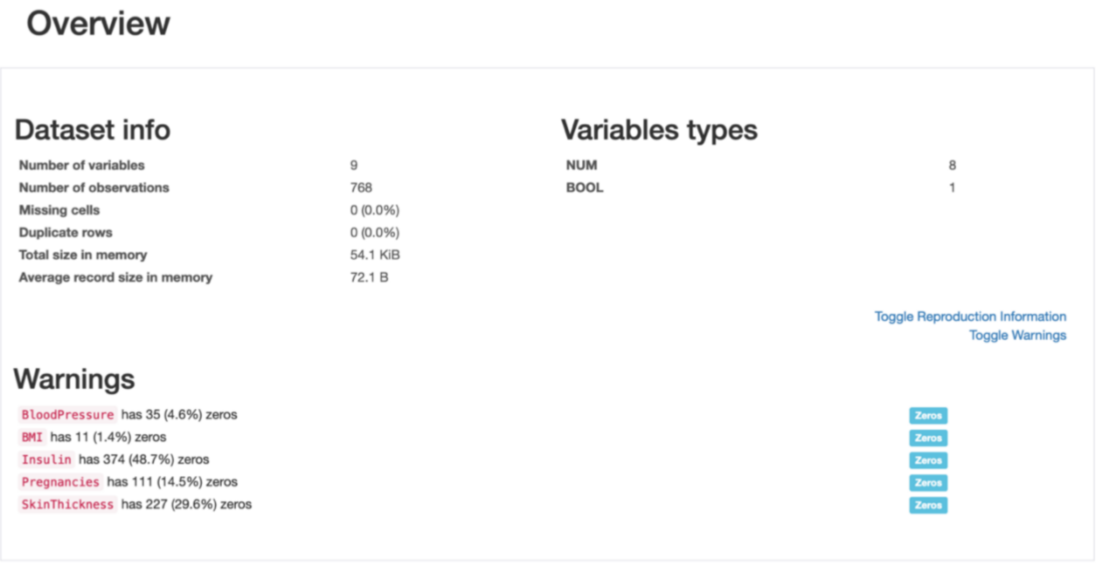
Project Proposal: diabetes

Diabetes is a chronic condition in which the body develops a resistance to insulin, a hormone which converts food into glucose. Diabetes affect many people worldwide and is normally divided into Type 1 and Type 2 diabetes. Both have different characteristics. This article intends to analyze and create a model on the PIMA Indian Diabetes dataset to predict if a particular observation is at a risk of developing diabetes, given the independent factors. This article contains the methods followed to create a suitable model, including EDA along with the model.

Dataset

The dataset can be found on the Kaggle website. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases and can be used to predict whether a patient has diabetes based on certain diagnostic factors. Starting off, I use Python 3.3 to implement the model. It is important to perform some basic analysis to get an overall idea of the dataset.



Overview

Dataset info

Number of variables	9
Number of observations	768
Missing cells	0 (0.0%)
Duplicate rows	0 (0.0%)
Total size in memory	54.1 KiB
Average record size in memory	72.1 B

Variables types

NUM	8
BOOL	1

Warnings

BloodPressure has 35 (4.6%) zeros

BMI has 11 (1.4%) zeros

Insulin has 374 (48.7%) zeros

Pregnancies has 111 (14.5%) zeros

SkinThickness has 227 (29.6%) zeros

We can see the basic information about the dataset such as the size, missing values, etc. On the top right, we see 8 of numerical columns and 1 Boolean column (which is our dependent variable). In the lower panel, (%) of zeros are given in every column, which will be an useful information for us later. We do not have any categorical variable as an independent variable.

Question for DataSet:

What information should be added to the program?

Tools:

Python

Jupyter notebook

Numpy

Pandas