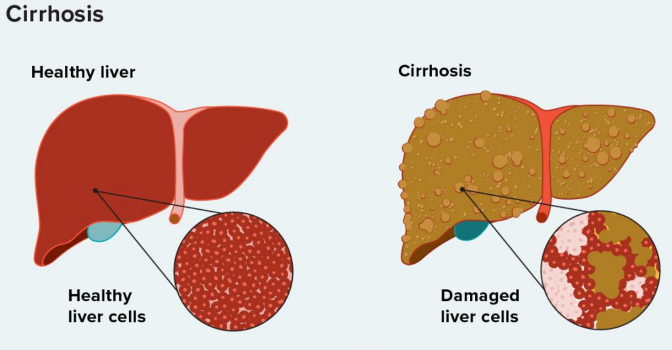
Cirrhosis Prediction

By: Justice Mansfield-Beaulieu

Cirrhosis Background

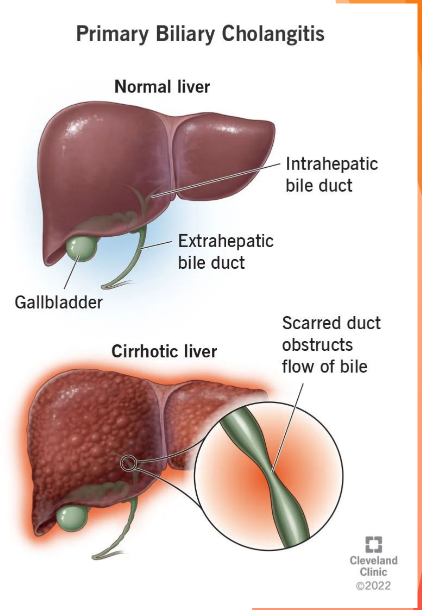
- What is Cirrhosis?
  - According to the National Institute of Health Cirrhosis is a chronic disease of the liver which your liver is scared and permanently damaged. When having cirrhosis, scar tissue replaces healthy liver tissue and prevents the liver from working normally. In later stages of cirrhosis the liver begins to fail.

- What causes Cirrhosis, and what are the symptoms?
  - The National Institute of Health also mention common causes is cirrhosis is alcohol abuse, fat build up on the liver, and chronic hepatitis C or B.
  - Symptoms in the early stages include; fatigue, poor appetite, weight loss, nausea/vomiting and discomfort or pain in abdomen.
  - In the later stages symptoms include; jaundice, darken color urine, severe itchy skin, ascites, edema, bruising/bleeding easily and memory loss.



Project Overview

- The purpose of the project is to predict different stages of cirrhosis based on data gather in the Mayo Clinic trial in primary biliary cirrhosis. The data gather in the trail was gathered for 10 years between 1974 and 1984. A total of 424 patients participated in this trial were randomly selected to be treated with a placebo or D-penicillamine, which is a drug used to treat biliary cirrhosis.



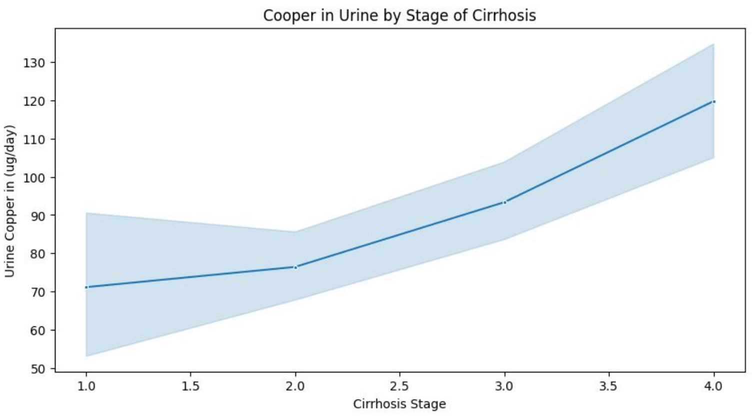
Data Description

- This data is from the Cirrhosis Prediction Dataset on Kaggle from owner: Fedesriano
- This dataset includes 20 features for predicting Cirrhosis stage.

N_Days: number of days between registration and the earlier of death, transplantation, or study analysis time in July 1986	**Cholesterol**: serum cholesterol in [mg/dl]
Status: status of the patient C (censored), CL (censored due to liver tx), or D (death)	**Albumin**: albumin in [gm/dl]
Drug: type of drug D-penicillamine or placebo	**Copper**: urine copper in [ug/day]
Age: age in [days]	**Alk_Phos**: alkaline phosphatase in [U/liter]
Sex: M (male) or F (female)	**SGOT**: SGOT in [U/ml]
Ascites: presence of ascites N (No) or Y (Yes)	**Triglycerides**: triglycerides in [mg/dl]
Hepatomegaly: presence of hepatomegaly N (No) or Y (Yes)	**Platelets**: platelets per cubic [ml/1000]
Spiders: presence of spiders N (No) or Y (Yes)	**Prothrombin**: prothrombin time in seconds [s]
Edema: presence of edema N (no edema and no diuretic therapy for edema), S (edema present without diuretics, or edema resolved by diuretics), or Y (edema despite diuretic therapy)	**Stage**: histologic stage of disease (1, 2, 3, or 4)
Bilirubin: serum bilirubin in [mg/dl]	**ID:** unique identifier

