Liver Disease Prediction

Nishaal Ajmera

05/06/2020

Introduction

Patients with liver disease have been on the rise. Patients often result in liver transplant or die of the disease. Factors contributing to this include increased alcohol consumption, drugs consumption, unhealthy and fatty foods and inhalation of toxins. An important part of understanding this is disease is to diagnose patients with this disease as early as possible accurately.This dataset was obtained patients records in the North East of Andhra Pradesh, India.

The key goals of this project are:

- to help doctors diagnose patients with liver disease
- to investigate best machine learning model to predict patients with liver disease accurately

Data Mining

```
#Getting the dataset
liverurl="https://raw.githubusercontent.com/nishaalajmera/Indian-Liver-Disease-Capstone-/master/indian_l
liverdis<-read_csv(url(liverurl)) #reading and saving file from url into workable format
```

Data Modification

The original dataset has 416 liver disease patients and 167 non-liver disease patients. The data has been modified to remove NA's and it contains 414 liver disease patients and 165 non-liver disease patients. The dataset contains 11 variables (Age,Gender,Total Bilirubin, Direct Bilirubin,Alkaline Phosphotase,Alamine Aminotransferase, Aspartate Aminotransferase, Total Proteins, Albumin, Albumin to Globulin Ratio). The `dataset` column shows patients with [1] or without liver disease [2].

Table 1: First five rows of the data

Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase
65	Female	0.7	0.1	187	16
62	Male	10.9	5.5	699	64
62	Male	7.3	4.1	490	60
58	Male	1.0	0.4	182	14
72	Male	3.9	2.0	195	27

Table 2: Summary stats of the dataset

Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumin_and_Globulin_Ratio	Dataset
18	6.8	3.3	0.90	1
100	7.5	3.2	0.74	1
68	7.0	3.3	0.89	1
20	6.8	3.4	1.00	1
59	7.3	2.4	0.40	1

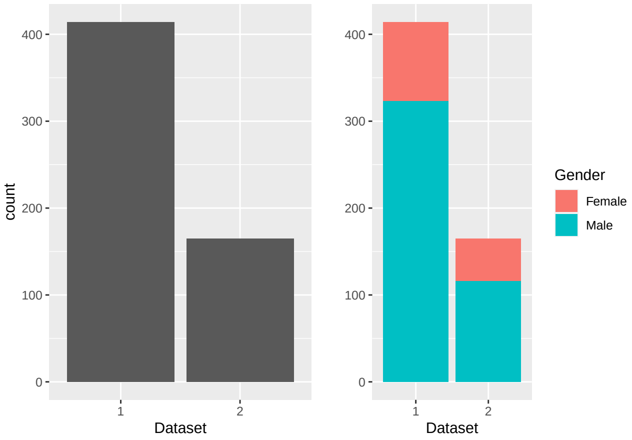
Age	Gender	Total_Bilirubin	Direct_Bilirubin
Min. : 4.00	Length:579	Min. : 0.400	Min. : 0.100
1st Qu.:33.00	Class :character	1st Qu.: 0.800	1st Qu.: 0.200
Median :45.00	Mode :character	Median : 1.000	Median : 0.300
Mean :44.78	NA	Mean : 3.315	Mean : 1.494
3rd Qu.:58.00	NA	3rd Qu.: 2.600	3rd Qu.: 1.300
Max. :90.00	NA	Max. :75.000	Max. :19.700

Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase
Min. : 63.0	Min. : 10.00	Min. : 10.0
1st Qu.: 175.5	1st Qu.: 23.00	1st Qu.: 25.0
Median : 208.0	Median : 35.00	Median : 42.0
Mean : 291.4	Mean : 81.13	Mean : 110.4
3rd Qu.: 298.0	3rd Qu.: 61.00	3rd Qu.: 87.0
Max. :2110.0	Max. :2000.00	Max. :4929.0

