

# Assignment No.2:

“Data exploration and enrichment for supervised classification ” -

**The Hepatocellular Carcinoma Dataset**

**EAICD** (1ºano, 2º semestre)  
**BioInformática**  
**FCUP**

Ferramentas:  
Pyzo or Pycharm  
Streamlit  
Google Collab



**Grupo 30 - BioInf:**  
Hugo Lameira - up202306642  
João Carneiro - up202306538



# Objetivos pretendidos no projeto

## Analisar e processar os Dados:

- Inicialmente analisar detalhadamente os dados - examinando que tipos de dados, as características, valores e distribuição desses mesmo dados;
- Realizar o processamento dos dados, onde vamos tratar dos valores que estejam ausentes ou errados e transformação esses dados para que não haja uma má interpretação;

## Modelagem de Dados (Aprendizagem Supervisionada):

- Definir o resultado pretendido e selecionar e preparar os algoritmos de aprendizagem a serem utilizados (Árvores de Decisão e KNN - uso do Scikit-learn);
- Avaliar o processo da Aprendizagem;

## Avaliação de Dados e Interpretação dos resultados:

- Comparar resultados de classificação usando diferentes métricas de avaliação, como matriz de confusão, ROC/AUC, precisão, recall e acurácia. Deverá ser através de tabelas e/ou gráficos (ex: bibliotecas Seaborn ou Matplotlib);
- Extrair insights significativos dos resultados obtidos, explique o comportamento dos modelos, e forneça recomendações para análises futuras;

## Elementos Extras:

- Explorar técnicas de imputação de dados ausentes, técnicas de balanceamento de classes, métodos adicionais de partição de dados, algoritmos adicionais e otimização de hiperparâmetros;
- Considerar a possibilidade de implementar uma aplicação web interativa para o projeto - aplicação Streamlit;

# Análise de Dados

	Gender	Symptoms	Alcohol	HBSAg	HBeAg	HBCAb	HCVAb	Cirrhosis	Endemic	Smoking	
0	Male	No	Yes	No	No	No	No	Yes	No	Yes	
1	Female	?	No	No	No	No	No	Yes	?	?	
2	Male	No	Yes	No	Yes	No	Yes	No	Yes	No	
3	Male	Yes	Yes	No	No	No	No	Yes	No	Yes	
4	Male	Yes	Yes	Yes	No	Yes	No	Yes	No	Yes	

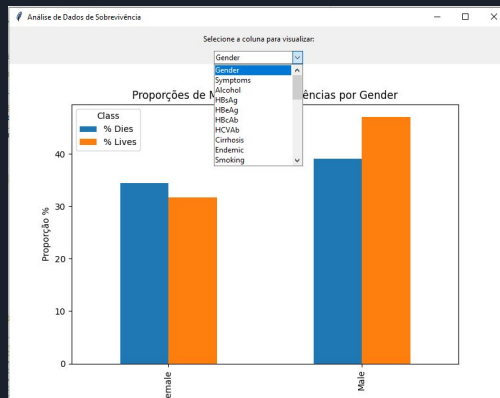
	...	ALP	TP	Creatinine	Nodules	Major_Dim	Dir_Bil	Iron	Sat	Ferritin	Class
0	...	158	7.1	0.7	1	3.5	0.5	?	?	?	Lives
1	...	?	?	?	1	1.8	?	?	?	?	Lives
2	...	109	7	2.1	5	13	0.1	28	6	16	Lives
3	...	174	8.1	1.11	2	15.7	0.2	?	?	?	Dies
4	...	109	6.9	1.8	1	9	?	59	15	22	Lives

	Gender	Symptoms	Alcohol	HBSAg	HBeAg	HBCAb	HCVAb	Cirrhosis	Endemic	Smoking	
0	Male	No	Yes	No	No	No	No	Yes	No	Yes	
1	Female	Yes	No	No	No	No	Yes	Yes	No	Yes	
2	Male	No	Yes	Yes	No	Yes	No	No	Yes	No	
3	Male	Yes	Yes	No	No	No	No	Yes	No	Yes	
4	Male	Yes	Yes	Yes	No	Yes	No	Yes	No	Yes	

