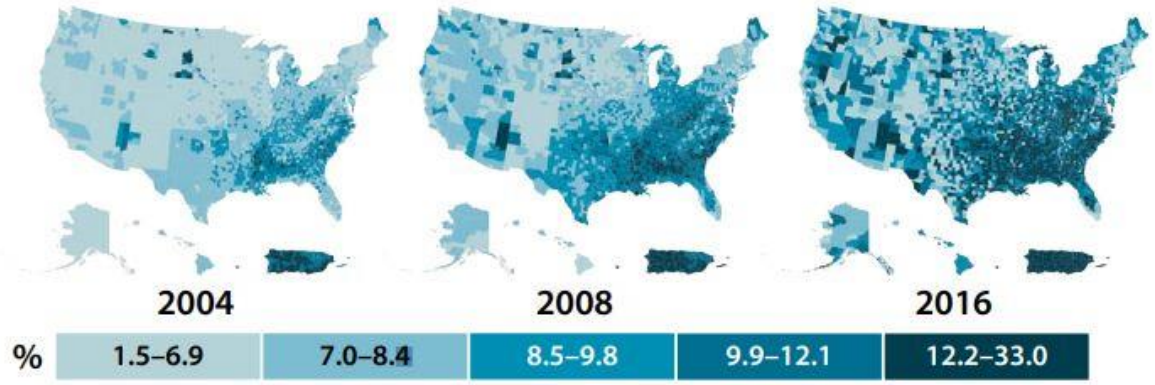
Diabetes Prediction

SOOYEON CHOI

Overview



Motivation

According to CDC's National Diabetes Statistics Report, 10.5% (34.2 million people) of USA's total population had diabetes in 2018's records. When looking at the reports over the years, it is clear that the rate is increasing yearly. In 2000, the percentage of people with diabetes was 4.4%, which means that is has more than doubled over 18 years.

Goals

1. Data Preparation : Cleaning data and creating visualizations for deeper understanding of the dataset.

2. Applying Classification Models : Logistic regression model and Decision Tree model used and evaluated to find the highest accuracy in predicting diabetes.

Research Question:

Diabetes prediction, which factor influences the cause of diabetes the most?

Dataset

The dataset is obtained from Kaggle (diabetes dataset), originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether a patient (female, over age 20) has diabetes, based on certain diagnostic measurements included in the dataset. It includes 2000 observations, each representing an individual. There are 9 Columns:

Pregnancies: Number of times pregnant

Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test

Blood Pressure: Diastolic blood pressure (mm Hg)

Skin Thickness: Triceps skin fold thickness (mm)

Insulin: 2-Hour serum insulin (mu U/ml)

BMI: Body mass index (weight in kg/(height in m)^2)