Total_Protiens	Albumin	Albumin_and_Globulin_Ratio	Dataset
Min. :2.700	Min. :0.900	Min. :0.3000	1:414
1st Qu.:5.800	1st Qu.:2.600	1st Qu.:0.7000	2:165
Median :6.600	Median :3.100	Median :0.9300	NA
Mean :6.482	Mean :3.139	Mean :0.9471	NA
3rd Qu.:7.200	3rd Qu.:3.800	3rd Qu.:1.1000	NA
Max. :9.600	Max. :5.500	Max. :2.8000	NA

Exploratory Analysis

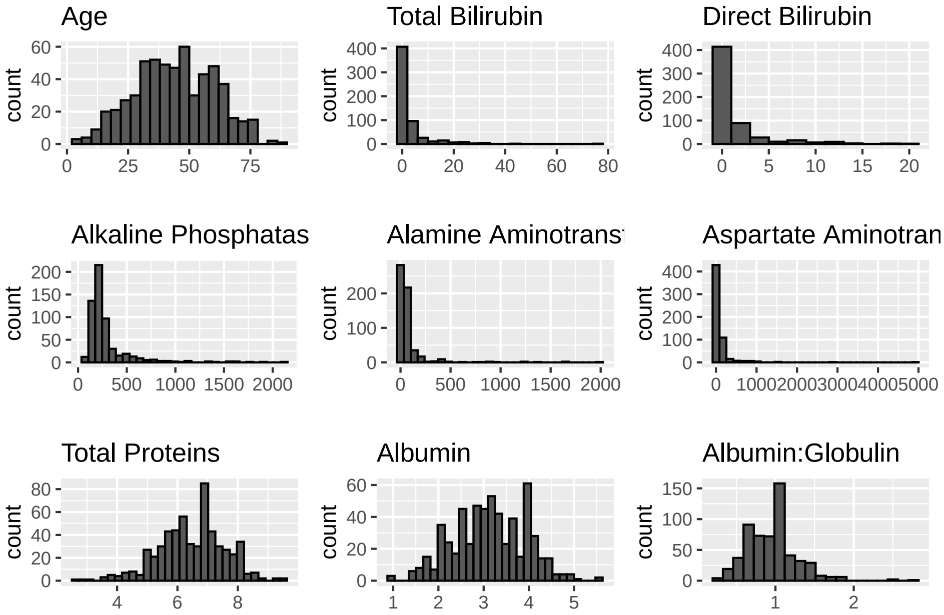
Liver patients and non-liver patients segregated by Gender



It is observed that in both groups there are less female patients compared to male patients

Distribution

The distribution of each continuous variable is shown below.



Some variables such as: Total Bilirubin, Direct Bilirubin, Alkaline Phophatase, Alamine Aminotransferase and Aspartate Aminotransferase have skewed distributions. This might be due to some clustering suggesting that the levels could be higher in one of the patient groups.

