Liver Disease Prediction

GROUP 8

Presentation Outline

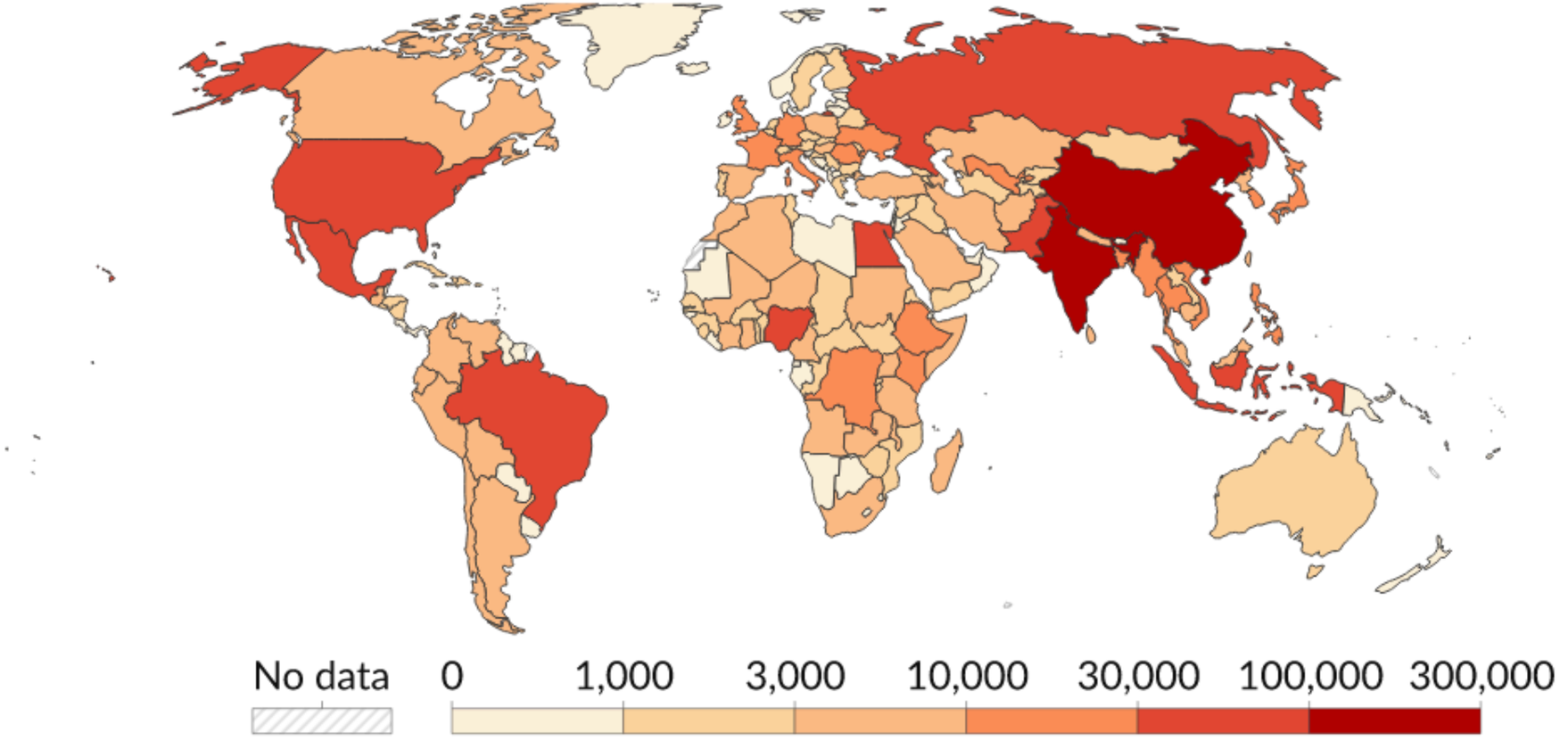
“Liver disease accounts for two million deaths annually and is responsible for 4% of all deaths”

-Global burden of liver disease: 2023 update

Deaths from liver disease, 2019

Annual number of deaths from cirrhosis and other chronic liver diseases.

Our World in Data



Data source: IHME, Global Burden of Disease (2019)

OurWorldInData.org/causes-of-death | CC BY

Liver Disease, an Indicator of Global Health

Currently liver disease pose a significant threat on the health of humans worldwide. Liver issues sometimes go unnoticed, and they are a silent killer with a huge impact.

Factors such as lifestyle choices, demographic variables and clinical variables have an effect on prevalence and severity of liver conditions.

Monitoring liver health and enabling early detection is vital for identifying health status and understanding trends.

MOTIVATION

- Optimal approach for identification is a liver biopsy, which is an invasive procedure and is infeasible to perform on all patients.
- In Sri Lanka constraints exist due to limited availability of equipment.
- Therefore, an accurate predictive model is essential, enabling early interventions and thus preventing patients deteriorating further.

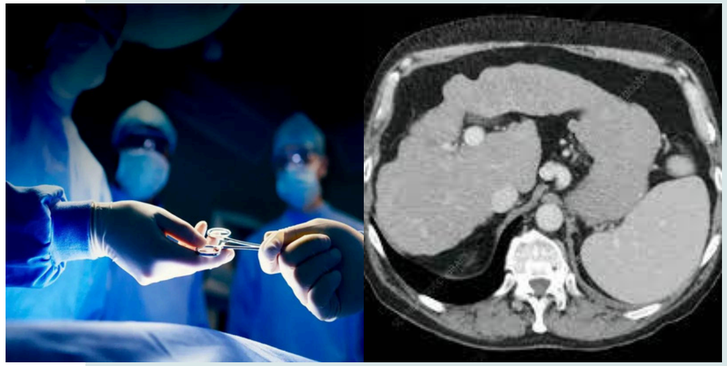
PREVIOUS RESEARCH INTO LIVER DISEASE PREDICTION

Serum biomarkers as an alternative for liver biopsy.
Blood tests performed to measure various levels and these used to calculate the biomarkers. Based on a threshold, occurrence of disease can then be identified.

Rahman, A. K. M. et.al. found a model with 75% accuracy using Logistic Regression.
Random forest approach with 75.7% accuracy was followed by Riya et.al.
Subabhrata fitted a model with 71.8% accuracy using support vector machines algorithm, while Pavan Prabu fitted a gradient boosting model with 81.35% accuracy.
The most accuracte model we came accorss in our research into prior literature was a model created by Ninad Bhase which was a random forest model with 99% accuracy.

ABOUT THE PROJECT

- We analyzed liver disease prevalence among patients using data from Kaggle platform.
- This project utilizes statistical methods and machine learning techniques to identify influential characteristics which are then used for liver disease prediction.
- Results show association between variables and the occurence of liver disease and multi-collinearity in the data
- Multiple models were evaluated to identify the optimal.



UNDERSTANDING OF KEY TERMS

Total Bilirubin

A combination of direct and indirect bilirubin

A protein found in all body tissues.

Alkaline phosphatase (ALP)

One of the two liver enzymes. It is also known as serum glutamic-oxaloacetic transaminase

Sgot Aspartate Aminotransferase

Albumin

A protein made by the liver

Sometimes referred to as conjugated, is the form of bilirubin which has been conjugated with glucoronic acid and is excreted in the bile.

Direct Bilirubin

Sgpt Alamine Aminotransferase

Amount of the enzyme alanine aminotransferase (ALT) in your blood

Measures albumin and globulin and is used to monitor nutritional status, immune function, and overall health

A/G ratio test

WHO WILL BENEFIT FROM OUR FINDINGS

- Patients
    - Healthcare Providers
        - Pharmaceutical Companies
            - Insurance Companies

RESEARCH QUESTIONS

What are the key patterns and trends in the data?

Which variables have a higher impact on accurately predicting Liver Disease and to what extent?

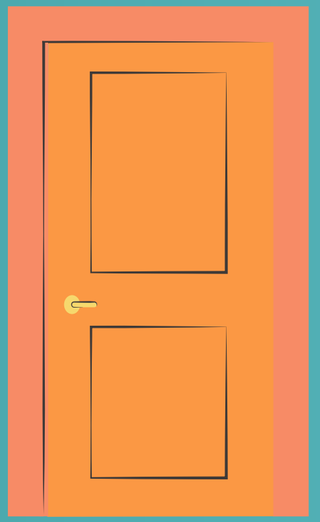
Are variables in the dataset are associated with each other?

What machine learning models are able to predict occurrence of Liver Disease?

What variables are highly associated with the response variable “Liver Disease”?

Which is the best model for the same?

INTRODUCTION TO DATASET



DESCRIPTION :

Liver Disease Patient Dataset obtained from the Kaggle website consists of data of different patients and the occurence of liver disease (Predictors were mainly clinical variables)

OBSERVATIONS: 30692

VARIABLES: 11

QUALITATIVE

Gender (Male/Female)

Result (Liver Patient/Non-Liver Patient)

QUANTITATIVE

TB Total Bilirubin

DB Direct Bilirubin

Alkphos Alkaline Phosphotase

Sgpt Alamine Aminotransferase

Sgot Aspartate Aminotransferase

TP Total Protiens

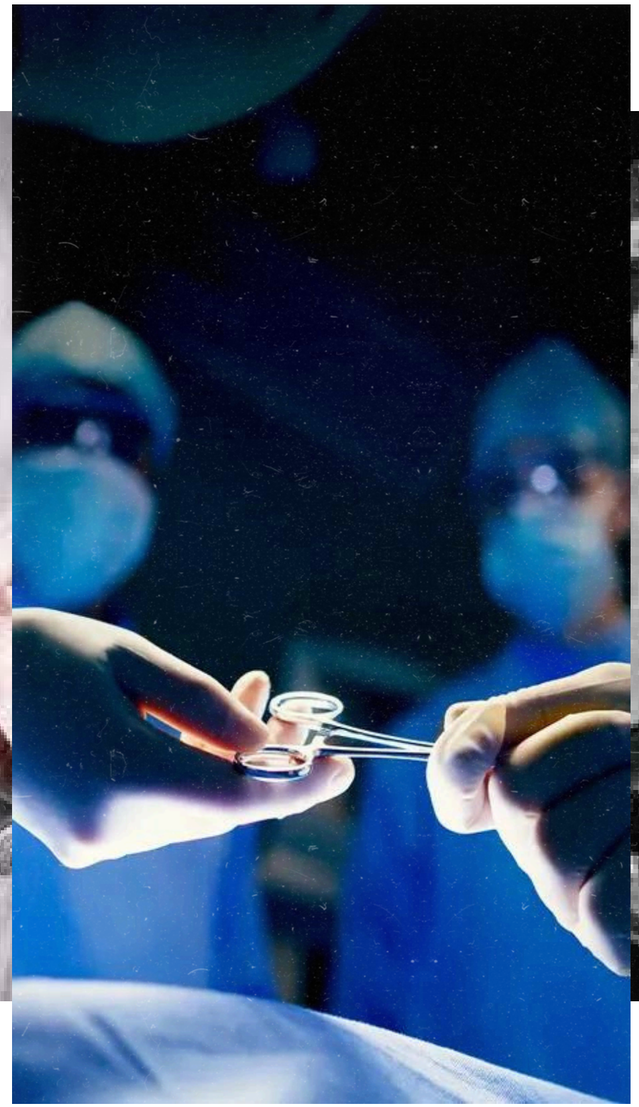
ALB Albumin

A/G Ratio Albumin and Globulin Ratio

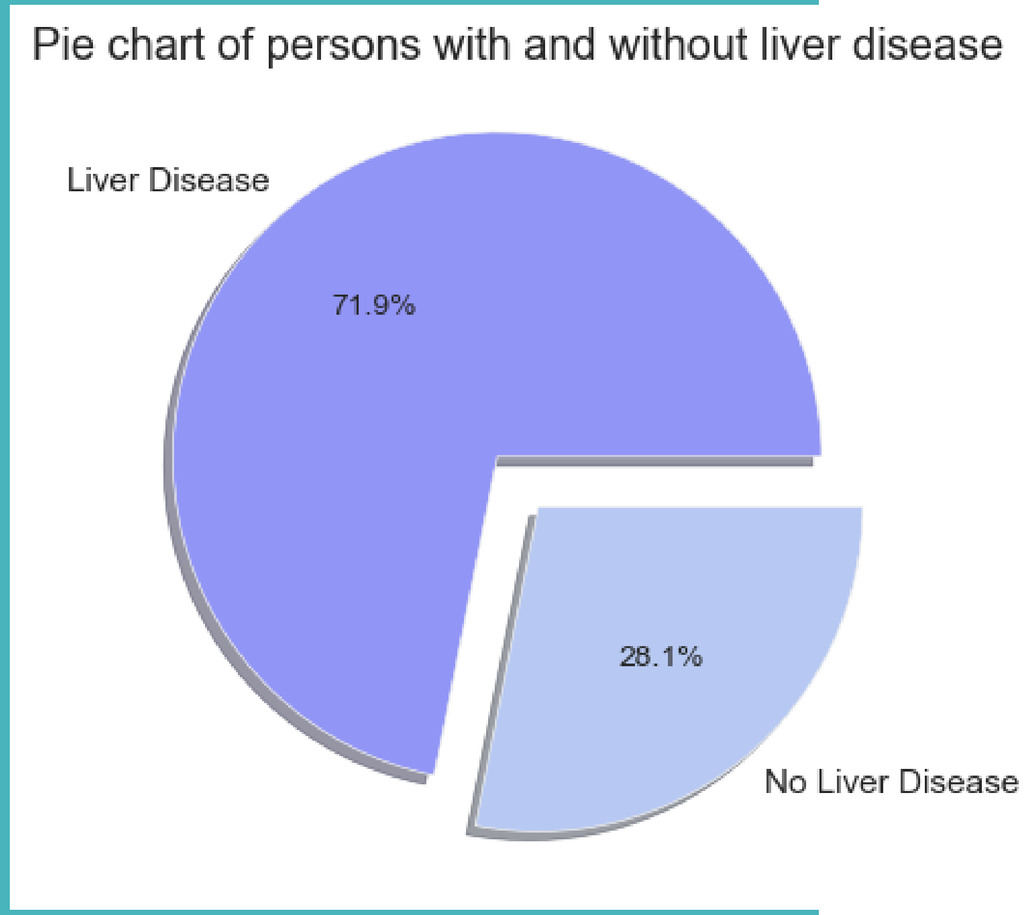
DATA PREPROCESSING

1. Rename Result Column as Liver_Disease
2. Encoded Liver_ Disease column of Liver Disease patients and Non liver Disease patients[1,2] as [1,0]
3. Encoded the 'Gender' column into numerical values using label encoding. Male is encoded as 1 and female as 0.
4. Remove the missing values and No duplicates are found
5. Scaling the variables to overcome the issue of having skewed variables

UNIVARIATE ANALYSIS

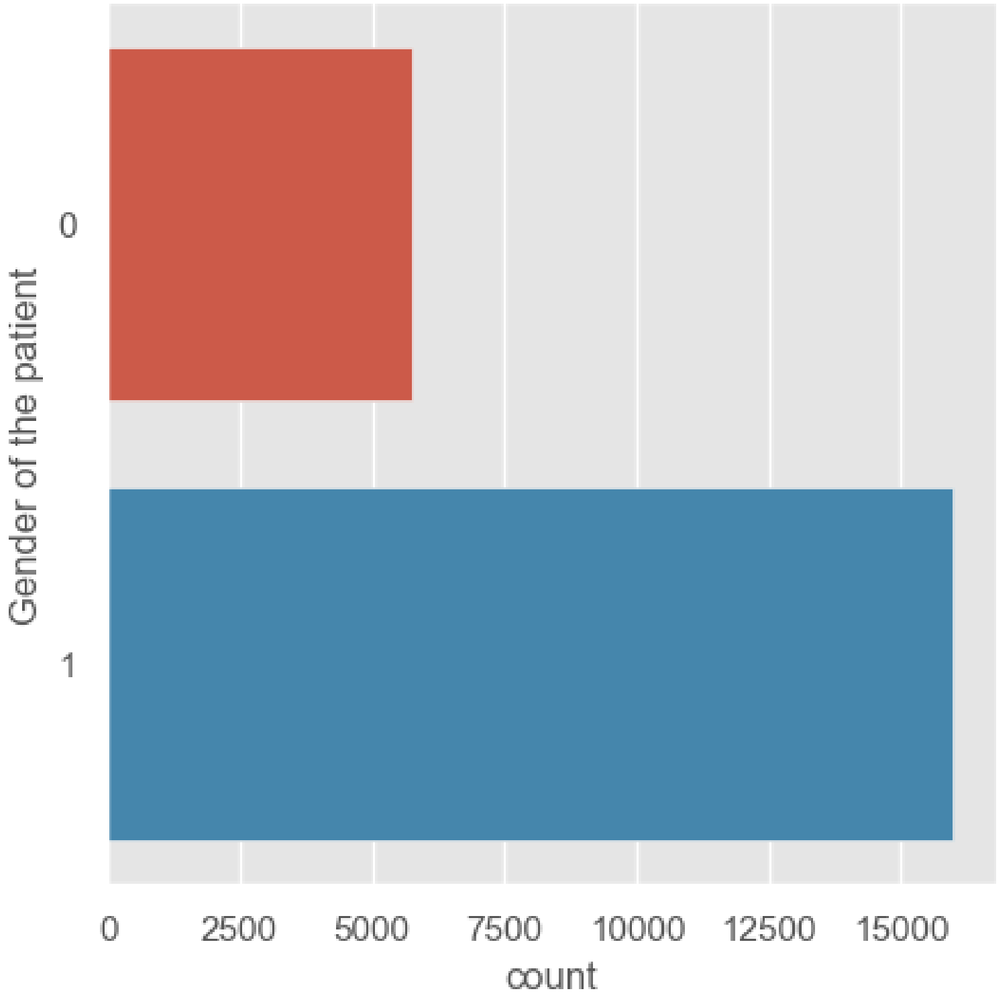


LIVER DISEASE



- Reveals a substantial prevalence of liver disease within the studied population, accounting for approximately 71.9%.

