Predicting Patients with Liver Disease

Capstone Project 1 Milestone Report, Springboard

Predicting whether a patient has liver disease or not based on set of records can significantly reduce burden on doctors in an effort to correctly identify liver disease. By applying predictive machine learning algorithms against the patients dataset, we can solve the problem of identifying such patients. In this first capstone project, I am going to use machine learning models and a dataset that contains 416 liver patient records and 167 non liver patient records collected from North East of Andhra Pradesh, India.

The dataset for patient records will be taken from UC Irvine (UCI) ML repository: (*Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science)*. The current dataset contains 10 columns with patient data and a resulting 'outcome' column with the patient liver diagnosis (where 1 means patient has a liver disease, 2 means no disease). Thus, this is a binary classification problem. The techniques used in the project can be applied to other datasets with liver patient records.

It is important to note that the dataset is relatively clean and easy to use. However, the dataset will still require some data cleansing, data wrangling and possibly feature engineering. The exploratory data analyses and statistical learning will be applied to find more insights from data. Based on the above mentioned details regarding the data, several relevant machine learning algorithms (such as logistic regression, random forest and others) will be applied for predicting a liver patient diagnosis.

Overall, liver disease caused death rate is among top 10 death causes in the lower-middle income and upper-middle income countries in the world. Thus, having an accurate model for predicting patient live disease on early stages based on their records can significantly improve the diagnosis and help in early disease preventive cares. Below are the charts provided by World Health Organization (WHO):

Top 10 causes of deaths in lower-middle-income countries in 2016
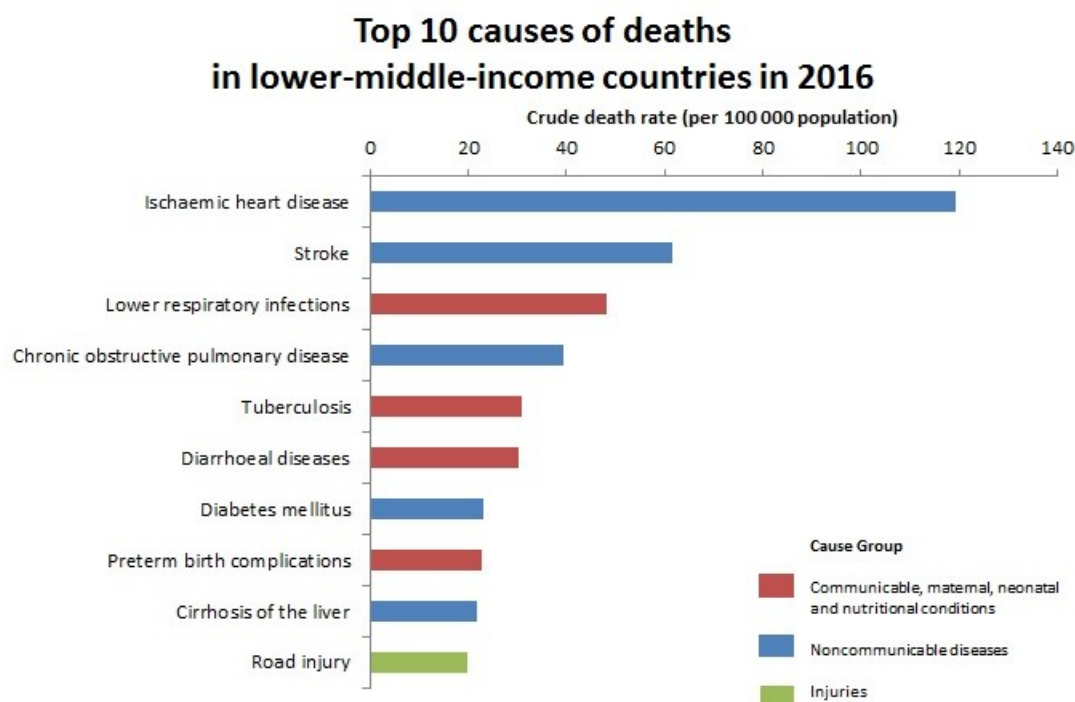
Top 10 causes of deaths in upper-middle-income countries in 2016