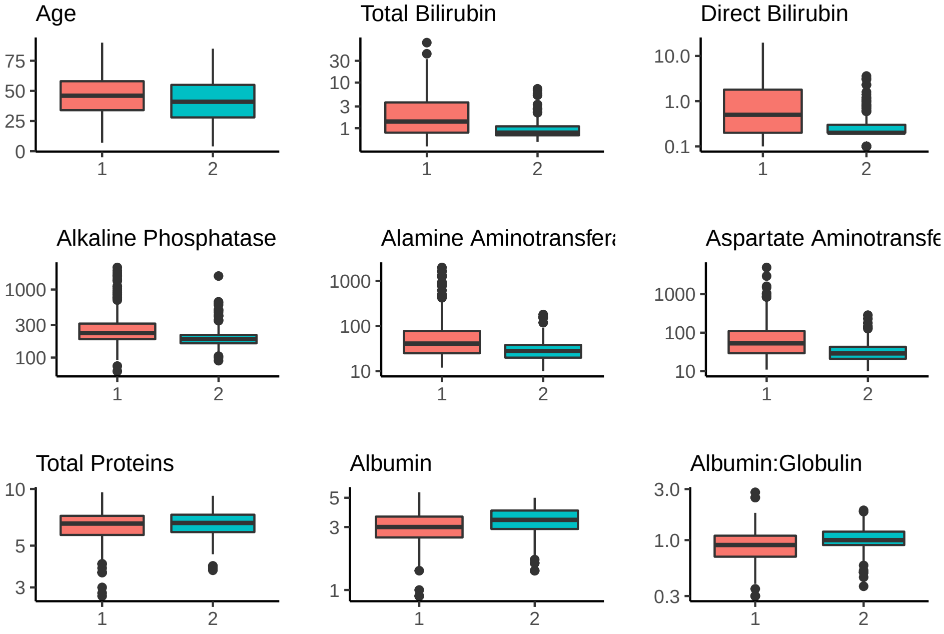
Normal Distribution Test

Shapiro-Wilk test is used here to check if continuous variables follow a normal distribution.
Null Hypothesis: Continuous variable follows a distribution pattern similar to normal distribution
Alternative Hypothesis: Continuous variable does not follow normal distribution

P-values for all variables are less than 0.05 therefore the null hypothesis is rejected. Going forward, non-parametric tests will be used to assess the statistical significance of the the data. Therefore median will be used as measure of central tendency and interquartile range will explain the variability of the data.

Data Analysis between the two groups of data; liver disease patients and non-liver disease patients

1 represents patients with liver disease and 2 represents patients with no liver disease. Below are boxplots to visualize any obvious differences. Log scale has been used to better visualize data.



The median of some variables show differences in the two groups. However this has to be further assessed.

Wilcoxon Signed Ranked Test

Wilcoxon Signed Ranked Test is a non-parametric test is used to compare two related samples.

All p-values except Total Proteins show that there is a significance between the liver disease and non-liver disease group. We will try some models without Total protein checking if it improves the accuracy.

Generating Training and Test Samples

```r
test_index<- createDataPartition(y=liverdis$Dataset,times=1,p=0.2,list=FALSE) #index of test set
train_set<- liverdis[-test_index,] #generating train set
test_set<- liverdis[test_index,] #generating test set
```

Model 1: Logistic Regression Model

This model uses logistic regression to predict the patient group. Here all the variables are used as predictors.

```r
#Model 1: Logistic Regression Model (all predictors)
fit_glm<- glm(Dataset~.,data=train_set,family="binomial") #Training algorithm
p_hat_glm<- predict(fit_glm,test_set,type = "response")
y_hat_glm<- ifelse(p_hat_glm>0.5,"1","2")
```

Table 6: Logistic Regression with all predictors

	x
Accuracy	0.25

It is seen that this model gives very poor accuracy. We wil try improving the model by removing some variables.

Model 2: Logistic Regression Model

In this model we are only using continuous variables to predict the dataset.

```
#Model 2: Logistic Regression Model (continuous variables)
fit_glm<- glm(Dataset~Age+ Total_Bilirubin + Direct_Bilirubin + Alkaline_Phosphotase + Alamine_Aminotran
p_hat_glm<- predict(fit_glm,test_set,type = "response")
y_hat_glm<- ifelse(p_hat_glm>0.5,"1","2")
```

Table 7: Logistic Regression with 9 predictors

	x
Accuracy	0.2413793

This model has reduced the accuracy.

Model 3: Logistic Regression

In this model, we will use the variables that have a skewed distribution. It is proposed that some the levels of some variables might be higher in a one group of patient.

```
#Model 3: Logistic Regression using variables that have a skewed distribution
fit_glm<- glm(Dataset~ Total_Bilirubin + Direct_Bilirubin + Alkaline_Phosphotase + Alamine_Aminotransfe
p_hat_glm<- predict(fit_glm,test_set,type = "response")
y_hat_glm<- ifelse(p_hat_glm>0.5,"1","2")
```

Table 8: Logistic Regression with 5 predictors

	x
Accuracy	0.2844828

There a has been a slight improvement of 13% compared to the first logistic regression model in the accuracy. We will try using a different model to improve the predictions.

