Prediction of Diabetes in PIMA Indians

NYU Data Science Bootcamp Project

Bhairavi, Mihir, Rugved, Sneha - November 30, 2023

Introduction

Who are Pima Indians?

- Pima Indians are a Native American tribe primarily residing in Arizona, United States, and parts of Mexico.
- Pima community has faced significant challenges, particularly related to health and socioeconomic factors. They have experienced high rates of diabetes and other health issues, which have been the subject of extensive study and research aimed at understanding the prevalence and causes of diabetes within this population.

Introduction

Purpose of the Analysis

- We utilize the dataset to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset.
- This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases.

About the Dataset

Columns

- **Pregnancies**: Number of times pregnant
- **Glucose**: Plasma glucose concentration after 2 hours in an oral glucose tolerance test
- **Blood Pressure**: Diastolic blood pressure (mm Hg)
- **Skin Thickness**: Triceps skinfold thickness (mm)
- **Insulin**: 2-Hour serum insulin (mu U/ml)

About the Dataset

Columns (Contd.)

- **BMI**: Body mass index (weight in kg/(height in m)^2)
- **Diabetes Pedigree Function**: Diabetes pedigree function (a function that scores likelihood of diabetes based on family history)
- **Age**: Age in years
- **Outcome**: Class variable (0 or 1), where 1 indicates diabetes presence and 0 indicates absence