Crude death rate (per 100 000 population)
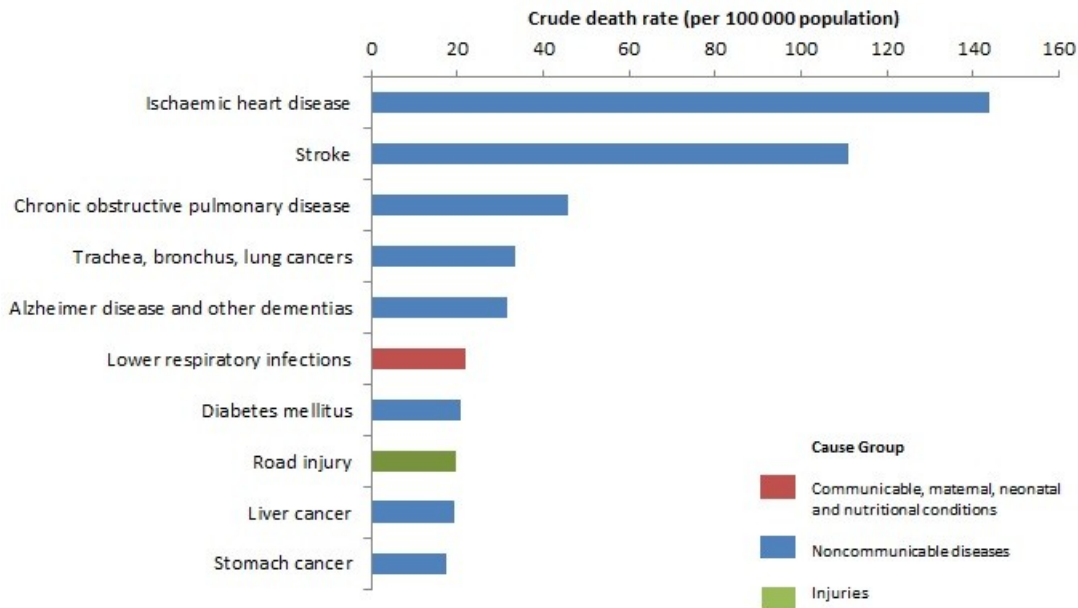
(Source: Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016. Geneva, World Health Organization; 2018.)

https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

Data Wrangling

Description and Objectives

The main goal of this project is to build a model that will accurately predict the liver disease for patients. Currently, the liver disease is one of the top 10 causes of deaths in developing countries, which is caused by a number of factors, such as diet, alcohol consumption and smoking. In worst cases, the disease can develop into cirrhosis. An early stage diagnosis of liver associated complications in patients can reduce the mortality rate. The machine learning techniques can help in predictions of liver disease and reduce burden on doctors.

Data Source

The dataset contains 416 liver patient records and 167 non liver patient records collected from North East of Andhra Pradesh, India[1]. The dataset consists of 10 columns and a resulting 'Dataset' column that contains the patient liver diagnosis (where 1 means patient has a liver disease, 2 means no disease). Also, the dataset contains 441 male patient records and 142 female patient records.

Data info

```
data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 583 entries, 0 to 582
Data columns (total 11 columns):
Age                          583 non-null int64
Gender                       583 non-null object
Total_Bilirubin              583 non-null float64
Direct_Bilirubin             583 non-null float64
Alkaline_Phosphotase         583 non-null int64
Alamine_Aminotransferase     583 non-null int64
Aspartate_Aminotransferase   583 non-null int64
Total_Protiens               583 non-null float64
Albumin                      583 non-null float64
Albumin_and_Globulin_Ratio   579 non-null float64
Dataset                      583 non-null int64
dtypes: float64(5), int64(5), object(1)
memory usage: 50.2+ KB
```

There are missing values in Albumin_and_Globulin_Ratio column. They will be filled with median value of the Albumin_and_Globulin_Ratio column.

The Gender column is categorical, meaning that the values from that column are neither continues nor discrete. The hot encoding technique will be applied to split that column into the two new columns (Female, Male) and then the Gender column will be dropped.

