PREDICTION OF LIVER DISEASE AND IDENTIFYING THE KEY FACTORS

Kusara Udayana - 16073
Vidusara Vimukthi - 16074
Samudika Wanasinghe - 16075
Kavindu Weerasekara - 16076
Sadini Thiranja - 16263
Shashin - 16375

1. Introduction

With millions of people affected and a heavy weight on healthcare systems, liver disease is a major global health concern. The liver's healthy operation is key to general health because it plays a crucial role in metabolism, detoxification, and the creation of necessary proteins. Prompt identification and treatment of liver illness can prevent serious consequences, lower medical expenses, and enhance patient outcomes. The mild beginning and course of liver illnesses, however, frequently make prompt identification and therapy difficult.

Predictive modeling based on medical and lifestyle data has become a potent technique for liver disease early detection in this context. The likelihood of liver disease can be predicted and the underlying elements that contribute to its development can be identified by utilizing statistical and machine learning approaches. In addition to supporting public health activities targeted at illness prevention, this strategy helps with individual-level diagnosis**.**

Problem Statement:

Predicting liver disease using a dataset that includes medical and lifestyle data is the main goal of this work. The secondary objective is to determine and examine the risk factors for liver disease, providing information that can direct clinical procedures and lifestyle modifications.The dataset was obtained from the UCI machine learning repository (Ramana, & B. & Venkateswarlu, n.d.)y

Data Description:

The study's dataset has 583 observations with 11 attributes that reflect biochemical, medical, and demographic information. Each feature is described in full below:

1. **Age**: This number variable shows the individuals' ages in years. A major demographic determinant, it is frequently linked to the occurrence of chronic illnesses, such as liver disease.
2. **Gender**: A variable that is categorical and indicates whether the people are male or female. According to medical research, there are gender disparities in the incidence and course of liver disease.
3. **TB (Total Bilirubin)**: 2. A continuous variable that gauges blood levels of bilirubin overall. Liver dysfunction and associated disorders are indicated by elevated bilirubin levels.
4. **DB (Direct Bilirubin):** A continuous variable that gauges bilirubin levels directly. This measure aids in distinguishing between different liver disease types.
5. **Alkphos (Alkaline Phosphatase):** This number denotes the concentrations of alkaline phosphatase, an enzyme whose high levels frequently indicate biliary blockage or liver damage.

6. **Sgpt (Alanine Aminotransferase):** A number that indicates the amount of the enzyme alanine aminotransferase, which is released into the bloodstream when the liver is injured.
7. **Sgot (Aspartate Aminotransferase):** The levels of aspartate aminotransferase, another important enzyme associated with liver health, are measured by this numerical variable, which is comparable to SGPT.
8. **TP (Total Proteins):** A continuous measure of the blood's overall protein content. It displays the capacity of the liver to produce proteins.
9. **ALB (Albumin):** A continuous variable measuring the albumin levels, which provide an indication of liver function and nutritional status.
10. **A/G Ratio (Albumin to Globulin Ratio):** This variable measures the ratio of albumin to globulin in the blood. Deviations from the normal range may point to liver or kidney dysfunction.
11. **Selector:** A binary target variable indicating the presence (1) or absence (0) of liver disease. This is the variable of interest for prediction and classification tasks.

Significance of the Study:

The goal of this analysis is to close the gap between data-driven insights and medical research. The project aims to forecast liver disease and identify important risk variables in order to:

- Help medical professionals make well-informed choices.
- Make it possible for at-risk persons to get early interventions.
- Raise awareness of the medical and lifestyle variables that affect liver health.

The knowledge gained from this research could improve individualized treatment and support more comprehensive public health initiatives for the management and prevention of liver disease.

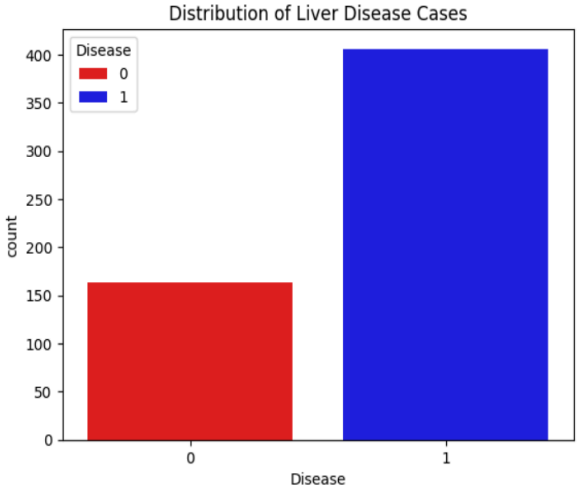
2. Exploratory Data Analysis (EDA) on Indian Liver Disease

2.1 Data Preprocessing

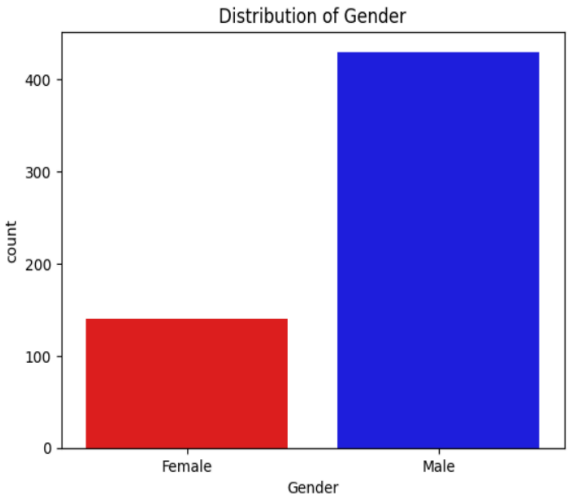
- The dataset contains 583 observations, but after removing duplicates (keeping only the first occurrence), the number of unique observations was reduced to 570.
- There are 4 missing values in the A/G ratio column and imputed with the median since it shows a skewed distribution.

2.2 Exploring the Distribution and Characteristics of Key Variables

- The bar chart below shows that, among the 570 observations in this dataset, 406 individuals have liver disease, while 164 do not. This highlights that the majority of people in the dataset are affected by liver disease.



