Supervised Learning Project

Prepared by: Chloe Phuong

Project Goals

The primary objective of this project is to analyze the Diabetes dataset using various supervised learning models. By comparing different models, we aim to extract valuable insights from the data and effectively communicate these findings to stakeholders. The project will employ appropriate visualizations and metrics to present the insights and facilitate informed decision-making in response to specific business questions.

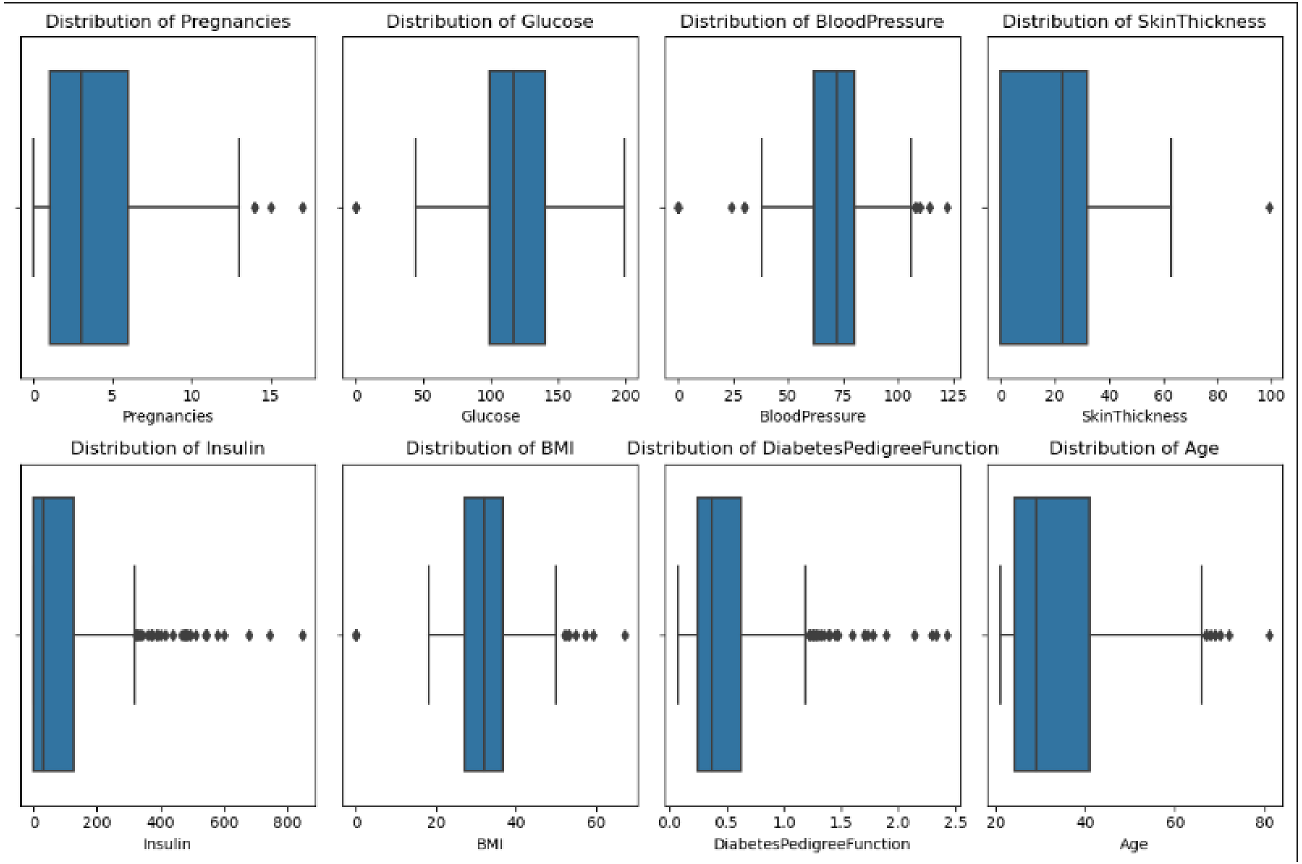
Process

1. EDA - Exploratory Data Analysis
2. Preprocessing & Feature Engineering
3. Training ML Model
4. Conclusion

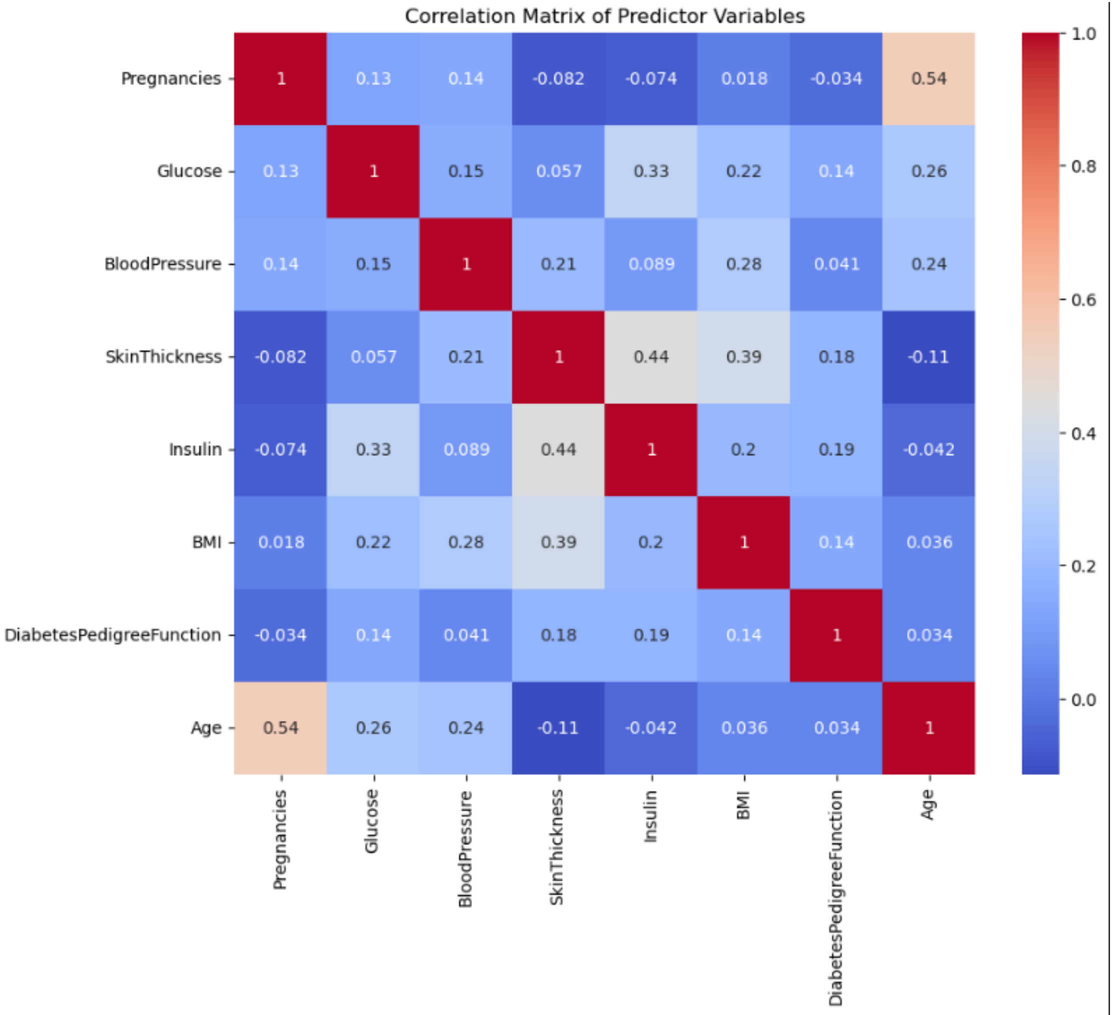
What I have discovered

The distribution of each predictor variable

Many outliers in the variables’ distributions



How are the predictor variables related to each other?

