Diabetes Predictor

Assignment 1
GEN 511 - Machine Learning
International Institute of Information Technology
Bangalore

IMT2017016 - Eric John
IMT2017024 - Kaushal Mittal
IMT2017045 - Vemuru Srihari

I. Introduction

We are trying to build a model which can predict whether a patient is diabetic or not based on different parameters. We use different methods to understand the given data, preprocess it and handle the missing as well as incorrect data.

II. Problem Definition

Given the Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function and Age predict accurately whether or not the patient has diabetes.

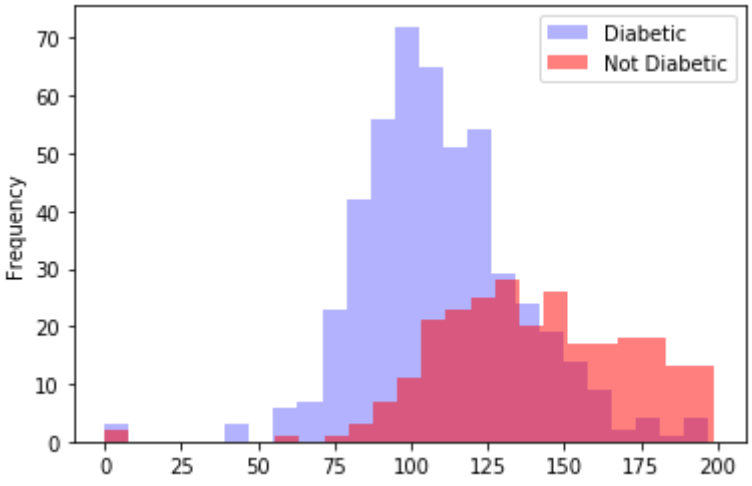
III. Data Processing and Missing data handling

Firstly, we detect the data points which have invalid values (such as negative pregnancies, or zero values for BMI and other columns) and replace them with NaN.

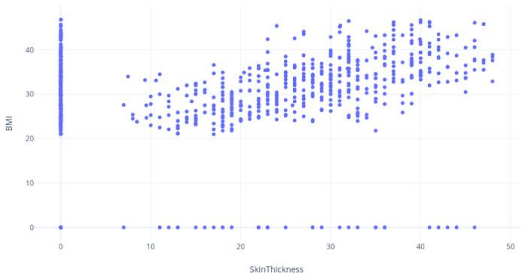
We find that SkinThickness, Glucose and Insulin columns have the most NaN values. We fill the values in other columns which are NaN with column average.

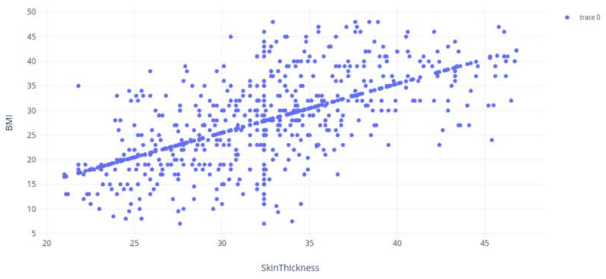
We remove top 2 percentile and bottom 5 percentile of data points as they are highly likely to be outliers.

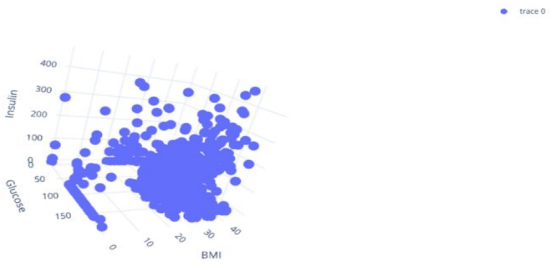
By looking at the median for Glucose after grouping it by Outcome, we clearly observe that diabetic patients have a higher median for Glucose. Hence, we replace the NaN values with median of corresponding group.

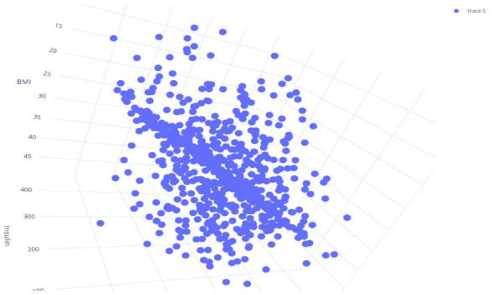


Then using correlation matrix for SkinThickness and Insulin columns we find the columns on which they depend the most. Then we plot these and observe they roughly are in linear pattern. Then we make a linear regression model to fill these, SkinThickness with respect to BMI and Insulin with respect to Glucose and BMI.









IV. Exploratory Data Analysis and Feature Extraction

We use Principal Component Analysis(PCA) to understand the dimensionality of the data. Since the variance of every principal component is greater than 3%, we do not remove any component and take projection of the normalised data points along the principal axes.

V. Algorithm and Model Building

During data preprocessing the Insulin and Skin Thickness columns contained many invalid and NaN data points, and using its correlation with other features (as described earlier) we built a linear regression model to predict the approximate values.

Since our outcome is binary which depends on multiple dimensions we use logistic regression model. We build a logistic regression model from the Scikit learn library. We train the model with the projections along principal components of data.

VI. Result

We use K-fold cross validation to evaluate our model. We choose k=10 which shuffles the data and divides the data into 10 parts, training and testing data as 90% and 10% respectively.

After taking an average of all the K-fold validation scores we obtain 76% accuracy.