Model 4: KNN model 1 (continuous variables)

The K-nearest neighbours model will be applied here to all the continuous variables. It is a non-parametric machine learning algorithm that is easy to apply to multiple dimensions.

```r
#Model 4: KNN Model 1
#Used all continuous variables
control<- trainControl("cv",number=10,p=.9)
train_knn<- train( Dataset ~ Age+ Total_Bilirubin + Direct_Bilirubin + Alkaline_Phosphotase + Alamine_Am
                   data = train_set, method = "knn",
                   tuneGrid= data.frame(k=seq(9,51,2)),
                   trControl = control)
```

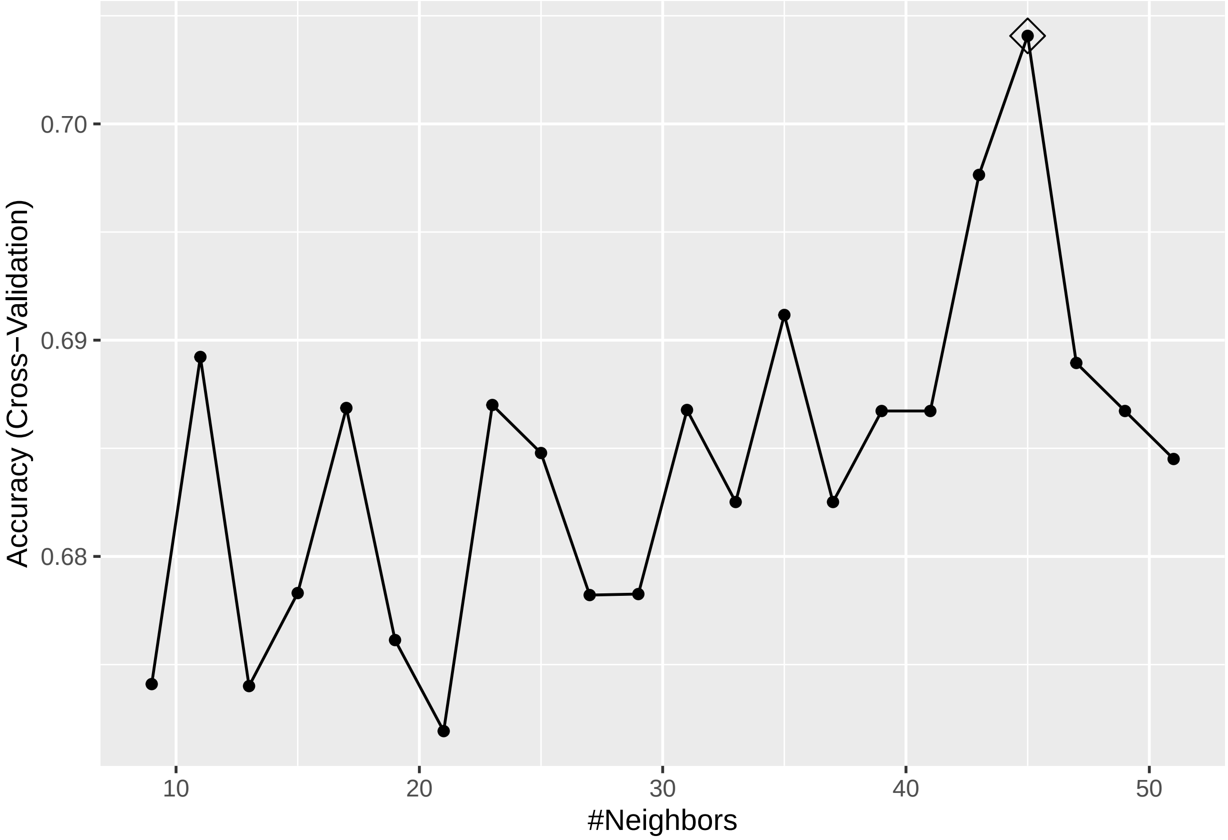


Table 9: Optimal K-nearest neighbours

	k
19	45

Table 10: KNN with 9 predictors

	x
Accuracy	0.7068966

It is seen that the accuracy improves greatly. However will try and get the accuracy as close to 100%.

