Final Project: Indian Liver Patient Data Analysis

Tiffany Chen
STAT 448
May 1, 2024

1. Introduction

The dataset that will be analyzed in this report is the Indian Liver Patient Dataset from UC Irvine's Machine Learning Repository. It contains records for 584 patients from India, with variables for whether a patient was a liver patient, patient age, patient gender, and measurements of the following biochemical features: total Bilirubin, direct Bilirubin, Alkaline Phosphatase, Alamine Aminotransferase, Aspartate Aminotransferase, total proteins, Albumin, and Albumin Globulin ratio.

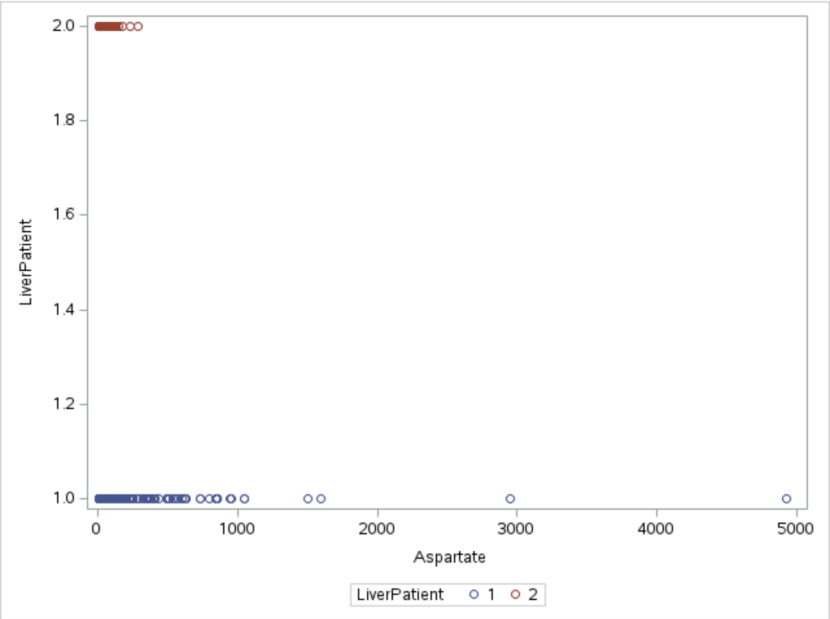
This analysis and report intends to provide insight into the differences in features between liver patients and non-liver patients, how liver patient status can be predicted using clinical measurements and age, how total protein levels can be predicted using clinical measurements, age, and gender, the similarities and differences present in clinical measurements between liver patients and non-liver patients, and how patients may be classified by liver patient status and gender based on clinical measurements and age.

2. Descriptive Analysis

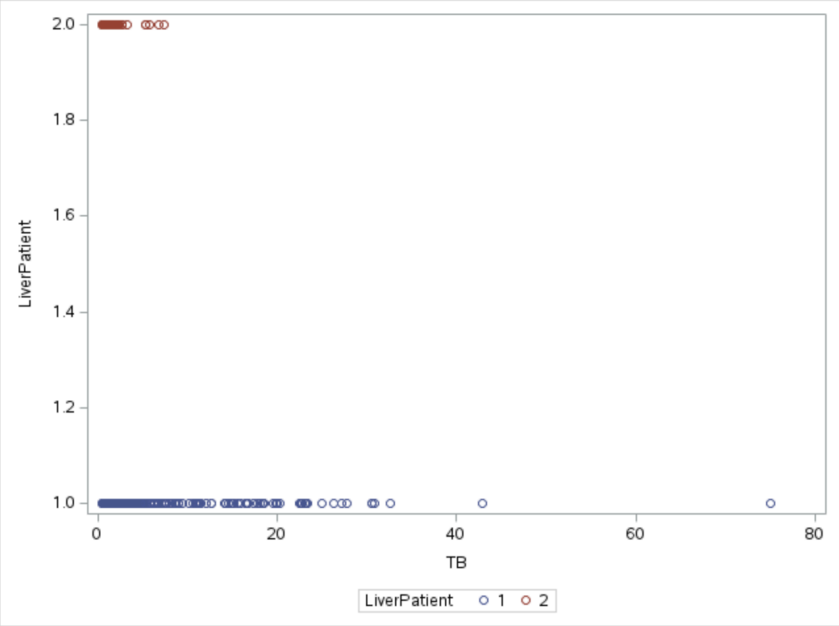
When comparing clinical measurements and ages of liver patients and non-liver patients, the following significant results were found: mean ages for liver patients (46.15) were 4.91 years higher than mean ages for non-liver patients (41.24). The median value for total Bilirubin in liver patients was 1.4, and in non-liver patients was 0.8; The median value for direct Bilirubin in liver patients was 0.5, and in non-liver patients was 0.2; The median value for Alkaline Phosphatase measurements in liver patients was 229, and in non-liver patients was 186; The median value for Alamine in liver patients was 41, and in non-liver patients was 27; The median value for Aspartate in liver patients was 52.5, and in non-liver patients was 29. The mean value for Albumin in liver patients was 3.06, and in non-liver patients was 3.34. The mean value for Albumin Globulin ratio in liver patients was 0.9, and in non-liver patients was 1.03. Overall, liver patients showed significantly higher levels of total Bilirubin by 0.6, direct Bilirubin by 0.3, Alkaline Phosphatase by 43, Alamine by 14, and Aspartate by 23.5. Liver patients showed lower levels of Albumin by 0.28, and Albumin Globulin ratio by 0.13.

