Capstone 1 Project: Liver Disease Prediction
Data Wrangling/Cleaning

Nils Madsen

Description of 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 (ALP), alanine aminotransferase (ALT), aspartate aminotransferase (AST), total protein, albumin, and albumin-globulin ratio (a-g ratio). The final column is the group membership of the patient (i.e. liver patient or non-liver patient).

Data Types
The data types in the raw data already look clean. There are no issues like numeric variables being stored as strings. For performance purposes, the sex and group columns could be converted to categorical type. However, I am storing the cleaned data as another csv file, so making that conversion here would be premature. Once the cleaned data is loaded into another python script for the purposes of data exploration or machine learning, these columns can be properly cast as categorical variables.

Missing Values
There are 4 observations that have incomplete data; specifically, these patients are missing data for a-g ratio. Since there are only 4 observations with incomplete information, dropping these observations is unlikely to bias the overall dataset in a meaningful way. However, since only one field is missing from these observations, dropping the entire row seems like a waste of data.

Another option is to drop the entire variable. However, this option is usually used in cases where a large portion of the field is missing. Also, removing an entire variable might reduce the effectiveness of machine learning.

Given the small number of missing values, and the goal of predicting the presence of liver disease, I chose to fill the missing values by imputing from the mean of the relevant liver disease group. This should give the machine learning algorithm slightly more data to learn from, while preserving the a-g ratio distributions of the two groups.

I elected to use the mean rather than the median for the imputation because there were only a few outliers on a-g ratio, and the mean and median were very similar for both groups.

Duplicate Rows

There are 13 duplicate rows in the data set. Unfortunately, there is no column giving a unique identifier for each patient, so there is no way to determine whether these rows are actually duplicated, or if they refer to multiple patients who happened to have exactly the same measures along all the variables. Because of this, I have elected to trust the original data set and include the duplicate rows.

Outliers

An inspection of the variable distributions reveals a significant amount of outliers present in total bilirubin, direct bilirubin, ALP, AST, and ALT. While this might be alarming in other types of data, the presence of outliers in these measures is not surprising. In fact, it is expected in data on liver disease.

The molecules being measured by these tests are normally kept at low concentrations in the blood. When the liver becomes damaged, these measures can rise by orders of magnitude, because there is a failure of the body's ability to keep them low. In addition, the scale of the increase in certain markers is used by doctors to predict what kind of liver disease is present. For example, an ALT result 2-3x higher than the normal maximum might indicate alcoholic or fatty liver, while a measure 20x higher could point to viral hepatitis or poisoning.

A small number of outliers are present in total protein, albumin, and a-g ratio. It is not clear what justification could be made for removing them; there is no way to judge whether or not the measurements are legitimate or errors. Also, the outliers are not dramatically larger or smaller in these fields, as are the outliers for bilirubin and the liver enzymes.

For these reasons, I have chosen not to remove any outliers from the data set.