Capstone 1: Liver Disease Prediction
Milestone Report

Nils Madsen

Introduction

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 set of blood tests and other diagnostic measures. These measurements are much less invasive and costly than a liver biopsy, since at most they require blood to be drawn for a blood test.

The prediction of liver disease from these metrics can help physicians in multiple ways during the diagnosis process. If the algorithm can give a confident positive determination (that a patient has liver disease), it may allow doctors to reach a faster diagnosis, and allow patients to avoid unnecessary testing for other diseases. If the algorithm is able to make a confident negative determination (that a patient does not have liver disease), doctors can avoid putting patients through costly and invasive liver biopsies. In any case, this algorithm could help reduce some of the uncertainty doctors face with ambiguous or mixed test results, by learning from a much larger number of cases than could ever be seen by a single physician.

This algorithm may also be useful to insurance companies during the prior-authorization process for invasive medical tests. This could help to protect patients from unnecessary risk, and lower the burden of 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 controls. The data set has 10 diagnostic metrics: age, sex, 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 control).

The data set was discovered on Kaggle, and 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)#

This approach is not restricted simply to liver disease, however. Any disease that has blood tests or other metrics used in its diagnosis could be approached in this way.

Data Cleaning

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 patients with incomplete information, dropping these observations entirely would be unlikely to bias the dataset in a meaningful way. However, since only one field is missing, dropping entire rows seems like a waste of data. Therefore, I elected to fill the missing values by imputing from the mean of the population. This should give the model more information to learn from, while not affecting the population distribution of a-g ratio.

Duplicate Rows
There are 13 duplicate rows in the data set. Unfortunately, there is no 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 have the same measures across all the variables. Because of this, no duplicate rows have been removed.

Outliers
An inspection of the variable distributions (see Appendix) 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.

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 several fold, 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 is more specific 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.

Initial Findings

The diagnostic measures appear to have independent predictive value
An investigation of the correlation structure of the predictive measures in the data set indicates that most of the measures are not highly correlated to each other (Figure 1), but almost all are related to liver disease diagnosis in a statistically-significant way (Table 1).

diagnostic metric	liver patient mean	control mean	t-statistic	p-value
age	46.153846	41.239521	3.342369	8.840632e-04
total bili	4.164423	1.142515	5.441441	7.801431e-08
direct bili	1.923558	0.396407	6.118790	1.734103e-09
ALP	319.007212	219.754491	4.534141	7.027487e-06
ALT	99.605769	33.652695	3.992646	7.371781e-05
AST	137.699519	40.688623	3.705217	2.313901e-04
total protein	6.459135	6.543114	-0.844354	3.988191e-01
albumin	3.060577	3.344311	-3.941755	9.074361e-05
a-g ratio	0.914337	1.028588	-3.965115	8.251144e-05

Table 1: T-test comparisons of the numeric diagnostic metrics between controls and liver patients

Age, total and direct bilirubin, ALP, ALT, and AST are all higher in liver patients than in controls. Albumin and a-g ratio are lower in liver patients. Only total protein is not significantly different between controls and liver patients, indicating that this measure may have little utility in training the learning model.

Total and direct bilirubin are highly correlated, which is not surprising given direct bilirubin makes up a portion of total bilirubin. Similarly, albumin and a-g ratio are positively correlated, since albumin is the numerator in the albumin-globulin ratio.

