SUPERVISED LEARNING - DIABETES PREDICTION

OVERVIEW OF THE DATASET

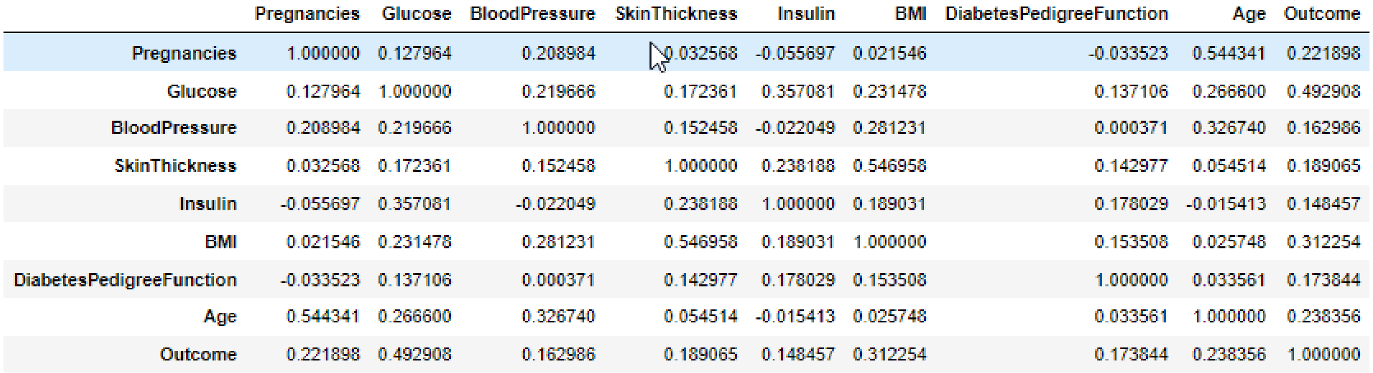
In this write-up, I will be analyzing the contributing factors to diabetes, and predicting if a person is likely to have diabetes based on the parameters available.

The dataset used is the Diabetes dataset from Kaggle where you can see all the columns and what they represent.
The factors considered are

- **Pregnancies**
- **Glucose**
- **Blood Pressure**
- **Skin Thickness**
- **Insulin**
- **BMI**
- **Diabetes Pedigree Function**
- **Age**
- **Outcome**