	...	ALP	TP	Creatinine	Nodules	Major_Dim	Dir_Bil	Iron	Sat	Ferritin	Class
0	...	158	7.1	0.7	1	3.5	0.5	94	25	48	Lives
1	...	109	7.3	0.7	1	1.8	0.3	94	25	48	Lives
2	...	109	7	2.1	5	13	0.1	28	6	16	Lives
3	...	174	8.1	1.11	2	15.7	0.2	94	25	48	Dies
4	...	109	6.9	1.8	1	9	0.3	59	15	22	Lives

- Como vamos interpretar os dados? Será que estão corretos e completos?
- O que vamos fazer com os dados problemáticos? Para preencher os dados que faltavam. Troca de os termos desconhecidos “?” pela média para colunas numéricas, e moda para as colunas categóricas.
- Quais são os dados mais importantes? Certos dados têm mais relevância, pois tornam-se mais importantes no contexto do diagnóstico.



```

Gender: ['Male' 'Female']
Symptoms: ['No' '?' 'Yes']
Alcohol: ['Yes' 'No']
HBSAg: ['No' 'Yes' '?']
HBeAg: ['No' '?' 'Yes']
HBCAb: ['No' 'Yes' '?']
HCVAb: ['No' 'Yes' '?']
Cirrhosis: ['Yes' 'No']
Endemic: ['No' '?' 'Yes']
Smoking: ['Yes' '?' 'No']
Diabetes: ['Yes' 'No' '?']
Obesity: ['?' 'No' 'Yes']
Hemochro: ['Yes' 'No' '?']
AHT: ['No' 'Yes' '?']
CRF: ['No' 'Yes' '?']
HIV: ['No' '?' 'Yes']
NAASH: ['No' '?' 'Yes']
Varices: ['Yes' 'No' '?']
Splenomeg: ['No' 'Yes' '?']
PHT: ['No' 'Yes' '?']
PVT: ['No' 'Yes' '?']
Metastasis: ['No' 'Yes' '?']
Hemoglobin: ['Yes' 'No' '?']
Age: [57 62 78 77 76 75 49 61 58 43 41 74 66 56 63 72 60 64 71 73 84 80 45 57
20 70 59 86 52 58 27 51 81 65 82 68 40 88 23 83 69 79 87 93 85 55 46 25
36 47 44]
Grams_day: ['137' '0' '50' '48' '100' '?' '200' '500' '80' '100' '150' '60' '20'
'120' '75' '70' '300' '90' '96' '250']
Packs_year: ['15' '?' '50' '30' '0' '20' '32' '60' '78' '47' '16' '67.5' '2' '8'
'40' '10' '44' '48' '34.5' '33' '1' '7.5' '43' '23' '52.5' '510' '12'
'37' '18' '25']
PS: ['Active' 'Ambulatory' 'Restricted' 'Selfcare' 'Disabled']
Encephalopathy: ['Grade I/II' '?' 'Grade III/IV']
Ascites: ['Mild' 'Moderate/Severe' '?']
INR: ['1.53' '?' '0.96' '0.95' '0.94' '1.58' '1.4' '1.46' '3.14' '1.12' '1.05'
'1.33' '1.2' '1.25' '1.61' '2.14' '1.13' '1.44' '1.29' '1.06' '1.27'
'4.82' '1.74' '1.38' '1.37' '1.3' '1.23' '1.19' '1.32' '1.24' '1.28' '2'
'1.09' '1.42' '1.18' '1.17' '1.71' '1.64' '1.49' '0.97' '1.65' '1.35'
'3.56' '1.96' '1.11' '1.94' '1.57' '1.34' '2.07' '1.01' '1.15' '1.92'
'3.16' '1.45' '2.08' '1.63' '1.48' '1.1' '1.47' '1.88' '1.23' '1.41'
'1.22' '1.39' '1.6' '1.67' '1.16' '1.04' '1.55' '1.93' '1.54' '1.26'
'1.87' '1.36' '2.42' '1.14' '1.56' '1.66' '1.07' '2.5' '1.03' '1.02'
'1.798' '1' '1.68' '1.52' '0.84' '1.79']
AFP: ['95' '?' '5.8' '2440' '49' '110' '138.9' '8660' '8.8' '1.8' '100889' '86'
'68' '6.8' '29' '4.6' '9.2' '34' '19.6' '3.9' '1975' '185' '5532' '13327'
'5.9' '3255' '1.9' '11' '1237' '7.7' '266' '5689' '14.2' '3.1' '633'
'5.4' '479' '19' '2.8' '185203' '5' '237' '16' '153' '20' '46' '41470'
'4.7' '7.3' '1898' '77' '2.7' '2.6' '12' '608' '41' '2' '7' '13' '1.7'
'249' '66' '358' '1818346' '33582' '2.5' '2269' '4181' '5.1' '345' '2.9'
'5.04' '3.7' '2.75' '42' '457' '123' '8.7' '225' '2159' '48' '2.4' '64'
'5.5' '8.5' '2785' '6574' '5.7' '39' '32' '7.6' '173' '2.3' '114' '18'
'28274' '736' '1009' '22475' '5.2' '14177' '50655' '1.2' '657' '421500'
'472' '2.1' '4.2' '7.5' '152' '811' '2689' '4.9' '248' '180' '15' '24'
'9.4' '4.8' '92421' '1.5' '9204' '10' '695' '33' '615' '1671' '975'
'1713' '4887' '75' '94964' '44340']
Hemoglobin: ['13.7' '?' '8.9' '13.4' '14.3' '10.4' '10.8' '11.9' '11.8' '13' '15.7'
'13.3' '13.5' '10.2' '12.1' '10.3' '14.9' '15.9' '11.7' '16.4' '10.7'
'13.1' '13.6' '15.5' '12.2' '9.9' '14.8' '11.3' '13.9' '15' '15.1' '15.6'
'16.6' '14' '10.6' '10.5' '12.6' '15.3' '12.4' '7.3' '10.9' '18.7' '14.6'
'11.5' '9.5' '12.7' '14.4' '9.1' '9.8' '15.8' '10.1' '14.1' '12' '5']

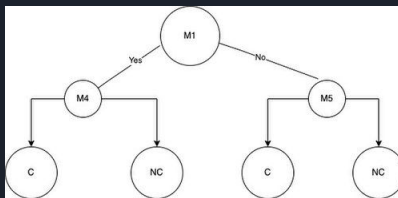
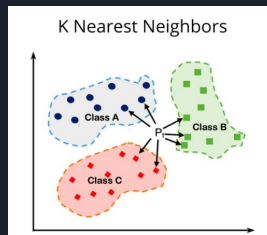
```

