Project Report - Predicting Liver Disease

Bhathiya Maneendra Pilanawithana

2023-04-07

Introduction

The project is aimed at developing a predictive algorithm for liver disease using patient records from India, which are publicly available on Kaggle.com through the link "https://www.kaggle.com/datasets/uciml/indian-liver-patient-records"".

Liver disease is a significant health problem worldwide, and its early detection is crucial because it allows for prompt treatment, which can prevent the progression of the disease and potentially save lives. The liver is a vital organ that performs numerous functions in the body, such as filtering toxins from the blood, producing bile, and metabolizing drugs. Liver disease can develop over time and may not present any symptoms until it has progressed to an advanced stage. Early detection can help identify the disease before symptoms become severe and irreversible damage has occurred. Furthermore, some types of liver disease, such as viral hepatitis, can be highly contagious, making early detection even more important in preventing the spread of the disease to others. Additionally, early detection of liver disease can enable healthcare professionals to closely monitor and manage the condition, which can help reduce the risk of complications and improve overall health outcomes. This may involve lifestyle changes, medication, and in some cases, surgery or liver transplantation.

In this project, various machine learning will be applied and fine tuned to predict the likelihood of liver disease in patients.

Dataset Description

This is extracted from Kaggle.The data set contains 416 liver patient records and 167 non liver patient records collected from North East of Andhra Pradesh, India. Any patient whose age exceeded 89 is listed as being of age "90".

Columns:

1. Age of the patient
2. Gender of the patient
3. Total Bilirubin
4. Direct Bilirubin
5. Alkaline Phosphotase
6. Alamine Aminotransferase
7. Aspartate Aminotransferase
8. Total Protiens
9. Albumin
10. Albumin and Globulin Ratio
11. Dataset: field used to split the data into two sets (patient with liver disease, or no disease)

Analysis

First the csv file stored in “Raw-Dataset” subfolder is loaded in to a dataframe using following code.

```
dat <- read.csv("./Raw-Dataset/indian_liver_patient.csv")
```

It was observed that presence of liver disease is stated in the “Dataset” column existence of liver disease is denoted by 1 and non-existence is denoted by 2. Since this notation is counter intuitive, “Dataset” Column is renamed to “Disease” and existence and non-existence of liver disease is denoted by 1 and 0 respectively. This is achived by the following code chunk.

```
dat <- dat %>% mutate(Disease = ifelse(Dataset==1,1,0)) %>% select(-Dataset)
```

Now, the existence of liver disease is noted in “Disease” column intuitively using 1 and 0.