Supervised learning Project

by

Cynthia Okaja

22nd August 2023

LIGHTHOUSE LABS

Project Goals

To use supervised learning techniques to build a machine learning model that can predict whether a patient has diabetes or not.

Project Execution

- Exploratory Data Analysis
- Data Preprocessing
- Modelling

EDA

- Getting familiar with the data

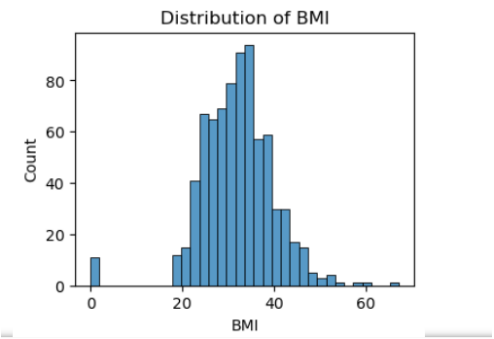
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	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

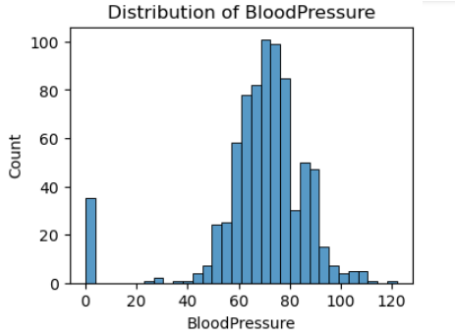
- Info

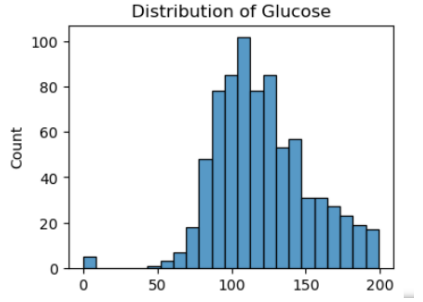
```
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
```

