Data exploration and enrichment for supervised classification

Adriano Pires and Miguel Costa

1 - Specification of the work

- **Objective of the project:** The aim of this project is to carry out a comprehensive and professional analysis of the data related to patients with hepatocellular carcinoma in order to develop a machine learning pipeline capable of determining the survival of patients one year after diagnosis (i.e. "live" or "die").

- This work will use a holistic approach and advanced data analytics and machine learning techniques to provide valuable information on the factors that influence the survival of patients with HCC. The combination of detailed data exploration, robust pre-processing, advanced modelling and comprehensive evaluation will ensure reliable results. The implementation of the solution in an interactive web application will make the results accessible and easy to understand for end-users.

2 - Related works

- **KAGGLE:** Kaggle is a web platform that brings together the world's largest Data Science community, with more than 536,000 active members in 194 countries, receiving more than 150,000 posts per month, providing all the most important tools and resources to make the most progress in data science (i.e. Titanic Disaster competition).

- **Santos, M. et al. “A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients.” Journal of biomedical informatics 58 (2015): 49– 59**

- **Chicco, D. et al. “Computational intelligence identifies alkaline phosphatase (ALP), alphafetoprotein (AFP), and hemoglobin levels as most predictive survival factors for hepatocellular carcinoma.” Health Informatics Journal 27.1 (2021).**

