**Multi-Class Prediction of Cirrhosis Outcomes.**

**Developed by**

Yash Raj

B.Tech CSE 2024

Central University of Haryana.

Contact no: +91-7644849436

Email: [yashraj3376@gmail.com](mailto:yashraj3376@gmail.com)

*Kaggle Playground Series - Season 3, Episode 26*

**Objective:**

My task is was to use a multi-class approach to predict the outcomes of patients with Cirrhosis*. (* A degenerative disease of the liver resulting in scarring and liver failure.)

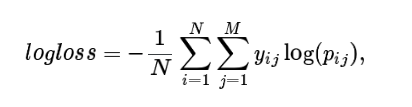
**Problem statement : Cirrhosis Patient survival prediction.**

Develop a machine learning model using 17 clinical features to predict the survival state of patients with liver cirrhosis. The survival states, categorized as 0 = D (death), 1 = C (censored), and 2 = CL (censored due to liver transplantation), will be inferred from data sourced from a Mayo Clinic study on primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984.

**Evaluation :**

Submissions are evaluated using the multi-class logarithmic loss. Each id in the test set had a single true class label, Status. For each id, you must submit a set of predicted probabilities for each of the three possible outcomes, e.g., Status\_C, Status\_CL, and Status\_D.

The metric is calculated



where **N** is the number of rows in the test set, **M** is the number of outcomes (i.e., 3), **log** is the natural logarithm, **yij** is 1 if row **i** has the ground truth label **j** and 0 otherwise, and **Pij** is the predicted probability that observation **i** belongs to class **j**.

The submitted probabilities for a given row are not required to sum to one because they are rescaled prior to being scored (each row is divided by the row sum). In order to avoid the extremes of the log function, predicted probabilities are replaced with

**max( min( p , 1-10-15 ) , 10-15 ).**

**Dataset Description :**

The Liver Cirrhosis Survival Prediction dataset comprises clinical features gathered from a Mayo Clinic study conducted between 1974 and 1984. This dataset aims to facilitate the prediction of survival outcomes for patients diagnosed with liver cirrhosis, specifically focusing on primary biliary cirrhosis (PBC).

Dataset: <https://www.kaggle.com/competitions/playground-series-s3e26/data>

*Purpose :*

The dataset was created to investigate the survival outcomes of patients with liver cirrhosis, particularly those diagnosed with primary biliary cirrhosis. Liver cirrhosis, characterized by extensive scarring resulting from prolonged liver damage, is often associated with conditions such as hepatitis or chronic alcohol consumption.

*Funding :*

The creation of this dataset was funded by the Mayo Clinic.

*Instances :*

Each instance in the dataset represents an individual patient diagnosed with liver cirrhosis, primarily focusing on those with primary biliary cirrhosis (PBC).

*Additional Information :*

During the study period, a total of 424 PBC patients referred to the Mayo Clinic qualified for the randomized placebo-controlled trial involving the drug D-penicillamine. Out of these, 312 patients initially participated in the trial and provided comprehensive data. The remaining 112 patients did not participate in the clinical trial but consented to record basic metrics and undergo survival tracking. However, six of these patients became untraceable after diagnosis, leaving data available for 106 individuals in addition to the 312 who were part of the randomized trial.

*Files :*

**train.csv**: The training dataset containing clinical features and survival status information.

Features: 17 clinical features used for prediction.

Target: 'Status' is the categorical target variable indicating the survival status of patients.

Target Categories:

* C: Indicates the patient was alive at N\_Days (censored).
* CL: Indicates the patient was alive at N\_Days due to liver transplantation.
* D: Indicates the patient was deceased at N\_Days.

