Data exploration and enrichment for supervised classification

Assignment No. 2 Checkpoint n°1

Daniel Gomes up202306411
Gonçalo Pereira up202304057
Inês Castro up202304060

Work performed

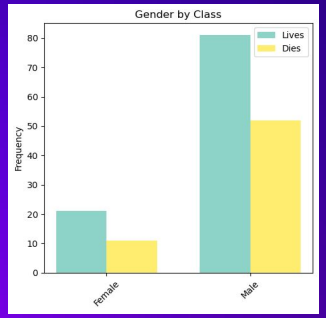
Using the Hepatocellular Carcinoma (HCC) dataset, we want to develop a machine learning pipeline capable of determining the survivability of patients at 1 year after diagnosis, for example “lives” or “dies”.

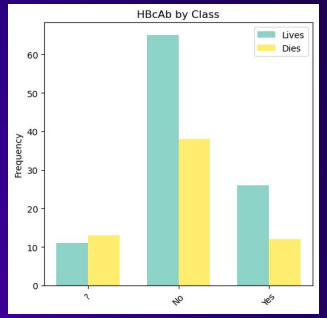
Then we followed the following steps:

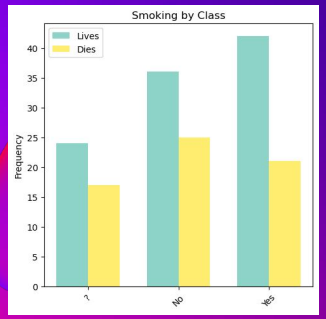
- Data Exploration;
- Data Preprocessing;
- Data Modeling and Training;
- Data Evaluation;
- Interpretation of Results;

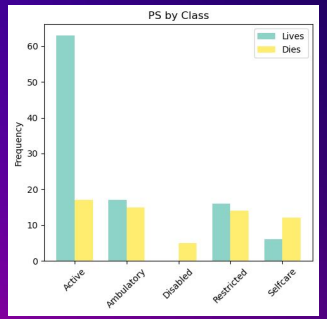
Data Exploration

Attribute Analysis:
-Categorical Attributes;

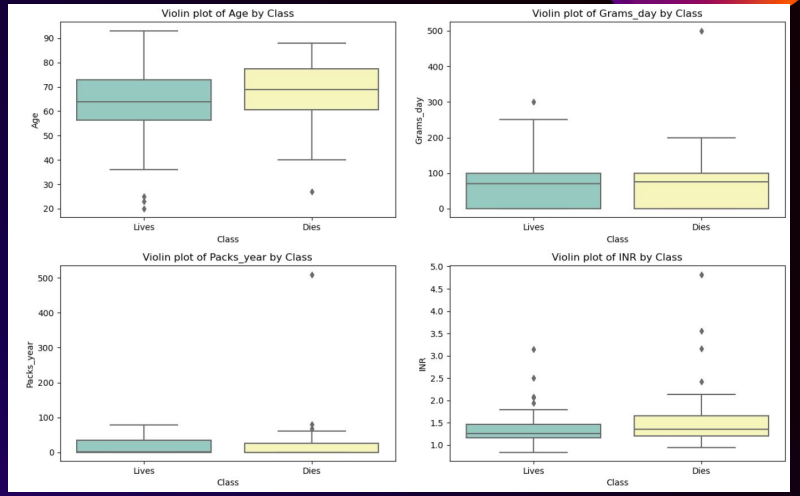








-Numerical Attributes;



Data Preprocessing

This stage involves applying a series of techniques and transformations to the raw data with the aim of preparing it for analysis and modeling.

- We exclude the Variable 'Dir_Bil' due to: High correlation;
- Imputation of missing values “?” by “NaN”;
- Replacing of the categorical variables by numerical values;
- Exchanging the NaN’s by mean, mode, maximum value and minimum value. In the categorical attributes it doesn't make sense using the mean, therefore we have to separate in two cases. The first one is the categorical attributes where mode, maximum value and minimum value are used. The second one is the rest of the attributes where mean, mode, maximum value and minimum value are used. This process will create about 12 different data sets.

