neighbors (KNN)

Classification report of Logistic Regression Model

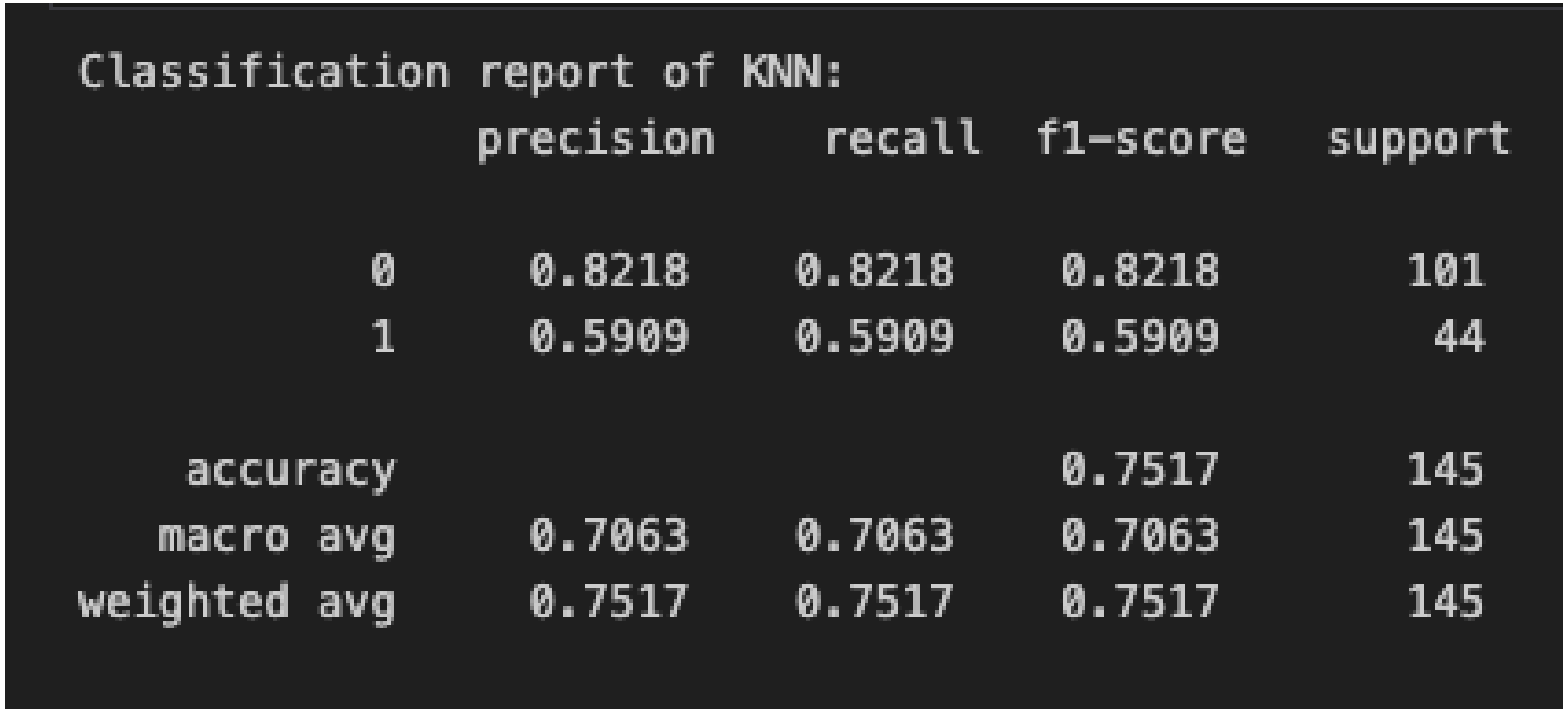


```
Classification report of Logistic Regression:
              precision    recall  f1-score   support

           0     0.8108    0.8911    0.8491       101
           1     0.6765    0.5227    0.5897        44

    accuracy                         0.7793       145
   macro avg     0.7436    0.7069    0.7194       145
weighted avg     0.7700    0.7793    0.7704       145
```