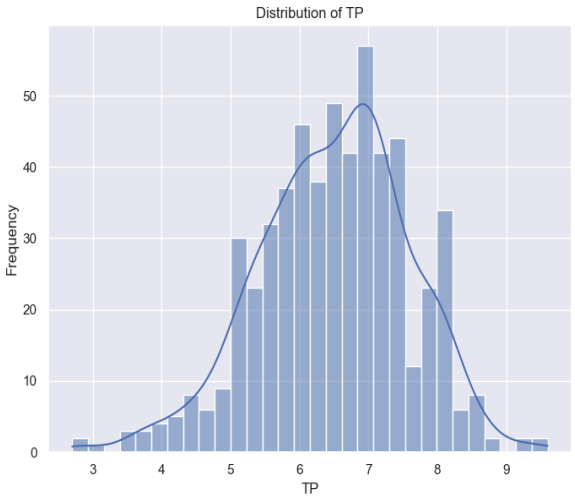
- The bar chart below shows the gender distribution among the individuals in the dataset. The majority are males (430), while the remaining 140 are females.



- Numerical summaries of potential variables related to liver disease, including age, bilirubin levels (direct and total), alkaline phosphatase concentration, and other relevant continuous variables.

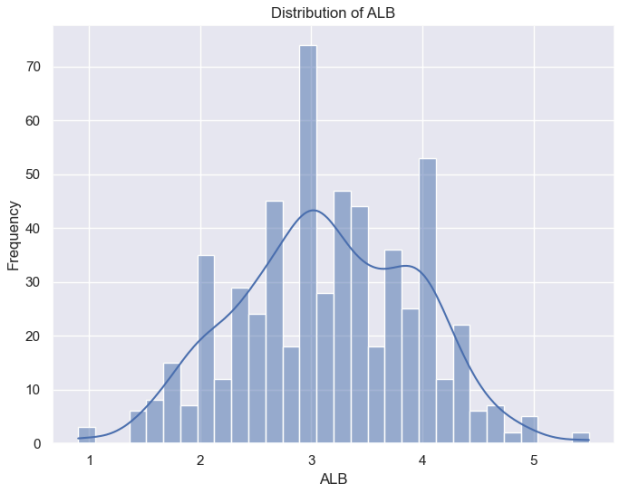
	Age	TB	DB	Alkphos	Sgpt	Sgot	TP	ALB	A/G Ratio
count	570.000000	570.000000	570.000000	570.000000	570.000000	570.000000	570.000000	570.000000	570.000000
mean	44.849123	3.321754	1.497544	291.750877	79.728070	109.380702	6.496316	3.148947	0.948018
std	16.242182	6.267941	2.833231	245.291859	181.471697	290.880671	1.088300	0.796813	0.318510
min	4.000000	0.400000	0.100000	63.000000	10.000000	10.000000	2.700000	0.900000	0.300000
25%	33.000000	0.800000	0.200000	176.000000	23.000000	25.000000	5.800000	2.600000	0.700000
50%	45.000000	1.000000	0.300000	208.000000	35.000000	41.000000	6.600000	3.100000	0.950000
75%	58.000000	2.600000	1.300000	298.000000	60.000000	86.750000	7.200000	3.800000	1.100000
max	90.000000	75.000000	19.700000	2110.000000	2000.000000	4929.000000	9.600000	5.500000	2.800000

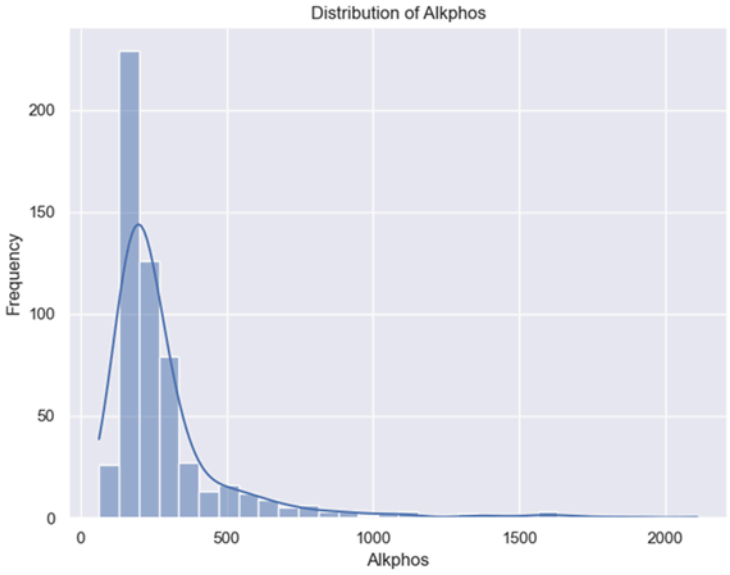
- The histogram for age, with a range of 86 (minimum age: 4, maximum age: 90), along with the Kernel Density Estimation (KDE) curve, forms a bell-shaped distribution, indicating that age is approximately normally distributed.



- Histogram with the kernel density estimation overlay of variable Total protein (TP) in blood shows a slight left-skewed distribution with few outliers in both edges.

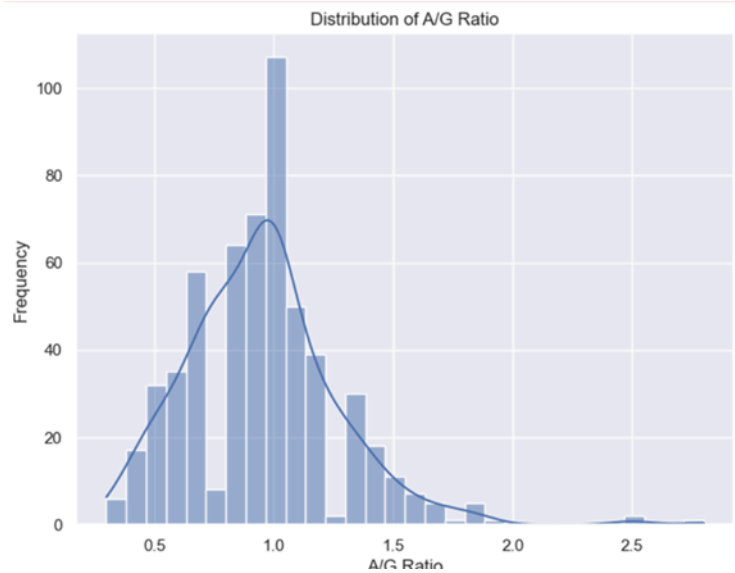
- Histogram with the kernel density estimation overlay of variable Albumin (ALB) shows an approximately normal distribution with few outliers in both edges.