Significance of these values was determined by a Wilcoxon Rank Sum Test due to the non-normal nature of each variable's distribution. Total protein levels were not found to be significantly different when comparing liver patients and non-liver patients.

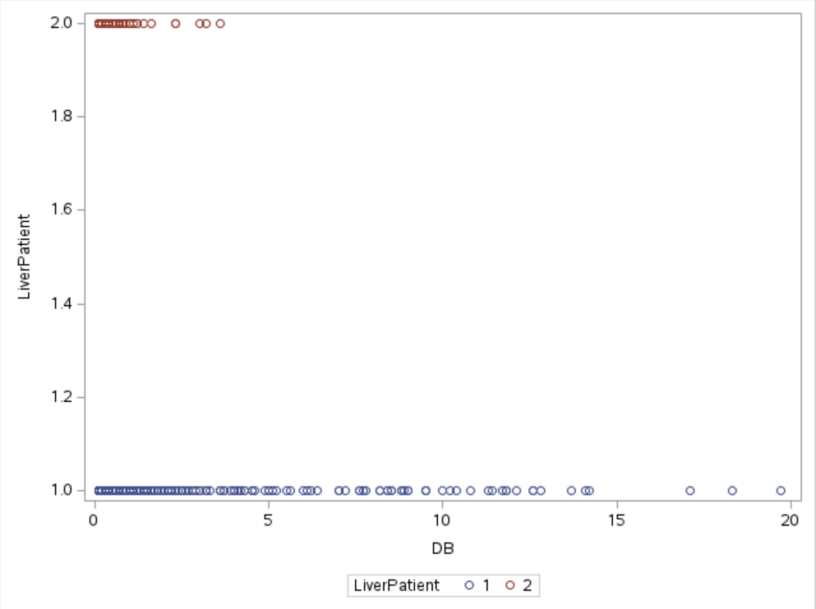
By performing a correlation test, I found that there is a correlation between liver patient status with age and nearly all clinical measurements in the dataset. The strongest correlations are visualized as follows.



