**Multi-Class classification of Liver Cirrhosis Outcomes**

Project Report By:

Ashwanth - AXR230028 Abishek - AXB220123 Damian Thomas - DXT230012 Thirunarayanan - TXR220025 Vinayak - VXS230023

**Table of Contents**

[Introduction 2](#_bookmark0)

[Problem Statement 2](#_bookmark1)

Data Description.............................................................................................................. 2

[Data Exploration and Preprocessing](#_bookmark5) 3

Baseline Models: Auto ML.............................................................................................. 8

Model Building................................................................................................................ 8

Model Evaluation.......................................................................................................... 10

Conclusion................................................................................................................................. 10

**Introduction**

Liver cirrhosis, a global health challenge, entails liver function decline due to scar tissue formation. Outcomes range from stable conditions to severe complications like liver failure or hepatocellular carcinoma. Predicting these outcomes is crucial for patient care, treatment planning, and resource allocation. The Multiclass Cirrhosis Outcome Classification project applies machine learning to predict diverse outcomes, using patient data like demographics, medical history, and imaging. This aids in stratifying patients and developing tailored interventions. By leveraging predictive models, healthcare providers can adjust treatments, enhance clinical decisions, and improve patient outcomes, ultimately advancing liver disease management.

**Problem Statement**

Problem statement: Liver cirrhosis is a challenge to predict outcomes, from stable conditions to severe complications accurately. Current prognostic methods often lack precision and do not consider the disease's diversity. The project aims to develop reliable predictive models using multiclass classification methods to classify patients with cirrhosis into different outcome categories. Challenges include handling heterogeneous patient data, ensuring model generalizability, and maintaining clinical interpretability. Successful implementation will improve the management, resource allocation, and clinical decision-making of patients with cirrhosis.

**Data Description**

This dataset was synthetically generated using a DL model and provided by Kaggle as a part of the 2023 Playground Series competition.

No. of records

* Train:7905
* Test: 5271

No. of attributes: 20

* Numeric: 12
* Categorical: 8

The target variable Status contains the following labels:

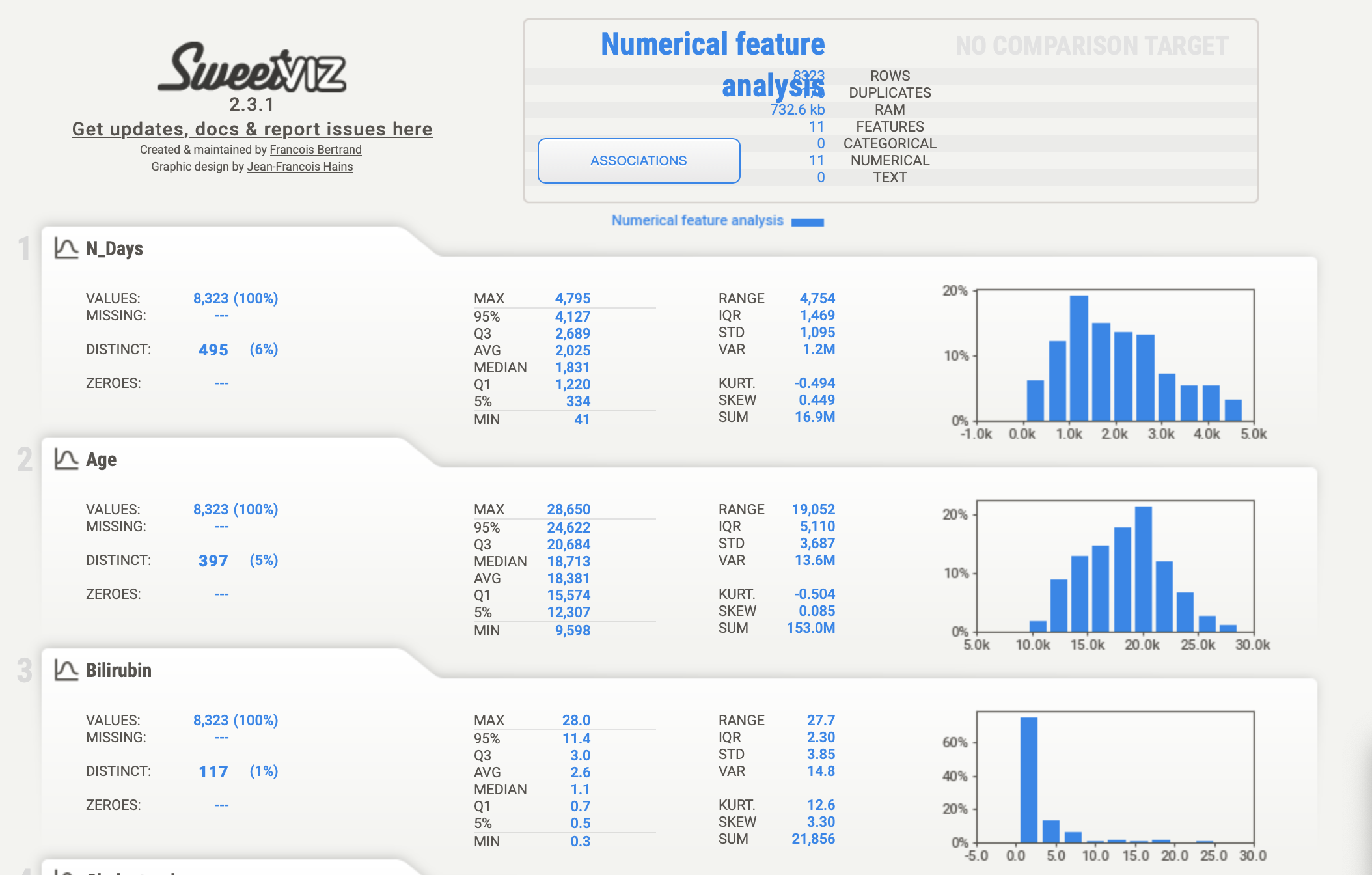
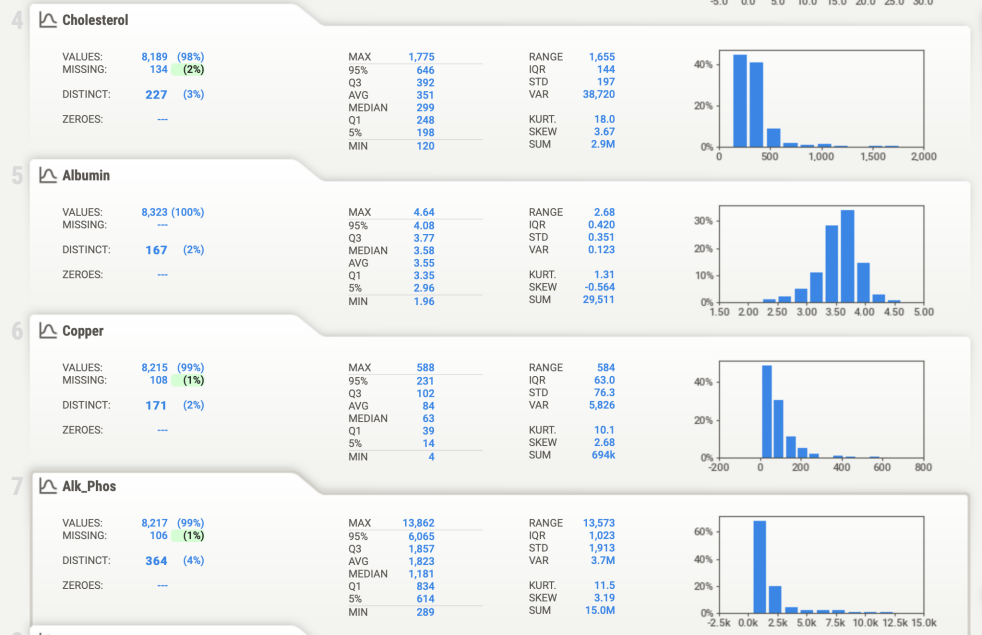
* D = 0 (Death)
* C = 1 (Censored due to lost follow-up)
* CL = 2 (Censored due to Liver transplant)

