**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) 4

Baseline Models: Auto ML.............................................................................................. 7

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

Model Evaluation.......................................................................................................... 9

Conclusion................................................................................................................................. 11

Introduction

Liver cirrhosis is a major global health problem characterized by the progressive deterioration of liver function due to the formation of scar tissue. The consequences of liver cirrhosis vary widely, ranging from relatively stable conditions to life-threatening complications such as liver failure, portal hypertension, and hepatocellular carcinoma. Prediction of various outcomes associated with liver cirrhosis is crucial to optimize patient care, treatment planning, and resource allocation in healthcare systems. The Multiclass Cirrhosis Outcome Classification project addresses this critical need by applying machine learning techniques to develop predictive models that can stratify patients based on the likelihood of different outcomes. Using a variety of patient data, including demographics, medical history, laboratory test results, imaging studies and lifestyle factors, the project aims to identify patterns and associations that may inform outcome. The multicategory classification method allows the classification of liver cirrhosis results into several categories that cover different possibilities from stable progression of the disease to various complications and side effects. By training machine learning algorithms on large heterogeneous datasets containing real patient data, the project aims to create robust predictive models that can accurately classify patients into different outcome classes. By developing and validating these predictive models, healthcare providers can gain valuable information about the individual risk profiles of patients with cirrhosis. This allows doctors to adjust treatment strategies, implement targeted interventions and optimize patient treatments based on predicted outcomes. In addition, the predictive models generated by this project have the potential to improve clinical decision-making, facilitate early intervention and improve patient outcomes by identifying high-risk individuals who may benefit from closer monitoring or aggressive treatment. In general, the "Multiclass Classification of Hepat Cirrhosis Outcomes" project represents an important advance in the treatment of liver diseases. It provides a data-driven approach to predict and share outcomes for patients with cirrhosis. Using the power of machine learning and predictive analytics, this project aims to help develop personalized medical strategies that improve the quality of care and outcomes for patients with cirrhosis.

Problem Statement

Problem statement: Liver cirrhosis is a challenge to accurately predict outcomes, from stable conditions to severe complications. Current prognostic methods often lack precision and do not take into account the diversity of the disease. The project aims to develop reliable prognostic 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 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:

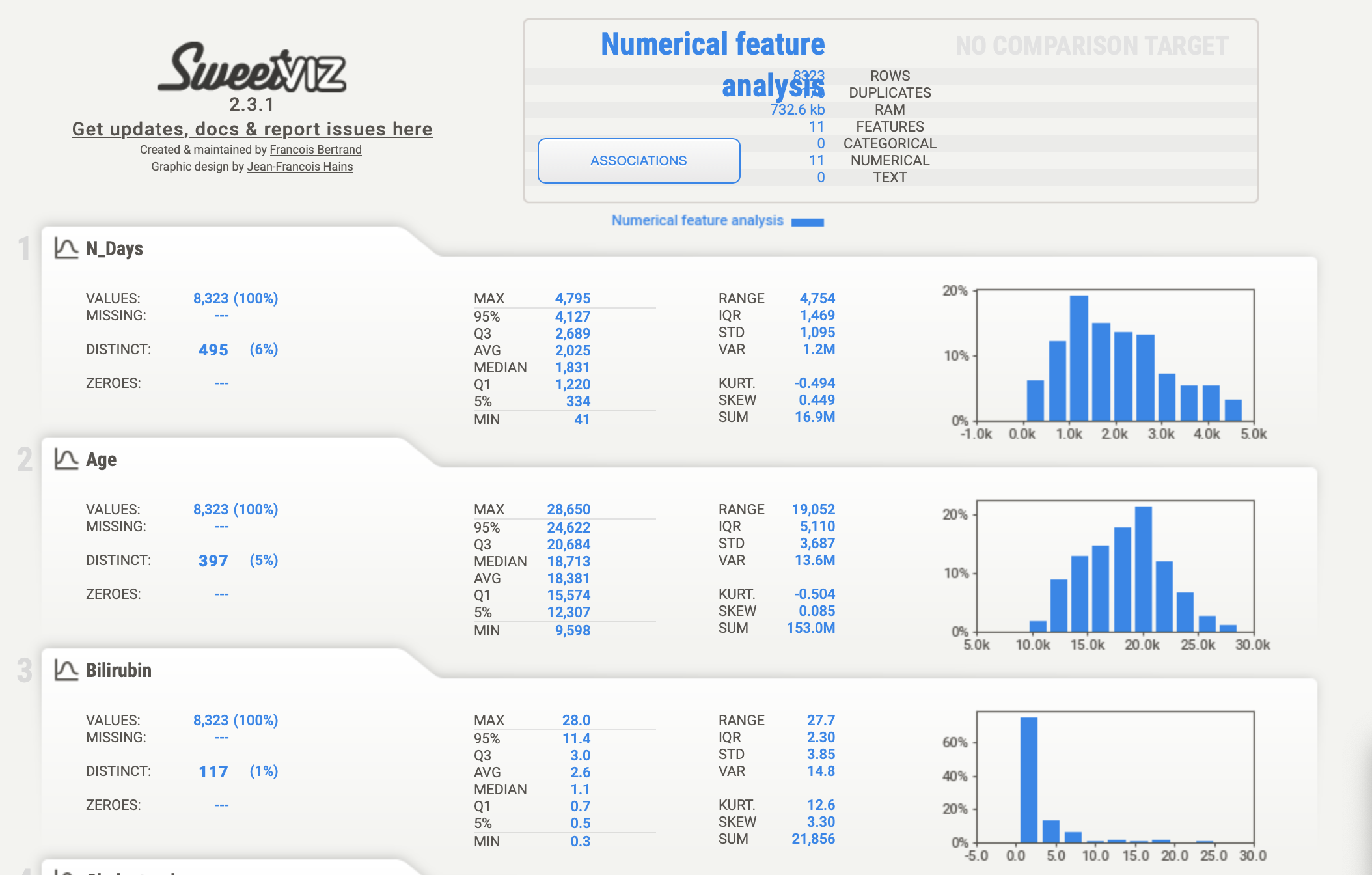
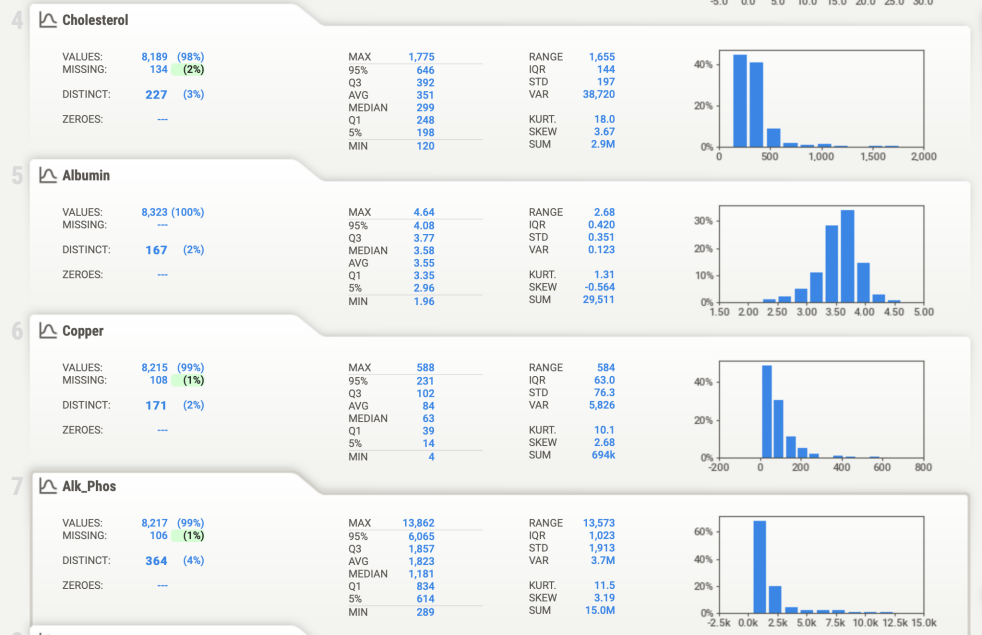
* D = 0 (Death)
* C = 1 (Censored due to lost follow-up)
* CL = 2 (Censored due to Liver transplant)

