Machine Learning Engineer Nanodegree

Capstone Project Proposal

Classifying Liver Disease dataset

MEKALA NAGA HARISH YADAV

February 6th,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.researchgate.net/publication/319983998_Analysis_of_classification_algorithms_for_liver_disease_diagnosis

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 for this problem is the ILPD (Indian Liver Patient Dataset) taken from the UCI Machine Learning Repository

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

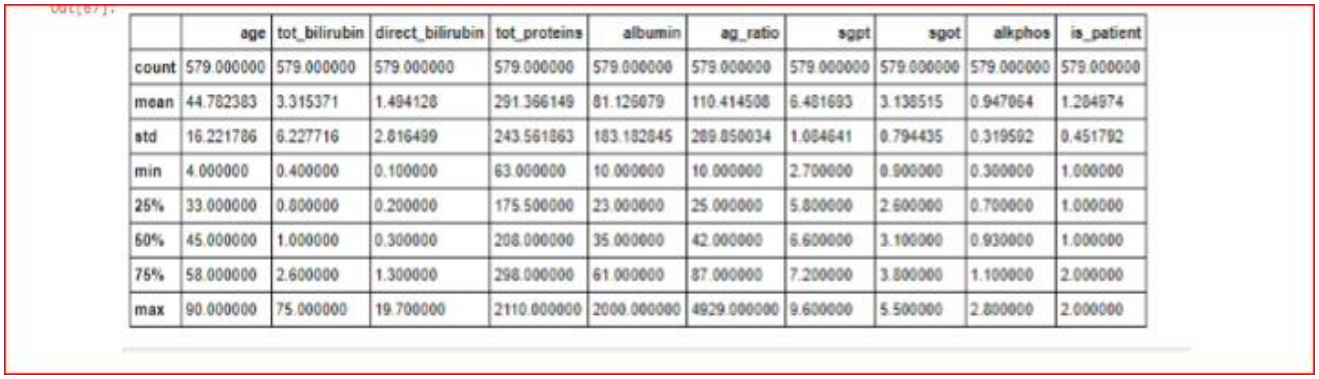
Based on chemical compounds(bilrubin,albumin,protiens,alkaline phosphatase) present in human body and tests like SGOT , SGPT the outcome mentioned whether person is patient ie needs to be diagnosed or not.

.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. ‘dataset‘ label '1' representing presence of disease and '2' representing absence of disease.

Here this data set contains 416 liver patient records and 167 non liver patient records collected from North East of Andhra Pradesh, India.

So,the data set is an unbalanced data set

Example of the data set:



	age	tot_bilirubin	direct_bilirubin	tot_proteins	albumin	ag_ratio	sgpt	sgot	alkphos	is_patient
count	579.000000	579.000000	579.000000	579.000000	579.000000	579.000000	579.000000	579.000000	579.000000	579.000000
mean	44.782383	3.315371	1.494128	291.366149	81.126079	110.414508	6.481693	3.138515	0.947064	1.284974
std	16.221786	6.227716	2.816499	243.561863	183.182845	289.850034	1.084641	0.794435	0.319592	0.451792
min	4.000000	0.400000	0.100000	63.000000	10.000000	10.000000	2.700000	0.900000	0.300000	1.000000
25%	33.000000	0.800000	0.200000	175.500000	23.000000	25.000000	5.800000	2.600000	0.700000	1.000000
50%	45.000000	1.000000	0.300000	208.000000	35.000000	42.000000	6.600000	3.100000	0.930000	1.000000
75%	58.000000	2.600000	1.300000	298.000000	61.000000	87.000000	7.200000	3.800000	1.100000	2.000000
max	90.000000	75.000000	19.700000	2110.000000	2000.000000	4929.000000	9.600000	5.500000	2.800000	2.000000

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.

I will be trying out Logistic Regression, knearest neighbours and one ensemble method. (Adaboost) for this project

Different combinations of hyperparameters for individual algorithms , like kernel, degree and C for SVM and weights, n_neighbours and algorithms for k-Nearest Neighbours will be tried across the training sets. Depending on their respective performances on the cross-validation sets, the best algorithm with appropriate hyperparameter tuning will be finalised as the solution.

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 Naïve bayes,SVM, ensemble methods etc. If i classifies the data applying on different algorithms we got the accuracy_score and f-score with minimum 50% 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 ,f-score,precision and recall as evaluation metric for prediction of rate of liver disease. Here I am predicting the accuracy score and f-score for the selected models. Here accuracy score and f-score which model have the high value it is selected as the best model. Additionally, we will use the F-scores which take both precision and recall into account, like the F-beta score: $F\beta = (1 + \beta2)(\text{Precision} \times \text{Recall})/(\beta2 \times \text{Precision} + \text{Recall})$

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 fit the data into the models which I was selected (SVM,LogisticRegression and AdaBoost) and finding the accuracy score and f-score for the models. Then, I want to choose a Benchmark model which will at least gives testing accuracy score and f-score around 50 % .

Finally, the best performing algorithm will be tested on the testing dataset and evaluation metrics will be calculated to witness the results. Here,comparing the accuracy score and f-score of the three models ,which model got the highest fscore and accuracy score that model is selected as the best model for predicting the liver disease