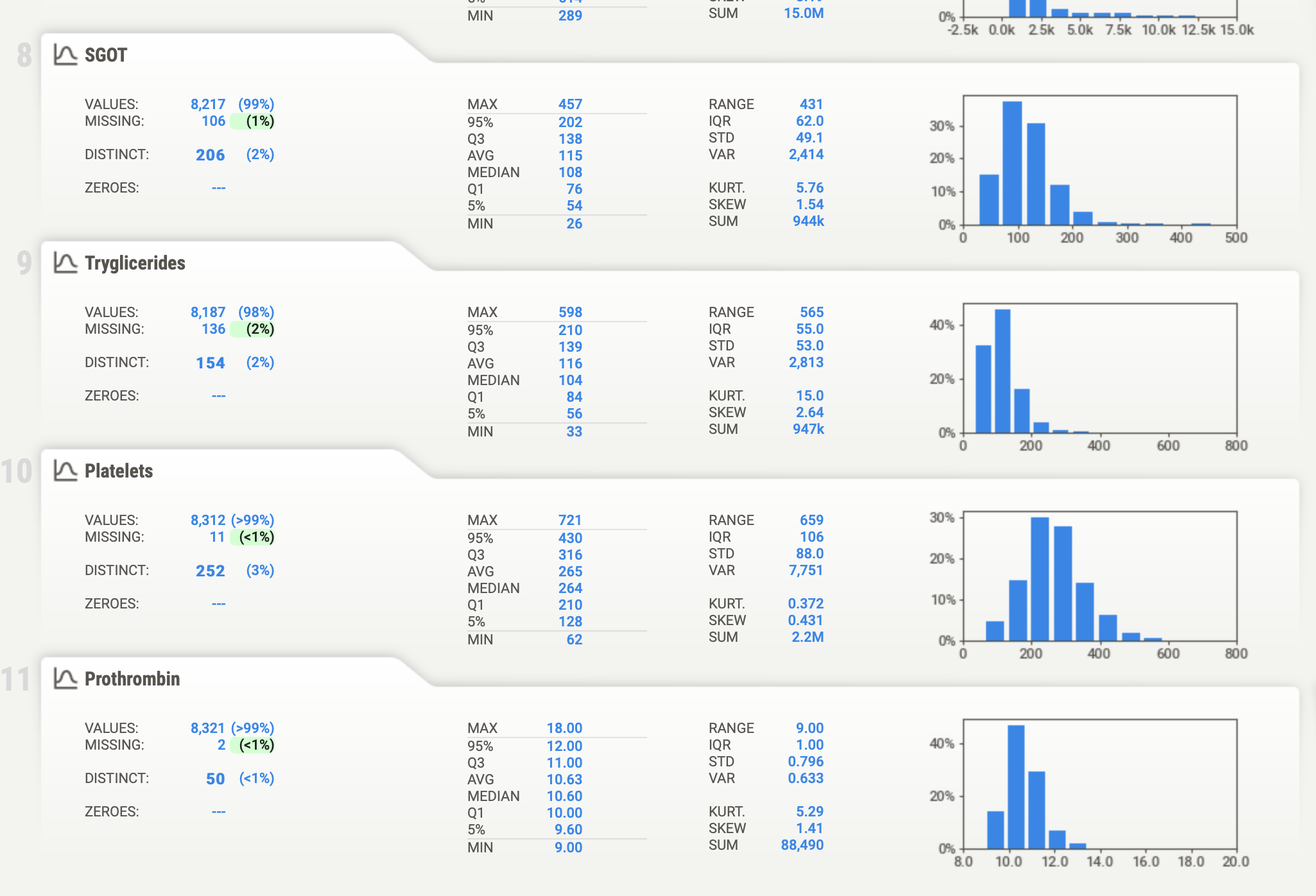
The below features are in the dataset.

