SEN4018 Project

Pima Indians Diabetes

Afaf Alalwan (1901077)
Charaf-Eddine M'rah (1900298)

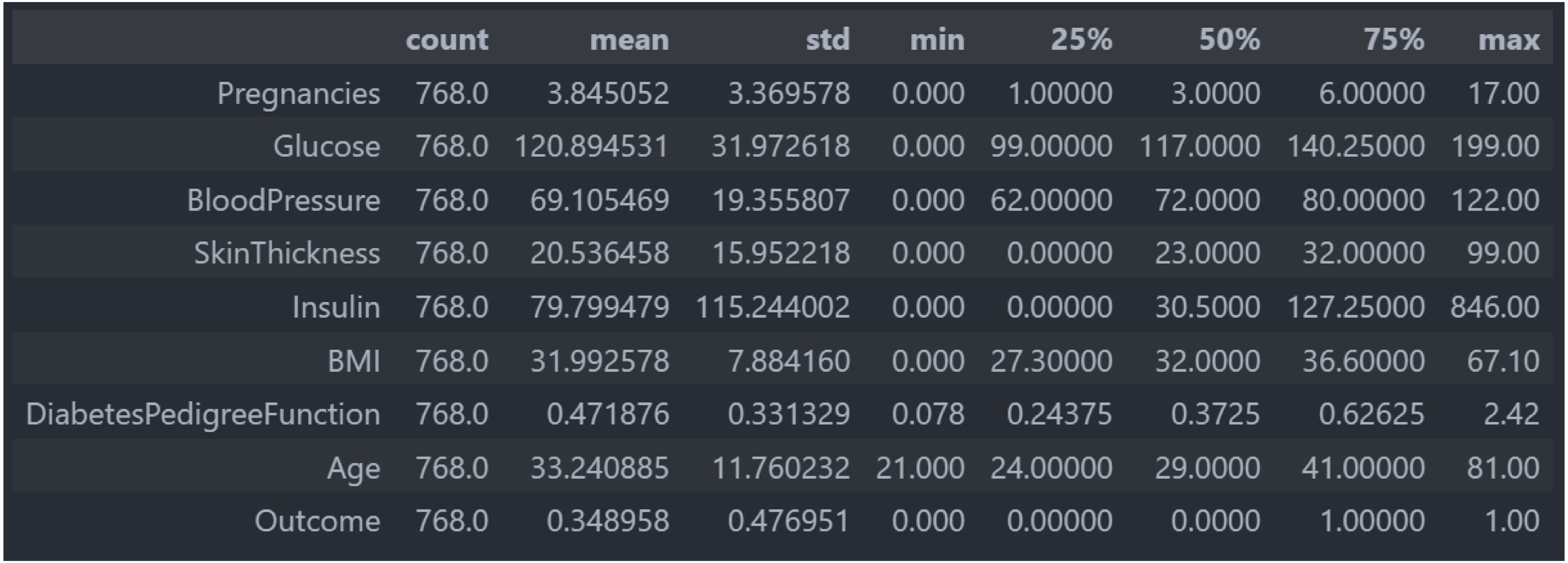
Dataset Description

The Pima Indians Diabetes Databas is provided by The National Institute of Diabetes and Digestive and Kidney Diseases. This dataset is a subset of the larger dataset. In this dataset, all of the patients, are Pima Indian women who are at least 21 years old. The dataset contains 8 medical predictor factors.

Medical Factors:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)

Statistical Description



	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

Data Preprocessing

01

Step 1

Visualize raw data

02

Step 2

Preprocess data

03

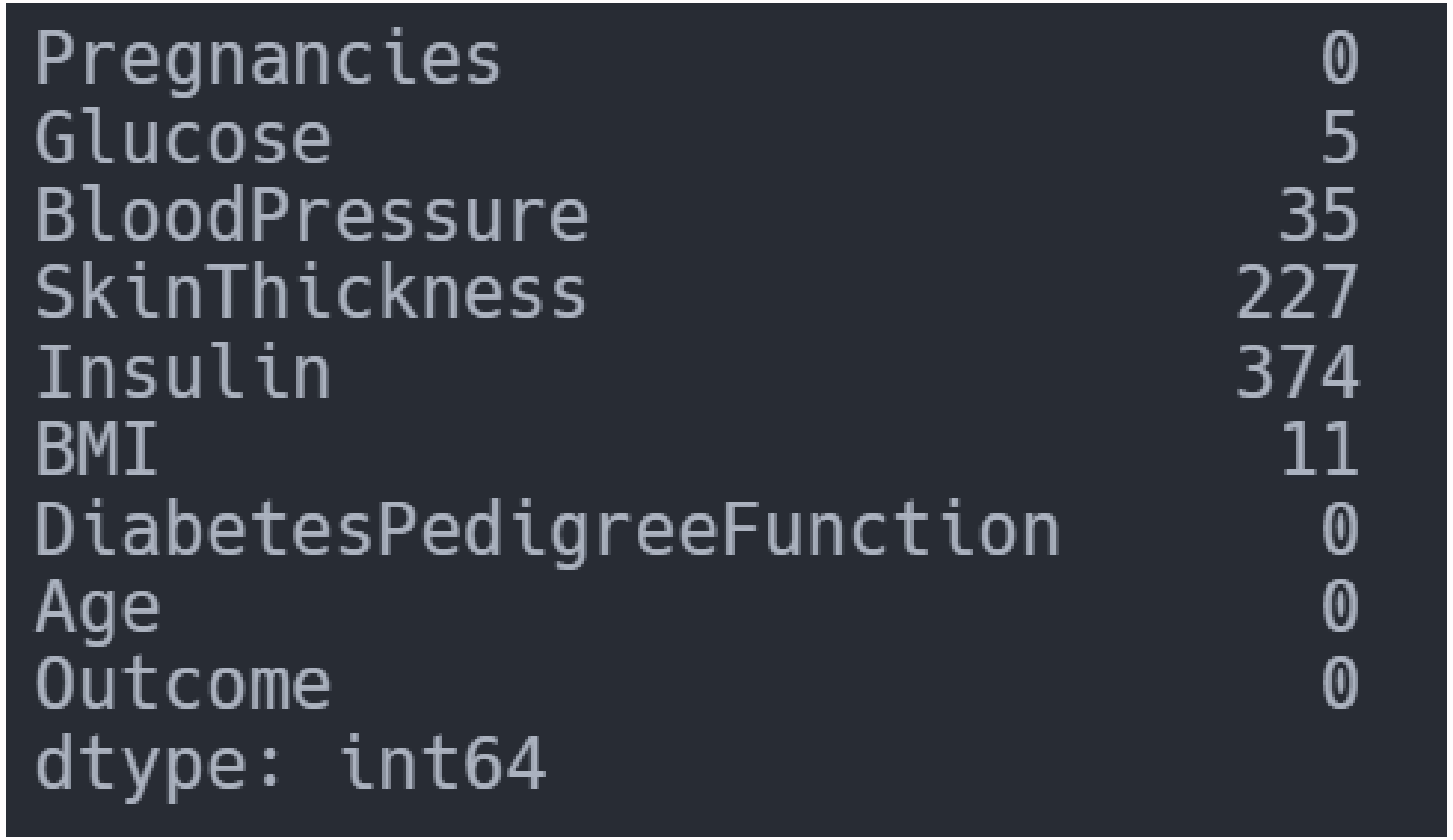
Step 3

Visualize Preprocessed data

Missing Data

These values can't be zero, so missing data is converted to NAN:

- Glucose
- BloodPressure
- SkinThickness
- Insuling
- BMI

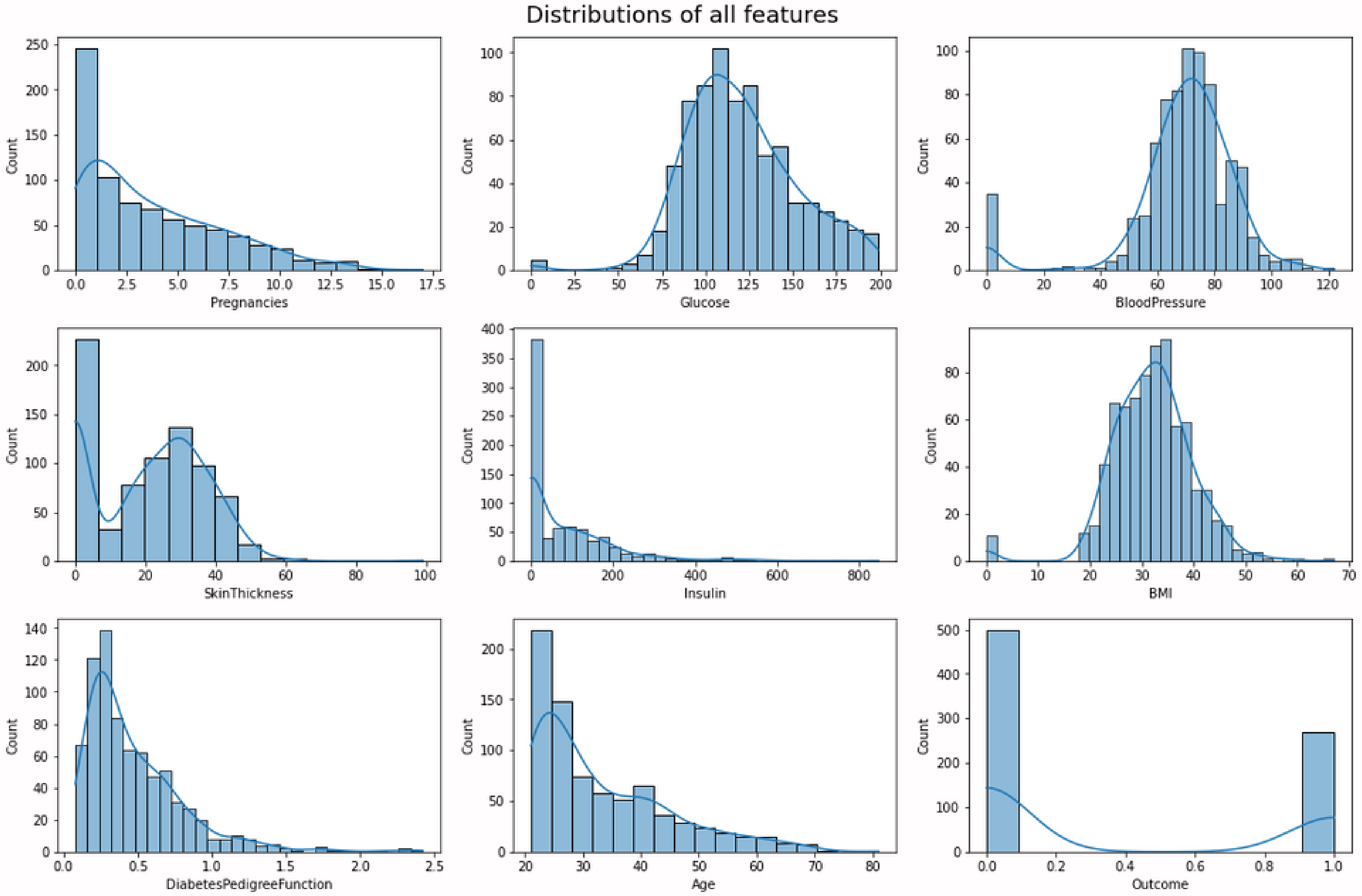


```
Pregnancies                   0
Glucose                       5
BloodPressure                35
SkinThickness               227
Insulin                     374
BMI                          11
DiabetesPedigreeFunction      0
Age                           0
Outcome                       0
dtype: int64
```