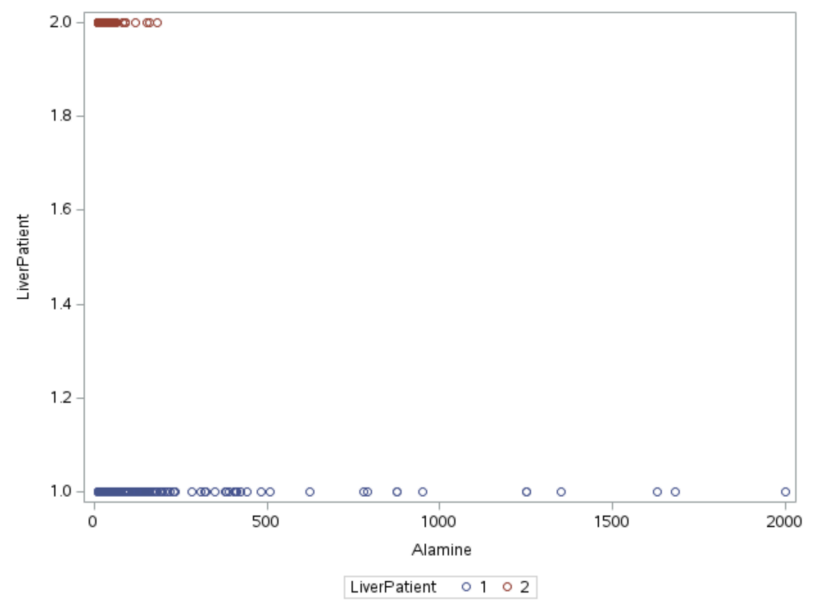
A one unit decrease in Aspartate correlates to a 0.309 times increased chance of the individual being a liver patient. There are significantly more liver patients with higher levels of Aspartate. Liver patients are indicated in blue and correspond to a value of 1 on the graph.



A one unit decrease in total Bilirubin correlates to a 0.304 times increased chance of the individual being a liver patient. There are significantly more liver patients with higher levels of total Bilirubin. Liver patients are indicated in blue and correspond to a value of 1 on the graph.



A one unit decrease in direct Bilirubin correlates to a 0.297 times increased chance of the individual being a liver patient. There are significantly more liver patients with higher levels of direct Bilirubin. Liver patients are indicated in blue and correspond to a value of 1 on the graph.

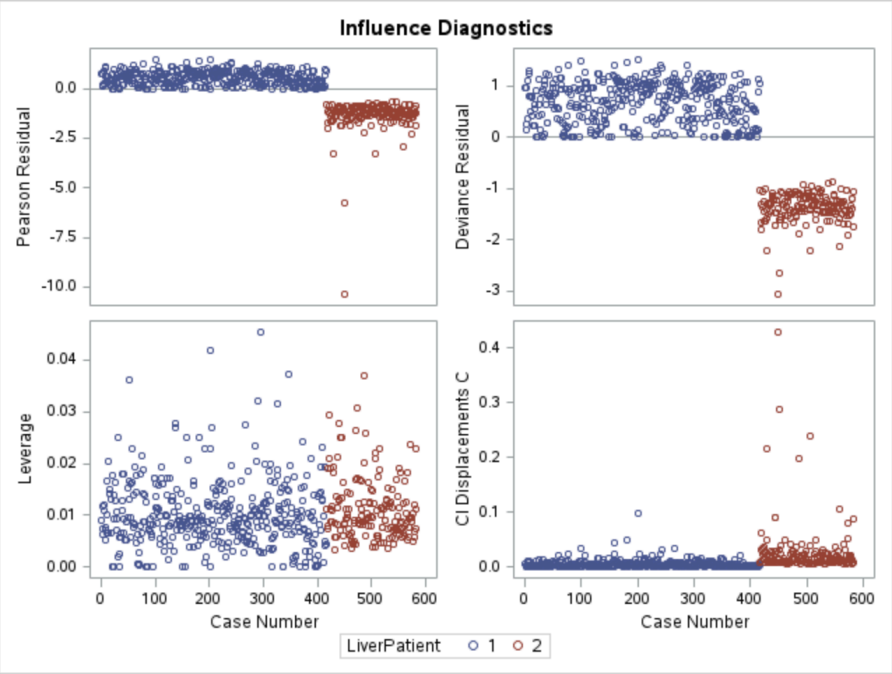


A one unit decrease in Alamine correlates to a 0.291 times increased chance of the individual being a liver patient. There are significantly more liver patients with higher levels of Alamine. Liver patients are indicated in blue and correspond to a value of 1 on the graph.