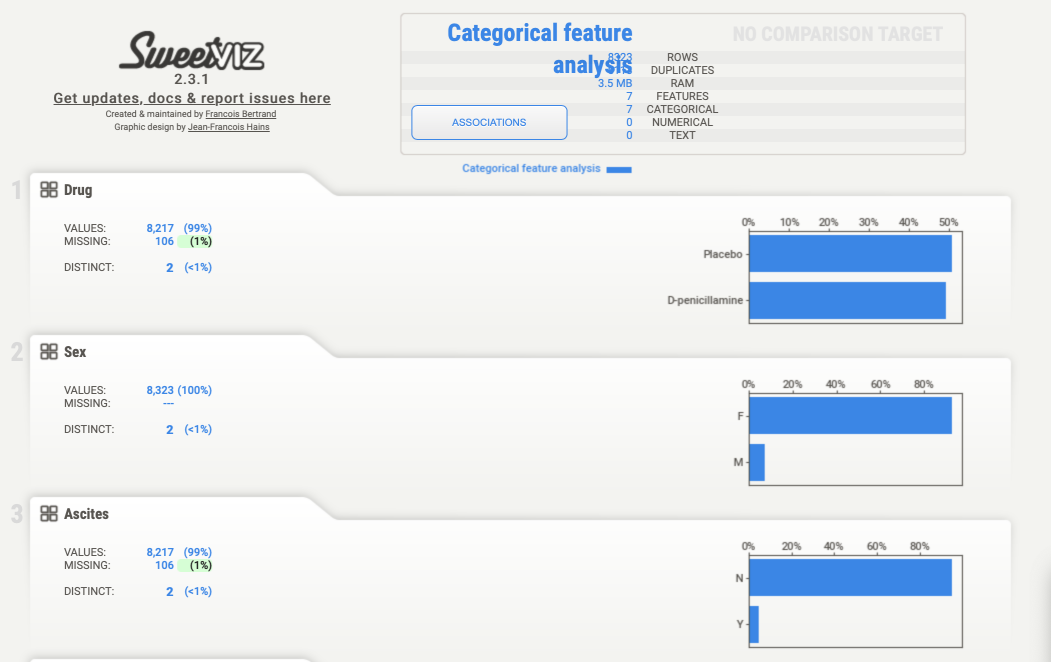
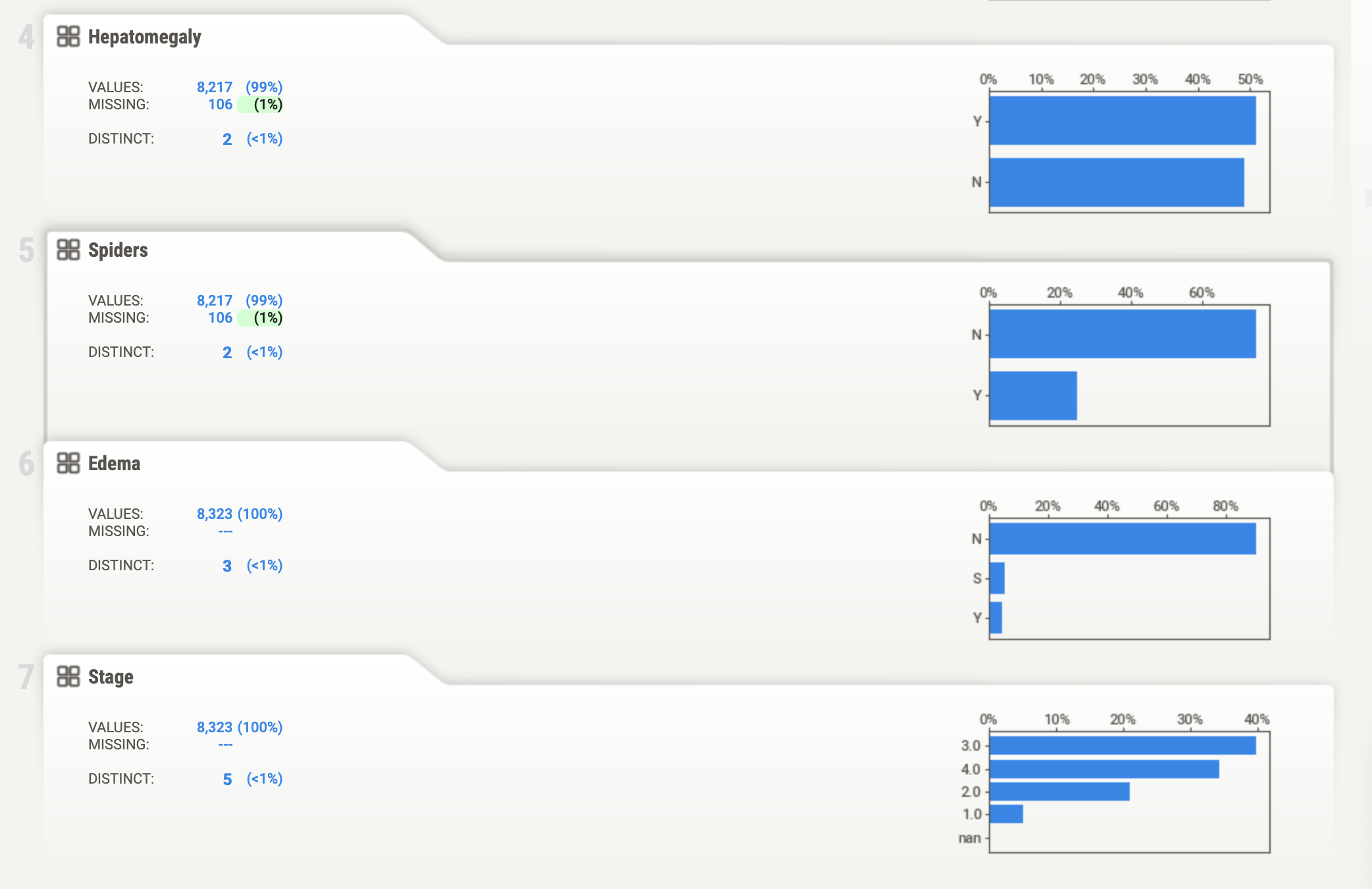
* **Drug**: The specific medication or treatment regimen administered to the patient during the monitoring period.
* **Bilirubin**: A key blood marker indicating liver function; elevated levels may suggest liver dysfunction.
* **Cholesterol**: The cholesterol level in the blood, a lipid profile marker relevant to cardiovascular and metabolic health.
* **Albumin**: A protein synthesized by the liver; low levels may indicate liver disease and affect overall protein balance
* **Hepatomegaly**: Enlargement of the liver is an important clinical indicator of liver health.
* **Spiders**: The presence of spider angiomas or spider nevi, visible vascular lesions on the skin associated with liver disease.
* **Edema**: The abnormal fluid accumulation, typically manifested as swelling, and often linked to liver-related complications.
* **Triglycerides:** A type of fat in the blood that influences metabolic and cardiovascular health.
* **Platelets**: Blood cell fragments, with abnormal levels potentially indicating liver dysfunction.
* **Prothrombin**: A blood clotting factor with variations affecting the coagulation process.
* **Stage**: The stage of liver disease, likely determined by specific diagnostic criteria and indicators.
* **Status**: The target variable indicating the patient's outcome, classified as censored (C), alive due to liver transplant (CL), or deceased (D).
* **N\_Days**: The duration in days a patient has been monitored, providing a temporal context for the observation.
* **Age**: The patient's age during observation contributes to the demographic profile.
* **Ascites**: The presence or absence of fluid accumulation in the abdominal cavity is a critical clinical sign often associated with liver diseases.
* **SGOT** (Serum Glutamic Oxaloacetic Transaminase): An enzyme indicating liver and heart health; elevated levels may suggest liver damage.

**Data Exploration and Preprocessing**

We use Sweetviz for ED.Sweetviz is a Python library designed to visualize and compare datasets during exploratory data analysis (EDA). It creates high-density visualizations with minimal code and helps data scientists and analysts quickly gain insights into their data. The library is handy for understanding the distribution of variables, identifying patterns, and comparing datasets before and after transformations. Critical features of Sweetviz include Dataframe and Series Analysis: Sweetviz can analyze both entire data frames and individual series (columns) within a data frame. This flexibility allows for comprehensive exploration.