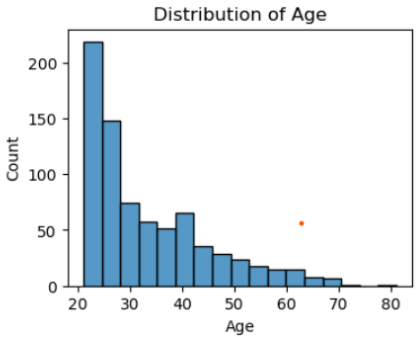
EDA

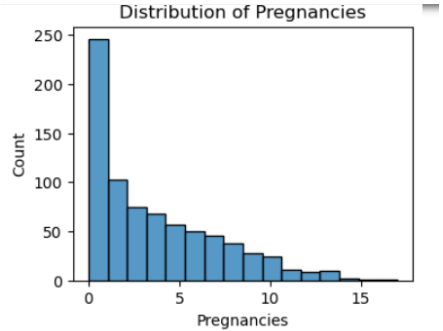
Histogram of predictor variables

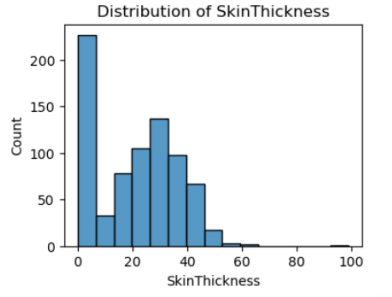






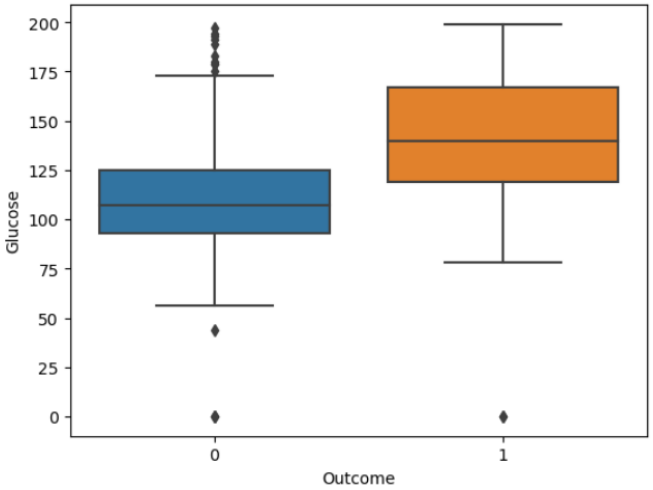




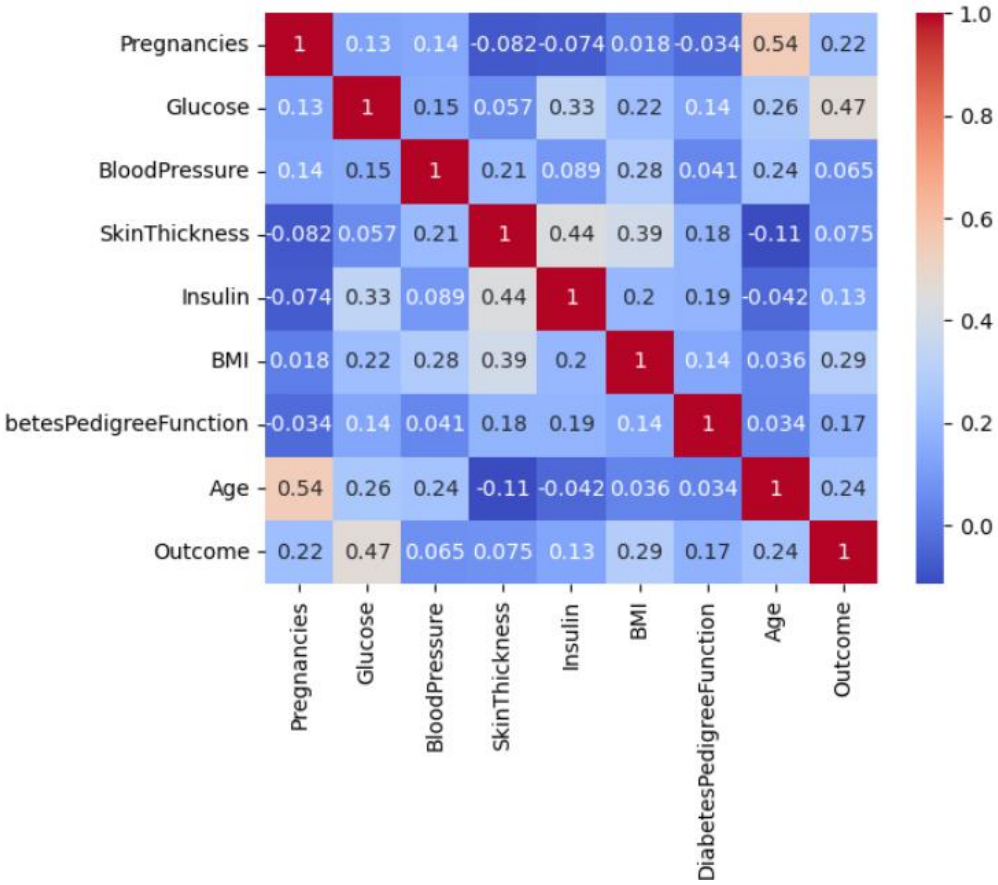


EDA

- Identifying outliers based on the target



Most people with glucose level more than 125 have diabetes, apart from a few outliers observed between 175 - 200



Heatmap

A strong correlation between the variables cannot be established. Age and Pregnancies have the highest correlation value of 0.54

