Predicting Histological Stage of Disease for Cirhosis

Yiyu Yao, Chuan Lin, Lena Wang, Matt Wang

1 Problem

This project aims to create a predictive model that accurately estimates the stage of Cirrhosis based on factors such as age, drug administration, cholesterol levels, and various other health factors. The model will help analyze 1). the effectiveness of the drug D-penicillamine (whether drug administration is a feature that would significantly impact the prediction) and 2). the relationship between other health factors and the Cirrhosis disease.

2 Importance

Cirhosis of the liver can be a life-threatening disease and is caused by long-term liver damage. Being able to identify whether D-penicillamine administration highly impacts the stage of Cirrhosis provides insight on whether this could be a potential effective medication. Additionally, liver biopsy and imaging are common methods for staging cirrhosis, but they are invasive abd expensive. A predictive model based on non-invasive data could reduce the need for invasive procedures by providing a reliable and convenient alternative

3 Dataset

The "Cirrhosis Patient Survival Prediction" dataset from UCI Machine Learning Repository (link) is readily accessible and offers a detailed view of health and medical factors from 312 patients, supporting immediate use in this analysis. Key variables include drug administration and presence of other symptoms, like ascites and spiders, days after drug administration, and other health factors like cholesterol levels.

Preprocessing will be necessary to handle several aspects of data quality. First, missing data points across patients will need to be addressed. It is likely that the data are not missing completely at random. To deal with missing data, we could employ methods like inverse weighting, imputation or maximum likelihood inference to ensure a complete dataset for analysis. Additionally, the dataset includes a mix of numerical and categorical data types, requiring encoding for categorical variables. Additionally, this is a relatively small dataset (only 312 rows with 18 features), which makes it a challenge to learn the complexities behind the features. To make the most use of the data, we could try to use cross-validation when training and testing. Bagging can also be applied by sampling with replacement, where the models outlined below would be less prone to outliers. To select features that impacts the prediction the most, we could employ feature ranking methods: (Reference to a research that did something similar: link).

4 Method and Model

Prior to training, we will randomly sample out our testing data. We will also perform exploratory data analysis (EDA), feature ranking methods, and cross-validation to identify the most relevant features. For prediction, we will employ multiple models, including Logistic Regression, Decision Trees, and Multilayer Perceptron (MLP). By combining these models into a hybrid approach, we aim to identify the configuration that achieves the highest predictive accuracy.

1. Logistic Regression will serve as a baseline model due to its simplicity and interpretability. As a baseline, it establishes a reference for other model performance.

2. Decision trees will be used to capture complex relationships without making assumptions about data distribution. To further enhance predictive accuracy, we may also incorporate random forests as an ensemble of decision trees.

3. MLP: The MLP will be used to capture complex, non-linear patterns between the features and Cirrhosis stage. By tuning hyperparameters, we will explore MLPs with multiple hidden layers, selecting the configuration with the highest accuracy to contribute to the hybrid model.