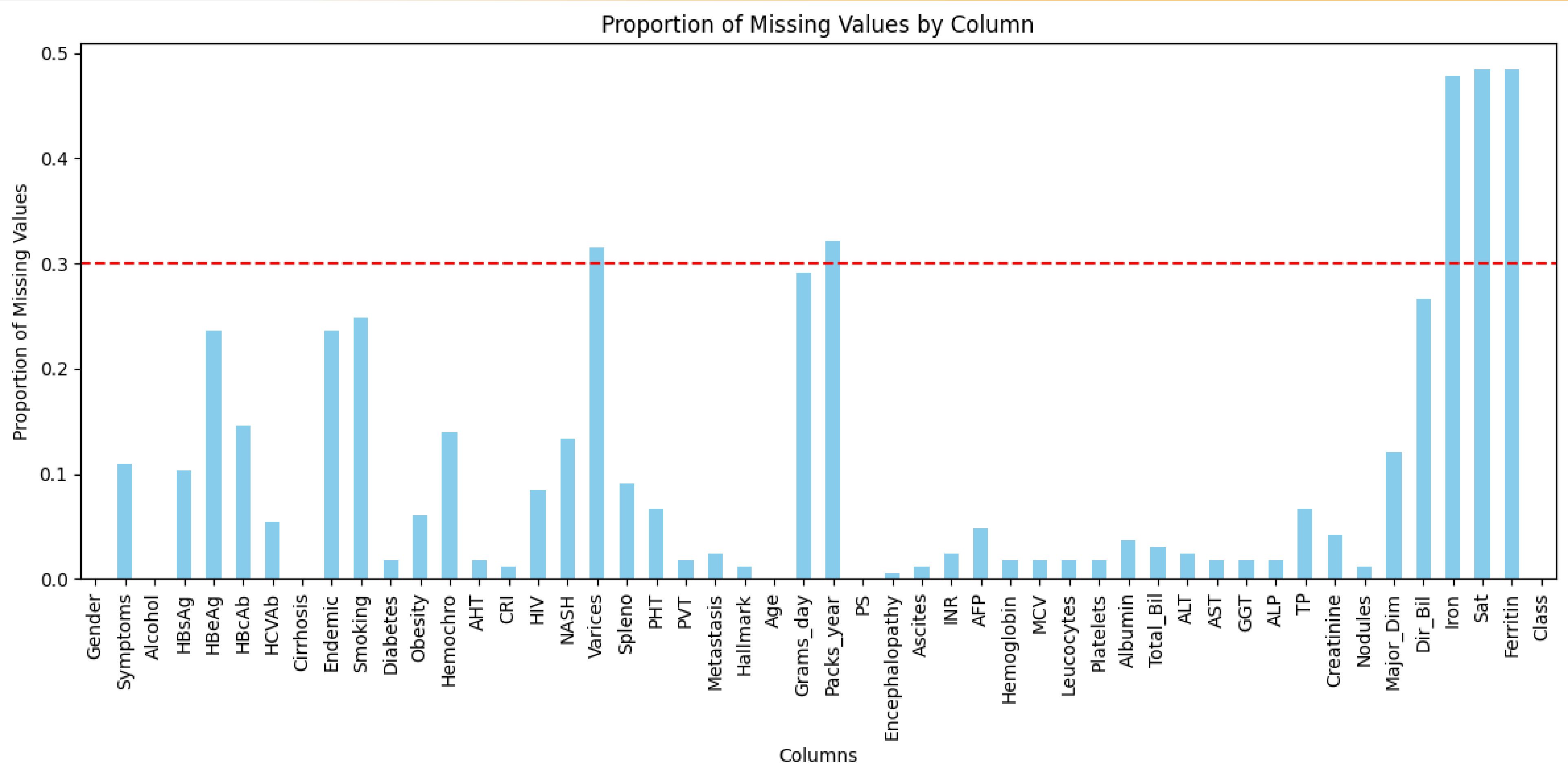
3 - Tools and algorithms

- **Programming language:** Python
- **Data analysis libraries:** Pandas, NumPy
- **Visualisation libraries:** Matplotlib, Seaborn - Machine learning libraries: Scikit-learn
- **Decision tree algorithms:** DecisionTreeClassifier, RandomForestClassifier - KNN Algorithm: KNeighborsClassifier
- **Class balancing techniques:** SMOTE, RandomUnderSampler
- **Evaluation metrics:** accuracy_score, precision_score, recall_score, f1_score, roc_auc_score, confusion_matrix
- **Validation techniques:** train_test_split, train_test_split, roc_auc_score, confusion_matrix
- **Hyperparameter optimisation:** GridSearchCV, RandomizedSearchCV, BayesianOptimization
- **Implementation tools:** Streamlit
- **Other algorithms:** LogisticRegression, SVC, GradientBoostingClassifier, XGBClassifier, MLPClassifier

4 - Data exploration

- The HCC database is made up of 50 variables (columns) and multiple patient records (rows). Each row represents a patient diagnosed with hepatocellular carcinoma, and the columns contain information on the demographic, clinical and laboratory characteristics of these patients.

- We differentiate between categorical and numerical variables.
- We calculate the proportion of null values per column.
- Using the correlation matrix we can see some significant correlations, which may indicate redundancies or important relationships.



5 - Data Preprocessing

- Columns with more than 30 per cent missing values were removed.
- Missing values in the remaining columns were filled in with the median(numerical variables) or mode (categorical variables).
- We have detected the outliers and applied the Winsorizing technique.

Winsorizing is a statistical technique used to reduce the impact outliers in a data set. The central idea is to limit the values of a distribution to reduce the influence of outliers, replacing them with less extreme values. This can be particularly useful when you want the outliers not to distort the descriptive statistics and the results of the analyses

- We create new variables by linking some variables that are highly correlated.
- We use a **Random Forest model** to identify the most important characteristics.

