Indian Statistical Institute

203, B.T. Road, Kolkata - 700108, India

Predicting Liver Disease With Logistic Regression

Project Report

Statistical Methods IV

Submitted by:
Himadri Mandal
B.Stat II Year
Roll No: BS2327

Under the supervision of:
Dr. Ayanendranath Basu
Professor (HAG), ISRU
Applied Statistics Division
ISI Kolkata

Contents

Abstract

In this project, we use Logistic Regression to predict if patients have Liver Disease. We also look into the reliability of the and try to understand what parameters are the most important in the model.

1 Introduction

Liver disease affects millions worldwide and often goes undetected until it's advanced, making early risk-stratification essential. In this project, we use logistic regression—a straightforward yet powerful tool—to predict the presence of liver disease from routine clinical and laboratory measurements. After cleaning and exploring our patient dataset, we build and validate the model, then evaluate its accuracy, sensitivity and specificity. By highlighting the most influential predictors, our goal is not only to flag high-risk individuals but also to shed light on key factors driving liver-disease risk. The following sections detail our data, methods, results and take-home insights.

2 Data Exploration

2.1 Dataset

The Dataset was obtained from Kaggle [1]. The Dataset contains a '.csv' file with the following features:

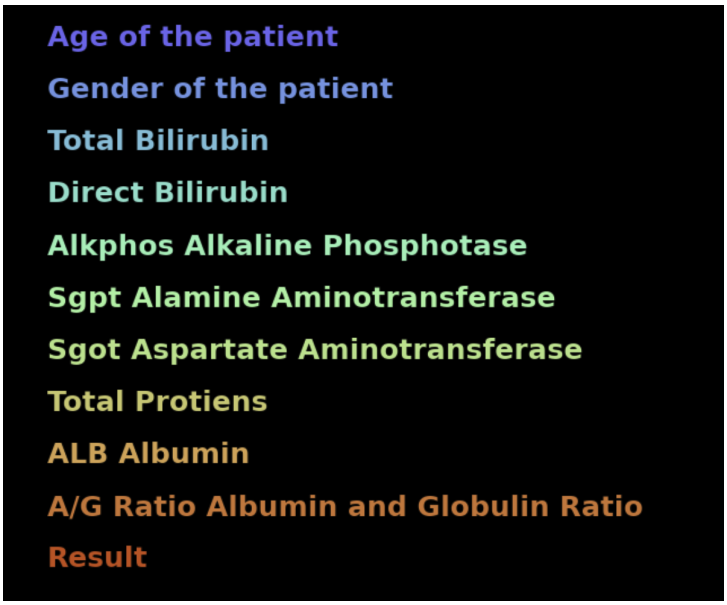

Figure 1: Features in the Dataset

- **Age of the patient**: The age of the patient, typically measured in years.
- **Gender of the patient**: The gender of the patient (male or female).
- **Total Bilirubin**: The total amount of bilirubin in the blood, which helps assess liver function.
- **Direct Bilirubin**: The portion of bilirubin that is directly excreted by the liver.
- **Alkphos Alkaline Phosphotase**: An enzyme found in the liver, bone, and other tissues; elevated levels may indicate liver disease.

- **Sgpt Alamine Aminotransferase**: An enzyme that helps in protein metabolism; higher levels may indicate liver injury.
- **Sgot Aspartate Aminotransferase**: Another enzyme found in liver cells, elevated levels are often associated with liver damage.
- **Total Proteins**: The total concentration of proteins in the blood, which helps evaluate liver function.
- **ALB Albumin**: A protein produced by the liver, important for maintaining blood volume and pressure.
- **A/G Ratio Albumin and Globulin Ratio**: The ratio of albumin to globulin in the blood, which can indicate liver or kidney issues.
- **Result**: The outcome of the test, indicating whether or not the patient has liver disease. Here 1 represents people who have liver disease diagnosed and 2 represent people who don't have liver disease diagnosis.

Because the $\{1, 2\}$ classification is a mess in **Result**, we make the values in $\{0, 1\}$. With 0 representing liver disease, and 1 not.