Data Modeling

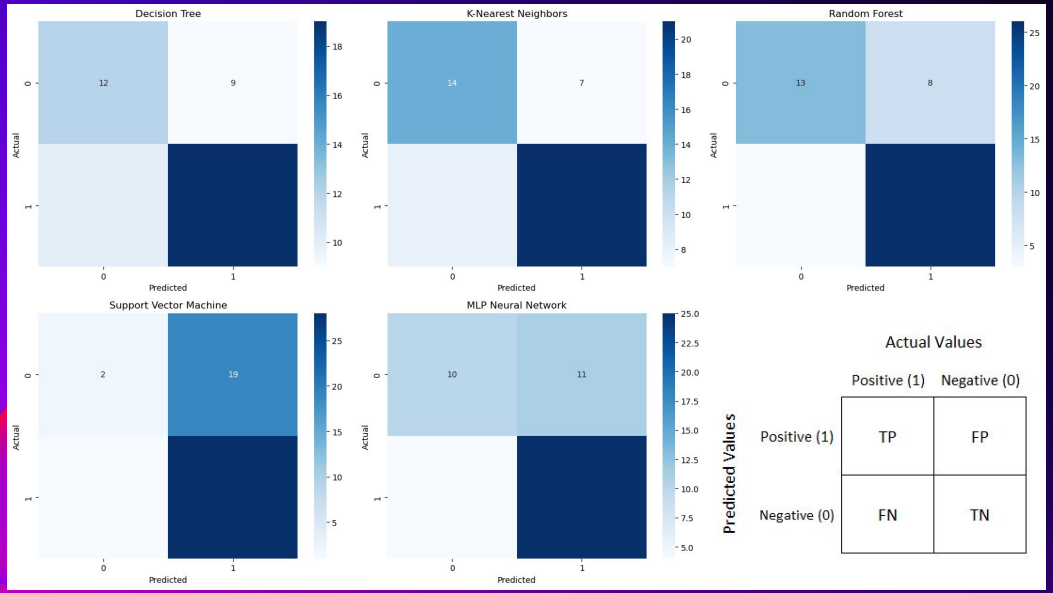
The decision tree, KNN ,Random Forest ,Support Vector Machine and Multi-layer perceptron neural network with dropout model are being compared based on accuracy, but also the different data sets created previously.

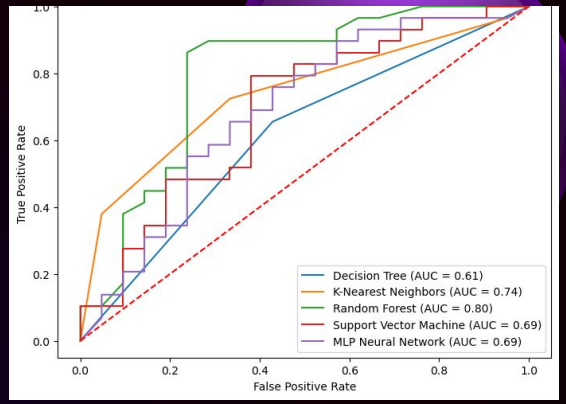
	Mode and Mean	Mode and Mode	Mode and Maximum	Mode and Minimum	Maximum and Mean	Maximum and Mode	Maximum and Maximum	Maximum and Minimum	Minimum and Mean	Minimum and Mode	Minimum and Maximum	Minimum and Minimum
Decision Trees	0.64	0.68	0.66	0.64	0.56	0.74	0.58	0.66	0.66	0.72	0.7	0.66
KNN	0.62	0.66	0.58	0.66	0.62	0.66	0.58	0.66	0,62	0.66	0.58	0.66
Random Forest	0.78	0.74	0.72	0.76	0.78	0.74	0.68	0.78	0.74	0.74	0.72	0.76
Support Vector Machine	0.56	0.58	0.56	0.58	0.56	0.58	0.56	0.58	0.56	0.58	0.56	0.58
Multi-layer perceptron neural network with dropout	0.62	0.62	0.62	0.60	0.62	0.62	0.62	0.60	0.62	0.62	0.62	0.60
Mean	0.64	0.656	0.63	0.648	0.628	0.668	0.604	0.656	0.645	0.664	0.636	0.652

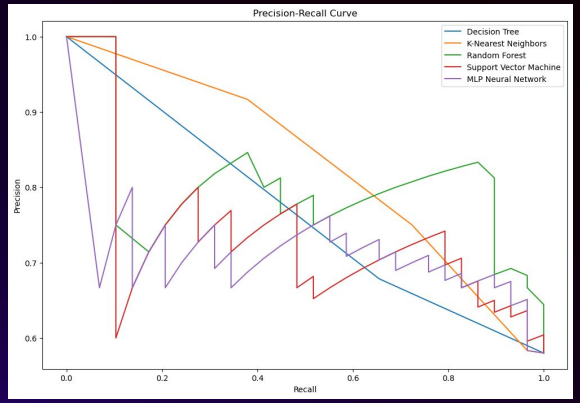
The accuracy from all these different datasets will be obtained from the 5 different supervised learning algorithms. Using the maximum value and the mode in all the NaN values provides in mean the most accurate results.

Data Evaluation

- Precision;
- Recall;
- ROC/AUC;
- Confusion matrix.

