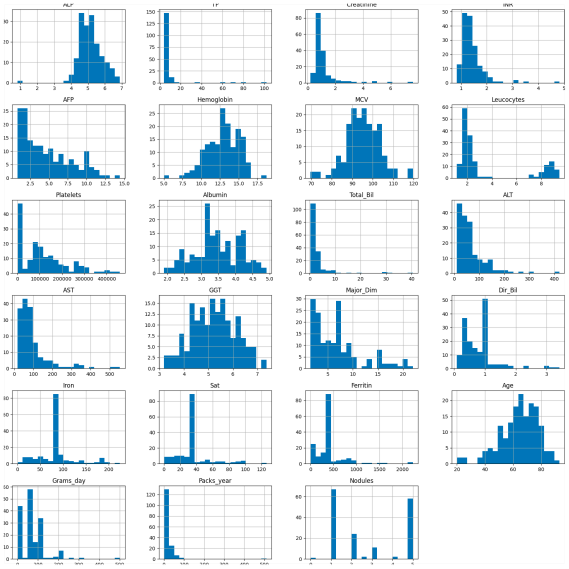
Data exploration and enrichment for supervised classification

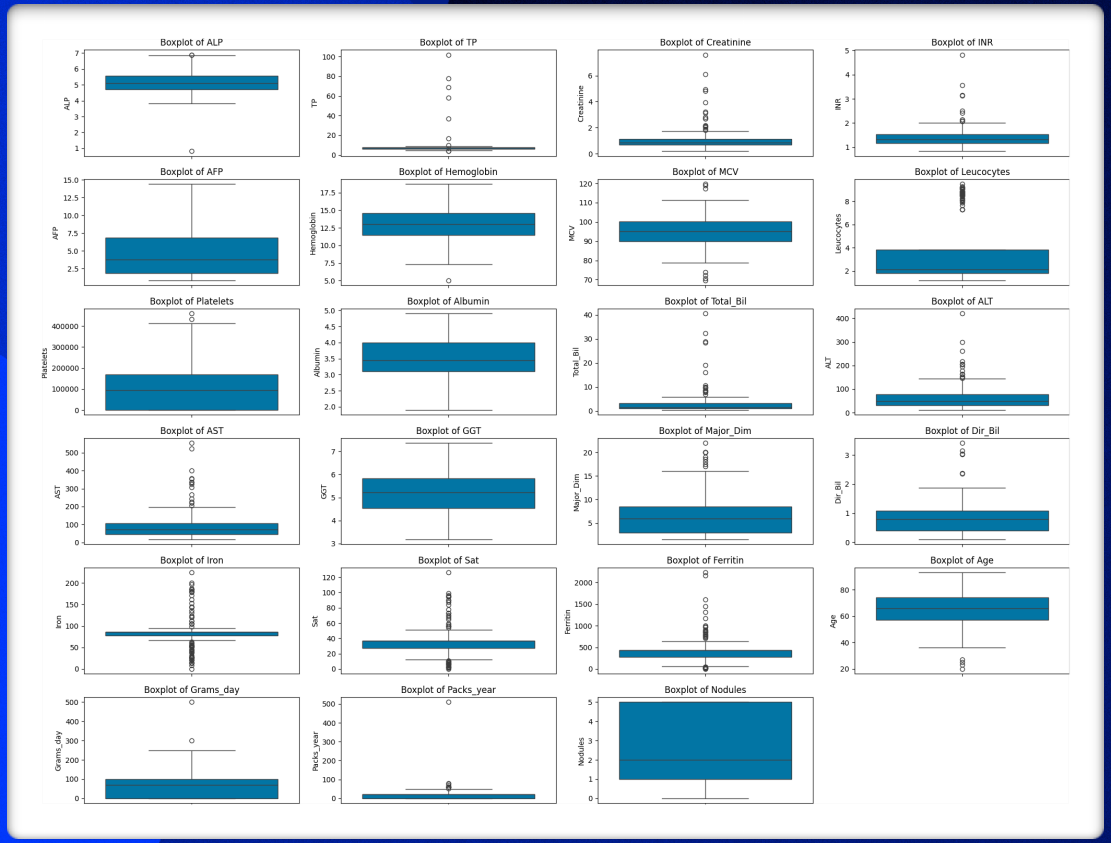
The Hepatocellular Carcinoma Dataset

Work realised by: Catarina Abrantes, Liliana Silva and Mariana Fonseca