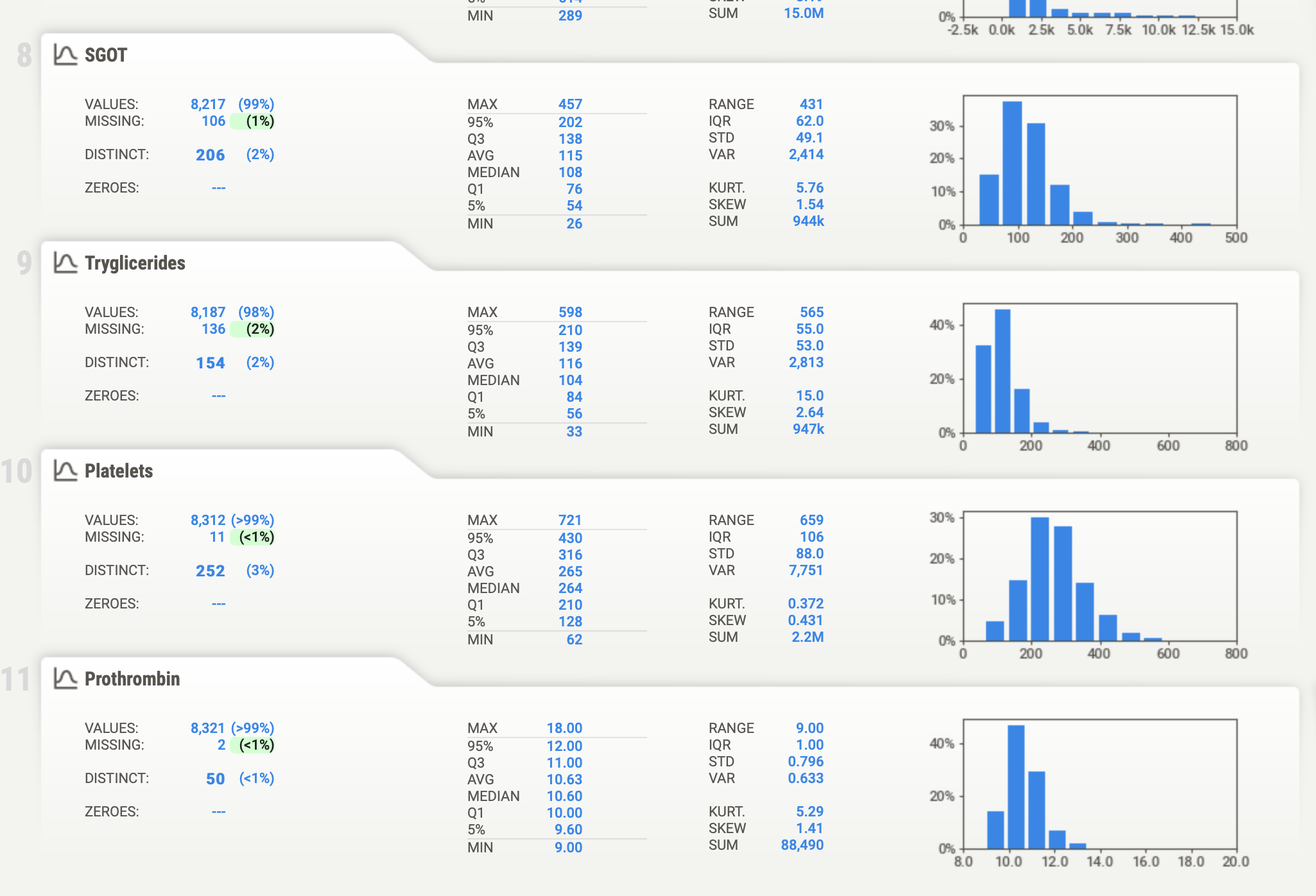
The features in the dataset are: N\_Days', 'Drug', 'Age', 'Sex', 'Ascites', 'Hepatomegaly', 'Spiders', 'Edema', 'Bilirubin', 'Cholesterol', 'Albumin', 'Copper', 'Alk\_Phos', 'SGOT', 'Tryglicerides', 'Platelets', 'Prothrombin', 'Stage', 'Status'.

