Predicting Hepatitis C: Revealing Patterns in Liver Panel Tests for Hep C Diagnosing and Cirrhosis Staging

Springboard Data Science
Capstone Two
Naomi Lopez

Problem Identification

Context

Hepatitis C stands as a prominent contributor to liver failure on a global scale. Despite the availability of a cure in the market, its accessibility remains constrained by its exorbitant cost. The CDC advocates for a singular, lifelong screening for individuals aged 18 and above, with the option for more frequent screenings if specific risk factors are present. Transmission of Hep C primarily occurs through contact with an infected person's blood, including practices such as intravenous and intranasal drug use, as well as blood transfusions.

Initial screening for Hep C relies on Hepatitis C Antibody testing, while the definitive confirmation of infection is accomplished through Hep C RNA testing, considered the gold standard. Timely and accurate screening is paramount in identifying Hep C early, thus averting the progression to terminal liver failure and curbing further transmission. This not only significantly impacts patient outcomes but also serves to mitigate the spread of infection within communities.

In instances where screening is overlooked or incorrectly conducted during healthcare visits, a liver panel test can offer valuable indications of potential Hep C infection. This prompts healthcare providers to initiate Hep C antibody screening, ensuring timely detection and intervention.

Criteria for success

Using pattern recognition from blood work, a model can forecast whether a patient is positive for Hepatitis C and ascertain their current level of cirrhosis, ranging from stages 1 to 3..

Scope of solution space

Data analysis using python will be conducted to compare the liver panel blood work of 615 patients.

Constraints

The data set is small with only 615 patients and it does not include lifestyle indicators such as number of alcoholic beverages consumed per week, BMI, and history of congenital or paroxysmal liver disease. All of the stated can mimic elevated liver enzymes as seen in patients with hepatitis.

Stakeholders

Medical providers such as Doctors, Nurse Practitioners, Physician Assistants
Directors of governmental agencies such as the NIH and CDC
Biomedical companies specializing in Hepatitis C treatment

Key data sources
Hepatitis C dataset from Kaggle

Methods

Data Wrangling

- Clean the data by transforming the csv into a readable data frame using Pandas.
- Assess the number of Null or NA values in data and determine if columns or rows need to be deleted.

Exploratory Data Analysis

- Utilize PCA to assess variation among variables.
- Implement a heat map to identify highly correlated variables.

Preprocessing & Training

- Create a linear regression and use cross validation and test splitting to see how it performs
- Generator a random forest regressor and compare how it does in comparison to the linear regression

Modeling

- Model the trained algorithm to determine if it can accurately predict Hep C diagnosis and stage of cirrhosis

Deliverables

- Written documentation of the Capstone
- A presentation designed for stakeholders in order to convey results