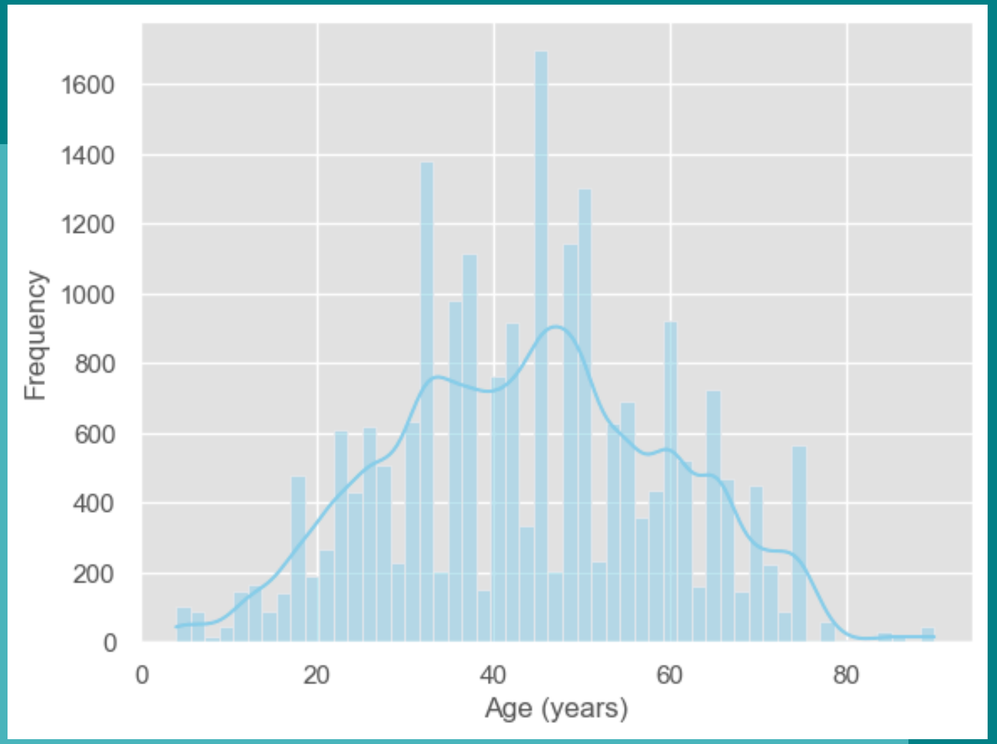
- Vividly depicts the substantial class imbalance within our dataset.

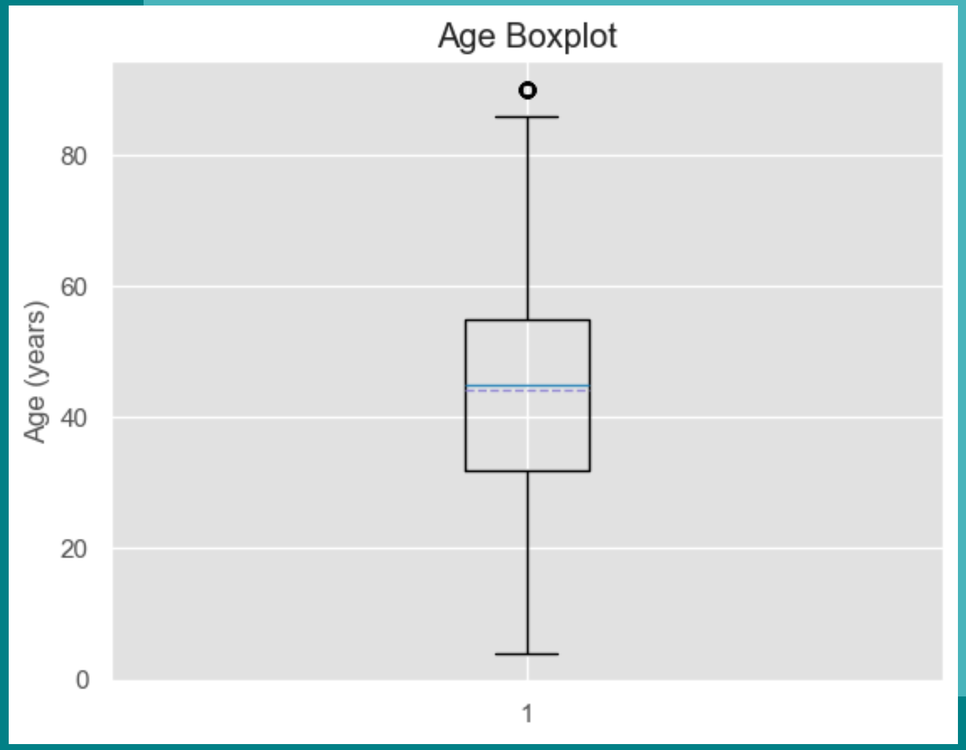


GENDER

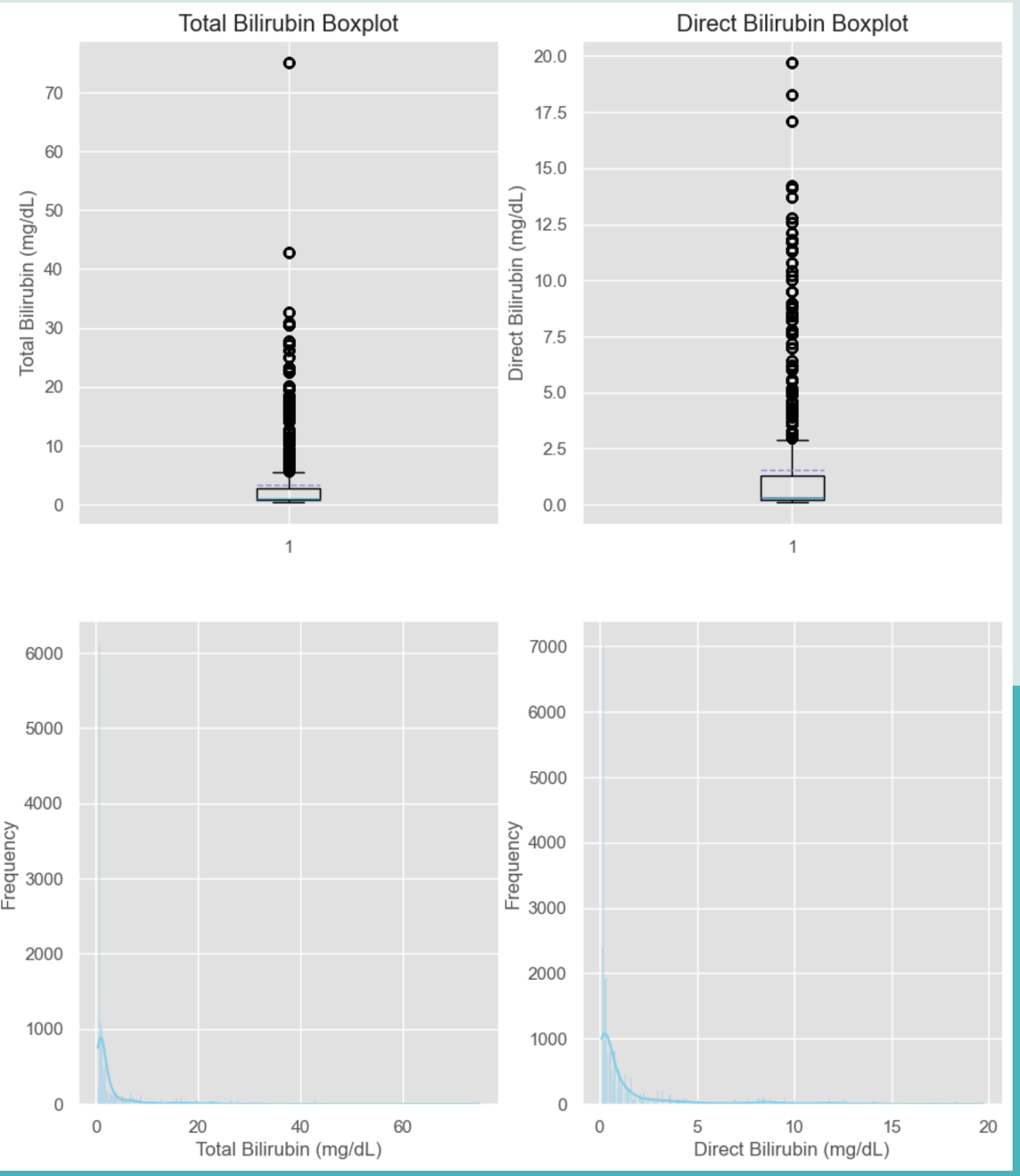
The data reveals a significant gender disparity among the patients, with a majority being male. This aligns with the findings of previous researchers.

AGE DISTRIBUTION





- The data reveals the majority of data falls between ages 30 and 60.
- Median age is around 45.
- Outlier exists beyond age 80.



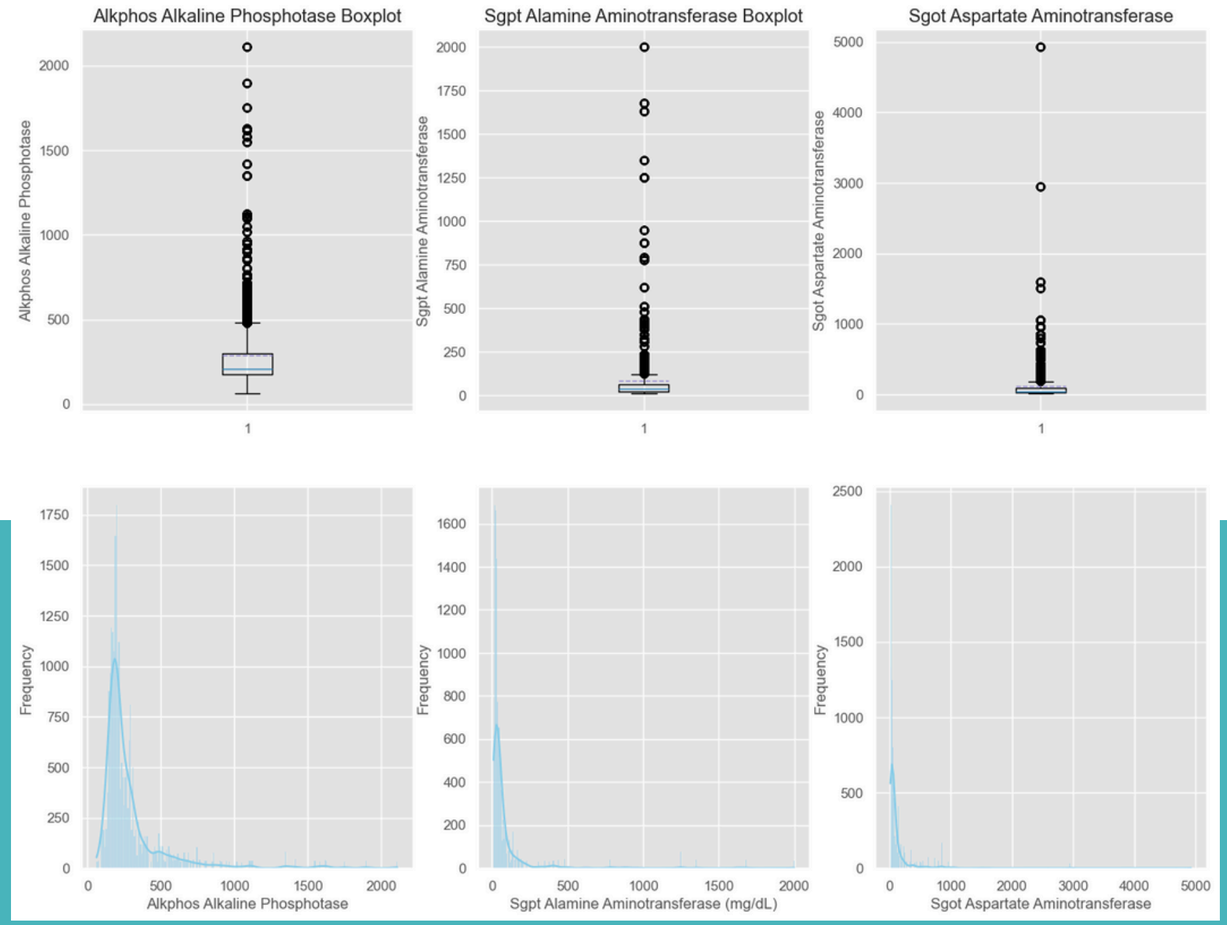
TOTAL BILIRUBIN AND DIRECT BILIRUBIN

- The majority of data falls within the IQR (between 10 and 20 mg/dL for Total Bilirubin and between 0.5 and 2.5 mg/dL for Direct Bilirubin) .
- There’s a prominent peak at the very beginning of the both histogram graph, close to zero on the x-axis indicating that most values for **both total bilirubin and direct bilirubin are low.**
- Right-tailed

ALKPHOS ALKALINE PHOSPHOTASE
SGPT ALAMINE AMINOTRANSFERASE &
SGOT ASPARTATE AMINOTRANSFERASE

- Most Alkphos Alkaline Phosphatase, SGPT, SGOT values are within their respective IQR.

- Most values are near the lower end for all three histograms (right-skewed), suggesting that lower alkaline phosphatase levels, SGPT, and SGOT values are more common.

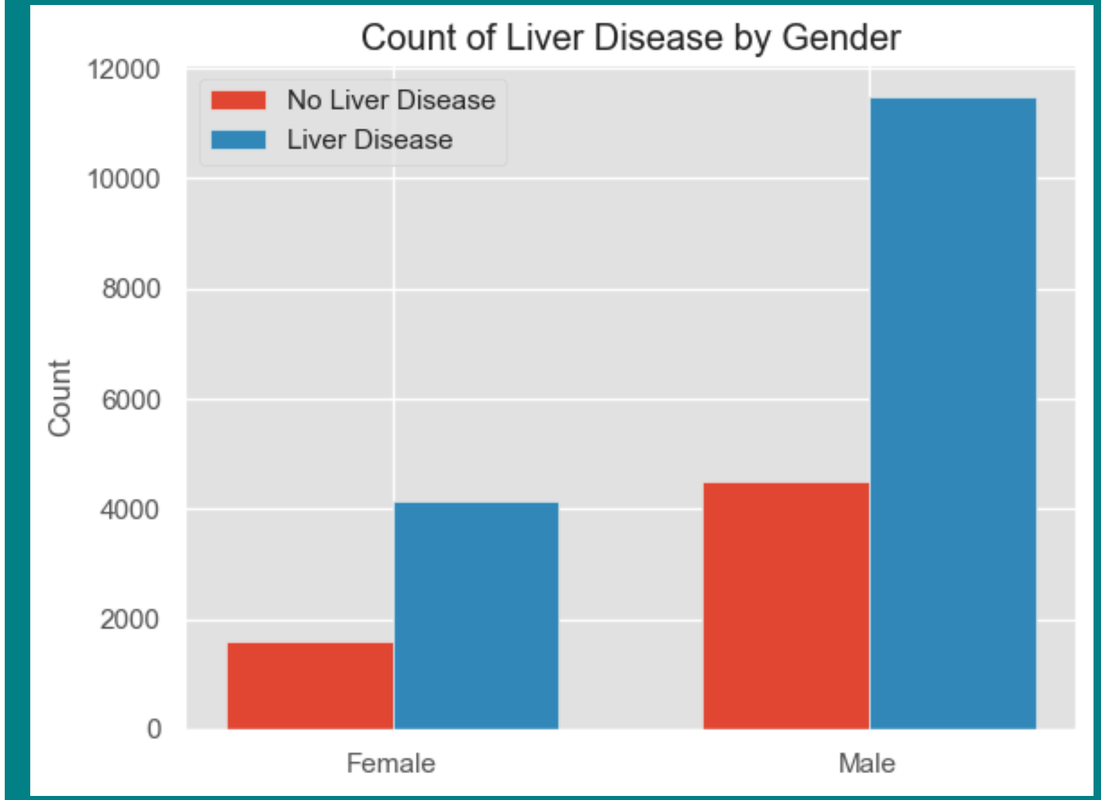


BIVARIATE ANALYSIS

GENDER VS LIVER DISEASE

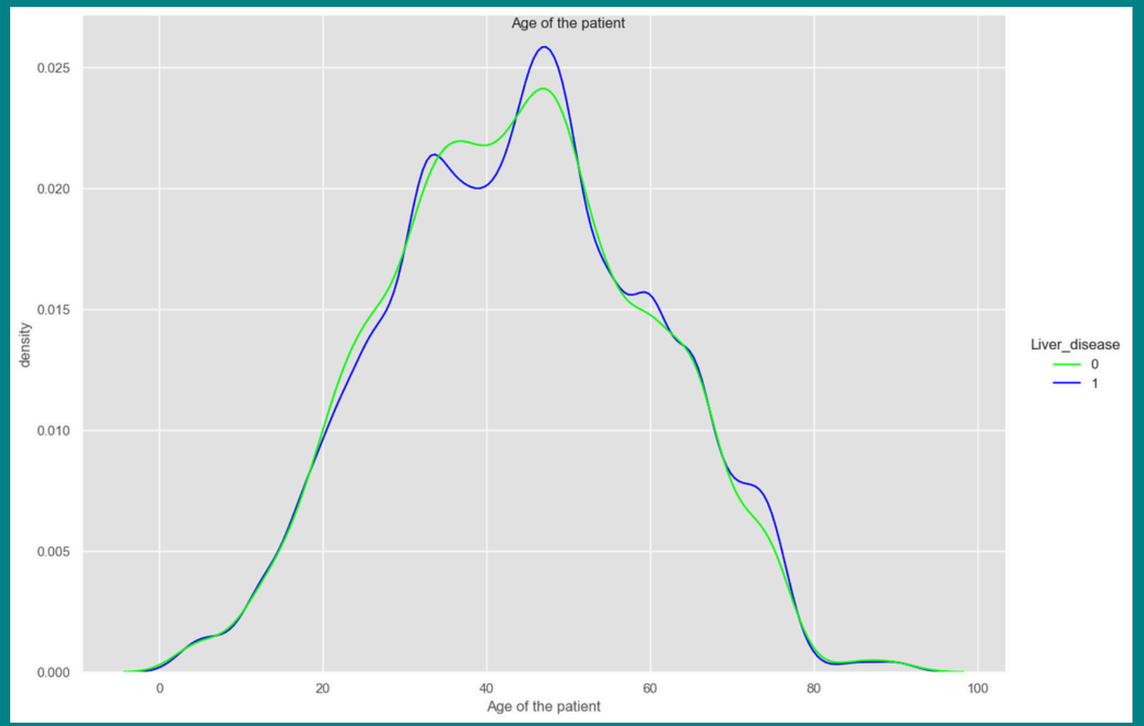
Males appear to be more affected by liver disease than females in this dataset.