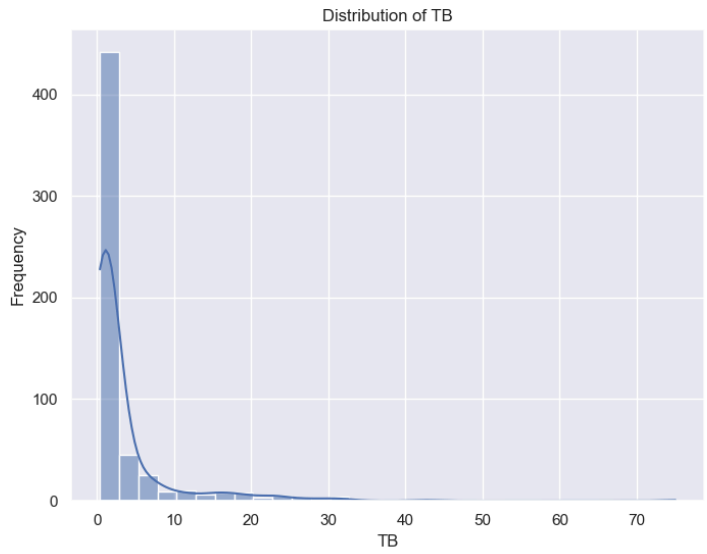


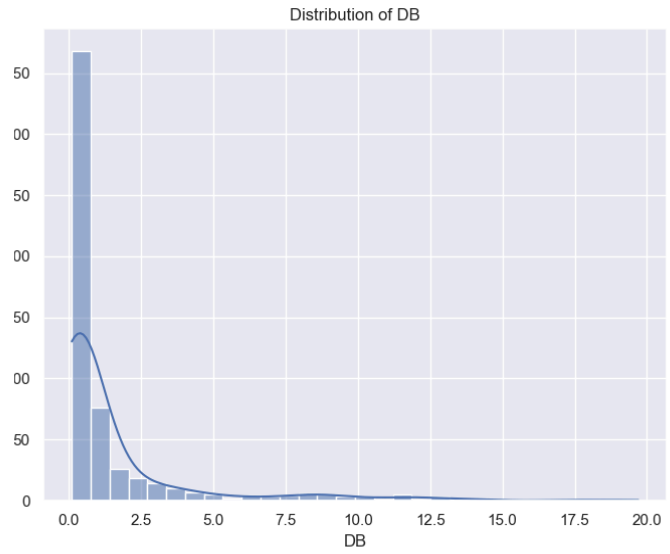


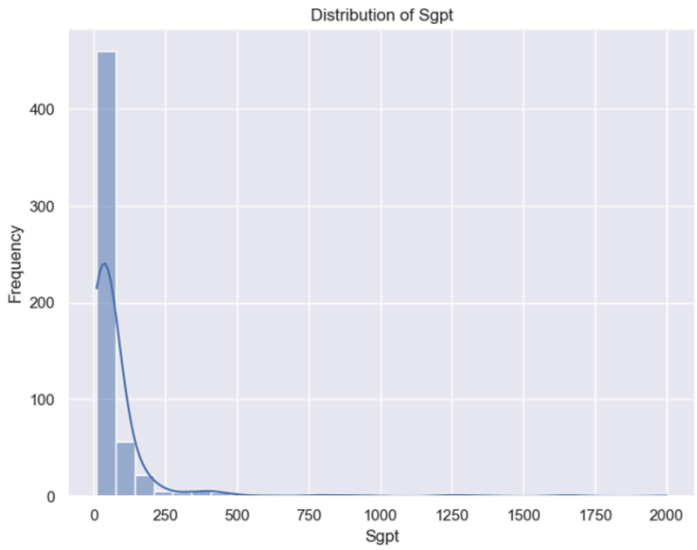
- Histogram with the kernel density estimation overlay of variable Alkphos shows a right-skewed distribution with majority of values concentrated around 210.

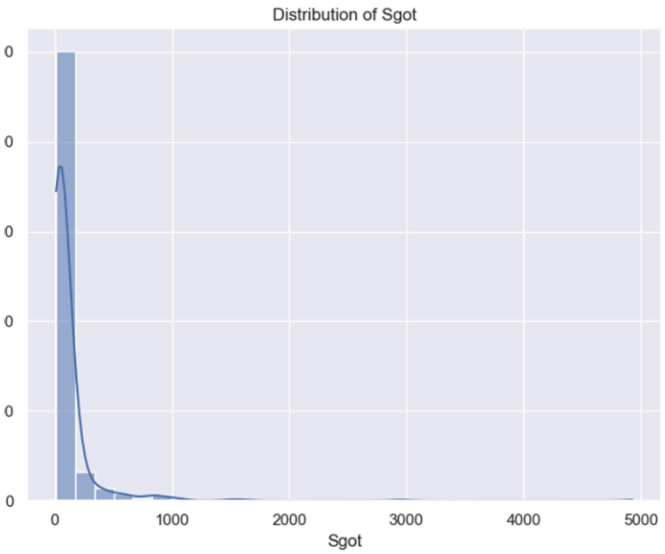
- Histogram with the kernel density estimation overlay of variable Albumin and Globulin Ratio (A/G ratio) shows a slight right skewed distribution with few extreme outliers in the right tail.