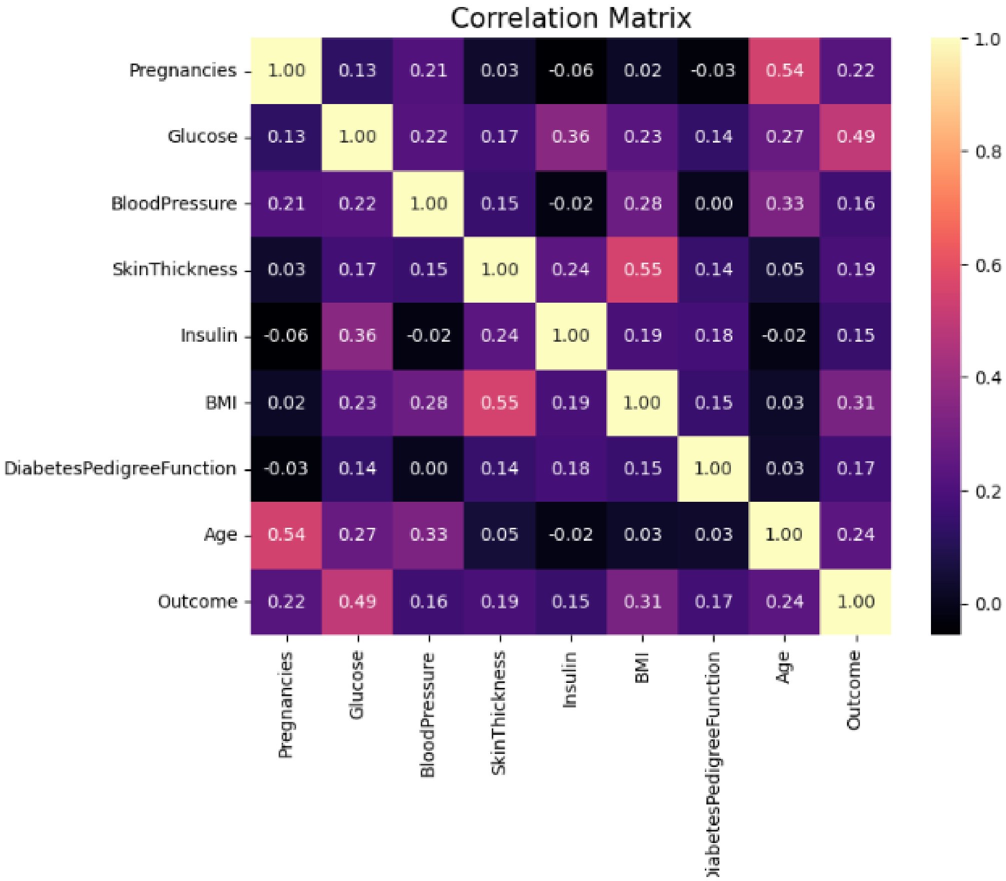
EXPLORATORY DATA ANALYSIS

1. **The dataset used has a shape of (768, 9): To be able to properly predict if a person has diabetes or not we would need more data to split into the test and train set, but for the purpose of this research we will make do with the data available.**
2. **Investigating the Correlation between the features of the dataset**



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.127964	0.208984	0.032568	-0.055697	0.021546	-0.033523	0.544341	0.221898
Glucose	0.127964	1.000000	0.219666	0.172361	0.357081	0.231478	0.137106	0.266600	0.492908
BloodPressure	0.208984	0.219666	1.000000	0.152458	-0.022049	0.281231	0.000371	0.326740	0.162986
SkinThickness	0.032568	0.172361	0.152458	1.000000	0.238188	0.546958	0.142977	0.054514	0.189065
Insulin	-0.055697	0.357081	-0.022049	0.238188	1.000000	0.189031	0.178029	-0.015413	0.148457
BMI	0.021546	0.231478	0.281231	0.546958	0.189031	1.000000	0.153508	0.025748	0.312254
DiabetesPedigreeFunction	-0.033523	0.137106	0.000371	0.142977	0.178029	0.153508	1.000000	0.033561	0.173844
Age	0.544341	0.266600	0.326740	0.054514	-0.015413	0.025748	0.033561	1.000000	0.238356
Outcome	0.221898	0.492908	0.162986	0.189065	0.148457	0.312254	0.173844	0.238356	1.000000