Data Cleaning

Handling Missing Values

```
dataset.describe()
```

	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 Cleaning

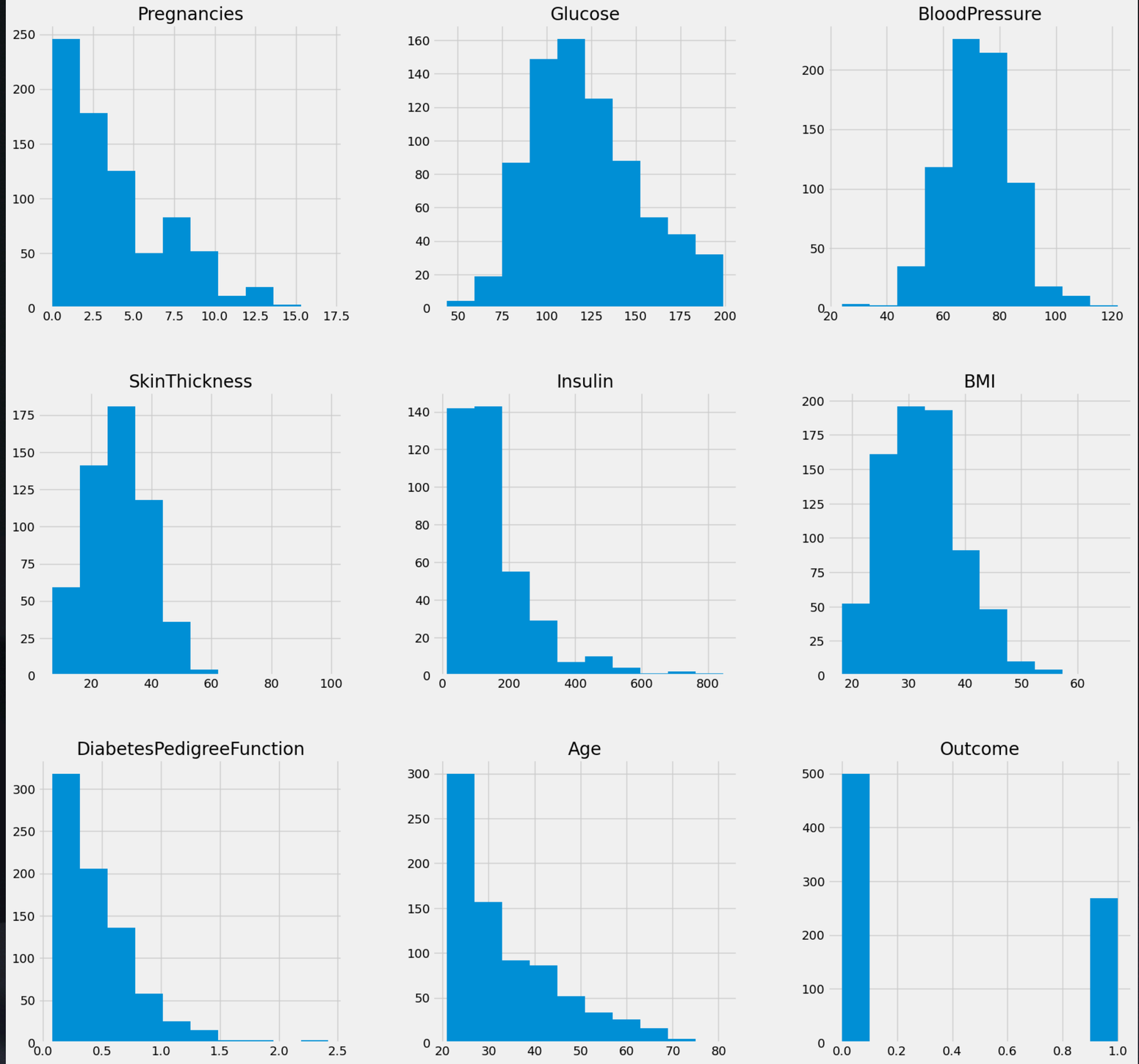
Handling Missing Values

- In our analysis of the Pima Dataset, notable observations reveal that the minimum values recorded for 'Glucose,' 'BloodPressure,' 'SkinThickness,' 'Insulin,' and 'BMI' are zero, which inherently contradicts physiological plausibility. This suggests the presence of missing or invalid entries, given that these measurements cannot realistically be zero in a human context.

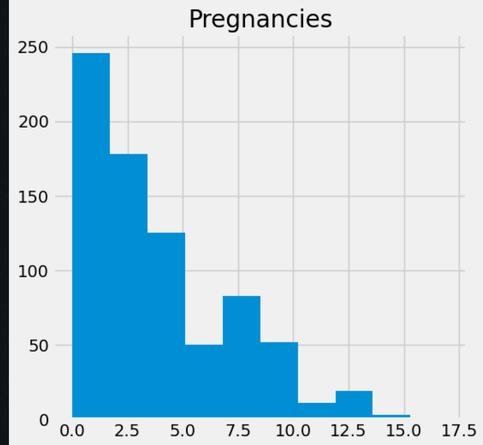
- To facilitate improved predictive accuracy and model reliability, it is imperative to address these zero values by employing suitable imputation methods to ensure the dataset's integrity and the efficacy of subsequent predictive modeling.

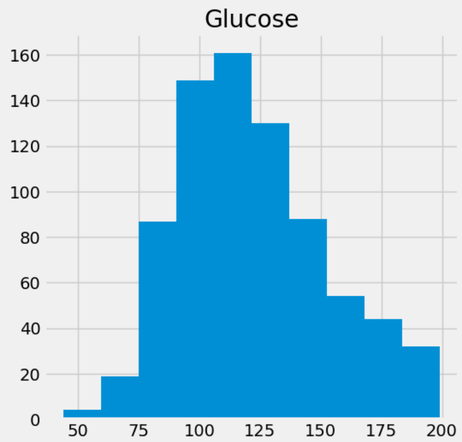
```
Pregnancies                 0
Glucose                     5
BloodPressure              35
SkinThickness             227
Insulin                   374
BMI                        11
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
```

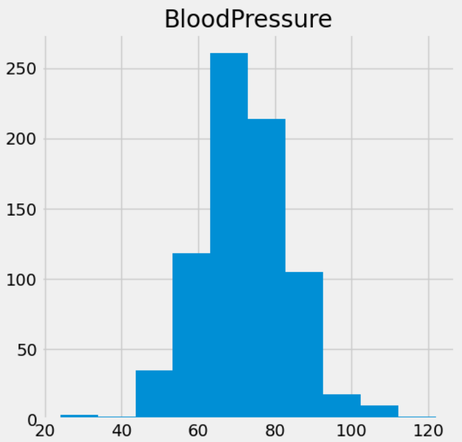
- In order to address these missing values, it is essential to comprehend the distribution characteristics inherent within the dataset.

