Machine Learning Engineer Nanodegree

Capstone Project Proposal

R Leela Samyuktha

February 6th, 2019

Proposal:

Classifying Pima Indians Diabetes Dataset

Domain Background:

History:

Diabetes is a metabolic disorder characterized by high blood sugar. Diabetes can lead to many complications, several of which can lead to an earlier death. These complications include heart disease, strokes, kidney disease, and blindness. The number of diabetics worldwide is increasing steadily, and developing effective treatment for diabetics has become important recently.

Considering in 1930, the diabetics detection was a big issue. IT is estimated that by 2030, there will be 366 million people affected by diabetics mellitus. It can be solved via looking at the suspected parameters and accumulating information on the ones parameters, then predicting the usage of supervised classification algorithms, the presence or absence of diabetes.

The related academic work is found at http://www.ijera.com/papers/Vol8_issue1/Part-2/C0801020913.pdf

Problem Statement:

Given a dataset containing various attributes of 768 Indian patients, define classification algorithms. Applying different classification algorithms on the Indian patient diabetes disease dataset and choose the best algorithms based on the accuracy which can identify whether a person is suffering from diabetes disease or not.

Datasets and Inputs:

The dataset that I am working is downloaded from

https://www.kaggle.com/uciml/pima-indians-diabetes-database

The number of instances are 768. It is a multivariate data set, contain 9 variables that are Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age and Outcome. All values are real integers. This data set contains 500 non diabetic patient records and 268 diabetic patient records. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. Outcome is target variable used to divide into groups (diabetic patient or not). In particular, all patients here are females at least 21 years old of Pima Indian heritage. Non diabetic is labelled as 0 and diabetic is labelled as 1.

Solution Statement:

To solve this problem, I will be using one or more classification algorithms covered in the Udacity Machine Learning . First explore the data set and using visualizations which helps me to better understand the solution. Then we will find the accuracy score for each classification model then find best classification algorithm for Diabetes disease.

Benchmark model:

However the problem lies in finding a dataset where the results are given in such a fashion which is easily comparable with our classification values. In datasets it is intrinsically difficult to compare the scores given with our outputs. Therefore, we will use a simple algorithm like SVM as our benchmark model and try to improve upon its performance by using other algorithms like Decision Trees, KNN, Logistic Regression etc. If I classify the data applying different algorithms we got the accuracy_score with minimum 60% accuracy.

Evaluation metric:

Since it is a problem of disease classification we will generate a confusion matrix so that we can know the False Positives as well as the False negative and Here I am also going to use accuracy score , precision and f-score as evaluation metric for the selected models. Here f-score which model have the high value it is selected as the best model.

Project Design:

First of all, dataset will be accessed using Pandas and data exploration and visualization will be carried out.

Project is composed of different steps as follows.

Pre-Processing:

The first task is to read the data after that clean the data that is removing the data or treatment for missing values or removing duplicates etc. The missing values are replaced with their corresponding mean or median values. After that we will visualize the data to get some deep insights for interpretation. Then check the correlation among the features. Now the data is divided into features and labels. After that the features are scaled using MinMaxScaler to get them on the same level.

Training the data:

After completion of prepocessing steps, the data splits into training and testing datasets to train the classifier and check it's accuracy. So we import cross validation fom sklearn and thus it splits the data into training and testing datasets accordingly. Here I use the classification models Logisti Regression, Decision trees, Naive Bayes, AdaBoost, KNN. After training the data we test each model with testing data. Then we find out the f-score for each model. Finally I declare the model with high f-score as best model for the classification of Diabetes disease.