Model 5: KNN Model 2

In this KNN model we have used on the variables that we significant in the Wilcoxon Signed Rank Test.

```
#Model 5: KNN Model 2
#Used significant variables (removed Total_Protiens)
control<- trainControl("cv",number=10,p=.9)
train_knn<- train( Dataset ~ Age+ Total_Bilirubin + Direct_Bilirubin + Alkaline_Phosphotase + Alamine_A
                   data = train_set, method = "knn",
                   tuneGrid= data.frame(k=seq(9,51,2)),
                   trControl = control)
```

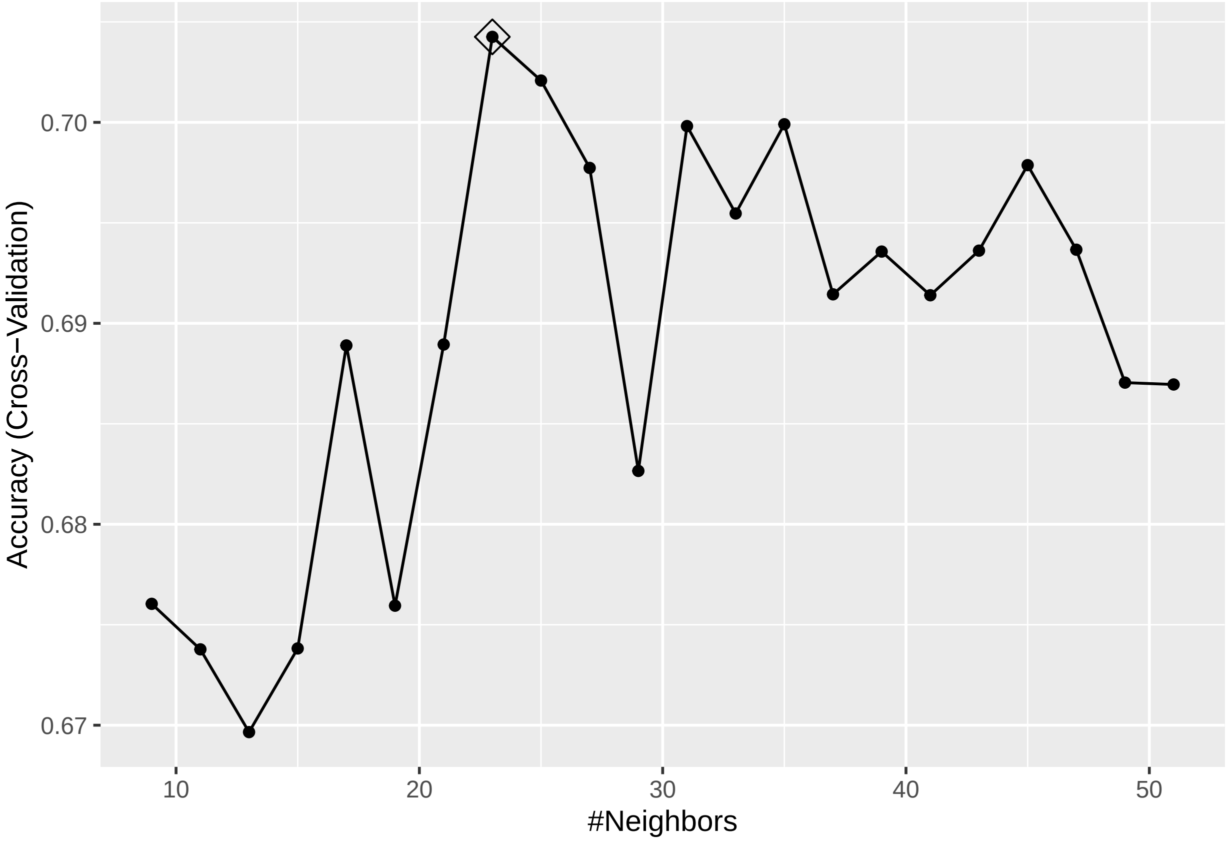


Table 11: Optimal K-nearest neighbours

	k
8	23

Table 12: KNN with 8 predictors

	x
Accuracy	0.6810345

The accuracy remains the same. Therefore, we will further add some change to improve it.

