Exploratory Data Analysis Report

Project: Liver Cirrhosis Stage Prediction

Prepared by: Eman Abdallah Yosif Maharik
Dataset Source: Kaggle - Cirrhosis Prediction Dataset

---

1. Introduction

The goal of this analysis is to explore and understand the structure, quality, and patterns in a dataset related to liver cirrhosis patients. The ultimate aim is to support building predictive models to classify the stage of liver cirrhosis.

---

2. Data Loading and Description

The dataset was loaded using pandas and is assumed to be in CSV format named liver_cirrhosis.csv.

Initial Exploration:

- **Shape of dataset**: The dataset contains *n rows* and *m columns* (exact numbers parsed from the data).
- **Basic Info**: Displayed data types and null counts per column.
- **Sample data**: df.head() was used to preview the first 5 rows.

---

3. Data Cleaning

Steps Taken:

- Checked and renamed column names for readability and consistency.
- Handled missing values by identifying them and taking suitable actions (e.g., imputation or row dropping).
- Verified and corrected inconsistent or invalid data entries if any.

Missing Values:

- Visualized using heatmap and bar charts.
- Columns with high null percentages (e.g., over 40-50%) may have been removed or treated.

---

4. Univariate Analysis

Examined individual variables to understand distributions, types, and unusual behavior.

Numerical Columns:

- Histograms and KDE plots were used to explore distributions (e.g., Age, Albumin, Bilirubin).
- Summary statistics included: mean, median, standard deviation, and range.

Categorical Columns:

- Bar plots and value counts used for variables such as Gender, Stage, Drug type.
- Checked balance or imbalance in the target variable (Cirrhosis Stage).

---

5. Bivariate & Multivariate Analysis

Categorical vs Target:

- Used count plots and stacked bar charts to compare features like Sex, Drug, or Ascites against the cirrhosis stage.

Numerical vs Target:

- Boxplots and violin plots used to compare distributions of numeric features across cirrhosis stages.
- Observed potential trends or thresholds (e.g., Albumin decreasing with stage severity)

---

6. Correlation Matrix

A heatmap using Pearson correlation was plotted to visualize the relationship between numeric variables:

- Positively or negatively correlated features with Stage were noted.

- Helped in identifying multicollinearity or important predictors.

---

7. Outlier Detection

Outliers were analyzed using:

- Box plots
- Z-score or IQR method

Features Analyzed:

- Bilirubin
- ALT/AST
- Prothrombin

Outliers were either retained, capped, or removed depending on their impact.

---

8. Missing Value Treatment

- Features like Prothrombin, Cholesterol, or Albumin had some missing values.
- Techniques applied:
  - Mean/Median imputation
  - KNN imputation (if implemented)
  - Row deletion if values were excessive

---

9. Feature Engineering

(if applied, otherwise skip)

- Derived features or transformations might include:
  - Binning Age
  - Creating binary indicators for lab values (e.g., high/low bilirubin)

---

10. Conclusions & Recommendations

Key Observations:

- Several lab test variables (e.g., Albumin, Bilirubin, AST) show strong correlation with cirrhosis stages.
- Stage 4 (Severe cirrhosis) has distinct patterns compared to earlier stages.
- Some features had high missing values and should be considered cautiously for modeling.

Recommendations:

- Consider feature selection or dimensionality reduction techniques.
- Normalize/standardize numerical values before modeling.
- Apply techniques to handle class imbalance if it exists in the target variable.