



Diabetes Supervised Learning project

Nasser Othman





Agenda

Introduction

Primary goals

Areas of growth

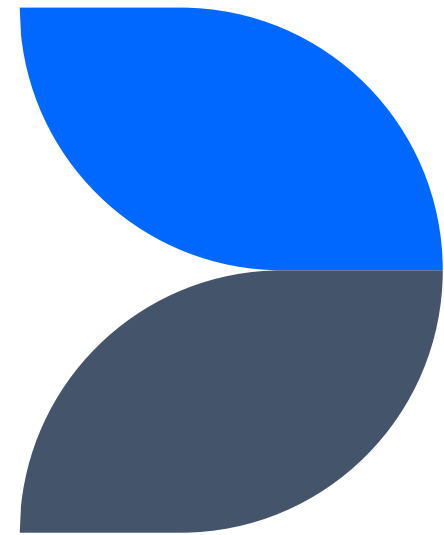
Timeline

Summary

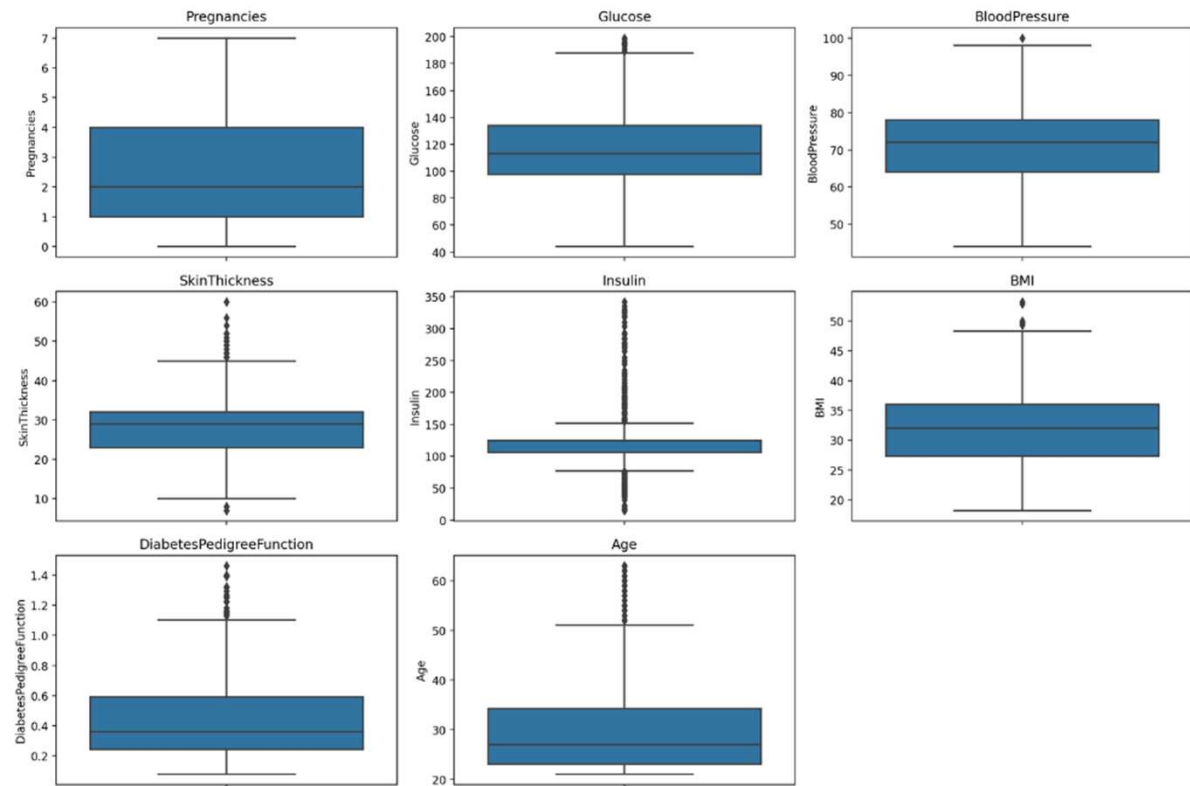
Introduction

Build a machine learning model that can predict whether a patient has diabetes or not, based on certain diagnostic measurements. This project involves three main parts: exploratory data analysis, preprocessing and feature engineering, and training a machine learning model.

