Classification of Diabetic patients: The case of Pima Indian Diabetes

Introduction

Diabetes is one of the most dreadful disease that has become common these days among men and women irrespective of the ages. This use case primarily focuses on understanding and classifying certain patient records into either diabetic or not. Diabetes patient records were obtained from two sources: an automatic electronic recording device and paper records.

The objective of the usecase is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. All patients are females at least 21 years old of Pima Indian heritage

About the dataset

The datasets consist of several medical predictor (independent) variables and one target (dependent) variable, Outcome. Independent variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

Dataset Description

The various features involved in the dataset has been described as below:

1. *Pregnancies*: Number of times pregnant
2. *Glucose* :Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. *Blood Pressure*: Diastolic blood pressure (mm Hg)
4. *SkinThickness:* Triceps skin fold thickness (mm)
5. *Insulin:* 2-Hour serum insulin (mu U/ml)
6. *BMI:* Body mass index (weight in kg/(height in m)^2)
7. *DiabetesPedigreeFunction:* Diabetes pedigree function
8. Age: Age (years)
9. *Diabetes:* Class variable (0 or 1/pos or neg) 268 of 768 are 1(pos), the others are 0(neg)

Objectives

1. To explore the dataset to study and understand various factors responsible for diabetes in a patient.
2. To build a classification model using machine learning and to predict whether or not a patient is diabetic.