# Algoritmos: Decision Tree e KNN

## K-Nearest Neighbors (KNN):

O KNN é um classificador de aprendizagem supervisionado, que usa proximidade para fazer classificações ou previsões sobre o agrupamento de um ponto de dados individual. É um dos classificadores de classificação e regressão mais populares e simples usados no aprendizado de máquina atualmente.

**Vantagens:** Simples, intuitivo, e com boa performance para dados pequenos.



## Decision Tree:

Um algoritmo de árvore de decisão é um algoritmo de aprendizado de máquina que usa uma árvore de decisão para fazer previsões. Segue um modelo de árvore de decisões e suas possíveis consequências. O algoritmo funciona dividindo recursivamente os dados em subconjuntos com base no recurso mais significativo em cada nó da árvore.

**Vantagens:** Fácil de interpretar e visualizar - pode lidar com dados categóricos e numéricos.

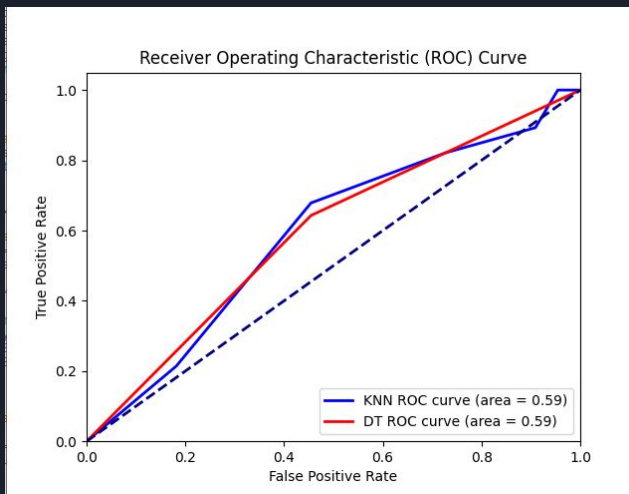
**Desvantagens:** Propenso a overfitting, especialmente com árvores profundas.

# Modelagem dos dados

Após o processamento dos dados, aplicamos os dois tipos de modelos para que possamos interpretá-los e obter resultados.