Data Preprocessing

- Replaced 0 with the mean of some columns

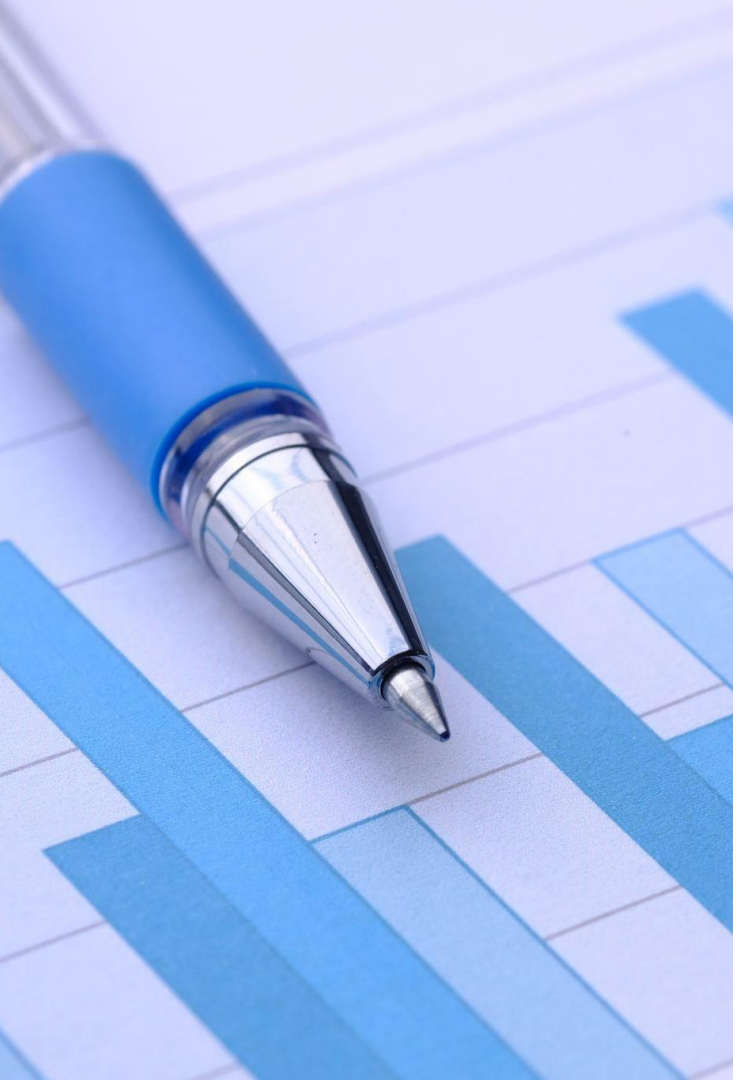
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6.000000	148.0	72.0	35.000000	79.799479	33.6	0.627	50	1
1	1.000000	85.0	66.0	29.000000	79.799479	26.6	0.351	31	0
2	8.000000	183.0	64.0	20.536458	79.799479	23.3	0.672	32	1
3	1.000000	89.0	66.0	23.000000	94.000000	28.1	0.167	21	0
4	3.845052	137.0	40.0	35.000000	168.000000	43.1	2.288	33	1

- Remove outliers in Insulin and Diabetes Pedigree Function
- Scaled and Normalized the predictor variables

Modelling

- Split dataset into Train and Test
- Performed Logistic Regression
- Performed Gradient Boosting Classifier
- Results:

Model Evaluation	Logistic Regression	Gradient Boosting Classifier
Accuracy	0.7987012987012987	0.7597402597402597
Precision	0.7380952380952381	0.64
Recall	0.6078431372549019	0.6274509803921569
F1_score	0.6666666666666666	0.6336633663366336
roc_auc	0.7505235103750239	0.7263468494193793



Conclusion

Both models have their strengths and weaknesses:

Logistic Regression has a higher accuracy, precision, and ROC-AUC score, indicating better overall performance in terms of correctly predicting positive cases and distinguishing between classes.

Gradient Boosting has a slightly higher recall, meaning it's slightly better at identifying actual positive cases, but it sacrifices some precision compared to Logistic Regression.