**Project Report**

**on**

**Analyzing Patterns and Predictors in Indian Liver Patient Record**

Submitted By:

SHIVANGI PATEL

Under The Guidance Of:

Prof. ALVIN CHIN

November, 2023

**CONTENTS**

[1. Introduction 1](#_Toc128488027)

2. Problem ………………………………………………………………………………………

3. Dataset Description …………………………………………………………………………

4. Procedure ……………………………………………………………………………………

5. Data Visualization …………………………………………………………………………..

5.1 Distribution of Target……………………………………………………………….

5.2 Distribution of Gender for Target………………………………………………….

5.3 Histogram……………………………………………………………………………

5.4 Plot Grid (Box plot)………………………………………………………………….

5.5 Density Plot…………………………………………………………………………

5.6 Scatterplots…………………………………………………………………………

5.6.1 Direct Bilrubin vs Total bilrubin………………………………………..

5.6.2 AlamineAT vs AspartateAT…………………………………………….

5.6.3 AlamineAT vs Alkaline phosphate…………………………………….

5.6.4 Albumin vs Total Proteins………………………………………………

5.6.5 Albumin Globulin ratio vs Albumin……………………………………

5.6.6 Albumin Globulin ratio vs Total Proteins…………………………….

5.7 Correlation Matrix…………………………………………………………………

6. Results of Logistic Regression model……………………………………………………

6.1 Full Model…………………………………………………………………………..

6.1.1 Confusion Matrix and Statistics ………………………………………

6.2 Variable Selection using Backward Step Model…………………………………

6.3 Fitting Final Model………………………………………………………………….

6.3.1 Confusion Matrix and Statistics ………………………………………

6.4 ROC Curve………………………………………………………………………….

6.5 Deviance Plot……………………………………………………………………….

6.6 Influential Points and Residuals………………………………………………….

6.7 Clean Model……………………………………………………………………….

6.8 Test for Comparison of Models…………………………………………………

6.9 Performance Metrics for Final model……………………………………………

7. Results of Decision Trees…………………………………………………………………

7.1 Confusion Matrices and Statistics………………………………………………

7.2 Variable importance in Decision Trees………………………………………..

7.3 New Decision Tree Model………………………………………………………

8. Conclusion…………………………………………………………………………………

9. Group Roles………………………………………………………………………………..

10. Future Work………………………………………………………………………………..

1. **INTRODUCTION**

The prevalence of liver disease has been on a significant rise, largely attributed to lifestyle and environmental factors such as excessive alcohol consumption, inhalation of harmful gasses, consumption of contaminated food and pickles, and drug intake. This escalating trend poses a substantial challenge to healthcare systems globally. In response, the utilization of predictive algorithms presents a promising avenue. By leveraging a comprehensive dataset that encapsulates a range of health metrics, these algorithms aim to facilitate the early detection and diagnosis of liver diseases. The primary objective is to lessen the diagnostic burden on medical professionals, thereby streamlining the healthcare process and potentially improving patient outcomes. This initiative represents a crucial step towards integrating advanced data analytics in medical diagnostics, offering a proactive approach in the battle against the rising tide of liver diseases.

1. **PROBLEM**

In the context of healthcare data analytics, our primary objective is to develop a predictive model capable of determining the likelihood of liver diseases in individuals. This model will analyze various chemical levels in the body, such as liver enzymes, bilirubin, albumin, and other relevant biomarkers. By scrutinizing these chemical indicators, the model aims to identify patterns and correlations that are indicative of liver diseases. The ultimate goal is to facilitate early diagnosis and intervention by accurately predicting the presence of liver conditions based on the analyzed chemical levels in the body. This approach is expected to enhance healthcare outcomes by enabling timely treatment and management of liver diseases.

1. **DATASET DESCRIPTION**

The dataset used for this project will be ‘Indian Liver Patient Records’ from Kaggle. This dataset contains health indicators related to liver such as bilirubin levels, proteins, enzymes, and more for patients, alongside their age and gender. Patients with Liver disease have been continuously increasing because of excessive consumption of alcohol, inhale of harmful gasses, intake of contaminated food, pickles and drugs. This dataset contains 416 liver patient records and 167 non liver patient records collected from Northeast of Andhra Pradesh, India. There are total 11 variables in our chosen dataset, and they are briefly explained below-

1. Age: The age of the patients, typically measured in years.

2. Gender: The sex of the patients, indicating whether they are male or female.

3. Total Bilirubin: The total amount of bilirubin in the blood, measured in milligrams per deciliter (mg/dL). Elevated levels may indicate liver dysfunction.

4. Direct Bilirubin: The amount of conjugated bilirubin in the blood, measured in mg/dL. Elevated levels can be a sign of liver or bile duct problems.

5. Alkaline Phosphatase: The level of alkaline phosphatase enzyme in the blood, measured in International Units per liter (IU/L). Elevated levels may suggest liver or bone issues.

6. Alamine Aminotransferase (ALT): The concentration of the ALT enzyme in the blood, measured in IU/L. Elevated ALT levels can indicate liver damage.

7. Aspartate Aminotransferase (AST): The concentration of the AST enzyme in the blood, measured in IU/L. Elevated AST levels can be a sign of liver or heart problems

8. Total Proteins: The total protein concentration in the blood, measured in grams per deciliter (g/dL). It includes various proteins, such as albumin and globulins.