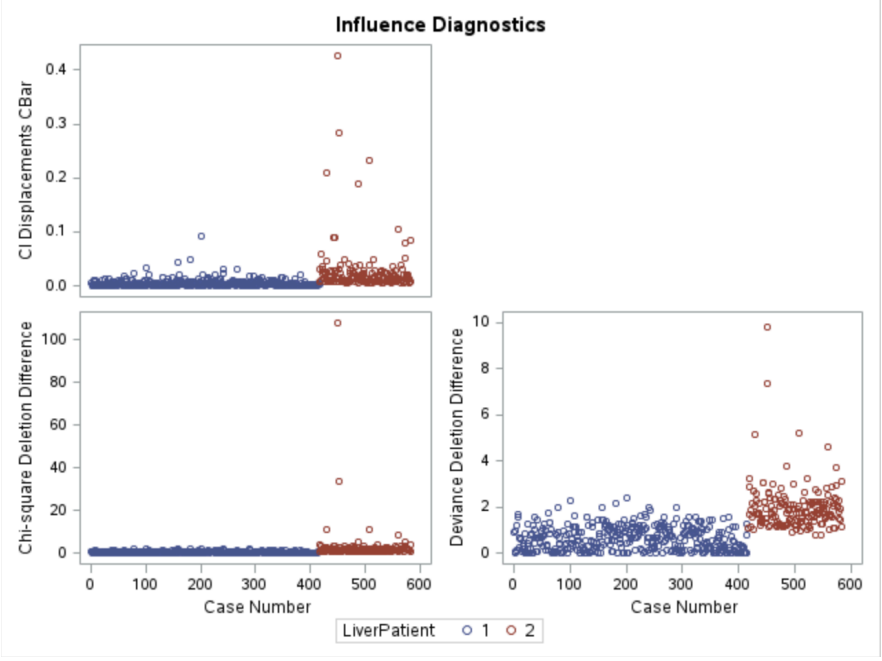
3. Modeling Liver Patient Status with Clinical Measurements and Age

To determine a model that will predict whether a patient has liver disease using the patient's age and clinical measurements provided, I utilized three different selection methods for logistic regression (accounting for the binary response of the LiverPatient variable). The best

model was selected using backward selection and takes the age, direct Bilirubin, Alamine, total protein, and Albumin predictors into account. This model has the lowest AIC value of those returned by forward, backward, and stepwise selection—590.53, 586.52, and 589.96 respectively. Diagnostics for the model are shown in the plots below. Liver patients are indicated in blue with the value 1 and non-liver patients are indicated in red with the value 2:



Overall, liver patients tend to be overestimated by the model, while non-liver patients tend to be underestimated by the model as seen in the Pearson residuals and deviance residuals. Residuals for non-liver patients tend to be larger, which may be due to the dataset containing a greater proportion of liver patients than non-liver patients.



The Cbar plot indicates some slightly influential observations within the non-liver patients, but there are no concerning highly influential points that need to be removed.

Below is the frequency table of the model's predicted responses against the true observation values. Out of 416 liver patients, the model accurately predicted 384 and out of 167 non-liver patients, the model accurately predicted 42. There is room for improvement as the model predicts non-liver patients accurately at a rate less than chance.

The FREQ Procedure

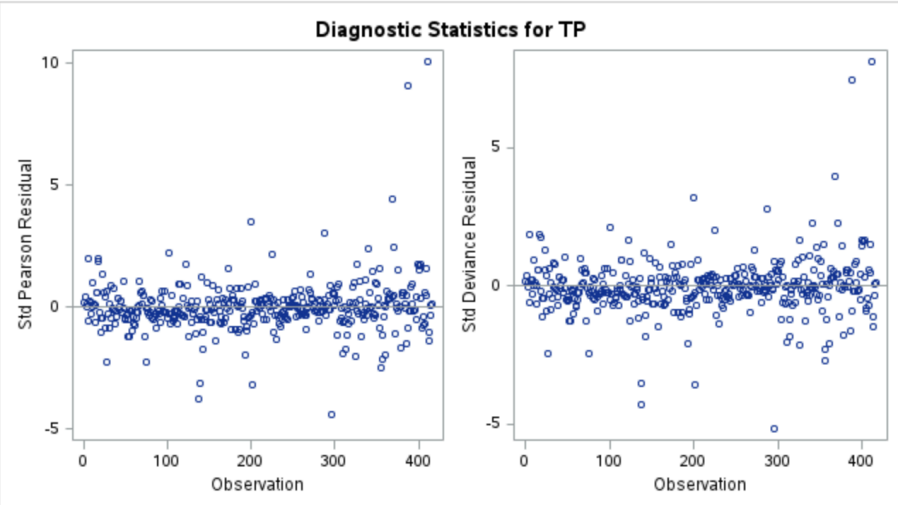
Frequency	Table of LiverPatient by _INTO_		
	INTO(Formatted Value of the Predicted Response)		
LiverPatient	1	2	Total
1	384	32	416
2	125	42	167
Total	509	74	583