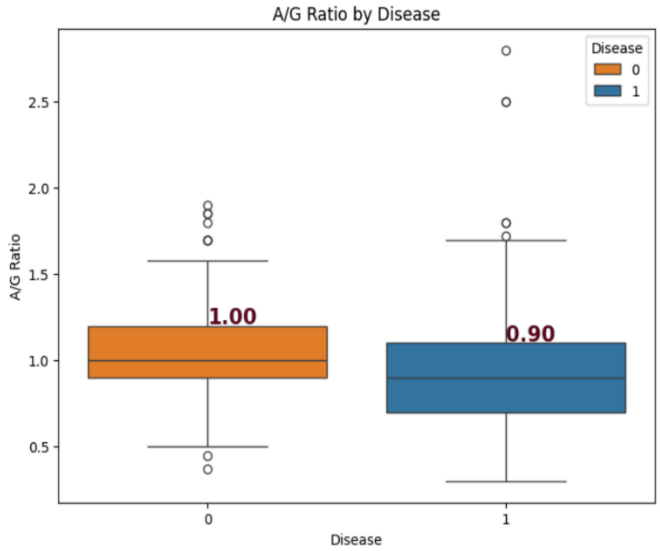




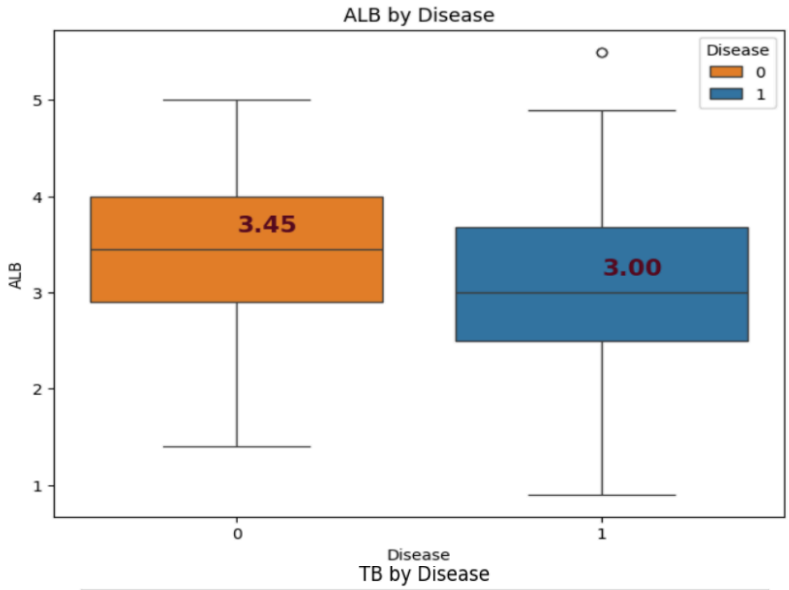


- The histograms with Kernel Density Estimation (KDE) curves for bilirubin (total and direct) and enzyme counts (SGPT and SGOT) show right-skewed distributions with potential outliers.

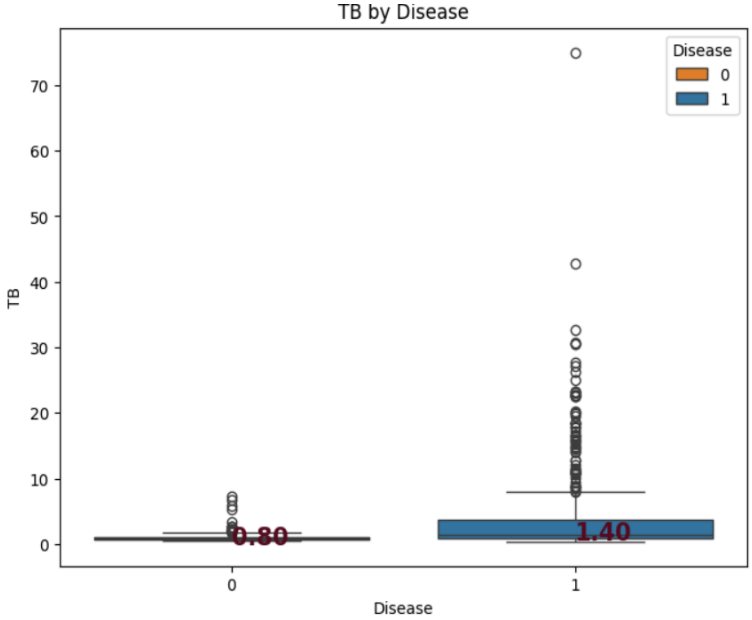
2.3 Exploring Trends and Patterns in Disease Status and Clinical Factors



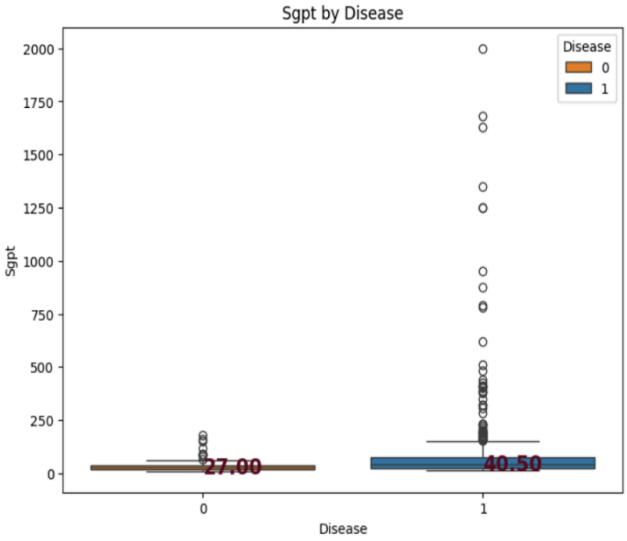
- The median A/G ratio for individuals with the disease is 0.9, which is lower compared to the median of 1 for those without the disease. Additionally, both boxplots show a few outliers, indicating some extreme values in both groups.

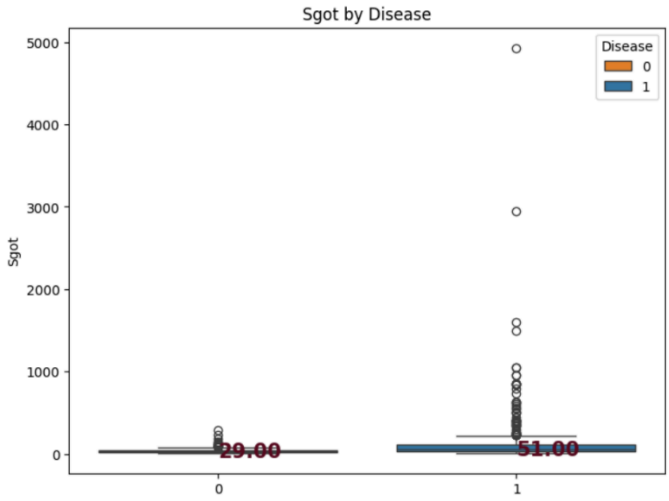


- The median Albumin level for individuals with the disease is 3.0, which is lower compared to the median of 3.45 for those without the disease.