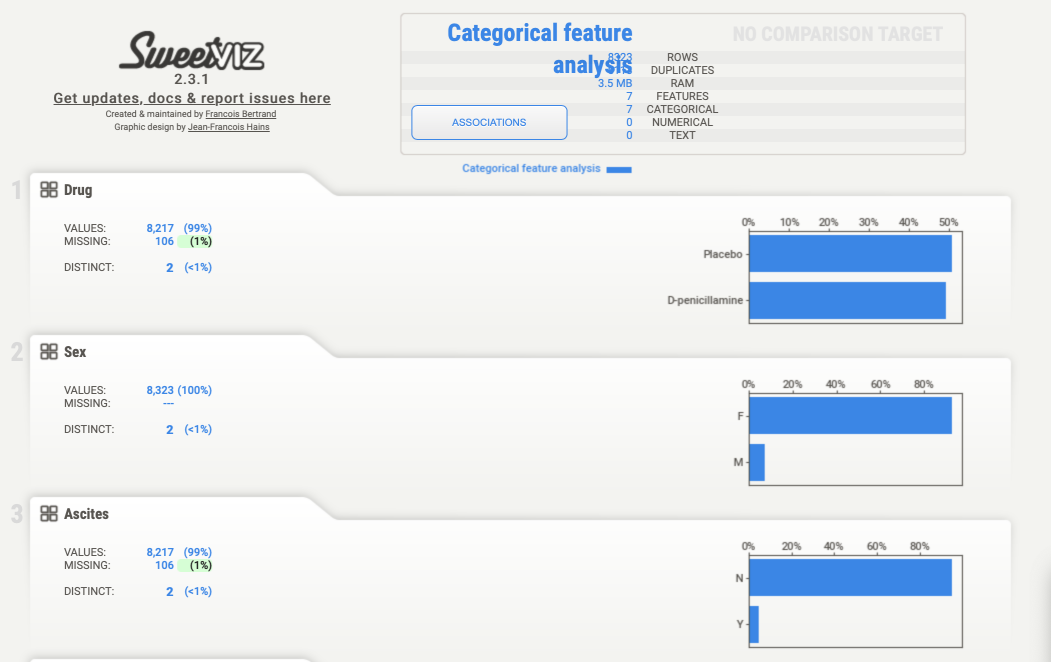
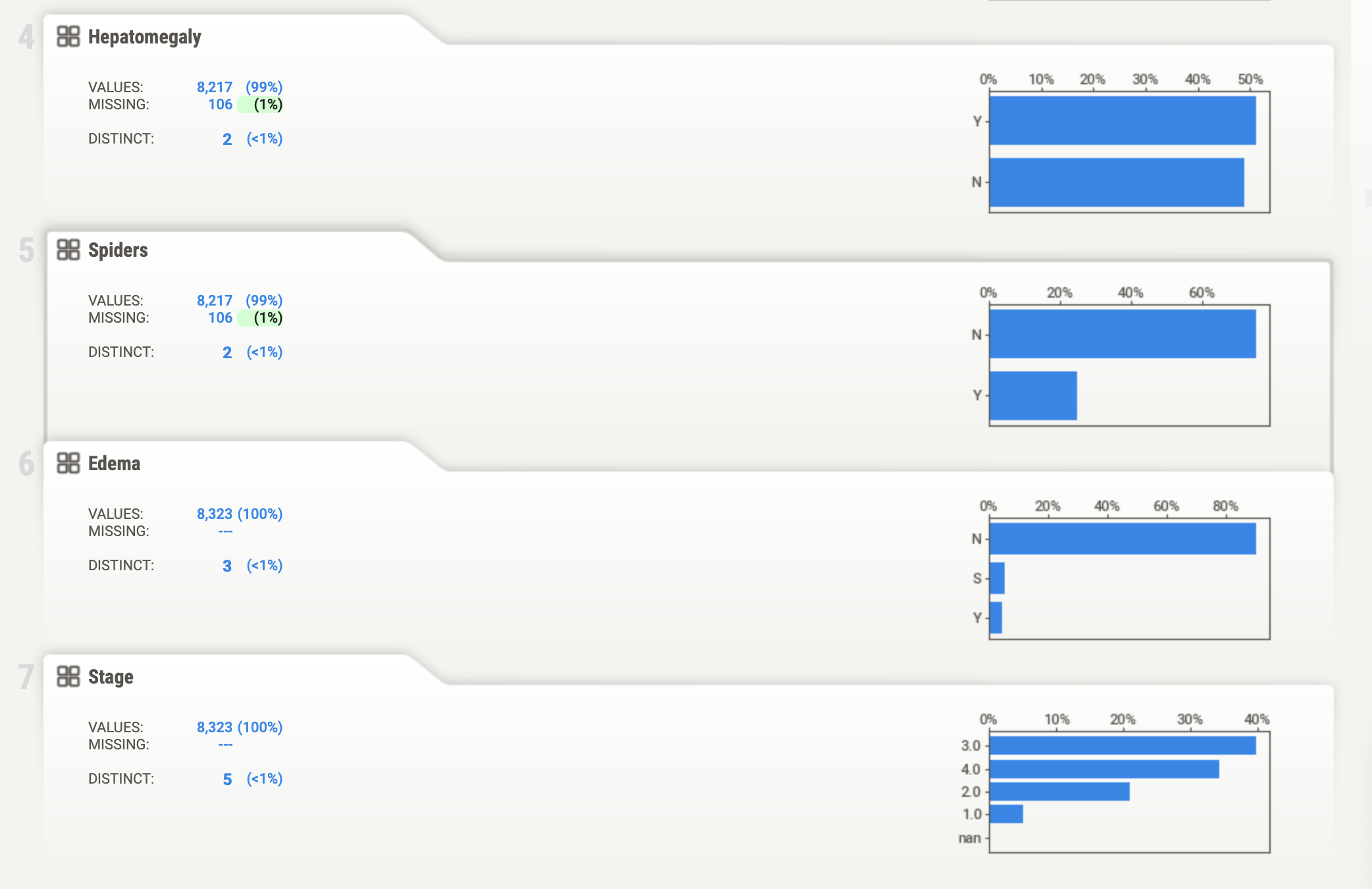
* Drug: The specific medication or treatment regimen administered to the patient during the monitoring period.
* Bilirubin: A key blood marker indicating liver function; elevated levels may suggest liver dysfunction.
* Cholesterol: The level of cholesterol in the blood, a lipid profile marker relevant 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, serving as 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 accumulation of fluid, typically manifested as swelling, and often linked to liver-related complications.
* Tryglicerides: A type of fat present in the blood, influencing 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 for which a patient has been monitored, providing a temporal context for the observation.
* Age: The age of the patient at the time of observation, contributing to the demographic profile.
* Ascites: Presence or absence of fluid accumulation in the abdominal cavity, a key 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 for visualizing and comparing 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 particularly useful for understanding the distribution of variables, identifying patterns, and comparing datasets before and after transformations. Key features of Sweetviz include: Dataframe and Series Analysis: Sweetviz can analyze both entire dataframes and individual series (columns) within a dataframe. This flexibility allows for comprehensive exploration.

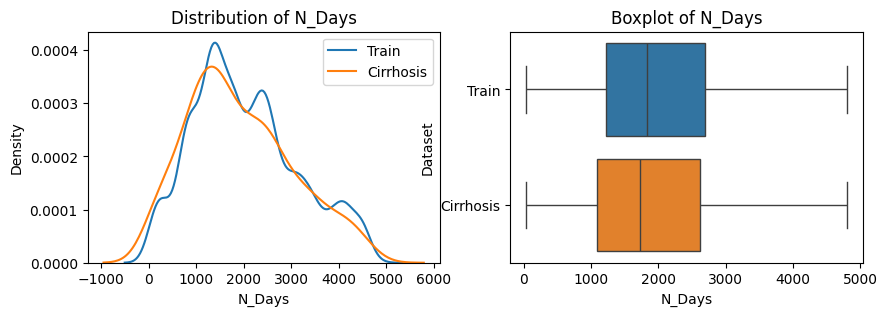
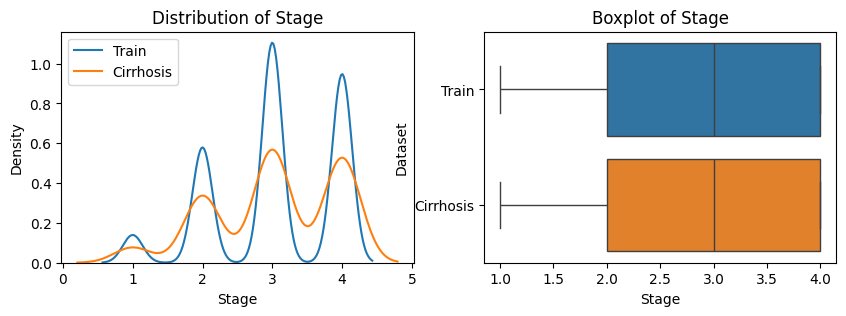
****