Sabemos que os modelos de Decision Tree e KNN foram treinados e avaliados com base em métricas de classificação.

Obtemos o primeiro resultado:



Observações:

- Decision Tree - AUC-ROC de 0.5942;
- KNN - AUC-ROC de 0.5901;
- As curvas ROC mostram que ambos os modelos têm um desempenho razoável;
- Leitura e análise conforme esperado;
- Baixa precisão;
- O que podemos fazer para aumentar?



# O que podemos fazer para aumentar a precisão?

Decidimos pesquisar diferentes formas para melhorar o desempenho dos modelos. Vamos explorar duas principais: a otimização de hiperparâmetros e técnicas avançadas de pré-processamento de dados.

## Database

Se conseguirmos adicionar mais informação à nossa base de dados, neste caso, mais diagnósticos, o algoritmo irá usufruir de mais informação que ajuda a melhorar o seu desempenho.

## Otimização de Hiperparâmetros

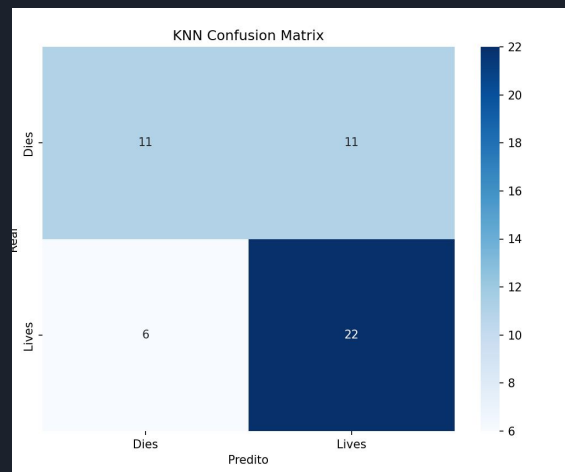
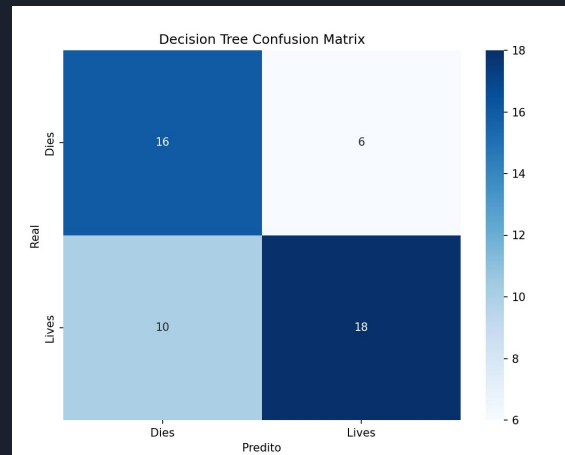
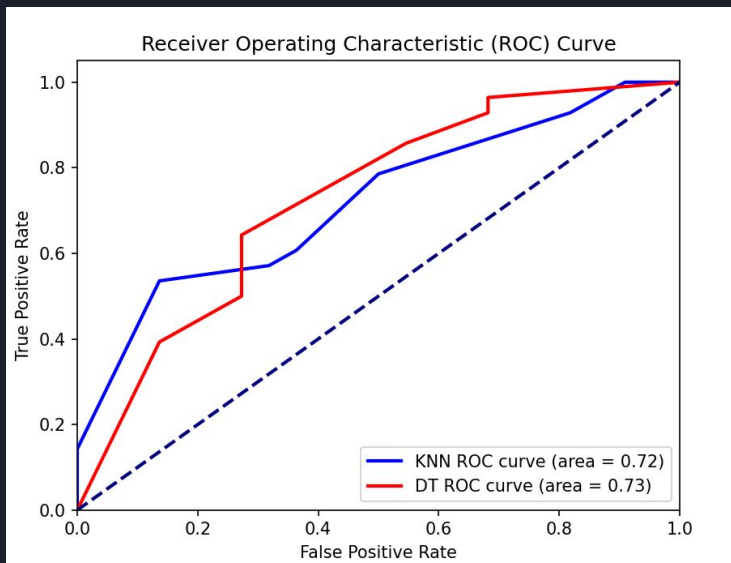
Envolve ajustar os parâmetros do modelo para melhorar seu desempenho. Encontramos que isso podia ser feito usando técnicas como “GridSearchcv” ou “RandomizedSearchcv” da biblioteca scikit-learn.