2.2 Understanding the Demographics of the Dataset

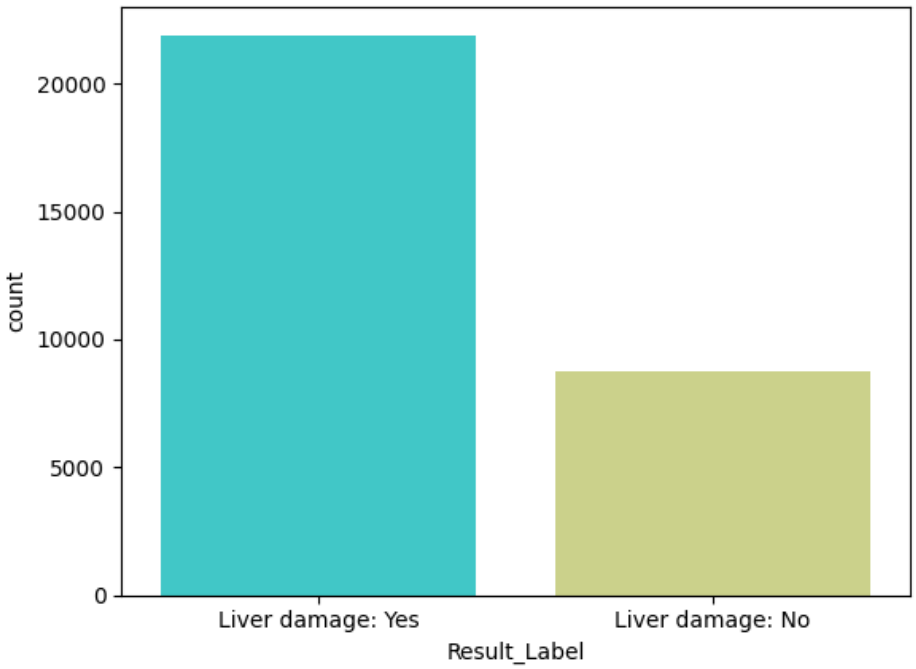

Figure 2: Patient Distribution

Understanding the demographics of the dataset helps us better understand and evaluate the results of the analysis. In this section, we plot 4 figures:

- **Patient Distribution**: A barplot showing the distribution of the patients across the gender dimension.
- **Categorical plot**: A Violin plot depicting the distribution of male and female Liver damage patients across age and gender.
- **Gender and Age**: Histograms depicting the distribution of the patients across age and gender.
- **Gender and Bilirubin**: Plot showing the relation between direct and total Bilirubin in the patients. A small trend is visible here.