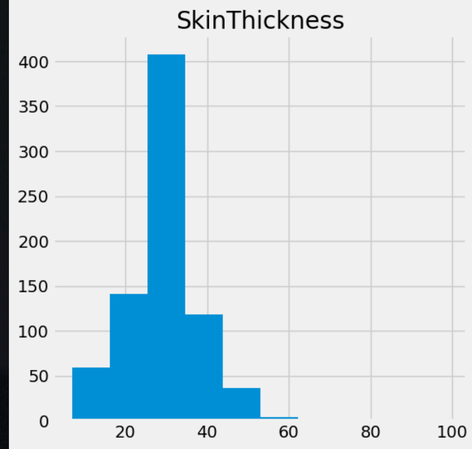


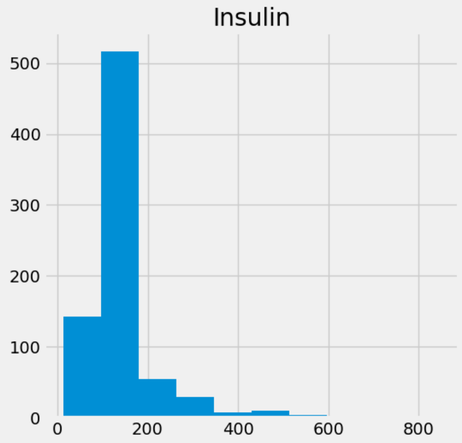
- We've chosen to impute missing values differently based on the distribution characteristics of the variables.
- For 'Glucose' and 'BloodPressure,' which exhibit a normal distribution, we've opted to replace missing values with their respective means. However, considering the right-skewed distributions observed in 'SkinThickness,' 'Insulin,' and 'BMI,' we've used the mean as the replacement value for their missing entries.