9. Albumin: The concentration of albumin in the blood, measured in g/dL. Albumin is an important protein produced by the liver, and its levels can reflect liver function.

10. Albumin and Globulin ratio (A/G ratio): The ratio of albumin to globulins in the blood, providing additional information about liver and kidney function.

11. Dataset (Changed to Target): This attribute might indicate whether the patient has liver disease or not, serving as the target variable or label for classification tasks. It is typically binary, with values like 1 (indicating liver disease) and 0 (indicating no liver disease).

The dataset consists of 4 NA values which were cleaned and omitted. For better understanding, the dataset variable named dataset changed to target in the further analysis of data.

1. **PROCEDURES**

**Logistic Regression :**Logistic regression is designed for binary classification tasks. The target variable in this dataset indicates the presence (1) or absence (0) of liver disease, making it a binary outcome. It allows for the modeling of the probability that a patient has liver disease based on the independent variables (health indicators). Health indicators are not likely to be linearly related to the likelihood of liver disease. Logistic regression does not assume a linear relationship between the dependent and independent variables. Instead, it uses a logistic function to model a binary outcome.

**Decision trees:** A decision tree Algorithm for comparing it with Logistic regression as it could be used for the dataset to depicted the confusion matrices because it can handle non-linear relationships between features, which might improve the low sensitivity observed in the current models.

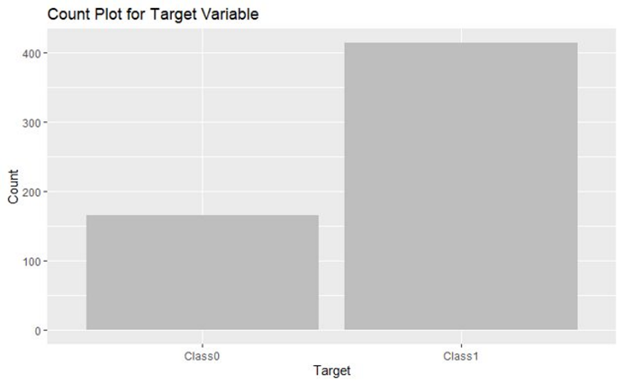
1. **DATA VISUALIZATIONS**

Figure 1 Distribution of Target

**5.1 DISTRIBUTION OF TARGET:**

The x-axis denotes the two categories of the 'Target' variable, denoted by 'CLASS0' and 'CLASS1'. The y-axis represents the frequency for each category. The Class0 has a significantly lower count than Class1.

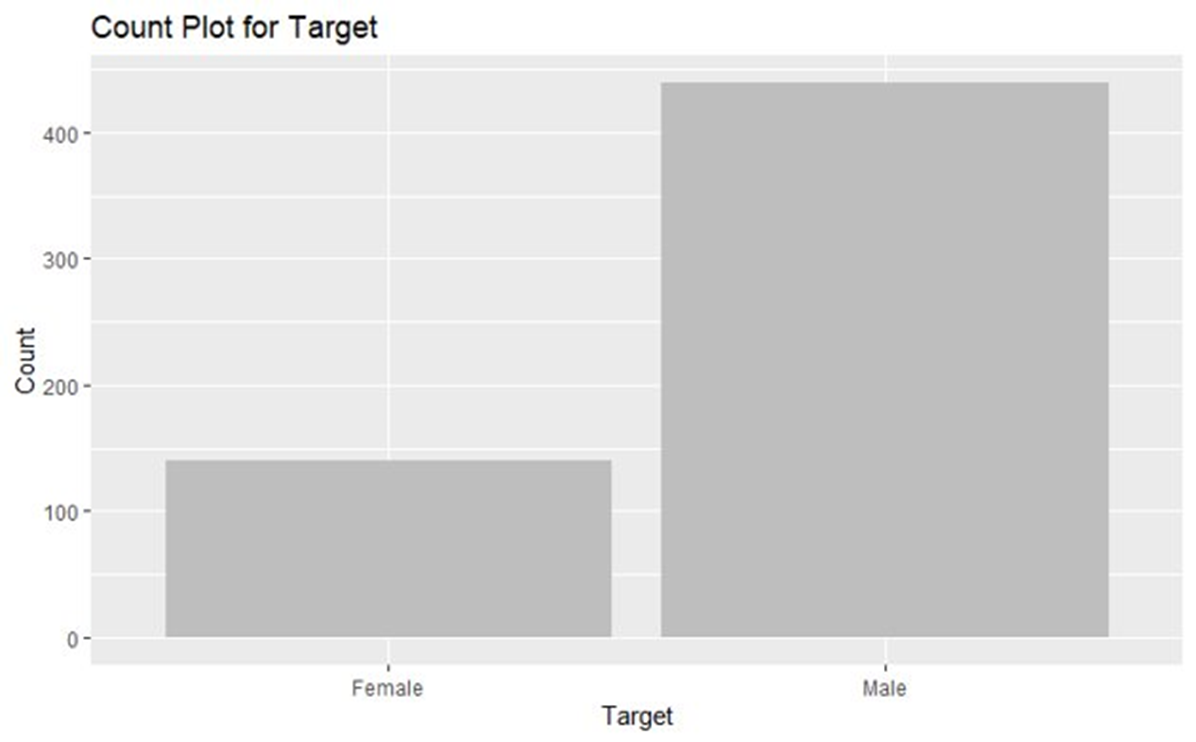
**5.2 DISTRIBUTION OF GENDER FOR TARGET:**

Figure 2 Distribution of Gender for Target

This visual suggests that whatever the "Target" represents, it is more common in males than females within the dataset from which this chart was generated.