Number of NAN values for each feature

Features Distribution

This is the distributions of all features **before** imputation:



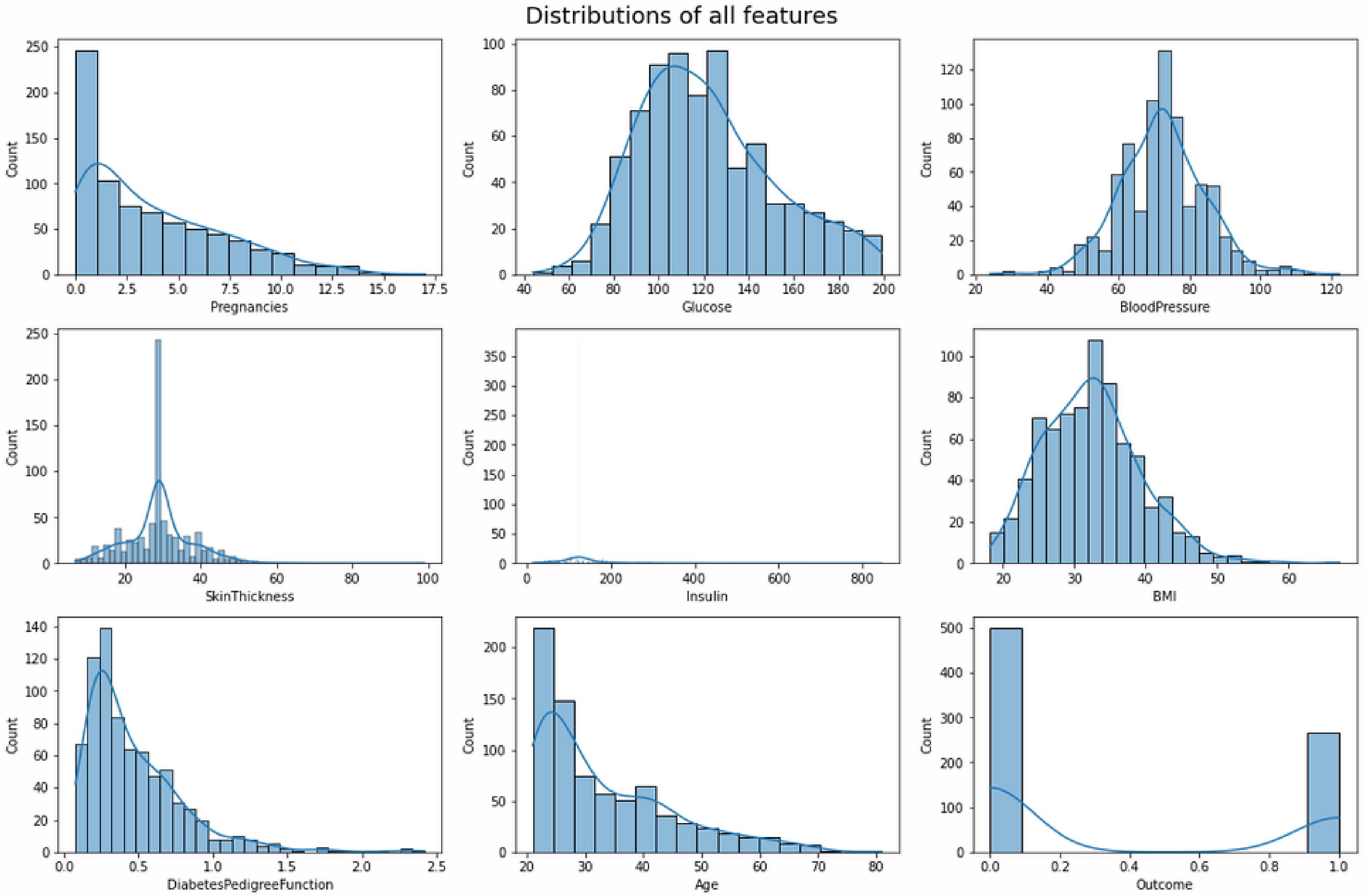
Features Distribution

NAN values were replaced with more suitable values:

- Glucose -> mean
- BloodPressure -> mean
- SkinThickness -> median
- Insulin -> mean
- BMI -> median

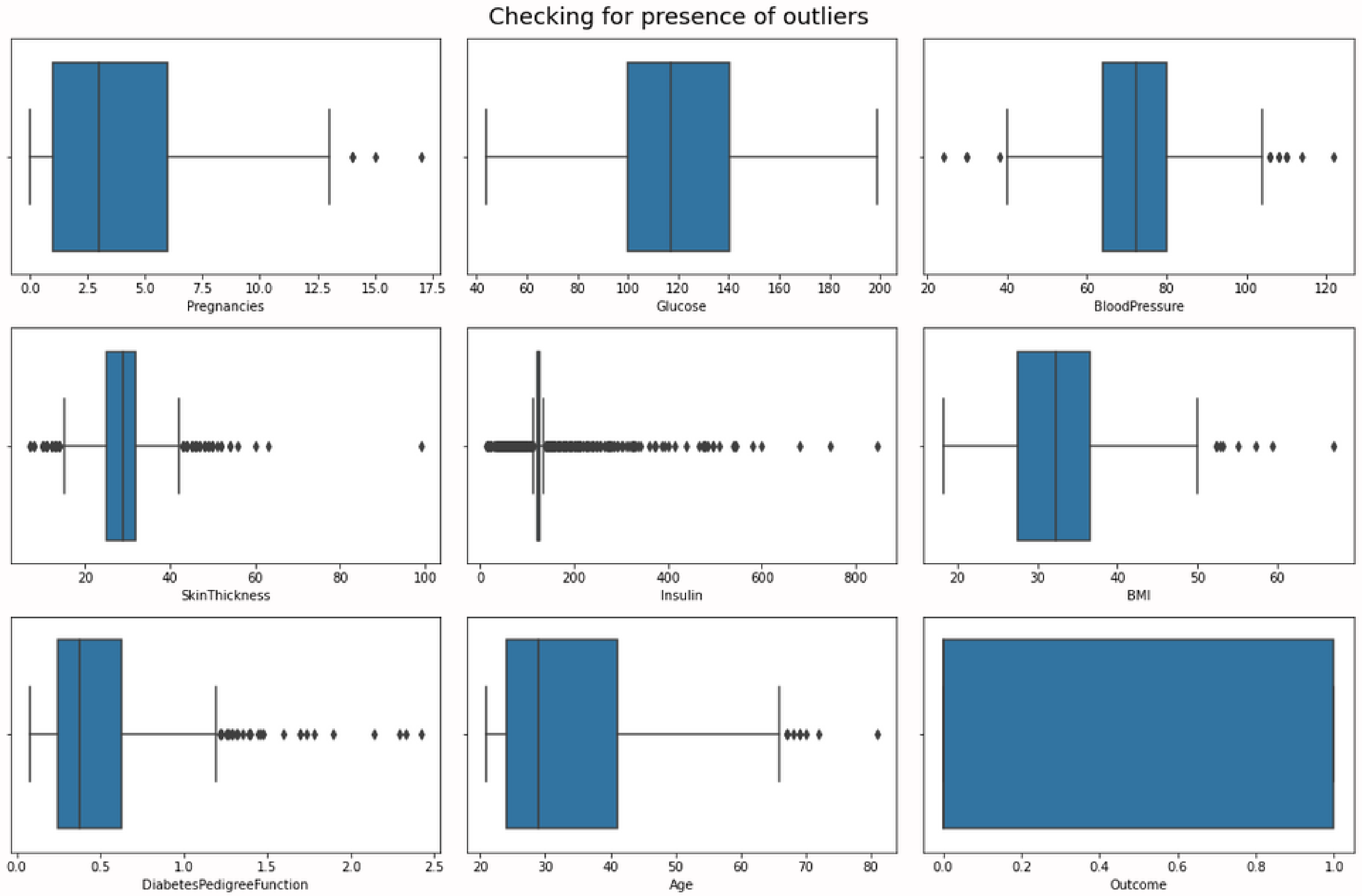
Features Distribution

This is the distributions of all features **after** imputation:



Checking for Outliers

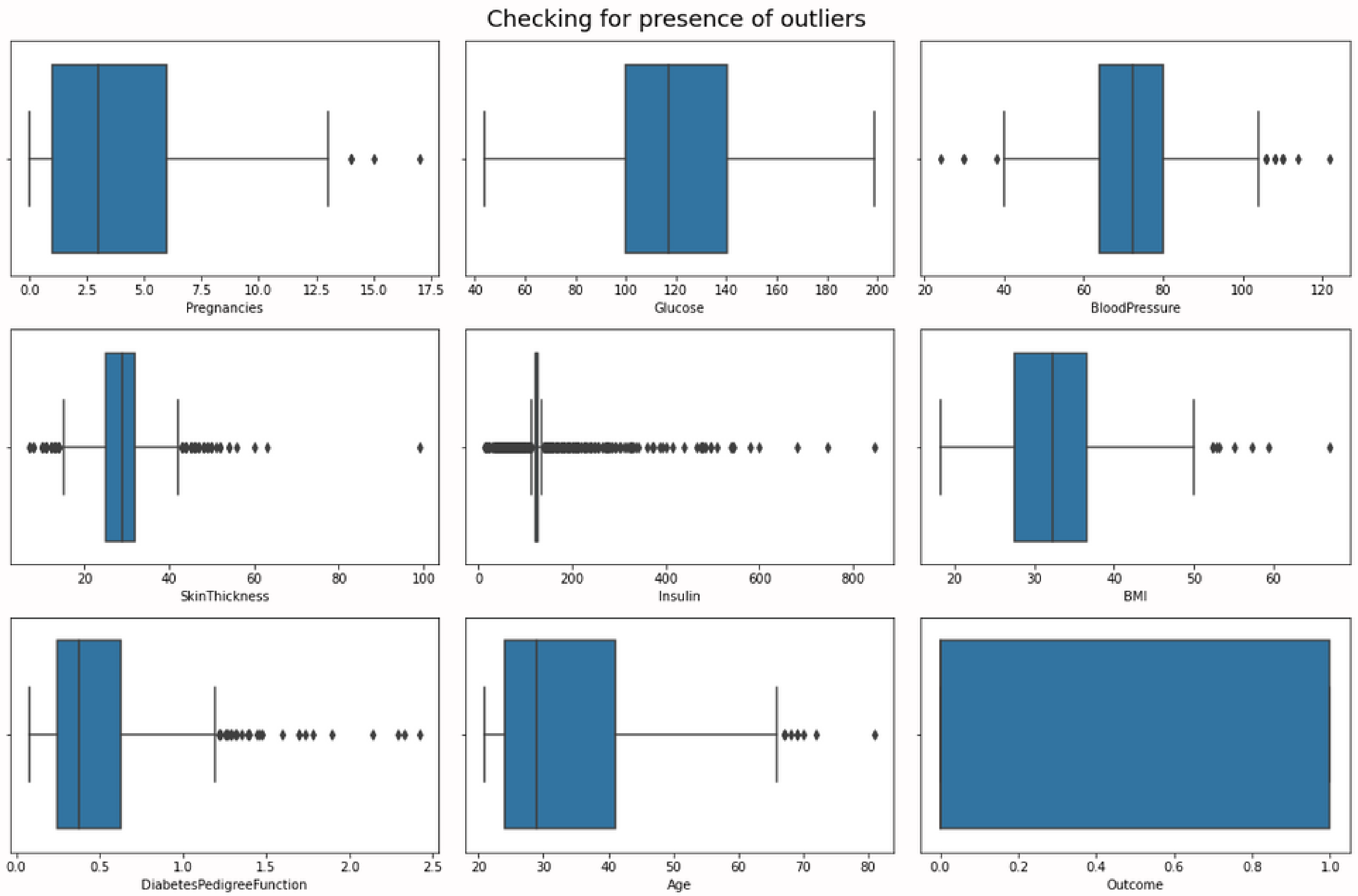
We then used Box plots to visualize the outliers in our dataset.

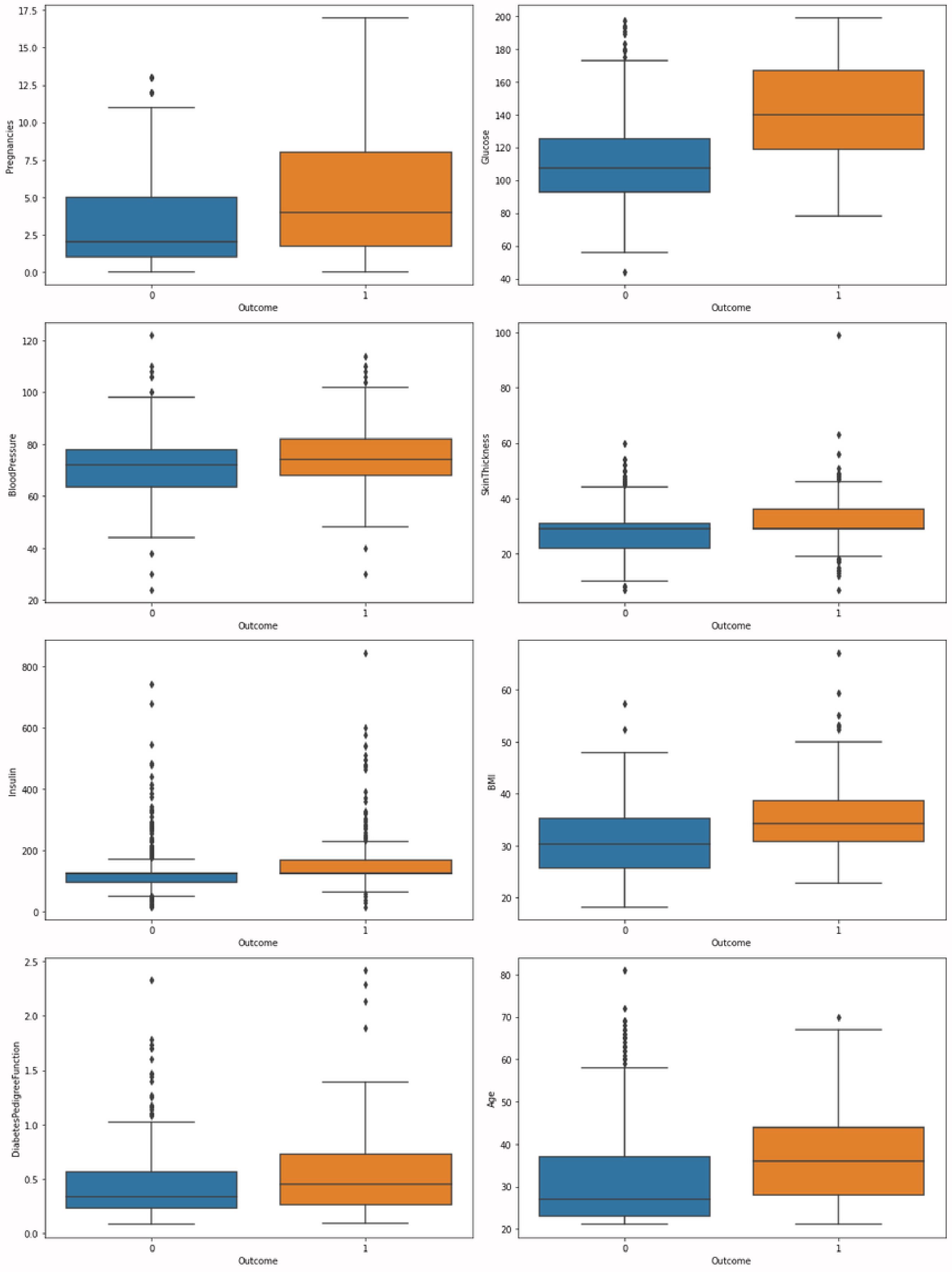


Checking for presence of outliers

Checking for Outliers

In our case, the outliers help improve the prediction accuracy of the logistic regression model, therefore we do not remove them.



