Efraín Pérez
September 21th, 2021

Capstone Project

Machine learning Engineer Nanodegree Program

Using Machine Learning to classify if a patient has diabetes

Project overview

Three years ago, I was diagnosed with pre-diabetes at 22 years old. It was a big shock for me, because being young you never have thoughts about any kind of disease or illness. But shortly after being diagnosed, I had so much interest and enthusiasm in the field of machine learning and was always wondering, how could I use this knowledge to help people like me and not have the same scenario like me?. So, with this project I have my opportunity!.

Diabetes is in the top 10 causes of death in my country, Panama. So for that, we have to take actions against this disease and prevent any future patients.

Problem Statement

The goal is to create a binary classifier using machine learning to predict whether or not a person is diabetic. The tasks are:

1. Download the dataset from kaggle
2. Split the dataset into 80% train, 20% test
3. Choose the machine learning model

4. Test the data
5. Show the metric (accuracy)

Metrics

A common metric for binary classification is accuracy. Using accuracy is enough to classify a possible patient of diabetes

$$accuracy = \frac{true\ positive + false\ positive}{total\ size\ of\ dataset}$$

Data set and inputs

We are going to use the Diabetes Kaggle dataset to achieve this project https://www.kaggle.com/uciml/pima-indians-diabetes-database. In this dataset all patients here are females at least 21 years old of Pima Indian heritage.
This data has many medical predictors as independent variables and one dependent variable.

Variables:

- **Pregnancies:** Number of times pregnant
- **Glucose:** Plasma glucose concentration 2 hours in an oral glucose tolerance test.
- **BloodPressure:** Diastolic blood pressure (mm Hg)
- **SkinThickness:** Triceps skin fold thickness (mm)
- **Insulin:** 2-Hour serum insulin (mu U/ml)
- **BMI:** Body mass index (weight in kg/(height in m)^2)
- **DiabetesPedigreeFunction**: Diabetes pedigree function
- **Outcome:** Is diabetic person or not

Efraín Pérez
September 21th, 2021

Benchmark model

This is a kaggle competition, so there is no benchmark model. I am going to create a neuronal network using tensorflow and compare it with a naive Bayes model, to get a better approach treating this kind of problem.

Project design

- Download the data set
- Visualize variables to get a better understanding of the data
- Split the dataset into train and test sets
- Create the neural network and machine learning model
- Print any report and compare the results

References

MINSA. (2018). *LA DIABETES, SEXTA CAUSA DE MUERTE EN PANAMÁ*. MINSA. http://www.css.gob.pa/web/6-julio-2018au.html