Supervised Learning Project

Samuel Croft

LHL Data Science (May 29th Cohort)

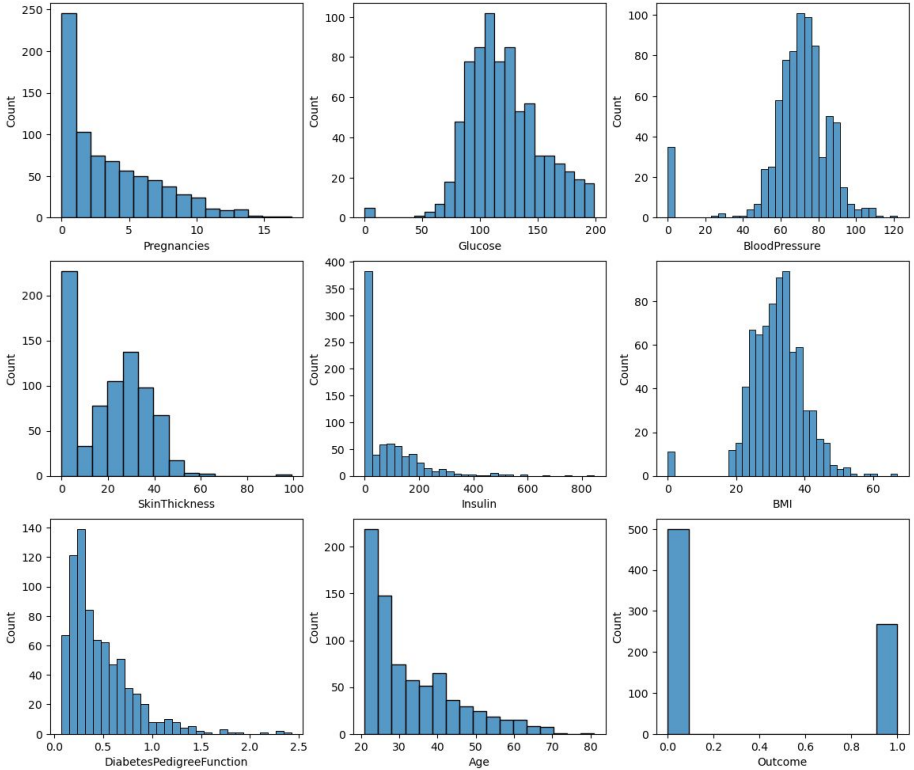
Project Goals

1. Perform EDA on the "Diabetes" dataset from the National Institute of Diabetes and Digestive and Kidney Diseases

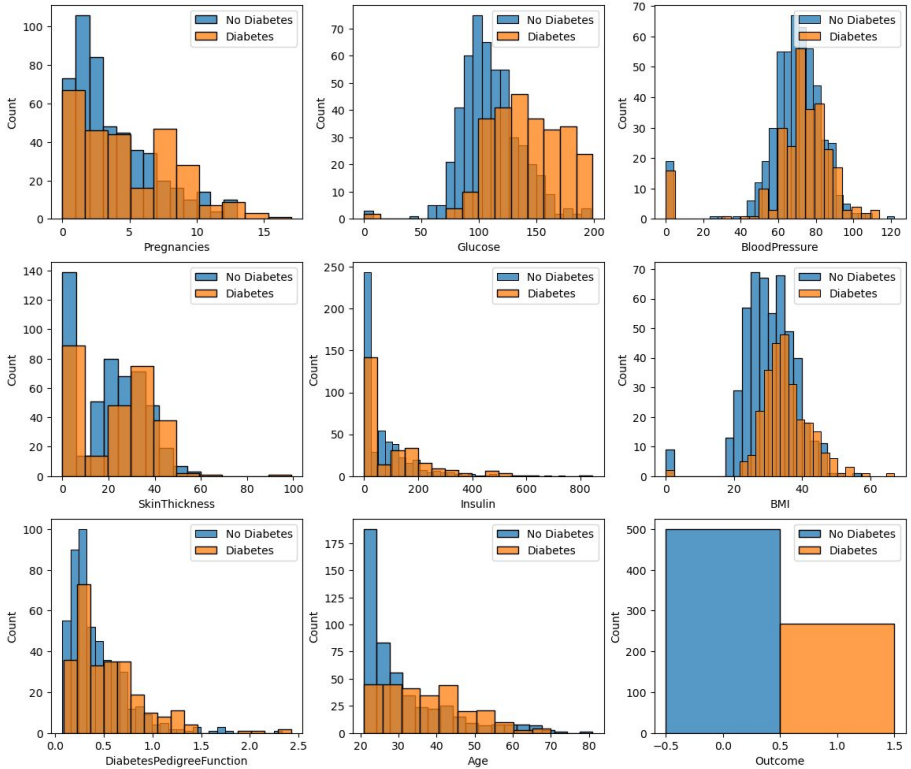
2. Perform Data Cleaning, Preprocessing, and Feature Engineering on the data set as needed