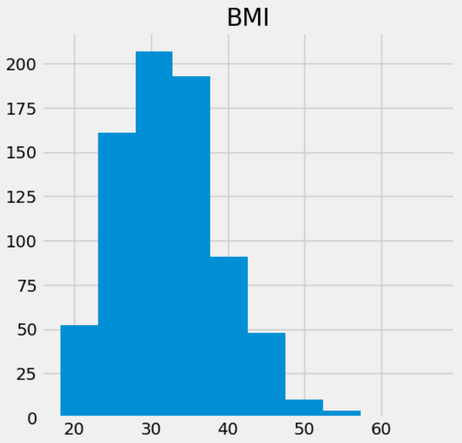


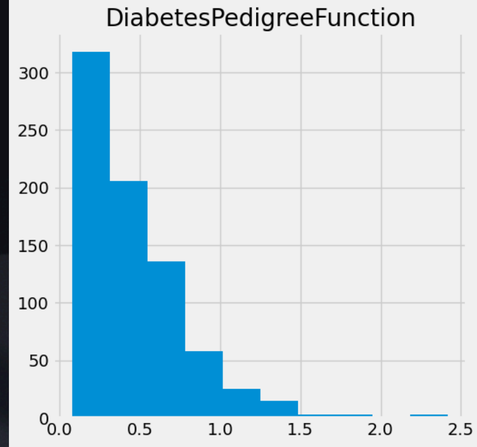


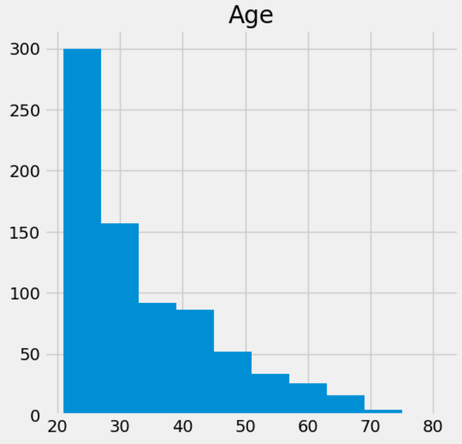


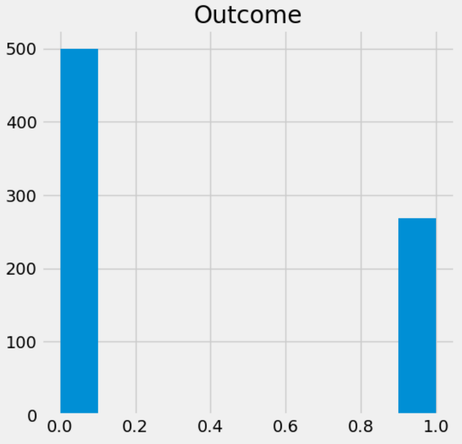








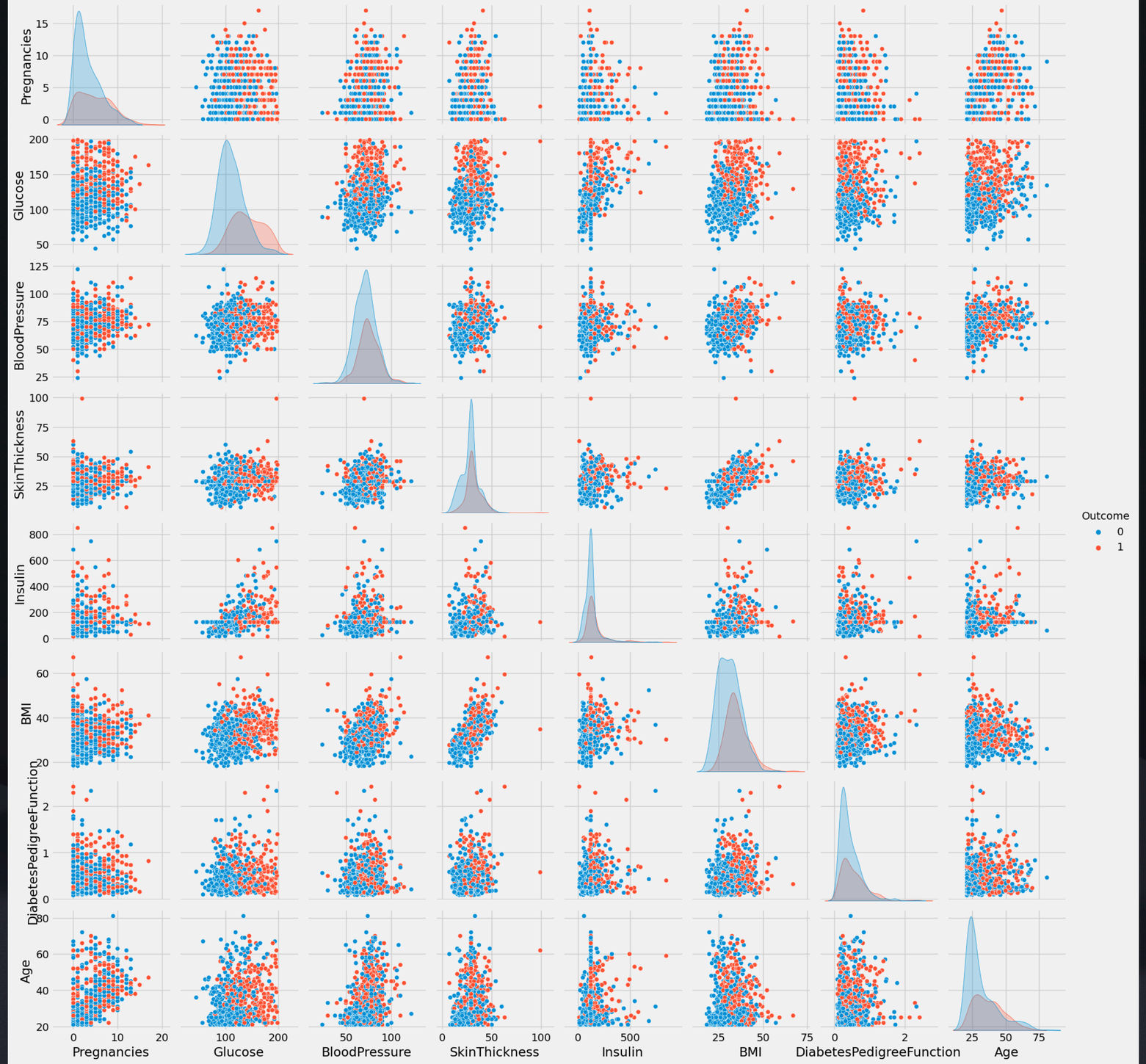


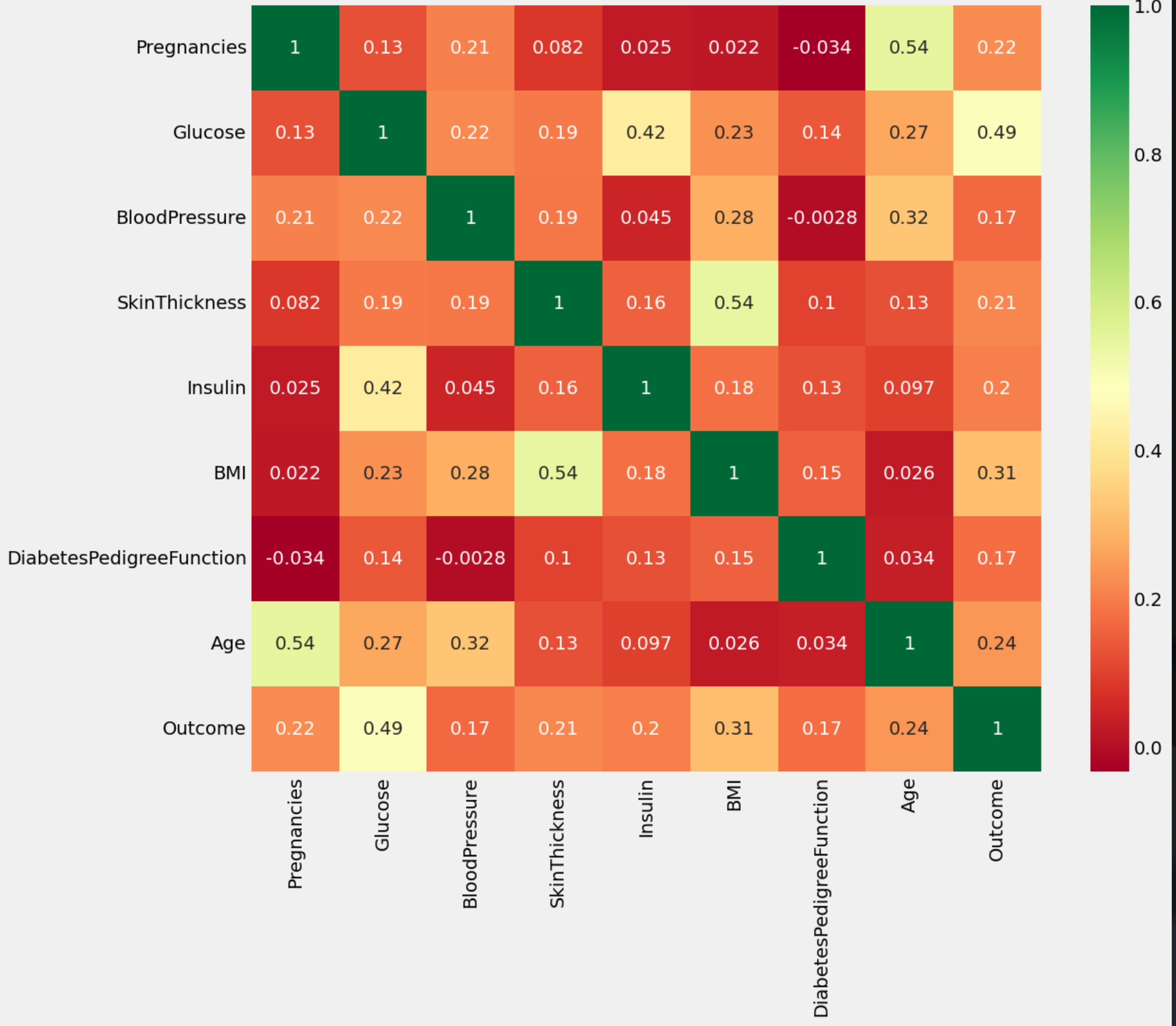


Exploratory Data Analysis (EDA)

Data Visualization

- Engaging in Exploratory Data Analysis (EDA) allows us to intimately understand the nuances within our dataset before delving into deeper analysis.
- During this phase, we explore the data's fundamental characteristics, assessing its structure, size, and the nature of variables at hand.


Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age
Outcome
0
1


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
Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age
Outcome
1.0
0.8
0.6
0.4
0.2
0.0

Exploratory Data Analysis (EDA)

Data Visualization

- EDA highlights a strong correlation between diabetic cases and elevated glucose levels, indicating the crucial role of glucose values in diabetes.
- High BMI shows a notable correlation with the presence of diabetes, emphasizing its relevance in understanding diabetic conditions.
- Age emerges as the third most correlated factor among individuals affected by diabetes, signifying its significance in association with this health condition.

Modelling

Scaling and Splitting Dataset

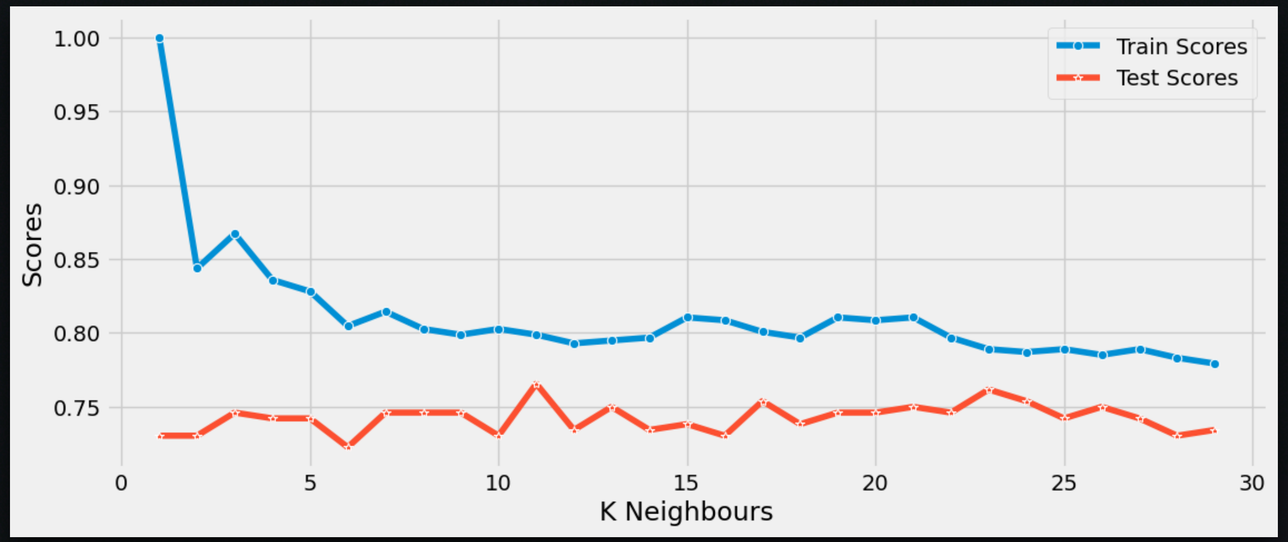
- Employed the StandardScaler module from the ‘sklearn’ library to standardize the dataset, ensuring uniformity in feature scales for optimal model performance.
- Utilized the train-test split function available in the library to effectively partition the dataset into separate training and testing subsets, crucial for model training and evaluation processes.