The cell entries of contingency table represent the count of patients falling into specific combinations of liver disease presence and gender.




```
Contingency Table:
Liver_disease              0      1
Gender of the patient
0                       1604   4130
1                       4503  11489

Chi-Square Statistic: 0.062142952205099765
p-value: 0.8031404249782508
Degrees of Freedom: 1
Expected Frequencies:
[[ 1611.78026328  4122.21973672]
 [ 4495.21973672 11496.78026328]]
```

Based on the Chi-Square test, there is no significant association between liver disease and gender in this dataset (since the p-value > 0.05).

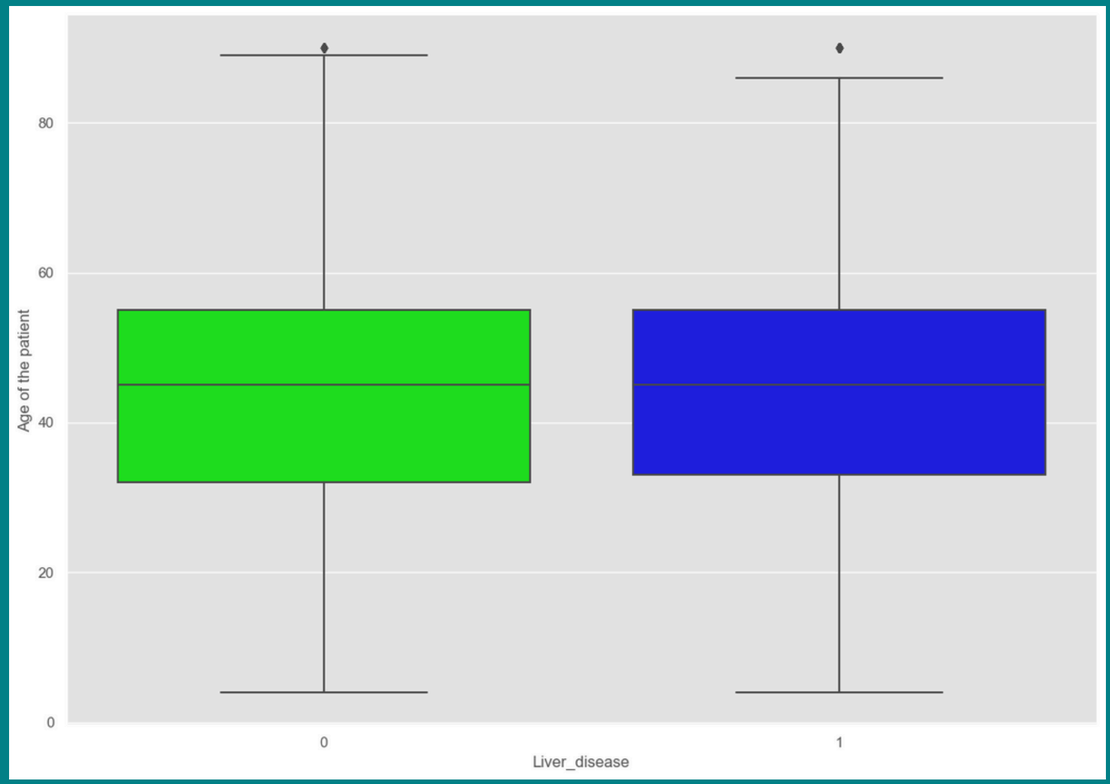


AGE VS LIVER DISEASE

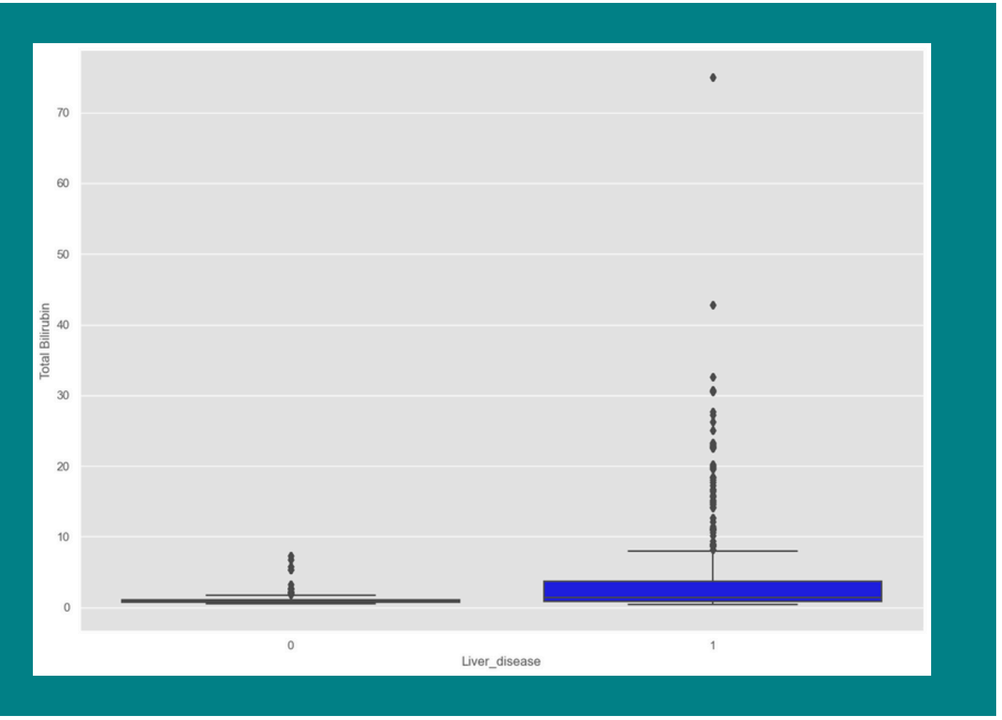
Both plots show the age range of patients from 0 to approximately 90 years.

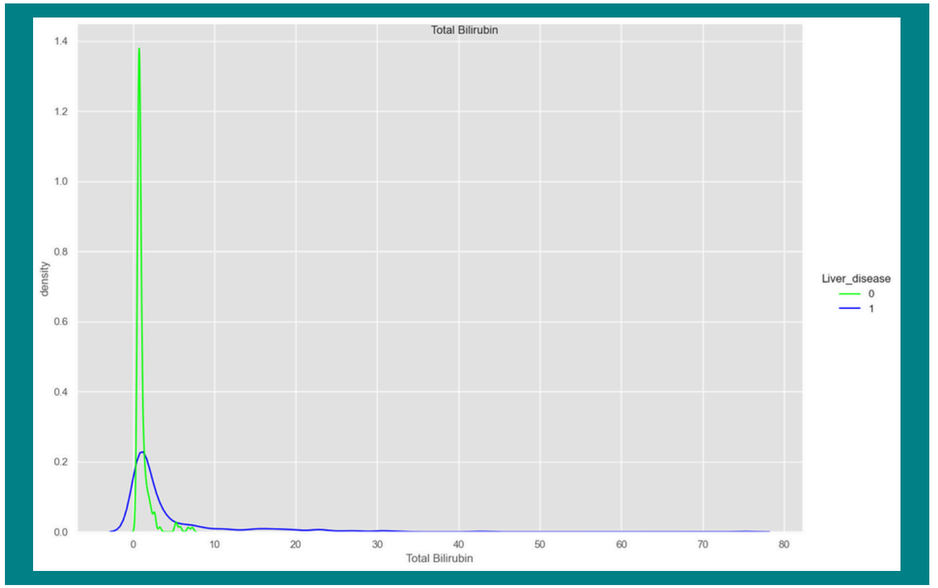
The length of the whiskers and the size of the boxes show there is a high variability for age distribution among the patients who doesn't have liver disease.

Median age of the group is approximately same for both groups.

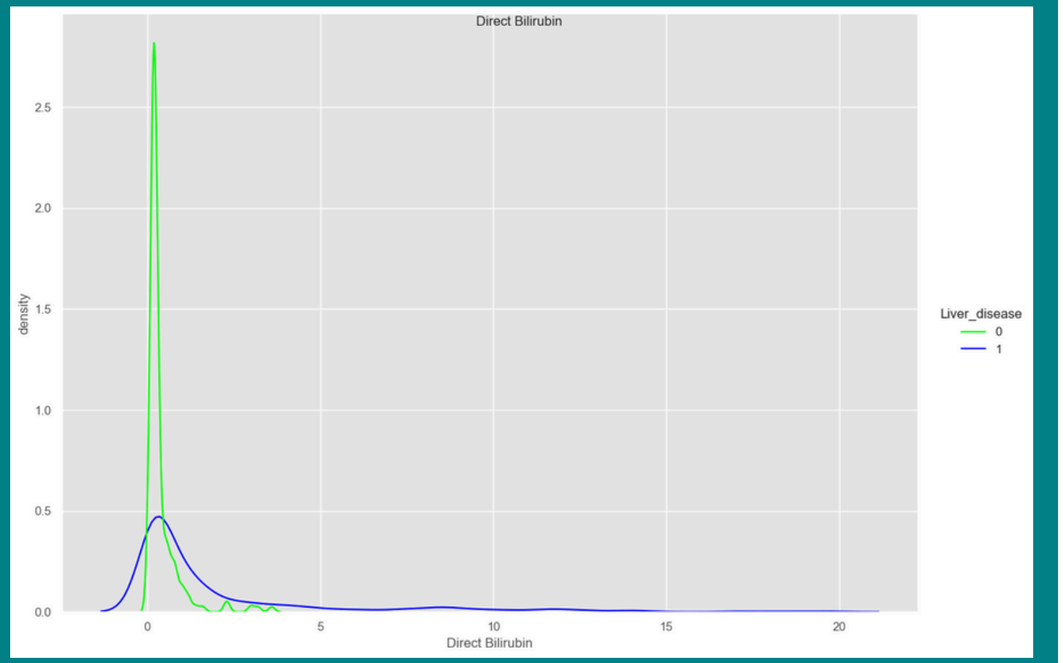


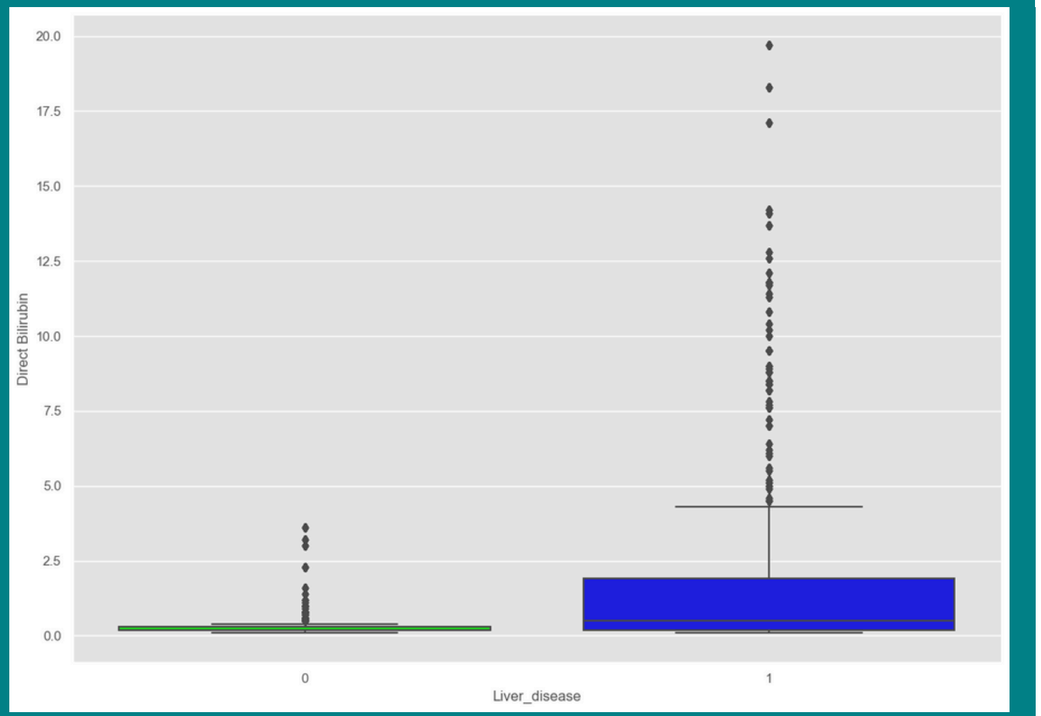
LIVER DISEASE VS TOTAL BILIRUBIN





The Non Liver Disease category has a higher concentration of low bilirubin levels, while the Liver Disease category has a wider range of values, indicating more variability in bilirubin levels among individuals.





DIRECT BILIRUBIN VS LIVER DISEASE

The Non Liver Disease category has a higher concentration of low Direct bilirubin levels, while the Liver Disease category has a wider range of values, indicating more variability in Direct bilirubin levels among individuals.

For the non Liver Disease Group

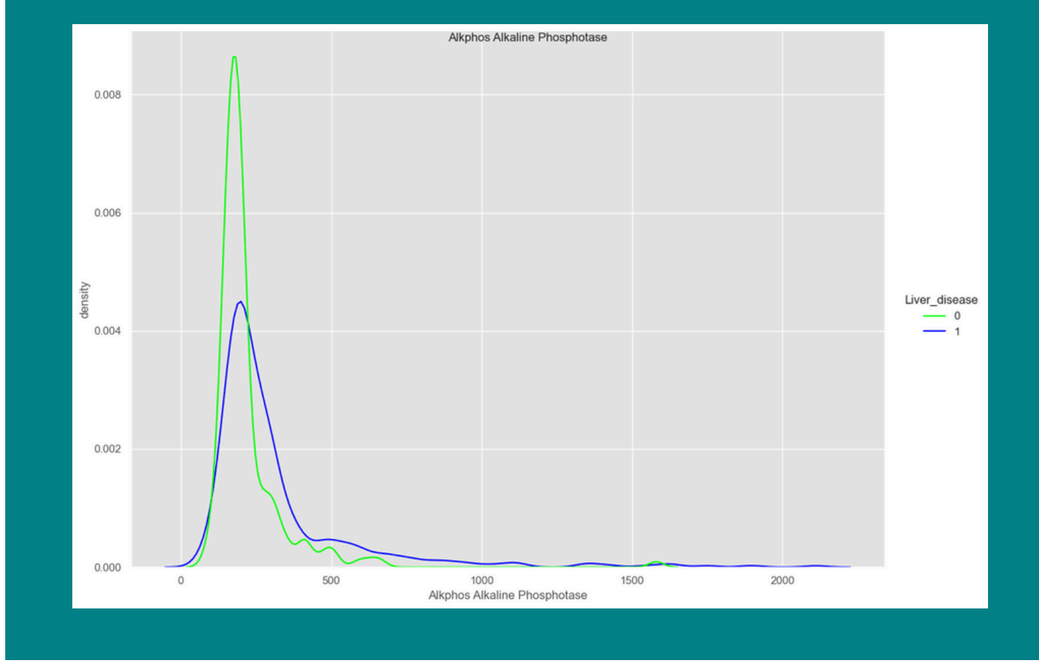
- Indicates a small interquartile range (IQR), suggesting that most individuals have similar Direct Bilirubin levels.
- The median line within the box is lower, indicating that the typical Direct Bilirubin level is relatively low.
- only a few outliers, which means that extreme values are rare in this group.

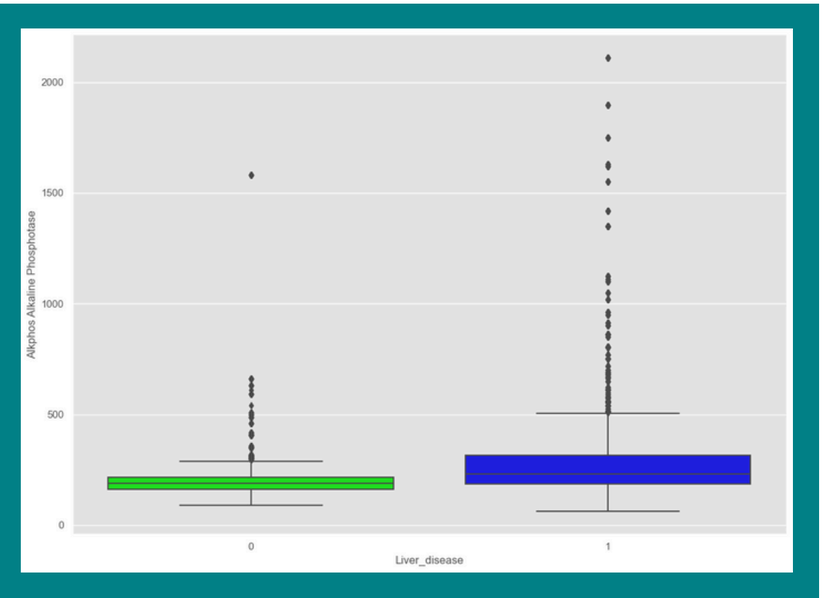
For the Liver Disease Group

- Indicates a large interquartile range (IQR), suggesting that more variability
- The median line is higher, showing that the typical Direct Bilirubin level is elevated compared to non Liver Disease Group.
- A significant number of outlier suggest that high Direct Bilirubin levels are more common in this group

LIVER DISEASE VS ALKPHOS ALKALINE PHOSPHOTASE

The Non Liver Disease category has a higher concentration of low Alkphos Alkaline Phosphotase levels, while the Liver Disease category has a wider range of values, indicating more variability in Alkphos Alkaline Phosphotase levels among individuals



