Supervised Learning Project

By

Zhiyun(Jenny) Liang

Project Objective

- 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.

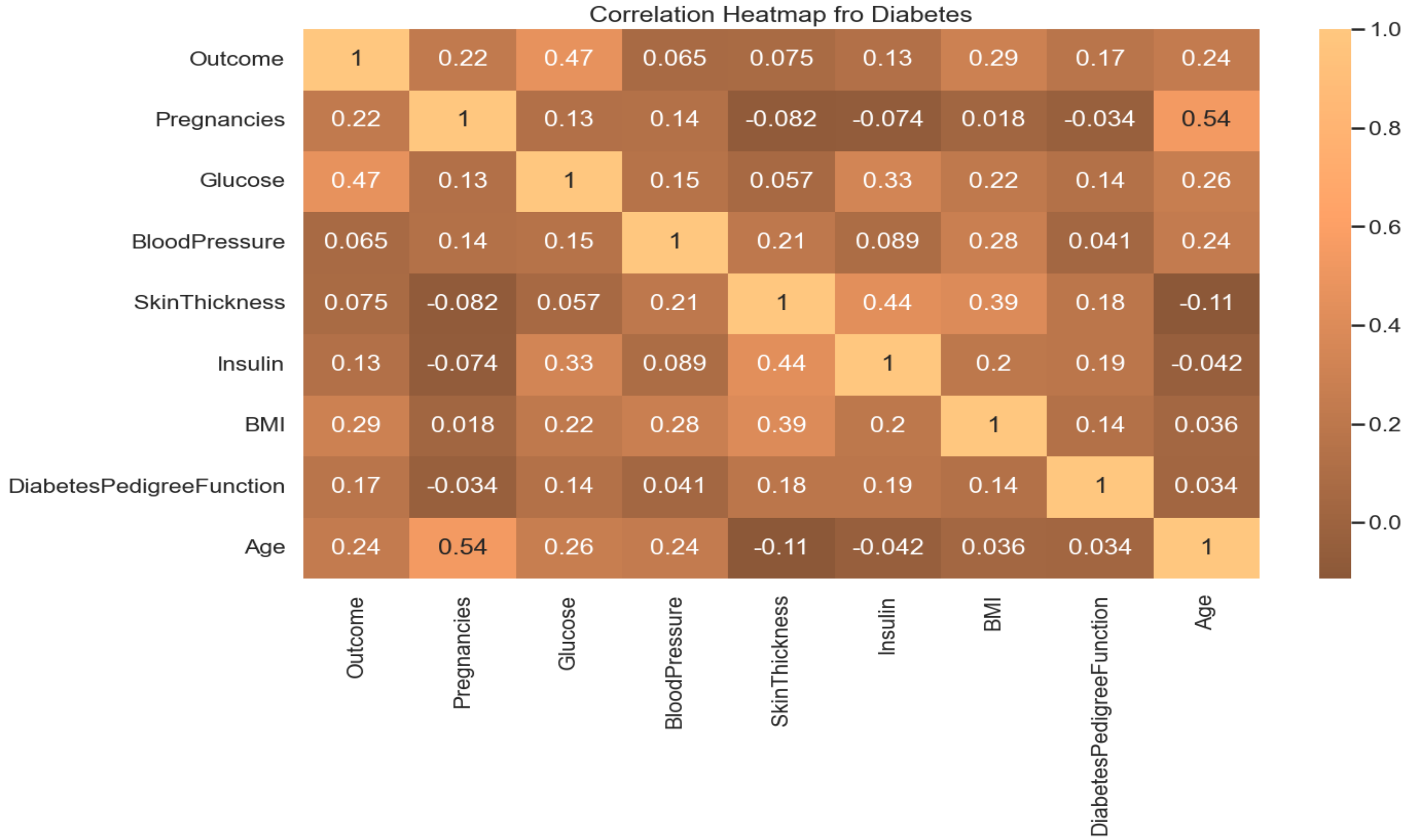
Project Flow Structure

- Exploratory data analysis
- Preprocessing and feature engineering
- Training machine learning models
- Results and Discussion