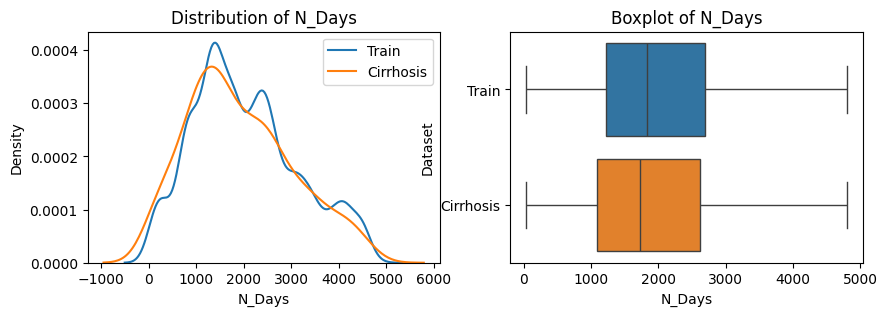
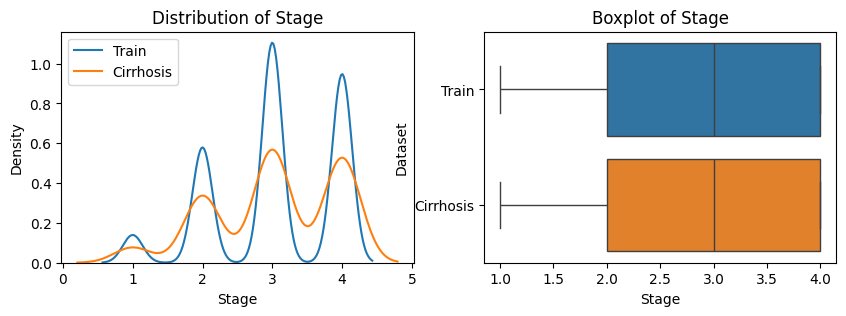
****

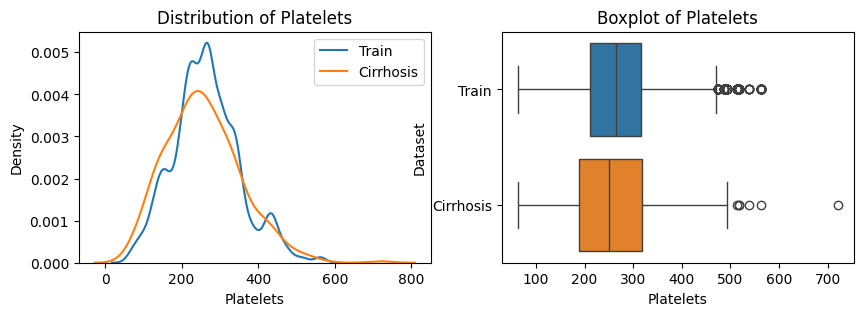


The dataset showcases outliers in critical biomarkers such as Tryglicerides, Alk\_Phos, etc., necessitating the use of RobustScaler to mitigate their influence. Notably, class C dominates the distribution, indicating an imbalanced dataset. Age emerges as a significant factor, showcasing a clear correlation with cirrhosis prevalence and subsequent patient outcomes. Biomarkers like Bilirubin, Albumin, and Copper further illuminate the severity of liver damage, aligning closely with clinical expectations. Alk\_Phos and SGOT levels provide insights into liver function and disease progression, while the thrombocytopenia indicator is a crucial marker for assessing potential clotting complications. The meticulous comparison of synthesized data features with clinical norms bolsters the dataset's credibility for modeling cirrhosis outcomes, enabling more informed and precise patient care interventions.

**EDA-Training Data Augmentation**



Analyzing the compatibility of the original cirrhosis dataset involves examining its structure, quality, and modeling adequacy. Key steps include estimating feature distributions, comparing them across datasets, and deciding to merge based on similarities. Merging requires addressing data integrity and consistency issues through preprocessing. The goal is to improve model performance while maintaining data integrity and relevance to the problem area.

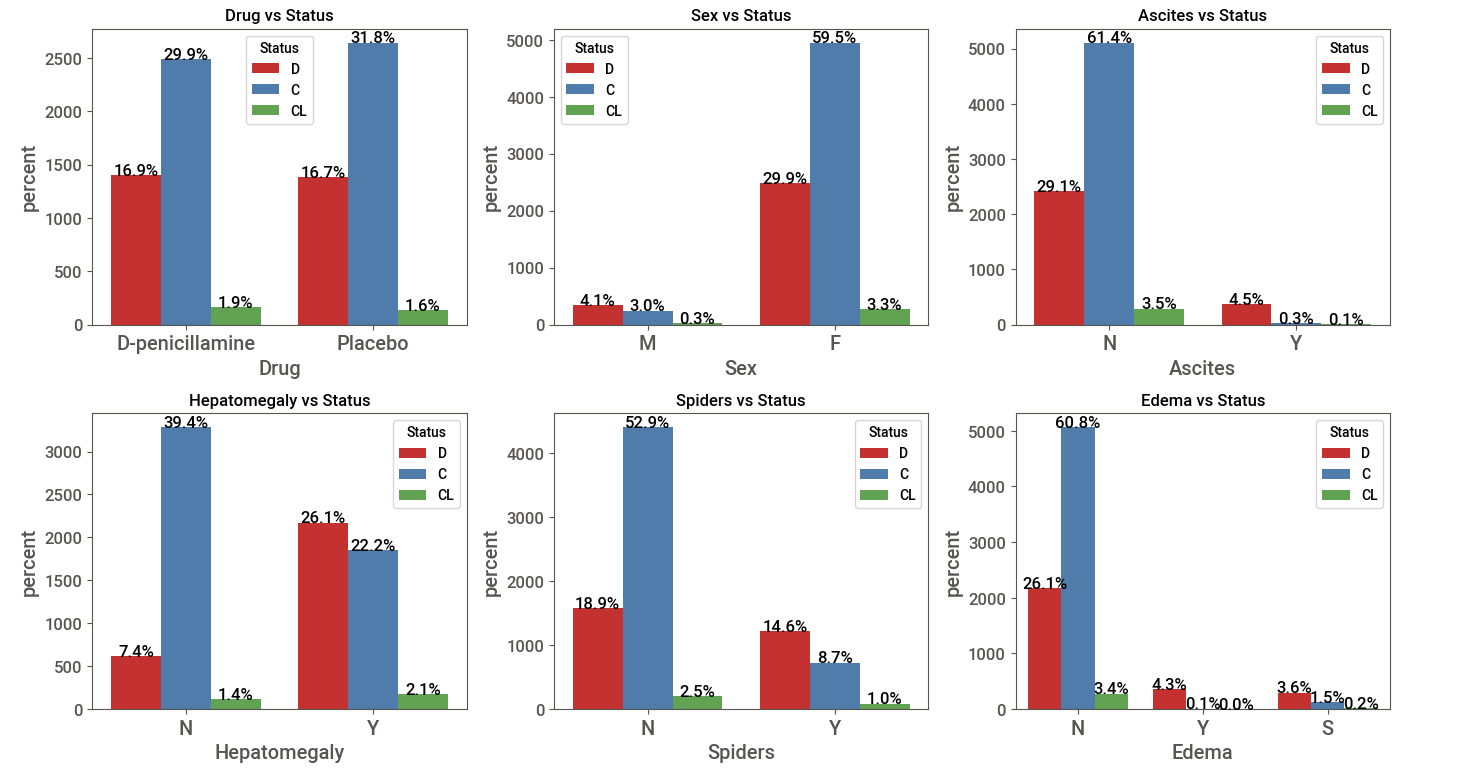
**EDA: Distribution of numerical variables across output classes**