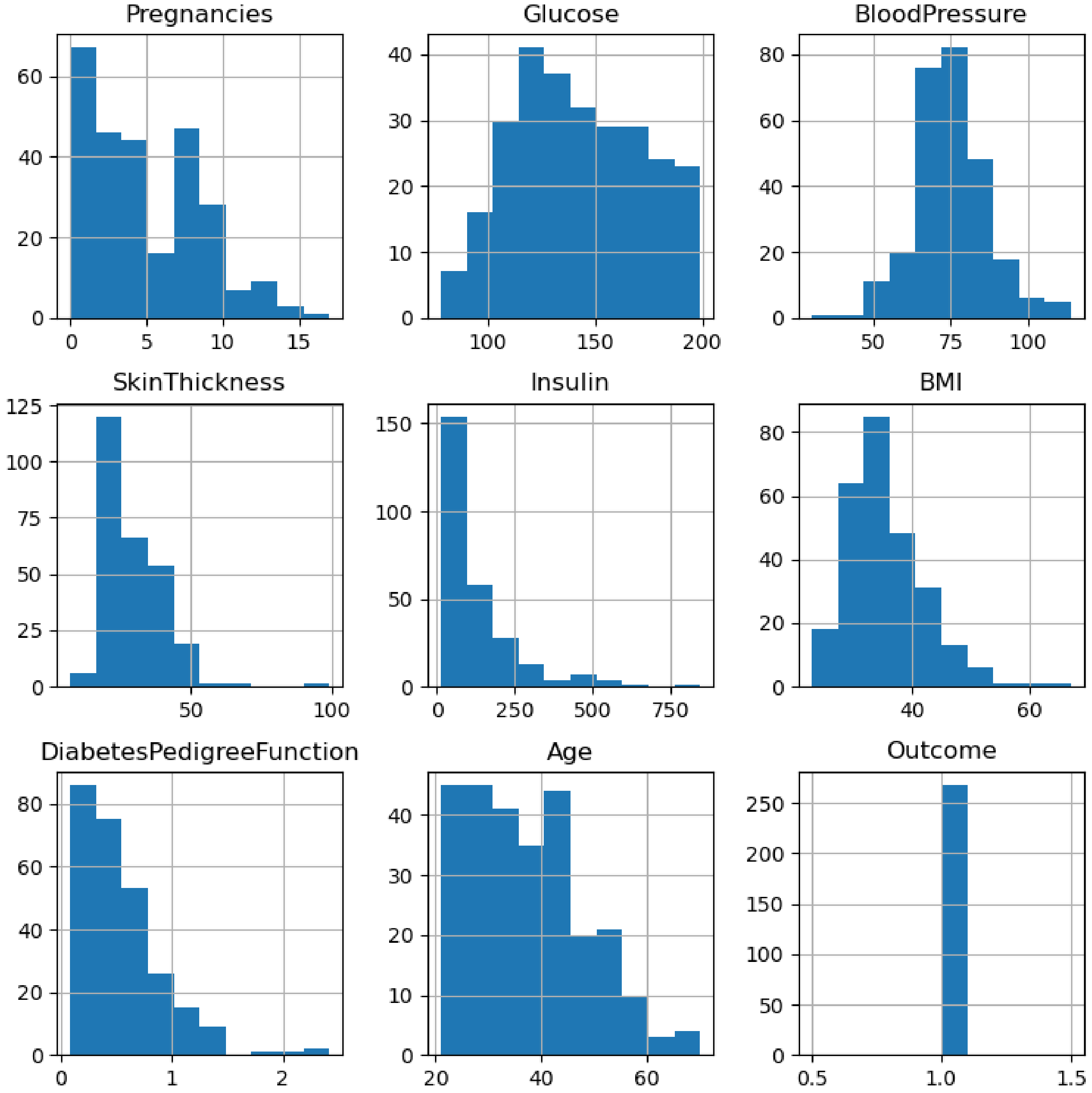
3. **And the heatmap for the correlation**