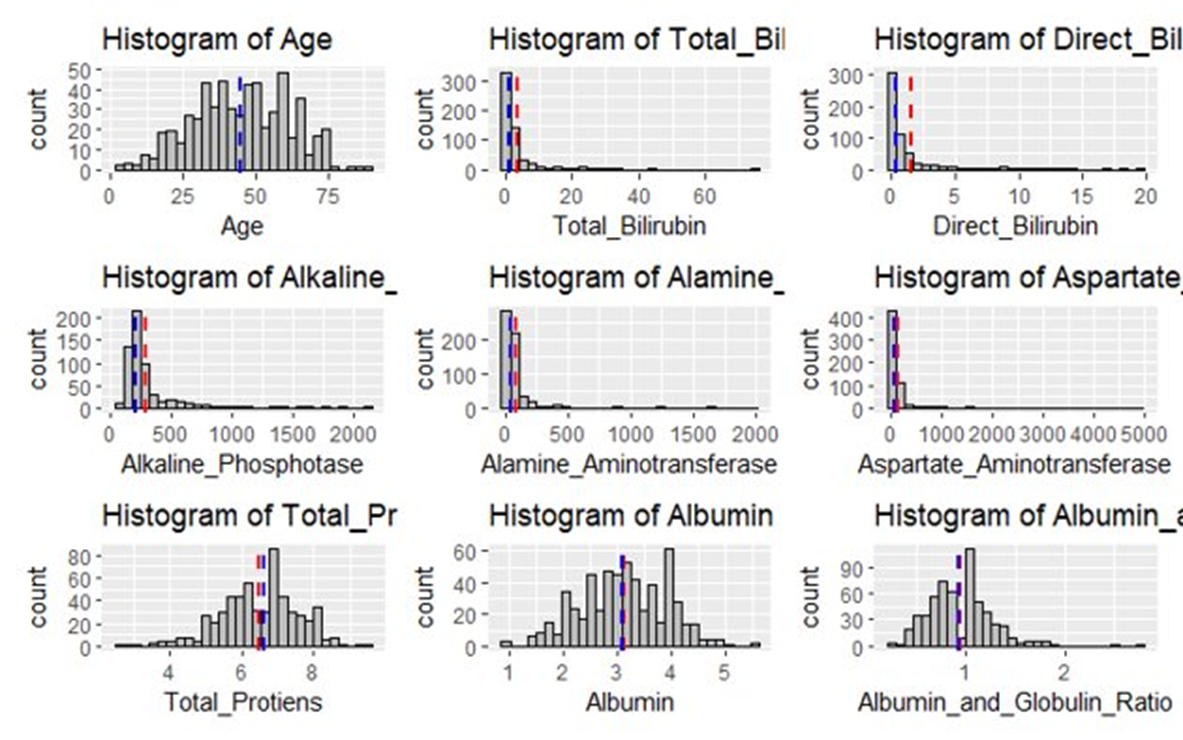


Figure 3 Histograms for all Variables

**5.3 HISTOGRAM:**

Most individuals are clustered between 25 to 75 years with a median age around 50.The majority of Total\_Bilirubin values are concentrated close to 0, with a few extending up to 60. The concentration of Alkaline\_Phosphotase is below 500, but some values spread out to 2000.Most of Alamine\_Aminotransferase data points are clustered around 0-500, with some outliers reaching 2000.A high concentration of Aspartate Aminotransferase is noted near 0, with a few values stretching to 5000.The distribution of Total\_Proteins is relatively uniform between 4 to 8, with a median around 6.Data of Albumin is spreaded between 1 and 5, peaking around values 3 and 4.

The majority of Albumin\_and\_Globulin\_Ratio observations lies between 0.5 to 2 with a peak near 1.