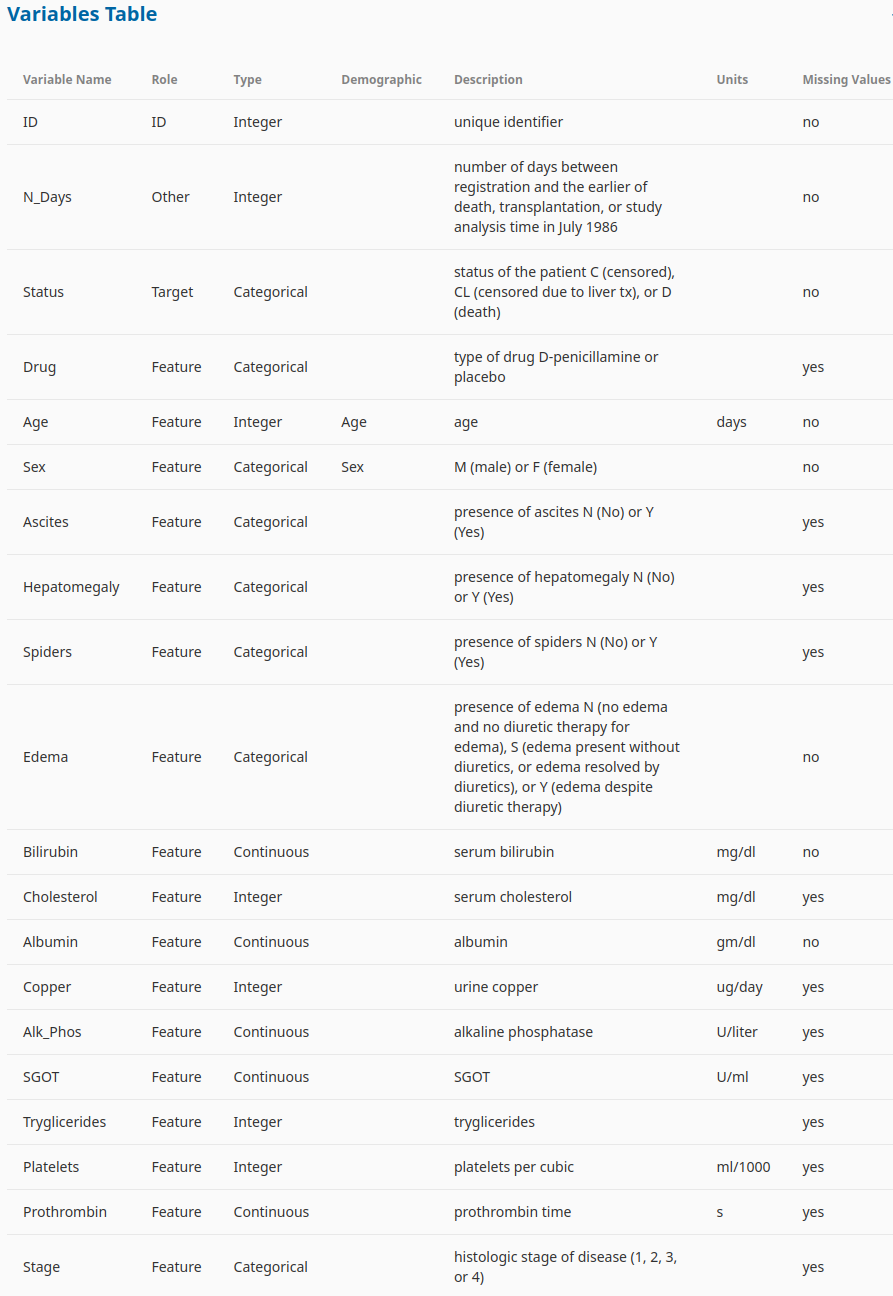
Note: N\_Days represents the duration of survival.

**test.csv**: The test dataset for predicting the probability of each status value.

Columns: Same as train.csv without the 'Status' column.

**sample\_submission.csv**: A sample submission file in the correct format.

Columns: 'Patient\_ID': Unique identifier for each patient in the test dataset. 'Status\_C', 'Status\_CL', 'Status\_D': Predicted probabilities for each status category.



**Citation:**

Walter Reade, Ashley Chow. "Multi-Class Prediction of Cirrhosis Outcomes." Kaggle, 2023. [Online]. Available: <https://kaggle.com/competitions/playground-series-s3e26>

**Approach or methodology for the given problem:**

This project adopts a systematic approach to develop a predictive model for classifying cirrhosis, a serious liver condition, using machine learning technique.

The methodology begins with data preprocessing, where the dataset undergoes cleaning and transformation to ensure its suitability for modeling.

Feature engineering techniques are then employed to extract relevant information from the dataset and create new informative features.

Subsequently, feature selection methods are applied to identify the most predictive features while eliminating redundant ones, optimizing the model's performance.

Next, a variety of machine learning algorithms, including XGBoost, LightGBM, and CatBoost, are considered and trained on the selected features to learn the intricate patterns within the data. Ensemble modeling is utilized to combine the predictions from multiple models, enhancing predictive accuracy and robustness.