https://github.com/CatarinaAbrantes/Trabalho_HCD/tree/main

Data exploration

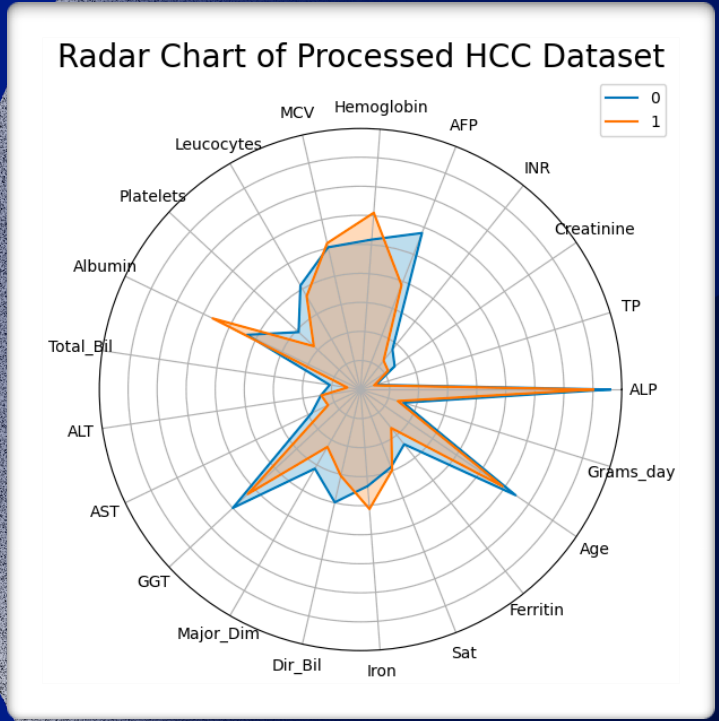


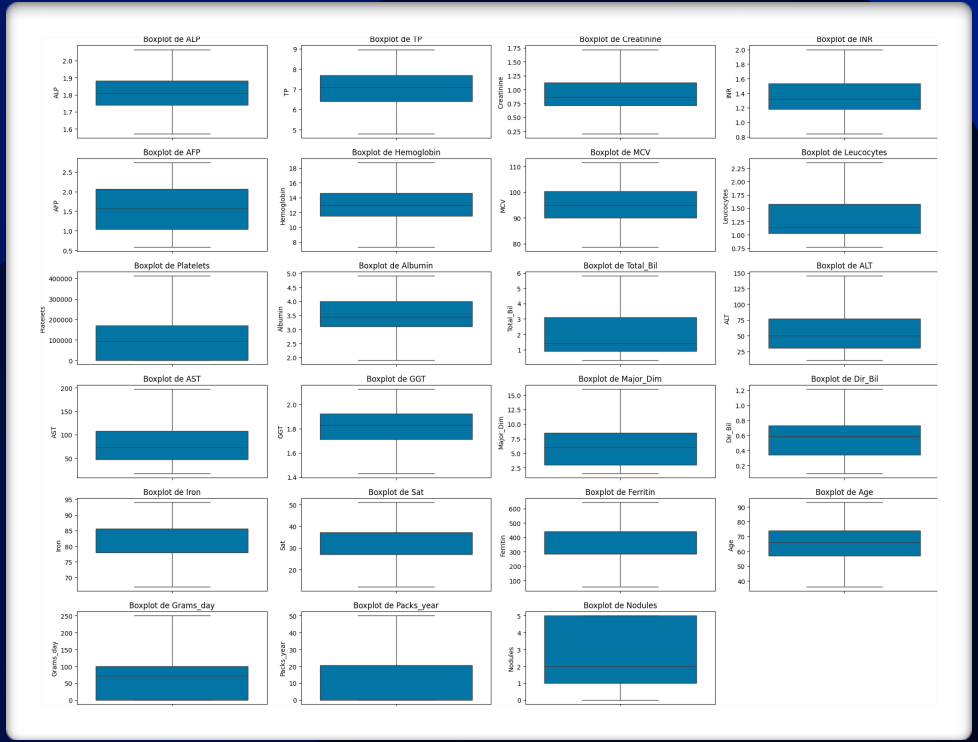


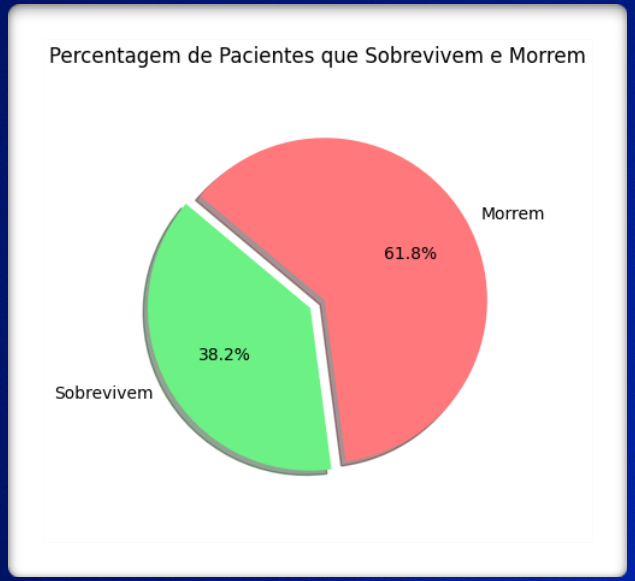
- Elderly Patients
- Male
- Cirrhosis and associated symptoms, such as varices and spleno.
- Significant alcohol and tobacco consumption.
- Approximately 62% of individuals survive.

Data Preprocessing

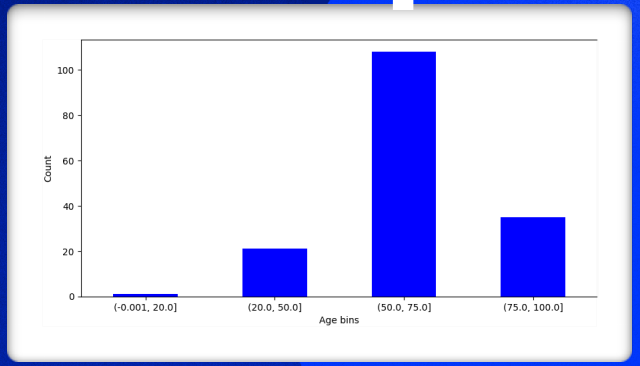
1. Handle missing values
2. Remove outliers