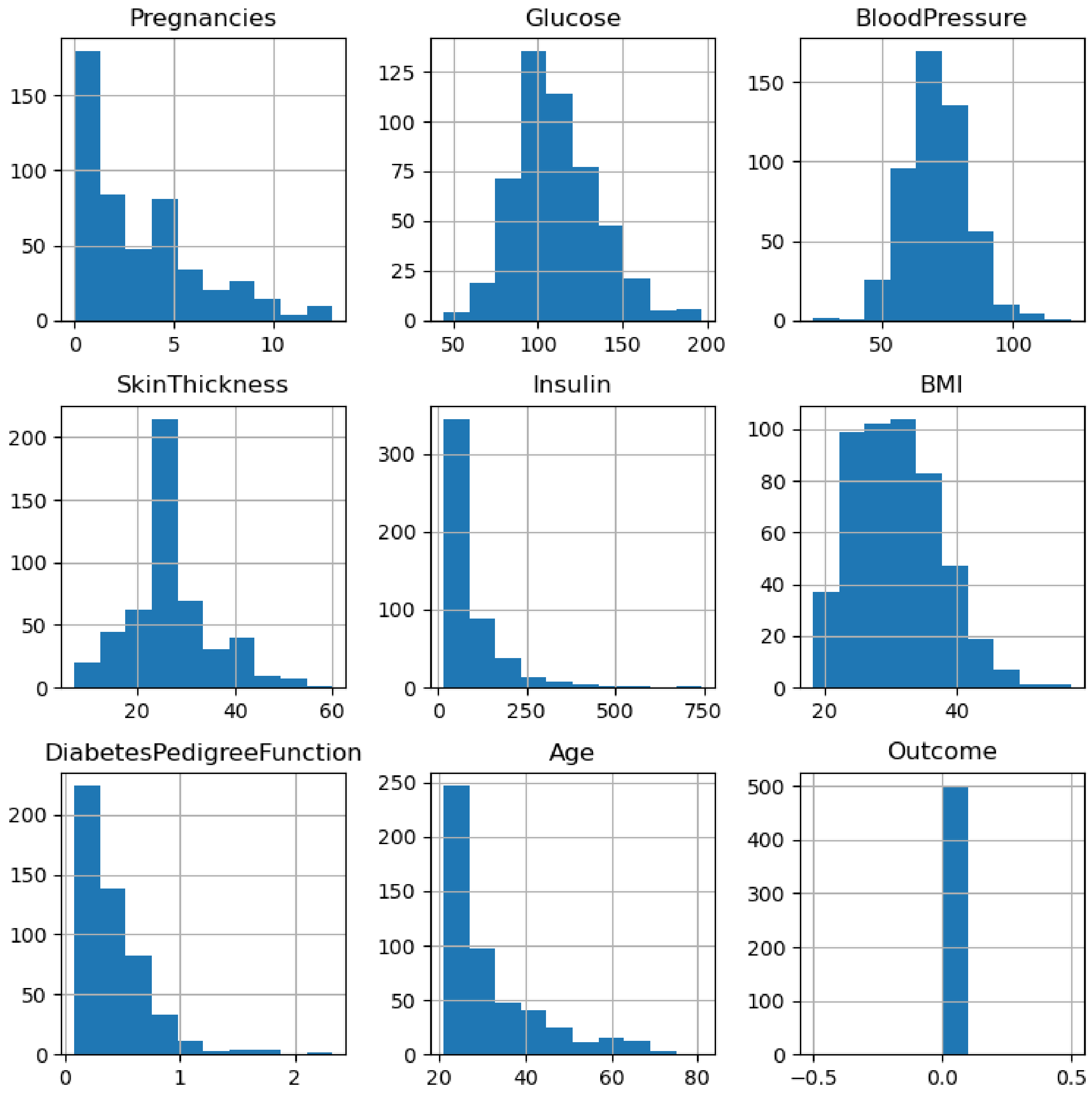


Exploratory data analysis revealed patterns and correlations within the diabetes dataset. For instance, there was a positive correlation between

glucose level and diabetes diagnosis, indicating that higher glucose levels are associated with an increased risk of diabetes.

4. **Comparing the distribution of the dataset for those withe diabetes and those without diabetes.**





5. **Conclusions**

EDA: When analyzing the data, it is crucial to pay attention to zero values. It is important to understand the units of each column and determine the normal range to gain insights into potential outliers. Before decided the outlier, I searched up if the max value of Insulin and Glucose are reasonable

Data Patterns: Exploratory data analysis revealed patterns and correlations within the diabetes dataset. For instance, there was a positive correlation between glucose level and diabetes diagnosis,

indicating that higher glucose levels are associated with an increased risk of diabetes. Additionally, factors like BMI and age showed some degree of correlation with diabetes occurrence.

In our research, we designed a system, which can predict diabetes with high accuracy.

All models provided an accuracy greater than 70%. LR and SVM provided approximately 77%–78% accuracy for both train/test split.

Feature Importance: The Random Forest model identified certain features as significant predictors of diabetes. These features, such as glucose level, body mass index (BMI), and age, exhibited higher importance in influencing the model's predictions

It is important to note that machine learning models are not perfect and may not always provide accurate predictions. It is also important to ensure that the data used to train the model is representative of the population being studied.

Comparing the three models we used, we have successfully built a machine learning model, specifically a Random Forest classifier,that exhibits a high level of accuracy in predicting diabetes or not¶