Classification report of KNN Model

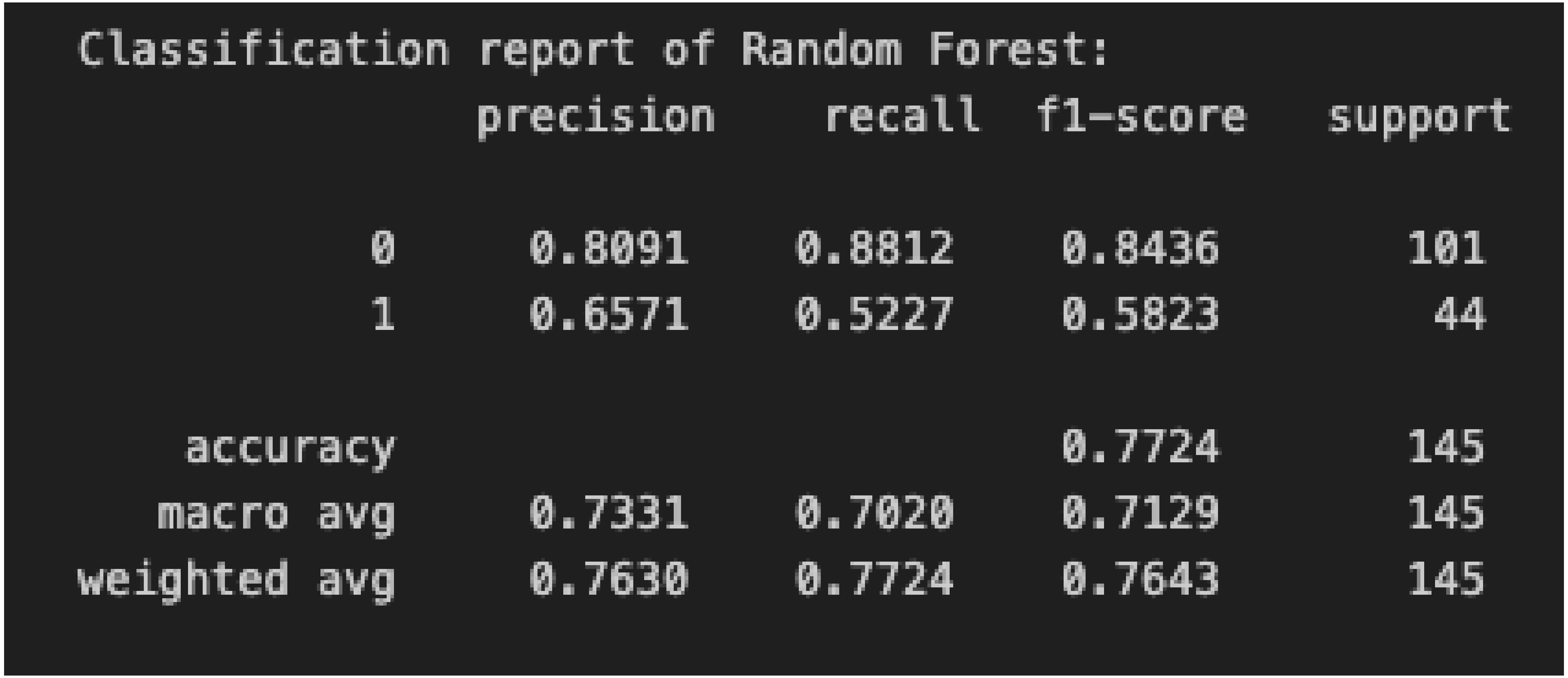


```
Classification report of KNN:
              precision    recall  f1-score   support

           0     0.8218    0.8218    0.8218       101
           1     0.5909    0.5909    0.5909        44

    accuracy                         0.7517       145
   macro avg     0.7063    0.7063    0.7063       145
weighted avg     0.7517    0.7517    0.7517       145
```

Classification report of Random Forest Model



```
Classification report of Random Forest:
              precision    recall  f1-score   support

           0     0.8091    0.8812    0.8436       101
           1     0.6571    0.5227    0.5823        44

    accuracy                         0.7724       145
   macro avg     0.7331    0.7020    0.7129       145
weighted avg     0.7630    0.7724    0.7643       145
```

Conclusion

- Our predictive models demonstrated promising results in diabetes prediction, achieving over 75% accuracy on the test set.
- The models showed consistent performance in differentiating individuals with and without diabetes, indicating their potential as valuable diagnostic tools.

01

Important features

- "Glucose" and "BMI" emerged as the most critical features contributing significantly to diabetes prediction.
- The "Glucose" level exhibited a strong positive correlation with the presence of diabetes, while "BMI" played a vital role in distinguishing diabetic and non-diabetic individuals.

02

Challenges

imbalance in some features may have affected model performance and generalizability

03

Future Work

Conduct Model fine tuning by optimizing hyperparameters
Explore feature selection techniques to identify Most Informative features

Thank You