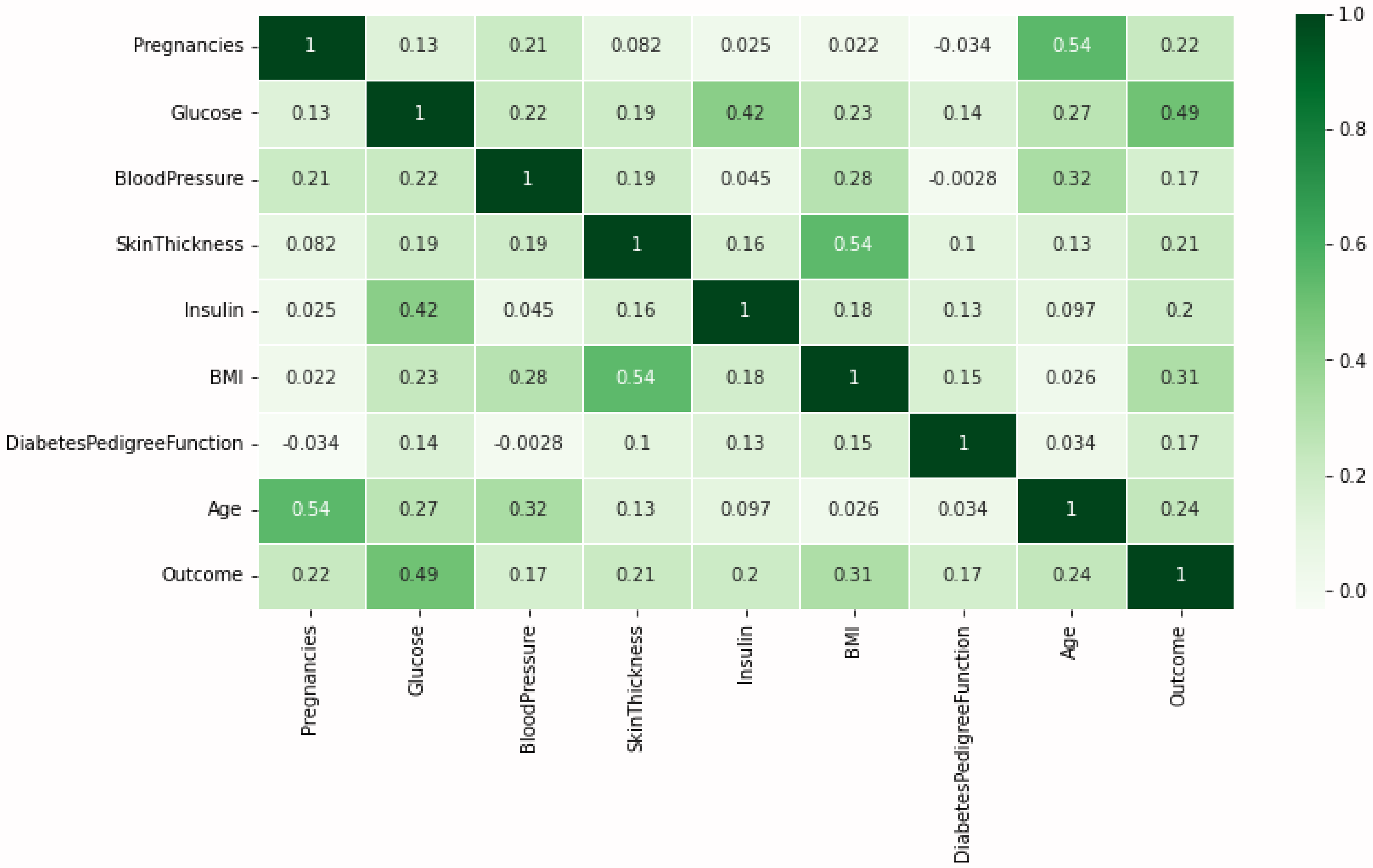


Predictor Features

We then plotted the predictor features against the dependent variable (Outcome) to check for correlations

Heatmap

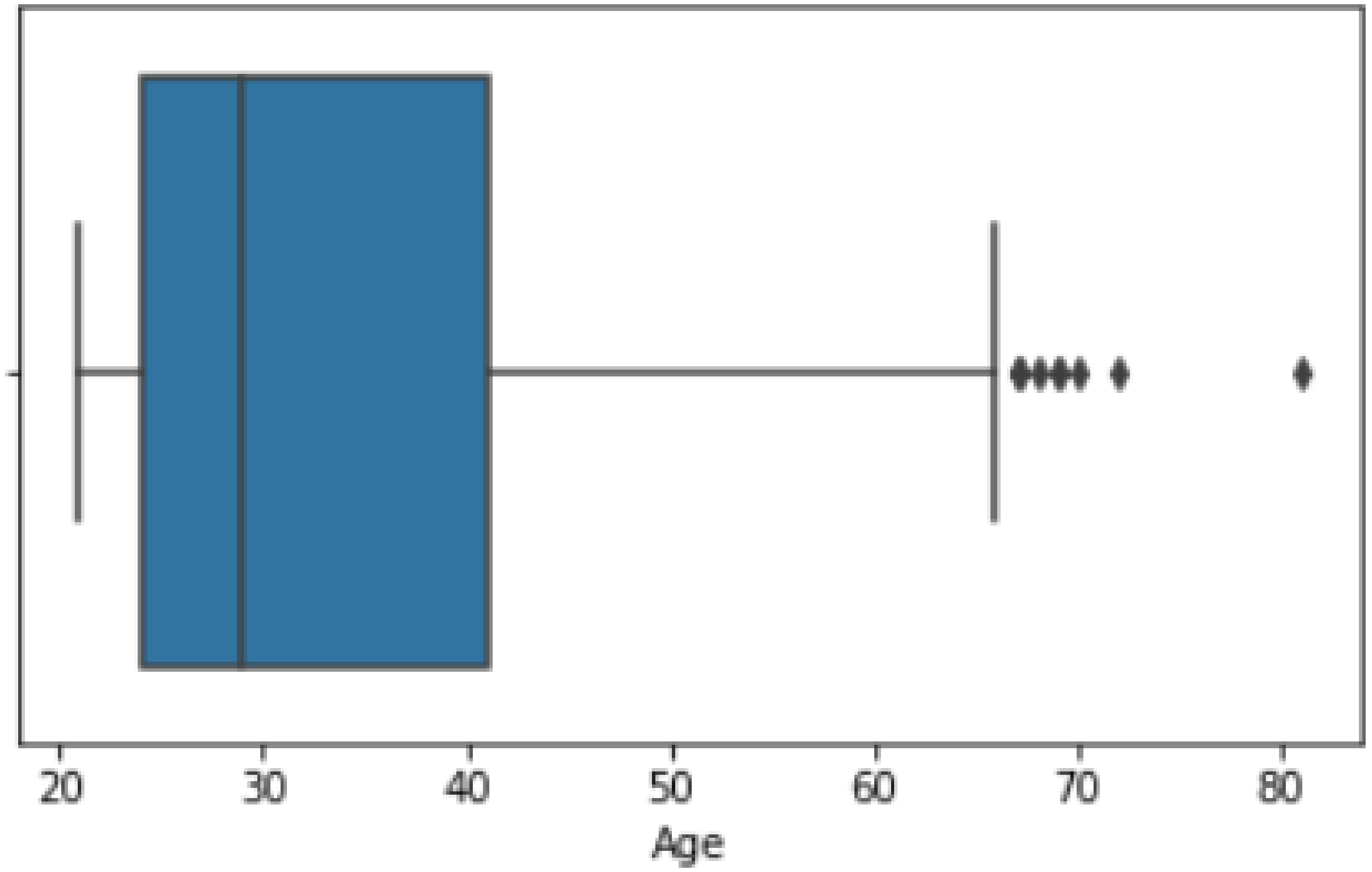
A heatmap of the correlation matrix:



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1	0.13	0.21	0.082	0.025	0.022	-0.034	0.54	0.22
Glucose	0.13	1	0.22	0.19	0.42	0.23	0.14	0.27	0.49
BloodPressure	0.21	0.22	1	0.19	0.045	0.28	-0.0028	0.32	0.17
SkinThickness	0.082	0.19	0.19	1	0.16	0.54	0.1	0.13	0.21
Insulin	0.025	0.42	0.045	0.16	1	0.18	0.13	0.097	0.2
BMI	0.022	0.23	0.28	0.54	0.18	1	0.15	0.026	0.31
DiabetesPedigreeFunction	-0.034	0.14	-0.0028	0.1	0.13	0.15	1	0.034	0.17
Age	0.54	0.27	0.32	0.13	0.097	0.026	0.034	1	0.24
Outcome	0.22	0.49	0.17	0.21	0.2	0.31	0.17	0.24	1