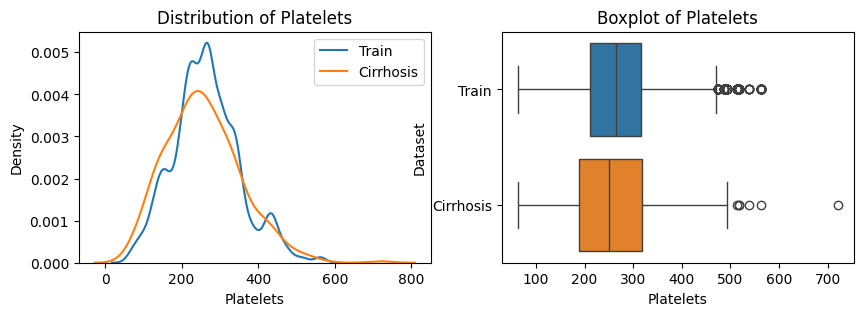


The dataset showcases outliers in key biomarkers such as Tryglicerides, Alk\_Phos, etc., necessitating the use of RobustScaler to mitigate their influence. Notably, class C dominates the distribution, indicating an imbalanced dataset. Age emerges as a significant factor, showcasing a clear correlation 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 serves as 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, thereby enabling more informed and precise patient care interventions.