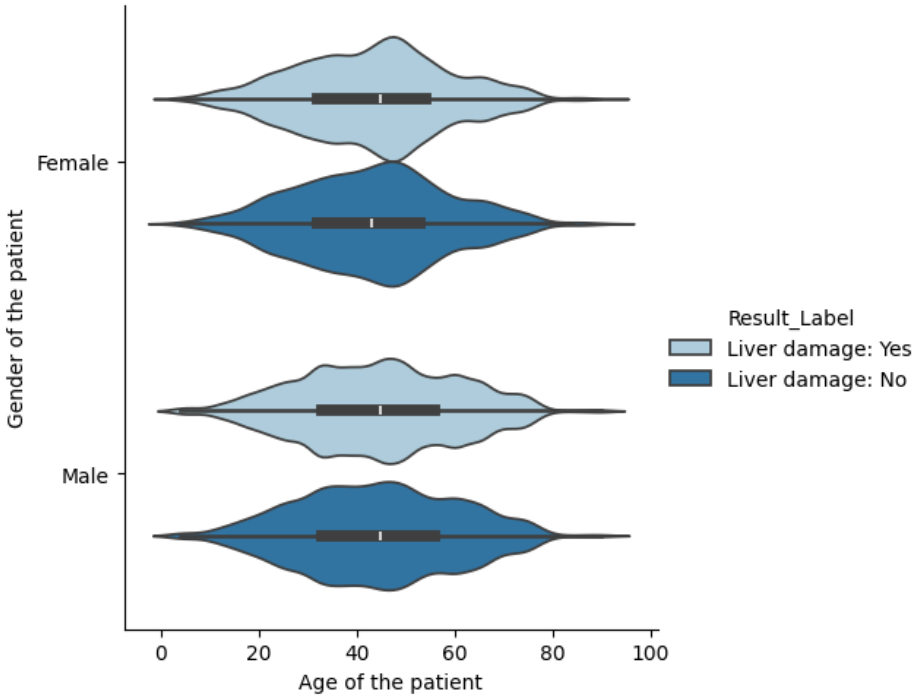

Figure 3: Categorical plot

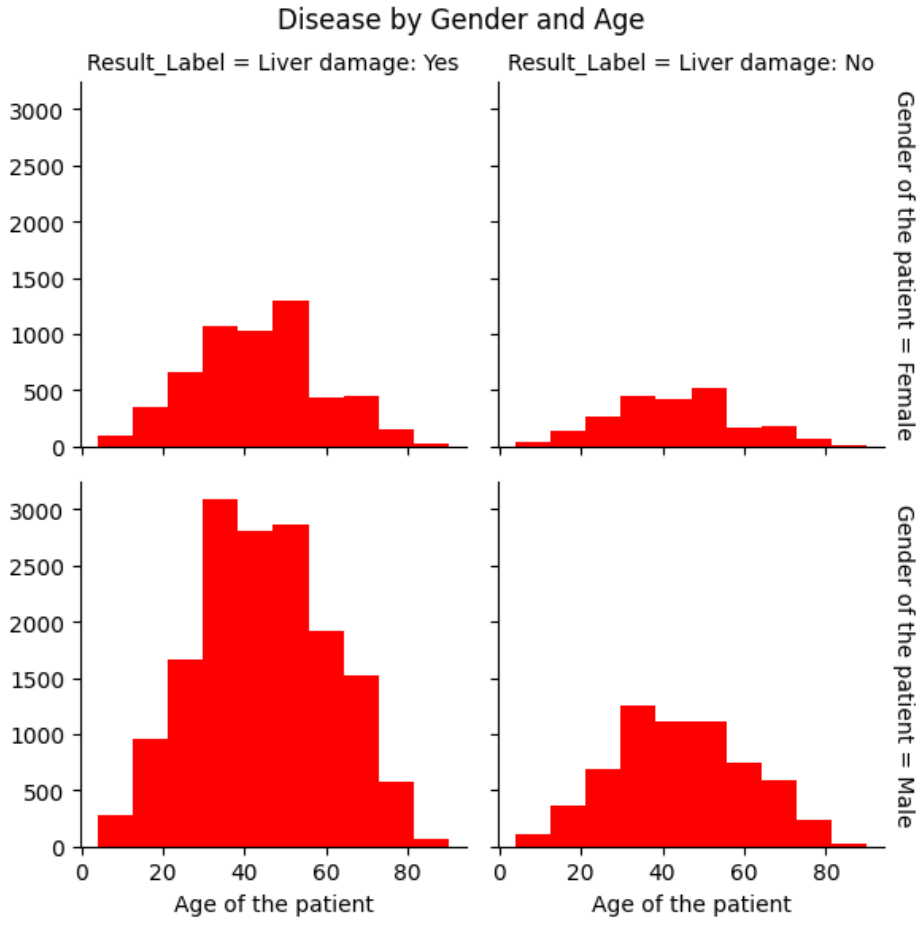

Figure 4: Gender and Age

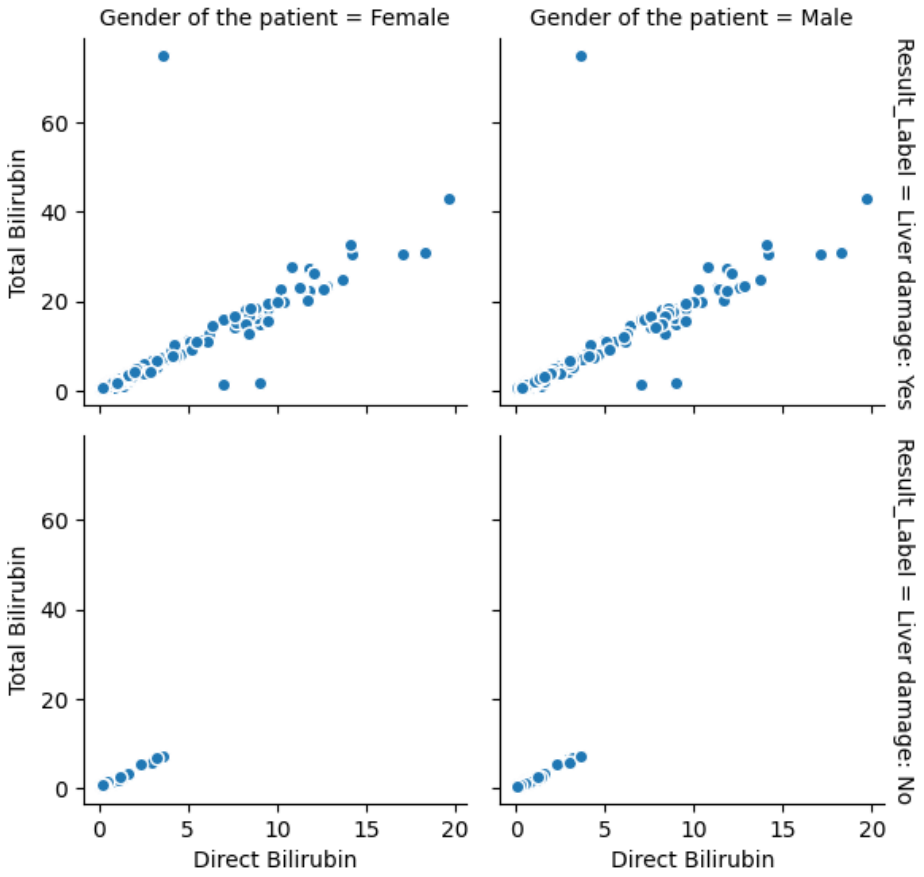

Figure 5: Gender and Bilirubin

2.3 Understanding the Relation between the Features

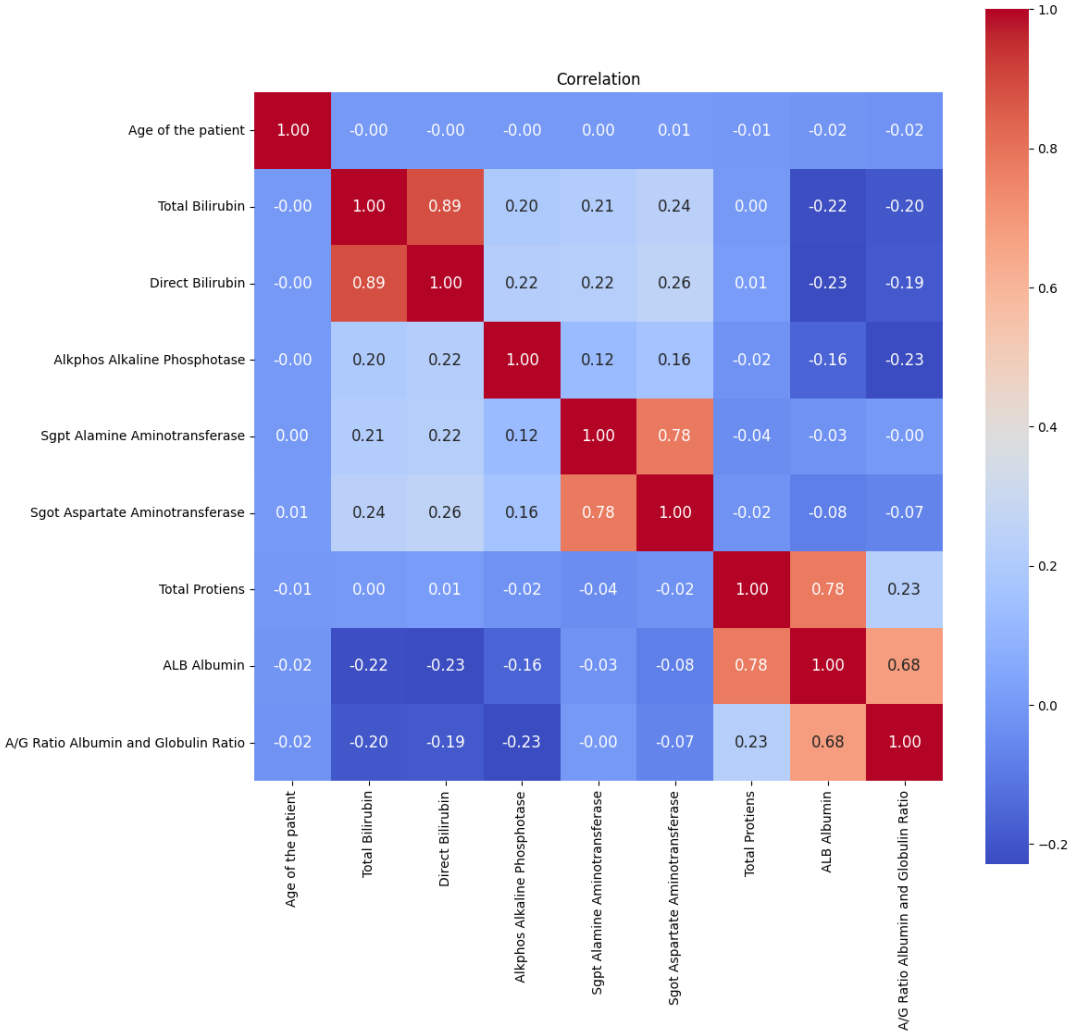

Figure 6: Correlation between the Features

From here we understand that although most features have insignificant correlations, there are a few highly correlated features:

- **Sgot Aspartate Aminotransferase** and **Sgot Alamine Aminotransferase**: 0.78
- **Total Proteins** and **Albumin**: 0.78
- **A/G Ratio** and **Albumin**: 0.68

This could be useful in further analysis.

3 Logistic Regression

3.1 Methodology

We use the **sklearn** library's **linear_model.LogisticRegression** class to create the predictive model. Here is a brief summary on the description of that class:

```
class sklearn.linear_model.LogisticRegression(penalty='l2', *, dual=False, tol
   =0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None,
    random_state=None, solver='lbfgs', max_iter=100, multi_class='deprecated',
    verbose=0, warm_start=False, n_jobs=None, l1_ratio=None)
```

Description

Logistic Regression (aka logit, MaxEnt) classifier. This class implements regularized logistic regression using the 'liblinear' library, 'newton-cg', 'sag', 'saga' and 'lbfgs' solvers. Note that regularization is applied by default. It can handle both dense and sparse input. Use C-ordered arrays or CSR matrices containing 64-bit floats for optimal performance; any other input format will be converted (and copied).

We divide the dataset into **X_train, X_test, y_train, y_test** where **X,y** represent the **feature** and **result** variables. And then use the **LogisticRegression** class to create a model. Then, we fit the model to **(X_train, y_train)**.