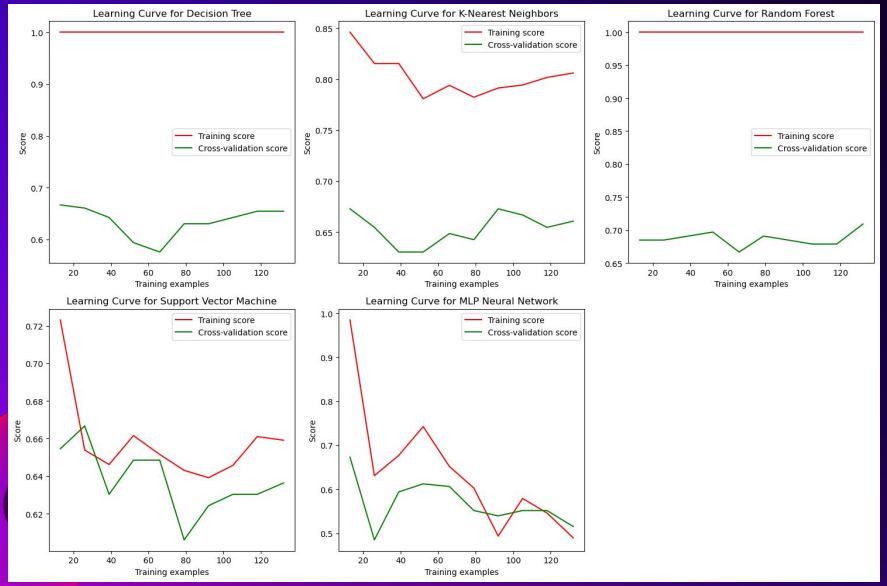






Data Evaluation

- Feature Importance;
- Performance During Learning;



	Decision Tree	K-Nearest Neighbors	Random Forest	Support Vector Machine	MLP Neural Network
AFP	0.244900	0.044	0.076202	0.018	-2.000000e-02
Leucocytes	0.111458	-0.008	0.047101	0.000	-8.000000e-03
Iron	0.097412	0.000	0.042955	0.000	-2.600000e-02
Total_Bil	0.078683	0.000	0.038311	0.000	0.000000e+00
MCV	0.072803	0.000	0.040890	0.000	-2.200000e-02
Ferritin	0.067961	0.002	0.053006	0.000	8.000000e-03
INR	0.053358	0.000	0.034471	0.000	-1.000000e-02
ALT	0.045494	0.000	0.025243	0.000	0.000000e+00
Albumin	0.036675	0.000	0.040281	0.000	2.000000e-03
TP	0.035080	0.000	0.032500	0.000	-2.000000e-03
PS	0.034829	0.000	0.020031	0.000	-1.000000e-02
Packs_year	0.034623	0.000	0.016609	0.000	-4.000000e-03
Encephalopathy	0.033587	0.000	0.007962	0.000	-4.000000e-03
Hemoglobin	0.028131	0.000	0.059249	0.000	-6.000000e-03
Hallmark	0.025005	0.000	0.006232	0.000	0.000000e+00
Ascites	0.000000	0.000	0.016506	0.000	-8.000000e-03
Sat	0.000000	0.000	0.031381	0.000	0.000000e+00
Creatinine	0.000000	0.000	0.023722	0.000	-4.000000e-03
Major_Dim	0.000000	0.000	0.037373	0.000	-1.200000e-02
Platelets	0.000000	0.040	0.045749	0.016	-1.600000e-02
AST	0.000000	0.000	0.032451	0.000	1.600000e-02
GGT	0.000000	0.024	0.028369	0.000	-1.200000e-02
ALP	0.000000	0.022	0.066927	0.000	-1.200000e-02
Nodules	0.000000	0.000	0.010855	0.000	-8.000000e-03
Gender	0.000000	0.000	0.006821	0.000	-4.000000e-03
Symptoms	0.000000	0.000	0.006692	0.000	2.220446e-17
Obesity	0.000000	0.000	0.001970	0.000	-4.000000e-03
Alcohol	0.000000	0.000	0.007580	0.000	-6.000000e-03
HBsAg	0.000000	0.000	0.006699	0.000	0.000000e+00
HBeAg	0.000000	0.000	0.003112	0.000	0.000000e+00
HBcAb	0.000000	0.000	0.006277	0.000	0.000000e+00
HCVAb	0.000000	0.000	0.009799	0.000	-2.800000e-02
Cirrhosis	0.000000	0.000	0.002520	0.000	-4.000000e-03
Endemic	0.000000	0.000	0.007206	0.000	-8.000000e-03
Smoking	0.000000	0.000	0.002969	0.000	-8.000000e-03
Diabetes	0.000000	0.000	0.007142	0.000	-2.200000e-02
Hemochro	0.000000	0.000	0.005072	0.000	4.000000e-03
Age	0.000000	0.000	0.031171	0.000	-6.000000e-03
AHT	0.000000	0.000	0.006345	0.000	-3.000000e-02
CRI	0.000000	0.000	0.003934	0.000	-2.000000e-03
HIV	0.000000	0.000	0.002940	0.000	-6.000000e-03
NASH	0.000000	0.000	0.004257	0.000	0.000000e+00
Varices	0.000000	0.000	0.005569	0.000	-6.000000e-03
Spleno	0.000000	0.000	0.005896	0.000	-1.600000e-02
PHT	0.000000	0.000	0.005514	0.000	-1.200000e-02
PVT	0.000000	0.000	0.005566	0.000	-1.200000e-02
Metastasis	0.000000	0.000	0.011153	0.000	-1.800000e-02
Grams_day	0.000000	-0.002	0.009421	0.000	-1.200000e-02