6 - Data Modeling

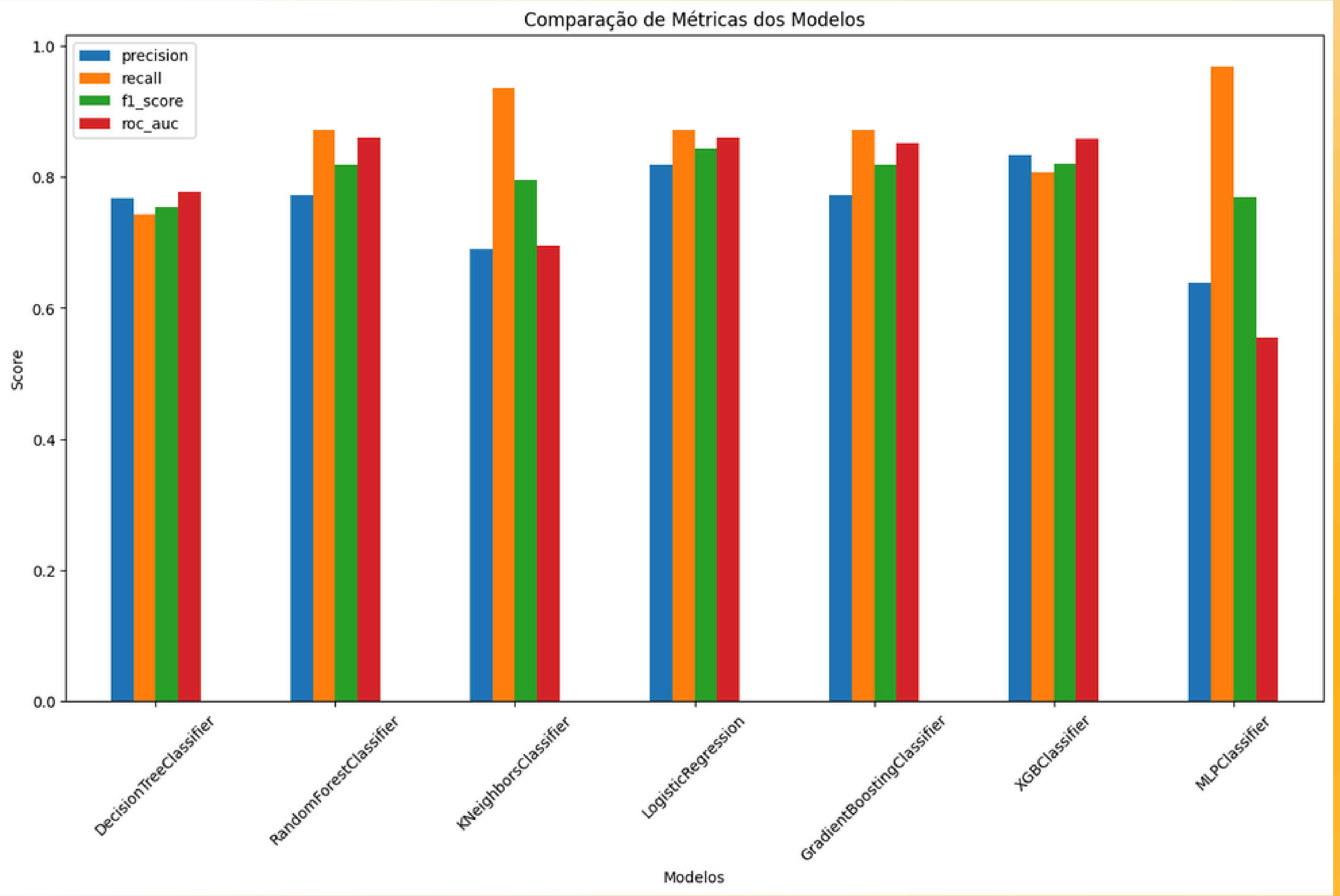
Let's proceed with the modelling step . Here are the steps we follow:

- **Model selection:** Use the following initial models: DecisionTreeClassifier, RandomForestClassifier, KNeighborsClassifier, LogisticRegression, SVC, GradientBoostingClassifier, XGBClassifier, MLPClassifier.
- **Model training:** Train each model using the training sets.
- **Hyperparameter optimisation:** Use GridSearchCV, RandomizedSearchCV and Bayesian Optimisation to find the best hyperparameters.