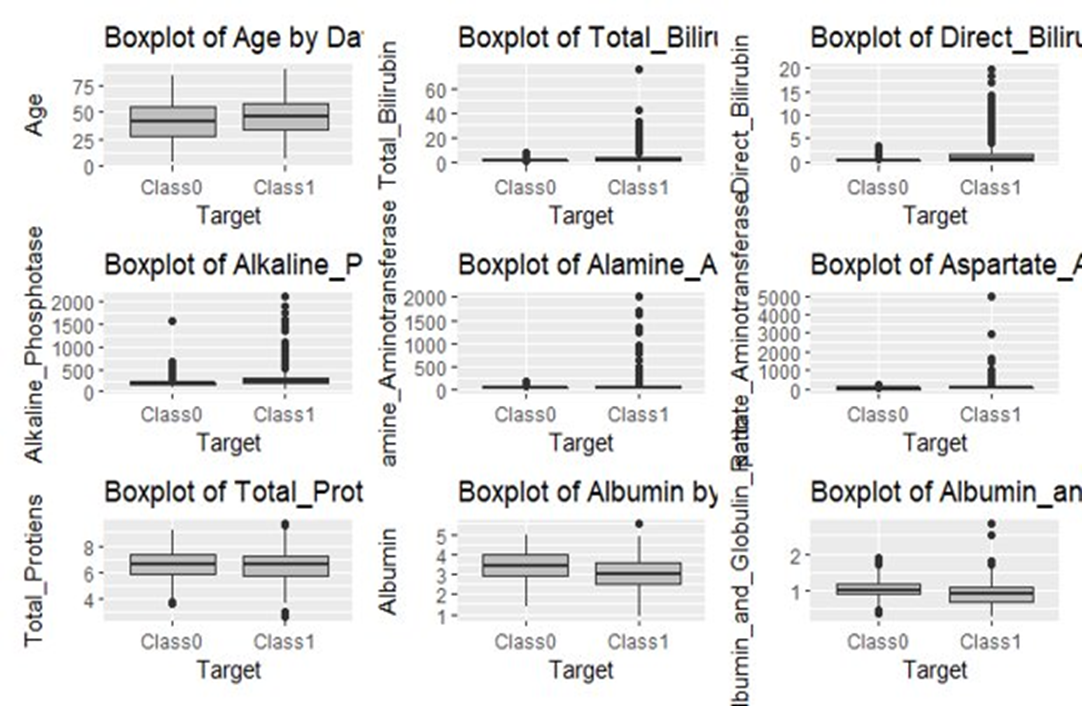


Figure 4 Box plots for all Variables

**5.4 PLOT GRID (Box plot):​**

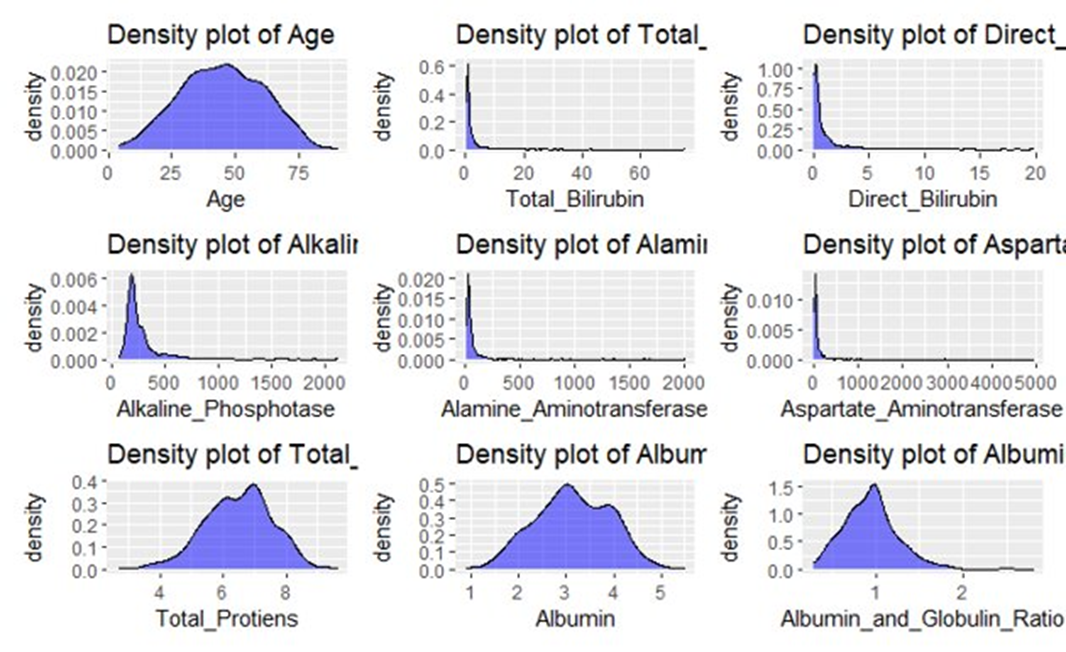
These boxplots show medical features categorized by a binary "Target." Group 1 consistently exhibits higher values and more outliers for liver-related metrics, indicating potential liver issues, while both groups have similar distributions for other attributes. The presence of outliers, especially in group 1, suggests possible severe medical conditions or data anomalies, hinting at the potential for classifying individuals based on these medical attributes.

Figure 5 Density plot for all Variables

**5.5 DENSITY PLOT:​**

The density plots depict the distribution of clinical metrics for a population. Age is centered around middle adulthood. Most biochemical markers are right-skewed, indicating a few individuals with high values. Total proteins and albumin have nearly normal distributions, suggesting typical ranges for most, while the albumin and globulin ratio is mostly below 1. These plots indicate a range of health statuses, with most within normal bounds but some potentially showing clinical abnormalities.​

