Prediction of Liver Disease From Blood Markers

Joe King

30 December, 2020

Contents

1 Introduction and Background

The purpose of this report is to examine the Indian Liver Patient data available at https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset) to determine a model for predicting the presence of liver disease from patients' blood enzyme and protein levels.

For non-clinicians it is useful to include a little background about how levels of proteins and enzymes in the blood are used by clinicians as indicators of liver malfunction.

Elevated levels of the enzymes alamine aminotransferase (ALT, also known as alanine transaminase), aspartate aminotransferase (AST, also known as aspartate transaminase) and alkaline phosphotase (APT) are considerered markers for liver disease.

So too are depleted levels of total proteins and albumin, and, on the other hand, elevated levels of billirubin.

Also used by clinicians to distinguish between types of liver disease are the AST/ALT ratio with high values indicating possible alcohol-related disease.

The Albumin/Globulin ratio (AGR) is a measure which in conjunction with the absolute levels of proteins in the blood can indicate various conditions including liver disease, if the ratio is significantly different from the normal level of just over 1.

Sources:

Mayo Clinic https://www.mayoclinic.org/tests-procedures/liver-function-tests/about/pac-20394595)
Wikipedia https://en.wikipedia.org/wiki/AST/ALT_ratio

2 Data Analysis

2.1 Data Breakdown and Cleaning

Table 1: Summary of Indian Liver Patient Data

No of Patients	Total Males	Total Females	Males With Liver Disease	Females With Liver Disease
583	441	142	324	92

The above table shows the breakdown of the patients included in the Indian Liver Patient data. For each patient the following data items are included:

Age
Gender
Total Billirubin (TBIL)
Direct Billirubin (DBI)
Alkaline Phosphotase (APT)
Alamine Aminotransferase (ALT)
Aspartate Aminotransferase (AST)
Total Proteins (TP)
Albumin (ALB)
Albumin/Glolbulin Ratio (AGR)
Selector For Liver Disease

Inspection of the data revealed only four patients with missing data: for these, no Albumin/Globulin Ratio (AGR) was present. Instead of ignoring these patients, the Albumin/Globulin Ratio was estimated using the formula:

$$\mathbf{AGR} = \frac{\mathbf{ALB}}{\mathbf{TP} - \mathbf{ALB}}$$

This is not as accurate as for the other patient records as TP and ALB data are already rounded to 2 significant figures, but it is the best estimate.

The AST/ALT Ratio is not included with the patient records, but for the purposes of this study it was calculated to see if it can help in predictions.

2.2 Division into Training and Test Datasets

The original dataset was divided randomly into training and test partitions. 90% of the original dataset was allocated to training data used for developing the predictive model and the remaining 10% to test data for testing the developed model. Table 2 shows the breakdown.

Table 2: Summary of Indian Liver Patient Data

Dataset	No of Patients	Total Males	Total Females	Males With Liver Disease	Females With Liver Disease
Training	524	396	128	292	82
Testing	59	45	14	32	10

2.3 Data Visualisation

The data is categorised by age, gender and whether or not a patient has been diagnosed as having liver disease. For this reason, the approach used in visualising data was to use age as the x-axis and the different blood markers as the y-axis.

2.3.1 Normalising the Data

Firstly, the data was normalised by calculating the means and standard deviations for healthy patients (i.e. those without liver disease) for each of the blood markers treating males and females separately.

Table 3: Healthy Base Levels of Blood Enzymes

Gender	Mean ALT	Std Dev ALT	Mean AST	Std Dev AST	Mean APT	Std Dev APT
Female	30.50000	24.68535	32.71739	19.92169	205.2609	83.12252
Male	36.11538	26.30416	46.11538	42.19725	230.5288	164.02971

Table 4: Healthy Base Levels of AST/ALT Ratio

Gender	Mean	Std Dev
Female	1.271087	0.8556225
Male	1.380673	0.9981957

Table 5: Healthy Base Levels of Blood Proteins

Gender	Mean TP	Std Dev TP	Mean ALB	Std Dev ALB	Mean TBIL	Std Dev TBIL	Mean DBI	Std Dev DBI
Female	6.532609	1.124881	3.317391	0.8127605	0.9130435	0.4674321	0.2717391	0.2500338
Male	6.547115	1.056736	3.331731	0.7821137	1.2721154	1.2022194	0.4721154	0.6204378

Table 6: Healthy Base Levels of Albumin/Globulin Ratio

Gender	Mean	Std Dev
Female	0.9995652	0.2712437
Male	1.0286538	0.2890159

Using these "base" levels, we can calculate z-values for each patient record and blood marker. For example, the z-value of ALT for patient i of gender g is given by the formula:

$$\textbf{zALT}_i = (\textbf{ALT}_i - \widehat{\textbf{ALT}}_g)/\textbf{sd(ALT)}_g$$

where $\widehat{\textbf{ALT}}_g$ is 30.5 for females or 36.1153846 for males, and $\textbf{sd(ALT)}_g$ is 24.6853533 for females or 26.304161n for males.