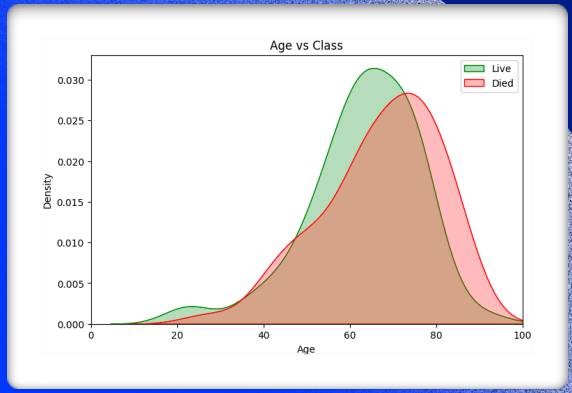


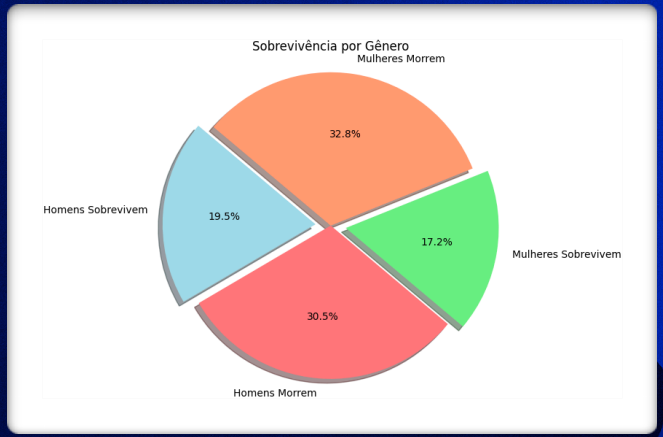


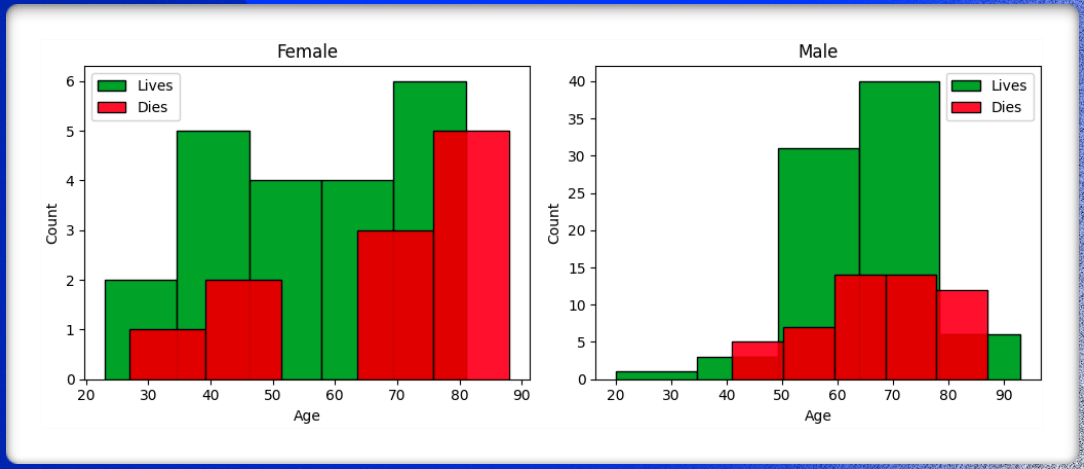


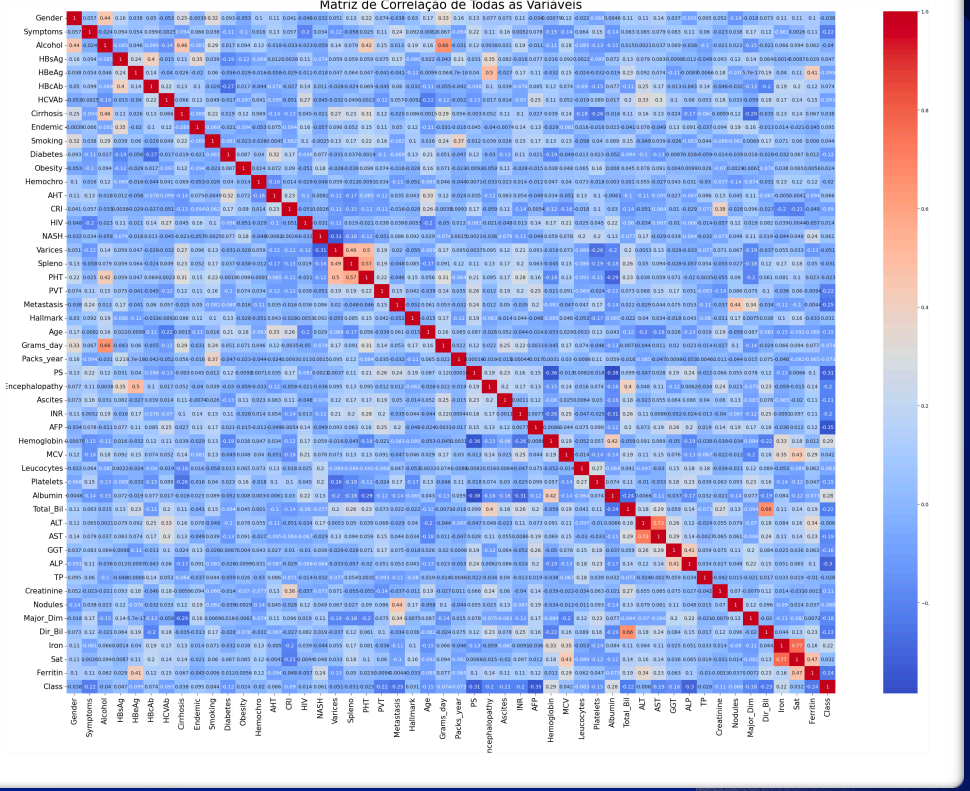
Most patients are in this age range.