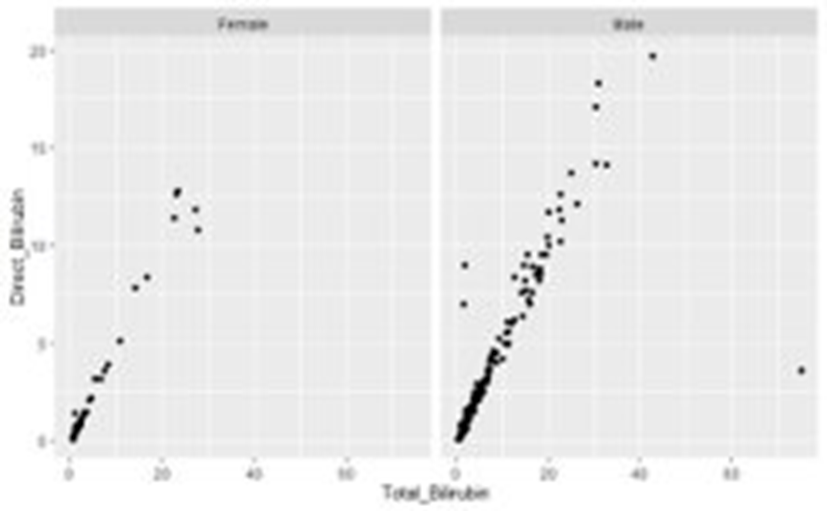
**5.6 SCATTERPLOTS**

Figure 6 Direct Bilrubin vs Total Bilrubin

**5.6.1 DIRECT BILIRUBIN VS. TOTAL BILIRUBIN:**

While both genders show a positive correlation between total and direct bilirubin, males seem to have a wider range of total bilirubin values, and the distribution of their data points is different from that of females. The presence of outliers in both groups suggests unique cases or potential data anomalies.

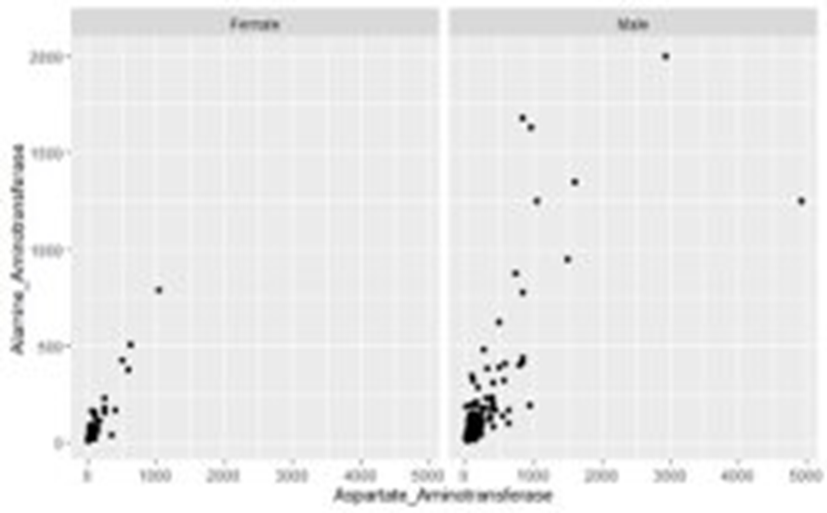
**5.6.2 ALAMINE AMINOTRANSFERASE VS. ASPARTATE AMINOTRANSFERASE :**

Figure 7 AlamineAT vs AspartateAT

While both genders generally show low levels of enzymes, males display a slightly broader distribution, with a few having significantly elevated enzyme levels. The scatter pattern, especially for males, suggests diverse conditions or factors affecting enzyme levels in different ways.

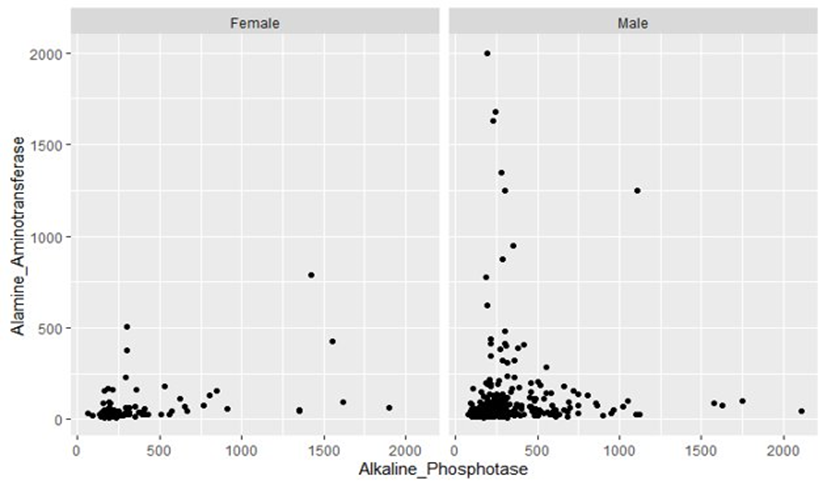


Figure 8 AlamineAT vs Alkakine phosphate

**5.6.3 ALAMINE AMINOTRANSFERASE VS. ALKALINE PHOSPHATE:**

For both genders, most individuals have low levels of both enzymes. There are only a few instances in either gender where elevated levels of enzymes are observed. However, males seem to have a slightly more pronounced clustering in the lower levels of alkaline phosphatase as compared to females. It has rare exceptions of elevated levels.

