Machine Learning Engineer Nanodegree
Capstone Project Proposal
Karna Venkata Triveni
Feburary2nd,2019
Proposal:
Classifying Heart Disease Dataset
Domain Background:
History:
Across the World, Dead due to Heart disease is very common. About 610,000 people die of heart disease in the United States every year–that's 1 in every 4 deaths. Heart disease is the leading cause of death for both men and women. More than half of the deaths due to heart disease in 2009 were in men.

Heart disease is considered as one of the top preventable causes of the death in the United States. Some genetic factors can contribute, but the disease is largely attributed to poor life style habits

So,I am going to predict the rate of heart diseases in this project

One recent research paper based on the heart disease reference link: https://www.researchgate.net/publication/328031918_Machine_Learning_Classification_Techniques_for_Heart_Disease_Prediction_A_Review#pf7

Applications:

This project is predicting the heart rate of the dataset and it will helps for the government to take prevention methods i.e making awareness and how to follow the diet control programs etc.It helps for the total analysis of the rate of heart disease persons

Problem Statement:

The main aim of my project is to predict the rate of heart disease.For this I selected the data set compiled from a wide range of sources and made publicly available by the United States Department of Agriculture Economic Research Service (USDA ERS).So,My goal is to predict the rate of heart diseases in U.S.Here I am using the classification models to find the accuracy of each model and select the best model which will have high accuracy to predict the rate of heart diseases.Here the input parameters are the training data and the output will either 0 or 1i.e having heart heart disease or not

DataSets And Inputs:

This datasets contains 76 attributes, but all published experiments refer to using a subset of 14 of them. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4.Here the total samples in the data set is 6278(training data and testing data).Here the target variable is The data type of heart_diseases_mortality_per_100k is an

integer, Output is integer value.The data set consists of Nan values for some features.I will remove it(clean the data)

Attributes:
> 1. age
> 2. sex
> 3. chest pain type (4 values)
> 4. resting blood pressure
> 5. serum cholestoral in mg/dl
> 6. fasting blood sugar > 120 mg/dl
> 7. resting electrocardiographic results (values 0,1,2)
> 8. maximum heart rate achieved
> 9. exercise induced angina
> 10. oldpeak = ST depression induced by exercise relative to rest
> 11. the slope of the peak exercise ST segment
> 12. number of major vessels (0-3) colored by flourosopy
> 13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

Solution Statement:

Here, I trying to predict the rate of the heart disease for the selected data set.For predicting the rate of heart disease we want to use the different classification models.Then, we will find the accuracy score for each classification model. I explore the data set with opencv and matplotlib.pyplot libraries.in this project.By using visualization helps me to better understand the solutions

Benchmark Model:

This step will be important because compare your final model with some of them and see if it got better, same or worse. Here accuracy score will be compared between the models and select the best one

Evaluation Metrics:

I want to use accuracy score as evaluation metric for prediction of rate of heart disease. Here I am predicting the accuracy score fot the selected models.

In multilabel classification, this function computes subset accuracy: the set of labels predicted for a sample must *exactly* match the corresponding set of labels in y_true.

Here accuracy score which model have the high value it is selected as the best model

Project Design:

The project is composed of different steps as follows:

Pre-processing:

First task is to read the dataset and perform visualizations on it to get some insights about the data. After reading the data clean the data i.e. removing unwanted data or replacing null values with some constant values or removing duplicates. Then finding the correlation for each features with the heart disease target variable

After Data Exploration, I want to split the total data into training, validation and testing sets and normalize the data to make it suitable for .Then applying the Classifying models and then predicting the accuracy score to the selected models

First step in training:

First, I want to choose a Benchmark model which will at least gives testing accuracy score around 50 % accuracy score.

Second step in training:

I want to apply classification models of my own and use on the data. I want to apply Support Vector Machine and Logistic Regression model then find the Accuracy score for both the models

Finally, I will declare the model which highest accuracy score on both training and testing data sets concluded as the best model for detecting the rate of heart diseases