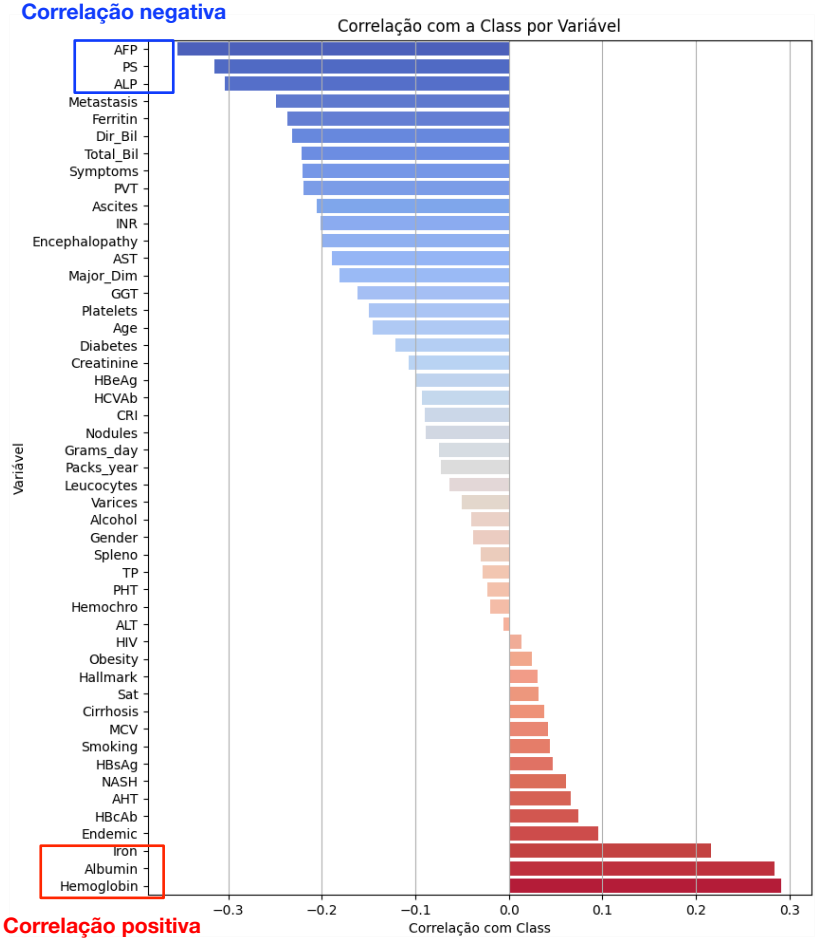




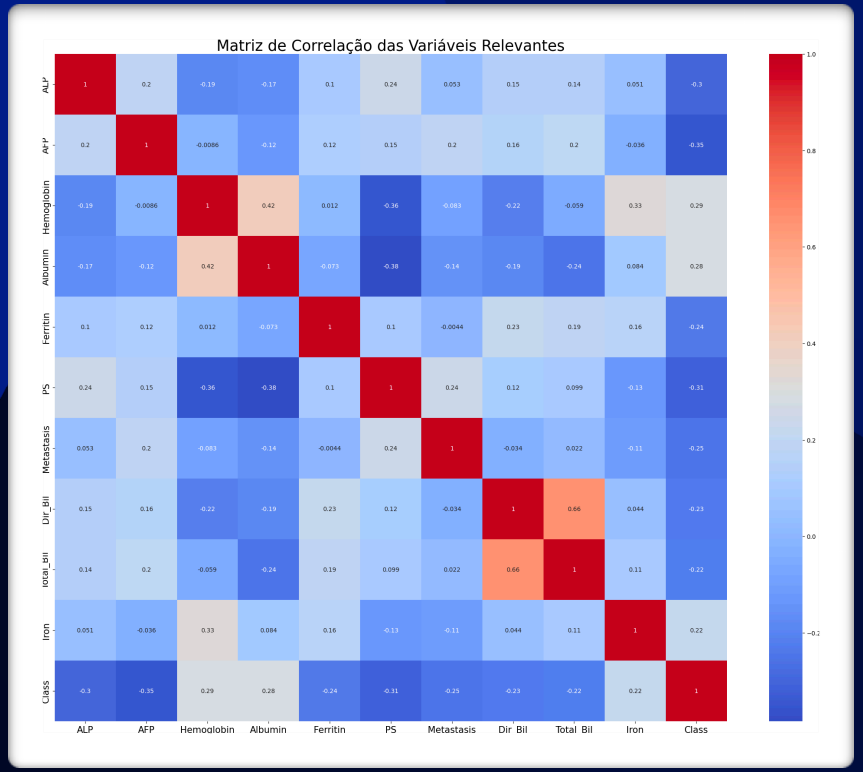


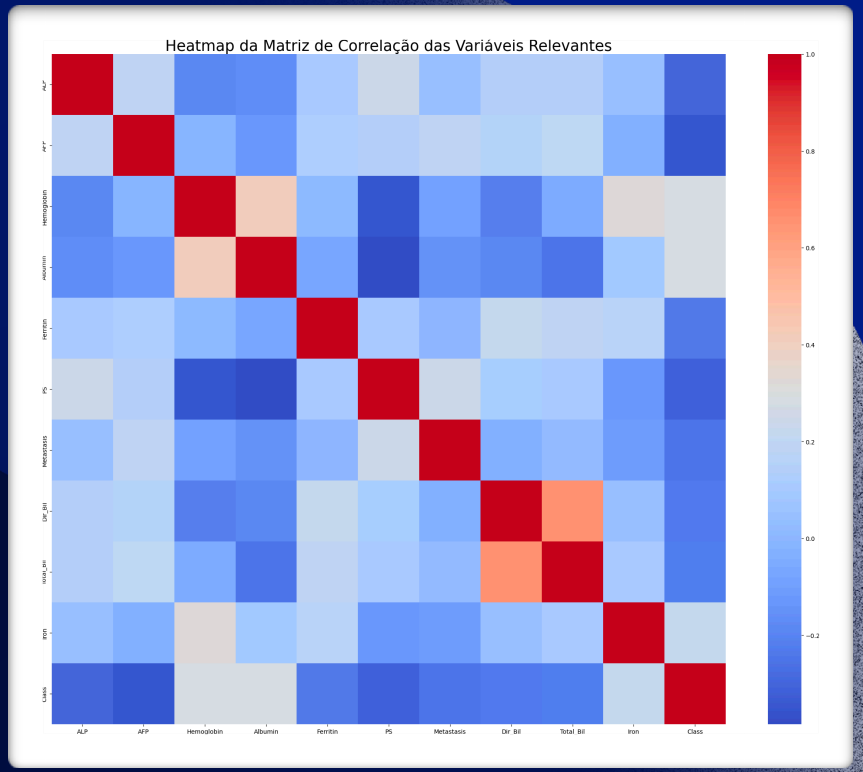






- **AST, ALT, GGT, Total_Bil, and Albumin have significant correlations with each other;**
- **Age and Class have a moderate correlation.**



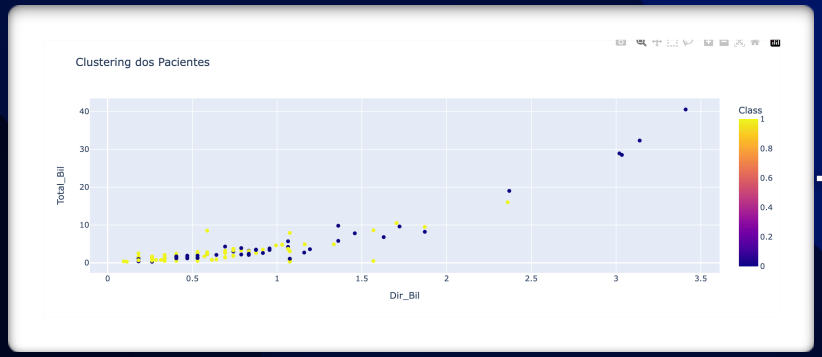


Highly Correlated Variables: Dir_Bil and Total_Bil provide redundant information due to their high correlation.

Important Variables for Survival: ALP, AFP, Hemoglobin, PS, Metastasis, and Albumin are crucial for predictive analysis of patient survival.