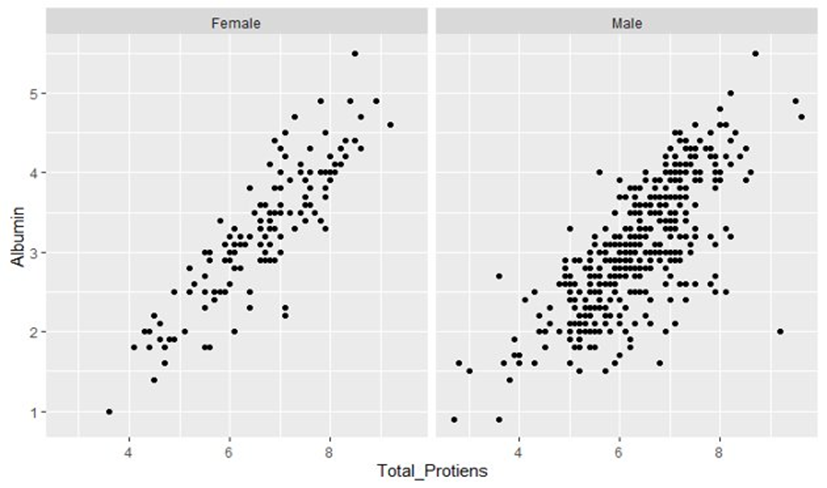
**5.6.4 ALBUMIN VS. TOTAL PROTEINS:**

Figure 9 Albumin vs Total Proteins

Both the total proteins and albumin are distributed across a range, with varying densities of data points across different regions of the plot for both genders. There's a positive correlation between total proteins and albumin. Males have a denser clustering of data points in the mid to high ranges, whereas females have a more spread-out distribution across the plot.

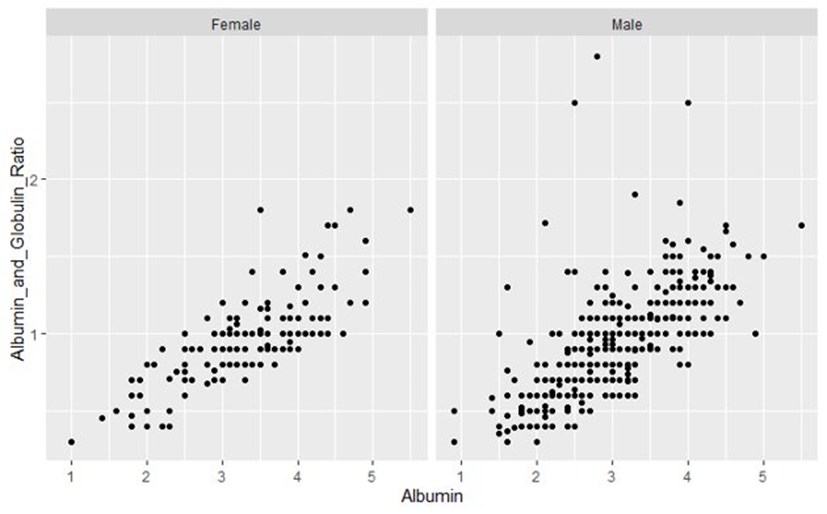
**5.6.5 ALBUMIN-GLOBULIN RATIO VS. ALBUMIN:**

Figure 10 AGR vs Albumin

Both albumin and albumin and globulin ratio plots indicate positive correlations for both genders, with higher albumin levels associated with higher albumin and globulin ratios. Males show a slightly more pronounced mid-range clustering than females, but both genders mostly have mid-range values for these metrics.

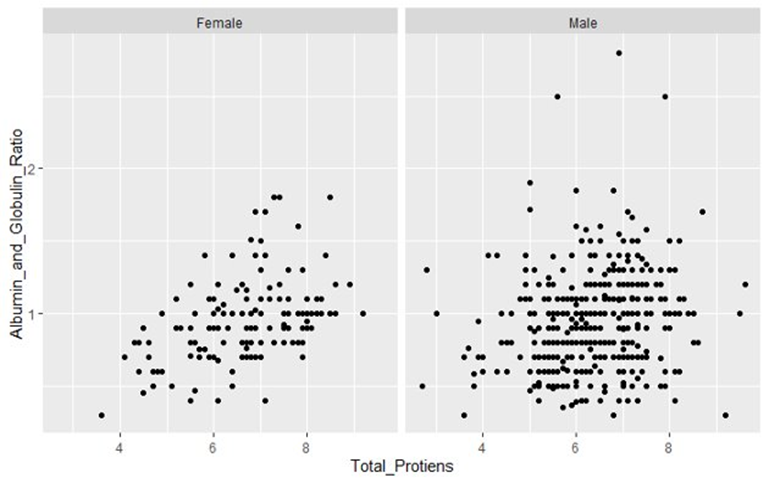
**5.6.6 ALBUMIN-GLOBUMIN RATIO VS. TOTAL PROTEINS:**

Figure 11 AGR vs Total Proteins

Both genders display a lack of a strong linear correlation between total proteins and albumin and globulin ratio, the distribution and density of data points differ slightly between the two. Males have a more dispersed distribution in the mid-ranges, while females have a more concentrated clustering in specific regions