7 - Model Evaluation

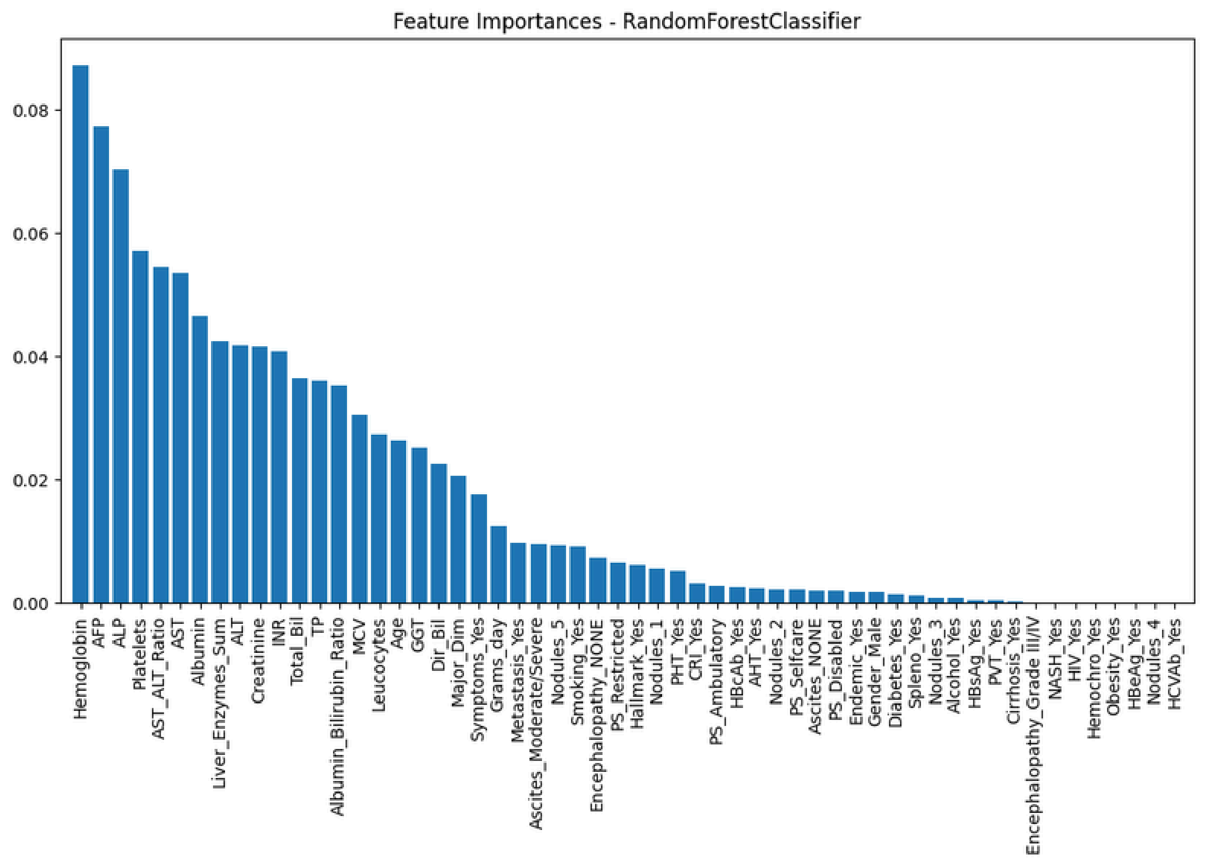
We have evaluated the models using the following metrics:

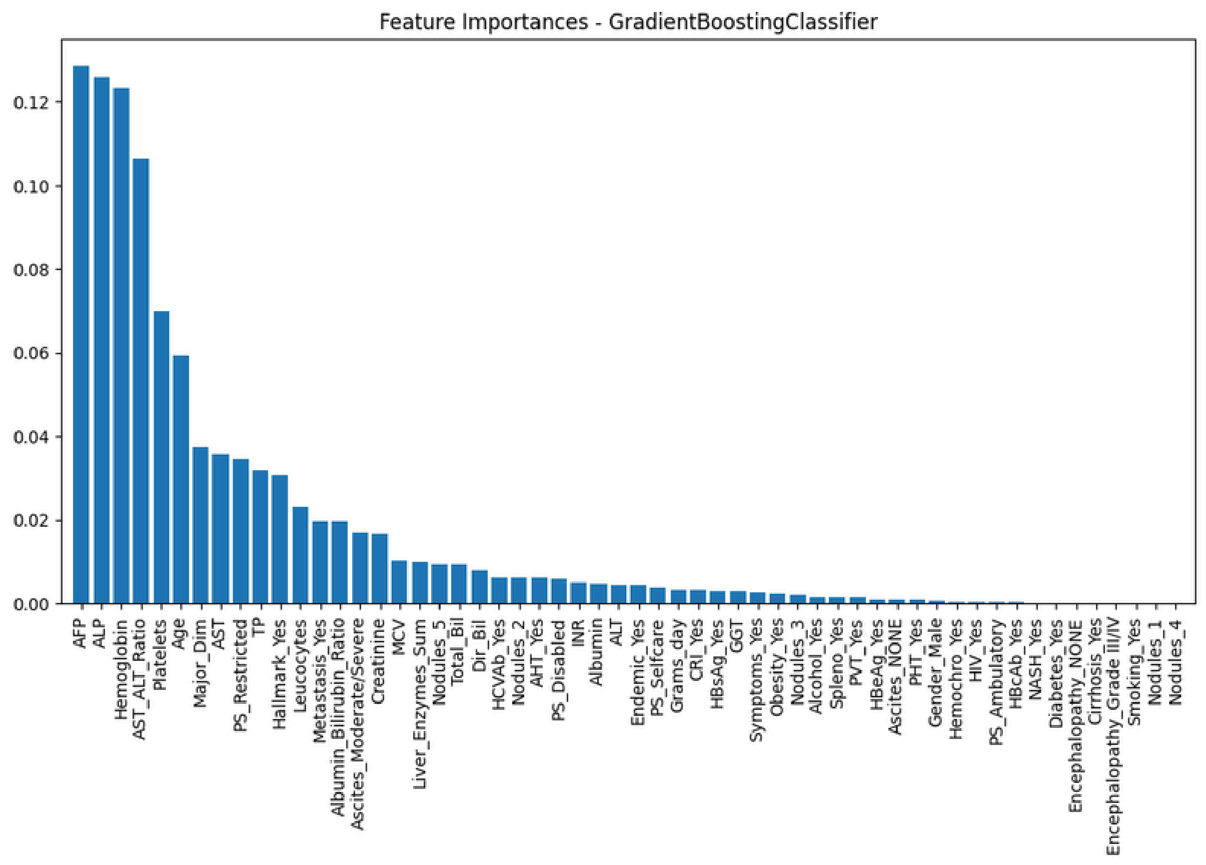
- **Accuracy:** Proportion of correct predictions out of total predictions. A basic metric, but can be misleading in unbalanced data sets.
- **Precision:** Proportion of true positives out of all positive predictions. High precision means few false positives.
- **Recall:** Proportion of true positives among all actual occurrences of the positive class. High recall means few false negatives.
- **F1-score:** Harmonic mean of precision and recall. A good balance between precision and revocation.
- **ROC AUC (Area under the ROC curve):** Measures the model's ability to distinguish between classes. Values closer to 1 indicate a better model.