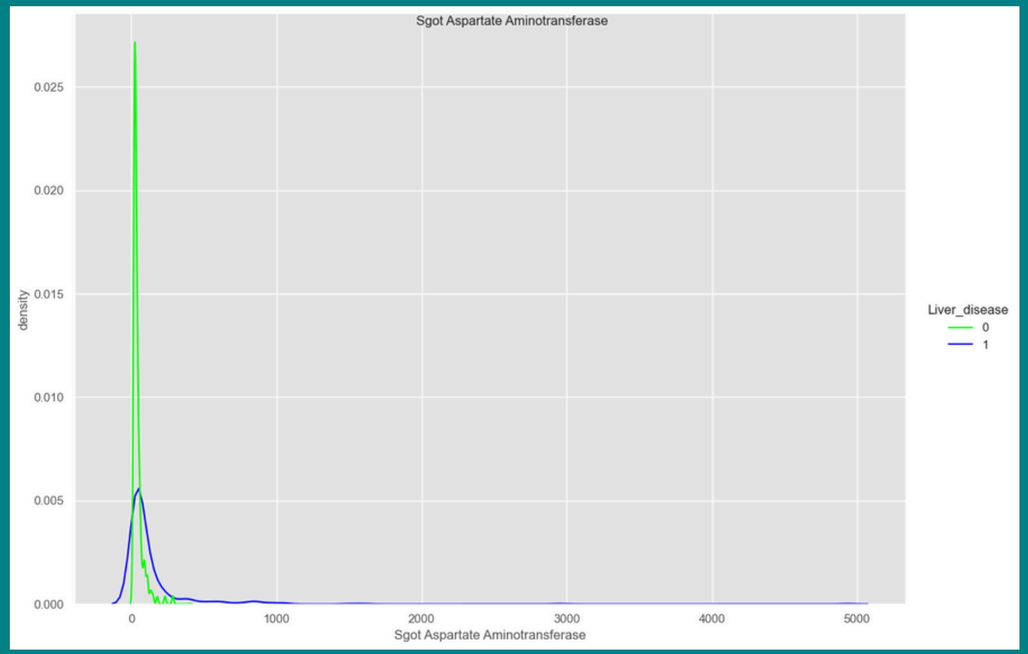


For the non Liver Disease Group Indicates

- a small (IQR), suggesting that most individuals have similar AAP levels.
- The median line within the box is lower, indicating AAP level is relatively low.
- only a few outliers

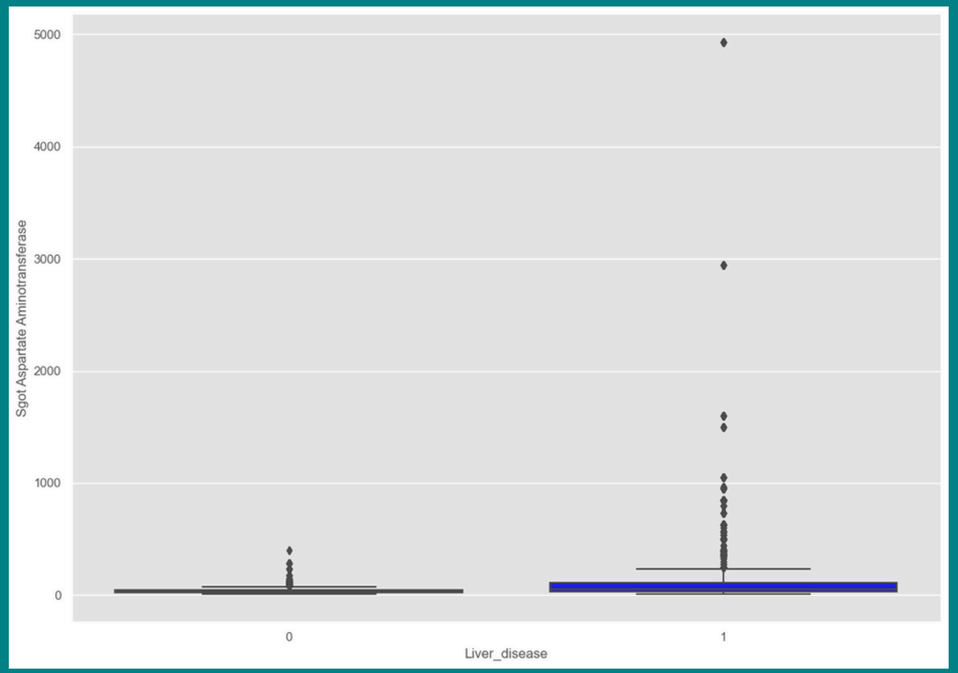
.For the Liver Disease Group

- a large (IQR), suggesting that more variability
- The median line is higher, showing AAP level is elevated compared to non Liver Disease Group.
- A significant number of outlier



SGOT ASPARTATE AMINOTRANSFERASE VS LIVER DISEASE

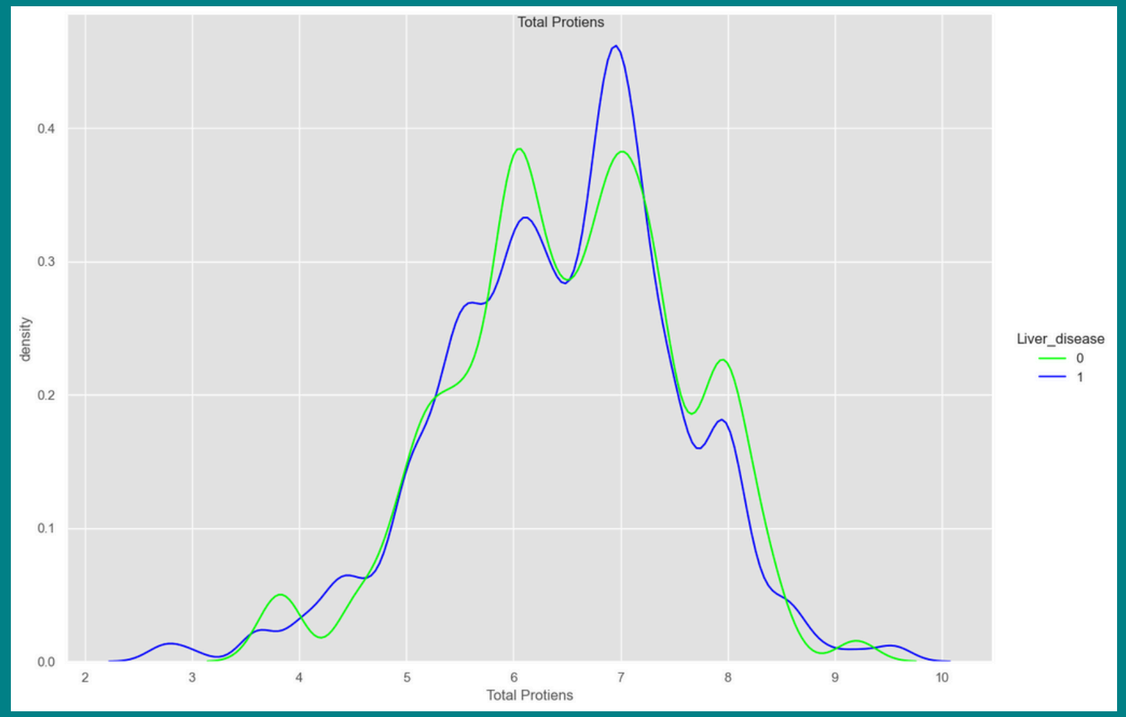
Both category has a higher concentration of low Sgot Aspartate Aminotransferase levels



For both categories

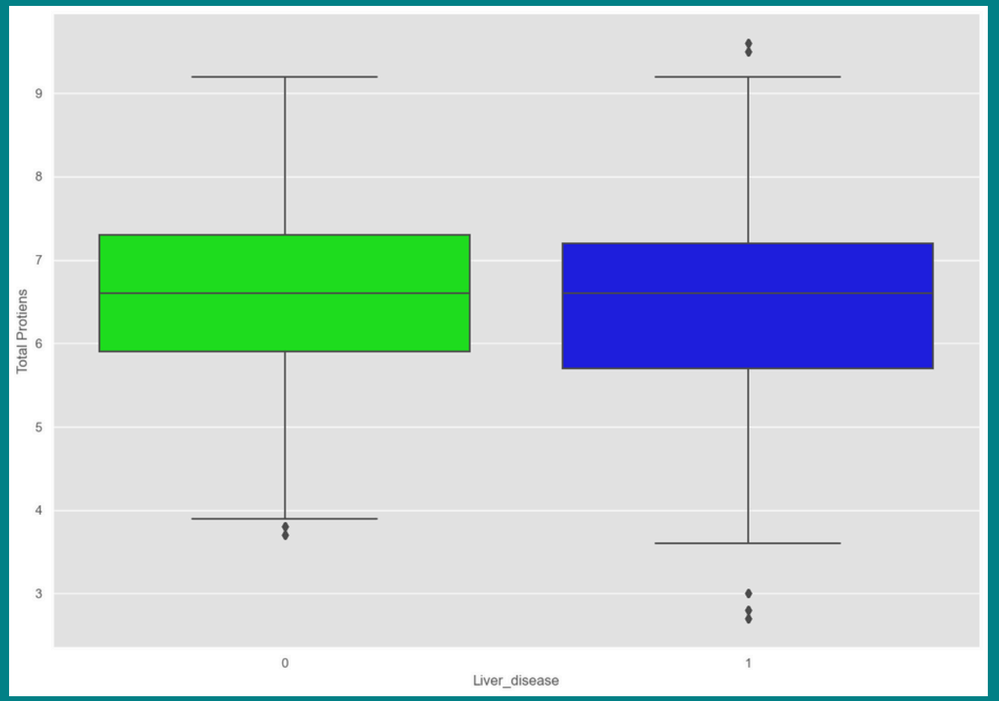
- a small (IQR), suggesting that most individuals have similar SAP levels.
- The median line within the box is lower, indicating SAP level is relatively low.
- only a few outliers for Non liver disease group whereas A significant number of outlier for other category

TOTAL PROTEINS VS LIVER DISEASE



Both plots show the Total Protein range of patients from 0 to approximately 10. The highest density of patients for both conditions is around 7 indicating this is the most common Total Proteins group.

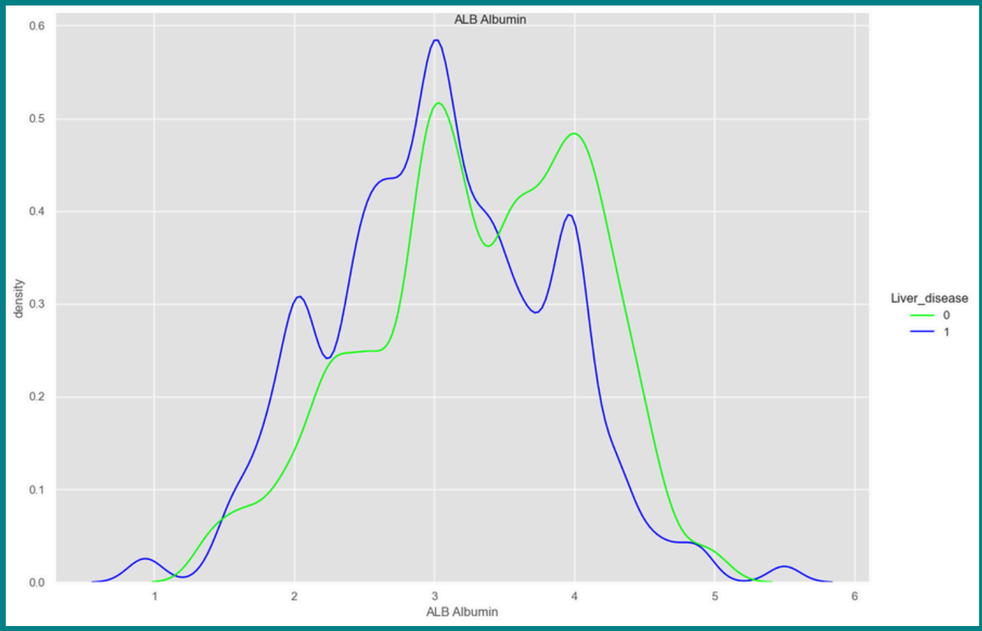
- The length of the whiskers and the size of the boxes show there is a high variability for Total Proteins distribution among the patients who have liver disease
- Median Total Proteins is approximately same for both groups

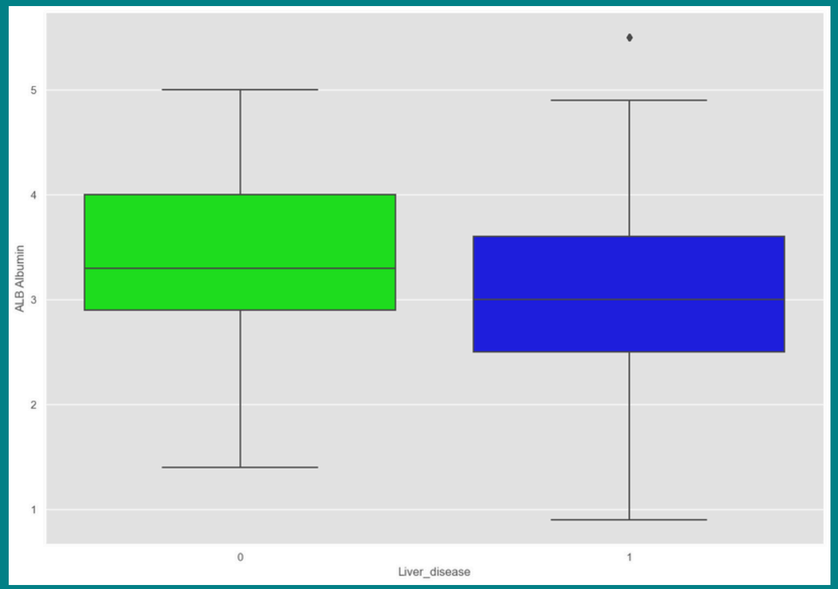


LIVER DISEASE VS ALB ALBUMIN

Both plots show the Albumin range of patients from 0 to approximately 6. The highest density of patients for both conditions is around 3

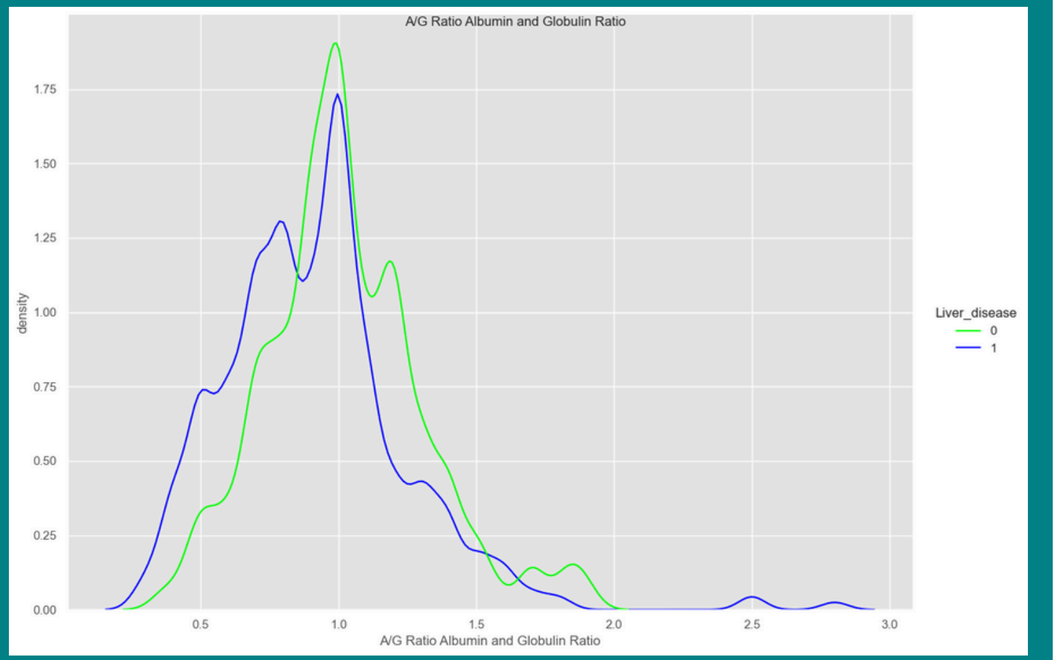
The length of the whiskers and the size of the boxes show approximately same variability for Albumin distribution for both categories





- **Median Albumin of the non liver disease group is grater** than that of liver disease group

- According to a **Medmastery article**, the continuous liver dysfunction often leads to a decrease in albumin synthesis, causing low levels of albumin in the body. which is frequently observed in individuals with chronic liver disease.