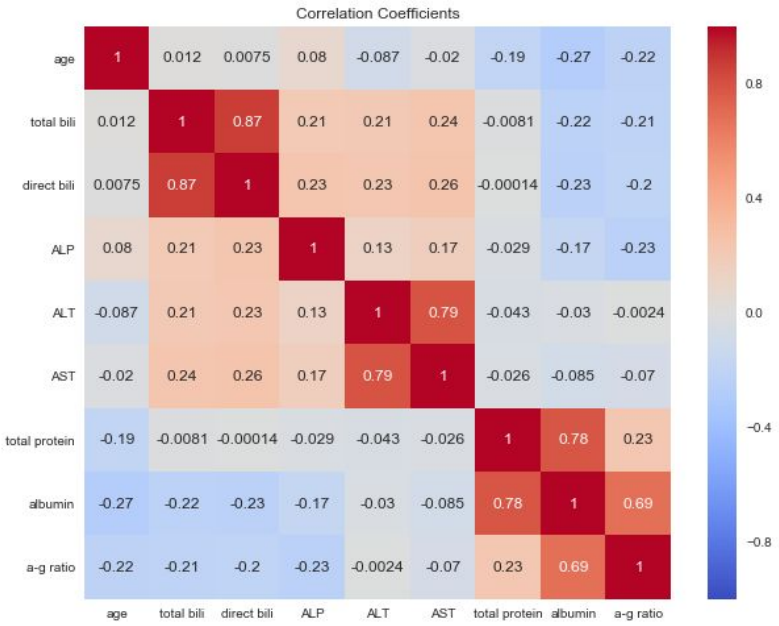


Figure 1: Correlations among the diagnostic measures (all subjects)

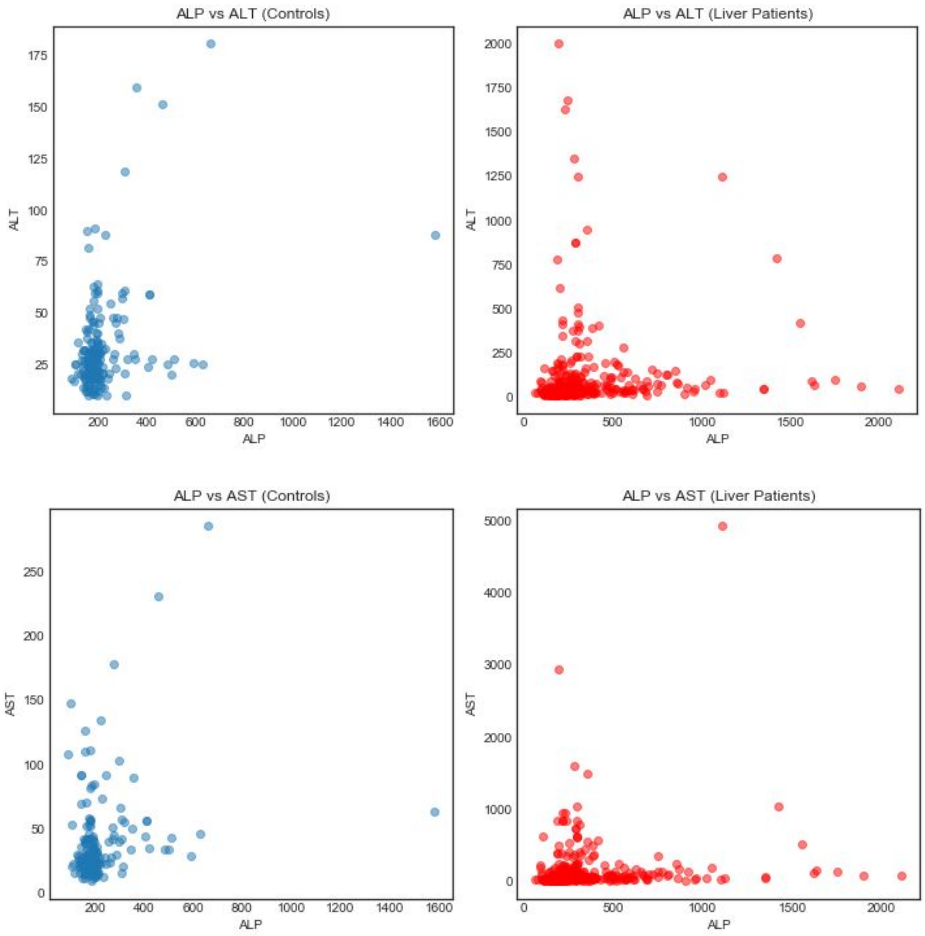


Figure 2: Distribution of ALP vs ALT and AST

AST and ALT are also highly correlated, indicating that they have overlapping (somewhat redundant) value in diagnosing liver disease.

ALP has only weak correlation with the other two liver enzymes (AST and ALT). Furthermore, very high levels of ALP in the blood are not often seen with very high levels of ALT or AST in liver patients (Figure 2), suggesting that ALP and ALT/AST may be elevated in different forms of liver disease.