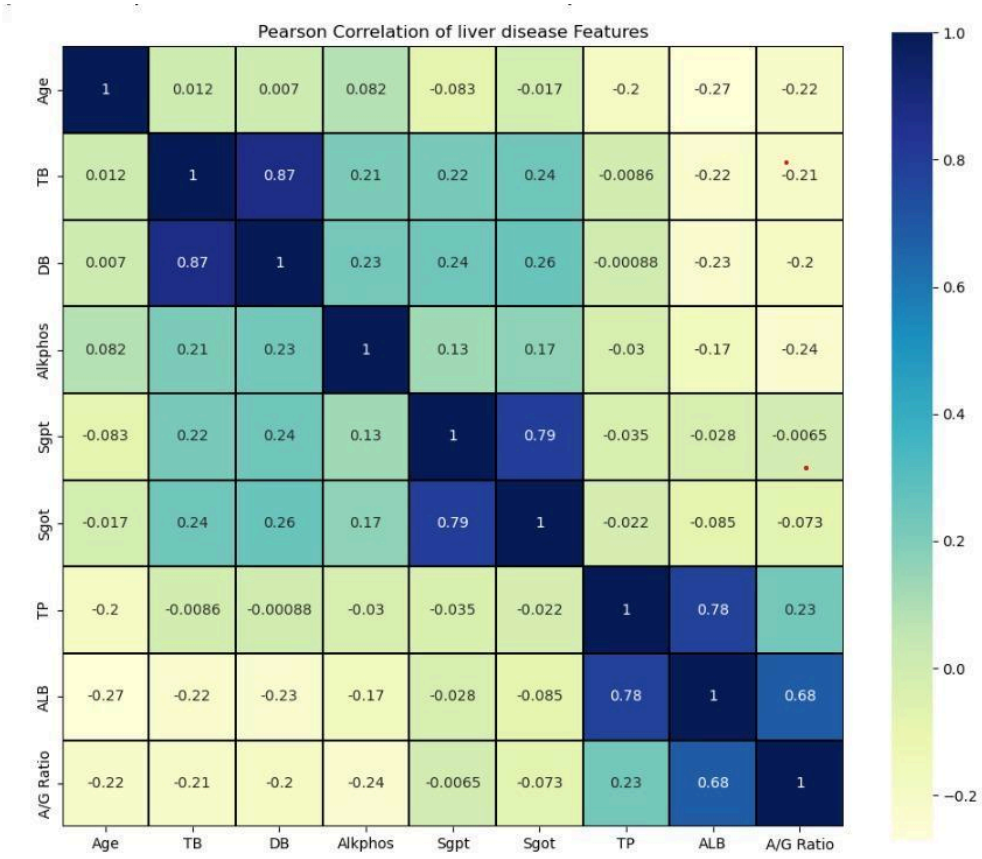


- Due to a considerable number of outliers, both boxplots are distorted. However, the median total bilirubin level for individuals with the disease is 1.40, which is higher than the median of 0.80 for those without the disease.
- A similar pattern is observed for the direct bilirubin levels in both groups, with individuals having the disease showing higher levels compared to those without the disease.





- The medians of alanine aminotransferase (SGPT) and aspartate aminotransferase (SGOT) levels are higher in the group with the disease compared to the group without the disease.



- There are strong positive correlations observed between direct bilirubin (DB), total bilirubin (TB), and the enzymes SGOT and SGPT, as well as between albumin (ALB) and total protein (TP).

3. Advanced Methodology

3.1. Choosing the statistical model

Initially a logistic regression model is selected to fit the model for the following reasons :

- Binary outcome handling : Logistic regression is chosen because the response variable is binary (having a disease or not having a disease)
- Easier interpretation : Logistic regression models allow for meaningful interpretation as the predator variables are linearly related with the log odds ratio. This allows to quantify how each predictor variable affects the likelihood of having a disease
- Ability to perform statistical inference : Logistic regression facilitates hypothesis testing , confidence interval estimation to assess the significance of the predictor variables
- Model fit assessment : Various statistical models can be used to help evaluate the model validity and the goodness of fit of the model

3.2 Model fitting and evaluation

A logistic regression model was fitted using statsmodels which is different from the logistic regression model that is obtained from scipy , since the statslearn gives the model coefficient estimates as well as the necessary calculations for statistical inference.
After fitting the model with the statsmodels , the following model was obtained.

```
Optimization terminated successfully.
         Current function value: 0.491107
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:          liver_disease   No. Observations:                  570
Model:                          Logit   Df Residuals:                      559
Method:                           MLE   Df Model:                           10
Date:                Thu, 30 Jan 2025   Pseudo R-squ.:                 0.1816
Time:                        14:17:29   Log-Likelihood:                -279.93
converged:                       True   LL-Null:                       -342.06
Covariance Type:            nonrobust   LLR p-value:                6.936e-22
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -3.0700      1.279     -2.401      0.016      -5.576      -0.564
Age            0.0181      0.006      2.820      0.005       0.006       0.031
Gender        -0.0169      0.233     -0.072      0.942      -0.474       0.441
TB             0.0151      0.107      0.141      0.888      -0.194       0.224
DB             0.5050      0.280      1.806      0.071      -0.043       1.053
Alkphos        0.0013      0.001      1.565      0.118      -0.000       0.003
Sgpt           0.0111      0.005      2.198      0.028       0.001       0.021
Sgot           0.0026      0.003      0.822      0.411      -0.004       0.009
TP             0.7752      0.361      2.147      0.032       0.068       1.483
ALB           -1.3991      0.699     -2.001      0.045      -2.769      -0.029
A/G Ratio      1.3339      1.063      1.254      0.210      -0.750       3.418
==============================================================================
```

3.2.1 Checking the goodness of fit of the model -

For checking the goodness of fit of the model, the residual deviance can be calculated which is -2 times the loglikehood of the model.

```
# Deviance is calculated as -2 times the log-likelihood
deviance = -2 * model.llf

print(f"Deviance of the fitted model: {deviance}")
```

```
Deviance of the fitted model: 559.8624360106223
```

The residual deviance follows a chi-squared distribution with degrees of freedom number of predictor variables which is 10. The critical value can be calculated as below.

```
import scipy.stats as stats

# Chi-square critical value for alpha = 0.05 and 10 degrees of freedom
alpha = 0.05
degrees_of_freedom = 10

chi_square_critical = stats.chi2.ppf(1 - alpha, degrees_of_freedom)

print(f"Chi-Square Critical Value (α=0.05, df=10): {chi_square_critical}")
```

```
Chi-Square Critical Value (α=0.05, df=10): 18.307038053275146
```

And the null hypothesis is the model fits the data well with the alternative hypothesis being, the model does not fit the data well. As the test statistic is greater than the critical value , there is enough evidence at 5% significance to reject Ho , indicating that the logistic regression model fitted with maximum likelihood estimation is not a good fit for the model.

3.2.2 Significant Variables

From the p-values of the model output , it can be seen that the p values of age, Sgpt , Tp and ALB are less than 0.05 indicating that the above variables have significant effect on the response variable.However this provides no a meaningless interpretation , as from the goodness of fit test of the model, it was obtained that the model does not fit the data well.