In the analysis, we delved into the dataset, examining how different characteristics were distributed across different outcome categories. By segmenting each class's data and characteristic distributions, we try to discover unique patterns that could improve the modeling process. This detailed investigation provided insights into potential relationships between characteristics and outcome categories that helped develop more accurate prediction models.

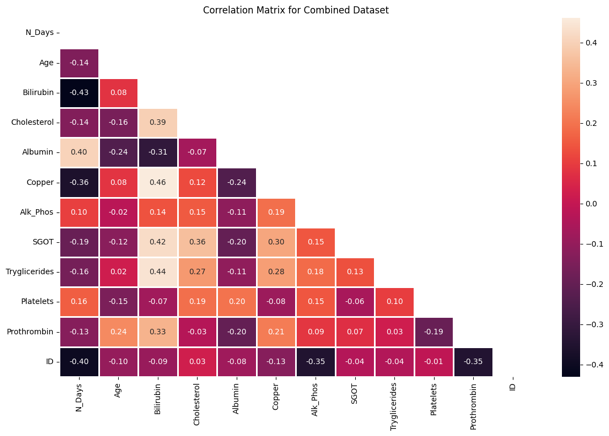
**A group of graphs showing different colored lines

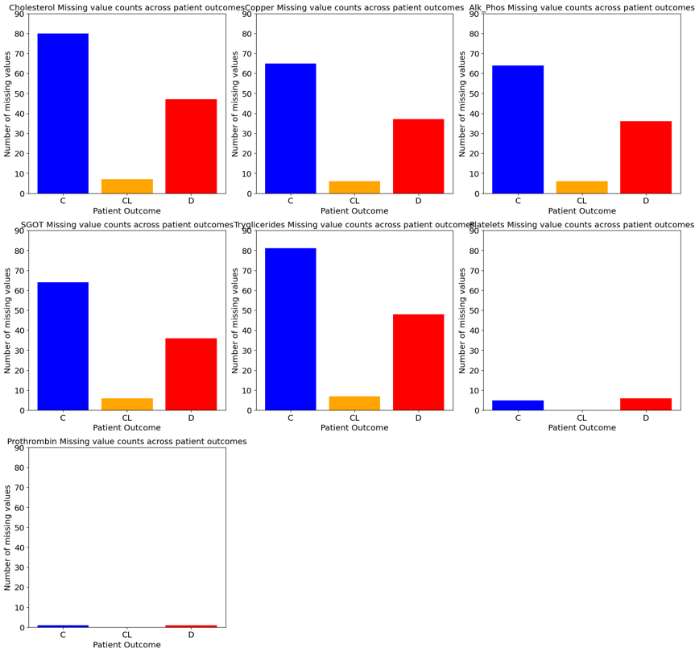
Description automatically generated**

**EDA: Distribution of categorical variables across output classes and output class distribution**



**EDA - Correlation and statistical significance**

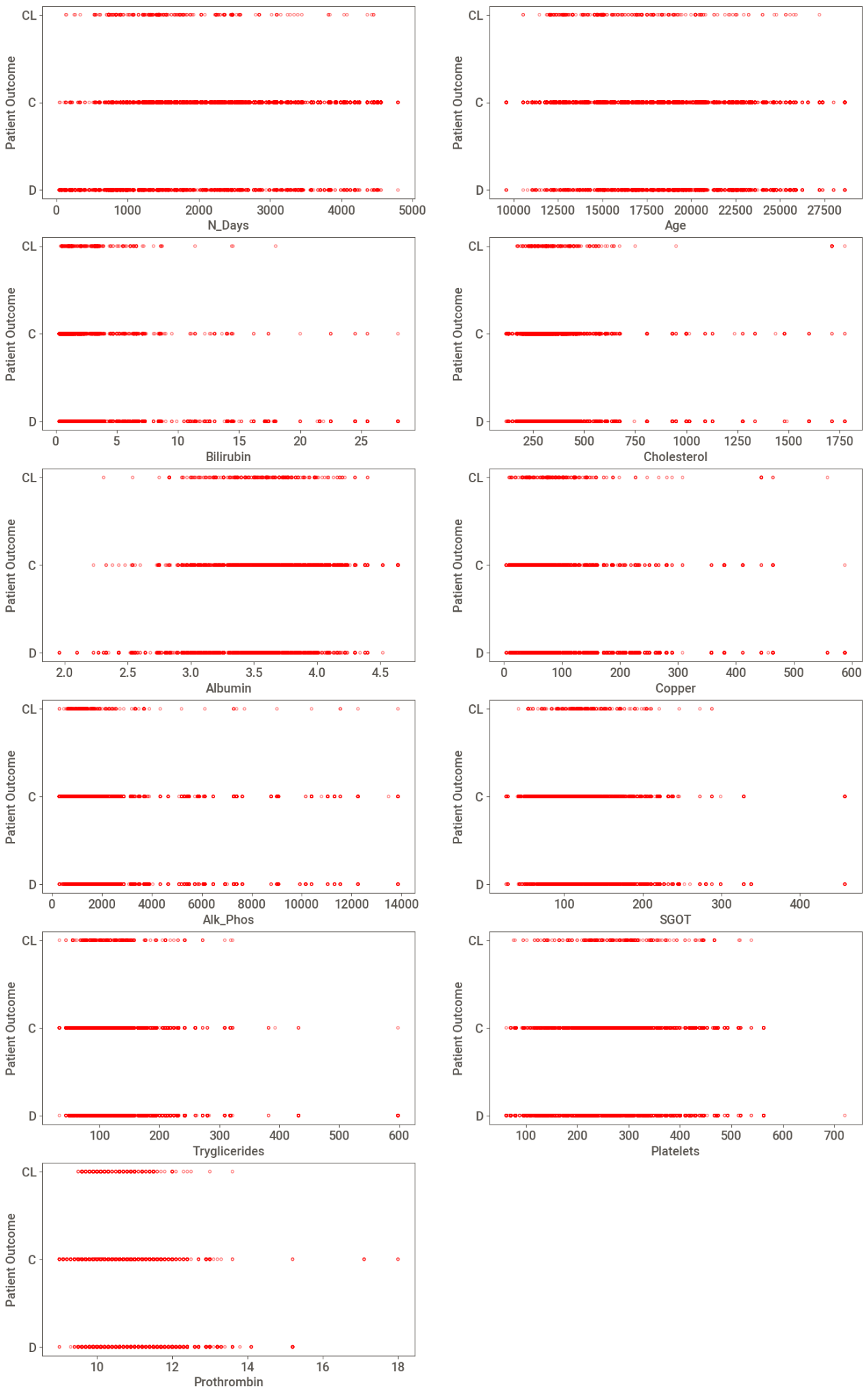
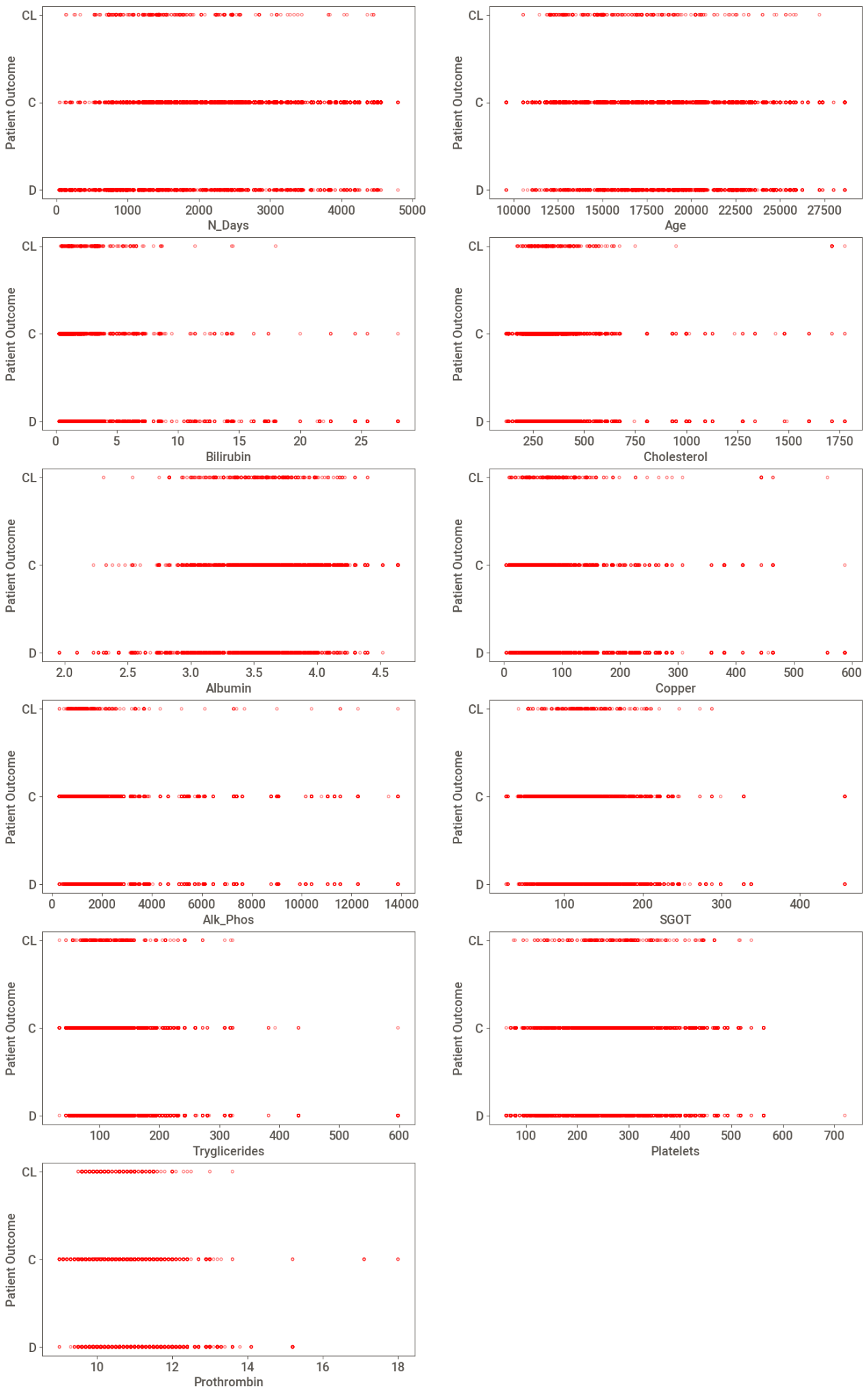
