Liver Cirrhosis Stage Detection - Project Report

# 1. Introduction

Liver cirrhosis is a chronic liver disease marked by degeneration of liver cells, inflammation, and fibrous thickening of tissue. Early and accurate detection of cirrhosis stages is critical for timely intervention. This project aims to build a machine learning system to predict the histologic stage of liver cirrhosis (Stage 1, 2, or 3) based on clinical and biochemical attributes from patient data.

# 2. Problem Statement

The goal is to develop a predictive model that can determine the stage of liver damage using various features from a dataset collected by the Mayo Clinic between 1974 and 1984. The system should accept patient information and return the predicted cirrhosis stage.

# 3. Dataset Overview

The dataset consists of medical records from a study on primary biliary cirrhosis (PBC). Key features include:

* - Demographic: Age, Sex  
  - Clinical symptoms: Ascites, Hepatomegaly, Spiders, Edema  
  - Biochemical markers: Bilirubin, Cholesterol, Albumin, Copper, Alk\_Phos, SGOT, Tryglicerides, Platelets, Prothrombin  
  - Outcome variables: Status, Drug, N\_Days, Stage

The target variable is Stage, with possible values:  
- Stage 1: Mild  
- Stage 2: Moderate  
- Stage 3: Severe

# 4. Data Preprocessing

The following preprocessing steps were applied in the notebook:

* - Unnecessary columns such as 'id' and 'status' were dropped  
  - Missing values were handled by removing incomplete records  
  - Categorical features (like sex, ascites, hepatomegaly, spiders, and edema) were encoded to numeric values  
  - Feature scaling was applied using StandardScaler to normalize the data

# 5. Exploratory Data Analysis (EDA)

The dataset was explored using visualizations and statistical methods. Count plots helped identify class distribution in the target variable. Distribution plots were used to understand the spread of key features like Bilirubin, Albumin, and Prothrombin. A heatmap of correlations revealed strong associations between certain features and disease stage.

# 6. Model Building and Evaluation

A Random Forest classifier was used to predict the stage of liver cirrhosis. The dataset was split into training and testing sets using an 80/20 ratio. The model was trained on the processed features, and its performance was evaluated using accuracy score, confusion matrix, and classification report. The Random Forest model achieved high accuracy and was able to correctly classify the stages in most cases.

# 7. Conclusion

This project successfully built a pipeline to predict liver cirrhosis stages using a Random Forest classifier. The model performed well after data cleaning, encoding, and scaling. Important features such as Bilirubin, Albumin, and Prothrombin played a significant role in predicting the correct stage of the disease.

# 8. Future Work

* - Tune hyperparameters of the Random Forest model to improve performance  
  - Explore feature importance and conduct deeper clinical analysis  
  - Consider deploying the model as a web-based application for healthcare providers  
  - Address potential class imbalance with techniques such as SMOTE