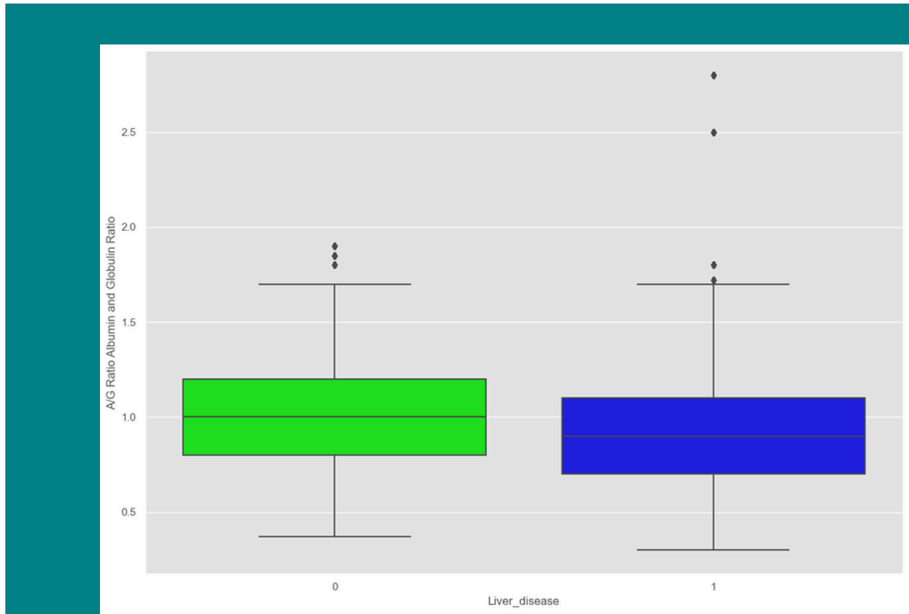
A/G RATIO- ALBUMIN AND GLOBULIN RATIO VS LIVER DISEASE

These plots show the A/G Ratio range of patients from 0 to approximately 3. The highest density of patients for both conditions is around 1

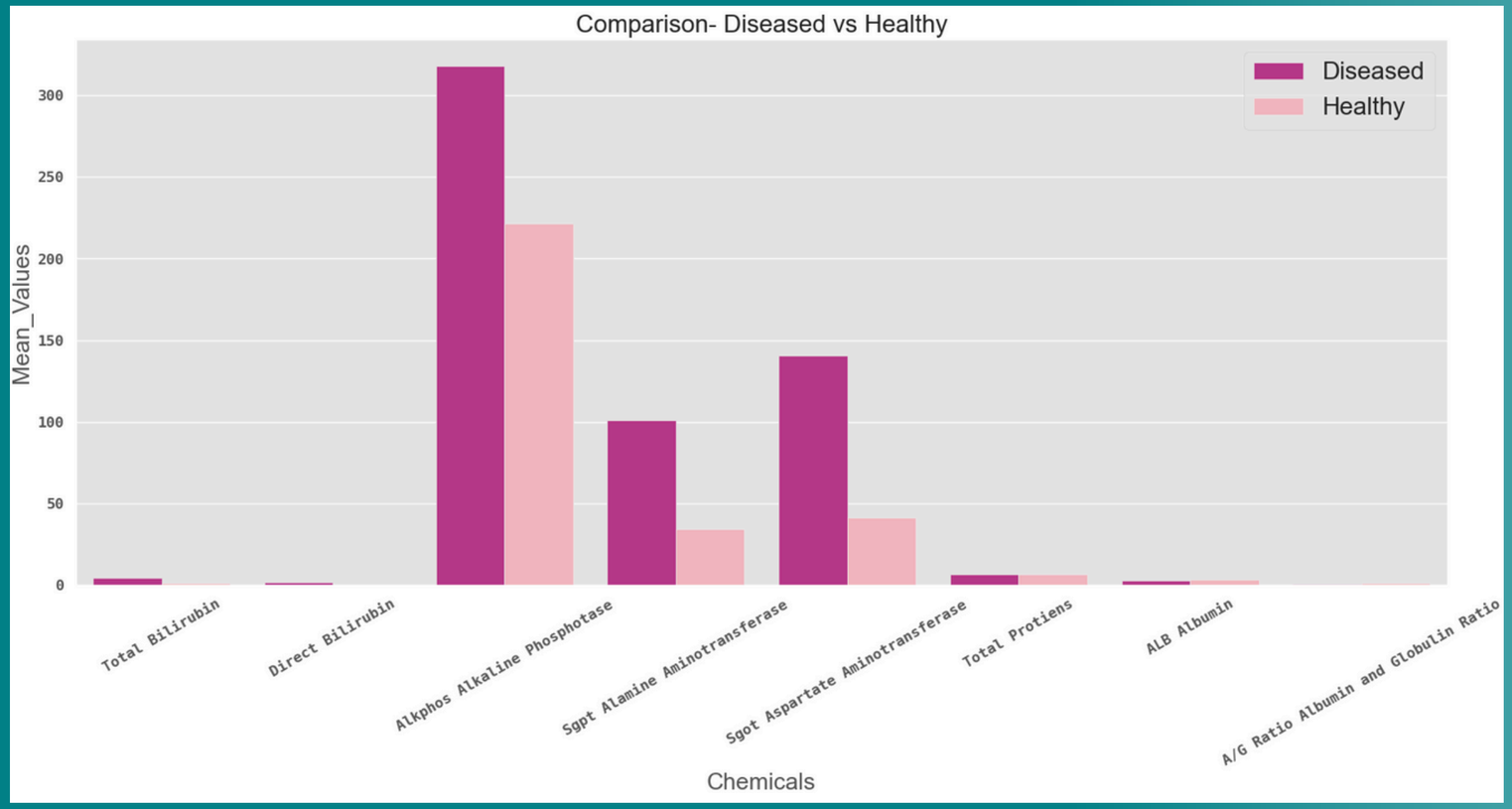
An article on **Healthline** states that the normal range for albumin/globulin ratio is , usually around 1 to 2.That's because there's a bit more albumin than globulin in protein.

- The length of the whiskers and the size of the boxes show approximately same variability for A/G Ratio distribution for both categories

- Median A/G Ratio of the non liver disease group is grater than that of liver disease group



MULTIVARIATE ANALYSIS

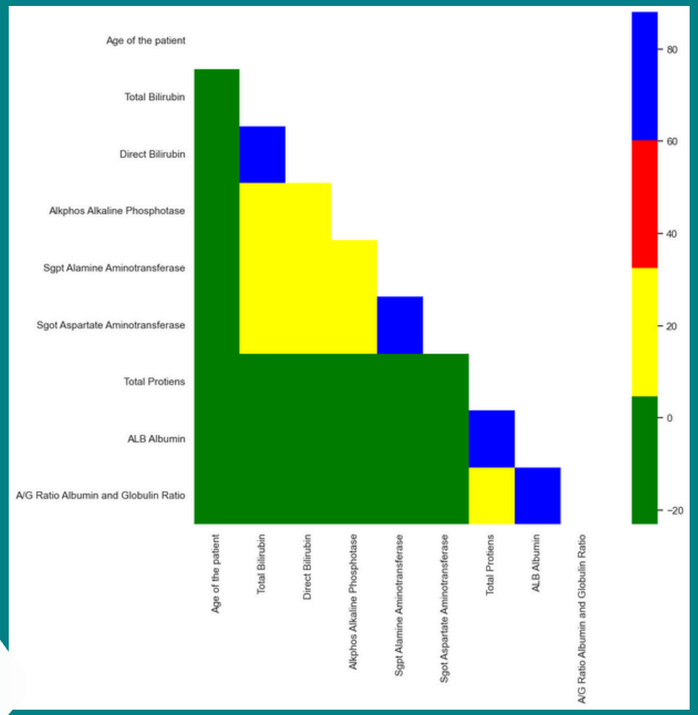


	Variables	Mean_Values	Status
0	Total Bilirubin	4.326103	Diseased
1	Direct Bilirubin	1.994993	Diseased
2	Alkphos Alkaline Phosphotase	318.034573	Diseased
3	Sgpt Alamine Aminotransferase	100.766630	Diseased
4	Sgot Aspartate Aminotransferase	140.717716	Diseased
5	Total Protiens	6.456751	Diseased
6	ALB Albumin	3.047058	Diseased
7	A/G Ratio Albumin and Globulin Ratio	0.911290	Diseased
8	Total Bilirubin	1.136565	Healthy
9	Direct Bilirubin	0.392353	Healthy
10	Alkphos Alkaline Phosphotase	221.156214	Healthy
11	Sgpt Alamine Aminotransferase	34.070575	Healthy
12	Sgot Aspartate Aminotransferase	41.187817	Healthy
13	Total Protiens	6.523268	Healthy
14	ALB Albumin	3.320386	Healthy
15	A/G Ratio Albumin and Globulin Ratio	1.022019	Healthy

- The dataset includes various biochemical variables related to liver function, such as Total Bilirubin, Direct Bilirubin.
- Comparing mean values between diseased and healthy individuals reveals significant differences in various biomarkers
- Levels of Alkphos (Alkaline Phosphotase),Sgpt Alamine Aminotransferase,Sgot Aspartate Aminotransferase are notably higher in diseased individuals compared to healthy ones, indicating potential liver dysfunction or disease.
- This suggests that these biomarkers could be useful indicators for diagnosing liver disease

CORRELATION

	Age of the patient	Total Bilirubin	Direct Bilirubin	Alkphos Alkaline Phosphotase	Sgpt Alamine Aminotransferase	Sgot Aspartate Aminotransferase	Total Protiens	ALB Albumin	A/G Ratio Albumin and Globulin Ratio
Age of the patient	1.000000	0.007648	0.007361	-0.001151	0.001036	0.010131	-0.006876	-0.017934	-0.022789
Total Bilirubin	0.007648	1.000000	0.879826	0.197495	0.207131	0.240529	0.000198	-0.224007	-0.201758
Direct Bilirubin	0.007361	0.879826	1.000000	0.222137	0.225658	0.260510	0.008018	-0.231126	-0.193652
Alkphos Alkaline Phosphotase	-0.001151	0.197495	0.222137	1.000000	0.126906	0.167112	-0.018666	-0.157353	-0.227013
Sgpt Alamine Aminotransferase	0.001036	0.207131	0.225658	0.126906	1.000000	0.780555	-0.046201	-0.032031	-0.005445
Sgot Aspartate Aminotransferase	0.010131	0.240529	0.260510	0.167112	0.780555	1.000000	-0.030471	-0.092181	-0.073979
Total Protiens	-0.006876	0.000198	0.008018	-0.018666	-0.046201	-0.030471	1.000000	0.777007	0.232695
ALB Albumin	-0.017934	-0.224007	-0.231126	-0.157353	-0.032031	-0.092181	0.777007	1.000000	0.683058
A/G Ratio Albumin and Globulin Ratio	-0.022789	-0.201758	-0.193652	-0.227013	-0.005445	-0.073979	0.232695	0.683058	1.000000



Both heatmaps and correlation values indicates there is high correlation between **Total Bilirubun & Direct Billirubin (0.879826)**

Sgpt Alamine Aminotransferase & Sgot Aspartate Aminotrsnasferase (0.780555)

ALB Albumin & Total Protiens (0.777007)

A/G ration Albumin and Globiumin Ration & ALB Albiumin (0.683058)

CORRELATION

	Variable	VIF
0	Age of the patient	7.611217
1	Total Bilirubin	5.705971
2	Direct Bilirubin	5.956732
3	Alkphos Alkaline Phosphotase	2.637626
4	Sgpt Alamine Aminotransferase	3.152047
5	Sgot Aspartate Aminotransferase	3.088881
6	Total Protiens	81.210175
7	ALB Albumin	111.068235
8	A/G Ratio Albumin and Globulin Ratio	22.321848

'Total Protiens', 'ALB Albumin', and 'A/G Ratio Albumin and Globulin Ratio' have very high VIF values, indicating high multicollinearity among these variables.

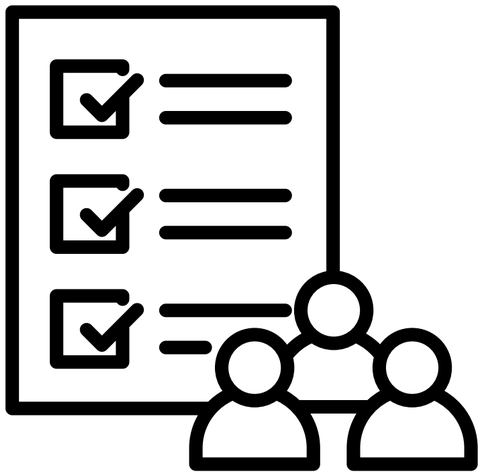
EVALUATION METRICS

The accuracy score measures the proportion of correct predictions out of the total number of predictions made by the model

Misclassification Rate representing the proportion of incorrect predictions out of the total predictions.

Recall: Proportion of actual positives which were correctly classified

F1 score A combination of precision and recall



1

```
Accuracy score of this model:  71.7 %
Misclassification rate of this model:  28.3 %

Report card of this model:
              precision    recall  f1-score   support

           0      0.479     0.075     0.129      1556
           1      0.728     0.968     0.831      3978

    accuracy                          0.717      5534
   macro avg      0.604     0.521     0.480      5534
weighted avg      0.658     0.717     0.634      5534

Confusion Matrix:
+-----------------+----------------------+----------------------+
|                 |   Predicted Negative |   Predicted Positive |
+=================+======================+======================+
| Actual Negative |                  116 |                 1440 |
+-----------------+----------------------+----------------------+
| Actual Positive |                  126 |                 3852 |
+-----------------+----------------------+----------------------+
```

2

3

4

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

CLUSTER ANALYSIS

- **Understanding Different Patients subgroups**: Liver disease affects people differently, with various causes and severity levels. Clustering helps us group patients based on these differences, so we can create better prediction models for each group.
- **Predicting Treatment Response:** Clustering helps predict how patients in each group will respond to different treatments. This helps doctors choose the best treatment for each patient, improving their chances of getting better.
- **Early Detection of High-Risk Patients**:clustering helps us find patients who are at high risk of liver disease early on. This allows us to keep a closer eye on them and take action to help them stay healthy.
- **Personalized Medicine**:enhances the accuracy and effectiveness of liver disease prediction and management strategies.
- **Improving Model Performance**: Clustering-based approaches can enhance the performance of liver disease prediction models

CLUSTERING APPROACHES

KMeans

We can use KMeans for mix type data as well by changing the distance mesures.We have used Hamming distance for catergorical variables and Euclidian distance for the numerical variables

KPrototype

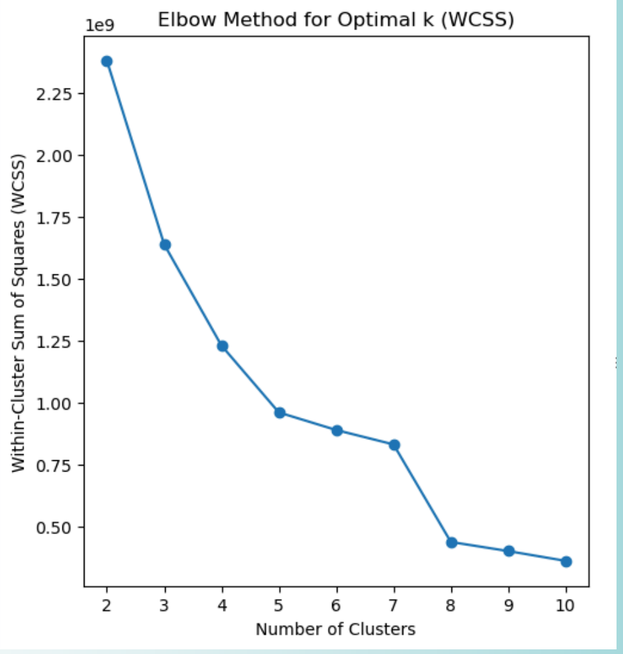
Since our data consist of both categorical and quantitative (mixed type) data

Kmedoids

We can use kmedoids for mix type data as well by changing the distance mesures.We have used Hamming distance for catergorical variables and Euclidian distance for the numerical variables

CLUSTERING APPROACHES- KPROTOTYPE

4 is the elbow point, implying that there is a noticeable decrease in WCSS up to k=4, after which the rate of decrease slows down,and also when K=4 it provides considerable high Silhoutte Scores suggesting that number of Clusters equal to 4

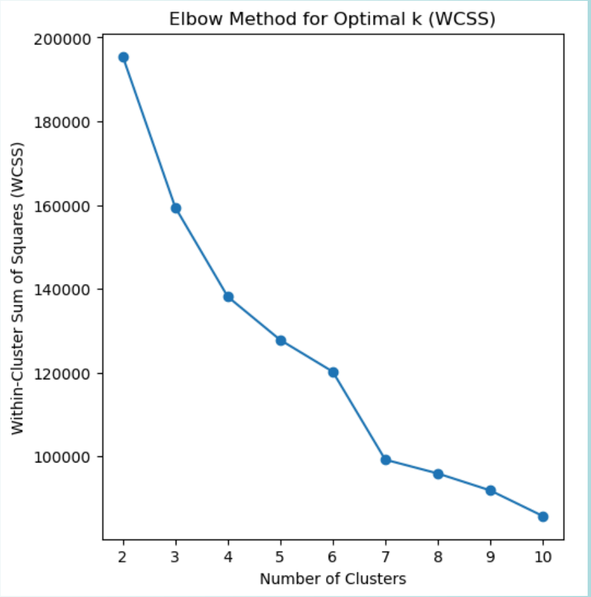


A Silhouette Score of 0.70 indicates a relatively good clustering result,suggests well-separated and that each data point is close to its own cluster's centroid . Some points in negative side indicate that data points may have been assigned to the wrong clusters.

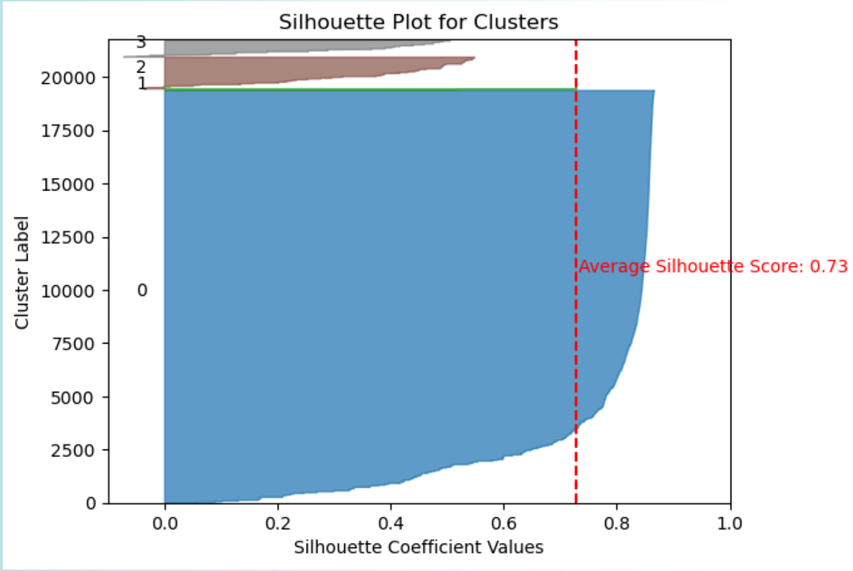
Kprotype Clustering method suggest there are clusters in the data set but when we considers the clusters ; Two classes(Liver Disease and Non liver Disease) are highly disproportionate suggesting inappropriateness of fitting models within clusters

Kprototypes			
Silhoutte Scores	0.7		
cluster	with Liver Disease	without Liver Disease	Total Count
0	0.6805	0.31948	18890
1	0.9472	0.0527	1366
2	1	0	1211
3	1	0	259

CLUSTERING APPROACHES- KMEANS



1st plot suggest 4 as the elbow point, it implies that there is a noticeable decrease in WCSS up to k=4, after which the rate of decrease slows down,and plot also when K=4 it provides considerable high Silhoutte Scores suggesting Number of Clusters equal to 4

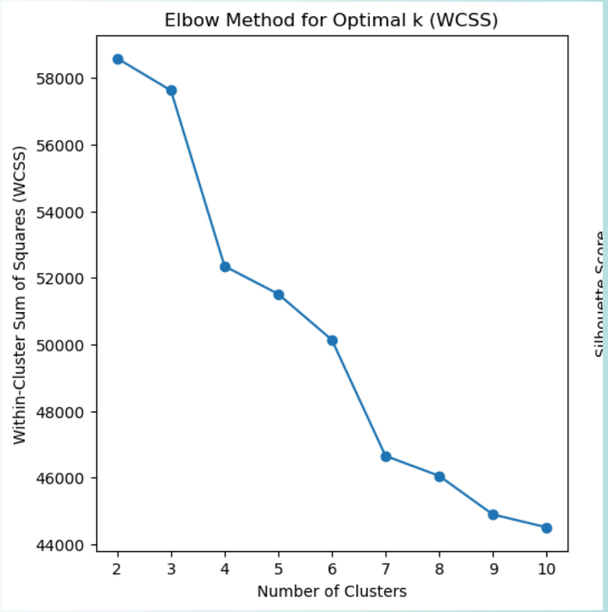


A Silhouette Score of 0.73 indicates a relatively good clustering result,suggests well-separated and that each data point is close to its own cluster's centroid .
Some points in negative side indicate that data points may have been assigned to the wrong clusters.