Similar formulae for the other blood markers define the z-values for patient i, namely $\textbf{zAST}_i$, $\textbf{zAPT}_i$, $\textbf{zAST:ALT}_i$, $\textbf{zTP}_i$, $\textbf{zALB}_i$, $\textbf{zAGR}_i$, $\textbf{zTBIL}_i$ and $\textbf{zDBI}_i$.

2.3.2 Age Distribution of Patients

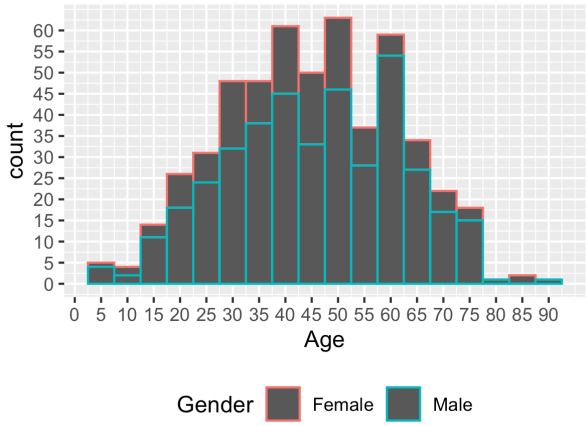

Figure 1: Age Distribution of Indian Liver Patients

As can be seen in Figure 1, there is limited data for those aged less than 18 and greater than 77, so in order not to skew the model, we will exclude these patients while developing the model.

The distribution also clearly shows that the data available for women are fewer than for men, so it may be that any model developed will not be as good at predicting the presence of liver disease for women as for men.

2.3.3 Blood Markers v Age

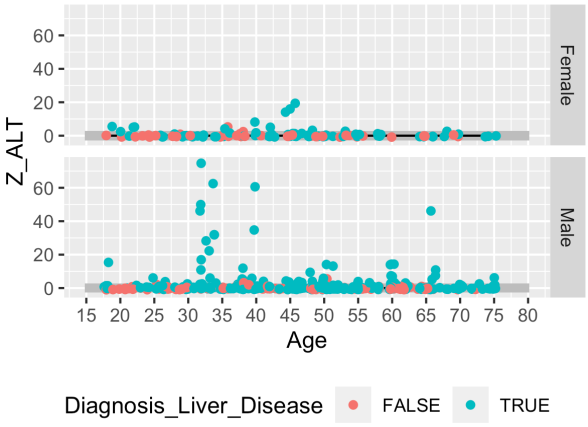

Figure 2: Alamine Aminotransferase (ALT) By Age

Figure 2 shows the training set patient data for $\mathbf{zALT}_i$. The grey region represents the 95% confidence interval for healthy patients with the black line representing the mean level, which is by definition zero, since these are standardised z-values. The dots represent individual patients, with the colour determining which are healthy and which are diagnosed with liver disease.

It is clear that ALT for almost all the healthy patients are clustered around the "normal" with only a few outliers. For the diagnosed liver disease patients, while many are within normal range, many, particularly males, are clearly way above the normal range.

So, as clinicians suggest, this is clearly a good marker for predicting liver disease, but it does not catch all liver disease.

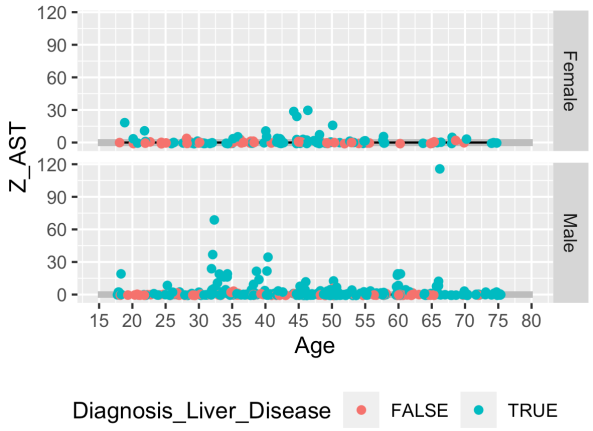

Figure 3: Aspartate Aminotransferase (ALT) By Age

Figure 3 shows a similar plot for **zAST**$_i$.