Exploratory Data Analysis

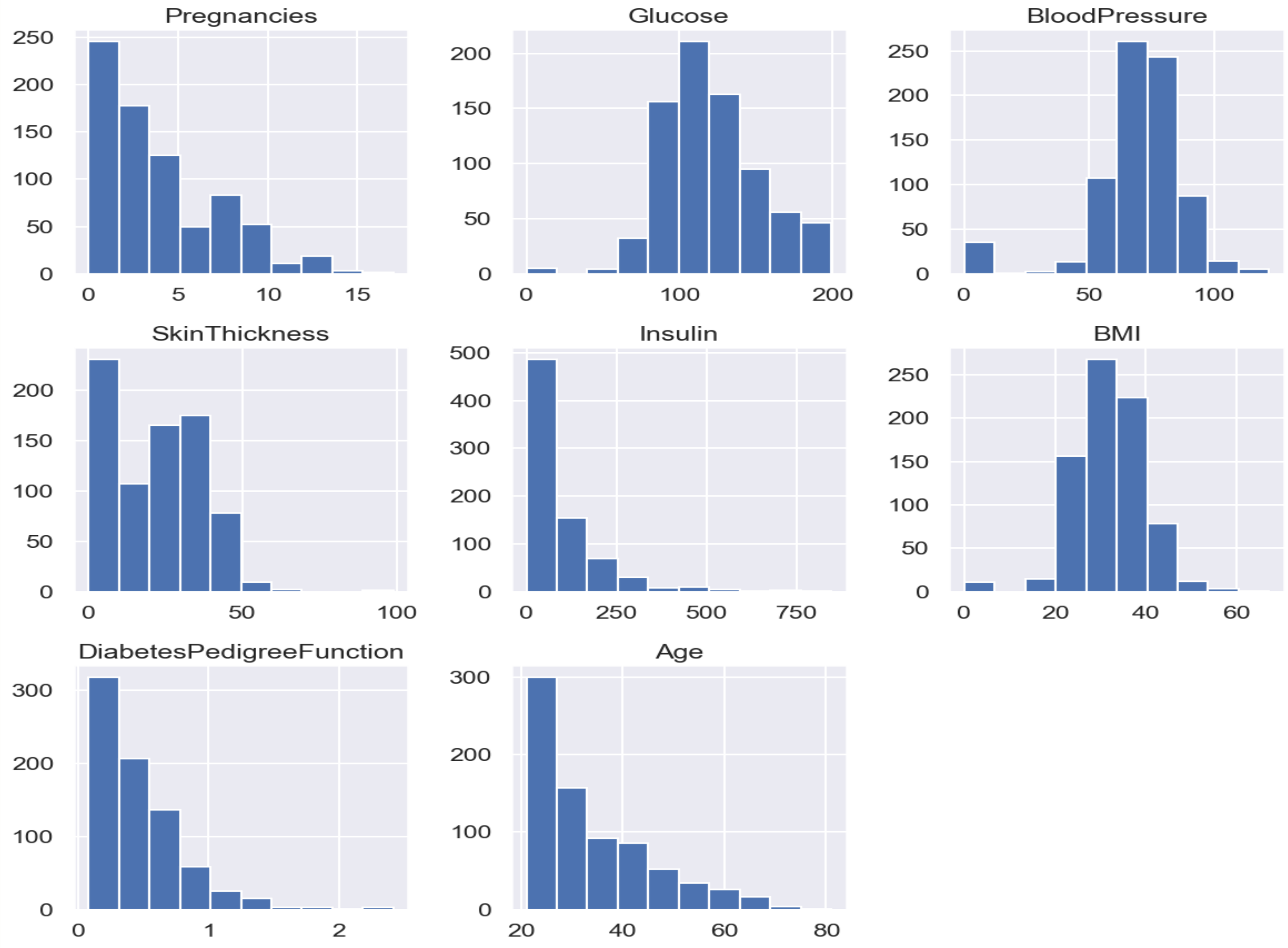
- Exploratory data (null value, outlier, etc.)
- Distribution of each predictor variable
- Correlation between the predictor variables