3. Build 2 different machine learning models to predict the presence of diabetes based on the available features

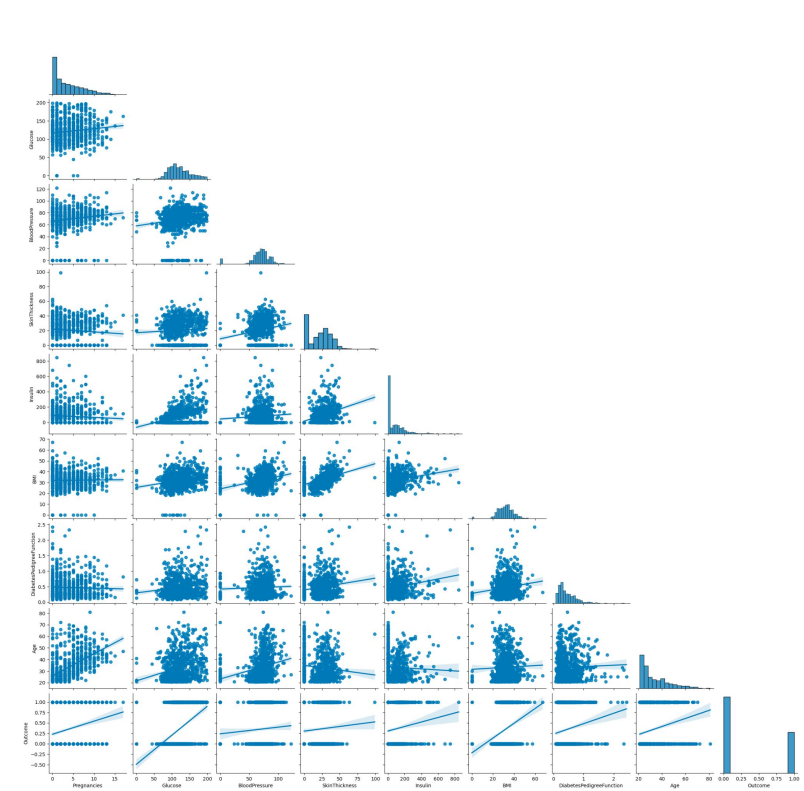
EDA – Distributions of Features

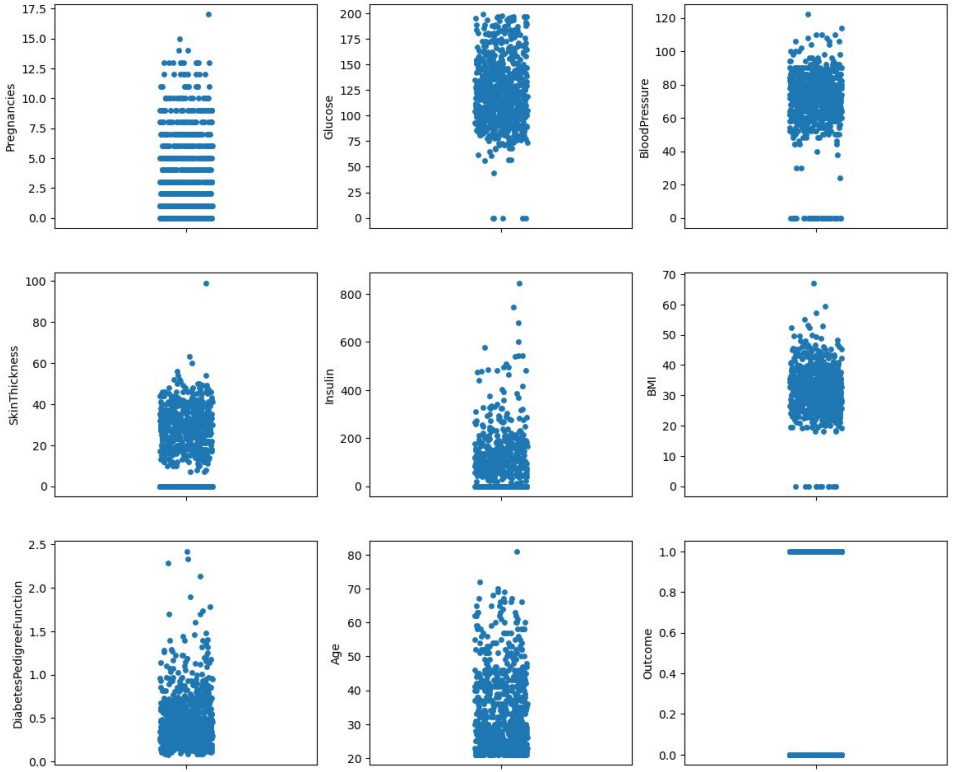


EDA – Distributions of Features by Outcome



EDA – Correlations and Strip Plots





EDA – Conclusions

1. There are some zeros in Glucose, BloodPressure, and BMI. These measurements can't be zero, so they will need to be converted before building the models

2. Insulin and SkinThickness both have many zeros, so these features will need to be removed