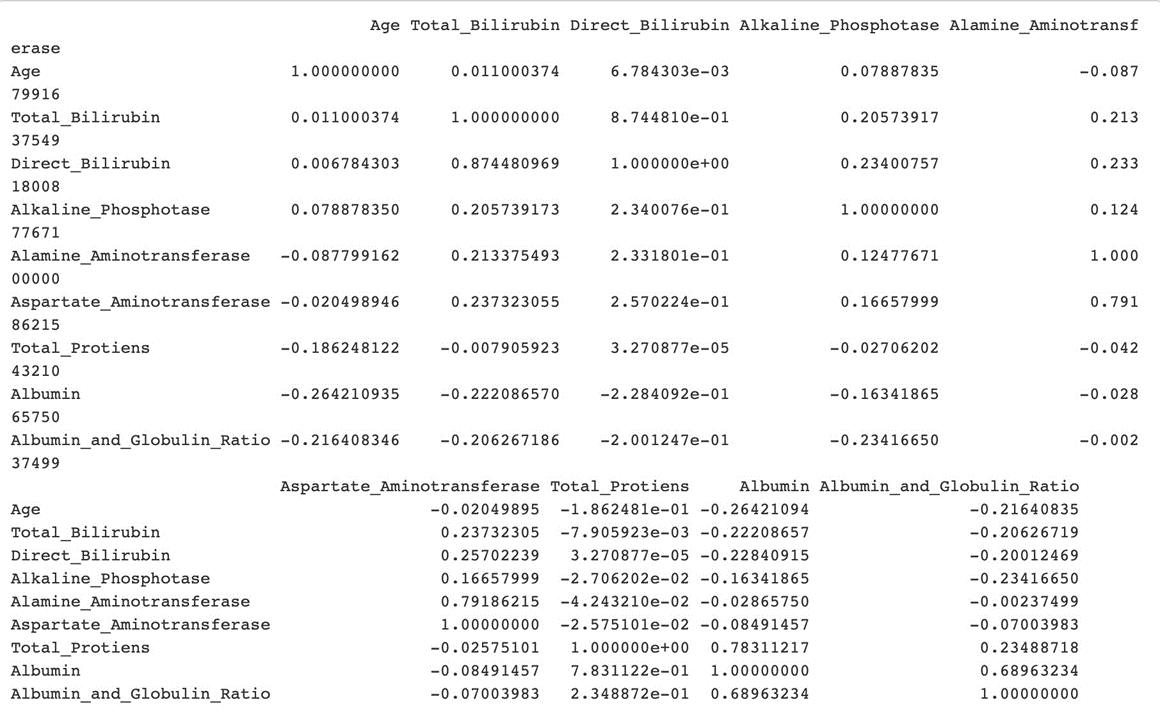
**5.7 CORRELATION MATRIX:**

Figure 12 Correlation matrix

1. **RESULTS OF LOGISTIC REGRESSION MODEL**

**6.1 FULL MODEL(ALL VARIABLES):**

****Logistic regression model identifies ‘alamine aminotransferase' and ‘total proteins' as significant predictors for the target outcome, with positive coefficients indicating a higher likelihood of the outcome as these values increase. The exclusion of ‘gender (male)' due to singularity suggests gender is a binary variable in this model. Other variables did not significantly predict the outcome at the 5% significance level. The model improves upon the null model and demonstrates adequate fit with convergence in 8 iterations. The Akaike information criterion(AIC) is 426.59 for all variables.

Figure 13 Logistic regression model with all variables

**6.1.1 CONFUSION MATRIX AND STATISTICS FOR FULL MODEL:**

A screenshot of a computer

Description automatically generated-True negatives (Class0 predicted as Class0): 10

Figure 14 Confusion matrix of full model

-False negatives (Class1 predicted as Class0): 39

-False positives (Class0 predicted as Class1): 8

-True positives (Class1 predicted as Class1): 116

This performance metrics of a binary classification model, revealing an accuracy of 72.83% and a low kappa value of 0.1726, suggesting minimal agreement beyond chance. The model exhibits high specificity (93.55%) but low sensitivity (20.41%), indicating it's better at predicting negative instances than positive ones. The McNemar's test p-value is very low (1.209e-05), showing a significant difference in the model's prediction ability for the two classes. The balanced accuracy stands at 56.98%, reflecting the model's moderate ability to balance between sensitivity and specificity.

**6.2 VARIABLE SELECTION USING BACKWARDS STEP MODEL**

The stepwise variable selection process for logistic regression identified a final model with age, direct bilirubin, alamine aminotransferase, total proteins, and albumin as predictors for the "Target" variable.The coefficients indicate that age, alamine aminotransferase, and total proteins positively influence the likelihood of the "Target" outcome, while direct bilirubin and albumin have negative effects. With a lower AIC of 418.7, the model demonstrates improved fit and simplicity, enhancing its utility in predicting the "Target" variable.

**A screenshot of a computer

Description automatically generated6.3 FITTING FINAL MODEL:**

The final model provided by step function has only significant predictors “Direct\_Bilirubin","Total\_Protiens",”Age”, "Alamine\_Aminotransferase",and "Albumin".

This model has the lowest AIC=418.68 . The model's adequacy is suggested by the reduction in deviance, yet the warning about probabilities being 0 or 1 suggests potential overfitting issues.With seven iterations of Fisher scoring for convergence, the model’s fit is decent but needs cautious interpretation due to the extremes in predicted probabilities.

Figure 15 Logistic regression model with significant variables

**6.3.1 CONFUSION MATRIX AND STATISTICS FOR FINAL MODEL:**

A screenshot of a computer

Description automatically generatedThis final model gives us better accuracy 73.41%.This model has a slightly higher kappa statistic (0.184 vs. 0.1726), indicating a marginal improvement in performance beyond random chance. The little lower p-value shows a higher statistical significance suggesting stronger evidence against the null hypothesis. The model has a high specificity of 94.35% but low sensitivity of 20.408.The balanced accuracy is 57.382%, reflecting moderate average performance of sensitivity and specificity. The model is better at predicting negative cases than positive cases. The 'positive' class for analysis is 'Class0'.

