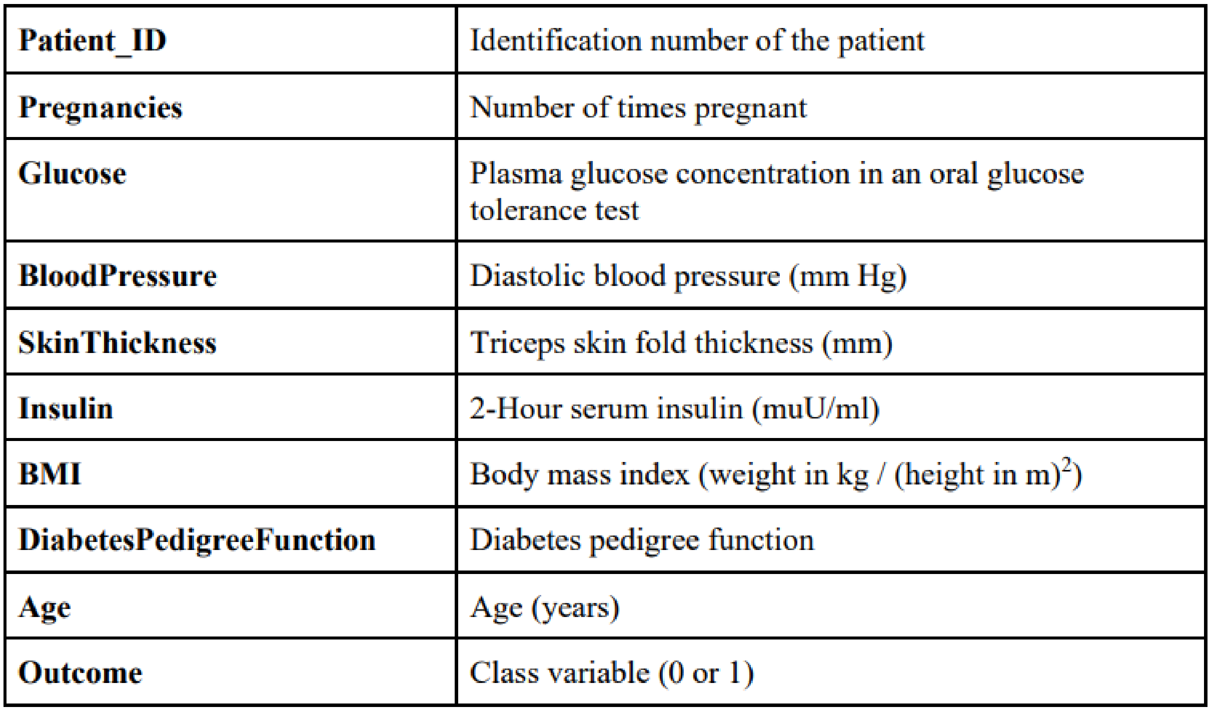
Description of Preprocessing Data Lab

The National Institute of Diabetes and Digestive and Kidney Diseases, United States, has collected a dataset regarding diabetes patients at Pima Woman's Hospital of Arizona. The objective of the dataset is to diagnostically predict whether a patient has diabetes based on specific diagnostic measurements included in the dataset. All the patients recorded in this dataset are females of Pima Indian heritage. The dataset consists of 8 medical predictor variables and one target variable, Outcome. Table 1 shows the description of each variable.



Patient_ID	Identification number of the patient
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration in an oral glucose tolerance test
BloodPressure	Diastolic blood pressure (mm Hg)
SkinThickness	Triceps skin fold thickness (mm)
Insulin	2-Hour serum insulin (muU/ml)
BMI	Body mass index (weight in kg / (height in m)^2)
DiabetesPedigreeFunction	Diabetes pedigree function
Age	Age (years)
Outcome	Class variable (0 or 1)

Table 1

In here diabetes.csv file,

- Removed duplicates if any.
- Filled missing values without deleting any record.
- Resolve out-of-range values.

Here, Pregnancies, Blood pressure, and Glucose values can't be negative. By the way, some doctors put 0 in some reports if the values are normal in Blood pressure and Glucose. But I didn't consider that in that case I used the median value.

All the outliers were detected, and those values were replaced appropriately.

Filled the null/nan values were appropriately detected and filled.

I mostly used **Google Collaboratory**.