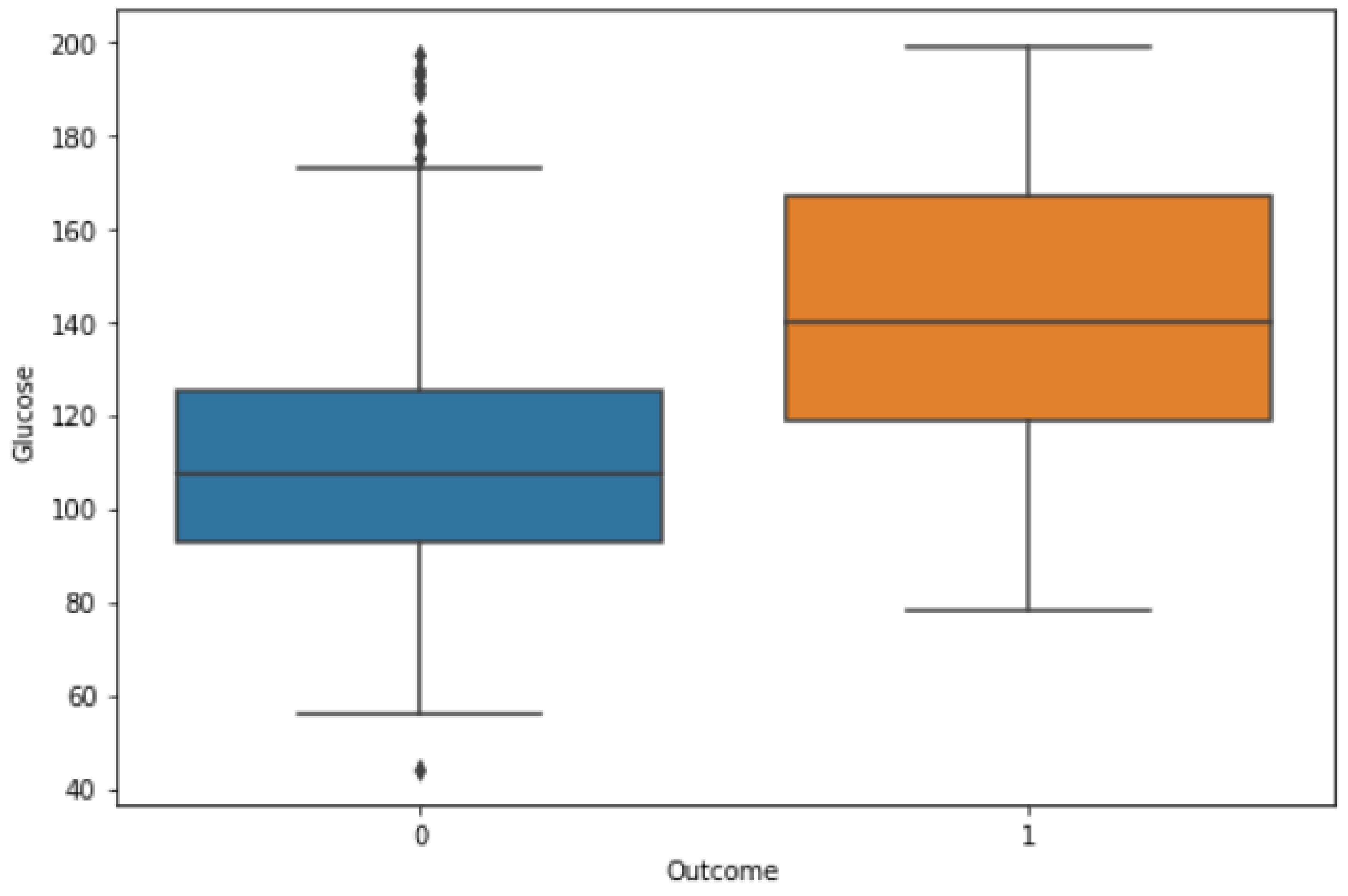
Major Findings

Outliers in terms of age, are usually women over 65.



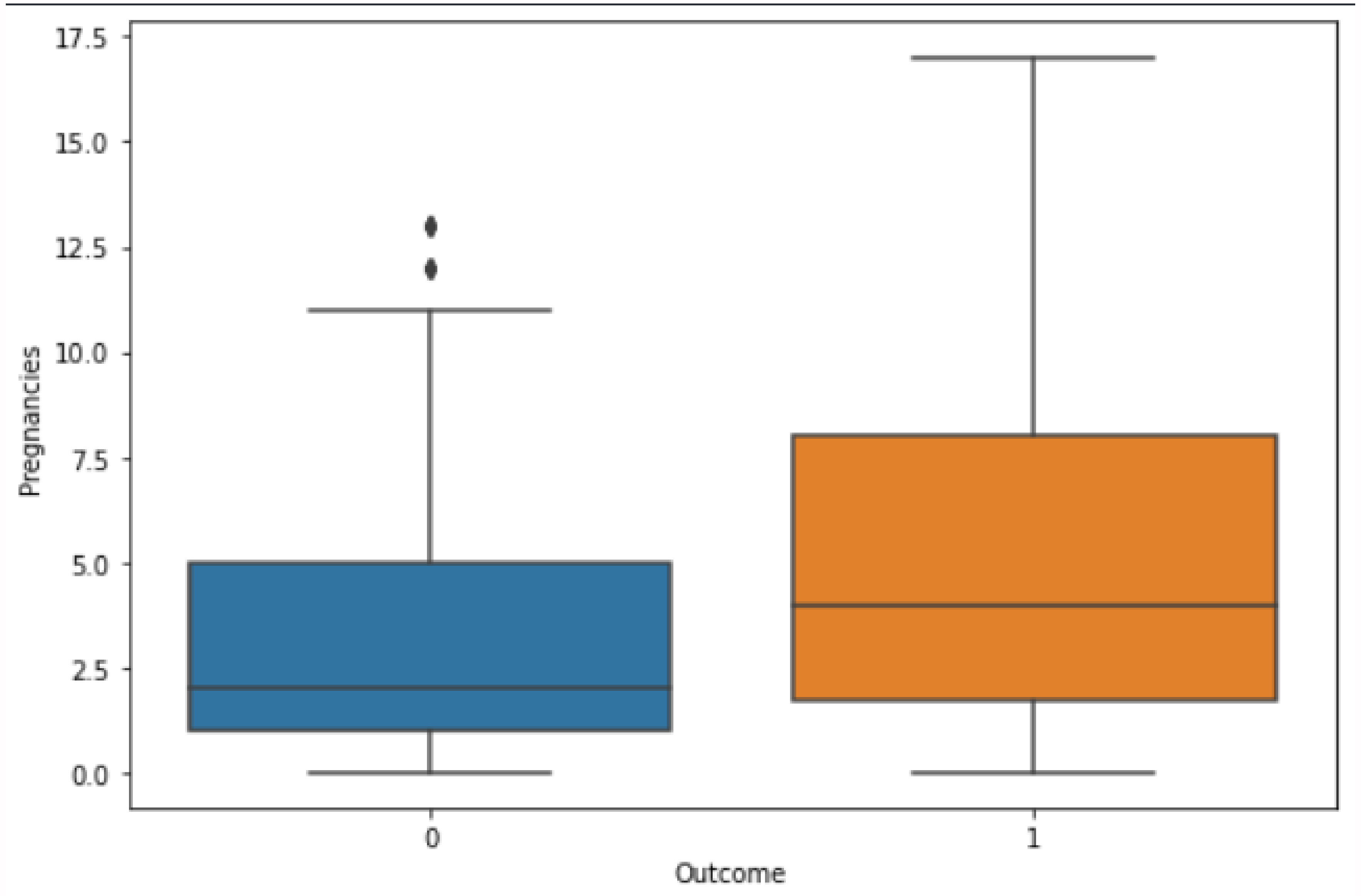
Major Findings

Glucose is a significant predictor for the outcome, especially positive cases.



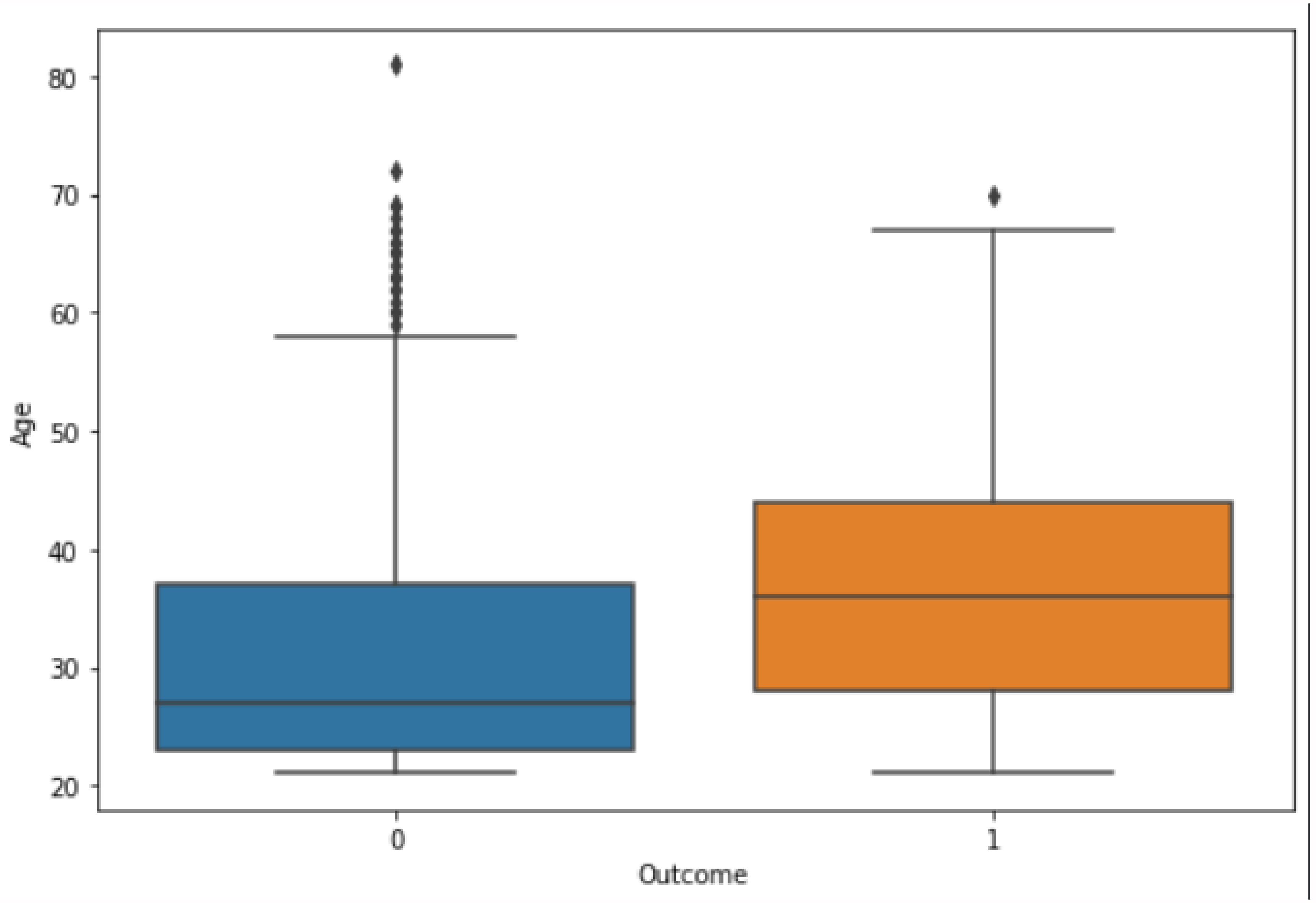
Major Findings

Number of pregnancies is a major predictor, especially when the number is high.