Based on this model, the odds ratios indicate that a one unit increase in the patient's age increases their odds of being a liver patient by about 1.018 times. Similarly, a one unit increase in the patient's direct Bilirubin levels increases their odds of being a liver patient by about 1.739 times. A one unit increase in the patient's Alamine levels increases their odds of being a liver patient by about 1.016 times. A one unit increase in the patient's total protein levels increases their odds of being a liver patient by about 1.542 times. A one unit increase in the patient's Albumin levels decreases their odds of being a liver patient by about 0.514 times. Thus, patients who are older and have higher levels of direct Bilirubin, Alamine, and total protein, and lower levels of Albumin are more likely to be liver disease patients.

4. Modeling Total Protein Values with Clinical Measurements, Gender, and Age

To create a model predicting the protein values of liver patients using the clinical measurements, gender, and age predictors, I selected variables in a gamma model with a log-link. Using Type 3 likelihood-ratio analysis statistics, I removed the variables that were shown to be insignificant in the following order: Alkaline Phosphatase, gender, age, total Bilirubin. The final variables selected for the model as a function of total protein levels were: direct Bilirubin, Alamine, Aspartate, Albumin, and Albumin Globulin ratio.

Below are the Pearson and deviance residual diagnostics for the model. Overall, there are no major underlying issues and the residuals appear to be evenly distributed for most observations.



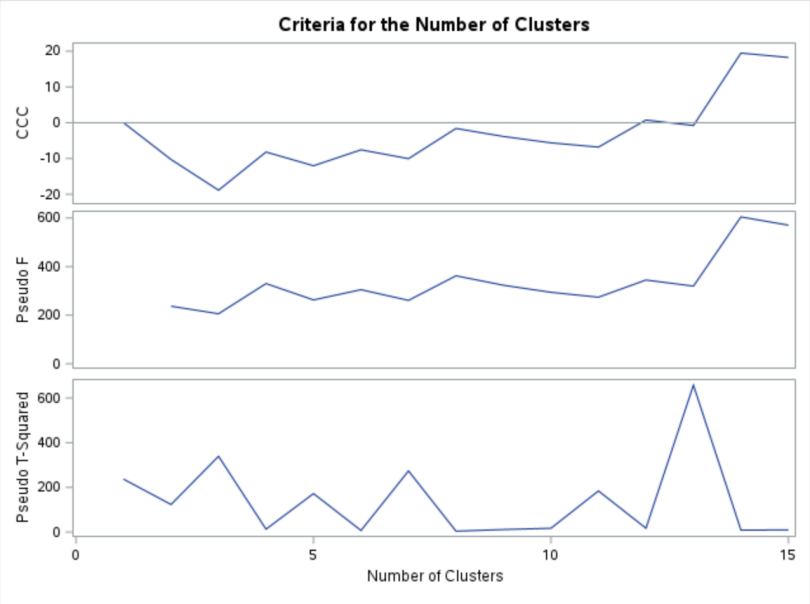
With this model, we can determine the multiplicative expected changes in total protein levels based on the significant predictor parameter estimates. A one unit increase in total protein levels for a liver patient corresponds with an expected exp(0.0121) more or 1.012% higher levels of direct Bilirubin. A one unit decrease in total protein levels for a liver patient corresponds with an expected exp(0.0001) more or 1.000% higher levels of Alamine. A one unit increase in total protein levels for a liver patient corresponds with an expected exp(0.0001) more or 1.000% higher levels of Aspartate. A one unit increase in total protein levels for a liver patient corresponds with an expected exp(0.2522) more or 1.287% higher levels of Albumin, and a one unit decrease in total protein levels for a liver patient corresponds with an expected exp(0.2536) more or 1.289% higher Albumin Globulin ratio.

Direct Bilirubin, Albumin, and Albumin Globulin ratio are the most significant predictors in the model. Thus, higher direct Bilirubin levels, higher Albumin levels, and a lower Albumin Globulin ratio are more indicative of higher estimated total protein levels for liver patients.