From the above exploration, there appear to be 5 meaningful clusters of correlated measures in the dataset: age, the bilirubin measures (total and direct), ALP, the transaminases (ALT and AST), and the albumin measures (albumin and a-g ratio). However, there are only weak correlations between these clusters, likely indicating a good amount of independent predictive power in this set of measures. This is a good thing, and should allow the machine learning algorithm to predict liver disease better than any of the individual measures in isolation.

The blood tests have low sensitivity for liver disease

One notable observation from close inspection of the diagnostic measure distributions (appendix 1) is that many liver disease patients have blood tests that are normal or close to normal. This indicates that these blood tests have relatively low sensitivity for liver disease; a normal result for these blood tests cannot rule out a liver disease diagnosis.

On the other hand, it appears that some of the blood tests have good specificity, or in other words a blood test result above a certain value virtually guarantees the presence of liver disease. These two conclusions, taken together, suggest that a machine learning algorithm trained on this data will be capable of high specificity but not high sensitivity.

The high specificity and low sensitivity somewhat reduce the value of this machine learning model; if the model cannot rule out liver disease, then it cannot help patients and doctors avoid unnecessary liver biopsies. However, a model with high specificity and low sensitivity still has value, because its use as a 'rule-in' test for liver disease could save some patients from undergoing unnecessary tests for other diseases in their doctor's differential diagnosis.

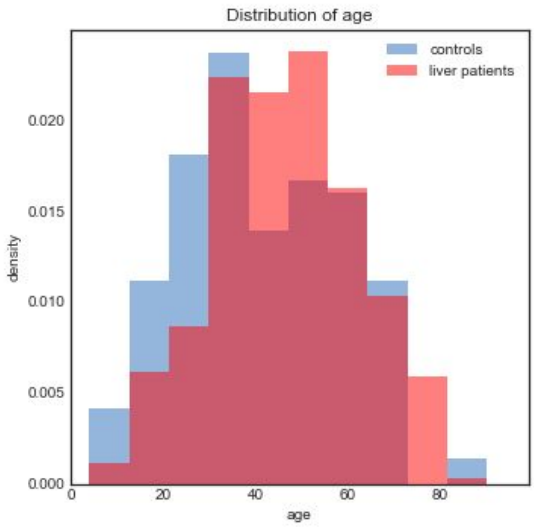
Sex is significantly associated with liver disease in this dataset

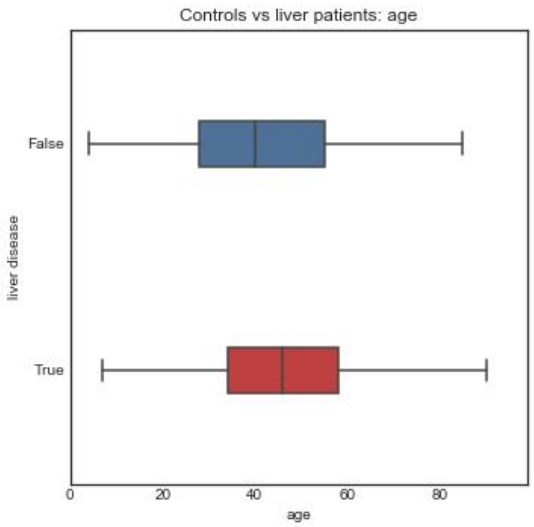
A proportions z-test indicates that there is a significantly greater proportion of men among the liver patients than among the controls (0.78 vs. 0.70; $p < 0.05$). This may suggest that male sex is associated with a higher risk of liver disease. However, this could also be a result of sampling bias, if for some reason male liver patients were easier to recruit than female liver patients, or female controls were easier to recruit than male controls. Large and carefully-collected data sets are needed to draw strong conclusions on this kind of question.

