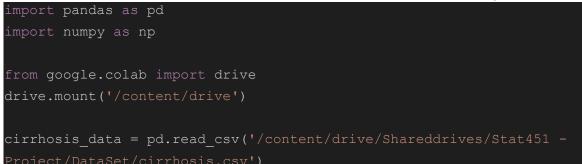
Group 21 - Cirrhosis: Proposal

Cirrhosis Patient Survival Prediction (kaggle.com)

Read Data:

The <u>dataset</u> is going to be analyzed using Google Colab implementation of the iPython (jupyter notebook) environment import. This allows us to read the .csv file from Google Drive.



Questions:

- 1. Which of the features best predicts the stage of the cirrhosis?
 - This allows us to see which one of the variables has a bigger effect on the stage of the cirrhosis.
- 2. What kind of regression model best fits this data?
 - With this we can see the which model can be used in the future for novel samples
- 3. Does the drug have any effect on the status of the patient? Does it improve their condition?
 - The intention of this question is to analyze the effect of the drug implemented in the trials.

Methods:

- Feature selection: we will determine which features best predict the cirrhosis stage in patients
- Classification (kNN, decision tree, logistic regression).
- The dataset contains missing values, so we will use different data imputation techniques to find if these methods improve the prediction accuracy or worsen it.

Variable	Description	Data Type
ID	Unique identifier	Integer
N_Days	Num days between registration and the earlier or death, transplantation	Integer
Status	C(Censored),CL(Censored due to liver tx),D(death)	Categorical
Drug	type of drug D-penicillamine or placebo	Categorical
Age	age	Integer
Sex	M (male) or F (female)	Categorical
Ascites	presence of ascites N (No) or Y (Yes)	Categorical
Hepatomegaly	presence of hepatomegaly N (No) or Y (Yes)	Categorical
Spiders	presence of spiders N (No) or Y (Yes)	Categorical

Variables:

Edema	Presence of edema N (no edema and no diuretic therapy for edema) S (edema present w/o diuretics or edema resolved by diuretics) Y (edema despite diuretic therapy)	Categorical
Bilirubin	Serum bilirubin [mg/dl]	Continuous
Cholesterol	Serum cholesterol [mg/dl]	Integer
Albumin	Albumin [mg/dl]	Continuous
Copper	Urine copper [ug/day]	Integer
Alk_Phos	Alklaline Phosphatase [u/liter]	Continuous
SGOT	SGOT [u/ml]	Continuous
Tryglicerides	tryglicerides	Integer
Platelets	platelets per cubic	Integer
Prothrombin	prothrombin time	Continuous
Stage	histologic stage of disease (1, 2, 3, or 4)	Categorical