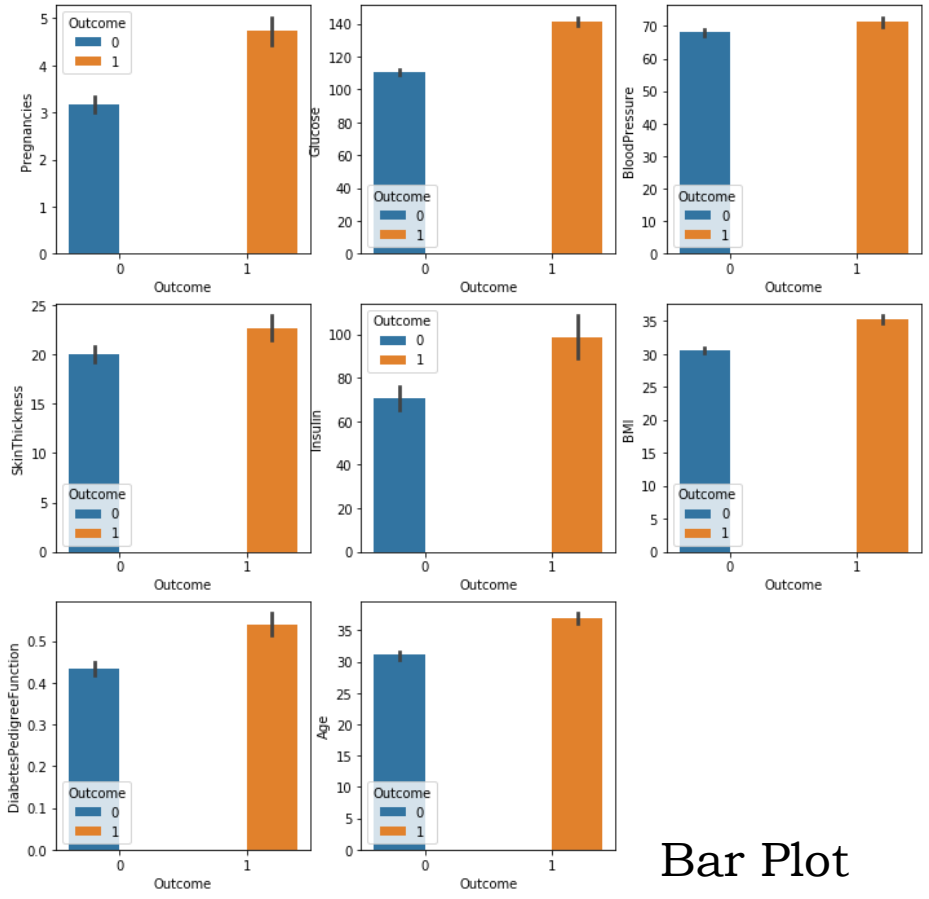
Diabetes Pedigree Function: Diabetes pedigree function

Age: Age (years)

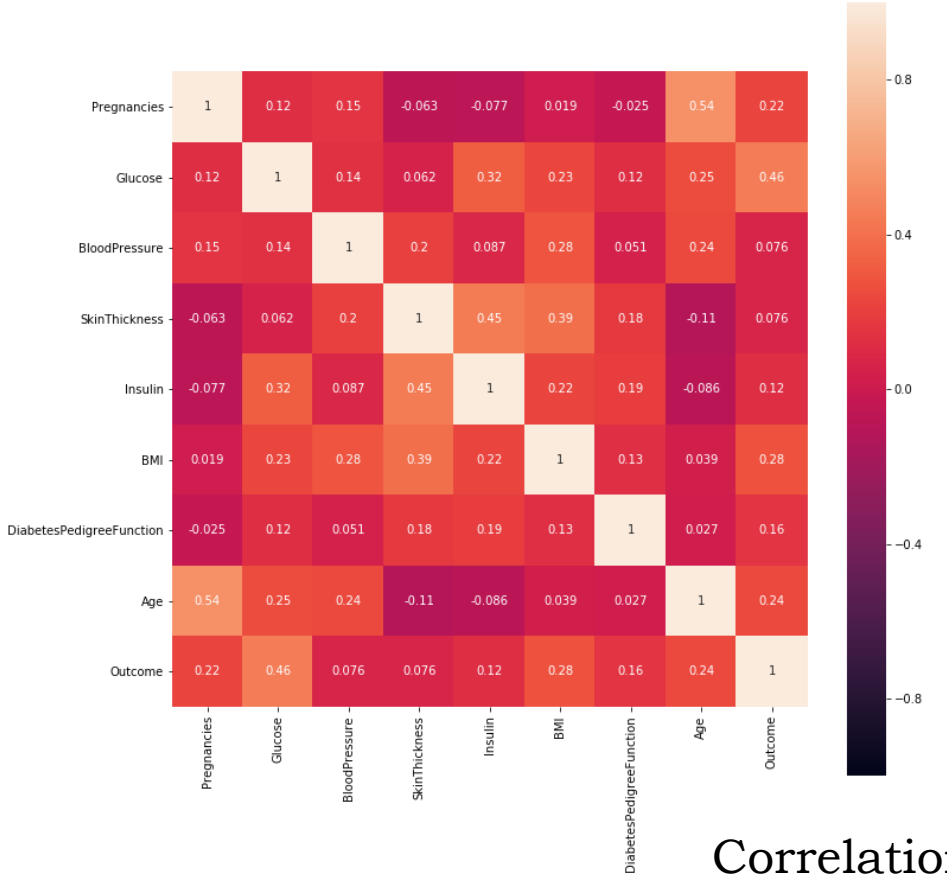
Outcome: Class variable (0 or 1)