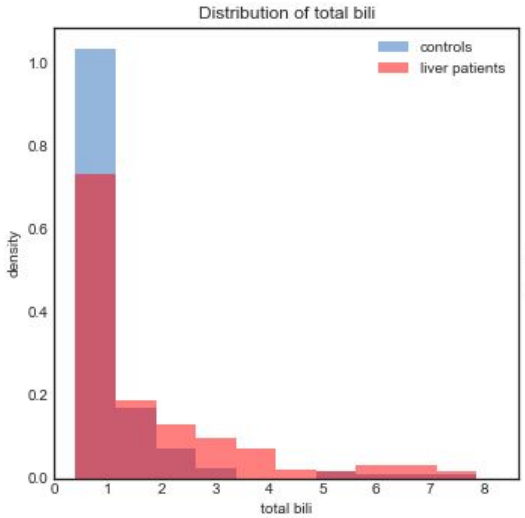
Next Steps

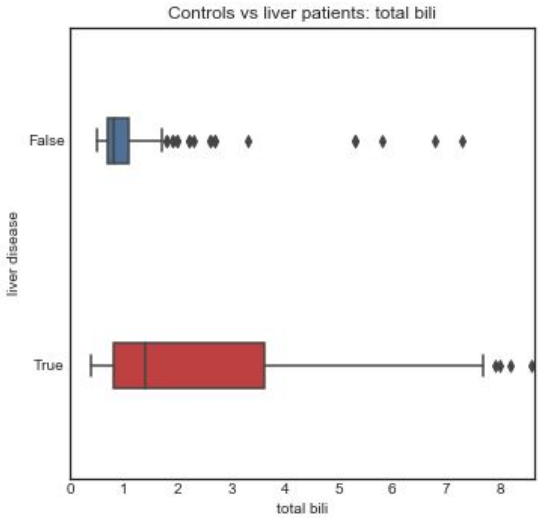
Training the machine learning model is the next step for this project. This is a supervised classification task, so Random Forest and Support Vector Machine are two algorithms that would be well-suited. A comparison of performance between these algorithms will be performed to find the best possible model. 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.

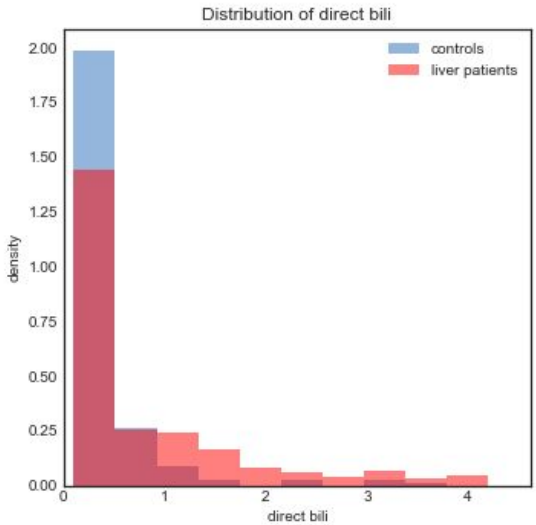
Appendix: Distributions of the diagnostic measures

