

Introduction

This project aims to identify diagnostic parameters that can be used to predict diabetes. In addition, it also checked if few variables are correlated.

Dataset

It utilized data from the National Institute of Diabetes and Digestive and Kidney Diseases, India (Source: Pima Indians Diabetes Database | Kaggle). Dataset included 8 predictor variables such as number of times pregnant, Glucose levels, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes pedigree function, and Age. Outcome variable was Boolean (0 for diabetes and 1 for non-diabetic).

Outcomes of the Exploratory Data Analysis

Logistical Regression was performed for predicting diabetes with glucose, blood pressure, insulin, BMI, and Age as predictor variables. P values indicated that Glucose, Blood Pressure, BMI, and Age are statistically significant in predicting diabetes if we consider alpha to be 5%. Accuracy rate of the model was observed to be >99%.

Blood Pressure and Age are positively correlated but strength of relationship is weak. They can be non-linearly correlated.

Blood Pressure and Insulin levels seem to be positively correlated but strength of relationship is very less. They can be non-linearly correlated.

Missed During the Analysis

Data appears to be of only female Indian population. Therefore, it is not representative of even Indian population. More comprehensive data collection and analysis is required to check a statistically significant connection.

Normalcy of only one variable was checked. More variables are to be checked if they are normally distributed.

Were there any variables you felt could have helped in the analysis

Male data should have helped in the analysis. Also, people's location provides lifestyle information, e.g. people living in rural communities vs in cities.

Any assumptions made you felt were incorrect

I made assumptions about the applicable variables and not going by solid medical evidence may have resulted in selection of wrong or missing relevant variables.

Challenges

No challenge was faced during the analysis.