Data Evaluation

Test 1 - Use only the top 10 variables with the best feature value.
Test 2 - Use all variables except the 10 with the worst feature value.
Test 3 - Use only the 20 variables with the best feature value.
Test 4 - Use only the 5 variables with the best feature value.

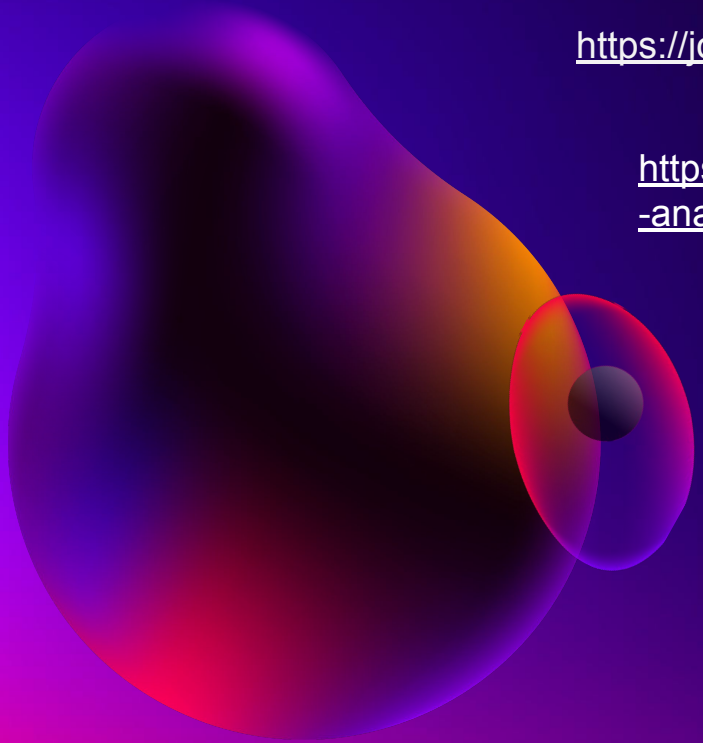
	Base case	Test 1	Test 2	Test 3	Test 4
Decision Trees	0.74	0.62	0.70	0.64	0.62
KNN	0.66	0.70	0.64	0.64	0.64
Random Forest	0.74	0.78	0.74	0.70	0.66
Support Vector Machine	0.58	0.6	0.58	0.58	0.60
Multi-layer perceptron neural network with dropout	0.62	0.68	0.48	0.62	0.68
Mean	0.670	0.676	0.630	0.636	0.640

Interpretation of results:

- Random Forest is the algorithm that provides the best accuracy values overall.
- Since the dataset with the highest mean of accuracy values was the one from test 1 this will be the one will the focus of this analysis.

- Using the 10 variables with the best feature value turned out to give off the best accuracy levels.
- On the ROC curve since the values are always in between 0.5 and 1, never being less than 0.5, it can be concluded that the model despite not being perfect ir is pretty efficient.
- In the precision-recall graph the Random Forest, Support Vector Machine and MLP Neural Network present inconsistent curves, while decision tree has the most linear one.
- In the confusion matrix, the Support Vector Machine presents a lot of False positives, which is the opposite of what was wanted for the model in study.
- In general, the variables with the most feature importance are 'AFP','Leucocytes','Age','Metastasis' and 'INR'. However using only these five variables does not improve the quality of the model.
- In the learning curve Support Vector Machine and MLP Neural Network are the ones where the curves are the closest.

Bibliographic search



https://journals.sagepub.com/doi/10.1177/1460458220984205

https://datascience.cancer.gov/training/learn-data-science/explore-analyze-data-basics

https://reintech.io/blog/create-machine-learning-algorithm-with-python-tutorial