After an in-depth analysis, it became clear that there are no significant relationships between the numerical variables included in the data. Consequently, the deletion of some numerical variables was not considered. Although this may seem ambiguous at first glance, it suggests that each numerical variable acts independently, forcing a reevaluation of feature selection and modeling. This highlights the importance of considering the unique contribution of each variable when optimizing forecast performance.

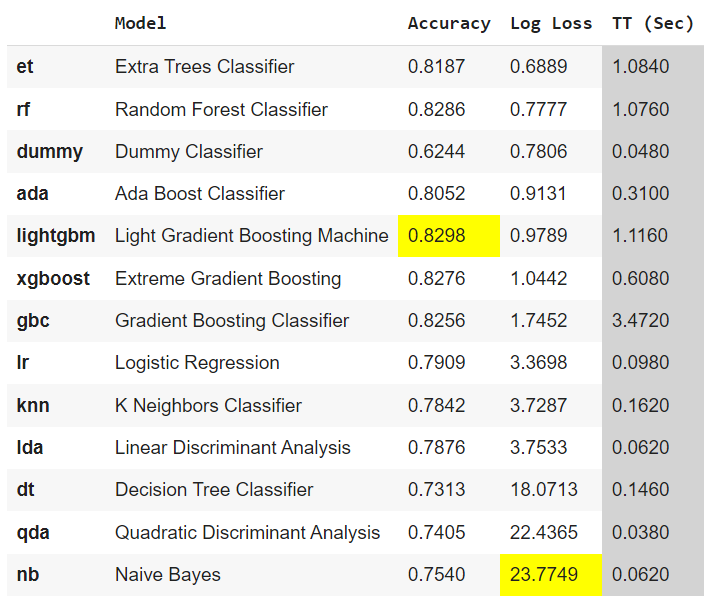
**EDA – Missing Values**

After analysis, we found no discernible pattern in the distribution of missing values in the feature spaces or outcome categories. This suggests that missing values occur randomly rather than systematically. Therefore, consistent handling of missing values during data processing becomes critical. This finding also highlights the importance of robust modeling techniques to accommodate missing data effectively. Although the lack of pattern may seem insignificant, it guides our approach to data processing and modeling and ensures the reliability of our predictive models.