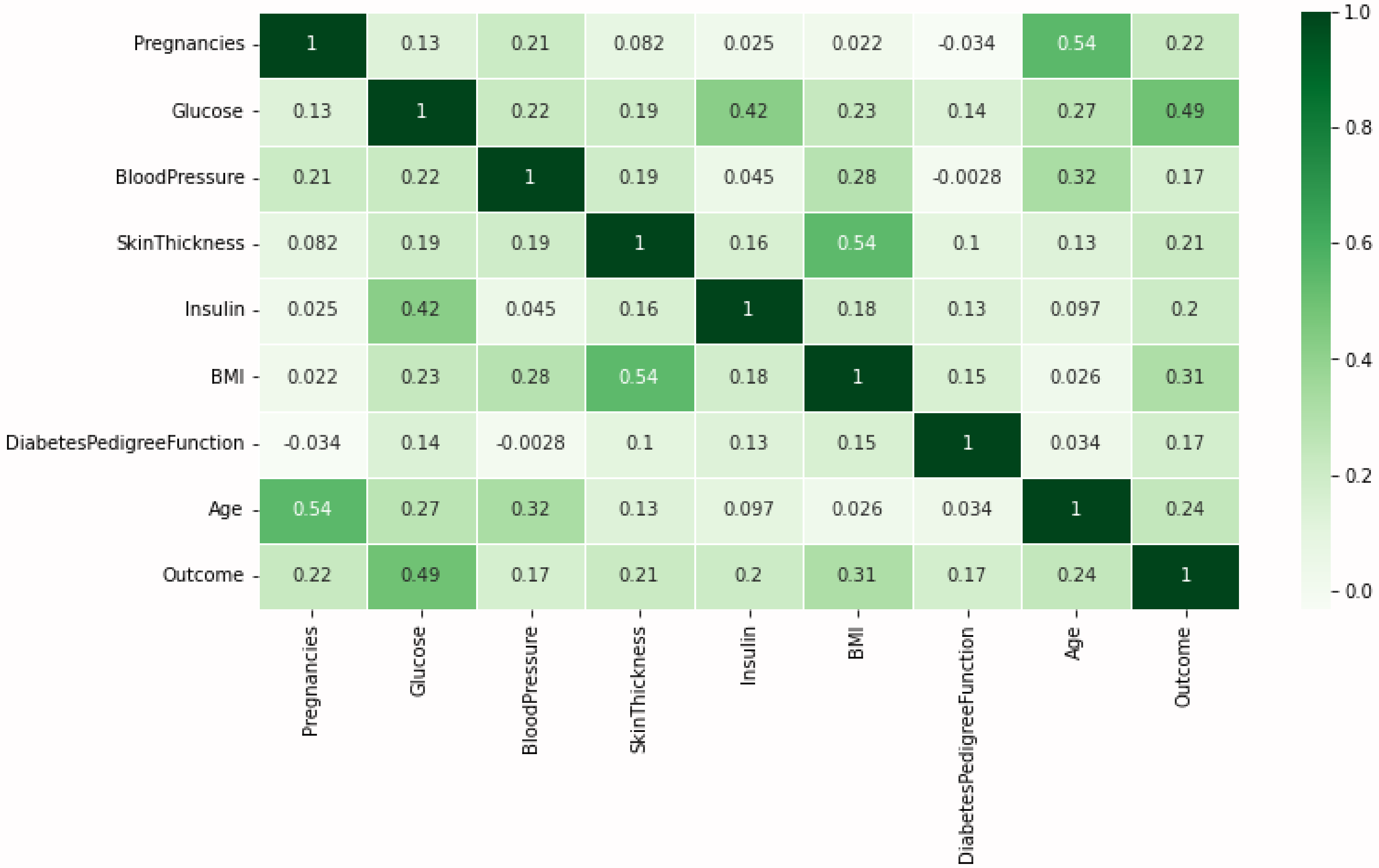
Major Findings

Age is a strong predictor. Women aged 38+ are more likely to be positive with an exception of numerous negative outliers.



Major Findings

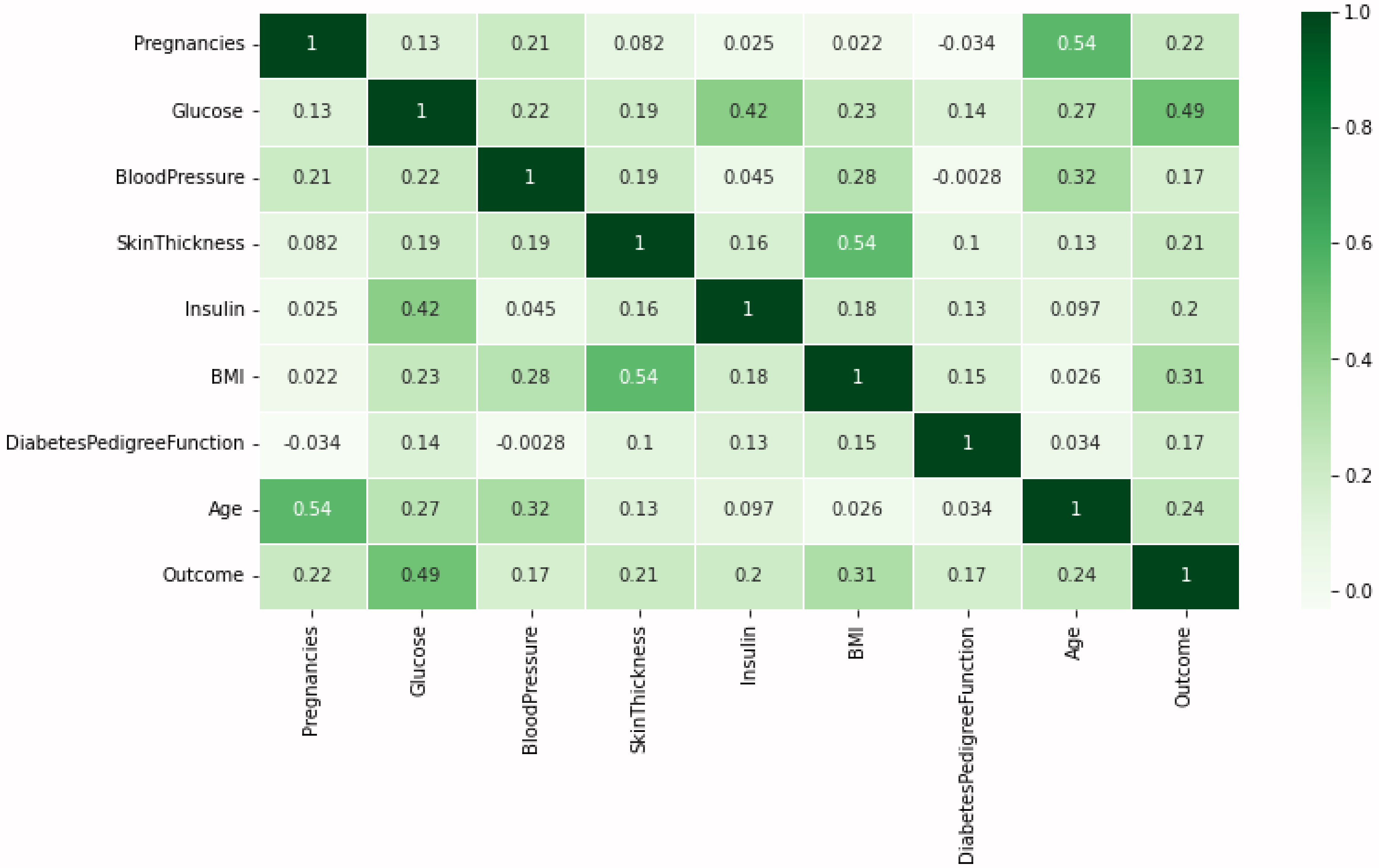
Skin thickness and BMI are **positively** correlated.



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1	0.13	0.21	0.082	0.025	0.022	-0.034	0.54	0.22
Glucose	0.13	1	0.22	0.19	0.42	0.23	0.14	0.27	0.49
BloodPressure	0.21	0.22	1	0.19	0.045	0.28	-0.0028	0.32	0.17
SkinThickness	0.082	0.19	0.19	1	0.16	0.54	0.1	0.13	0.21
Insulin	0.025	0.42	0.045	0.16	1	0.18	0.13	0.097	0.2
BMI	0.022	0.23	0.28	0.54	0.18	1	0.15	0.026	0.31
DiabetesPedigreeFunction	-0.034	0.14	-0.0028	0.1	0.13	0.15	1	0.034	0.17
Age	0.54	0.27	0.32	0.13	0.097	0.026	0.034	1	0.24
Outcome	0.22	0.49	0.17	0.21	0.2	0.31	0.17	0.24	1

Major Findings

Glucose and the outcome are **positively** correlated.