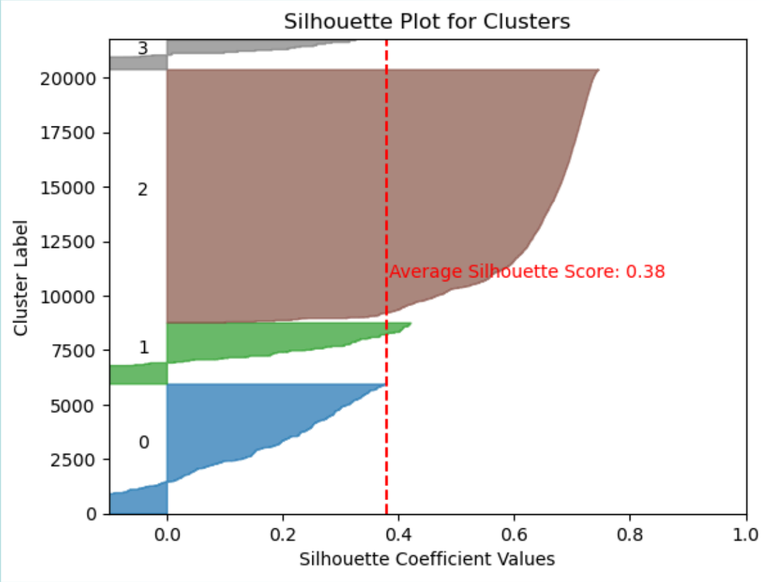
KMeans Clustering method suggest there are clusters in the data set but when we considers the clusters ; Two classes(Liver Disease and Non liver Disease) are highly disproportionate suggesting inappropriateness of fitting models within clusters

Kmeans			
Silhoutte Scores	0.73		
cluster	with Liver Disease	without Liver Disease	Total Count
0	0.6805	0.31948	18890
1	0.9472	0.0527	1366
2	1	0	1211
3	1	0	259

CLUSTERING APPROACHES- KMEDOIDS



1st plot suggest 4 as the elbow point, it implies that there is a noticeable decrease in WCSS up to k=4, after which the rate of decrease slows down,and also when K=4 it provides considerable high Silhoutte Scores suggesting Number of Clusters equal to 4



A Silhouette Score of 0.38 suggests a moderate clustering result, indicating some level of separation between clusters. While it implies that data points within each cluster are closer to each other than to points in other clusters, the degree of separation is not particularly strong. Additionally, Some points in negative side indicateindicate potential misassignments of data points to clusters

In KMedoids Clustering method ; Two classes(Liver Disease and Non liver Disease) are highly disproportionate suggesting inappropriateness of fitting models within clusters.

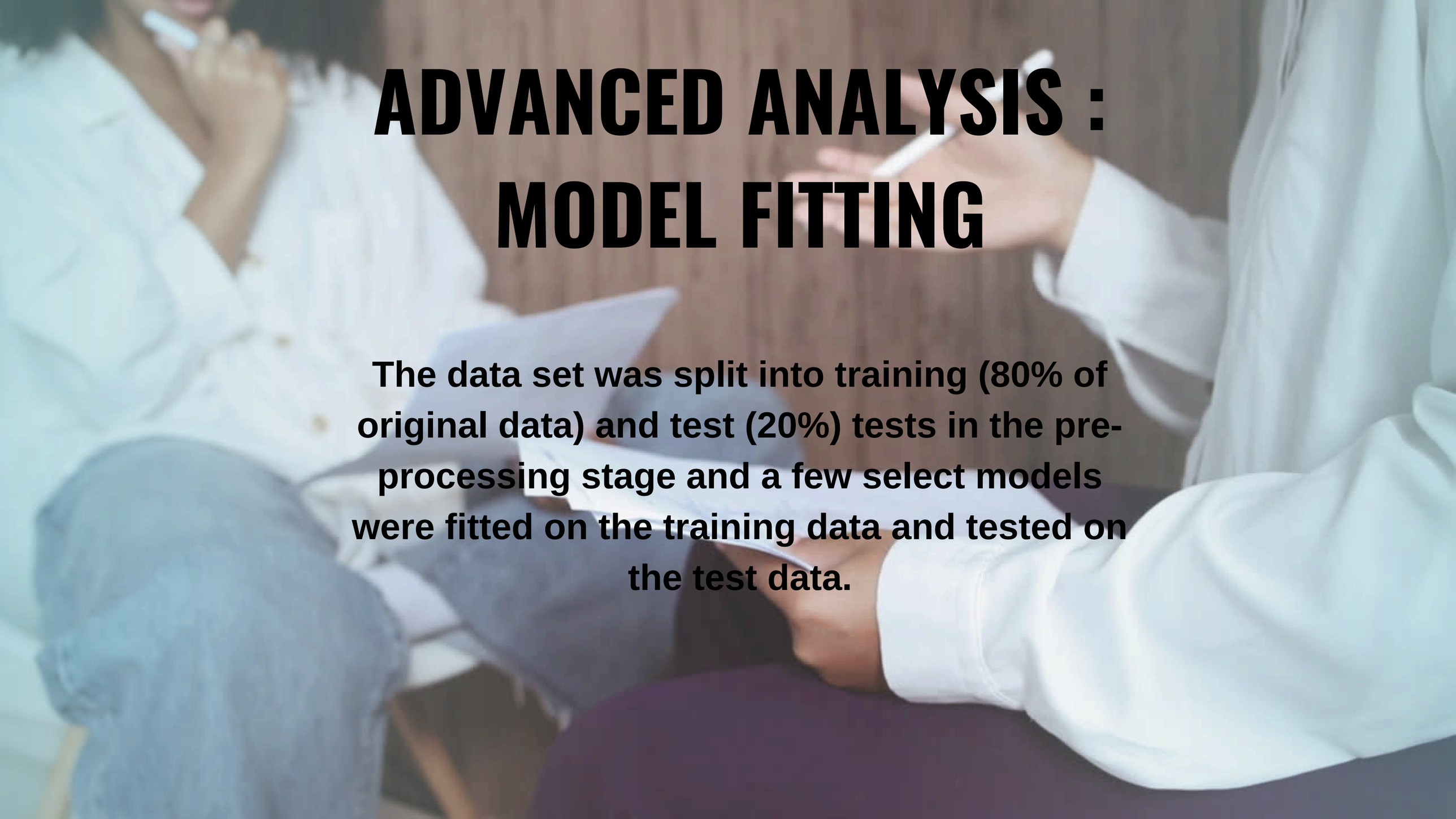
Kmedoids			
Silhoutte Scores	0.37		
cluster	with Liver Disease	without Liver Disease	Total Count
0	0.82919046	0.17080954	5954
1	0.893214286	0.106785714	2800
2	0.587267402	0.412732598	11608
3	1	0	1364

SUMMARY FOR CLUSTERING

- Each 3 types of Clustering Methods suggest the presence of 4 Clusters in the Data Set.

- Withing the clusters the 2 classes(Liver Disease and Non liver Disease) are highly disproportionate suggesting inappropriateness of fitting models withing the clusters as it may leads to biased or unreliable results

- Since our best model achieved 100% accuracy on the test set and 99.99% accuracy on the train set, there's no need to delve further into clustering with the aim of enhancing the model's predictive accuracy.



ADVANCED ANALYSIS : MODEL FITTING

The data set was split into training (80% of original data) and test (20%) tests in the pre-processing stage and a few select models were fitted on the training data and tested on the test data.

Considering the **ORIGINAL DATA** set After the preprocessing stage we have fitted Several Models

This suggests the best model is **RANDOM FOREST** with

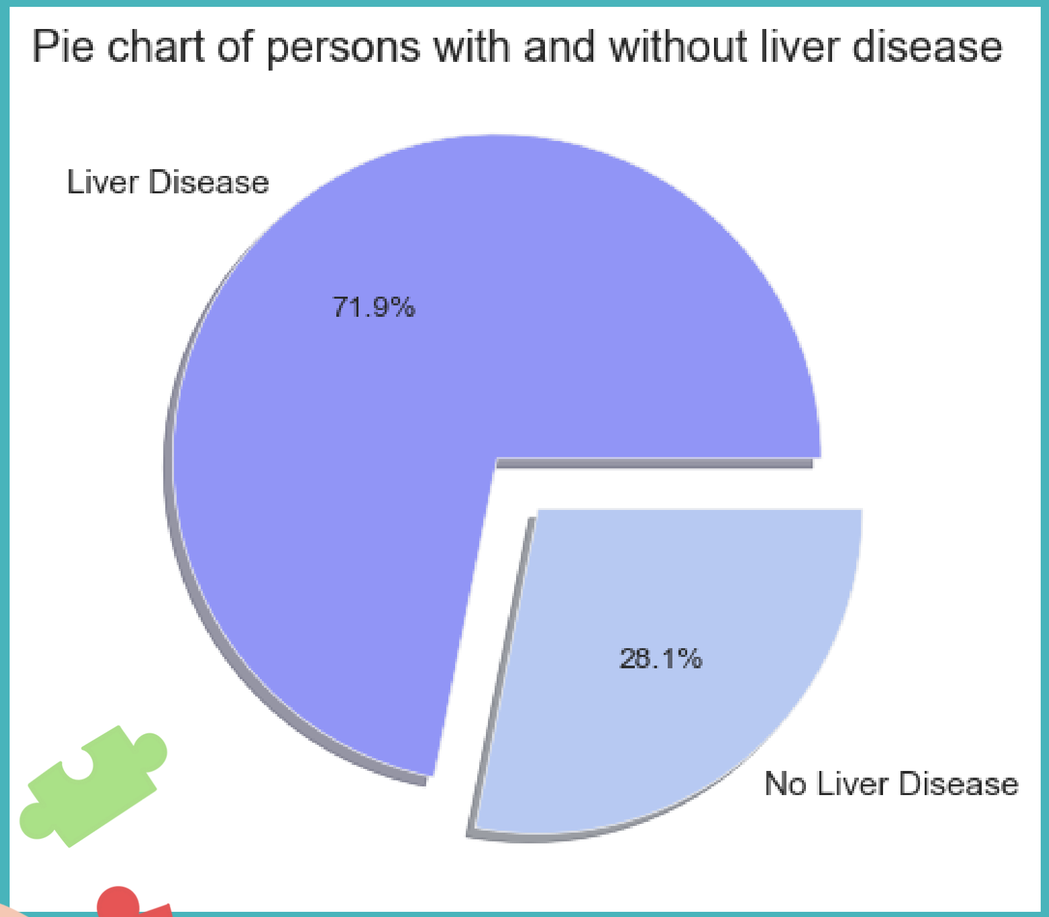
Accuracy (99.98%) demonstrates correctly predicting the outcome for almost all instances .

F1-score (100%):suggests that the model achieves a balance between precision (correctly identifying positive cases) and recall (capturing all positive cases) for both classes.

Recall (100%):, indicating that the model effectively captures all positive cases without missing any.

Models Fitted for the original training Set							
Model	Accuracy	f1-score	recall	Confusion Matix			
Logistic Regression	0.7243	0.558	0.564			Actual	
						Predicted Negative	Predicted Positive
				Predicted	Actual Negative	308	1248
					Actual Positive	278	3700
Support Vector Machine	0.7188	0.418	0.5			Actual	
						Predicted Negative	Predicted Positive
				Predicted	Actual Negative	0	1556
					Actual Positive	0	3978
Decision Tree	0.9996	1	1			Actual	
						Predicted Negative	Predicted Positive
				Predicted	Actual Negative	1555	1
					Actual Positive	1	3977
Random Forest	0.9998	1	1			Actual	
						Predicted Negative	Predicted Positive
				Predicted	Actual Negative	1555	1
					Actual Positive	0	3978
Xgboost	0.9995	0.9999	0.9999			Actual	
						Predicted Negative	Predicted Positive
				Predicted	Actual Negative	1554	2
					Actual Positive	1	3977
KNN	0.9736	0.968	0.972				
						Predicted Negative	Predicted Positive
				Predicted	Actual Negative	1508	48
					Actual Positive	98	3880
Adboost	0.8536	0.806	0.788				
						Predicted Negative	Predicted Positive
				Predicted	Actual Negative	994	562
					Actual Positive	248	3730
Gradient boost	0.9995	0.9999	0.9999				
						Predicted Negative	Predicted Positive
				Predicted	Actual Negative	15554	2
					Actual Positive	1	1977
MLP (multi-layer perceptron	0.9561	0.946	0.945				
						Predicted Negative	Predicted Positive
				Predicted	Actual Negative	1431	125
					Actual Positive	118	3860

LIVER DISEASE



- We have observed that a substantial class imbalance within our dataset indicating the prevalence of liver disease within the studied population, accounting for approximately 71.9%.
- We have applied
  - SMOTE
  - Upsampling
  - Downsampling

Considering the **SMOTE DATA** set After the preprocessing stage we have fitted Several Models

This suggests the best model is **XGboost**

Models Fitted for the SMOTE training Set							
Model	Accuracy	f1-score	recall	Confusion Matix			
Logistic Regression	0.6409	0.629	0.697			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1283	273
					Actual Positive	1714	2664
Support Vector Machine	0.6243	0.62	0.713			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1423	113
					Actual Positive	1946	2032
Decision Tree	0.9989	0.999	0.999			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1554	2
					Actual Positive	4	3974
Random Forest	0.9996	1	0.999			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1554	2
					Actual Positive	0	3978
Xgboost	0.9998	1	1			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1555	1
					Actual Positive	0	3978
KNN	0.9817	0.978	0.985			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1544	12
					Actual Positive	89	3889
Adboost	0.8211	0.804	0.858			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1464	92
					Actual Positive	898	3080
Gradient boost	0.9995	0.999	0.999			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1554	2
					Actual Positive	1	3977
MLP (multi-layer perceptror	0.9827	0.979	0.988			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1554	2
					Actual Positive	94	3884

Considering the **UPSAMPLED DATA** set After the preprocessing stage we have fitted Several Models

None of the models were suitable under Upsampling method Compared to other methods

Models Fitted for the UPSAMPLED training Set							
Model	Accuracy	f1-score	recall	Confusion Matix			
Logistic Regression	0.3554	0.302	0.29			Actual	
						Predicted Negativ	Predicted Positive
				Predictec	Actual Negative	219	1337
					Actual Positive	2230	1748
Support Vector Machine	0.3849	0.304	0.292			Actual	
						Predicted Negativ	Predicted Positive
				Predictec	Actual Negative	124	1432
					Actual Positive	1972	2006
Decision Tree	0.9156	0	0			Actual	
						Predicted Negativ	Predicted Positive
				Predictec	Actual Negative	1304	252
					Actual Positive	215	3763
Random Forest	0.04	0.001	0.001			Actual	
						Predicted Negativ	Predicted Positive
				Predictec	Actual Negative	1	1555
					Actual Positive	3977	1
Xgboost	0.05	0.001	0.001			Actual	
						Predicted Negativ	Predicted Positive
				Predictec	Actual Negative	2	1554
					Actual Positive	3977	1
KNN	0.0262	0.026	0.023				
						Predicted Negativ	Predicted Positive
				Predictec	Actual Negative	24	1532
					Actual Positive	3857	121
Adboost	0.1791	0.163	0.147				
						Predicted Negativ	Predicted Positive
				Predictec	Actual Negative	117	1439
					Actual Positive	3104	874
Gradient boost	0.04	0	0				
						Predicted Negativ	Predicted Positive
				Predictec	Actual Negative	1	1555
					Actual Positive	3977	1
MLP (multi-layer perceptror	0.0183	0.018	0.014				
						Predicted Negativ	Predicted Positive
				Predictec	Actual Negative	9	1547
					Actual Positive	3886	92

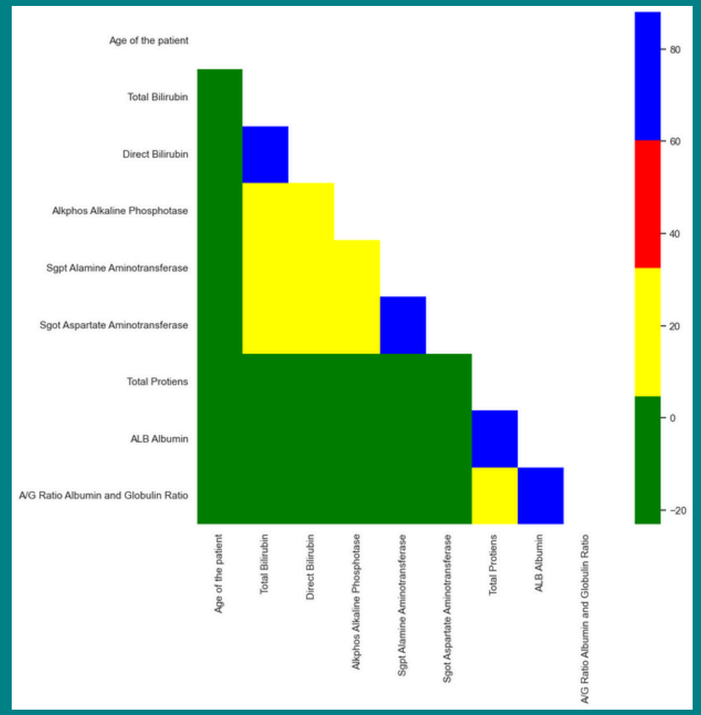
Considering the **DOWNSAMPLE DATA** set After the preprocessing stage we have fitted Several Models