Model 6: KNN Model 3

In this KNN model we will use the variables that have a skewed distribution.

```
#Model 6: KNN Model 3
#Used significant variables (using skewed distribution variables)
control<- trainControl("cv",number=10,p=.9)
train_knn<- train( Dataset ~ Total_Bilirubin + Direct_Bilirubin + Alkaline_Phosphotase + Alamine_Aminot
                   data = train_set, method = "knn",
                   tuneGrid= data.frame(k=seq(9,51,2)),
                   trControl = control)
```

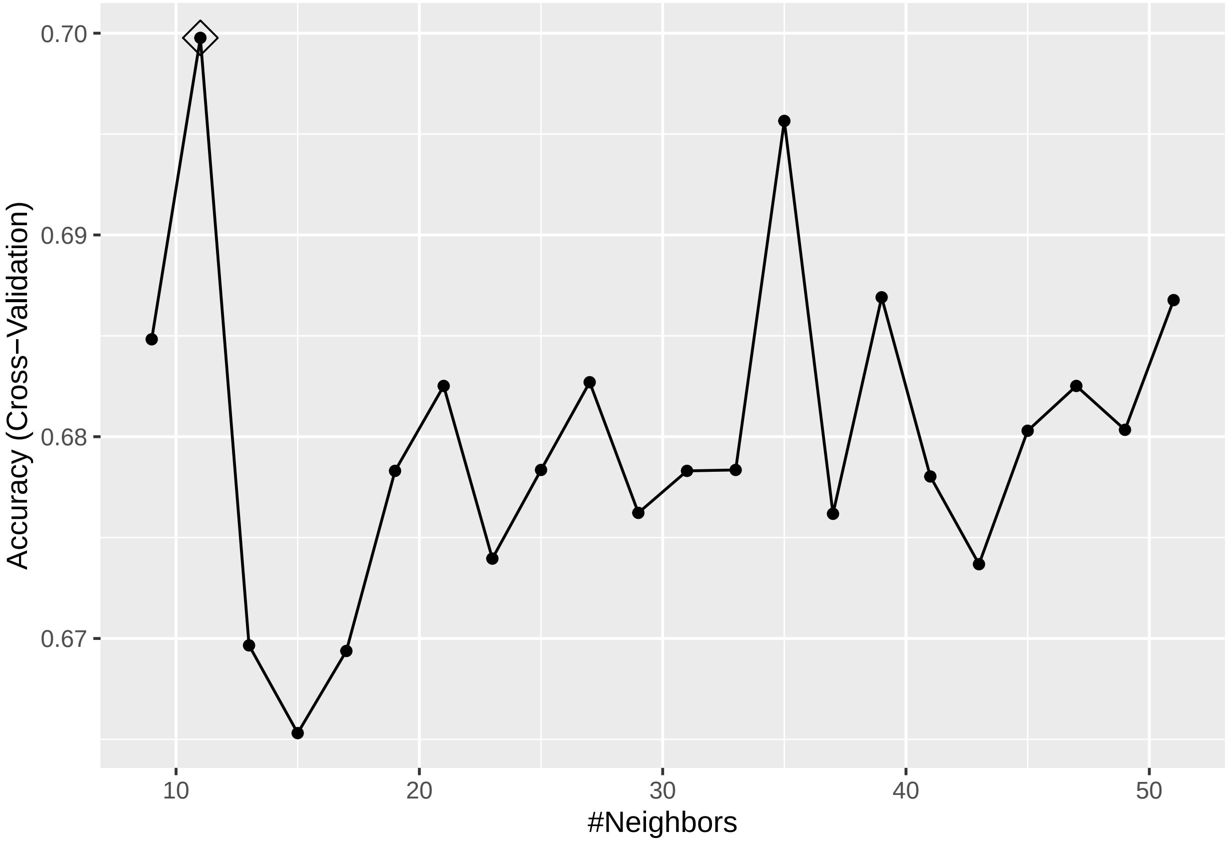


Table 13: Optimal K-nearest neighbours

	k
2	11

Table 14: KNN with 7 predictors

	x
Accuracy	0.6982759

It is observed that the accuracy still remains the same.