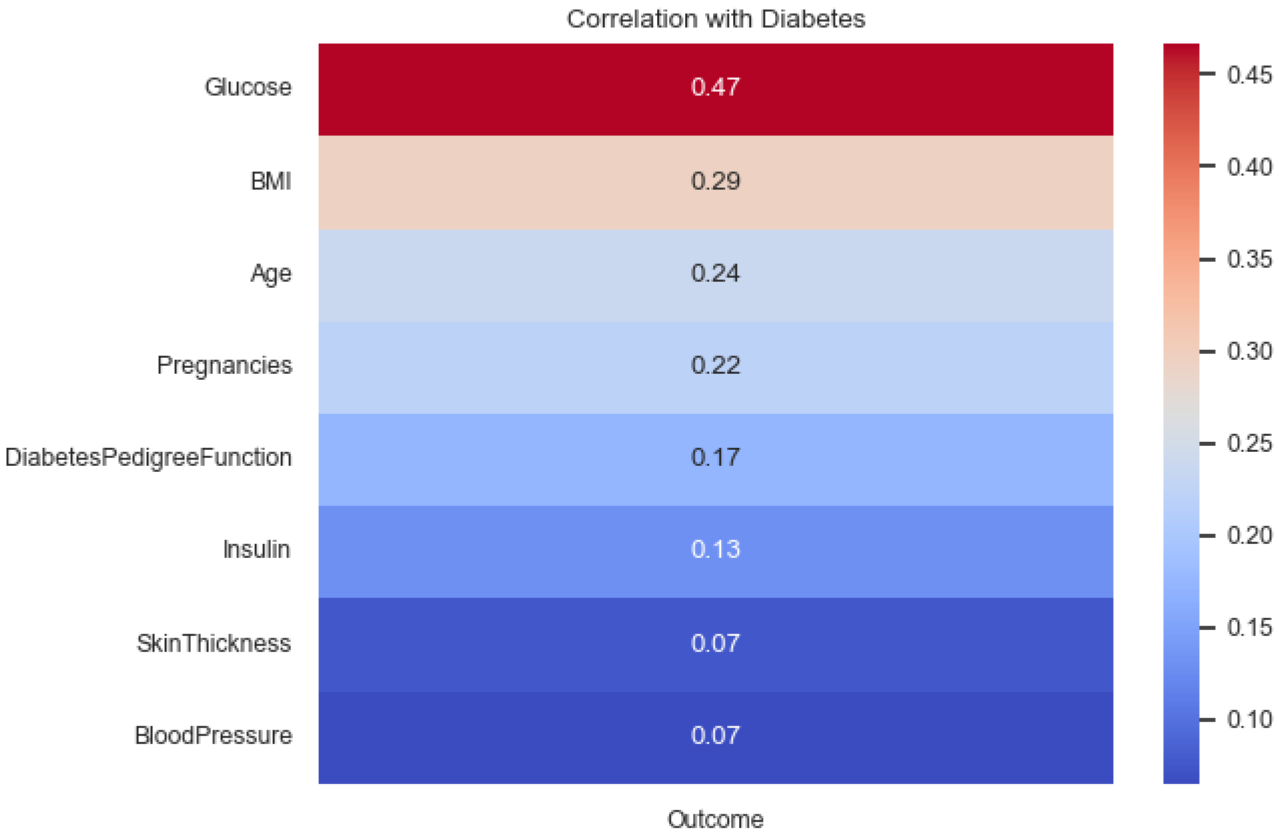


Correlation Matrix of Predictor Variables

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

Correlation with Diabetes

The Glucose level and BMI have a strong positive correlation with the presence of diabetes



The distribution of the predictor variables differ for individuals with diabetes and without diabetes

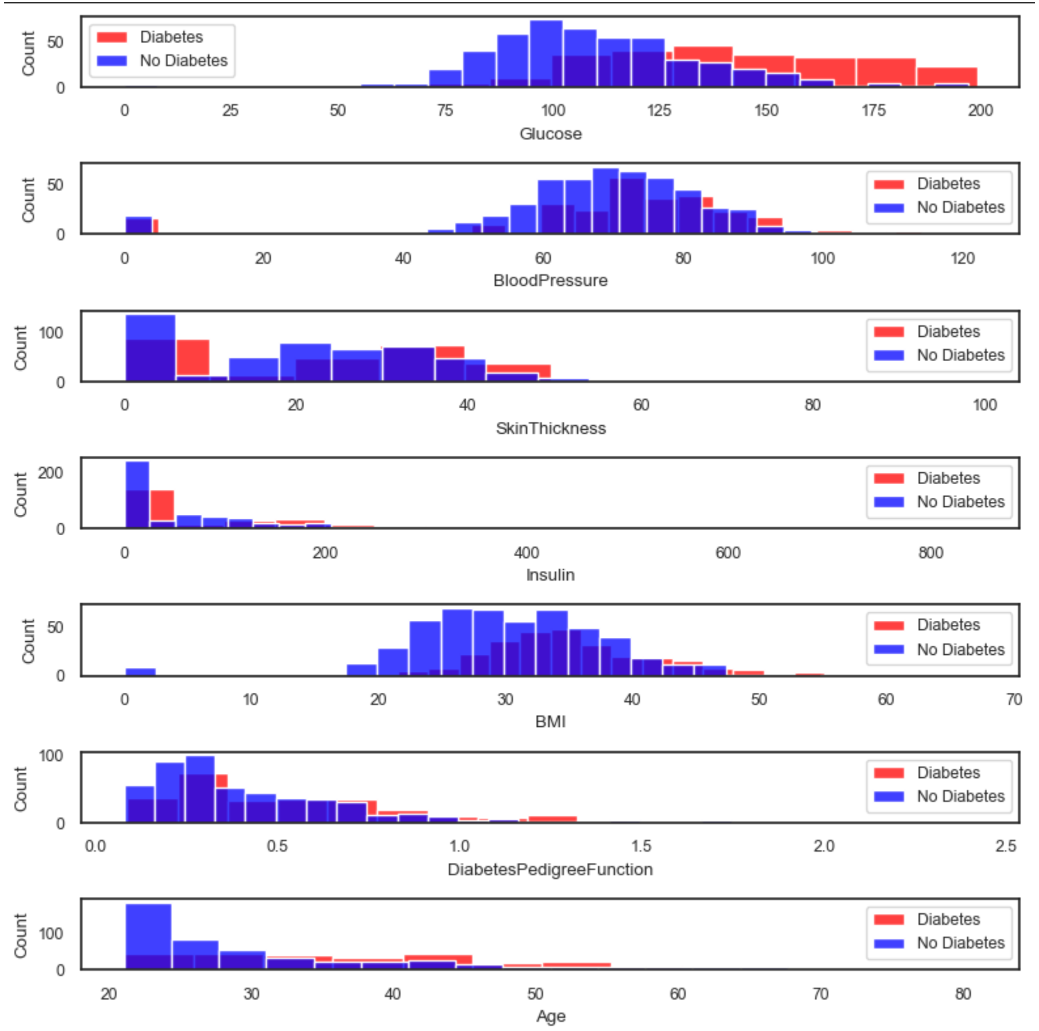
The average age of the individuals in the dataset is: 33.240885416666664