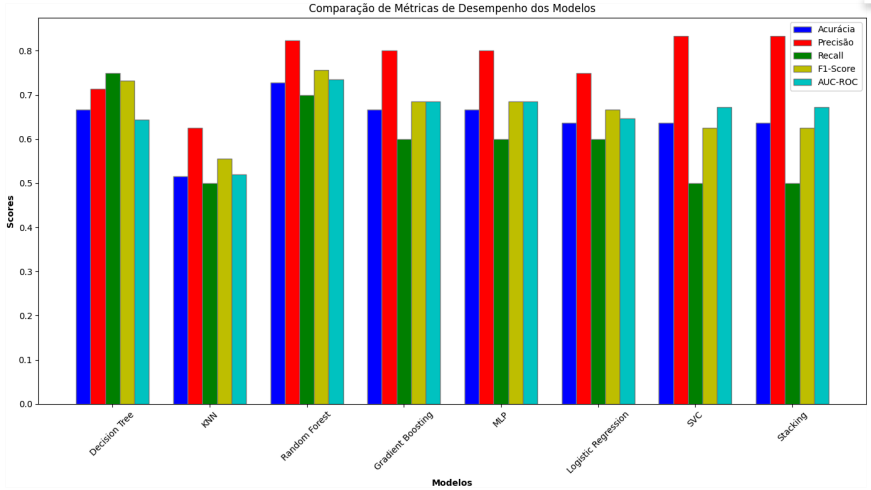
Data Modeling (Supervised Learning):

Model	Description
Decision Trees	Classify if the patient lives or dies based on rules derived from the data.
K-Nearest Neighbors (KNN)	Classify the patient based on the k nearest neighbors.
Random Forest	Build multiple decision trees and combine their results to improve accuracy.
Gradient Boosting	Improve prediction by sequentially correcting errors of previous models.
Multi-Layer Perceptron (MLP)	Neural network that learns complex representations to predict the outcome.
Logistic Regression	Model the probability of the patient living or dying.
Stacking Classifier	Combine predictions from various base models to improve final accuracy.
Support Vector Classifier (SVC)	Find the hyperplane that best separates the classes to predict the outcome.



Relate the two variables that have the highest correlations with each other and with the target column.

Data Evaluation



Best overall performance: Random Forest

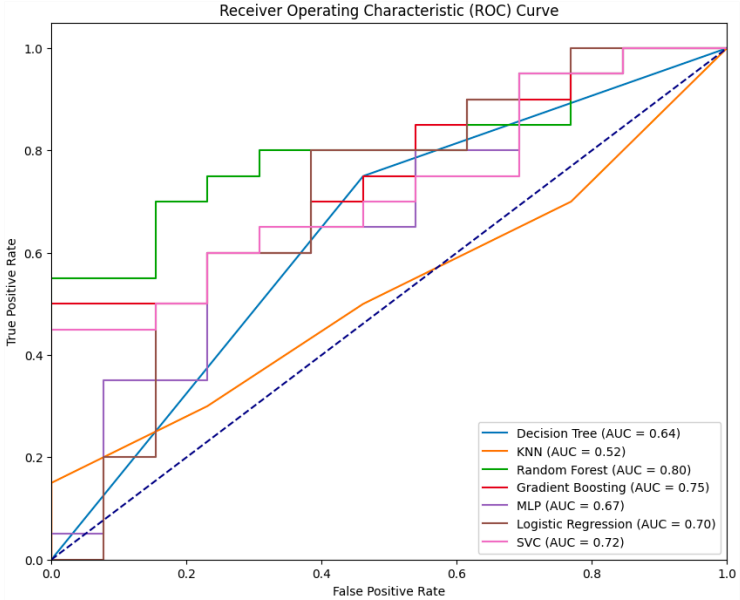
Second best option: Gradient Boosting

Third best option: K-Nearest Neighbors (KNN)

Fourth best option: Decision Tree

Priority on accuracy: Logistic Regression, Support Vector Classifier (SVC)