1. Grid Search: Testa todas as combinações de hiperparâmetros especificados - demorado para muita pesquisa de informação.
2. Randomized Search: Estuda aleatoriamente um conjunto dos hiperparâmetros, sendo mais eficiente caso haja muita informação, embora não garanta a melhor solução.

## Técnicas Avançadas de Pré-Processamento de Dados

1. Tratamento de Valores Ausentes: Além da imputação de média e moda, você pode explorar a imputação de valores ausentes usando modelos mais sofisticados, como KNNImputer ou IterativeImputer.
2. Feature Engineering: Criar novas features baseadas nas existentes pode melhorar o desempenho. Isso pode incluir interações entre variáveis, transformações não lineares, etc.
3. Seleção de Features: Use técnicas como `Recursive Feature Elimination (RFE)`, `SelectFromModel`, ou `PCA` para reduzir a dimensionalidade e remover features irrelevantes.

# Accuracy, Precision, ROC-AUC, Confusion Matrix



# Interpretação dos Resultados



## Resumo da Análise:

- ❖ Foi realizada uma análise comparativa de dois modelos de classificação (Decision Tree e KNN) utilizando o conjunto de dados da Hepatocellular Carcinoma. Os dados foram pré-processados para tratar valores ausentes, quer dos dados numéricos, quer das variáveis categóricas.
- ❖ Cada modelo teve o seu desempenho com a sua Accuracy, Precision, Recall, Confusion Matrix e AUC-ROC.
- ❖ AUC-ROC Scores: Ambos os modelos apresentam AUC-ROC scores próximos de 0.7 - indica que tem um bom desempenho e capacidade para distinguir as classes.
- ❖ Os Modelos precisam de ser mais otimizados ou talvez substituídos por técnicas mais avançadas como Random Forest, Gradient Boosting, ou redes neurais para melhorar a acurácia e a sua capacidade de distinção entre cada caso, isso se for necessário para aplicar na realidade.
- ❖ Análise de características importantes pode ajudar a entender melhor os fatores que influenciam as previsões e potencialmente levar a melhorias no modelo.



# Website - StreamLit

## Objetivos:

- O aplicativo oferece uma interface amigável para os usuários inserirem os dados de diagnóstico dos pacientes e selecionarem o modelo para fazer as previsões.
- O site pode ser usado para apresentar informações de uma forma clara e acessível, tornando os dados mais compreensíveis para o público-alvo.
- Os dados de entrada do usuário são pré-processados usando o pré-processador salvo antes de fazer previsões.
- Com base nos dados inseridos pelo usuário e no modelo selecionado, o aplicativo faz previsões e exibe os resultados.
- Os diagnósticos previstos podem ser registrados e visualizados em uma tabela.

## Conclusões:

- Ideal para contexto clínico - automatizar o trabalho - fácil para navegar, entender e interagir com as funcionalidades oferecidas. Cumpre a sua tarefa principal
- Necessita de uma grande base de dados;

**Parâmetros do Paciente**

Nome da Pessoa  
Maria

Idade  
13

Gênero  
Female

Sintomas  
No

Alcoólico  
No

HbSag  
No

HbHg  
No

HbC4D  
No

HCVAb  
No

**Aplicativo de Previsão de Diagnóstico de Hepatocarcinoma**

**Resultados do Diagnóstico**

Nome: Maria  
Diagnóstico Previsto: Lives

[Registrar Diagnóstico](#)

**Registros de Diagnóstico**

	Nome	Diagnosis	Model
0	Jonas	Dies	Decision Tree
1	Maria	Lives	KNN

[Resetar Tabela de Diagnóstico](#)

# GITHUB / Referências - Pesquisa Bibliográfica

## O NOSSO GITHUB:

<https://github.com/UP202306538/Assignment-2-Final-Delivery-May-21->

## Websites explicativos:

<https://www.geeksforgeeks.org/>

<https://scikit-learn.org/stable/>

<https://www.analyticsvidhya.com/>

<https://github.com/streamlit>

<https://stackoverflow.com/>

## Vídeo-aulas sobre o conteúdo:

<https://www.youtube.com/>

## Ferramentas de AI:

ChatGPT - OpenAI