**EDA – Outlier Capping**

In response to outliers in our data graphs, we developed a method to reduce their impact by focusing on valuable features within result categories. This strategy was used to improve the robustness of the model. We evaluate its effectiveness through quantitative and qualitative assessments, considering accuracy and model stability metrics. While promising, later iterations are designed to improve performance for different datasets and conditions. Overall, this approach represents a proactive step in improving data quality and model reliability.

**Baseline Models: Auto ML (PyCaret)** We initially evaluated machine-learning models with PyCaret's AutoML function, using stratified k-fold cross-validation for fairness. Log loss emerged as a critical metric for accurate probabilistic classification assessment. We benchmarked Extra Tree Classifier and Random Forest Classifier for their robustness in various tasks, providing a baseline for future advanced algorithms. We'll enhance these models by refining features, detecting outliers, and addressing data imbalances. Tuning hyperparameters is essential for optimization and adapting models to dataset nuances. Ensuring generalizability and reducing overfitting risks through proper validation methods is a priority. We'll also explore ensemble learning and feature selection for improved performance, aiming to develop highly accurate and scalable predictive models tailored to our needs iteratively.

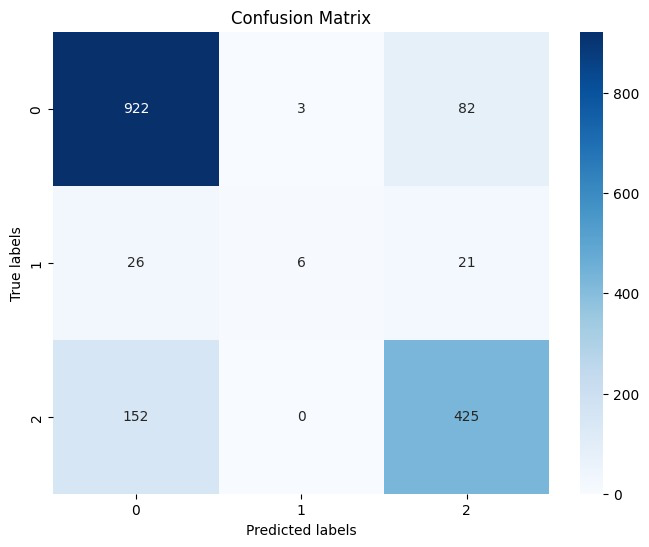
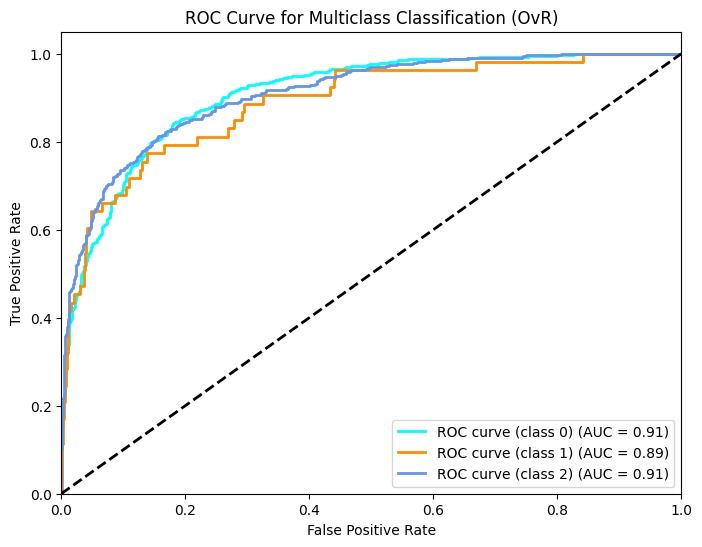
**Model Building**

**Iteration 1:**

The dataset underwent several preprocessing steps, including categorical encoding and data normalization. The Edema column was one-hot encoded, while the remaining columns were ordinal. A MinMax scaler was applied to normalize the data and ensure all features were within a consistent range. Additionally, custom outlier removal was performed using a selective capping approach, which involved adjusting extreme values to minimize their impact on the analysis while maintaining data integrity.

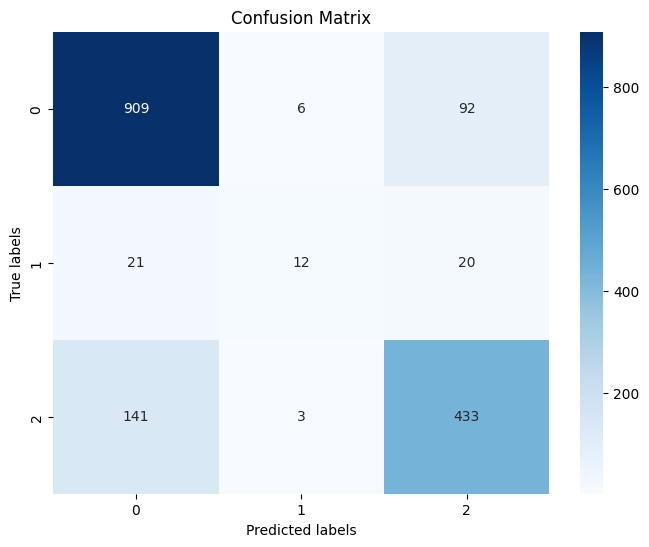
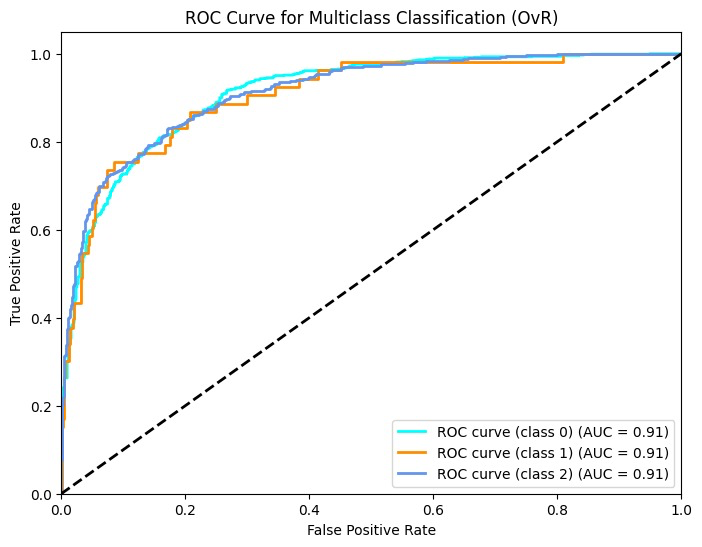
**Iteration 2:**