**EDA-Training Data Augmentation**

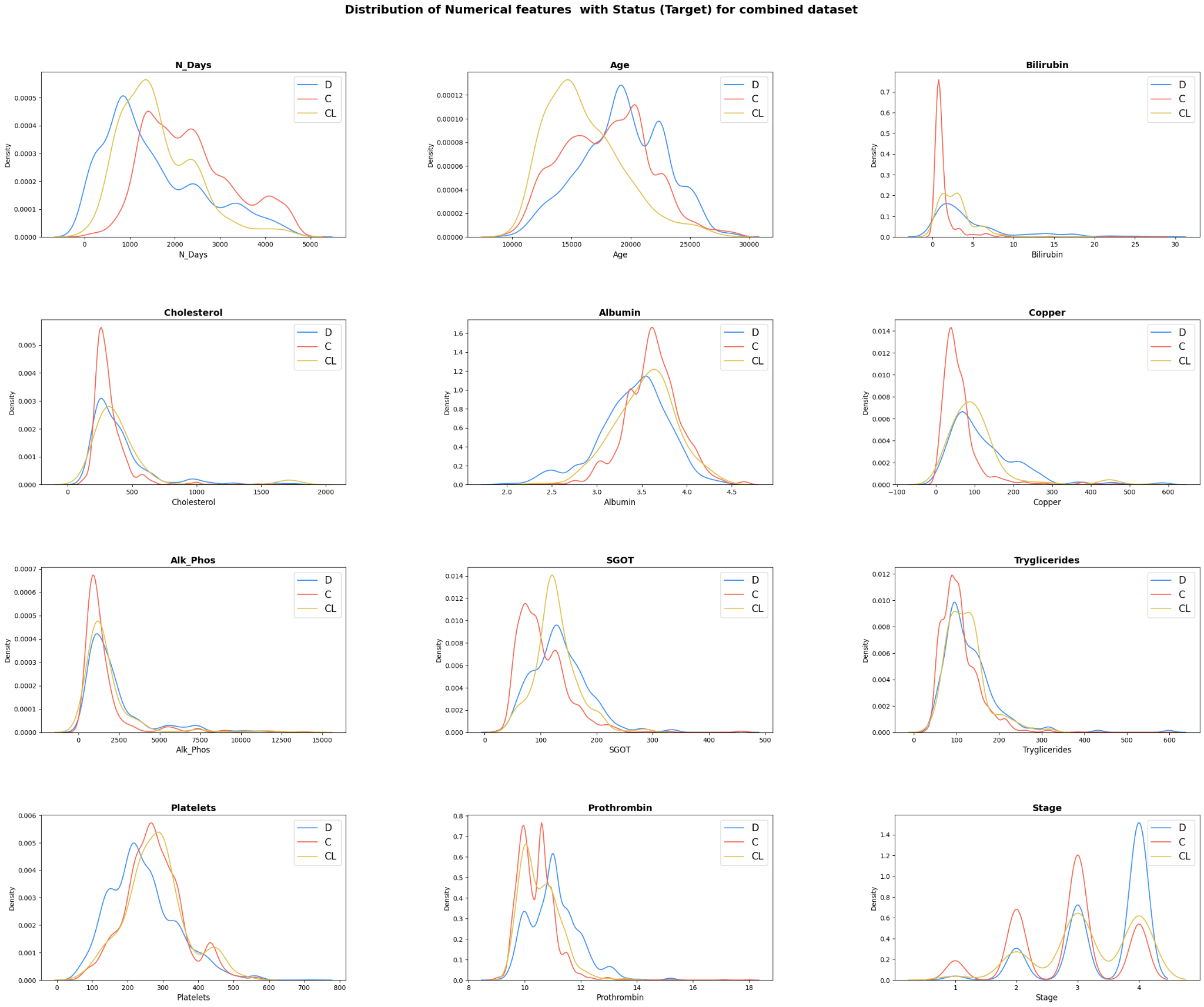
 



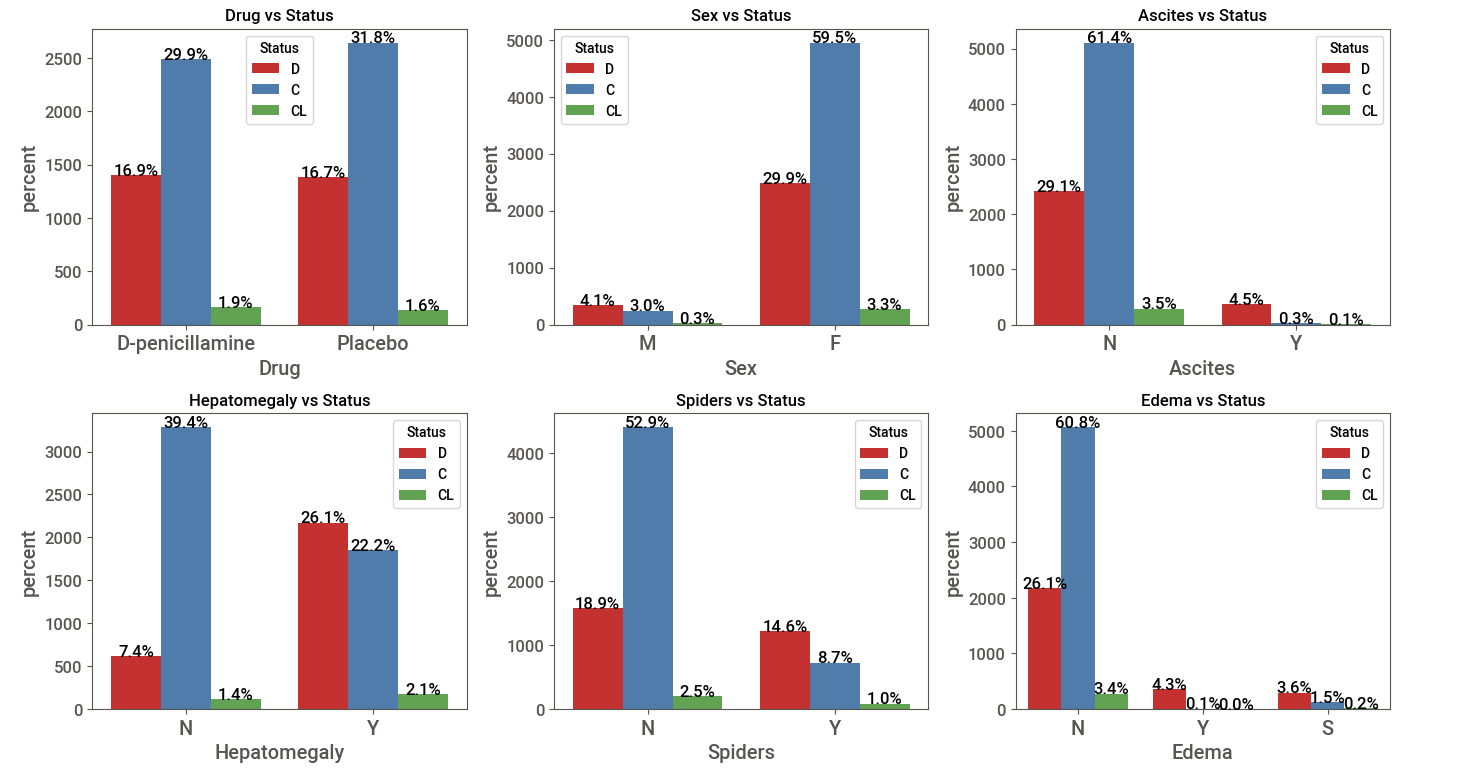
Analyzing the compatibility of the original cirrhosis dataset involves examining its structure, quality, and modeling adequacy. Key steps include estimating feature distributions, comparing them across datasets, and deciding to merge based on similarities. Merging requires addressing data integrity and 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**

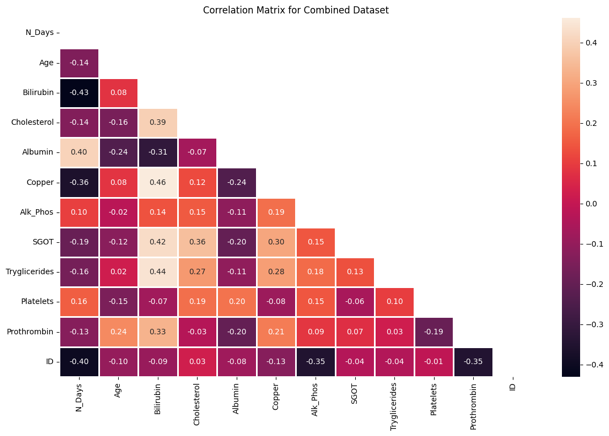
In the analysis, we delved into the dataset, examining how different characteristics were distributed across different outcome categories. By segmenting the data and characteristic distributions in each class, we try to discover unique patterns that could improve the modeling process. This detailed investigation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



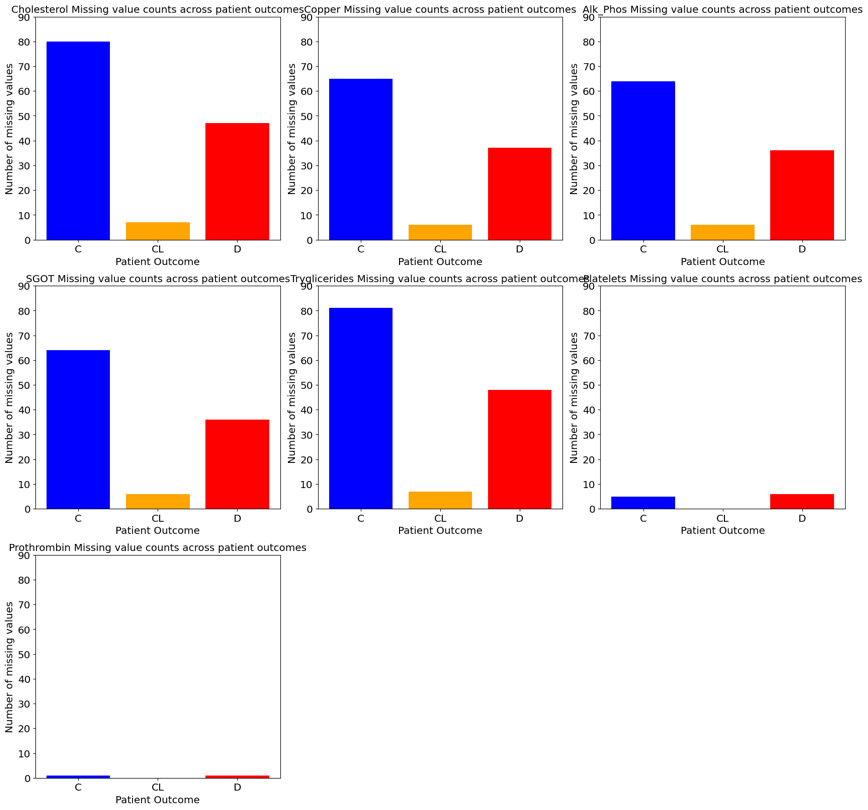
**EDA: Distribution of categorical variables across output classes and output class distribution**