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1	0.13	0.21	0.082	0.025	0.022	-0.034	0.54	0.22
Glucose	0.13	1	0.22	0.19	0.42	0.23	0.14	0.27	0.49
BloodPressure	0.21	0.22	1	0.19	0.045	0.28	-0.0028	0.32	0.17
SkinThickness	0.082	0.19	0.19	1	0.16	0.54	0.1	0.13	0.21
Insulin	0.025	0.42	0.045	0.16	1	0.18	0.13	0.097	0.2
BMI	0.022	0.23	0.28	0.54	0.18	1	0.15	0.026	0.31
DiabetesPedigreeFunction	-0.034	0.14	-0.0028	0.1	0.13	0.15	1	0.034	0.17
Age	0.54	0.27	0.32	0.13	0.097	0.026	0.034	1	0.24
Outcome	0.22	0.49	0.17	0.21	0.2	0.31	0.17	0.24	1

Dataset Splitting

01

Step 1

Splitting the dataset into dependent and independent features

02

Step 2

Scaling the independent features

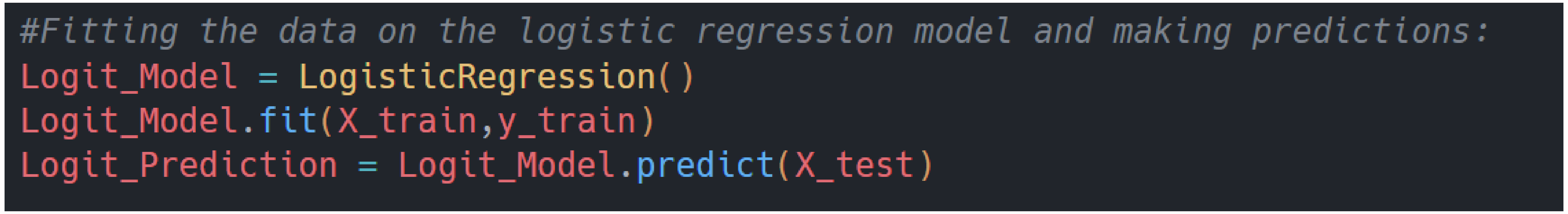
03

Step 3

Splitting the dataset into training and testing set

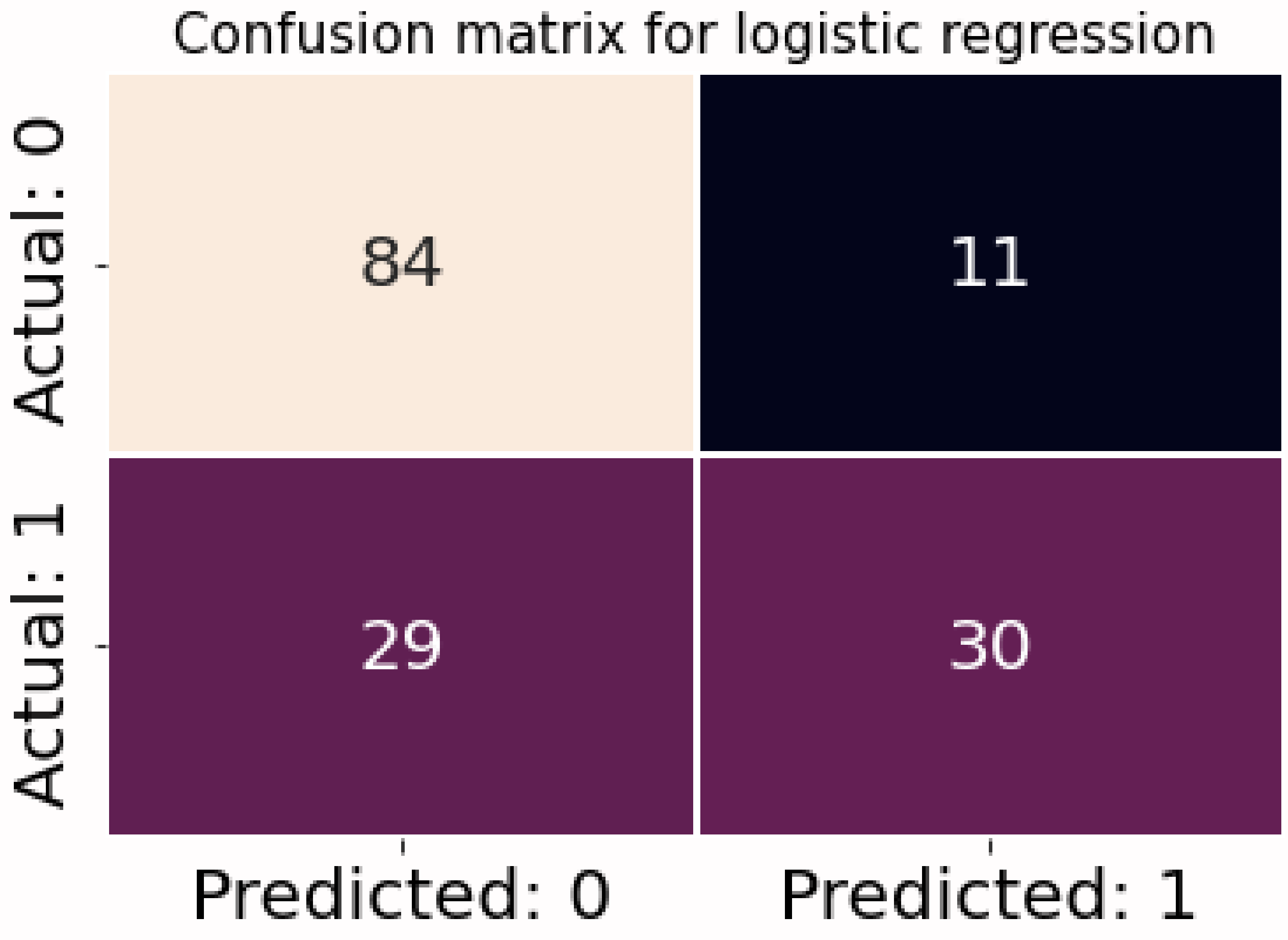
Modeling:

Since the dependent variable is binary in nature, ***logistic regression*** *would be a suitable model to train.*



```
#Fitting the data on the logistic regression model and making predictions:
Logit_Model = LogisticRegression()
Logit_Model.fit(X_train,y_train)
Logit_Prediction = Logit_Model.predict(X_test)
```

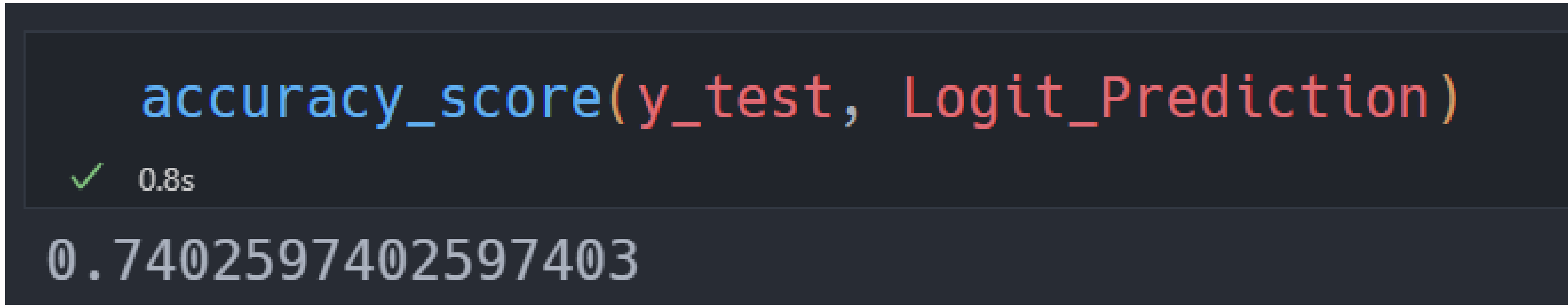
Confusion Matrix:



Confusion matrix for logistic regression

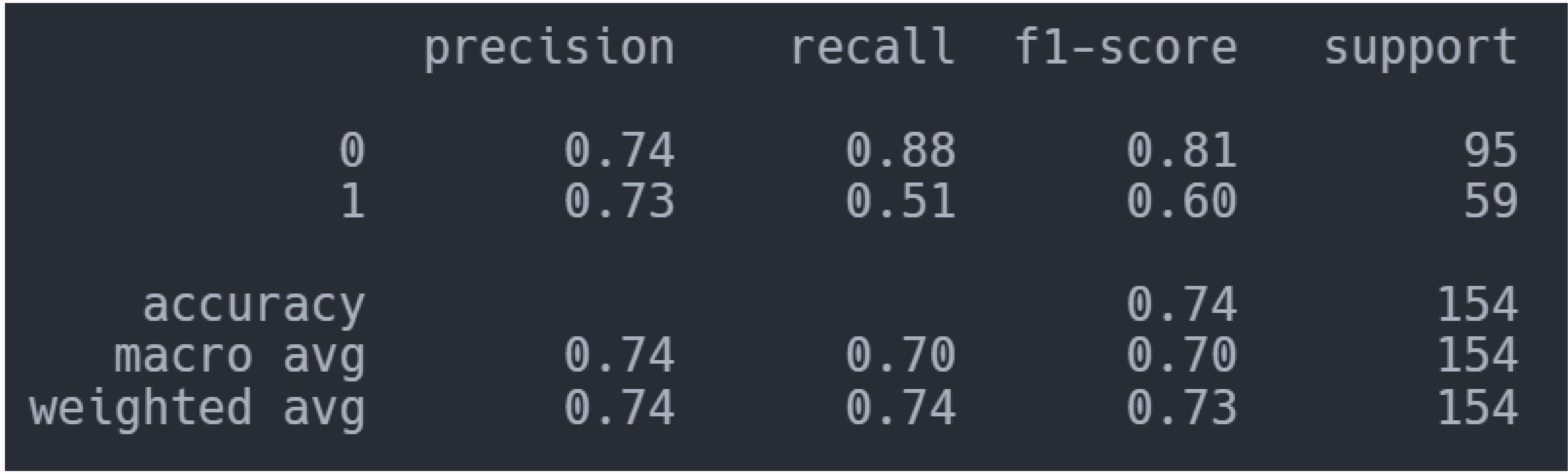
	Predicted: 0	Predicted: 1
Actual: 0	84	11
Actual: 1	29	30

Accuracy Score:



```
accuracy_score(y_test, Logit_Prediction)
✓ 0.8s
0.7402597402597403
```

Classification Report:



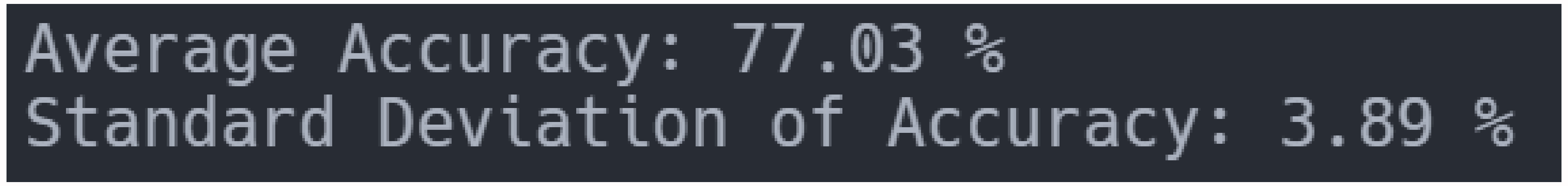
```
              precision    recall  f1-score   support

           0       0.74      0.88      0.81        95
           1       0.73      0.51      0.60        59

    accuracy                           0.74       154
   macro avg       0.74      0.70      0.70       154
weighted avg       0.74      0.74      0.73       154
```

K-Fold Cross Validation:

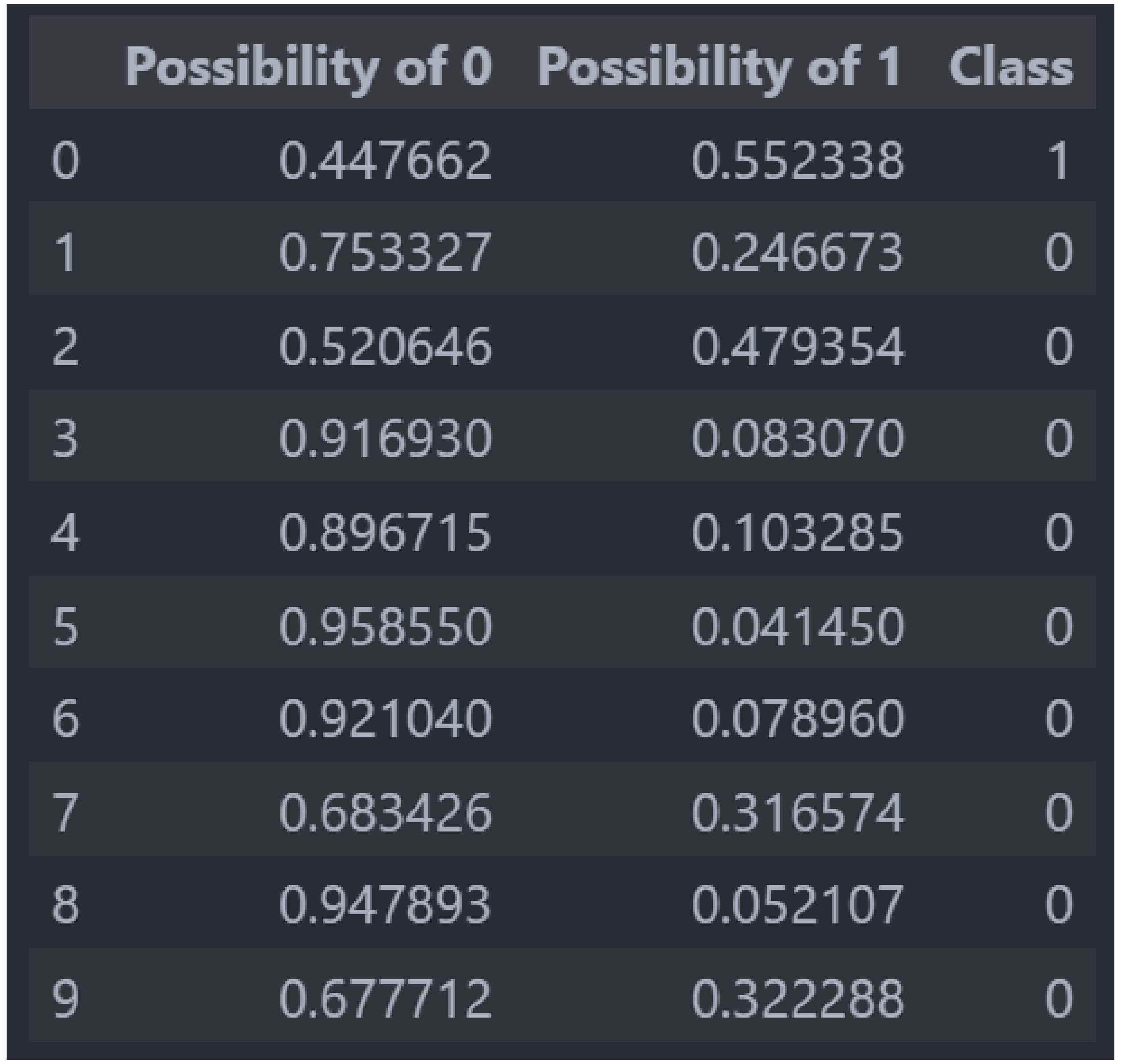
By using the K-Fold cross validation technique, we can see that the average accuracy of the logistic regression model is about 77.03% with a 3.89% standard deviation.



```
Average Accuracy: 77.03 %
Standard Deviation of Accuracy: 3.89 %
```

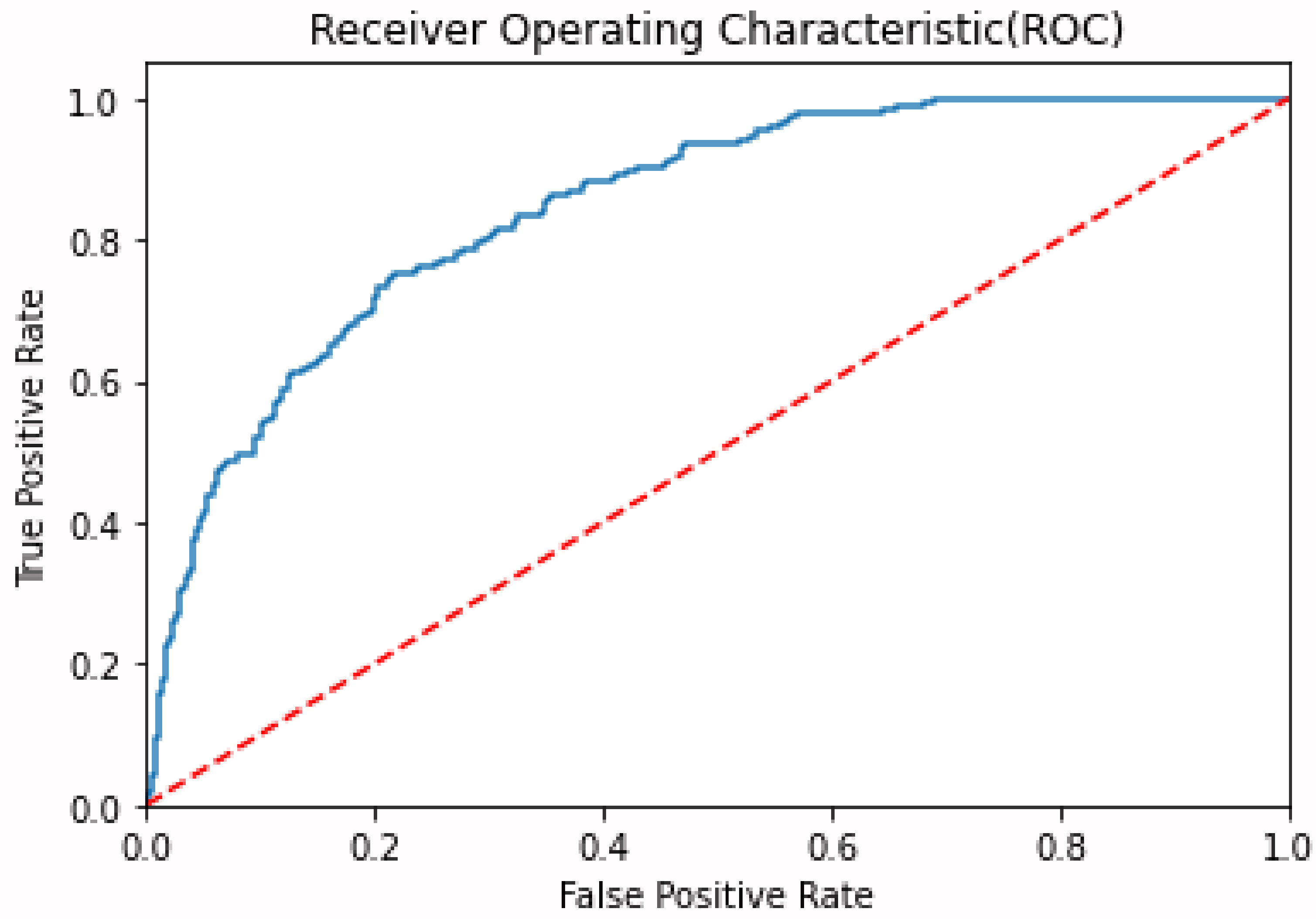
Test Data

Some outputs of applying the logistic regression model on the test data:



	Possibility of 0	Possibility of 1	Class
0	0.447662	0.552338	1
1	0.753327	0.246673	0
2	0.520646	0.479354	0
3	0.916930	0.083070	0
4	0.896715	0.103285	0
5	0.958550	0.041450	0
6	0.921040	0.078960	0
7	0.683426	0.316574	0
8	0.947893	0.052107	0
9	0.677712	0.322288	0

Receiver Operating Characteristic



THANK YOU

References:

Pima Indians Diabetes Dataset: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database