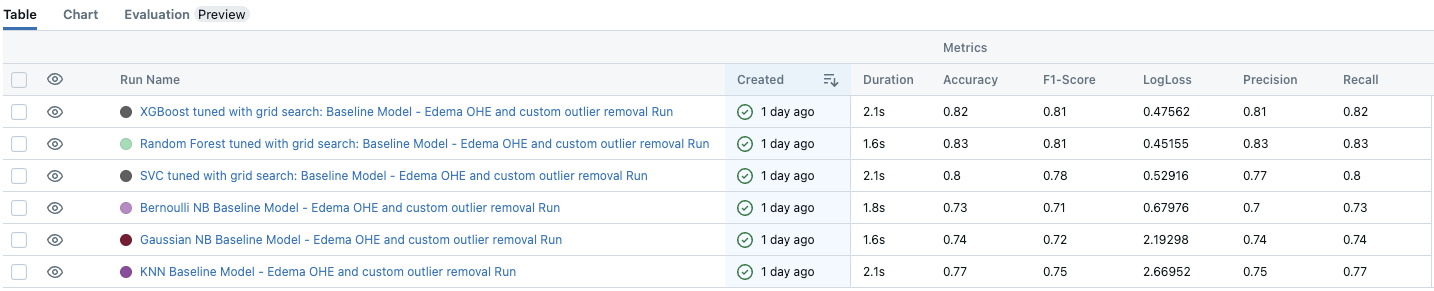
The dataset underwent several preprocessing steps, including categorical encoding, data scaling, and hyperparameter tuning. All columns were one-hot encoded except for the target column, which was ordinal encoded. To standardize the range of values across the data, the MinMax scaler was applied to the Age and N\_days columns, while the Robust scaler was used for all other columns to mitigate the influence of outliers. Additionally, hyperparameter tuning was performed using the Hyperopt optimization library to fine-tune the model and achieve optimal performance.

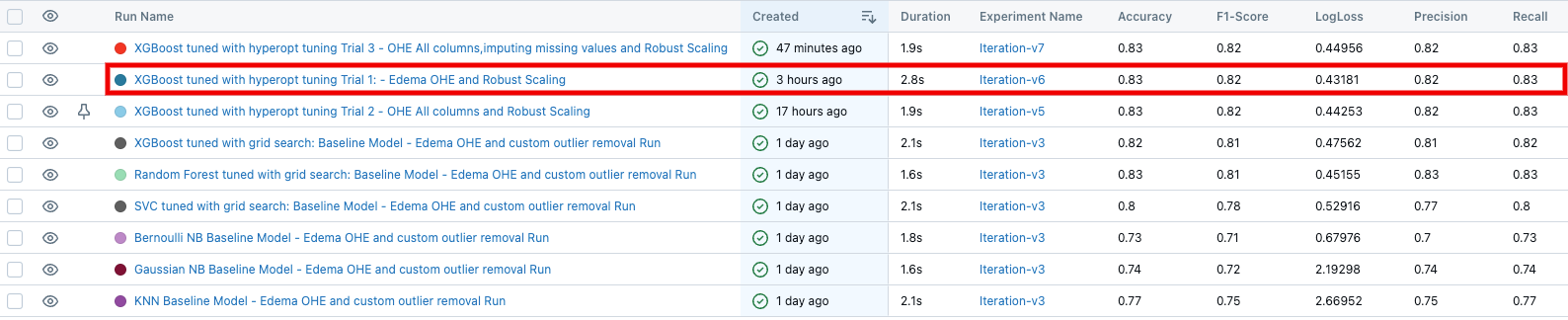
**Iteration 3:**

In preprocessing, categorical encoding involves utilizing one-hot encoding for the 'Edema' column and ordinal encoding for the 'Target' column. Scaling techniques are then applied, with MinMax scaling employed for 'Age' and 'N\_days,' while a Robust Scaler is utilized for all other columns. Finally, hyperparameter tuning uses the hyperopt() function to optimize model performance. These steps collectively enhance the data's suitability for machine learning algorithms, ensuring practical model training and accurate predictions.

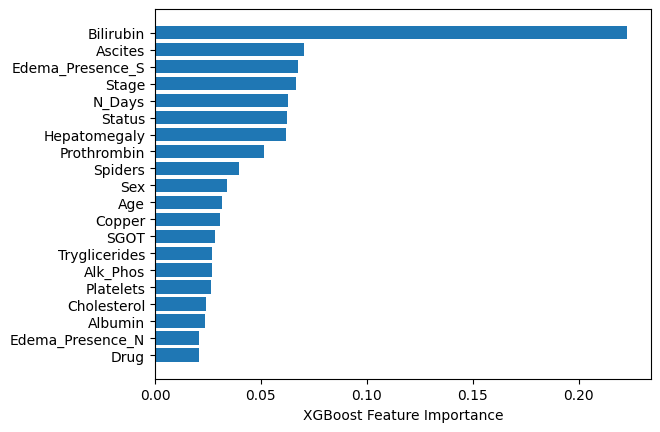
 



**Model Evaluation**



The table shows performance metrics of XGBoost models, including Accuracy, F1-Score, LogLoss, Precision, and Recall. The best model, "XGBoost tuned with hyperopt tuning Trial 3 - Edema OHE and Robust Scaling," has Accuracy, Precision, and LogLoss of 0.83, 0.83, and 0.43181, respectively. It also has high F1-Score and Recall (0.82 each), indicating a well-balanced model crucial for classification tasks.



The bar chart from an XGBoost model ranks variables by importance in predicting an outcome. Critical variables like **Bilirubin**, **Ascites**, and **Edema** are significant, while variables like Sex, Age, and Drug have minimal impact. This chart aids in refining models or focusing on impactful clinical assessments.

**Conclusion**

The project delved into the impactful role of predictive modeling in liver transplant patient selection, employing advanced analytics and patient factor analysis. Its primary contributions include improving patient selection by prioritizing candidates based on post-transplant survival prognoses, empowering informed decisions by healthcare professionals, and optimizing organ allocation for maximum benefit. Additionally, it explored the optimal timing of transplantation considering disease progression dynamics and organ availability, aiming to enhance patient outcomes and healthcare resource utilization. The project's implications extended to financial planning, with the predictive model estimating healthcare costs associated with various outcomes. It underscored the criticality of early intervention by proactively identifying at-risk patients, thus preventing adverse events, reducing costs, and enhancing patient quality of life. This initiative highlights the transformative potential of predictive modeling in liver transplantation, paving the way for more accurate, efficient, and patient-centered care strategies.