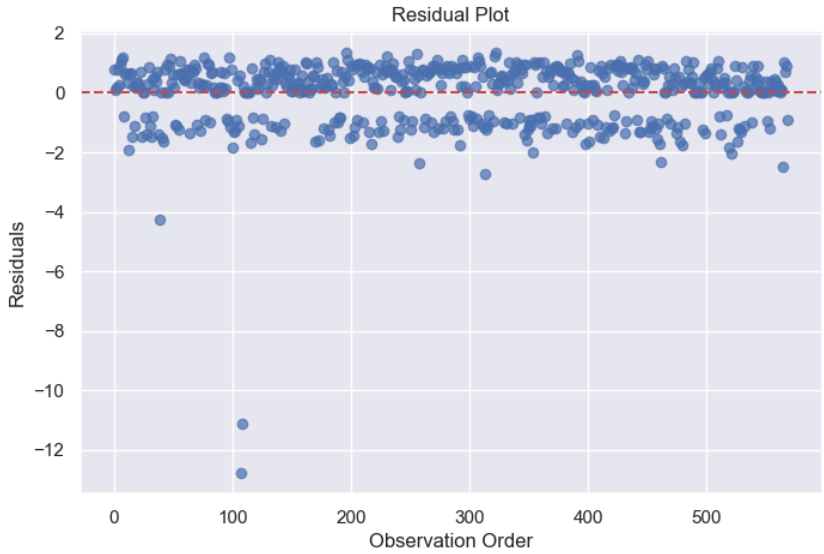
3.2.3 Checking for validity of assumptions

Assumptions of logistic regression are ,

- The response variable is binary
- The Binary response variable should be independent of each other
- No presence of multicollinearity among predictor variables
- No presence of extreme outliers in the predictor variables

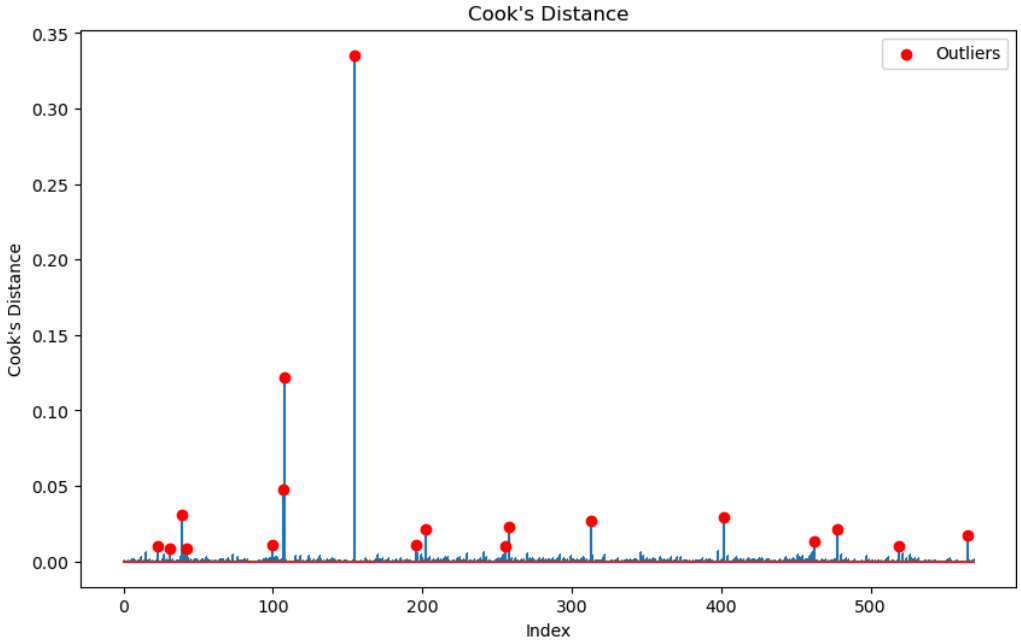
The independence of observations can be checked from the residual plot , with the residuals plotted with the order of time, and when it is plotted , the below figure is obtained

.



This plot shows that there is no pattern in the residual plot , since the majority of the residuals are scattered around the zero reference line indicating the independence of observations of residuals. However , few observations are greater than -10 in the residuals indicating that these might be outliers.and these could have high leverage

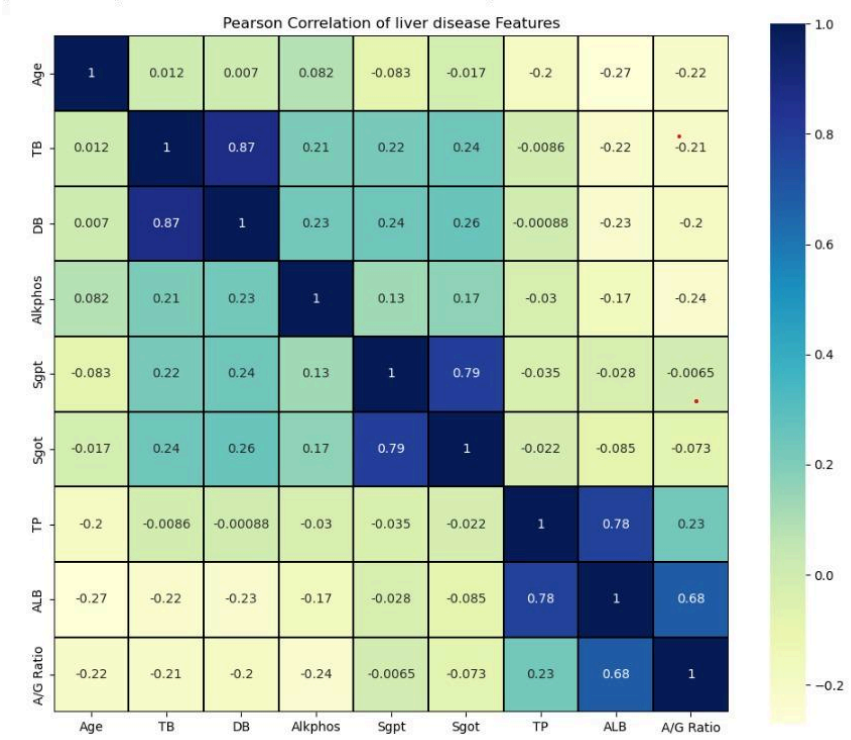
It can be checked with the plot of influence points vs observation order if there are influential points in the data . Influential points significantly affect the regression if they are excluded in the model.



From the above figure , it can be seen that there are 18 influential points in the dataset , and these points can cause the logistic regression to distort the model coefficients , causing the

model the estimated parameters to be inaccurate and mislead the statistical inferences conducted.

