Capstone 1 Project Proposal: Liver Disease Prediction

Nils Madsen

Problem

The liver performs many critical life functions in the context of digestion and detoxification of the blood. 'Liver disease' is a broad term which encompasses the many forms of liver malfunction that lead to illness. Liver disease needs to be diagnosed early and treated to prevent permanent liver damage (cirrhosis), which can lead to death or require drastic and risky medical intervention in the form of liver transplant.

The process of diagnosing liver disease begins with the identification of abnormal results on one or more blood tests of liver function, but blood tests alone are not considered sufficient to conclusively diagnose or rule out liver disease. The current gold standard for diagnosis of liver disease is liver biopsy, with inspection of the biopsied tissue under a microscope by a pathologist. However, liver biopsy is an invasive procedure which is not without risk, and can lead to substantial cost to the healthcare system.

Goal and Utility

The goal of this project is to predict the presence or absence of liver disease from a suite of blood tests and other diagnostic metrics. These measurements are much less invasive and much less costly than a liver biopsy, since they at most require blood to be drawn for a blood test. Thus, the prediction of liver disease from these metrics may help physicians make more informed decisions about whether or not a liver biopsy is justified in a given patient. This algorithm may also be useful to insurance companies during the prior-authorization process for liver biopsy. This could help to protect patients from unnecessary risk, and lower the burden of liver disease diagnosis on the healthcare system.

Data

The data set for this project consists of 583 observations of patients near Andhra Pradesh, India. The observations include 416 liver disease patients and 167 patients without liver disease. The data set has 10 diagnostic metrics: age, gender, total bilirubin, direct bilirubin, alkaline phosphatase, alanine aminotransferase, aspartate aminotransferase, total protein, albumin, and albumin-globulin ratio. The final column is the group membership of the patient (i.e. liver patient or non-liver patient).

The data set was discovered on Kaggle, and was gathered from the UCI Machine Learning Repository. This page contains information about the data set, as well as the academic source and relevant scientific publications:

https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)#

Approach

This is a supervised classification task. The 10 diagnostic metrics listed above will be used as the predictive variables, while the group membership of the patient will be the dependent variable. Random Forest and Support Vector Machine are two machine learning algorithms that are well-suited to classification tasks, and a comparison of performance between the two algorithms for this problem could be very informative and lead to the best outcome. Cross-validation, with multiple divisions of the data set into training and validation sets, will ensure that the outcome of the training is as reliable as possible.

Deliverables

The deliverables for this project will include the source code, data set, a slide deck, and a paper outlining the purpose, approach, and results of the project.