Average glucose level for individuals with diabetes: 141.25746268656715

Average glucose level for individuals without diabetes: 109.98

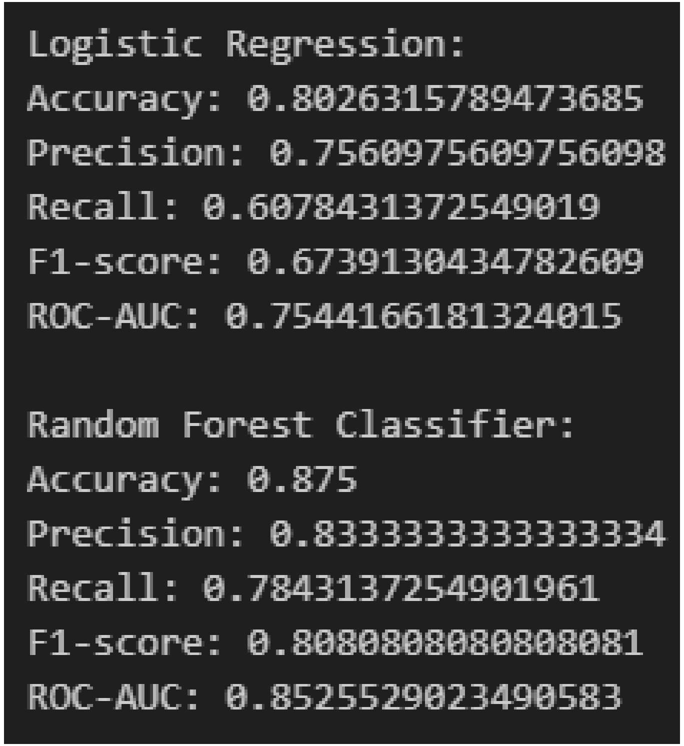
Average BMI for individuals with diabetes: 35.14253731343284

Average BMI for individuals without diabetes: 30.3042



Compare model performance

Based on the evaluation metrics, it can be concluded that the Random Forest Classifier outperformed the Logistic Regression model in predicting the outcome variable (presence of diabetes) for the given dataset.



```
Logistic Regression:
Accuracy: 0.8026315789473685
Precision: 0.7560975609756098
Recall: 0.6078431372549019
F1-score: 0.6739130434782609
ROC-AUC: 0.7544166181324015

Random Forest Classifier:
Accuracy: 0.875
Precision: 0.8333333333333334
Recall: 0.7843137254901961
F1-score: 0.8080808080808081
ROC-AUC: 0.8525529023490583
```

Key insights

- Based on the evaluation metrics, it can be concluded that the Random Forest Classifier outperformed the Logistic Regression model in predicting the outcome variable (presence of diabetes) for the given dataset.

- The Glucose level and BMI have a strong positive correlation with the presence of diabetes. This indicates that higher glucose levels and BMI are significant factors in predicting the likelihood of having diabetes.

- The average age of individuals in the dataset was approximately 33 years. This suggests that the dataset primarily consists of relatively young individuals, which could impact the generalizability of the findings to other age groups.

- The number of pregnancies was found to have a mild positive correlation with the presence of diabetes. This suggests that pregnancy history may play a role in diabetes risk, potentially due to hormonal changes and the impact on insulin resistance.

Thank you!