This suggests the best model is **RANDOM FOREST**

Models Fitted for the Doensampled training Set							
Model	Accuracy	f1-score	recall	Confusion Matix			
Logistic Regression	0.6449	0.635	0.71			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1335	221
					Actual Positive	1744	2234
Support Vector Machine	0.6122	0.608	0.703			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1419	137
					Actual Positive	2009	1969
Decision Tree	0.9989	0.999	0.999			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1555	1
					Actual Positive	3	3973
Random Forest	0.9998	1	1			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1555	1
					Actual Positive	0	3978
Xgboost	0.9996	1	0.999			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1555	1
					Actual Positive	1	3977
KNN	0.9275	0.916	0.948			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1547	9
					Actual Positive	392	3586
Adboost	0.8339	0.816	0.863			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1444	112
					Actual Positive	807	3171
Gradient boost	0.9993	0.999	0.999			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1554	2
					Actual Positive	2	3976
MLP (multi-layer perceptror	0.9389	0.927	0.946			Actual	
						Predicted Negativ	Predicted Positive
				Predicted	Actual Negative	1495	61
					Actual Positive	277	3701

CORRELATION

	Age of the patient	Total Bilirubin	Direct Bilirubin	Alkphos Alkaline Phosphotase	Sgpt Alamine Aminotransferase	Sgot Aspartate Aminotransferase	Total Protiens	ALB Albumin	A/G Ratio Albumin and Globulin Ratio
Age of the patient	1.000000	0.007648	0.007361	-0.001151	0.001036	0.010131	-0.006876	-0.017934	-0.022789
Total Bilirubin	0.007648	1.000000	0.879826	0.197495	0.207131	0.240529	0.000198	-0.224007	-0.201758
Direct Bilirubin	0.007361	0.879826	1.000000	0.222137	0.225658	0.260510	0.008018	-0.231126	-0.193652
Alkphos Alkaline Phosphotase	-0.001151	0.197495	0.222137	1.000000	0.126906	0.167112	-0.018666	-0.157353	-0.227013
Sgpt Alamine Aminotransferase	0.001036	0.207131	0.225658	0.126906	1.000000	0.780555	-0.046201	-0.032031	-0.005445
Sgot Aspartate Aminotransferase	0.010131	0.240529	0.260510	0.167112	0.780555	1.000000	-0.030471	-0.092181	-0.073979
Total Protiens	-0.006876	0.000198	0.008018	-0.018666	-0.046201	-0.030471	1.000000	0.777007	0.232695
ALB Albumin	-0.017934	-0.224007	-0.231126	-0.157353	-0.032031	-0.092181	0.777007	1.000000	0.683058
A/G Ratio Albumin and Globulin Ratio	-0.022789	-0.201758	-0.193652	-0.227013	-0.005445	-0.073979	0.232695	0.683058	1.000000

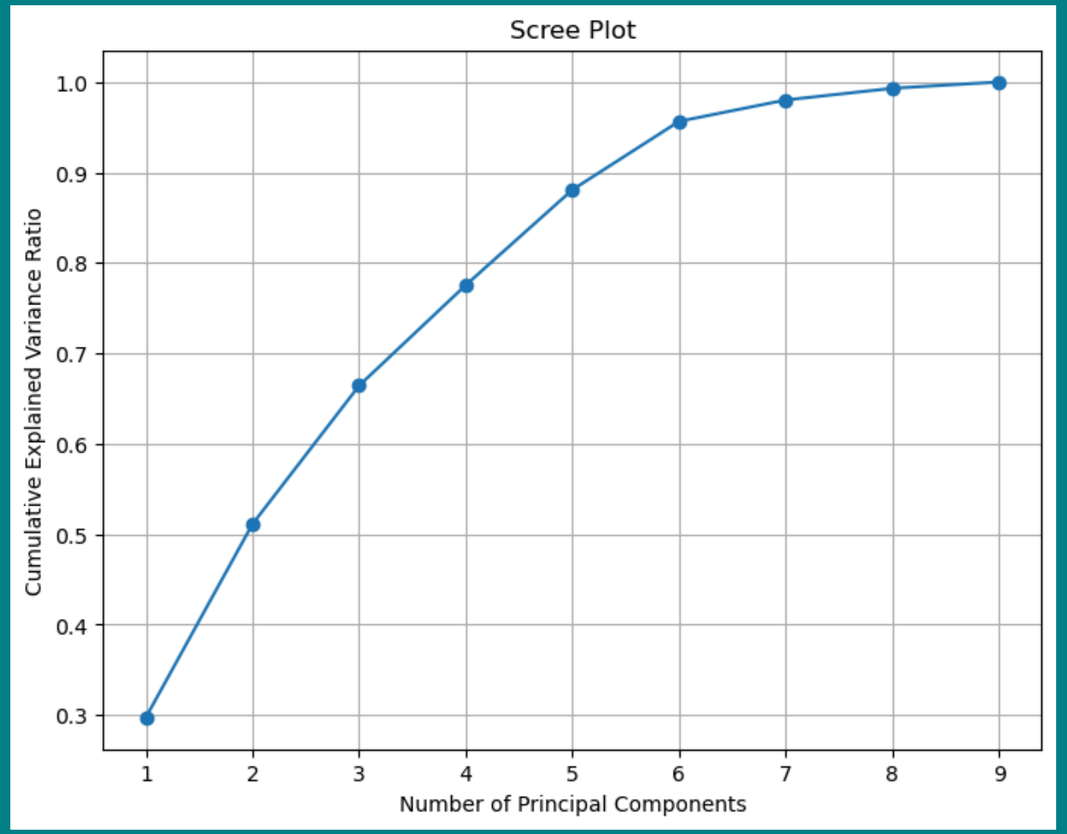


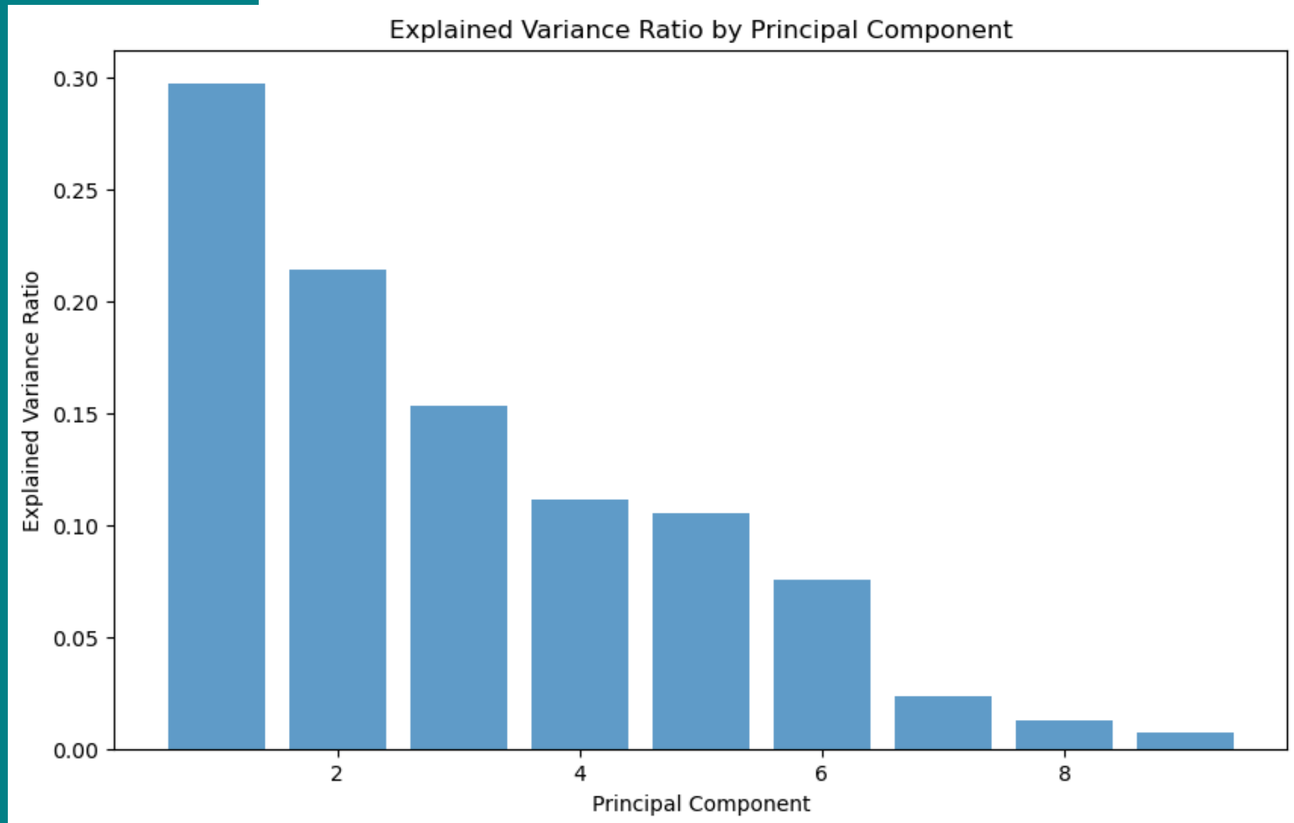
We have observed Multicolinearity exist in our Data set. So as a remedy we have applied PCA and fitted models and compared the accuracy of the models.

PCA

Each bar in the plot represents the proportion of the total variance explained by a single principal component

Indicating the elbow point as 5 PCs' where the explained variance ratio begins to level off or the rate of decrease slows down significantly





$$\text{Explained Variance Ratio}_i = \frac{\text{Variance explained by PC}_i}{\text{Total variance in the dataset}}$$

Indicating the elbow point as 5 PC's ,the point where the cumulative proportion reaches a 80% level at 5 PC's

Number of Principal Components = 5

RESULTS

After applying PCA number of variables have reduced to 6 and These are the values we have obtained under Original training set

This suggests the best model is Random Forest with 0.9812 accuracy but the accuracy is less compared to the previously obtained models

Models Fitted for the ORIGINAL training Set with PCA							
Model	Accuracy	f1-score	recall	Confusion Matix			
Logistic Regression	0.717	0.48	0.521			Actual	
						Predicted Negative	Predicted Positive
				Predictec	Actual Negative	131	1425
					Actual Positive	1867	2111
Support Vector Machine	0.7188	0.418	0.5			Actual	
						Predicted Negative	Predicted Positive
				Predictec	Actual Negative	60	1496
					Actual Positive	1636	2342
Decision Tree	0.9156	0.895	0.892			Actual	
						Predicted Negative	Predicted Positive
				Predictec	Actual Negative	768	788
					Actual Positive	3346	623
Random Forest	0.9812	0.977	0.974			Actual	
						Predicted Negative	Predicted Positive
				Predictec	Actual Negative	602	954
					Actual Positive	3601	377
Xgboost	0.9684	0.961	0.963			Actual	
						Predicted Negative	Predicted Positive
				Predictec	Actual Negative	592	964
					Actual Positive	3402	576
KNN	0.889	0.868	0.888			Actual	
						Predicted Negative	Predicted Positive
				Predictec	Actual Negative	317	1239
					Actual Positive	3248	730
Adboost	0.792	0.693	0.711			Actual	
						Predicted Negative	Predicted Positive
				Predictec	Actual Negative	185	1371
					Actual Positive	2343	1635
Gradient boost	0.9574	0.947	0.945			Actual	
						Predicted Negative	Predicted Positive
				Predictec	Actual Negative	580	976
					Actual Positive	3379	599
MLP (multi-layer perceptror	0.8285	0.886	762			Actual	
						Predicted Negative	Predicted Positive
				Predictec	Actual Negative	257	1299
					Actual Positive	2810	1168

RESULTS

After applying PCA number of variables have reduced to 6 and These are the values we have obtained under SMOTE training set

These models are not accurate as the models we have obtained under without Applying PCA

Model fitted for SMOTE training Set with PCA							
Model	Accuracy	f1-score	recall	Confusion Matix			
Logistic Regression	0.6032	0.6	0.7			Actual	
						Predicted Negative	Predicted Positive
				Predictec	Actual Negative	1432	124
					Actual Positive	2072	1906
Support Vector Machine	0.5654	0.682	0.565			Actual	
						Predicted Negative	Predicted Positive
				Predictec	Actual Negative	1474	82
					Actual Positive	2323	1655
Decision Tree	0.7248	0.681	0.696			Actual	
						Predicted Negative	Predicted Positive
				Predictec	Actual Negative	979	577
					Actual Positive	946	3032
Random Forest	0.7913	0.753	0.767			Actual	
						Predicted Negative	Predicted Positive
				Predictec	Actual Negative	1105	452
					Actual Positive	704	3274
Xgboost	0.7589	0.719	0.736			Actual	
						Predicted Negative	Predicted Positive
				Predictec	Actual Negative	1062	494
					Actual Positive	840	3138
KNN	0.0805	0.781	0.816			Actual	
						Predicted Negative	Predicted Positive
				Predictec	Actual Negative	1309	247
					Actual Positive	832	3146
Adboost	0.6682	0.657	0.727			Actual	
						Predicted Negative	Predicted Positive
				Predictec	Actual Negative	1339	217
					Actual Positive	1619	2359
Gradient boost	0.7472	0.705	0.72			Actual	
						Predicted Negative	Predicted Positive
				Predictec	Actual Negative	1025	531
					Actual Positive	868	3110
MLP (multi-layer perceptror	0.7476	0.724	0.766			Actual	
						Predicted Negative	Predicted Positive
				Predictec	Actual Negative	1255	301
					Actual Positive	1096	2882

RESULTS

After applying PCA number of variables have reduced to 6 and These are the values we have obtained under Upsampled training set

These models are not accurate as the models we have obtained under without Applying PCA

Models Fitted for the UPSAMPLED training Set with PCA							
Model	Accuracy	f1-score	recall	Confusion Matix			
Logistic Regression	0.4051	0.318	0.307			Actual	
						Predicted Negativ	Predicted Positive
				Predictec	Actual Negative	131	1425
					Actual Positive	1867	2111
Support Vector Machine	0.434	0.318	0.314			Actual	
						Predicted Negativ	Predicted Positive
				Predictec	Actual Negative	60	1496
					Actual Positive	1636	2342
Decision Tree	0.253	0.253	0.326			Actual	
						Predicted Negativ	Predicted Positive
				Predictec	Actual Negative	768	788
					Actual Positive	3346	623
Random Forest	0.1769	0.176	0.241			Actual	
						Predicted Negativ	Predicted Positive
				Predictec	Actual Negative	602	954
					Actual Positive	3601	377
Xgboost	0.2111	0.211	0.263			Actual	
						Predicted Negativ	Predicted Positive
				Predictec	Actual Negative	592	964
					Actual Positive	3402	576
KNN	0.1892	0.185	0.194			Actual	
						Predicted Negativ	Predicted Positive
				Predictec	Actual Negative	317	1239
					Actual Positive	3248	730
Adboost	0.3289	0.279	0.265			Actual	
						Predicted Negativ	Predicted Positive
				Predictec	Actual Negative	185	1371
					Actual Positive	2343	1635
Gradient boost	0.213	0.213	0.262			Actual	
						Predicted Negativ	Predicted Positive
				Predictec	Actual Negative	580	976
					Actual Positive	3379	599
MLP (multi-layer perceptro	0.2575	0.237	0.229			Actual	
						Predicted Negativ	Predicted Positive
				Predictec	Actual Negative	257	1299
					Actual Positive	2810	1168

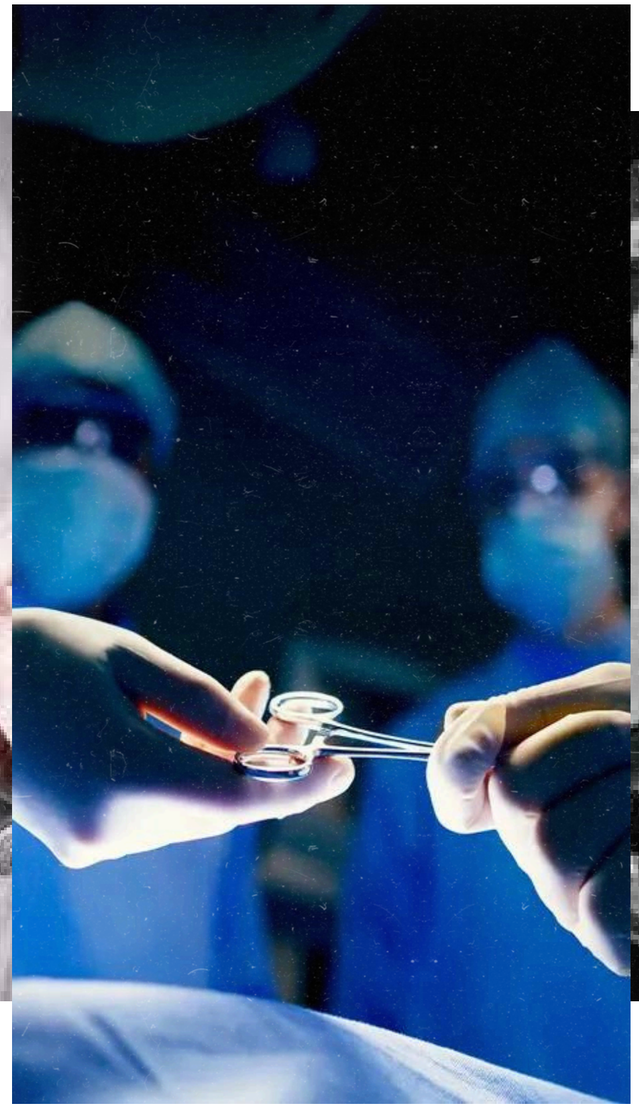
RESULTS

After applying PCA number of variables have reduced to 6 and These are the values we have obtained under Downsampledtraining set

These models are not accurate as the models we have obtained under without Applying PCA

Models Fitted for the Doensampled training Set with PCA							
Model	Accuracy	f1-score	recall	Confusion Matix			
Logistic Regression	0.595	0.592	0.693			Actual	
						Predicted Negative	Predicted Positive
				Predicted	Actual Negative	1426	130
					Actual Positive	2111	1867
Support Vector Machine	0.5617	0.561	0.683			Actual	
						Predicted Negative	Predicted Positive
				Predicted	Actual Negative	1496	60
					Actual Positive	2367	1611
Decision Tree	0.7315	0.682	0.691			Actual	
						Predicted Negative	Predicted Positive
				Predicted	Actual Negative	933	625
					Actual Positive	863	3115
Random Forest	0.8074	0.767	0.774			Actual	
						Predicted Negative	Predicted Positive
				Predicted	Actual Negative	1088	468
					Actual Positive	598	33800
Xgboost	0.7593	0.721	0.738			Actual	
						Predicted Negative	Predicted Positive
				Predicted	Actual Negative	1074	482
					Actual Positive	850	3128
KNN	0.7199	0.707	0.776			Actual	
						Predicted Negative	Predicted Positive
				Predicted	Actual Negative	1407	149
					Actual Positive	1401	2577
Adboost	0.6579	0.65	0.734			Actual	
						Predicted Negative	Predicted Positive
				Predicted	Actual Negative	1415	141
					Actual Positive	1745	2226
Gradient boost	0.7521	0.704	0.712			Actual	
						Predicted Negative	Predicted Positive
				Predicted	Actual Negative	967	589
					Actual Positive	783	3159
MLP (multi-layer perceptron	0.7479	0.729	0.782			Actual	
						Predicted Negative	Predicted Positive
				Predicted	Actual Negative	1338	218
					Actual Positive	1177	2801

