Dataset used

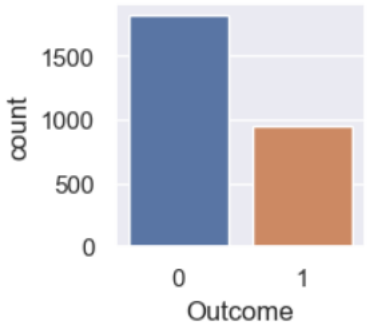
https://www.kaggle.com/datasets/nanditapore/healthcare-diabetes

Data Analysis and Visualization

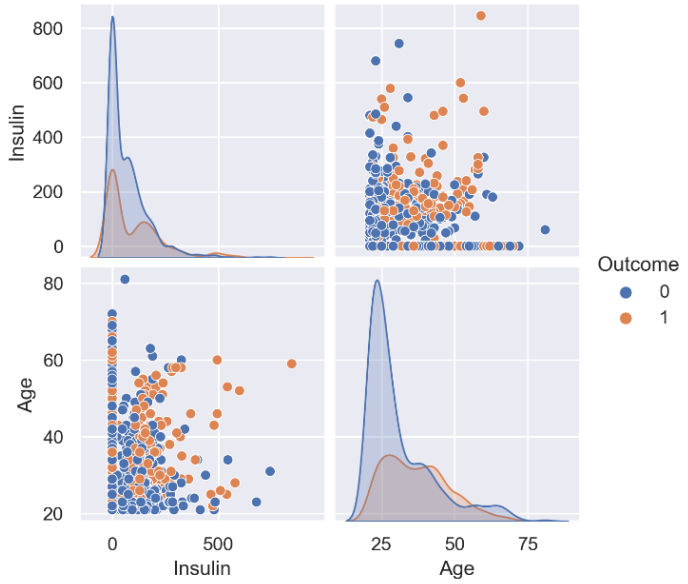
This dataset contains a range of attributes that can be used to predict whether an individual has diabetes. The outcome column in the dataset indicates the final diagnosis of the patients.

Our analysis will involve exploring the dataset, visualizing the data, and then building a prediction model using logistic regression.

1. The plot displays the count of individuals diagnosed with diabetes and those who tested negative for the condition, with a value of 1 representing a positive diagnosis and 0 representing a negative diagnosis.

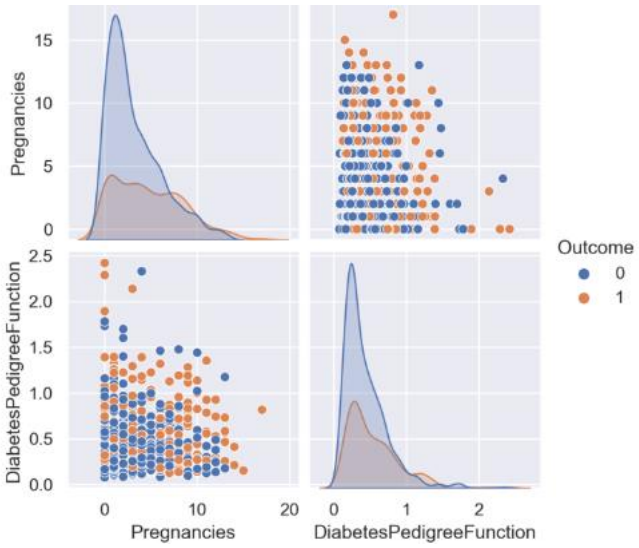


2. The pair plot of age against insulin reveals a noticeable trend in the likelihood of testing positive for diabetes. Specifically, individuals who are older or have elevated insulin levels appear to have a higher probability of testing positive.

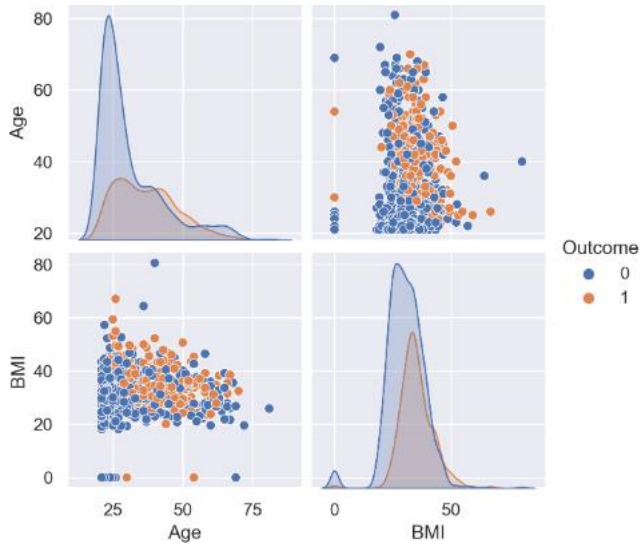


3. The scatter plot comparing the number of pregnancies with diabetes pedigree functions shows that the number of pregnancies does not impact the likelihood of testing positive for diabetes. Nonetheless, there is a trend indicating that higher

diabetes pedigree functions are associated with an increased likelihood of testing positive for diabetes.