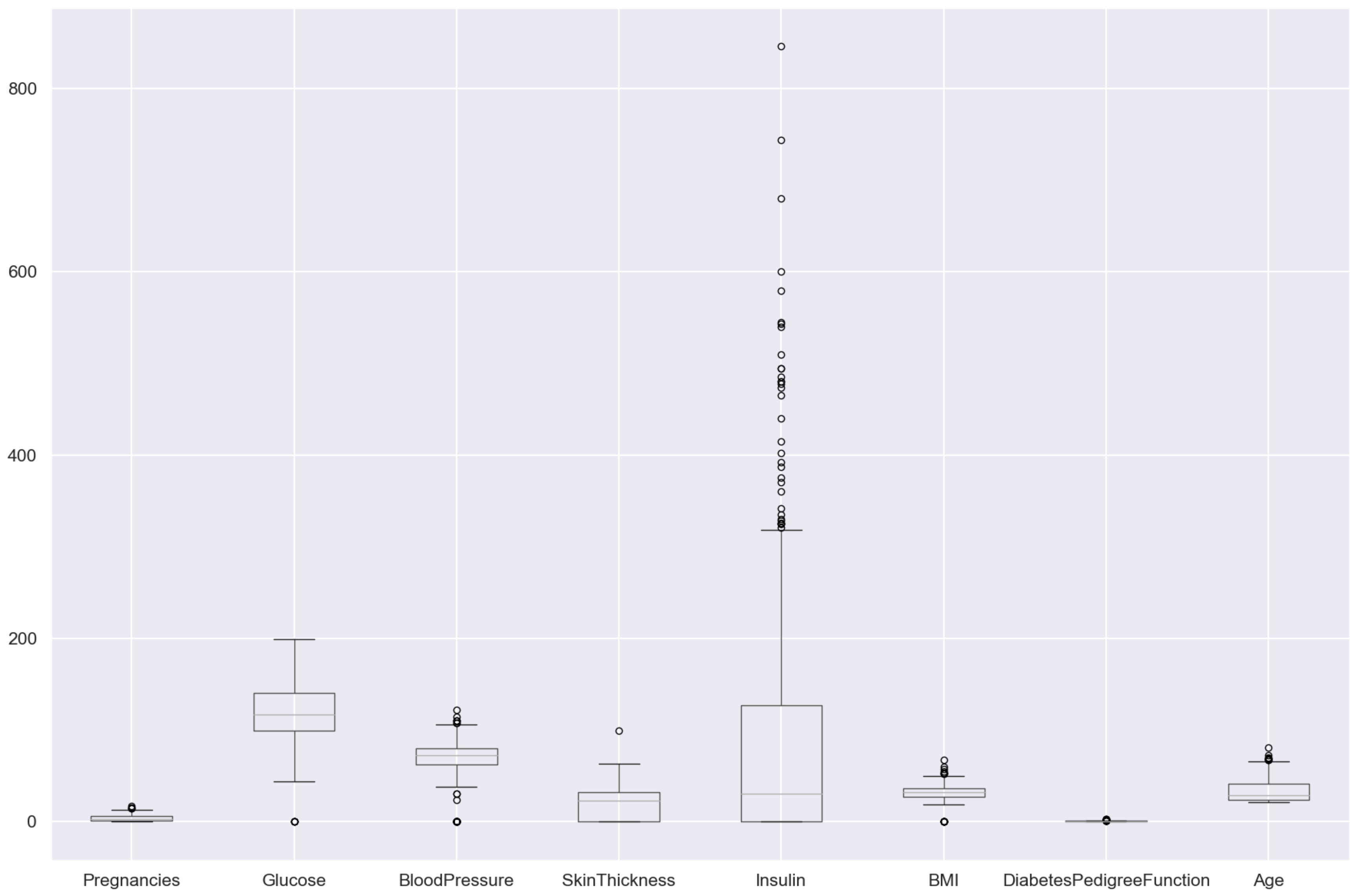
Heat Map



Correlation Heatmap fro Diabetes

	Outcome	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
Outcome	1	0.22	0.47	0.065	0.075	0.13	0.29	0.17	0.24
Pregnancies	0.22	1	0.13	0.14	-0.082	-0.074	0.018	-0.034	0.54
Glucose	0.47	0.13	1	0.15	0.057	0.33	0.22	0.14	0.26
BloodPressure	0.065	0.14	0.15	1	0.21	0.089	0.28	0.041	0.24
SkinThickness	0.075	-0.082	0.057	0.21	1	0.44	0.39	0.18	-0.11
Insulin	0.13	-0.074	0.33	0.089	0.44	1	0.2	0.19	-0.042
BMI	0.29	0.018	0.22	0.28	0.39	0.2	1	0.14	0.036
DiabetesPedigreeFunction	0.17	-0.034	0.14	0.041	0.18	0.19	0.14	1	0.034
Age	0.24	0.54	0.26	0.24	-0.11	-0.042	0.036	0.034	1


Pregnancies
250
200
150
100
50
0
0
5
10
15
Glucose
200
150
100
50
0
0
100
200
BloodPressure
250
200
150
100
50
0
0
50
100
SkinThickness
200
150
100
50
0
0
50
100
Insulin
500
400
300
200
100
0
0
250
500
750
BMI
250
200
150
100
50
0
0
20
40
60
DiabetesPedigreeFunction
300
200
100
0
0
1
2
Age
300
200
100
0
20
40
60
80


800
600
400
200
0
Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age

Preprocessing and Feature Engineering

- Handling missing values
- Handling outliers
- Scaling and normalization variable

Training ML Model

- Training Decision Tree Model
- Training Random Forest Model

Results and Discussion

- 'SkinThickness', 'Insulin' column around 1/3 values are missing, and 'Insulin' has a lot of outliers
- 'SkinThickness’, 'BloodPressure’ column has a lowest correlation with 'Outcome', which is 0.075, 0.065 respectively
- After Scaling the features, the model performance is improved
- The performance of Random Forest Model is better than Decision Tree Model