BEST 3 MODELS



Model 1: Random Forest Model Fitted on Original Data Set

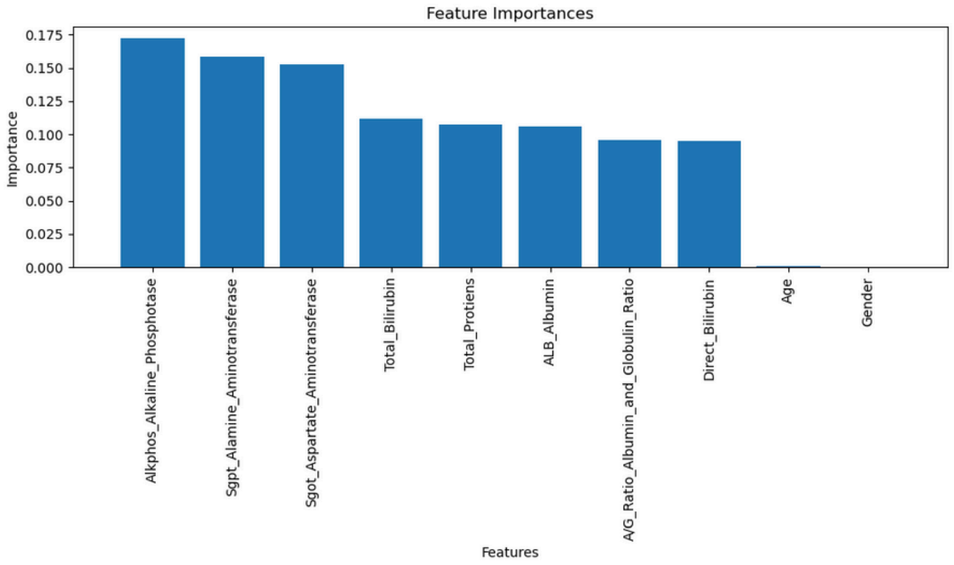
Training

```
Accuracy score of this model:  99.99 %
Misclassification rate of this model:  0.01 %

Report card of this model:
              precision    recall  f1-score   support

           0      1.000     1.000     1.000      6107
           1      1.000     1.000     1.000     15619

    accuracy                         1.000     21726
   macro avg      1.000     1.000     1.000     21726
weighted avg      1.000     1.000     1.000     21726

Confusion Matrix for y_pred_train:
+-----------------+----------------------+----------------------+
|                 |   Predicted Negative |   Predicted Positive |
+=================+======================+======================+
| Actual Negative |                 6104 |                    3 |
+-----------------+----------------------+----------------------+
| Actual Positive |                    0 |                15619 |
+-----------------+----------------------+----------------------+
```

Testing

```
Accuracy score of this model:  99.98 %
Misclassification rate of this model:  0.02 %

Report card of this model:
              precision    recall  f1-score   support

           0      1.000     0.999     1.000      1556
           1      1.000     1.000     1.000      3978

    accuracy                         1.000      5534
   macro avg      1.000     1.000     1.000      5534
weighted avg      1.000     1.000     1.000      5534

Confusion Matrix:
+-----------------+----------------------+----------------------+
|                 |   Predicted Negative |   Predicted Positive |
+=================+======================+======================+
| Actual Negative |                 1555 |                    1 |
+-----------------+----------------------+----------------------+
| Actual Positive |                    0 |                 3978 |
+-----------------+----------------------+----------------------+
```

IMPORTANT FEATURE SELECTION UNDER 1ST MODEL



- According to this plot Gender and Age has least significance for predicting Liver Disease

Therefore we have selected only top 8 Important variables and fitted the same model again

Model 1: Random Forest Model Fitted on Original Data Set

Training

Accuracy score of this model: 99.99 %
Misclassification rate of this model: 0.01 %

Report card of this model:

	precision	recall	f1-score	support
0	1.000	1.000	1.000	6107
1	1.000	1.000	1.000	15619
accuracy			1.000	21726
macro avg	1.000	1.000	1.000	21726
weighted avg	1.000	1.000	1.000	21726

Confusion Matrix for y_pred_train:

	Predicted Negative	Predicted Positive
Actual Negative	6104	3
Actual Positive	0	15619

Testing

Accuracy score of this model: 100.0 %
Misclassification rate of this model: 0.0 %

Report card of this model:

	precision	recall	f1-score	support
0	1.000	1.000	1.000	1573
1	1.000	1.000	1.000	3859
accuracy			1.000	5432
macro avg	1.000	1.000	1.000	5432
weighted avg	1.000	1.000	1.000	5432

Confusion Matrix:

	Predicted Negative	Predicted Positive
Actual Negative	1573	0
Actual Positive	0	3859

Model 2: XGboost Model Fitted on SMOTE Data Set

Training

Accuracy score of this model: 99.99 %
Misclassification rate of this model: 0.01 %

Report card of this model:

	precision	recall	f1-score	support
0	1.000	1.000	1.000	6277
1	1.000	1.000	1.000	15855
accuracy			1.000	22132
macro avg	1.000	1.000	1.000	22132
weighted avg	1.000	1.000	1.000	22132

Confusion Matrix y_pred_train:

	Predicted Negative	Predicted Positive
Actual Negative	6275	2
Actual Positive	0	15855

Testing

Accuracy score of this model: 99.98 %
Misclassification rate of this model: 0.02 %

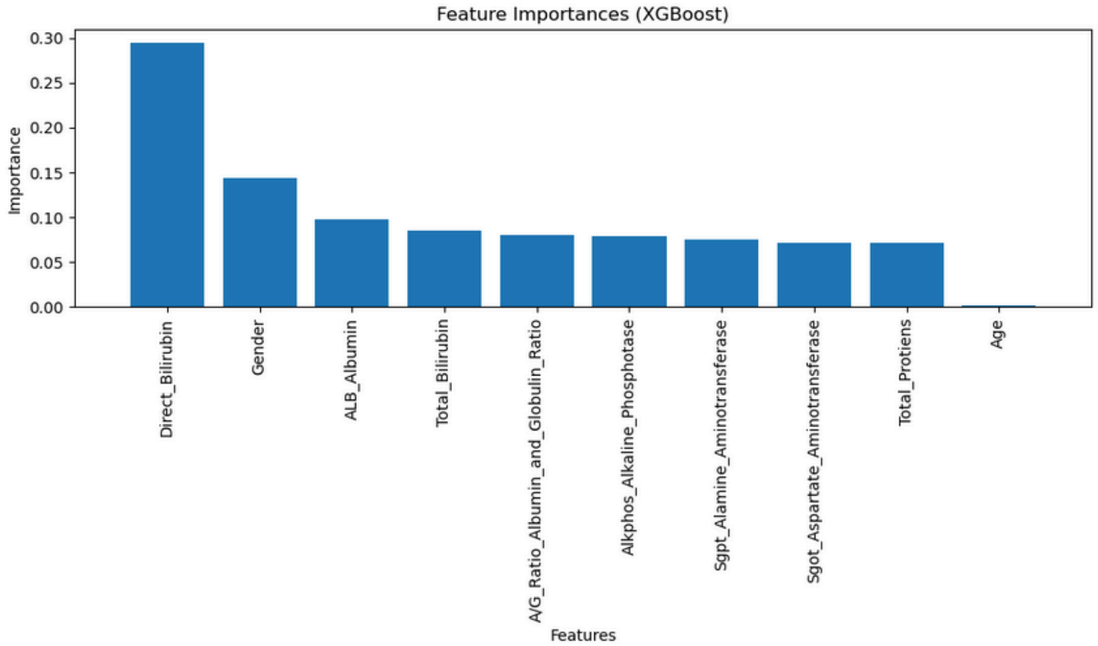
Report card of this model:

	precision	recall	f1-score	support
0	1.000	0.999	1.000	1556
1	1.000	1.000	1.000	3978
accuracy			1.000	5534
macro avg	1.000	1.000	1.000	5534
weighted avg	1.000	1.000	1.000	5534

Confusion Matrix:

	Predicted Negative	Predicted Positive
Actual Negative	1555	1
Actual Positive	0	3978

IMPORTANT FEATURE SELECTION UNDER 2ND MODEL



- According to this plot Age has least significance for predicting Liver Disease

Therefore we have selected only top 9 Important variables and fitted the same model again

Model 2: XGboost Model Fitted on SMOTE Data Set

Training

```
Accuracy score of this model:  99.99 %
Misclassification rate of this model:  0.01 %

Report card of this model:
              precision    recall  f1-score   support

           0      1.000     1.000     1.000      6277
           1      1.000     1.000     1.000     15855

    accuracy                         1.000     22132
   macro avg      1.000     1.000     1.000     22132
weighted avg      1.000     1.000     1.000     22132

Confusion Matrix:
+-----------------+----------------------+----------------------+
|                 |   Predicted Negative |   Predicted Positive |
+=================+======================+======================+
| Actual Negative |                 6275 |                    2 |
+-----------------+----------------------+----------------------+
| Actual Positive |                    1 |                15854 |
+-----------------+----------------------+----------------------+
```

Testing

```
Accuracy score of this model:  99.96 %
Misclassification rate of this model:  0.04 %

Report card of this model:
              precision    recall  f1-score   support

           0      0.999     0.999     0.999      1556
           1      1.000     1.000     1.000      3978

    accuracy                         1.000      5534
   macro avg      1.000     1.000     1.000      5534
weighted avg      1.000     1.000     1.000      5534

Confusion Matrix:
+-----------------+----------------------+----------------------+
|                 |   Predicted Negative |   Predicted Positive |
+=================+======================+======================+
| Actual Negative |                 1555 |                    1 |
+-----------------+----------------------+----------------------+
| Actual Positive |                    1 |                 3977 |
+-----------------+----------------------+----------------------+
```

Model 3: Random Forest Model Fitted on Downsampled Data Set

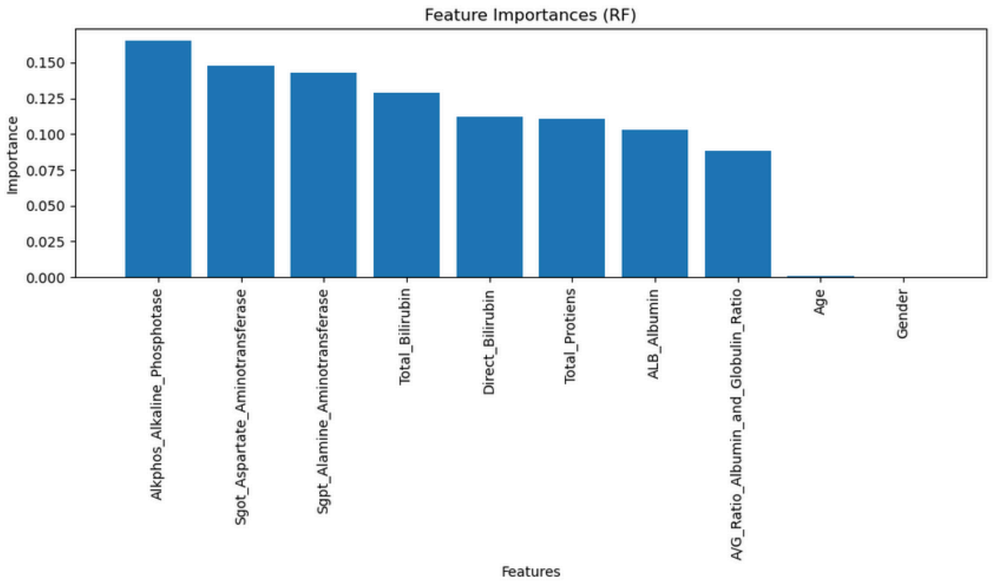
Training

```
Accuracy score of this model:  99.99 %
Misclassification rate of this model:  0.01 %

Report card of this model:
              precision    recall  f1-score   support

           0      1.000     1.000     1.000      6277
           1      1.000     1.000     1.000     15855

    accuracy                         1.000     22132
   macro avg      1.000     1.000     1.000     22132
weighted avg      1.000     1.000     1.000     22132

Confusion Matrix:
+-----------------+----------------------+----------------------+
|                 |   Predicted Negative |   Predicted Positive |
+=================+======================+======================+
| Actual Negative |                 6275 |                    2 |
+-----------------+----------------------+----------------------+
| Actual Positive |                    1 |                15854 |
+-----------------+----------------------+----------------------+
```

Testing

```
Accuracy score of this model:  99.98 %
Misclassification rate of this model:  0.02 %

Report card of this model:
              precision    recall  f1-score   support

           0      1.000     0.999     1.000      1556
           1      1.000     1.000     1.000      3978

    accuracy                         1.000      5534
   macro avg      1.000     1.000     1.000      5534
weighted avg      1.000     1.000     1.000      5534

Confusion Matrix:
+-----------------+----------------------+----------------------+
|                 |   Predicted Negative |   Predicted Positive |
+=================+======================+======================+
| Actual Negative |                 1555 |                    1 |
+-----------------+----------------------+----------------------+
| Actual Positive |                    0 |                 3978 |
+-----------------+----------------------+----------------------+
```

IMPORTANT FEATURE SELECTION UNDER 3RD MODEL



- According to this plot Gender and Age has least significance for predicting Liver Disease

Therefore we have selected only top 8 Important variables and fitted the same model again

Model 3: Random Forest Model Fitted on Downsampled Data Set

Training

```
Accuracy score of this model:  99.99 %
Misclassification rate of this model:  0.01 %

Report card of this model:
              precision    recall  f1-score   support

           0      1.000     1.000     1.000      6277
           1      1.000     1.000     1.000     15855

    accuracy                         1.000     22132
   macro avg      1.000     1.000     1.000     22132
weighted avg      1.000     1.000     1.000     22132

Confusion Matrix:
+-----------------+----------------------+----------------------+
|                 |   Predicted Negative |   Predicted Positive |
+=================+======================+======================+
| Actual Negative |                 6275 |                    2 |
+-----------------+----------------------+----------------------+
| Actual Positive |                    1 |                15854 |
+-----------------+----------------------+----------------------+
```

Testing

```
Accuracy score of this model:  99.98 %
Misclassification rate of this model:  0.02 %

Report card of this model:
              precision    recall  f1-score   support

           0      1.000     0.999     1.000      1556
           1      1.000     1.000     1.000      3978

    accuracy                         1.000      5534
   macro avg      1.000     1.000     1.000      5534
weighted avg      1.000     1.000     1.000      5534

Confusion Matrix:
+-----------------+----------------------+----------------------+
|                 |   Predicted Negative |   Predicted Positive |
+=================+======================+======================+
| Actual Negative |                 1555 |                    1 |
+-----------------+----------------------+----------------------+
| Actual Positive |                    0 |                 3978 |
+-----------------+----------------------+----------------------+
```

SUMMARY - BEST MODEL

	With All Variables			Number of Important variables	Only consisdering selected Variables		
	Evaluation Matrices	Training	Test		Evaluation Matrices	Training	Test
Model 1 Random Forest on ORIGINAL Training Set	Accuracr	0.9999	0.9998	8	**Accuracr**	**0.9999**	**1.0000**
	F1 Score	1.0000	1.0000		**F1 Score**	**1.0000**	**1.0000**
	Recall	1.0000	1.0000		**Recall**	**1.0000**	**1.0000**
Model 2 XGBoost on SMOTE Training Set	Accuracr	0.9999	0.9998	9	Accuracr	0.9999	0.9996
	F1 Score	1.0000	1.0000		F1 Score	1.0000	1.0000
	Recall	1.0000	1.0000		Recall	1.0000	1.0000
Model 2 Random Forest on DOWNSAMPLE Training Set	Accuracr	0.9999	0.9998	8	Accuracr	0.9999	0.9998
	F1 Score	1.0000	1.0000		F1 Score	1.0000	1.0000
	Recall	1.0000	1.0000		Recall	1.0000	1.0000

THE BEST MODEL IS RANDOM FOREST MODEL FITTED ON ORIGINAL DATA SET AFTER SELECTING TOP MOST 8 IMPORTANT VARIABLES

OUR DATA PRODUCT

LINK:

HTTPS://DRIVE.GOOGLE.COM/FILE/D/1B14DGZTGXWTJ3OOW7XY08T_YSZ0MNJN9/VIEW?PLI=1

OUR TEAM

Tharindu Fernando
15522

Hiruni Kudagama
15680

Kaveesha Vidushinie
15572

REFERENCES

[1] https://www.medmastery.com/guides/liver-lab-clinical-guide/everything-you-need-know-about-albumin-and-liver-function

[2]https://www.medchunk.com/tests/high-bilirubin-total-and-high-alkaline-phosphatase

[3]https://www.healthline.com/health/a-g-ratio-high

[4] Rahman, A. K. M. et.al. (2019). A Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms. 8. 419-422.

[5] Younossi, Z.M. et al. (2023) 'The global burden of liver disease', Clinical Gastroenterology and Hepatology, 21(8), pp. 1978–1991. doi:10.1016/j.cgh.2023.04.015.

[6] https://www.virogates.com/infomation/liver-function-tests-guide/

[7] Wu, X.-N. et al. (2024) 'Global burden of liver cirrhosis and other chronic liver diseases caused by specific etiologies from 1990 to 2019', BMC Public Health, 24(1). doi:10.1186/s12889-024-17948-6.

[8] Dhyani, A. et al. (2023) 'Comparative analysis of supervised machine learning algorithms for liver disease prediction with smote enhancement', 2023 3rd Asian Conference on Innovation in Technology (ASIANCON) [Preprint]. doi:10.1109/asiancon58793.2023.10270381.

[9] https://www.mayoclinic.org/diseases-conditions/liver-problems/symptoms-causes/syc-20374502

[10]https://mcpress.mayoclinic.org/women-health/my-liver-enzymes-are-elevated-now-what/

THANK YOU!