**EDA - Correlation and statistical significance**

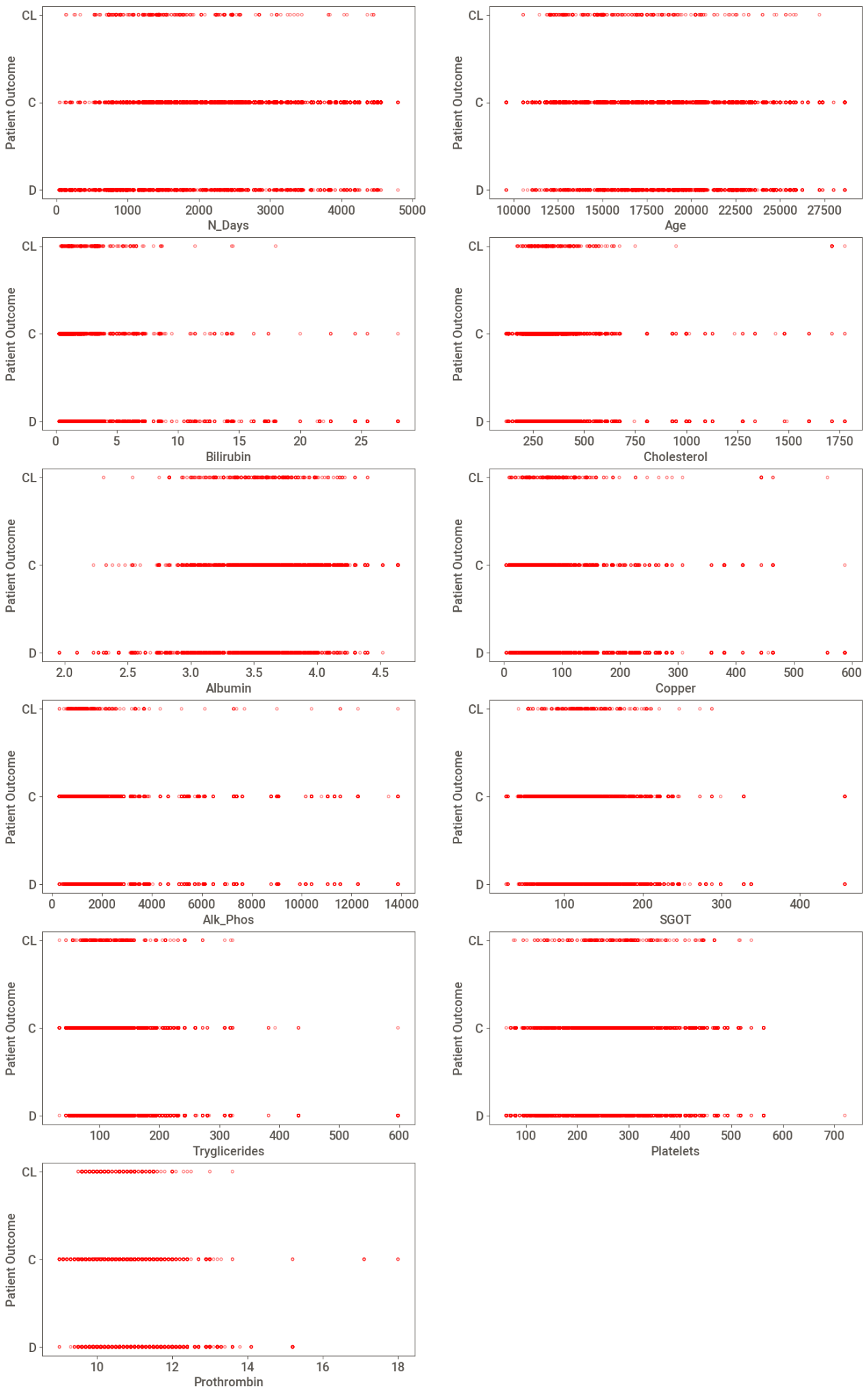
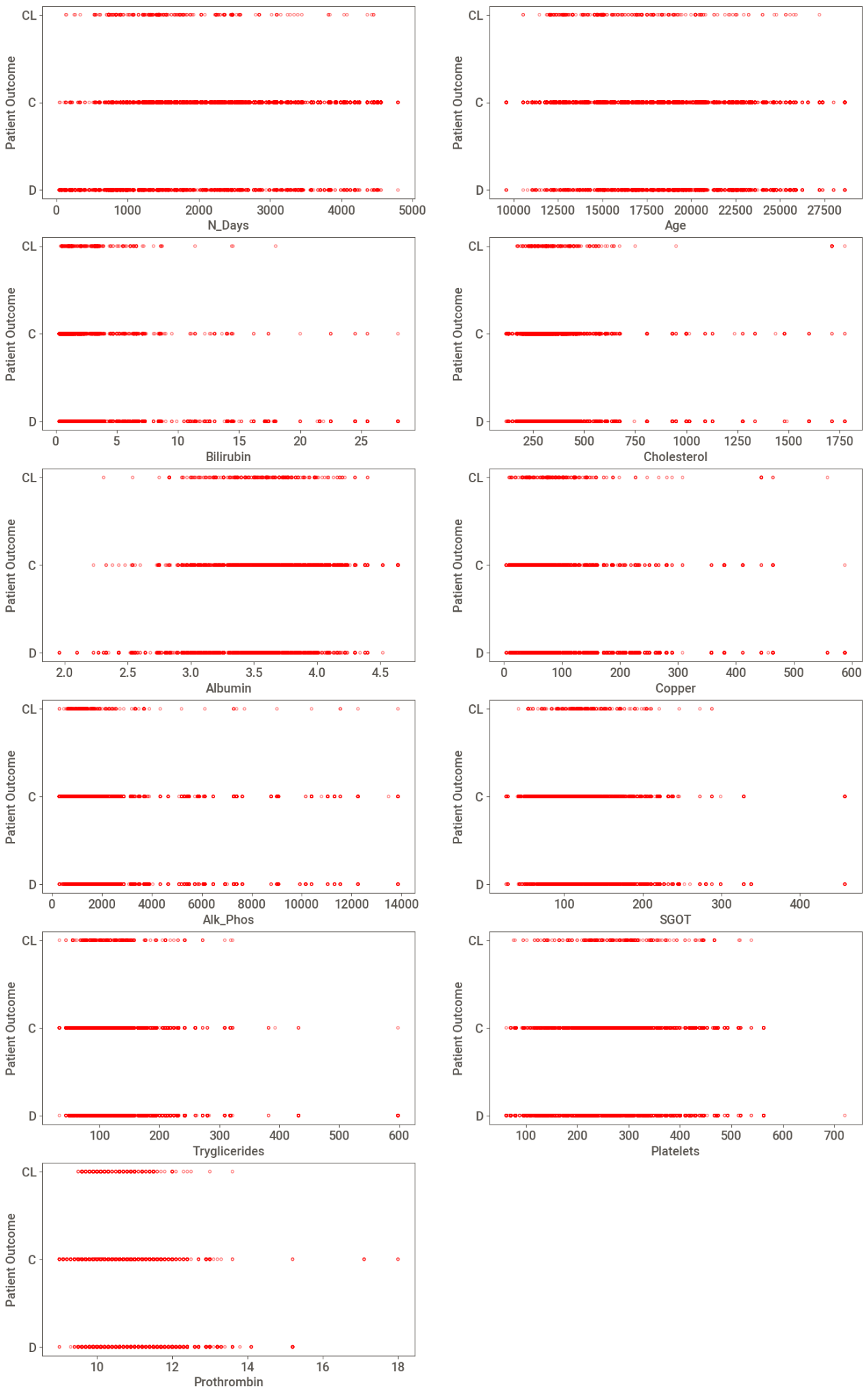
After an in-depth analysis, it became clear that there are no significant relationships between the numerical variables included in the data. Consequently, deletion of some numerical variables was not considered. Although this may seem ambiguous at first glance, it suggests that each numerical variable 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 using robust modeling techniques that can effectively accommodate missing data. Overall, 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.