The AST plot shows similar characteristics as for the plot for ALT. So again, this is a good marker for prediction, but not necessarily conclusive.

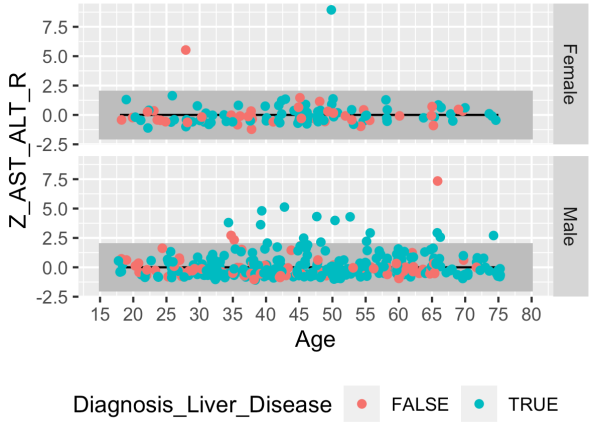

Figure 4: AST/ALT Ratio By Age

Figure 4 shows the plot for **zAST:ALT**$_i$, which as was mentioned in the introduction is sometimes used by clinicians to determine alcohol-related liver disease. The plot shows however, that without further data such as alcohol consumption data for patients, this is unlikely to be a useful marker in developing our predictive model, as the vast majority of patients fall within "normal" limits. An interesting observation is however that the plot may show that rather more men than women may be showing signs of heavy alcohol consumption, with more patients outside normal range and towards the upper end of normal limits.

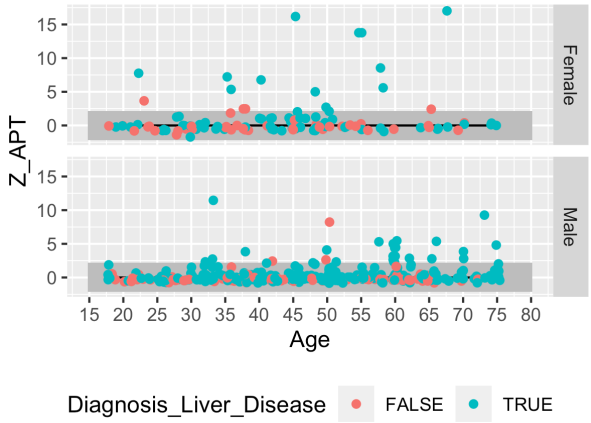

Figure 5: Alkaline Phosphotase By Age

Figure 5 shows the picture for **zAPT**$_i$. It can be seen that patients with elevated levels are present throughout the age range for females amongst those diagnosed with liver disease. For males, the picture is less clear but again for the higher ages APT may be a useful marker in predicting liver disease.

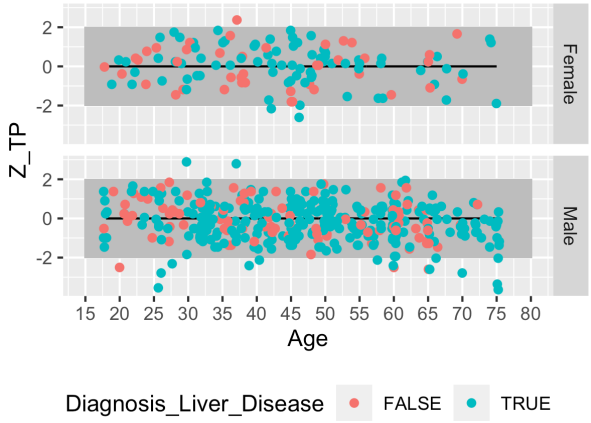

Figure 6: Total Proteins By Age

Figure 6 shows the patterns for **zTP**$_i$. Very few patients fall outside the 95% confidence interval, so, in itself, this marker is unlikely to prove useful as a predictor.