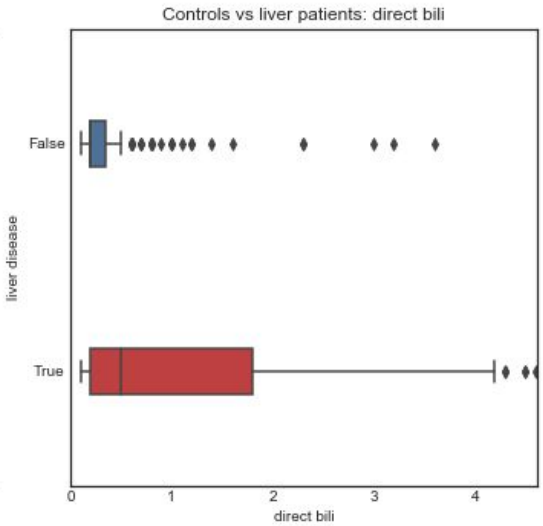


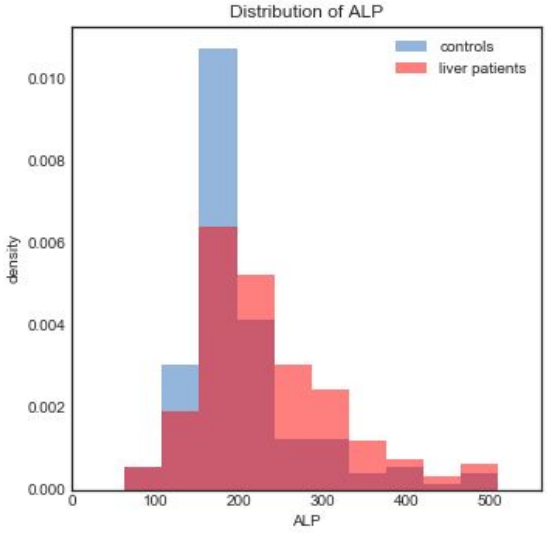


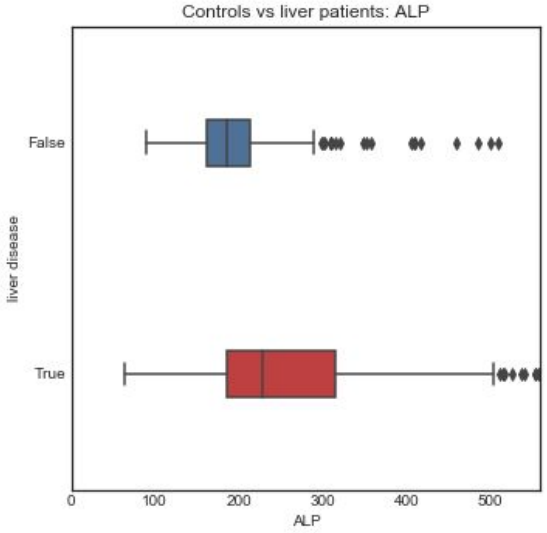


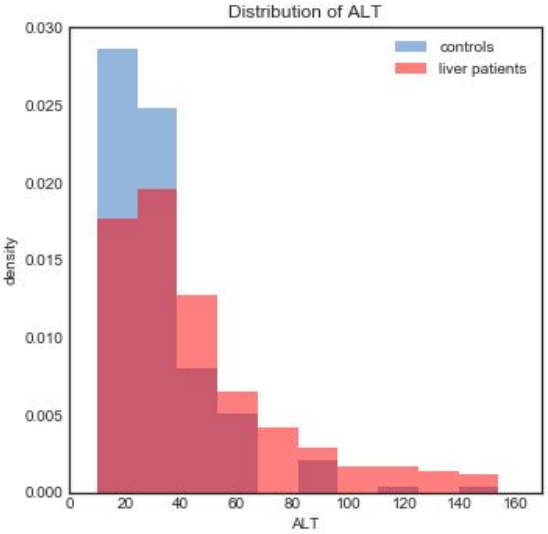


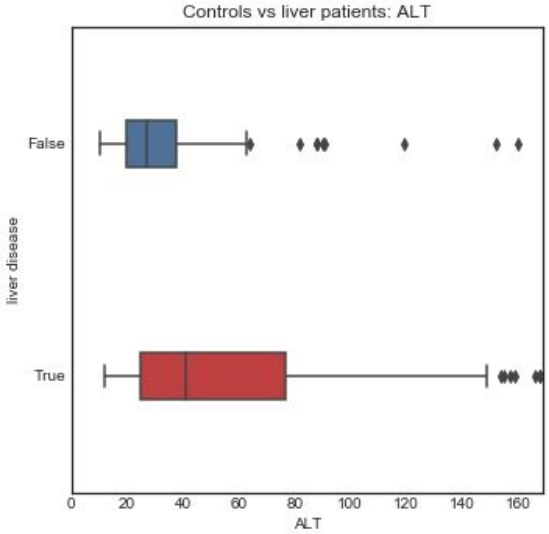


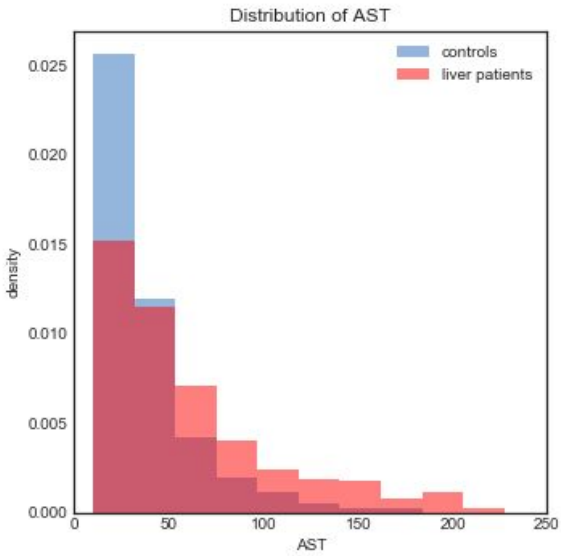


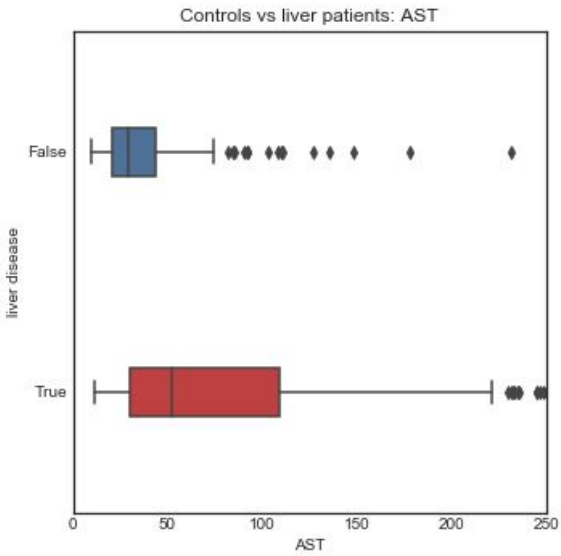

Distribution of ALP
controls
liver patients
density
ALP

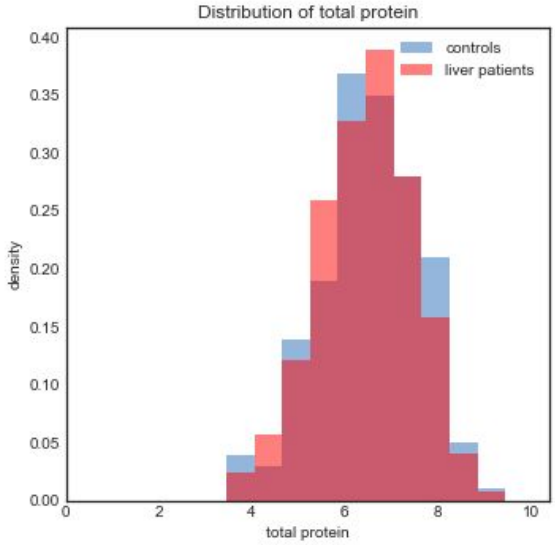

Controls vs liver patients: ALP
False
True
liver disease
