**2. EDA & Visualization Summary**

**🔹 EDA Objectives**

* Understand distribution of attributes
* Detect anomalies and outliers
* Identify relevant features for modelling

**🔹 Actions Taken**

| **Task** | **Description** |
| --- | --- |
| Missing Value Handling | Filled missing Albumin/Globulin ratios with mean |
| Encoding | Converted Gender (Male=1, Female=0) |
| Scaling | Normalized data for model input |

**🔹 Visualizations**

| **Chart Type** | **Feature(s) Visualized** |
| --- | --- |
| Histogram | Age, Bilirubin, Proteins |
| Correlation Heatmap | All numeric features |
| Count plot | Liver disease count by gender |
| Boxplot | Distribution of Albumin and A/G ratio |

**🔹 Key Insights**

* Patients with lower Albumin and A/G ratio often had cirrhosis
* Bilirubin levels also showed correlation with liver disease
* Dataset was balanced — no oversampling required

**🔹 Sample Insights**

* **Bilirubin Levels**: Cirrhosis-prone patients generally had **higher Total and Direct Bilirubin** levels.
* **Albumin and A/G Ratio**: These two were significantly **lower in cirrhotic patients**, making them key predictive features.
* **Alkaline Phosphotase and Aminotransferases**: Elevated values were common among diseased patients.
* **Gender Distribution**: Both genders were affected, with a **slightly higher representation of male patients** in the positive class.
* **Age Factor**: Middle-aged to older adults were more represented in the positive class.

**🔹 Conclusion**

* EDA played a critical role in:
* Improving data quality through cleaning
* Highlighting strong predictive features
* Deciding not to apply SMOTE/oversampling due to a relatively balanced dataset
* Informing decisions on which models to try first (Random Forest due to mixed feature types and interactions)