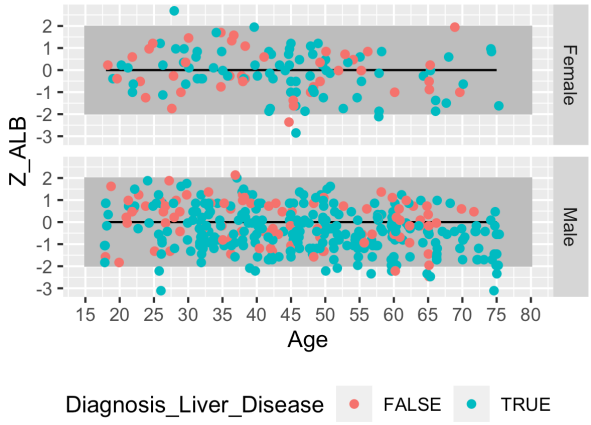

Figure 7: Albumin By Age

Figure 7 shows the patterns for **zALB**$_i$. Here, again most patients fall within the 95% confidence interval. So again, ALB in itself is unlikely to be a useful predictor of liver disease. However, it does seem that at least for males, the majority of diagnosed liver disease patients are in the lower ranges of normal with several falling out of range.

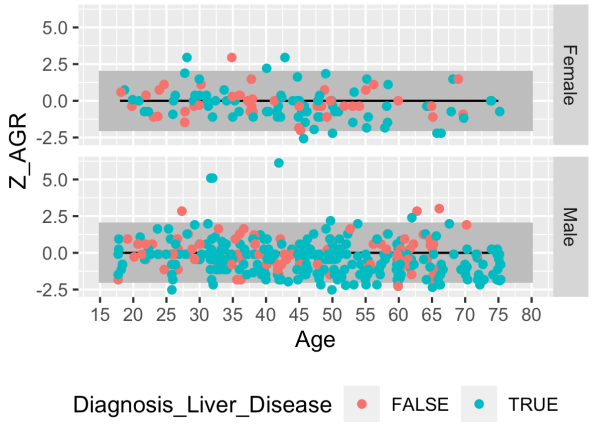

Figure 8: Albumin/Globulin Ratio By Age

Figure 8 shows the plot for **zAGR**$_i$. This shows a similar pattern to the Albumin plot, which is no surprise given that inherent link between the data.

Figure 9: Total Billirubin By Age

Figure 9 shows the plot for **zTBIL**$_i$. Here the picture is much clearer: there are very few outliers among the healthy patients, and very many diagnosed liver disease patients who are way above the 95% confidence interval for "normal". So TBIL is likely to be a significant predictor of liver disease.

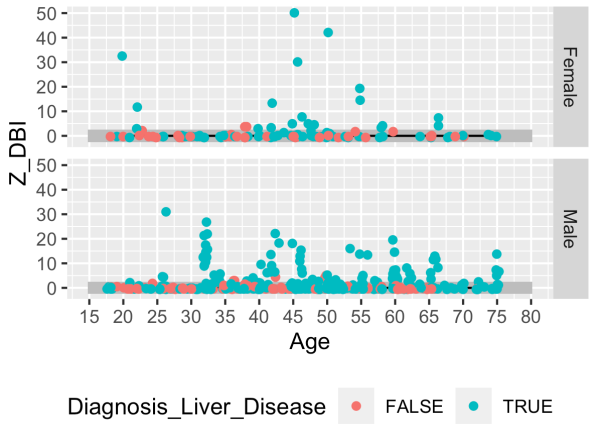

Figure 10: Direct Billirubin By Age

Figure 10 shows the plot for **zDBI**$_i$. The picture here is very similiar to the picture for TBIL and the same observations are relevant. It is probable that the use of either TBIL or DBI in predictions of liver disease will yield a similar outcome.

2.4 Developing the Predictive Model And Results

2.4.1 Strategy

The visualisation has shown that some of the blood markers are likely to provide more influence on predicting the presence of liver disease than others. In particular, the enzyme markers (AST, ALT, APT) seem more reliable than the protein markers (TP,ALB, AGR). The billirubin markers (TBIL and DBI) do seem to be reliable - intuitively this seems very plausible as they are responsible for the yellow colour associated with jaundice - a clear sign of problems with the liver.

The strategy will be to develop two classification models, the first using "K-Nearest Neighbours" (KNN) and the second using "Random Forest" (RF).

Within each model, a second subdivision will firstly be to use all z-value markers in the dataset (**Age**$_i$, **Gender**$_i$, **zALT**$_i$, **zAST**$_i$, **zAPT**$_i$, **zTP**$_i$, **zALB**$_i$, **zAGR**$_i$, **zTBIL**$_i$ and **zDBI**$_i$), and secondly to exclude the protein markers (i.e. all except **zTP**$_i$, **zALB**$_i$, **zAGR**$_i$).

