Analysing Diabetes Dataset

Manuel A. Castro R.

Last Updated: 15/01/22

Context

- Diabetes is one of the most frequent diseases worldwide and the number of diabetic patients is growing over the years. The main cause of diabetes remains unknown, yet scientists believe that both genetic factors and environmental lifestyle play a major role in diabetes.
- A few years ago, research was done on a tribe in America which is called the Pima tribe (also known as the Pima Indians). In this tribe, it was found that the ladies are prone to diabetes very early. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients were females at least 21 years old of Pima Indian heritage.

Text taken from the original Word Document

The Notebook

To access the notebook click here


```
import pandas as pd #Data manipulation
import numpy as np #Numerical - Statstical manipulation
from plotly.subplots import make_subplots #Plots library
import plotly.express as px
import plotly.graph_objects as go #these two libraries will help us to plot the data

# These libraries help us to build, train and test the logistic regression model
from sklearn import metrics
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, confusion_matrix
from sklearn.model_selection import train_test_split
```

The Dataset

Variable Name	Description	Measurement Unity	Type of varible
Pregnancies	Number of pregnancies		Numerical - Discrete
Glucose	Plasma glucose concentration over 2 hours in an oral glucose tolerance test		Numerical - Continuous
BloodPressure	Diastolic blood pressure	mm Hg	Numerical - Continuous
SkinThickness	Triceps skin fold thickness	mm	Numerical - Continuous
Insuline	2-Hour serum insulin	mu U/ml	Numerical - Continuous
BMI	Body mass index	weight in kg/(height in m)^2	Numerical - Continuous
DiabetesPedigreeFunction	A function which scores likelihood of diabetes based on family history.		Numerical - Continuous
Age	Age	Years	Numerical - Discrete
Outcome	Class variable (0: person is not diabetic or 1: person is diabetic)		Numerical - Discrete

Note: in this case, age variable is discrete even it could be continuous, just the years as whole number were taken.

The Dataset

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

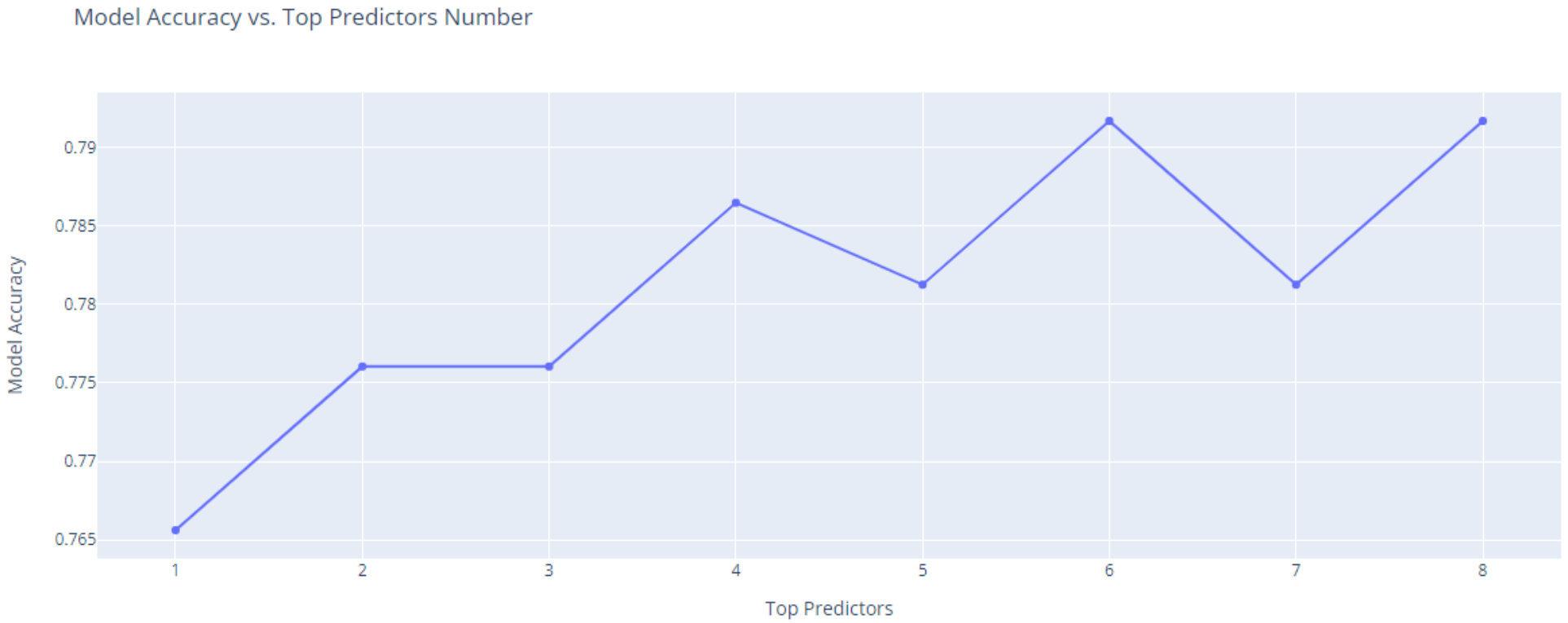
The Dataset

The variable of interest is the Outcome variable that indicates if a women is diabetic or not.

The correlation matrix suggest that there are not strong correlations, but Glucose, BMI and Age are the top 3 correlated variables that could help us to predict if a women is diabetic or not.

	Variable	Corr. index
8	Outcome	1.000000
1	Glucose	0.466581
5	BMI	0.292695
7	Age	0.238356
0	Pregnancies	0.221898
6	DiabetesPedigreeFunction	0.173844
4	Insulin	0.130548
3	SkinThickness	0.074752
2	BloodPressure	0.065068

Further Analysis



We can predict if a women would be diabetic with a 79% accuracy approx., but it is just one approach we can take from the dataset we have.

Resourses

[Repository link]

[Jupyter Notebook link]