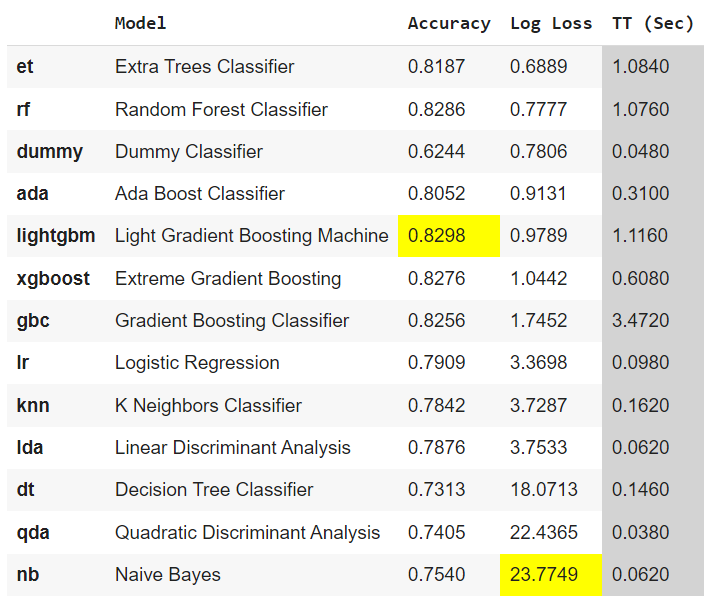
**EDA – Outlier Capping**

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, taking into account metrics such as accuracy and model stability. 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)

In our initial model evaluation with PyCaret, we used the AutoML function to comprehensively compare different machine learning models. To ensure fair evaluation, we used stratified k-fold cross-validation, which preserved the relationships between the resulting classes throughout the process. Log loss has emerged as a key evaluation metric due to its importance in probabilistic classification scenarios, allowing us to accurately assess forecast performance. As benchmarks, we chose Extra Tree Classifier and Random Forest Classifier, two widely used models known for their performance in various classification tasks. These models provide a solid starting point for measuring the effectiveness of more advanced algorithms and techniques.

In the future, we will focus on refining and improving these core models. This requires additional processing steps, including feature design, outlier detection, and resolution of potential unbalanced data problems. Tuning hyperparameters is also critical to optimizing model performance and fine-tuning their behavior to the nuances of the dataset. Ensuring the generalizability and robustness of the models remains a priority. The risks of over-installation are reduced by accurately validating invisible data and applying appropriate validation methods. In addition, we plan to explore advanced methods such as ensemble learning and feature selection to improve performance. Through iterative improvement and leveraging PyCaret's AutoML capabilities, we aim to develop highly accurate and robust predictive models specifically tailored to the requirements of our task. This iterative approach allows us to systematically improve model performance while ensuring scalability and maintainability throughout the development process.

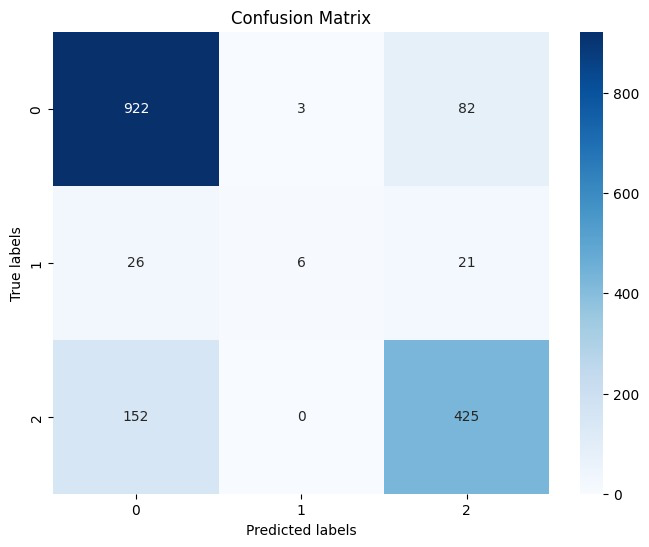
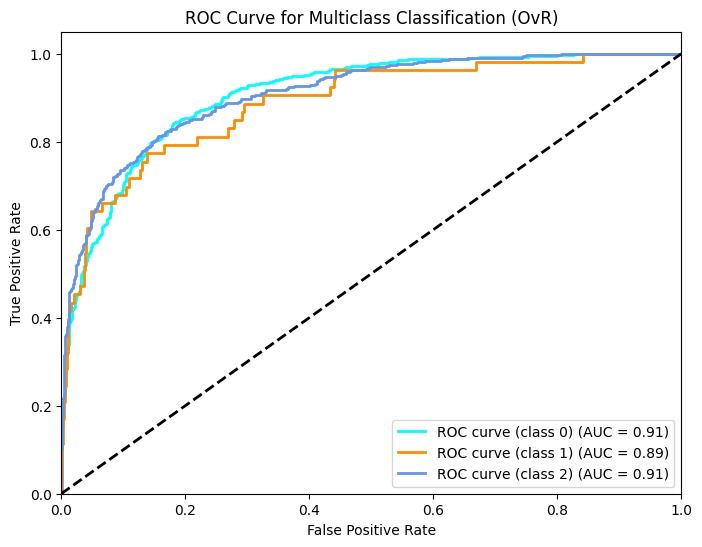
**Model Building**

**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 encoded. A MinMax scaler was applied to normalize the data and ensure all features are 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.

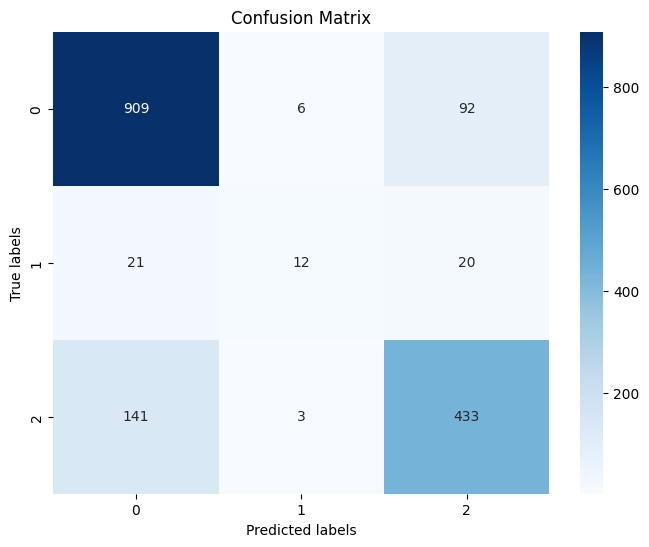
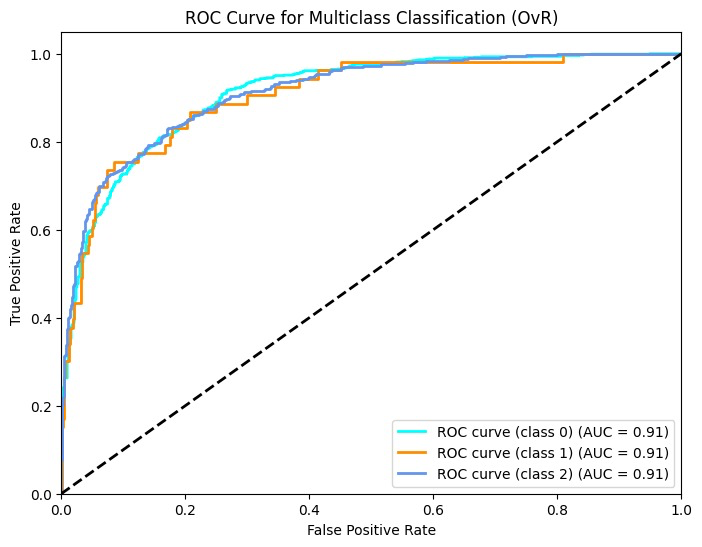
**Iteration 2:**