2.4.2 K-Nearest Neighbours (KNN)

The k-nearest neighbours model seeks to predict the classification for a particual row of data in the dataset, in our case representing a patient, by identifying the k rows of data in the dataset that are "nearest" in characteristics to that row. The "distance" between the markers is used to identify how near they are, where the k rows represent the neighbourhood of the particular row. The predicted classification is then the predominant prediction from the neighbourhood.

2.4.3 KNN All Markers Model

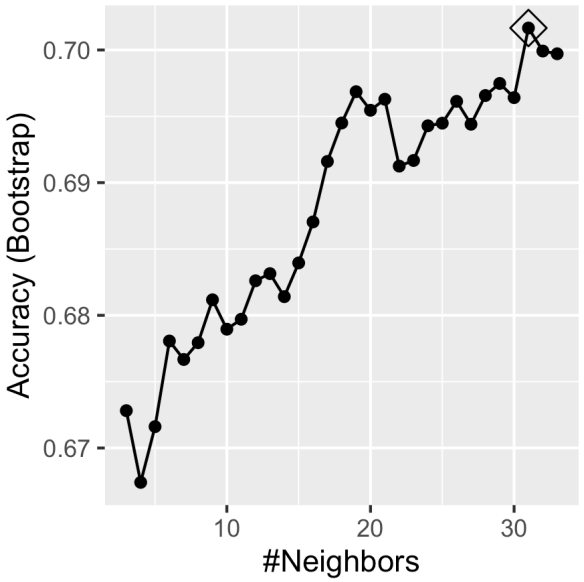

Figure 11: KNN All Markers - Best k

In this model, a range of values of k were considered between 3 and 33. As can be seen from the plot in figure 11, the optimal k was 31.

When applied to the test set data, this yielded an accuracy (% of correct predictions) of 67.7966%. The predictions are summarised in the following table:

```
##            Reference
## Prediction FALSE TRUE
##      FALSE     4    6
##      TRUE     13   36
```

This tells us that there are 13 false positives and 6 false negatives, with 40 matching predictions.

2.4.4 KNN Non-Protein Markers Model

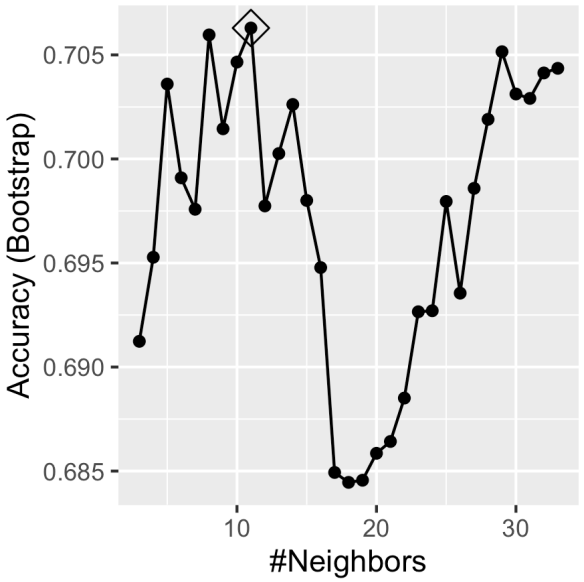

Figure 12: KNN Non-Protein Markers - Best k

In this model, a range of values of k were considered between 3 and 33. As can be seen from the plot in figure 12, the optimal k was 11.

When applied to the test set data, this yielded an accuracy (% of correct predictions) of 69.4915%. The predictions are summarised in the following table:

```
##            Reference
## Prediction FALSE TRUE
##      FALSE     8    9
##      TRUE      9   33
```

This tells us that there are 9 false positives and 9 false negatives, with 41 matching predictions.

2.4.5 Random Forest (RF)

The random forest model is essentially a development of a decision tree model.

In a decision tree, a prediction is made by looking at different data markers in turn. A binary choice is made at each node of the tree and eventually a prediction is made. The optimum order of markers is difficult to determine particularly where there are very many markers.

In the random forest method, the predictions from multiple randomly chosen decision trees are examined to give the "best" prediction. For classification models such as this one, the majority prediction amongst the decision trees is chosen.

Randomness is generated firstly by choosing multiple samples of N rows of the training dataset to create the trees, and randomly choosing the "features" (in our case the blood markers) included in each tree.

From the forest of decision trees the majority prediction is chosen for each row (or patient).