Data Statistics

```
data.describe().T
```

|	count	mean	std	min	25%	50%	75%	max
Age	583.0	44.746141	16.189833	4.0	33.0	45.00	58.0	90.0
Total_Bilirubin	583.0	3.298799	6.209522	0.4	0.8	1.00	2.6	75.0
Direct_Bilirubin	583.0	1.486106	2.808498	0.1	0.2	0.30	1.3	19.7
Alkaline_Phosphotase	583.0	290.576329	242.937989	63.0	175.5	208.00	298.0	2110.0
Alamine_Aminotransferase	583.0	80.713551	182.620356	10.0	23.0	35.00	60.5	2000.0
Aspartate_Aminotransferase	583.0	109.910806	288.918529	10.0	25.0	42.00	87.0	4929.0
Total_Protiens	583.0	6.483190	1.085451	2.7	5.8	6.60	7.2	9.6
Albumin	583.0	3.141852	0.795519	0.9	2.6	3.10	3.8	5.5
Albumin_and_Globulin_Ratio	579.0	0.947064	0.319592	0.3	0.7	0.93	1.1	2.8
Dataset	583.0	1.286449	0.452490	1.0	1.0	1.00	2.0	2.0

We observe from data statistics that features have different range of value. Thus, they need to be normalized before applying machine learning algorithms. This step is particularly important to avoid negative effect from the data on the performance of the algorithms. The log transform will be applied to

reduce any skewness in data and make it more normally distributed. Then, the min-max normalization will be applied to standardize the range of features of data.

References

[1] UC Irvine (UCI) ML repository: (Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science).