Model 7: Classification (Decision) Trees Model

We will use a different algorithm. Since the outcome is categorical we will use the classification (decision) trees model.

```
#Model 7- Classification (Decision) Trees Model
train_rpart <- train(Dataset ~ .,
                     method = "rpart",
                     tuneGrid = data.frame(cp = seq(0.0, 0.1, len = 25)),
                     data = train_set)
```

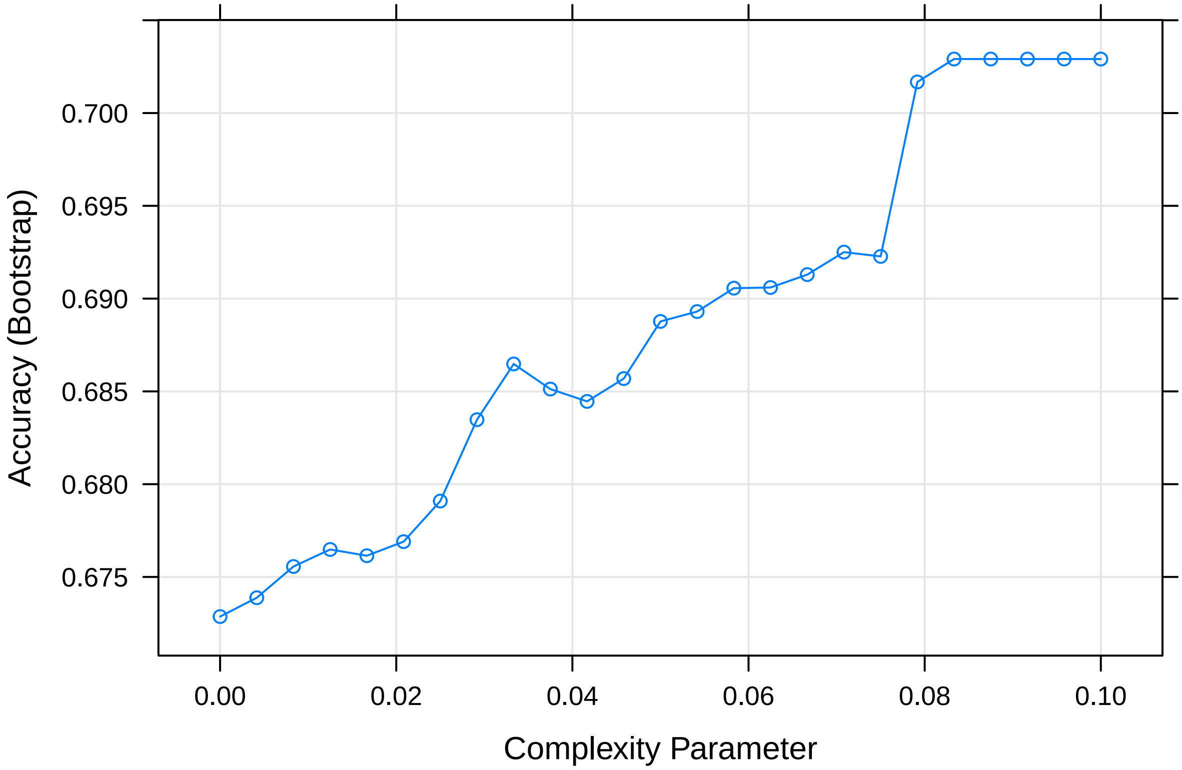


Table 15: Classification Decision Trees Model with all predictors

	x
Accuracy	0.7155172

The overall accuracy great improves. We will try improve this slightly more.

Model 8: Classification (Decision) Trees Model 2 (with significant variables)

In this model only the variables that had significant p-values in the Wilcoxon Signed Rank test will be used.

```
#Model 8- Classification (Decision) Trees Model with significant variables
train_rpart <- train(Dataset ~ Age+ Total_Bilirubin + Direct_Bilirubin + Alkaline_Phosphotase + Alamine_
                     method = "rpart",
                     tuneGrid = data.frame(cp = seq(0.0, 0.1, len = 25)),
                     data = train_set)
```