2.4.6 RF All Markers Model

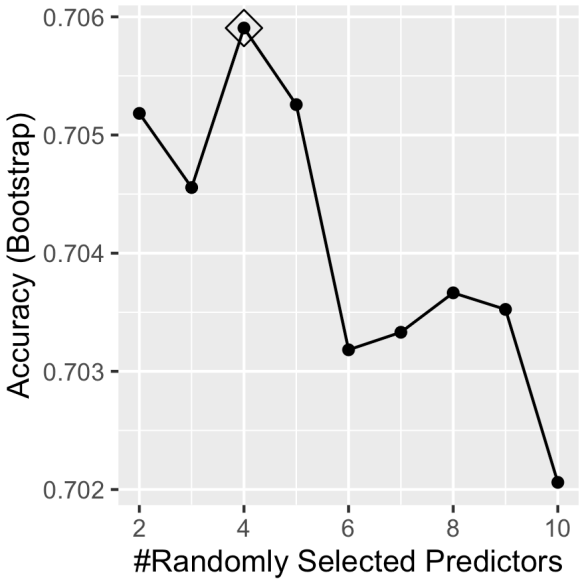

Figure 13: RF All Markers - Optimum No of Randomly Selected Predictors

In this model, we optimised for the best number of markers to include in each tree in the range 2 to 7. As can be seen from the plot in figure 11, the optimal number of markers is 4. We also set the node size to 14.

When applied to the test set data, this yielded an accuracy (% of correct predictions) of 69.4915%. The predictions are summarised in the following table:

```
##            Reference
## Prediction FALSE TRUE
##      FALSE     4    5
##      TRUE     13   37
```

This tells us that there are 13 false positives and 5 false negatives, with 41 matching predictions.

2.4.7 RF No Protein Markers Model

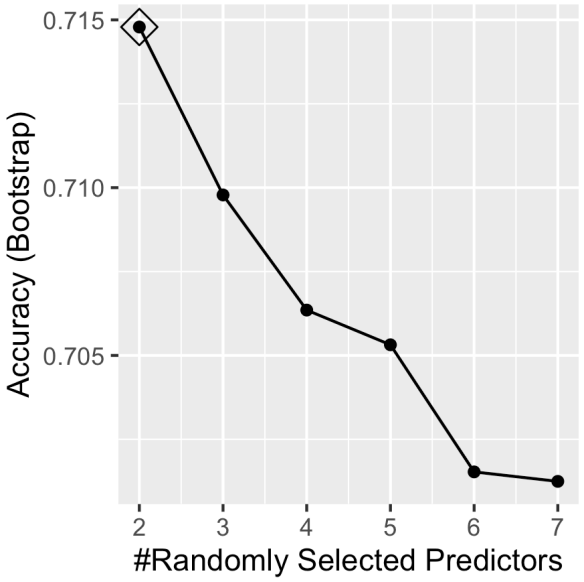

Figure 14: RF Non-Protein Markers - Optimum No of Randomly Selected Predictors

In this model, we optimised for the best number of markers to include in each tree in the range 2 to 7. As can be seen from the plot in figure 11, the optimal number of markers is 2. We also set the node size to 14.

When applied to the test set data, this yielded an accuracy (% of correct predictions) of 76.2712%. The predictions are summarised in the following table:

```
##            Reference
## Prediction FALSE TRUE
##      FALSE     8    5
##      TRUE      9   37
```

This tells us that there are 9 false positives and 5 false negatives, with 45 matching predictions.

3 Conclusion

It can be seen from the analysis and results in the last section that the most accurate model of the four examined is the Random Forest model where only the Non-Protein markers are considered.

It not only provides the highest accuracy for the test dataset at 76.2712%, but also has the smallest number of both false positives (9) and false negatives (5).

There is certainly scope for improving this model. Indeed as it would be used in clinical environments, it would be useful to have some parameters set as to what the criteria are for a "good" model. Is overall accuracy the metric to use, or would minimising false negatives be a better metric? Should the predictions include a "Not sure" diagnosis to cater for those patients where the decision is borderline?

It is certainly interesting that the blood protein markers, that is Albumin, Total Proteins and the Albumin/Globulin Ratio, does not seem to be useful in improving the acccuracy of the predictions.

In the introduction it was mentioned that the AST/ALT Ratio is used as an indicator of alcohol-related liver disease. Perhaps a future study should also gather data about lifestyle markers such as alcohol consumption to further to see if that would improve predictions.

The Indian Liver Patients dataset is relatively small, and as such the accuracy of a predictive model is unlikely to be sufficient in a clinical setting to correctly diagnose liver disease. As a tool to suggest further patient investigations it may be useful, but certainly should not be regarded as foolproof.