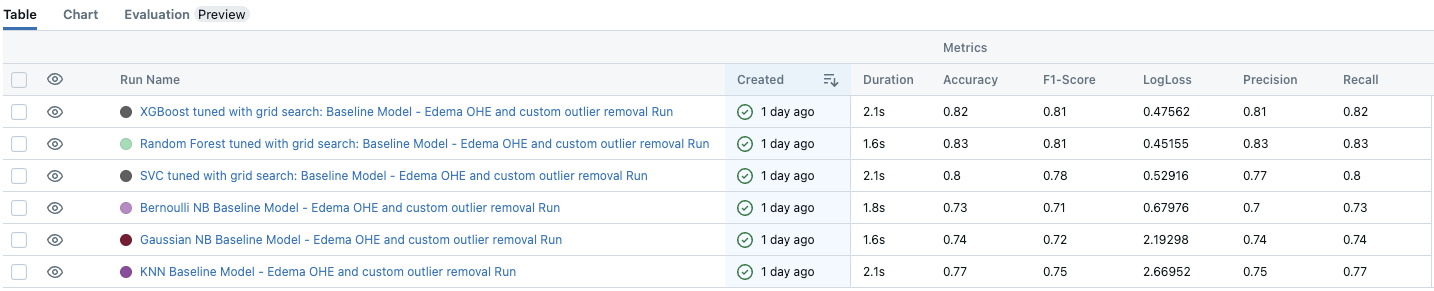
The dataset underwent several preprocessing steps, including categorical encoding, data scaling, and hyperparameter tuning. All columns were one-hot encoded, except for the target column, which was ordinal encoded. To standardize the range of values across the data, the MinMax scaler was applied to the 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.

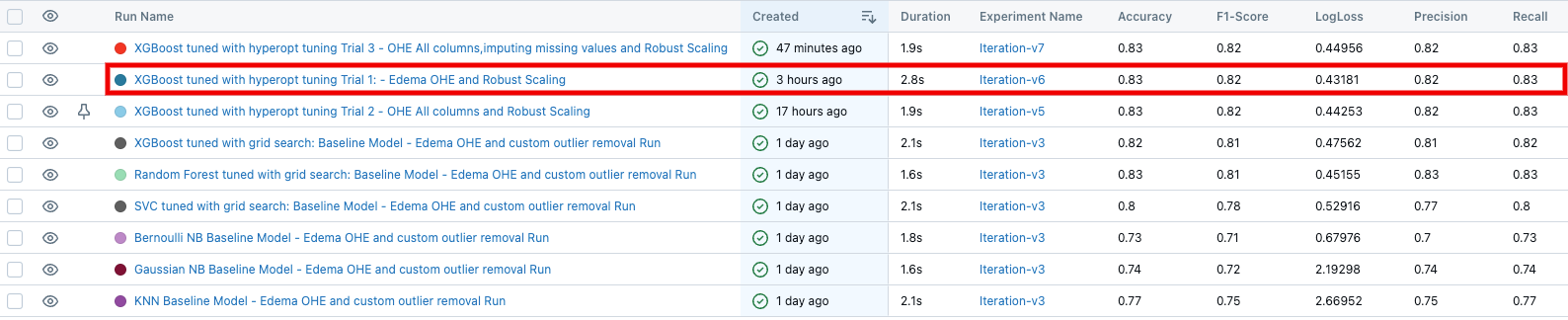
**Iteration 3:**

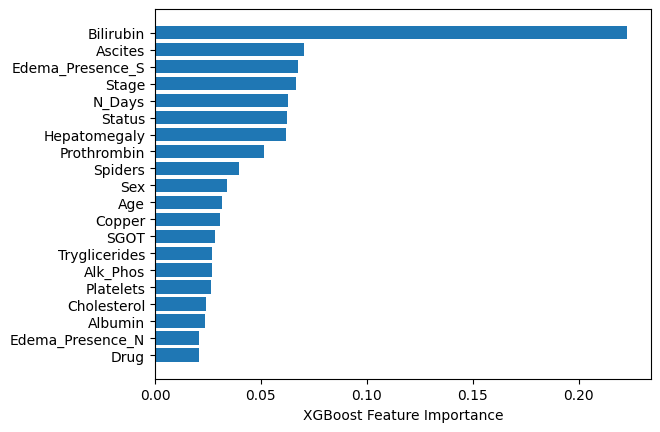
In preprocessing, categorical encoding involves utilizing one-hot encoding for the 'Edema' column and ordinal encoding for the 'Target' column. Scaling techniques are then applied, with MinMax scaling employed for 'Age' and 'N\_days', while a Robust Scaler is utilized for all other columns. Finally,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



**Model Evaluation**





**Conclusion**

At the conclusion of our project, we shed light on the multifaceted impact of predictive modeling in the selection of liver transplant patients. Through the integration of complex analysis of patient factors and advanced predictive algorithms, our project went beyond traditional approaches to facilitate a comprehensive framework for selecting ideal liver transplant candidates. One of the most important contributions of our project is to improve patient selection processes by prioritizing candidates based on post-transplant survival prognoses. By leveraging predictive modeling, we have empowered healthcare professionals to make informed decisions about patient eligibility and ensure that organs are allocated to those who will benefit the most, optimizing overall transplant outcomes. In addition, our project investigated the area of optimal timing of transplantation, taking into account the dynamic factors of disease progression and organ availability. This nuanced approach not only improves patient outcomes, but also promotes efficient use of healthcare resources, ultimately leading to better outcomes for transplant recipients. In addition to patient selection and transplant timing, our project extends its implications to the financial planning of healthcare organizations. Using the predictive power of our model, healthcare providers can estimate healthcare costs associated with various patient outcomes. This valuable information enables organizations to allocate resources efficiently, maximizing healthcare efficiency and ensuring financial sustainability. In addition, our project highlights the critical role of early intervention in improving patient outcomes. By using the model to identify patients at risk of worsening, healthcare professionals can proactively intervene with appropriate treatment, medication or lifestyle changes. This proactive approach not only prevents negative outcomes, but also reduces healthcare costs and improves patients' quality of life. Essentially, our project highlights the transformative potential of predictive modeling in liver transplantation. By integrating advanced analytics into healthcare decision-making processes, we have paved the way for more accurate, efficient and patient care. As we complete this project, we recognize its importance as a catalyst for continued innovation and progress in the field, with the ultimate goal of improving outcomes and improving the lives of liver transplant recipients worldwide.