Supervised Learning on Diabetes Dataset

Siyi Mo

Project purpose

The purpose of the project is to use supervised learning techniques to build a machine learning model that can predict whether a patient has diabetes or not, based on certain diagnostic measurements. The project involves three main parts: exploratory data analysis, preprocessing and feature engineering, and training a machine learning model.

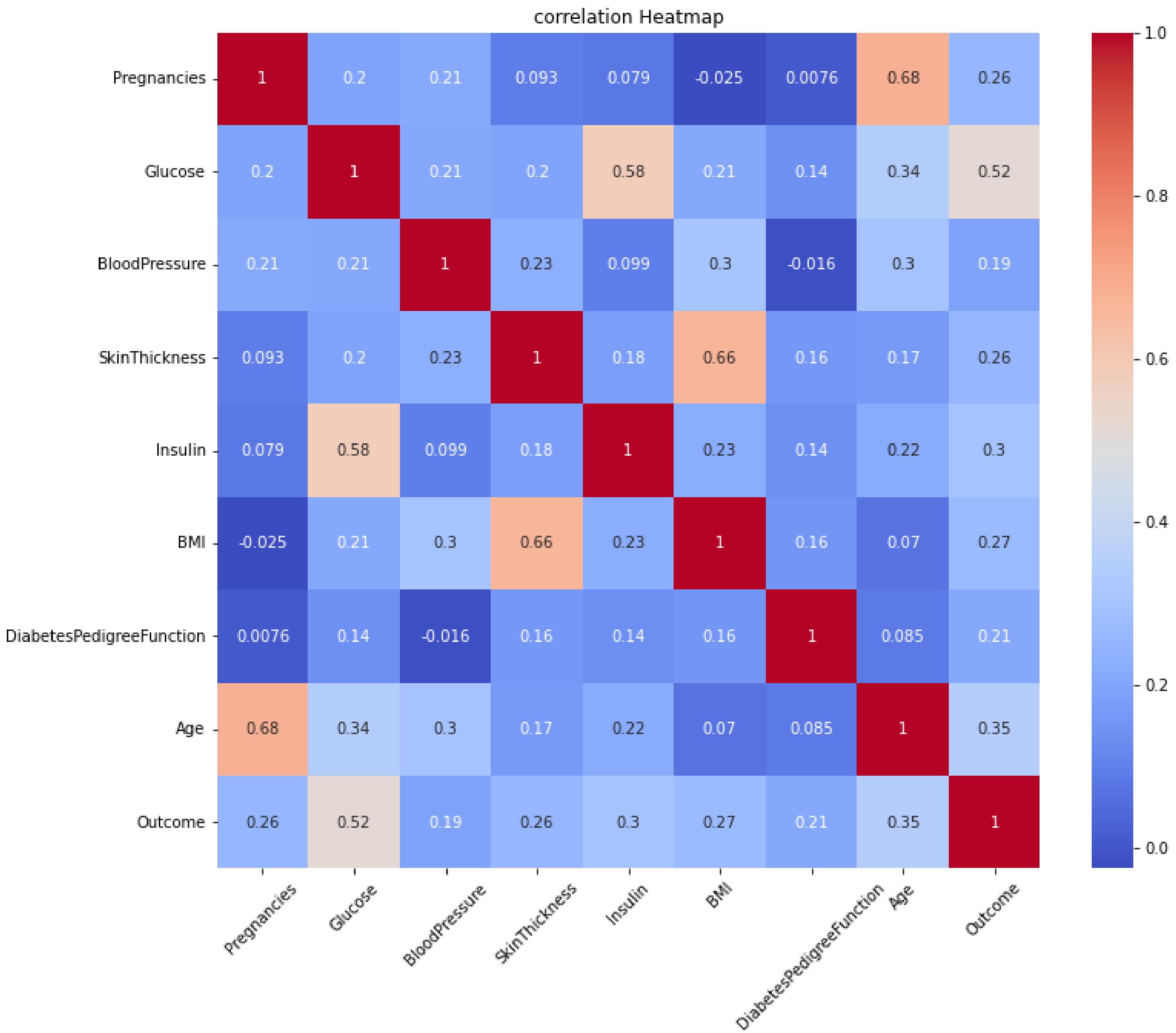
EDA Results

Glucose has the highest correlation to the target variables among all the predictive variables.

Age and insulin also have relatively high level correlation to the target variable.

Some of the predictive variables are not mutually exclusive instead there are some correlation between the variables

People with diabetes have higher mean value in each metrics. (pregnancies, glucoses, blood pressure, skin thickness, insulin, BMI and diabetes pedigree.)



correlation Heatmap

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1	0.2	0.21	0.093	0.079	-0.025	0.0076	0.68	0.26
Glucose	0.2	1	0.21	0.2	0.58	0.21	0.14	0.34	0.52
BloodPressure	0.21	0.21	1	0.23	0.099	0.3	-0.016	0.3	0.19
SkinThickness	0.093	0.2	0.23	1	0.18	0.66	0.16	0.17	0.26
Insulin	0.079	0.58	0.099	0.18	1	0.23	0.14	0.22	0.3
BMI	-0.025	0.21	0.3	0.66	0.23	1	0.16	0.07	0.27
DiabetesPedigreeFunction	0.0076	0.14	-0.016	0.16	0.14	0.16	1	0.085	0.21
Age	0.68	0.34	0.3	0.17	0.22	0.07	0.085	1	0.35
Outcome	0.26	0.52	0.19	0.26	0.3	0.27	0.21	0.35	1

Model Performance

	Random forest	Logistic regression
Accuracy score	0.77	0.71
Precision score	0.9	073
Recall score	0.46	0.41
F1 score	0.61	0.52
Roc auc score	0.71	0.66

Summary

The project is used random forest and logistic regression to predict diabetes. The process of the project consists of data cleaning (define missing values and remove outliers), data preprocessing (data scaling and transforming), modelling and model evaluation. Random forest performed better than logistic regression in this case based on the metrics scores returned by the evaluation metrics.