5. Comparing Clinical Measurements and Ages for Liver Patients and Non-Liver Patients

To group patient observations based on their age and clinical measurements, I first removed any outlier observations from the dataset.

When grouping the data, I determined that six clusters to be a relatively strong choice based on the pseudo t-squared, pseudo F-statistic, and CCC values shown below:



The pseudo t-squared value at the six cluster mark is at a low point right before a spike in the graph, indicating that it would be a plausible clustering choice. The pseudo F statistic value is also at a peak point along with the CCC.

The frequency table displays the distribution of clustering for liver patients and non-liver patients across all six groups. Most liver patients and non-liver patients are categorized into cluster 1, but some liver patients were clustered into the rest of the groups according to their clinical measurements and age, with the majority being in clusters 2 and 3.

The FREQ Procedure

Frequency

Table of CLUSTER by LiverPatient

CLUSTER	LiverPatient 1	2	Total
1	380	161	541
2	9	0	9
3	8	1	9
4	4	0	4
5	1	0	1
6	1	0	1
Total	403	162	565

Frequency Missing = 4

This may indicate that liver patients and non-liver patients generally share similar values for their measurements. with some slight differences uncovered by the cluster means.

CLUSTER=1

Variable	N	Mean	Std Dev	Minimum	Maximum
Age	541	44.9334566	16.2105135	4.0000000	90.0000000
TB	541	2.8391867	5.4598483	0.4000000	75.0000000
DB	541	1.2563771	2.3614146	0.1000000	14.2000000
Alkphos	541	267.0739372	165.2107467	63.0000000	1124.00
Alamine	541	53.4417745	59.3967734	10.0000000	482.0000000
Aspartate	541	70.9611830	88.4635538	10.0000000	630.0000000
TP	541	6.5070240	1.0561928	2.8000000	9.6000000
ALB	541	3.1754159	0.7792720	1.4000000	5.5000000
AGRatio	541	0.9605545	0.3164363	0.3500000	2.8000000

CLUSTER=2

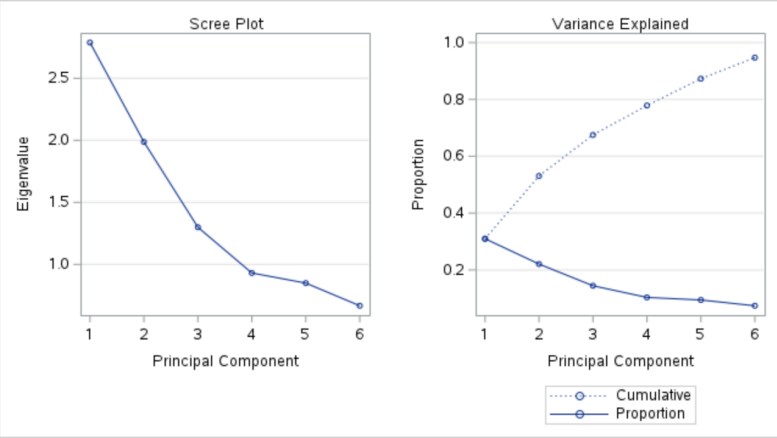
Variable	N	Mean	Std Dev	Minimum	Maximum
Age	9	41.2222222	15.1721602	18.0000000	60.0000000
TB	9	5.3111111	2.2340795	1.4000000	8.6000000
DB	9	2.5444444	1.0852547	0.6000000	4.0000000
Alkphos	9	245.5555556	49.1047633	186.0000000	308.0000000
Alamine	9	556.5555556	242.2788224	190.0000000	875.0000000
Aspartate	9	788.5555556	128.5127534	497.0000000	950.0000000
TP	9	6.0777778	1.3103223	4.0000000	7.4000000
ALB	9	2.7777778	0.6200358	1.7000000	3.6000000
AGRatio	9	0.8188889	0.2154324	0.6000000	1.1000000

