Machine Learning Engineer Nanodegree

Capstone Project Proposal

Classifying Liver Disease dataset

Ragala Sreekala

February 4th,2019

Proposal:

Classifying Liver Disease dataset

Domain Background:

History:

Problems with liver patients are not easily discovered in an early stage as it will be functioning normally even when it is partially damaged. An early diagnosis of liver problems will increase patient's survival rate. Liver failures are at high rate of risk among Indians. It is expected that by 2025 India may become the World Capital for Liver Diseases. The widespread occurrence of liver infection in India is contributed due to deskbound lifestyle, increased alcohol consumption and smoking. There are about 100 types of liver infections.

1. A patient going to a doctor with certain symptoms.

2. The doctor recommending certain tests like blood test, urine test etc depending on the symptoms.

3. The patient taking the aforementioned tests in an analysis lab.

4. The patient taking the reports back to the reports back to the hospital, where they are examined the disease is identified

Reference Link: https://www.irjet.net/archives/V5/i4/IRJET-V5I4896.pdf

Applications:

This project Classifying Liver Disease Data set can be applied in any medical hospital to check the person is infected by liver disease or not.

To serve the medicinal community for the diagnosis of liver disease among patients, a graphical user interface will be developed using python.

The GUI can be readily utilized by doctors and medical practitioners as a screening tool for the liver disease.

Problem Statement:

Given a dataset containing various attributes of 583 Indian patients, define a classification algorithms. To apply different classification algorithms on the Indian patient liver disease dataset than choose the best algorithms based on the accuracy which can identify whether a person is suffering from liver disease or not.

Datasets and Inputs:

The dataset that I am working is downloaded from

https://www.kaggle.com/sharadhiv/indian-liver-patient-dataset-ilpd

The number of instances are 583. It is a multivariate data set, contain 10 variables that are age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and Alkphos. All values are real integers. This data set contains 416 liver patient records and 167 non liver patient records.The data set was collected from north east of Andhra Pradesh, India. Selector is a class label used to divide into groups(liver patient or not). This data set contains 441 male patient records and 142 female patient records. ‘is_patient ‘ label '1' representing presence of disease and '2' representing absence of disease.

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 liver 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 Naïve bayes as our benchmark model and try to improve upon its performance by using other algorithms like SVM, ensemble methods etc. If i classifies the data applying on 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 calculate the accuracy score as evaluation metric for prediction of rate of liver disease. Here I am predicting the accuracy score for the selected models. Here accuracy 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. Any missing value or outlier will be suitably dealt with. Then, dataset will be split into training and testing set.Then, I want to choose a Benchmark model which will at least gives testing accuracy score around 50 % accuracy score.

Finally, the best performing algorithm will be tested on the testing dataset and evaluation metrics will be calculated to witness the results.