Priority on recall: Decision Tree

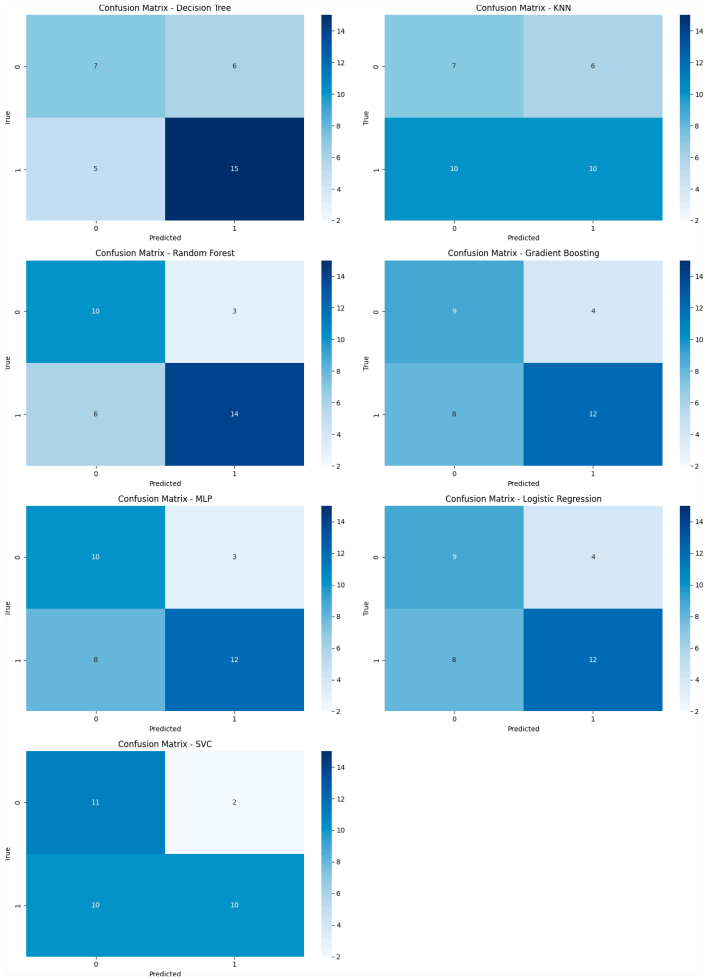


Best performance (based on ROC curve): Random Forest

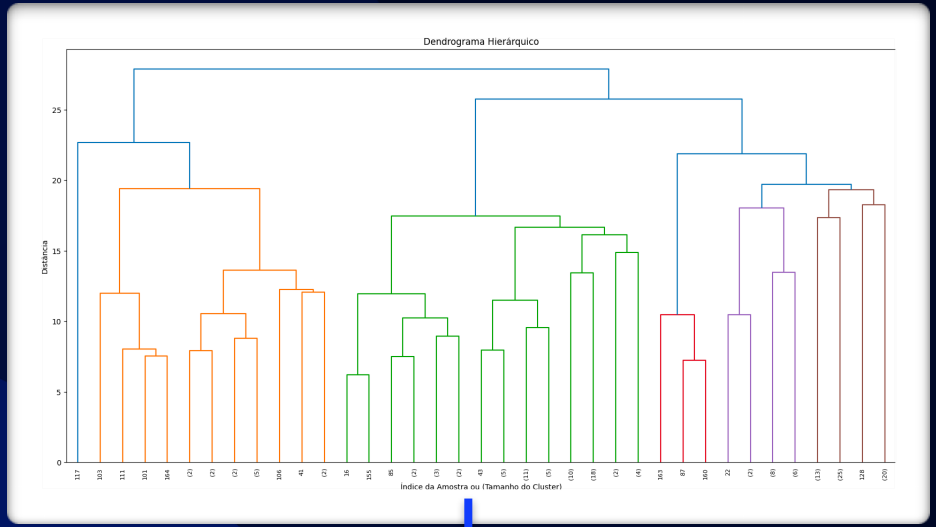
Second best performance: Gradient Boosting

Third best performance: Logistic Regression

Worst performance: Decision Tree, K-Nearest Neighbors (KNN)




```
Resultados da Validação Cruzada:
Random Forest: 0.74
Gradient Boosting: 0.70
SVC: 0.60
```



Y-Axis Proximity:

- **The closer on the Y-axis, the greater the similarity between clusters.**

Dendrogram Merging:

- **The closest clusters in the dendrogram are the first to be merged, forming larger groups.**

Random Forest: 🌟 The most robust model 📈 High accuracy rate ❌ No errors

KNN: 🚀 Promising performance ⚠️ Attention needed for false negatives

MLP: 🔧 Requires adjustments 📉 Needs improvement in performance

Interpretation of Results

- **Random Forest:** Showed the best performance with a high accuracy rate and no significant errors, making it the most robust model for this task.

- **KNN:** Demonstrated promising performance, although with some false negatives that need to be considered.

- **Gradient Boosting and Logistic Regression:** Both showed reasonable performance, with some true positives and few false negatives.

- **MLP:** Had inferior performance with a higher number of false negatives, indicating the need for adjustments.

- **Decision Tree:** Moderate performance, simple to interpret but less effective compared to ensembles.

- **Stacking Classifier:** Combined the strengths of multiple models, showing a good balance between accuracy and robustness.