CLUSTER=3

Variable	N	Mean	Std Dev	Minimum	Maximum
Age	9	53.8888889	11.9837853	33.0000000	73.0000000
TB	9	6.2888889	7.3961890	0.6000000	23.3000000
DB	9	3.1666667	4.0447497	0.1000000	12.8000000
Alkphos	9	1648.44	244.5133489	1350.00	2110.00
Alamine	9	110.3333333	119.7633082	48.0000000	425.0000000
Aspartate	9	142.8888889	142.2835588	57.0000000	511.0000000
TP	9	6.2222222	1.1177855	4.6000000	8.0000000
ALB	9	2.6666667	0.6946222	2.0000000	3.9000000
AGRatio	9	0.7055556	0.1666667	0.5000000	0.9500000

Comparing the mean values of the clinical measurements and ages across clusters 1, 2, and 3, some liver patients (those in clusters 2 and 3) tended to have higher total Bilirubin, direct Bilirubin, Alamine, and Aspartate levels, along with lower Albumin and Albumin Globulin ratios. Therefore, there may be slight overall differences between liver patients' Bilirubin, Alamine, Aspartate, Albumin, and Albumin Globulin and non-liver patients'.

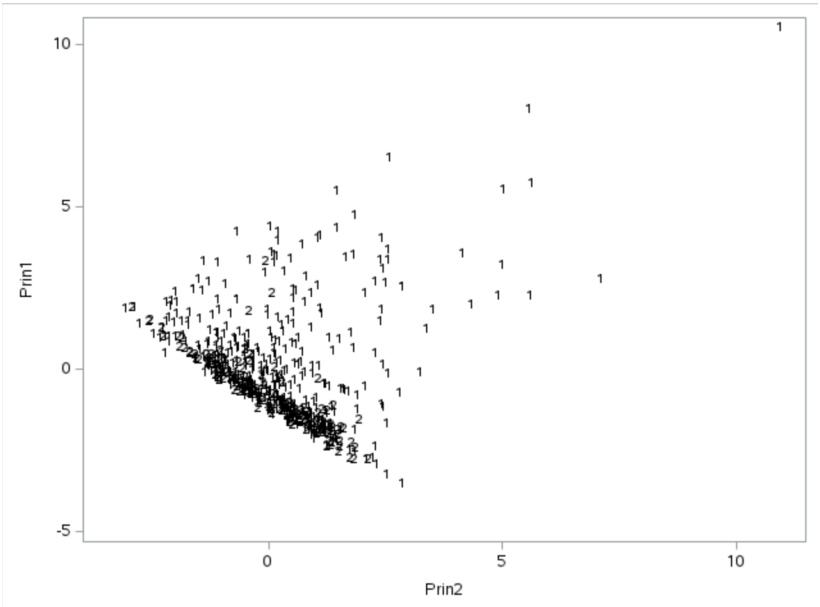
The most prominent features determining the clusters can be extracted using principal component analysis. Based on the scree plot below, three principal components is sufficient to accurately characterize the clusters and will describe approximately 67.5% of the variance in the clustering.



The first most prominent feature determining the clustering is a comparison between Albumin against total Bilirubin and direct Bilirubin measurements. Higher values of principal component 1 indicate higher Bilirubin measurements. The second feature is a comparison between age and Alamine, total proteins, and Albumin levels. Higher values of principal component 2 indicate higher Alamine, total protein, and Albumin measurements. The third feature is a comparison between Alamine and Aspartate against total Bilirubin and direct Bilirubin levels. Higher values of principal component 3 indicate higher Bilirubin levels.

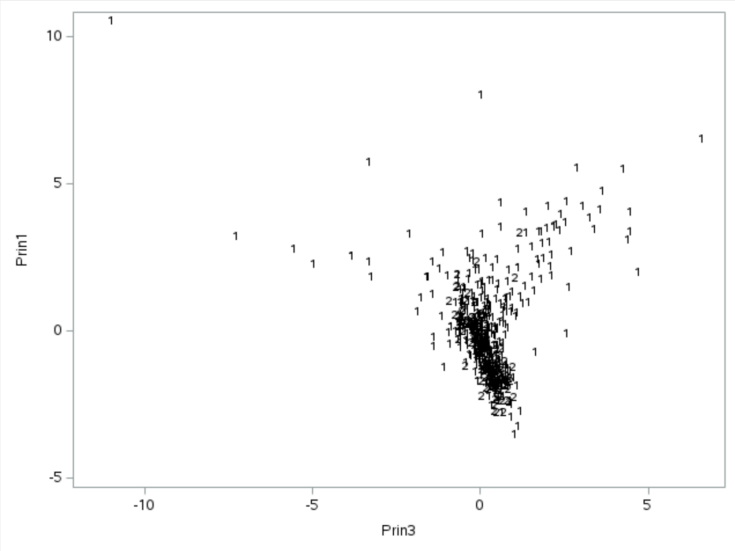
The principal components can provide insight into the similarities and differences in clinical measurements and ages of liver patients and non-liver patients.