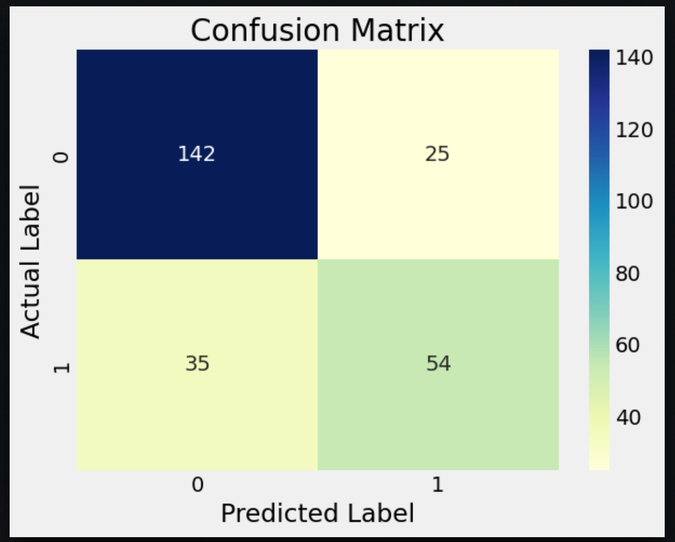
Model Training and Evaluation

K-Nearest Neighbors

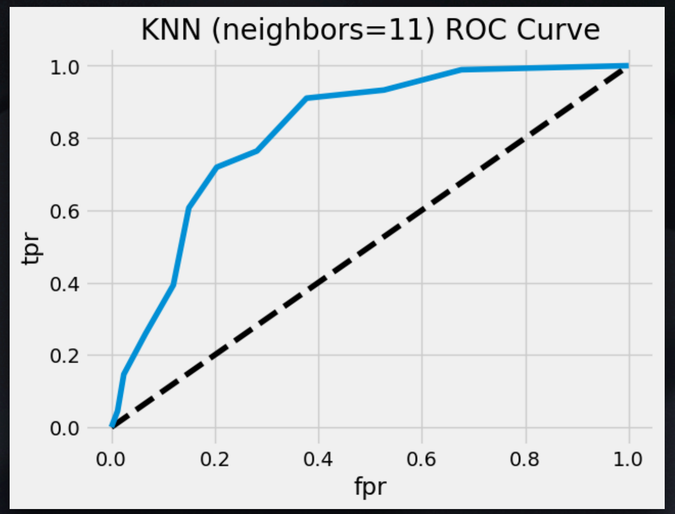
- K-Nearest Neighbors (KNN) model is selected for our dataset analysis due to its simplicity, making it suitable for initial exploration and intuitive understanding.
- Effective for both binary and multi-class classification, KNN accommodates the PIMA dataset's goal of categorizing individuals as diabetic or non-diabetic.
- Its ability to capture nonlinear relationships in data aligns with the potentially complex nature of the PIMA dataset.
- Utilizing the KNN model enables an exploration of the dataset’s intricacies, allowing us to comprehend its underlying patterns and relationships effectively while serving as a starting point for further model comparisons and analysis.

- At the K value of **11** within our K-Nearest Neighbors (KNN) model, we achieve the peak accuracy on our test dataset **76.56%**





- Precision Score : 68.35%
- Recall Score : 60.67%
- ROC AUC Score : 81.94%



Conclusion

- Our analysis of the dataset culminated in the utilization of a K-Nearest Neighbors (KNN) model, yielding an accuracy rate of 76.56%.
- This outcome serves as a foundational benchmark, indicating a substantial predictive capability for diabetes within the dataset.
- However, recognizing the potential for further accuracy enhancement, we acknowledge the necessity of exploring more intricate and sophisticated models beyond the scope of KNN.
- Throughout this analysis, we've gleaned insightful correlations and patterns crucial to understanding predictors of diabetes within the PIMA population. Moving forward, we recommend continued exploration involving advanced modeling techniques and meticulous refinement of feature engineering to fortify predictive accuracy.

Thank You!