3. The distributions of those with diabetes are further to the right than those without diabetes for all features

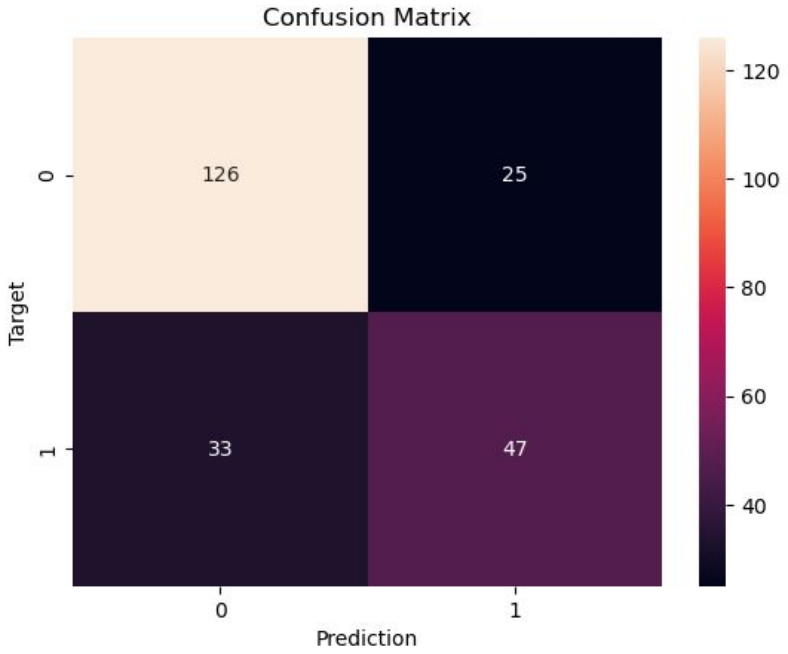
4. Pregnancies, Glucose, Insulin, BMI, DiabetesPedigreeFunction, and Age are all noticeably correlated with Outcome

5. Pregnancies, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age all have outliers that are much greater than the rest of the data points

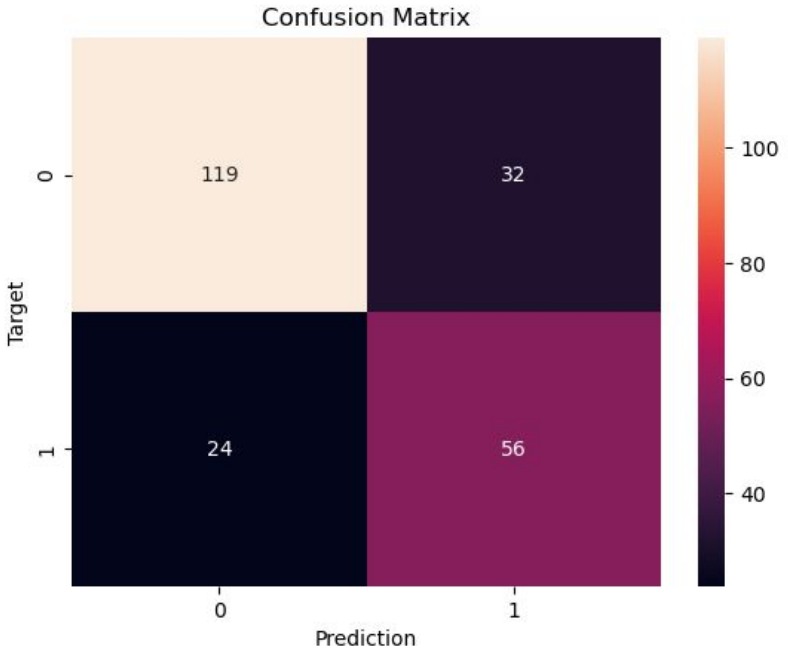
6. Since this is a binary classification problem, the two models used will be Logistic Regression and Random Forest Classification