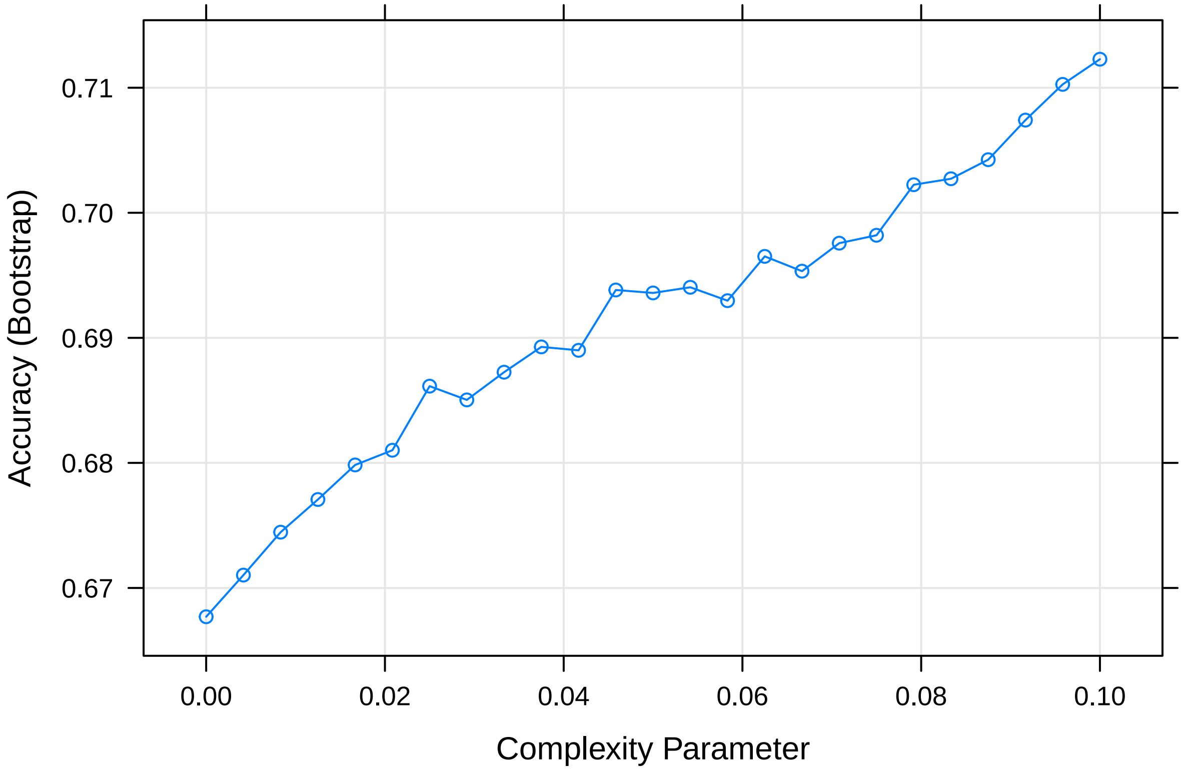


Table 16: Classification Decision Trees Model with 8 predictors

	x
Accuracy	0.7155172

Results

Table 17: Summary of model accurary

model	Accuracy
Logistic Regression with all predictors	0.2500000
Logistic Regression with 9 predictors	0.2413793
Logistic Regression with 5 predictors	0.2844828
KNN with 9 predictors	0.7068966
KNN with 8 predictors	0.6810345
KNN with 7 predictors	0.6982759
Classification Decision Trees Model with all predictors	0.7155172
Classification Decision Trees Model with 8 predictors	0.7155172

From the results we can see that the classification trees model performed the best. The predictors used is either all or 8 significant variables.

Conclusion

In this project various models have been tested to build a prediction algorithm for doctors to diagnose liver disease. The overall accuracy results show generally all the variables can be used as predictors. However it is suggested that the 8 significant ones are used for the algorithm. Overall, the Logistic Regression model did not perform well and the Classification Trees model performed the best. A limitation of the dataset is that it comes from a very niche group of people and the observations are very few. In future a larger dataset from people accross different regions can be obtained to improve the algorithm. Some kind of clustering categorization also could be added to see if there if any of the variables tend to cluster in a certain group of patients. Additional history of the patient diet and lifestyle could also be added to the clinical variables as it might enhance the prediction.