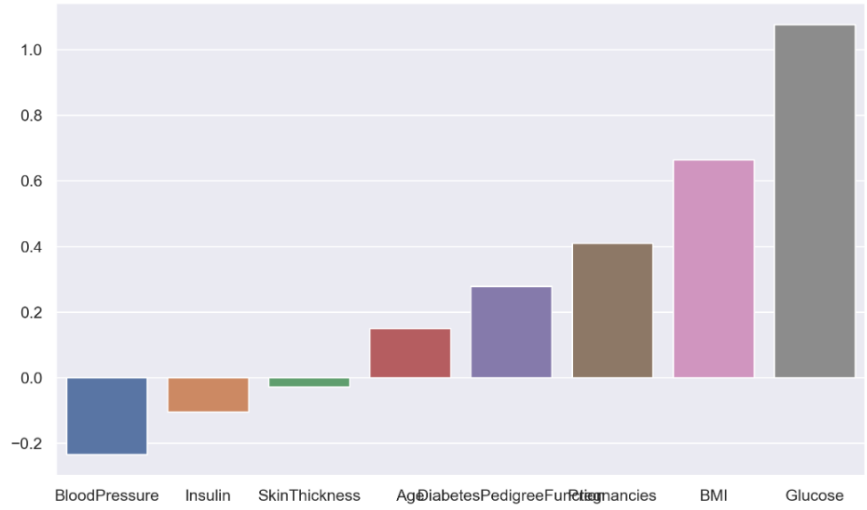


4. The pair plot of age against BMI also reveals a noticeable trend in the likelihood of testing positive for diabetes. Individuals older or with elevated BMI levels appear to have a higher probability of testing positive.



Machine Learning

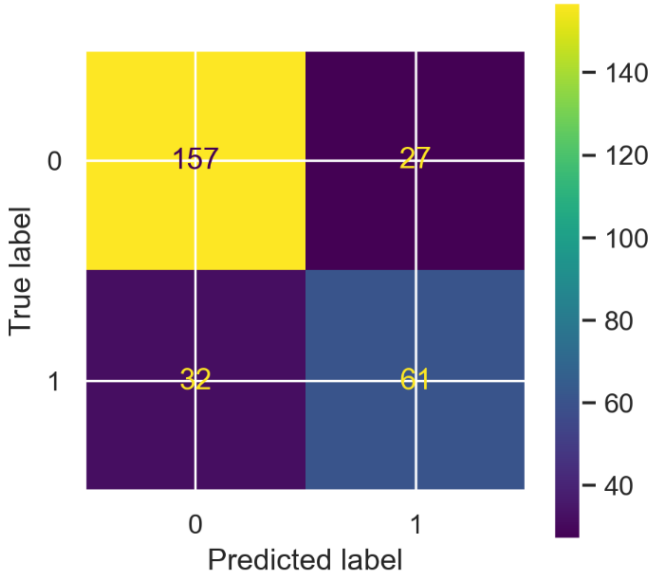
Now we use logistic regression to prepare a prediction model for diagnosing diabetes. This is done by splitting the data into train and test set. 90% of the data is used to train the model while remaining 10% is used to test the model.



Model Evaluation

We find that the model has an accuracy of 79%. This is not a good prediction accuracy and can be improved by trying out different models.

	precision	recall	f1-score	support
0	0.83	0.85	0.84	184
1	0.69	0.66	0.67	93
accuracy			0.79	277
macro avg	0.76	0.75	0.76	277
weighted avg	0.78	0.79	0.79	277



Finally, we test our model for a patient with the following findings.

Pregnancies:6
Glucose: 148
blood pressure: 72
Skin Thickness: 35
Insulin: 0
BMI: 33.6
Diabetes Pedigree Function: 0.627
Age: 50

Our model suggests that the patient has diabetes, which is a correct finding.

```
In [31]: log_model.predict(patient)
Out[31]: array([1], dtype=int64)
```