Figure 16 Confusion matrix of Final Model

**A graph of a curve

Description automatically generated6.4 ROC CURVE :**

Figure 17 ROC curve of final model

The ROC curve indicates that logistic regression model has a moderate-to-good performance on the dataset but it can be improved. The AUC value of 0.7605 suggests that the model has a good discrimination capacity, but there is still room for improvement. The shape of the ROC curve shows that the model has a good true positive rate for most thresholds.

A graph with lines and numbers

Description automatically generated**6.5 DEVIANCE PLOT :**

The residual deviance plot indicates that the logistic regression model is mis specified, exhibits non-linearity, and suffers from heteroscedasticity. It also identifies outliers and suggests poor predictive performance at the extremes. Addressing these issues can improve the model's reliability and effectiveness.

Figure 18 Deviance plot

**A screenshot of a computer

Description automatically generated6.6 INFLUENTIAL POINTS AND RESIDUALS:**

**A graph showing a number of dots

Description automatically generated with medium confidence**Observations 115 and 116 show significant standardized residuals with relatively high leverage, and observation 270 has a notable Cook's D value, indicating a potentially strong influence on the model. These points, particularly 270, may disproportionately affect the regression analysis and could be candidates for further investigation or removal to improve model accuracy.

Figure 19 Influential points and residuals

**A screenshot of a computer

Description automatically generated6.7 CLEAN MODEL(REMOVED INFLUENTIAL POINTS)**

Figure 20 Clean Model

Here is the model after removing influential points suggests that the removal has potentially improved the model's predictive reliability, as evidenced by the significant p-values and a lower AIC=409.06 indicates a better model fit.

**6.8 TEST FOR COMPARISON OF MODELS:A screenshot of a computer code

Description automatically generated**

The p-value of 0.8372 is greater than 0.05, which means that the difference in deviance between Model 1 and Model 2 is not statistically significant.

In other words, there is no evidence that Model 1 is a better fit for the data than Model 2.This suggests that the three predictor variables that were added to Model 2 (Gender, Alkaline\_Phosphotase, Aspartate\_Aminotransferase, and Albumin\_and\_Globulin\_Ratio) do not significantly improve the fit of the model.

Figure 21 Anova Test

**6.9 PERFOMANCE METRICS FOR FINAL MODEL:**

PRECISION : Precision measures the proportion of positive predictions that are actually correct. In this case, precision is 0.5882, which means that out of all the predictions your model made as positive, 58.82% were indeed correct.

RECALL : Recall measures the proportion of actual positives that are correctly identified by the model. In this case, recall is 0.2041, which means that only 20.41% of the actual positive instances were correctly classified by the model.

F-SCORE: The F1-score is a weighted average of precision and recall, and it represents a balance between the two. In this case, the F1-score is 0.3030, which indicates that the model has a relatively low overall performance.

1. A screenshot of a computer

   Description automatically generated**RESULTS OF DECISION TREES**

Figure 22 Confusion matrix and Statistics for decision tree full model.

**7.1 CONFUSION MATRICS AND STATISTICS:**

The statistics reveal that the model has an accuracy of 0.6185, indicating that it correctly predicts around 62% of cases. The kappa value of 0.2312 suggests fair agreement beyond chance. The model has relatively high sensitivity (0.6939), which means it is good at identifying true positives but lower specificity (0.5887), implying it's less effective at identifying true negatives. The positive predictive value is 0.4000, meaning that when it predicts class0, it's accurate only 40% of the time, and the negative predictive value is 0.8295, indicating a correct prediction of class1 about 83% of the time. Overall, the model shows potential but could benefit from improvements in specificity and precision.​

**A graph with text on it

Description automatically generatedA screenshot of a computer

Description automatically generated7.2 VARIABLE INPORTANCE FOR DECISION TREES:**

Figure 23 Variable importance for decision trees

Here, we can see only top three are the most important predictors in determining the outcome. Gender (Male) appears to have no influence, as its importance is rated zero, suggesting that biochemical markers are more indicative of the target outcome than demographic information in this model.

A screenshot of a computer

Description automatically generated**7.3 NEW DECISION TREE MODEL:**

Figure 24 Confusion matrix and Statistics for new decision tree model.

The model achieved an overall accuracy of 70.52% on the test dataset, with a 95% confidence interval indicating that the true accuracy is likely between 63.12% and 77.2%. The p-value of 0.66702 suggests some evidence that the model performs better than random guessing, but caution is warranted due to the relatively modest statistical significance. The kappa statistic of 0.3502 indicates a moderate agreement between the model's predictions and true class labels. The model demonstrated 67.35% sensitivity in correctly identifying positive class instances (class0) and 71.77% specificity in identifying negative class instances (class1). Class0 prevalence in the test set is 28.32%, and the model achieved a detection rate of 19.08% with a detection prevalence of 39.31% for class0 predictions.

1. A screenshot of a graph

   Description automatically generated**CONCLUSION**

Figure 25 Comparison of Logistic and Decision tree model

The logistic regression model is better than the decision tree model because it has a higher accuracy, specificity, positive predictive value, and negative predictive value. The decision tree model only has a higher sensitivity and detection rate. The logistic regression model is more interpretable than the decision tree model. This means that it is easier to understand how the model makes its decisions. In conclusion, the logistic regression model is a better model for predicting the class label of an instance. It is more accurate, has better class-specific performance, and is more interpretable than the decision tree model.