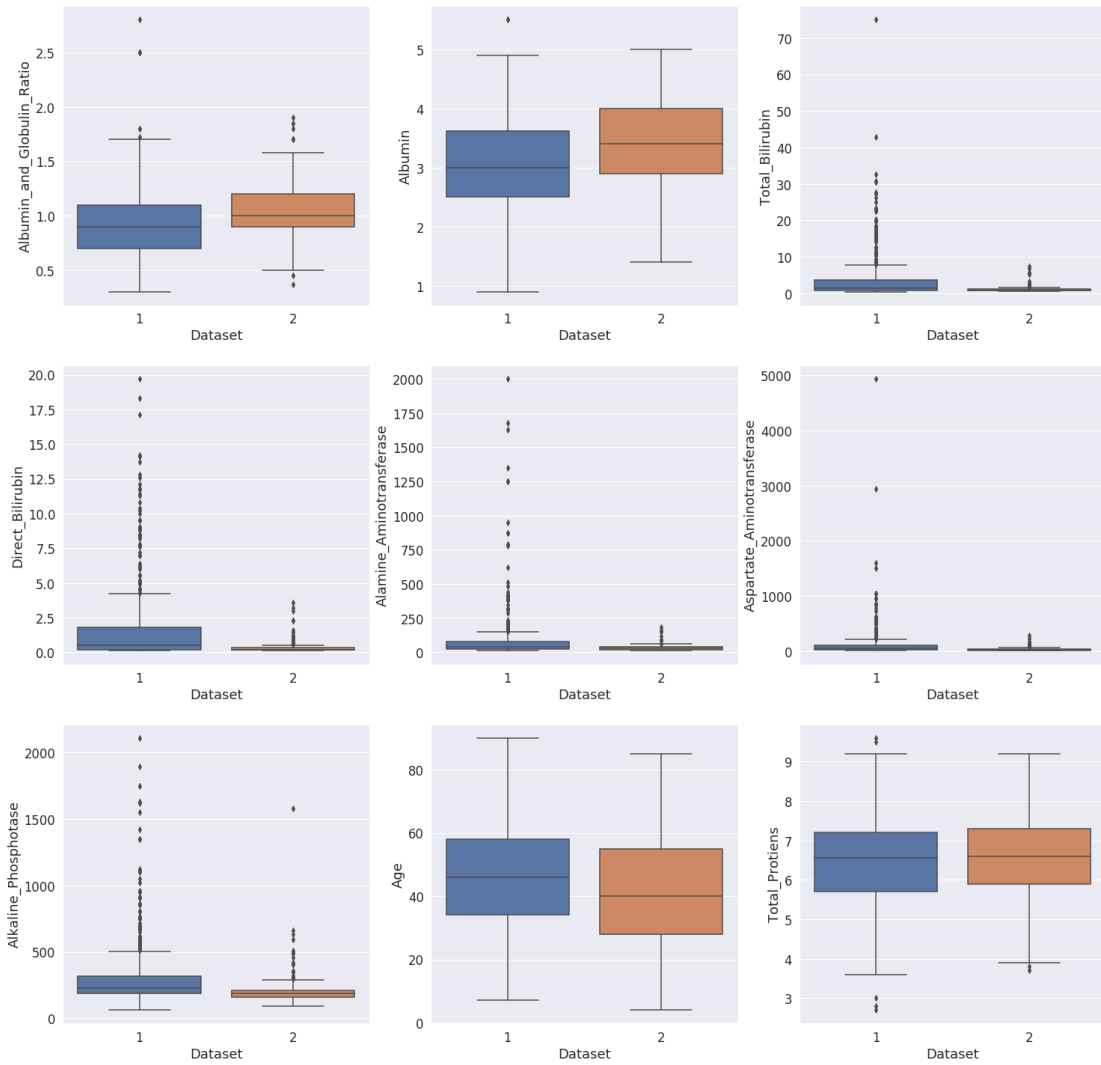# Exploratory Data Analyses (EDA):
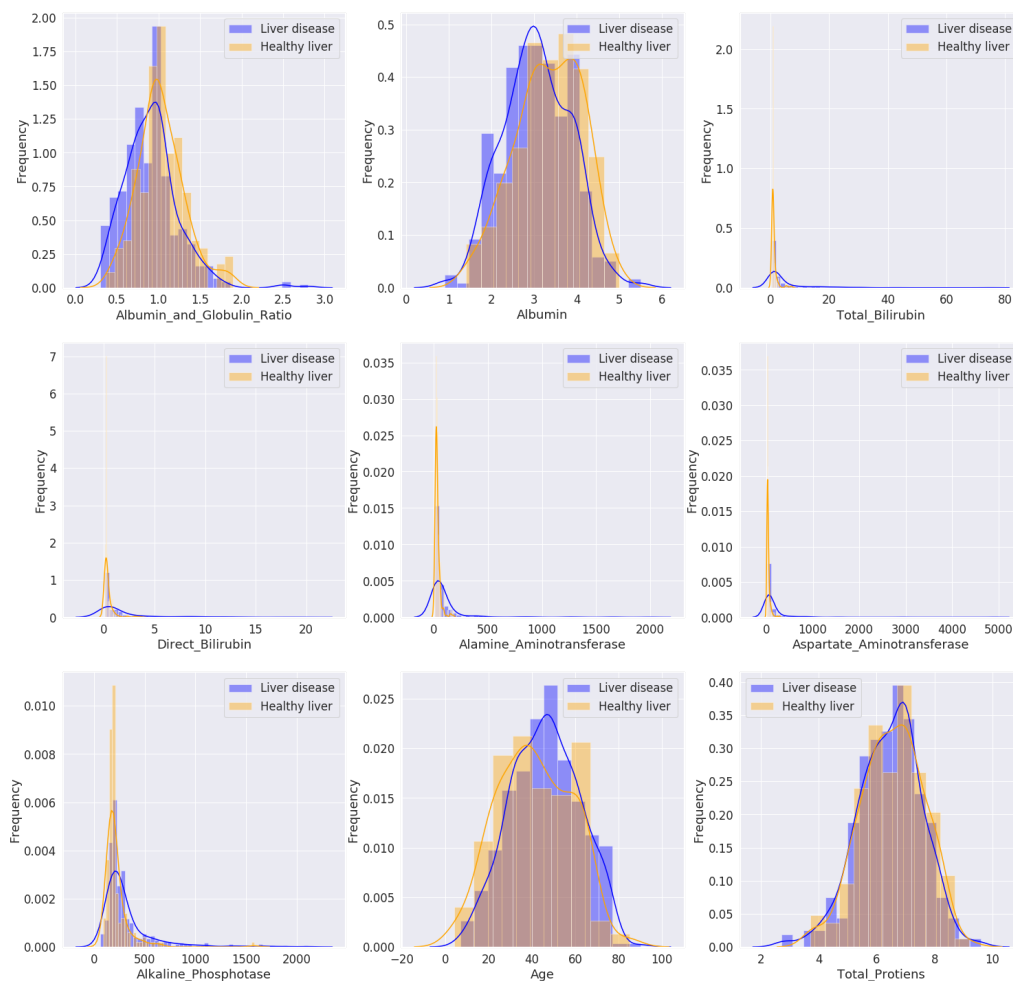
From the histogram plots we observe that the healthy patients have higher frequency of small values in narrow ranges compared to unhealthy patients.
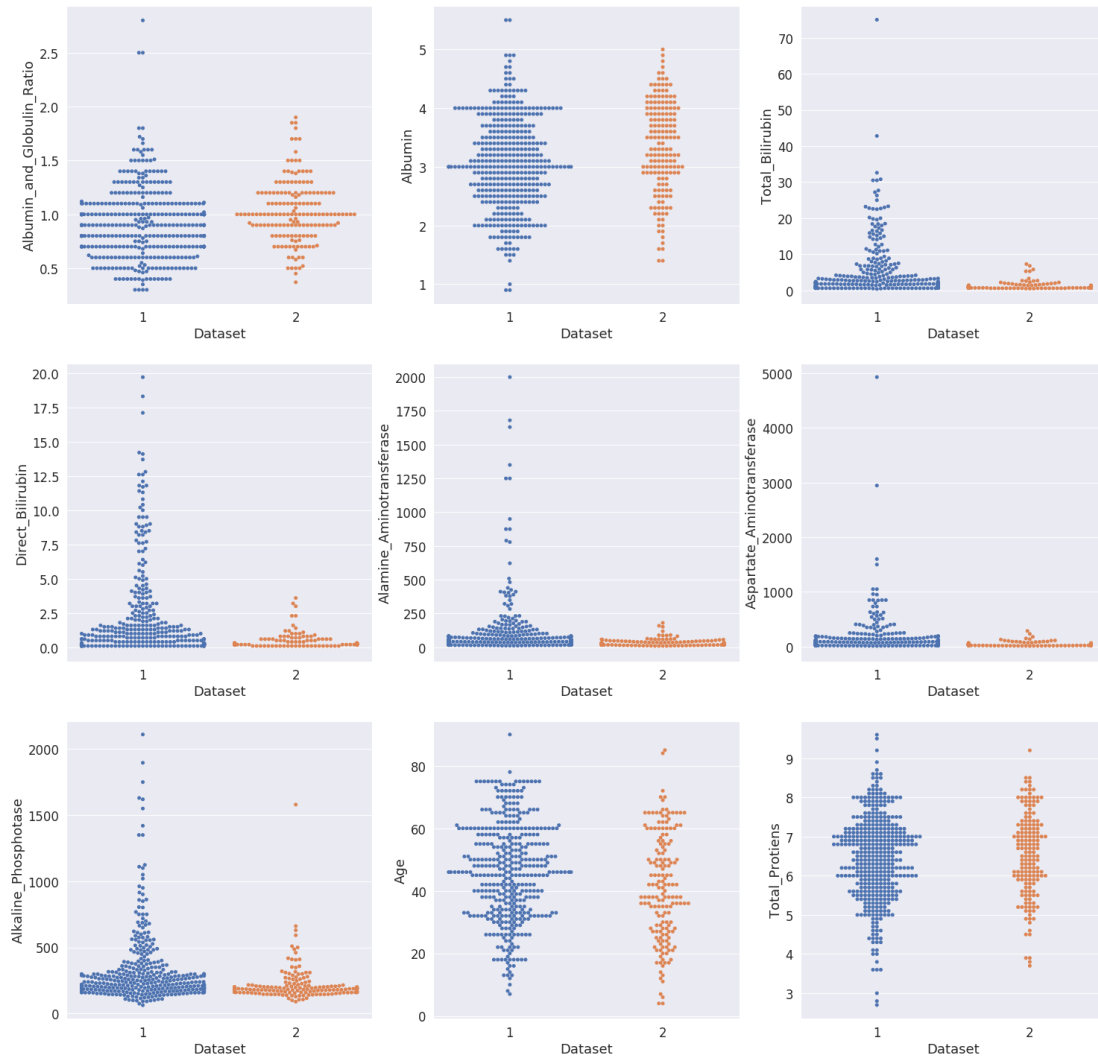
Swarm plots show outliers for some of the features for both healthy and unhealthy patients. However, we can not claim that these outliers represent erroneous data points.

From the above barplots, we can count outliers of some features for healthy patients to be within the quantile ranges of unhealthy patients. For example, Aspartate Aminotransfertase max value for healthy patients is within whisker extent of the corresponding unhealthy patients.