Results – Comparing Confusion Matrices

Logistic Regression



Random Forest Classification



Results - Comparing Classification Reports

Logistic Regression

	precision	recall	f1-score	support
0	0.83	0.79	0.81	151
1	0.64	0.70	0.67	80
accuracy			0.76	231
macro avg	0.73	0.74	0.74	231
weighted avg	0.76	0.76	0.76	231

Random Forest Classification

	precision	recall	f1-score	support
0	0.83	0.79	0.81	151
1	0.64	0.70	0.67	80
accuracy			0.76	231
macro avg	0.73	0.74	0.74	231
weighted avg	0.76	0.76	0.76	231

Results - Comparing Feature Importance

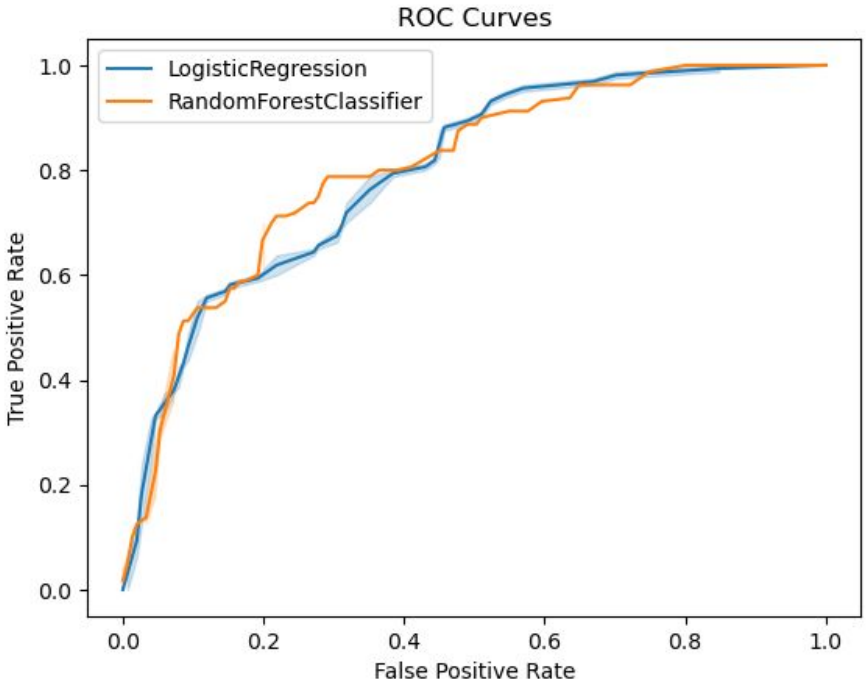
Logistic Regression

Feature	Coefficient	Absolute Coef	Importance Rank
Glucose	1.158306	1.158306	1.0
BMI	0.757252	0.757252	2.0
Age	0.386545	0.386545	3.0
Pregnancies	0.198387	0.198387	4.0
DiabetesPedigreeFunction	0.111745	0.111745	5.0
BloodPressure	-0.107364	0.107364	6.0

Random Forest Classification

Feature	Importance	Importance Rank
Glucose	0.320772	1.0
BMI	0.196453	2.0
Age	0.167552	3.0
DiabetesPedigreeFunction	0.132616	4.0
BloodPressure	0.097781	5.0
Pregnancies	0.084827	6.0

Results - Comparing ROC Curves and AUC



Area Under Curve (AUC)

Logistic Regression: 0.798
Random Forest Classifier: 0.805

Conclusions

1. The RandomForestClassifier performs better than the LogisticRegression in most metrics
    a. Area under curve
    b. All f1 metrics except for negative f1-score (tied)
    c. Precision on negatives
    d. Recall on positives

2. The LogisticRegression model has higher precision on positives, and higher recall on negatives.

3. Most important features for both models are Glucose, BMI, and Age.

4. RandomForestClassifier model should be chosen for predicting diabetes.