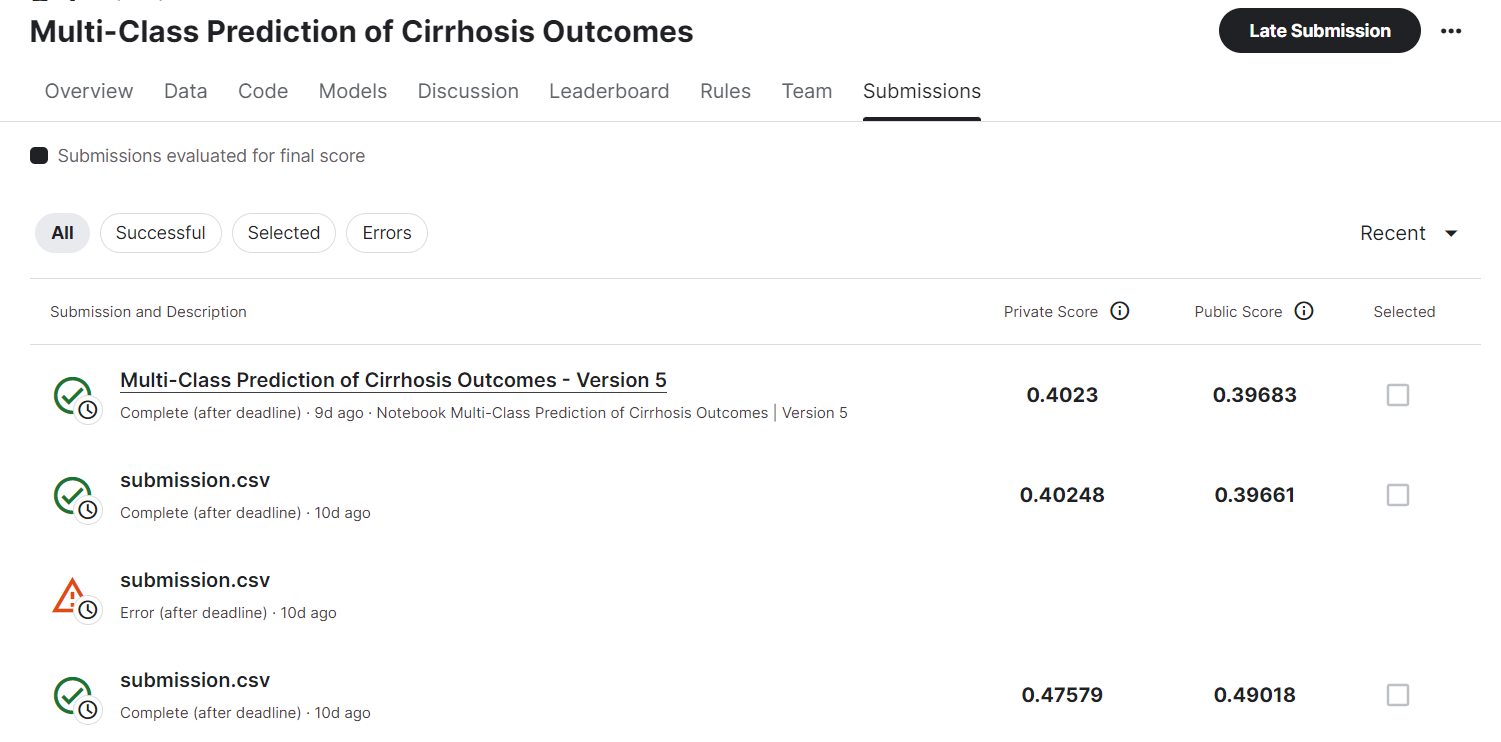
Rigorous evaluation, including cross-validation and hyperparameter tuning, ensures the model's reliability and generalization capability.

Finally, the model is tested on unseen data to assess its performance and generalization ability, aiming to provide clinicians with a valuable tool for early cirrhosis detection and intervention.

Through this approach, the project endeavours to contribute to improved healthcare outcomes and patient well-being.

**Achievement:**

Achievement is evidenced by the consistently high accuracy scores achieved during model evaluation, indicating the model's ability to correctly classify cirrhosis cases. Additionally, my models have demonstrated low log loss values, signifying minimal errors in probability estimates. Submission :



**Code:**

<https://www.kaggle.com/code/yashraj121/multi-class-prediction-of-cirrhosis-outcomes/notebook>

(Kaggle Notebook)

<https://drive.google.com/file/d/1w8Gf056bZZJTWjJX-eReGrYCQotaK20E/view?usp=sharing>

(pdf file)

<https://drive.google.com/drive/folders/1Fj9uYJvJq_U1E4eKytCJPEukG3CfayWB?usp=sharing>

(complete code)

**My resume:**

<https://drive.google.com/file/d/1DUSdiUajEbcKYzUEsMiggBvFlobJW6jZ/view?usp=sharing>