* The first 8 columns are factors that can influence whether one has diabetes or not. The last column 'outcome' shows if the patient has diabetes or not and is our target. So we will be using all columns.

Visualizations

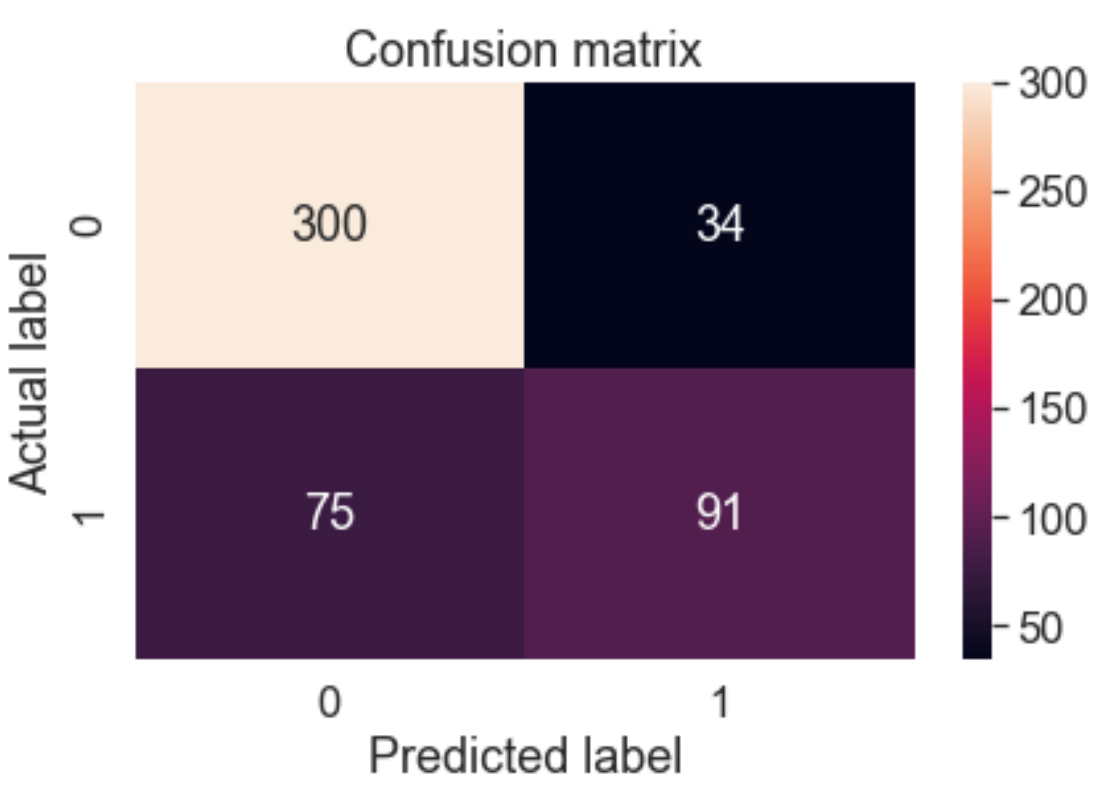


Bar Plot

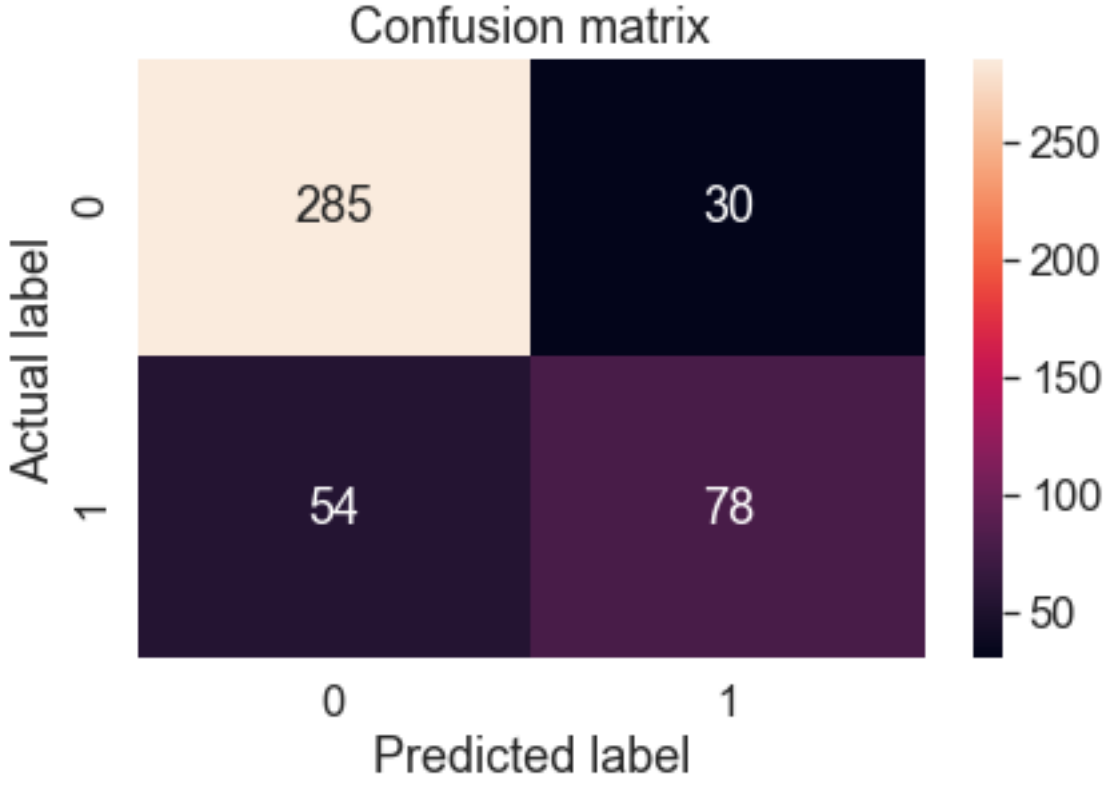


Correlation Heatmap

Visualizations



Logistic Regression



Decision Tree

Results

Correlation

Between features and Outcome

- Glucose: 0.46
- BMI: 0.28
- Age: 0.24
- Pregnancies: 0.22
- Diabetes Pedigree Function: 0.16
- Insulin: 0.12
- Blood pressure: 0.076
- Skin Thickness: 0.076

Logistic Regression

Accuracy

- Before removing outliers: 0.782
- After removing outliers: 0.812

Decision Tree

Accuracy

- Before removing outliers: 0.959
- After removing outliers: 0.970

Conclusion & Looking Forward

Through the correlation heatmap, we were able to see that **glucose had the highest correlation with our target goal, outcome**. This means that a patient's glucose level has the most influence on whether or not they have diabetes. For our machine learning models, both logistic regression model and decision tree model **performed better when the outliers were removed. Also, decision tree model had a much higher accuracy of 0.968, and it would be the best model to use when predicting diabetes**. It is important to keep the patients healthy all around, but doctors and the patients themselves should keep a more attention on their glucose level and BMI, especially for older patients, as they are the factors that can cause diabetes the most.

This project used machine learning algorithms learned only in class. Although the Decision tree model had a high accuracy, I would like to apply other models to compare the outcomes as I learn more techniques. Also, it would be interesting to investigate what kind of treatments or methods help patients lower the risk of diabetes, after finding related datasets.