Moreover , the presence of multicollinearity can be checked pairwise for predictor variables from the correlation heatmap of variables which can be seen in the below figure.



From the correlation heatmap , it can be seen that (TB, DB) , (Sgpt,Sgot) , (Tp,ALB) and ALB , A/G Ratio) are highly correlated. Moreover, VIF measures how much a predictor is explained by all the other predictors in the model, helping to detect collinearity among multiple variables.

	Feature	VIF
0	Age	7.764789
1	Gender	4.050951
2	TB	5.453792
3	DB	5.718159
4	Alkphos	2.635022
5	Sgpt	3.302523
6	Sgot	3.180294
7	TP	101.578685
8	ALB	132.375516
9	A/G Ratio	25.419165

From the above figure, it can be seen that VIF values > 5 , Age , TB , DB have moderate multicollinearity while TP and ALB have very multicollinearity. Multicollinearity causes the model to be unstable , therefore the coefficient estimates may not be accurate due to high variability of

the coefficient estimates. Moreover , since the variance of the estimated coefficients are inflated , the standard errors of the variables are inflated , this affects the p values of the variables thus making the statistical inferences unreliable.

Therefore, in order to solve the problem of multicollinearity and influential points, regularization techniques can be used.

However, Scikit-learn's logistic regression, by default, uses **L2 regularization** (also known as Ridge regularization). This technique penalizes large coefficients by adding a penalty term to the cost function, which helps reduce the impact of multicollinearity and prevents overfitting. L2 regularization works by shrinking the coefficients of correlated features toward zero, making them less influential in the model without completely eliminating them. Moreover , machine learning models such as Support vector machines , random forests , decision trees and XG boost are robust to multicollinearity and the impact of influential points is minimized.

3.3 Data Preprocessing for Machine Learning models

Before fitting machine learning models, it is required to scale the features since the **model's performance can be impacted by the differing scales of features**, especially for algorithms like SVM, KNN, which rely on distance calculations. Scaling ensures that all features contribute equally to the model.

```
from sklearn.preprocessing import StandardScaler

# Instantiate the StandardScaler
scaler = StandardScaler()

# Fit the scaler on the data
scaler.fit(X)

# Transform the features
X_scaled = scaler.transform(X)

# View the scaled features for the first instance
print(X_scaled[0])
X_scaled.shape
```

```
[ 1.24174059 -1.75254916 -0.41864741 -0.4937018  -0.42742097 -0.35148206
 -0.31442779  0.27928958  0.18973749 -0.15044652]
```

Due to the class imbalance between those who have a disease and those who do not, machine learning models may become biased toward the majority class, leading to poor accuracy in detecting the disease. Therefore, oversampling is done to balance the classes and improve model performance. And then the data is split with 70% of the data being split as the training set and 30% of the data being split as the test set. This ensures that the model learns patterns from a sufficient amount of data while keeping enough unseen data to evaluate its performance

accurately. This helps prevent overfitting and provides a reliable estimate of how well the model generalizes to new data.

```python
from imblearn.over_sampling import RandomOverSampler
from collections import Counter

# Initialize the oversampler
over_sampler = RandomOverSampler(random_state=42)

# Apply oversampling to your data
X_resampled, Y_resampled = over_sampler.fit_resample(X_scaled, Y)

# Check the new class distribution
print(f"Original class distribution: {Counter(Y)}")
print(f"Resampled class distribution: {Counter(Y_resampled)}")
```

```
Original class distribution: Counter({1: 406, 0: 164})
Resampled class distribution: Counter({1: 406, 0: 406})
```

```python
from sklearn.model_selection import train_test_split
# Split data into train and test
X_train_scaled, X_test_scaled, y_train, y_test =train_test_split(X_resampled, Y_resampled, train_size=.7, random_state=25)
# Check the splits are correct
print(f"Train size: {round(len(X_train_scaled) / len(X_resampled) * 100)}% \n\
Test size: {round(len(X_test_scaled) / len(X_resampled) * 100)}%")
print(y_train.value_counts())
```

```
Train size: 70%
Test size: 30%
liver_disease
0    286
1    282
Name: count, dtype: int64
```

Now Machine Learning models can be fitted

4. Results with Discussion

4.1 Logistic Regression

Since the logistic regression of the Scikit Learn by default uses regularization (L2 regularization) , it can handle the effect of multicollinearity , by minimizing the variance of coefficient estimates , thus more stable models will be produced.
The output of the coefficient estimates are given below

```
# use feature names from the original dataset
feature_names = list(x.columns)

# Get logistic regression coefficients
logistic_coefficients = logistic_regression.coef_[0]

# Print feature names with their corresponding coefficients
for feature, coef in zip(feature_names, logistic_coefficients):
    print(f"{feature}: {coef}")
```

```
const: 0.0
Age: 0.309779935929176
Gender: 0.012390457658617036
TB: 0.545902763328926
DB: 1.0947390194237976
Alkphos: 0.42490773402813814
Sgpt: 1.4708754335859695
Sgot: 0.7037995180166061
TP: 0.47235021730658727
ALB: -0.5760983729918365
A/G Ratio: 0.2002605207378667
```

It can be seen how the coefficient estimates are different from the logistic regression model in Statsmodels because of the regularization used.

Since the response variable is having liver disease (1) or not (0), the coefficients represent the log-odds change of having liver disease for a one-unit increase in each predictor, holding other variables constant.

- Age (0.310) : Older individuals are more likely to have liver disease.
- Gender (0.012) : Gender has a very small effect on liver disease likelihood.
- Total Bilirubin (TB) (0.546) : Higher TB levels increase the risk of liver disease.
- Direct Bilirubin (DB) (1.095) : A strong positive relationship; higher DB levels significantly increase the risk.
- Alkaline Phosphatase (Alkphos) (0.425) :Higher Alkphos is associated with increased liver disease risk.

- SGPT (1.471) : A strong effect; higher SGPT levels are highly associated with liver disease.
- SGOT (0.704) : Higher SGOT levels increase the likelihood of liver disease.
- Total Protein (TP) (0.472) : Higher TP is linked to an increased risk of liver disease.
- Albumin (ALB) (-0.576) : Higher albumin levels reduce the likelihood of liver disease.
- A/G Ratio (0.200) : A slight positive effect; higher A/G ratio increases the risk slightly.