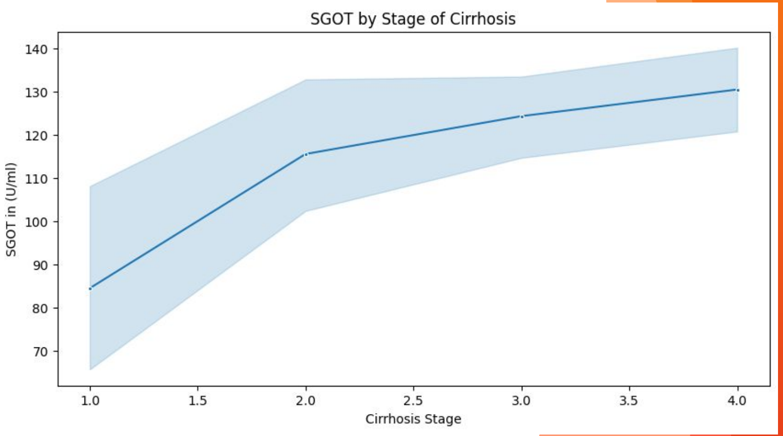
Key Findings

- Impact of Copper in Urine by microgram per day-
  - Copper had the highest correlation to our target. Looking at this graph we can see that as the stage of cirrhosis increases so does the amount of copper the patient has in their urine per day.
  - Once getting around levels of 95-115ug the patients are at stage 3 and 4, and levels around 70-75ug patients are at stage 1 nd 2

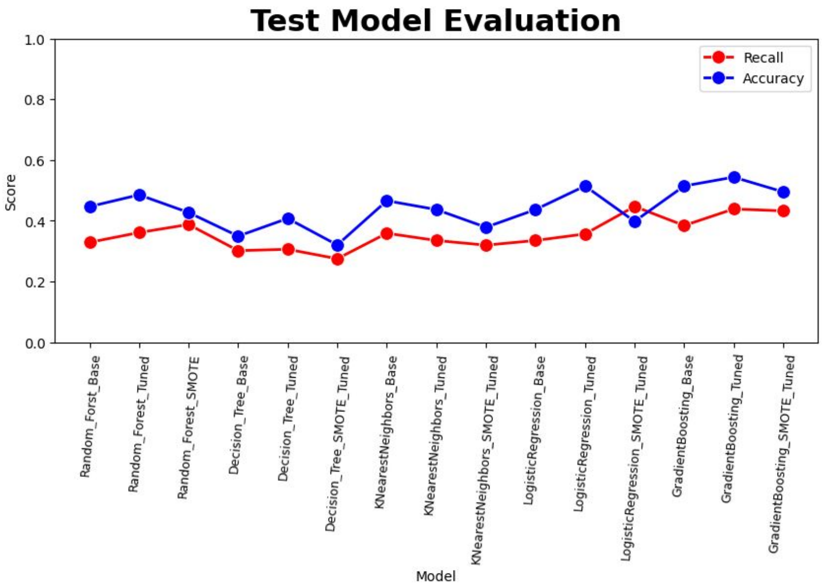


Key Findings

- Impact of serum glutamic-oxaloacetic transaminase(SGOT) per milliliter-
  - Levels of SGOT around 115 and lower will likely have stage 1 or 2 of cirrhosis. Those with levels around 125ml and up with likely have stage 3 or 4 cirrhosis. Higher the SGOT levels the higher the stage of the disease.



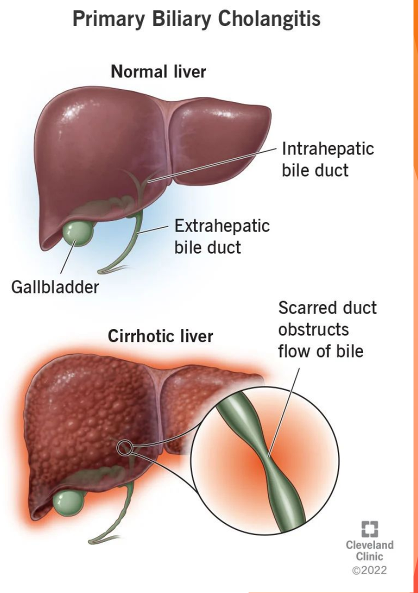
Model Evaluation



- The two metrics for evaluating our model are accuracy and recall. I chose these metrics because accuracy lets us know the correct predictions. Also choosing recall because we want minimize our false negatives(cirrhosis stages predicting at the wrong stage).
  - Gradient Boosting Tuned model
    - This model gave the best accuracy at 54% with a 43% recall which is only 1% off of the highest recall with the LogisticRegression SMOTE Tuned model.

Summary

- When predicting any form of disease we want to always make sure we handle our false negatives, in our case these types of errors could decrease the effectiveness to predict which patient has what stage of cirrhosis.
- Adding on, the false positive as well could be costly due to wasting resources on treating patients of lower stages of cirrhosis with the effort of those with higher stages and vice versa.
- Although Gradient Boosting Tuned model was out best model it was extremely overfit and didn't give the best results. In this case It would be best to explore more classification model and ways to tune them. Also, reverting to regression models because this dataset can be used those models as well and see if we will get better results.



Thanks!

Any questions?