Exploratory data analysis

Descriptive Analysis :

1. Dataset Overview

- **Dataset Used:** Indian Liver Patient Dataset (ILPD)
- **Instances (Rows):** 583
- **Features (Columns):** 10 features + 1 target column
- **Target Variable:**
  - 1: Liver Patient
  - 0: Not a Liver Patient

2. Summary Statistics

Feature	Mean	Std Dev	Min	25%	50%	75%	Max
Age	44.75	16.19	4	33.0	45.0	60.0	90
Total_Bilirubin	3.29	6.98	0.4	0.8	1.0	2.6	75.0
Direct_Bilirubin	1.48	3.59	0.1	0.3	0.4	1.3	19.7
Alkaline_Phosphotase	290.5	240.2	63	175.0	208.0	312.5	2110
Alamine_Aminotransferase	80.71	182.0	4	18.0	29.0	64.0	2000
Aspartate_Aminotransferase	109.3	226.0	5	25.0	40.0	98.0	4929
Total_Proteins	6.48	0.91	2.7	5.9	6.6	7.2	9.6
Albumin	3.13	0.89	0.9	2.5	3.2	3.7	5.5
Albumin_and_Globulin_Ratio	0.94	0.38	0.3	0.7	0.9	1.1	2.8

3. Demographic Insights

a) Age Distribution

- Average age: ~45 years
- Range: 4–90 years
- Most liver patients are in the **30–60** age group.

b) Gender Distribution

df['Gender'].value_counts()

- Males: ~441
- Females: ~142
- Liver disease is **more common in males** in this dataset.

4. Biochemical Marker Distribution

a) Total Bilirubin

- Skewed distribution
- Higher values generally correlate with liver damage

b) Alkaline Phosphotase, SGPT, SGOT

- High variability (presence of outliers)
- These are key enzymes affected in liver malfunction

c) Albumin & Protein Levels

- Low values associated with liver cirrhosis
- Albumin is a useful predictor in model building

5. Class Distribution (Target)

df['Target'].value_counts(normalize=True)

- Liver Patients: **416 (~71%)**
- Non-Liver Patients: **167 (~29%)**

6. Key Observations

Observation	**Insight**
Age	Older individuals are more at risk
Gender	Higher prevalence in males
Enzymes	SGOT, SGPT, and ALP are significantly elevated in liver patients
Protein & Albumin	Reduced in liver disease
Class imbalance	Will affect model performance if not handled

7. Suggested Preprocessing Based on Analysis

- Impute missing values (especially in Albumin_and_Globulin_Ratio)
- Normalize skewed features (e.g., Total_Bilirubin)
- Encode gender
- Apply class balancing during model training