3.2 Analysis

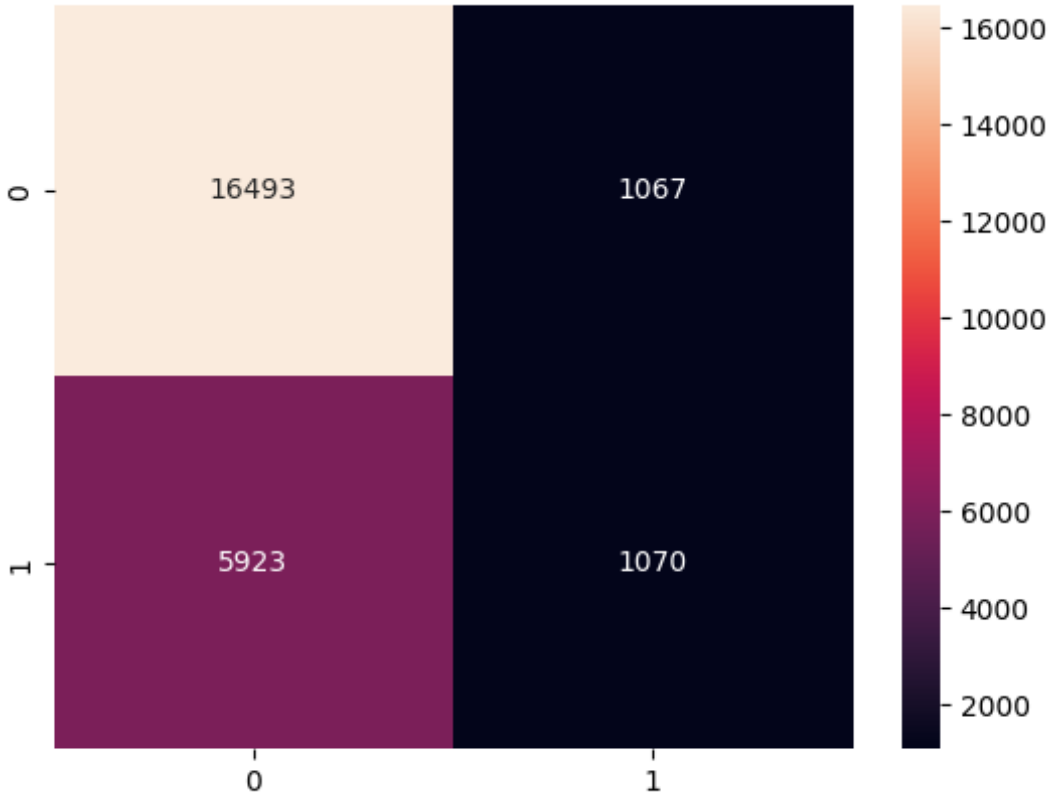

Figure 7: Confusion Matrix

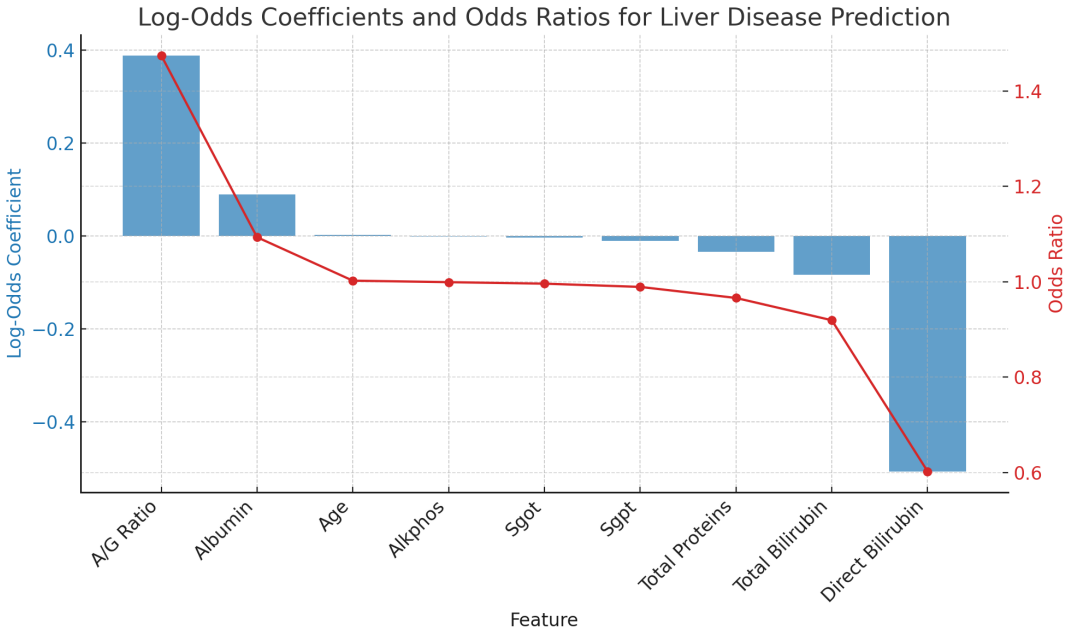

Figure 8: Odd Ratios

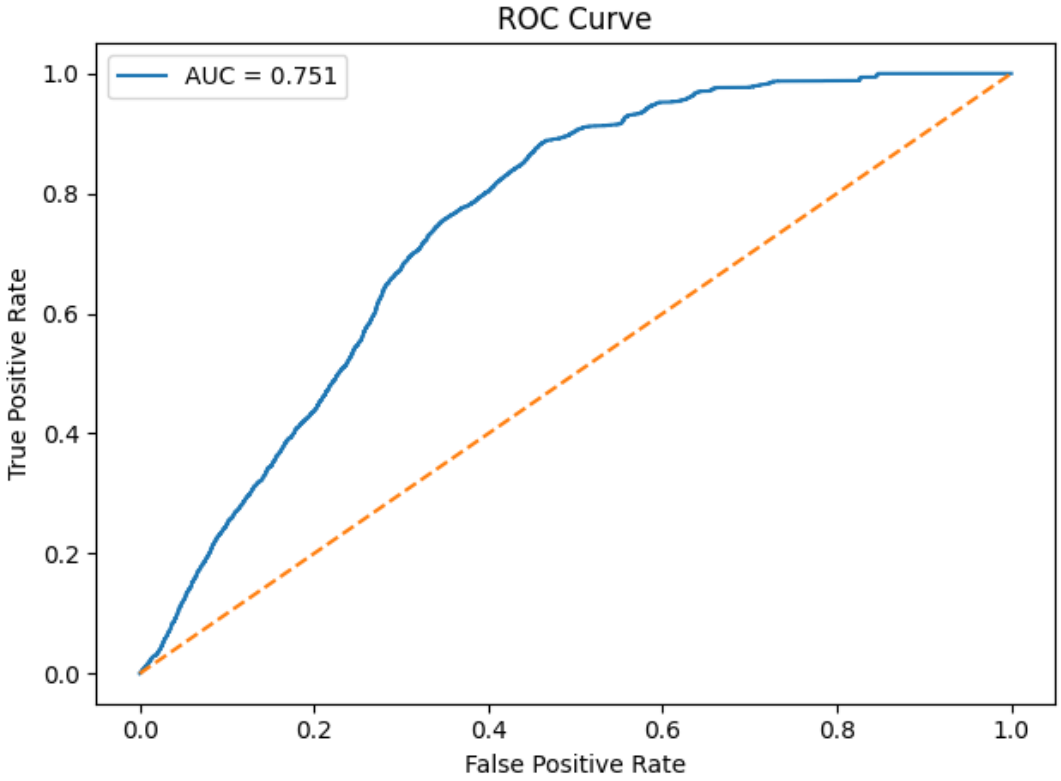

Figure 9: ROC Curve

4 Conclusion

4.1 Results

Upon testing the **Logistic** classifier on the **y_test**, we obtain the following results:

- **Accuracy** $= \dfrac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = \boxed{71.53\%}$
- **Sensitivity** $= \dfrac{\text{TP}}{\text{TP} + \text{FN}} = \boxed{15.30\%}$
- **Sensitivity** $= \dfrac{\text{TN}}{\text{TN} + \text{FP}} = \boxed{93.92\%}$

References

[1] Abhishek Shrivastava. *Liver Disease Patient Dataset*, 2022. https://www.kaggle.com/datasets/abhi8923shriv/liver-disease-patient-dataset. Accessed: 2025-04-01.