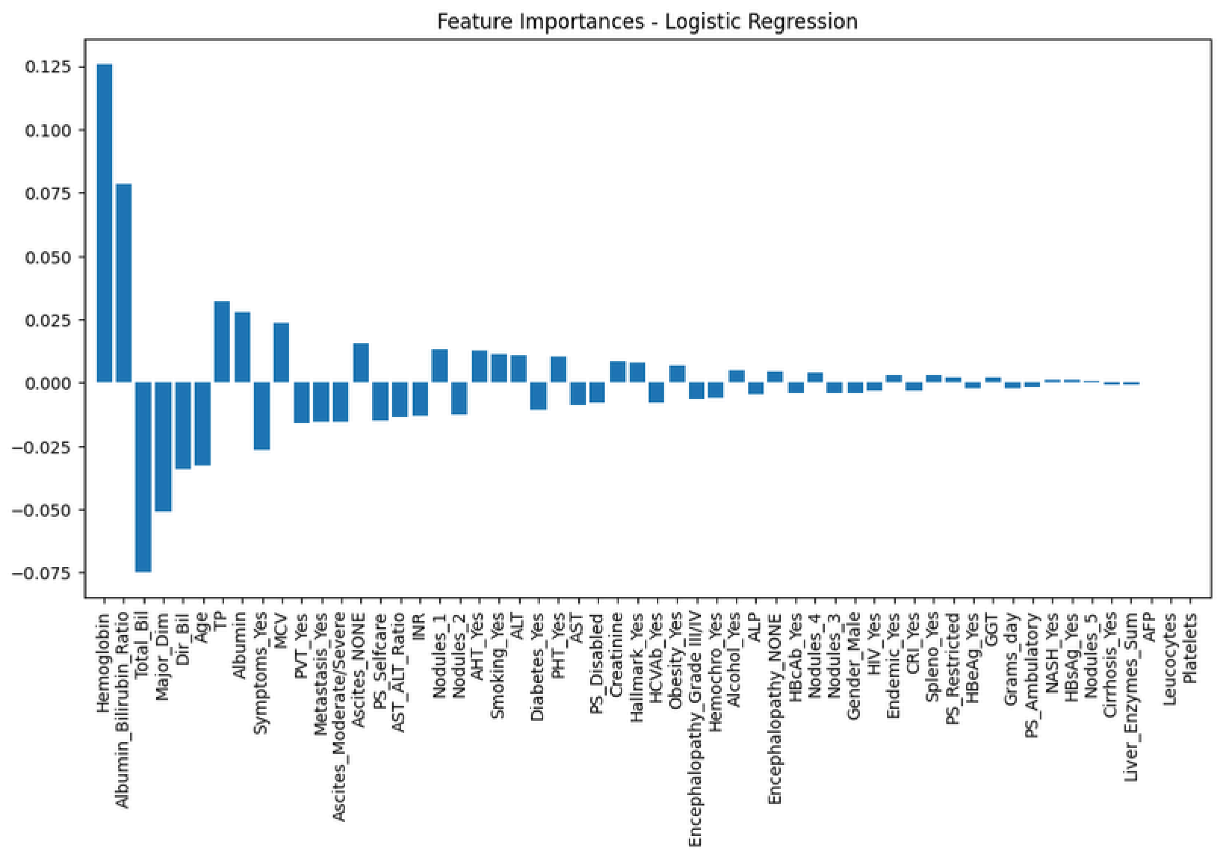


7 - Model Evaluation

- Based on performance metrics (precision, recall, F1-score, ROC AUC) and cross-validation, the best models were: **RandomForestClassifier**, **LogisticRegression** and **GradientBoostingClassifier**.
- All these models showed a good balance between precision, recall and F1-score, as well as an ROC AUC above 0.85.
- Looking at the feature significance plots for the three best models we see which variables are the most influential in predicting survival in patients with hepatocellular carcinoma (HCC).







8 - Interpretation of Results

- **Identification of Important Factors:** The most important variables identified by the models are Haemoglobin, AFP, ALP, AST_ALT_Ratio, Platelets, Albumin, Liver_Enzymes_Sum, ALT and Creatinine
- These variables should be the focus of future analyses to better understand how they influence the survival of HCC patients

- **Relevant insights:**

1. *Liver Function:* Variables related to liver function (AFP, ALP, AST, ALT, Albumin) are crucial for survival, which is in line with medical literature.
2. *General condition of the patient:* Indicators such as haemoglobin and creatinine, which reflect the general condition of the patient, are important.
3. *Tumour size:* The size of the largest liver nodule is a significant factor.