Most liver and non-liver patients share similar values of principal components 1 and 2, but liver patients have a slight tendency to have higher levels of both components. This means that liver patients generally have slightly higher Bilirubin, Alamine, total protein, and Albumin measurements than non-liver patients.



Values of 1 in the scatterplot represent observations for liver patients and values of 2 represent observations for non-liver patients.

Again, most liver and non-liver patients share similar values of principal components 1 and 3, but liver patients have a slight tendency to have higher levels of component 3 as well, and therefore generally have slightly higher Bilirubin than non-liver patients.



Values of 1 in the scatterplot represent observations for liver patients and values of 2 represent observations for non-liver patients.

6. Classifying Liver Patient Status and Gender based on Clinical Measurements and Age

When classifying observations into liver patient status and gender groups based on their clinical measurements and ages, I first performed a Chi-square test to test for equal covariance. The F-statistic of the test was shown to be significant at the 0.1 level, which indicates unequal covariances. Quadratic discriminant analysis will be utilized to classify the liver patient and gender groups.

After doing discriminant analysis on all clinical measurements and age, I found that the model with all predictors was able to correctly classify approximately 20.41% of non-liver patient females, 43.97% of non-liver patient males, 20.88% of liver patient females, and 43.05% of liver patient males. The overall classification error rate for the full model is 67.93%, meaning that 67.93% of all observations are estimated to be misclassified into liver patient status and gender groups by the model.

Using stepwise selection to select for the most significant predictors, I then classified liver patient statuses and genders with the direct Bilirubin, Alkaline Phosphatase, age, and Alamine variables. The MANOVA test statistics for the reduced model are significant at the 0.05 level, indicating that the direct Bilirubin, Alkaline Phosphatase, age, and Alamine variables may be able to discriminate between liver patient statuses/genders.

The classification table indicates that about 80% of non liver patient females, 29.91% of non liver patient males, 10.87% of liver patient females, and 35.80% of liver patient males were correctly classified:

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.LIVERGEN
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into patient					
From patient	NF	NM	YF	YM	Total
NF	40 80.00	9 18.00	0 0.00	1 2.00	50 100.00
NM	73 62.39	35 29.91	2 1.71	7 5.98	117 100.00
YF	47 51.09	15 16.30	10 10.87	20 21.74	92 100.00
YM	99 30.56	94 29.01	15 4.63	116 35.80	324 100.00
Total	259 44.43	153 26.24	27 4.63	144 24.70	583 100.00
Priors	0.08576	0.20069	0.1578	0.55575	

The overall classification error rate of this model is 65.52%, meaning that 65.52% of all observations are estimated to be misclassified into liver patient status and gender groups by the model. The reduced model performs slightly better overall, as it has a lower overall error rate compared to the 67.93% error of the full model.

Liver patient males, non liver patient males, and liver patient females tended to be overclassified into the non liver patient female group. This may indicate that all four groups share similarities in age, direct Bilirubin, Alkaline Phosphatase, and Alamine measurements. Furthermore, male liver patients were nearly just as likely to be classified as non liver patient males as they were to be classified as non liver patient females. This may indicate similarities in age, DB, Alkaline Phosphatase, and Alamine measurements across all males in general, regardless of their liver patient status.

7. Conclusion

In general, older patients with higher levels of direct Bilirubin, Alamine, and total protein, and lower levels of Albumin are more likely to be liver disease patients. Higher direct Bilirubin levels, higher Albumin levels, and a lower Albumin Globulin ratio also signify higher estimated total protein levels for liver patients.

Overall, clinical measurements and demographics of patients can be utilized successfully to model predictions of liver disease in patients and uncover similarities and differences between liver patients and non-liver patients, as well as genders.