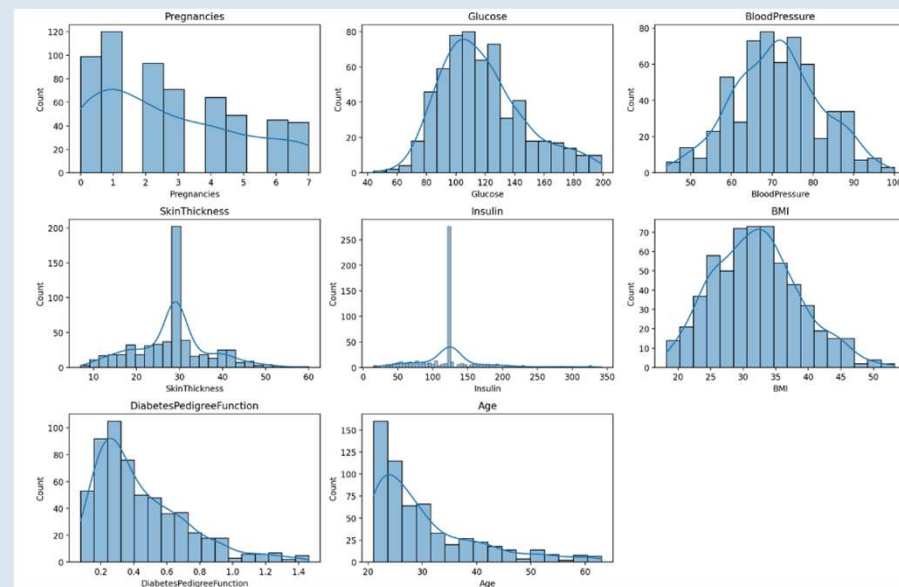
Exploratory data analysis



Boxplot to detect outliers

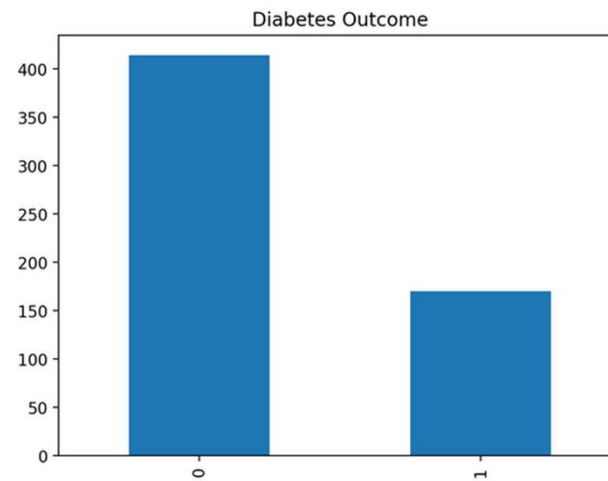


Distribution of Data



Diabetes outcome

```
0    414
1    170
Name: Outcome, dtype: int64
We can see that there are 414 individuals without diabetes and 170 with diabetes
```



Results of our model

```
Logistic Regression:  
Accuracy: 0.811965811965812  
Precision: 0.85  
Recall: 0.4722222222222222  
F1-score: 0.6071428571428571  
ROC-AUC: 0.8597393689986282
```

```
Random Forest:  
Accuracy: 0.811965811965812  
Precision: 0.8181818181818182  
Recall: 0.5  
F1-score: 0.6206896551724137  
ROC-AUC: 0.8655692729766804
```

Both models show reasonably good performance, with the Logistic Regression model slightly outperforming the Random Forest in terms of accuracy, precision, and F1-score. However, the Random Forest model has a slightly higher recall value, indicating its ability to correctly identify positive instances (diabetes) compared to the Logistic Regression model. The ROC-AUC values for both models are relatively high, indicating that both models have good discriminatory power to distinguish between the positive and negative classes.



Thank you

Nasser Othman