Logistic Regression Results:

	precision	recall	f1-score	support
0	0.70	0.90	0.79	120
1	0.87	0.63	0.73	124
accuracy			0.76	244
macro avg	0.78	0.76	0.76	244
weighted avg	0.79	0.76	0.76	244

From the classification report of the logistic regression , it can be seen that the model correctly classifies 76% of the cases. There is a high recall for class 0 which is not having a disease , which is that the model successfully identifies 90% of the actual non disease cases and the recall of 0.63 for class 1 shows that the model captures only 63% of the disease cases showing that the model may struggle for disease cases , that is some disease individuals may be misclassified as not having a disease. Therefore, when detecting a disease this is problematic and other models might perform better.

4.2 Decision Trees

	Feature	Importance
5	Alkphos	0.176570
4	DB	0.166559
1	Age	0.160840
6	Sgpt	0.155546
10	A/G Ratio	0.102323
7	Sgot	0.082810
8	TP	0.061352
9	ALB	0.048019
3	TB	0.039721
2	Gender	0.006260
0	const	0.000000

It can be seen that Alkphos is the most important feature for the prediction of liver disease while it has a moderate effect in predicting the disease according to the logistic regression. The variables DB, Age , Sgpt and A/G Ratio are of relative importance.It can be seen that these variables show importance in the logistic regression in varying terms. However, gender shows no influence in both models, showing that it has very small effect on predicting whether a person has a disease or not.

Decision Tree Results:

	precision	recall	f1-score	support
0	0.75	0.91	0.82	120
1	0.89	0.71	0.79	124
accuracy			0.81	244
macro avg	0.82	0.81	0.81	244
weighted avg	0.82	0.81	0.81	244

From the classification report of the decision tree , it can be seen that the model seen above, the model accuracy is 81% showing that the model classifies , 81% of the cases correctly showcasing that the model has a good performance.The recall scores shows that the model identifies 91% of the actual no disease correctly while the model identifies 71% actual disease cases correctly showcasing that although the decision tree performs better for disease prediction compared to logistic regression (71% vs 63%) , the model performs well for predicting not having a disease.

The model has a good balance between precision and recall scores for both cases with the model having a better recall score for disease cases when compared to logistic regression. This model is good for classification, however minor improvements can be made to the model to enhance its performance.

4.3 Random Forest

A random forest model can also be fitted and the feature importance can be seen in the output given below.

```
const: 0.0
Age: 0.12017178922673
Gender: 0.01604916555750463
TB: 0.12528290741693654
DB: 0.09122651815247712
Alkphos: 0.14506017515091865
Sgpt: 0.12732609362445846
Sgot: 0.12610668916245066
TP: 0.08216966625837946
ALB: 0.08963171676843525
A/G Ratio: 0.07697527868170943
```

It shows that the most important feature for predicting liver disease is Alkphos with SGPT, SGOT and TB being important as well. Age is of relative importance and it shows that older individuals have a notable association with having a liver disease. Moreover , it can be seen that gender is of the least importance for predicting whether someone has a liver disease or not.

In comparison to logistic regression , it can be seen that in logistic regression and random forest , age is of moderate importance and gender has no importance. DB and Alphos are considered important in prediction of liver disease but Random forest gives the highest importance to Alkphos while logistic regression assigns a moderate positive effect. However , although it can be seen from both random forest and decision trees, which features are important, logistic regression gives more quantitative insights and helps understand the direction of the relationship between the predictor variables and the liver disease.

The classification report of the random forest models is given in the figure below.

Classification Report:

	precision	recall	f1-score	support
0	0.82	0.88	0.85	120
1	0.88	0.81	0.85	124
accuracy			0.85	244
macro avg	0.85	0.85	0.85	244
weighted avg	0.85	0.85	0.85	244

It shows that the model accuracy is 85% , which means that 85% of all the cases were identified correctly. The precision scores for both class 0 and class 1 are slightly same with class 1 having higher precision , which means that out of all predictions the model predicted that respective class , 82% were correct for class 0 cases , and 88% were correct for class 1 cases. The recall score of 88% for class 0 says that out of all no liver disease cases , the model identified 88% of them while the recall score for class 1 shows that out of all liver disease cases , 81% of them were correctly identified by the model. Thus it shows that the model is good at identifying both liver disease cases and no liver disease cases and the overall model has the best performance out of both logistic regression and decision trees.

4.3.1 Improving the model performance

Moreover , for the model performance obtained from the above random forest model for prediction of liver disease , the model performance can still be improved with **hyper parameter turning.** This can be done using GridsearchCv by changing the parameters to see if the model improves its performance. Hyperparameters such as n_estimators, max_depth, min_sample_split , min_samples_leaf and max_features can be tuned to get optimal model performance.

Furthermore , feature selection can be done to improve the model performance , as it can be seen that the predictor variable gender is minimal to all the models and it can be seen if it improves the model performance by removing that feature. Additionally new features can also be tried by combining existing features to see if it improves model performance.

5. Conclusion

In the analysis, 3 machine learning models were evaluated - Logistic regression, Decision Trees and Random Forest for predicting the liver disease.

Logistic Regression with its regularization technique performed better than the traditional logistic regression obtained from statslearn with an accuracy of 76%. However, the recall for identifying the disease cases was lower than desired showing that it may sometimes predict someone who has liver disease as not having a liver disease. The Decision Tree model showed a better performance than logistic regression with having an accuracy of 81% and having a higher recall score for predicting the liver disease than logistic regression and it suggested Alkphos and DB as the strongest factors in the prediction of liver disease. Random Forest outperformed both logistic regression and decision trees in all scores of accuracy , prediction and recall , showing that random forest has the best model performance. **Although logistic regression provides insights on the direction of coefficients and enables interpretation of the results , the best model for accurate liver disease classification is Random Forests.**

Key takeaways from the random forest which is the model that has the best performance on liver disease prediction : **Alkphos , SGPT , SGOT , Total Bilirubin and Direct Bilirubin** are the strongest predictors of having a disease or not. Age has some influence on liver disease but other medical factors are more important. Gender has the minimal impact of having a liver disease or not. From the logistic regression sign of the coefficients , it can be seen that **Alkphos , SGPT , SGOT , Total Bilirubin and Direct Bilirubin ,** have strong positive correlations with liver disease , showing that higher levels of them has a risk for having a liver disease while the negative coefficient for Albumin shows that lower levels of it increases the risk of having a liver disease.

References

Ramana,, & B. & Venkateswarlu. (n.d.). *ILPD (Indian Liver Patient Dataset) [Dataset].* UCI Machine Learning Repository. https://doi.org/10.24432/C5D02C