ALP

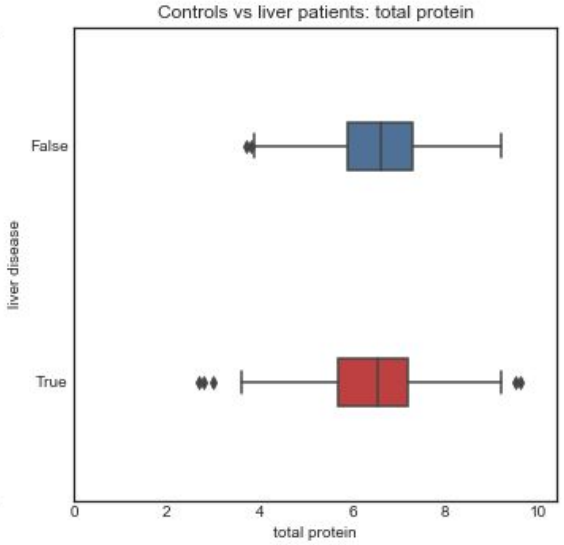

Distribution of ALT
controls
liver patients
density
ALT

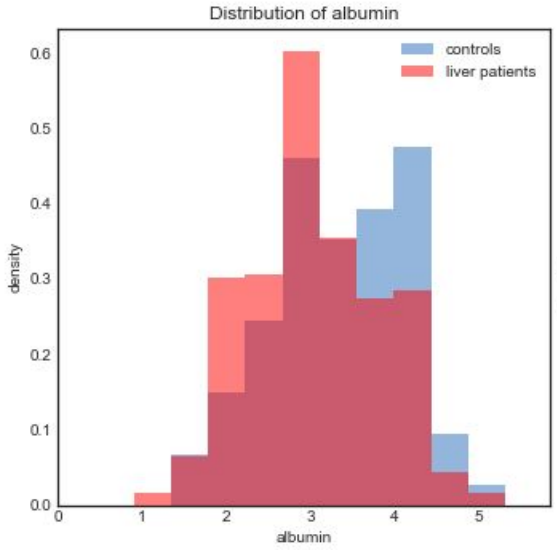

Controls vs liver patients: ALT
False
True
liver disease
ALT

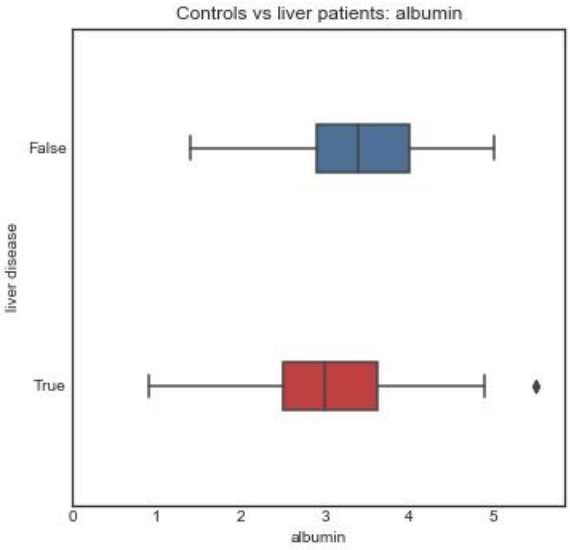

Distribution of AST
controls
liver patients
density
AST

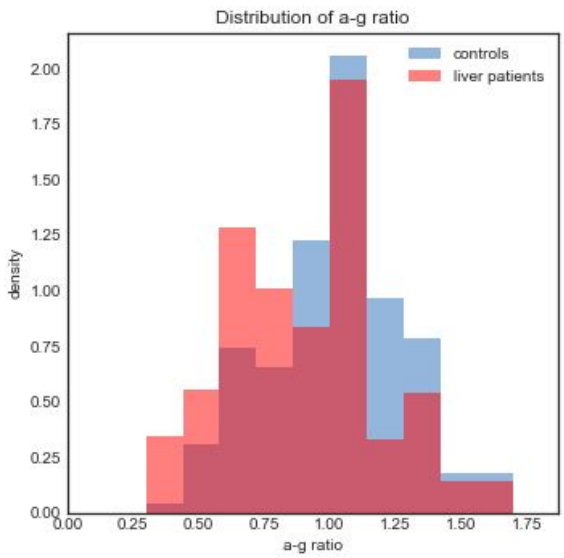

Controls vs liver patients: AST
False
True
liver disease
AST

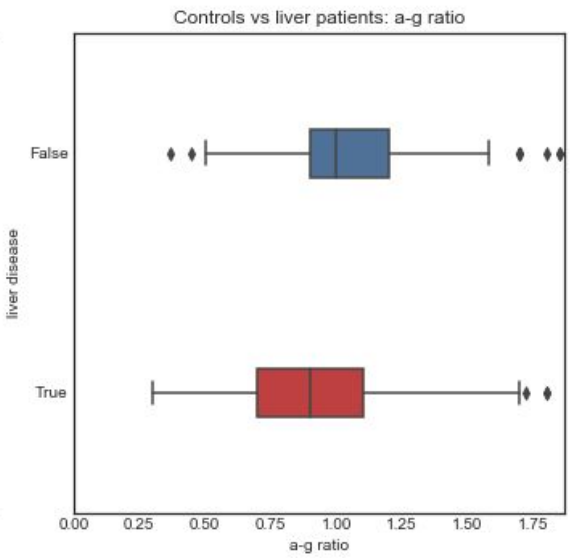

Distribution of total protein
controls
liver patients
density
total protein


Controls vs liver patients: total protein
False
True
liver disease
total protein


Distribution of albumin
controls
liver patients
density
albumin


Controls vs liver patients: albumin
False
True
liver disease
albumin


Distribution of a-g ratio
controls
liver patients
density
a-g ratio


Controls vs liver patients: a-g ratio
False
True
liver disease
a-g ratio