ASSIGNMENT NO. 2

DATA EXPLORATION AND ENRICHMENT FOR SUPERVISED CLASSIFICATION

Alunas

Ana Matilde Santos
Catarina Aguiar
Maria Leonor Carvalho

Cadeira

Inteligência Artificial e Ciência de Dados

Professores

Miriam Santos
Pedro Ferreira
Luís Paulo Reis

INTRODUÇÃO AO PROJETO

BREVE DESCRIÇÃO

*"Desenvolvimento de um pipeline completo de ciência de dados para **análise do conjunto de dados do Carcinoma Hepatocelular (HCC)**."*

ESPECIFICAÇÃO DO PROBLEMA DE MACHINE LEARNING A SER ABORDADO:

***Capacidade de sobrevivência** de pacientes diagnosticados com HCC após 1 ano do diagnóstico (por exemplo, "vive" ou "morre")."*

OBJETIVOS DO PROJETO:

- *Abordar um **caso real de uso da ciência de dados**, explorando dados clínicos reais de pacientes diagnosticados com HCC;*
- ***Desenvolvimento de um modelo de machine learning** para prever a taxa de sobrevivência dos pacientes diagnosticados, à um ano, com HCC.*

DATASET:

Conjunto de dados HCC, recolhidos no Centro Hospitalar e Universitário de Coimbra (CHUC) em Portugal.

ESPECIFICAÇÃO DO TRABALHO

Com este trabalho, pretendemos desenvolver um modelo de *Machine Learning* que preveja, com exatidão, a taxa de sobrevivência (vive ou morre) dos pacientes diagnosticados, à um ano, com HCC, tendo em conta os dados presentes no dataset disponibilizado.

DATASET

Conjunto de dados HCC, recolhidos no Centro Hospitalar e Universitário de Coimbra (CHUC) em Portugal.

	Gender	Symptoms	Alcohol	HBsAg	HBeAg	HBcAb	HCVAb	Cirrhosis	Endemic	Smoking	...	ALP	TP	Creatinine	Nodules	Major_Dim	Dir_Bil	Iron	Sat	Ferritin	Class
0	Male	No	Yes	No	No	No	No	Yes	No	Yes	...	150	7.1	0.7	1	3.5	0.5	?	?	?	Lives
1	Female	?	No	No	No	No	Yes	Yes	?	?	...	?	?	?	1	1.8	?	?	?	?	Lives
2	Male	No	Yes	Yes	No	Yes	No	Yes	No	Yes	...	109	7	2.1	5	13	0.1	28	6	16	Lives
3	Male	Yes	Yes	No	No	No	No	Yes	No	Yes	...	174	8.1	1.11	2	15.7	0.2	?	?	?	Dies
4	Male	Yes	Yes	Yes	No	Yes	No	Yes	No	Yes	...	109	6.9	1.8	1	9	?	59	15	22	Lives
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
160	Female	No	Yes	?	?	?	Yes	Yes	No	Yes	...	109	7.6	0.7	5	3	?	?	?	?	Lives
161	Female	Yes	No	?	?	?	?	Yes	No	No	...	280	6.7	0.7	1	2.2	2.3	?	?	?	Dies
162	Male	No	Yes	No	No	No	No	Yes	No	Yes	...	181	7.5	1.46	5	18.6	?	?	?	?	Lives
163	Male	No	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	...	170	8.4	0.74	5	18	?	?	?	?	Dies
164	Male	Yes	Yes	No	No	No	Yes	Yes	No	Yes	...	462	6.6	3.95	5	8.5	19.8	?	?	?	Dies

165 rows × 50 columns

Fig.1 Dataset

DATA EXPLORATION

CARACTERÍSTICAS DOS DADOS

- 165 Registos (linhas)
- 50 Atributos (colunas)
- Distribuição das ‘Classes’

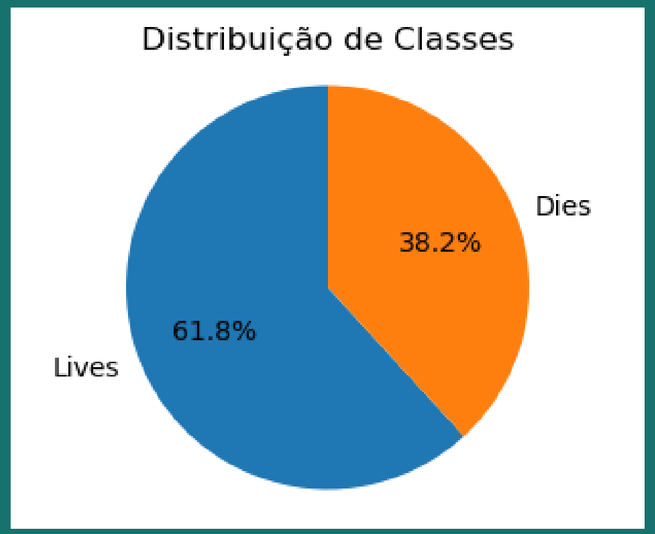


Fig.2 Gráfico com a distribuição de classes

- Dados categóricos distribuídos num ‘Violin plot’;

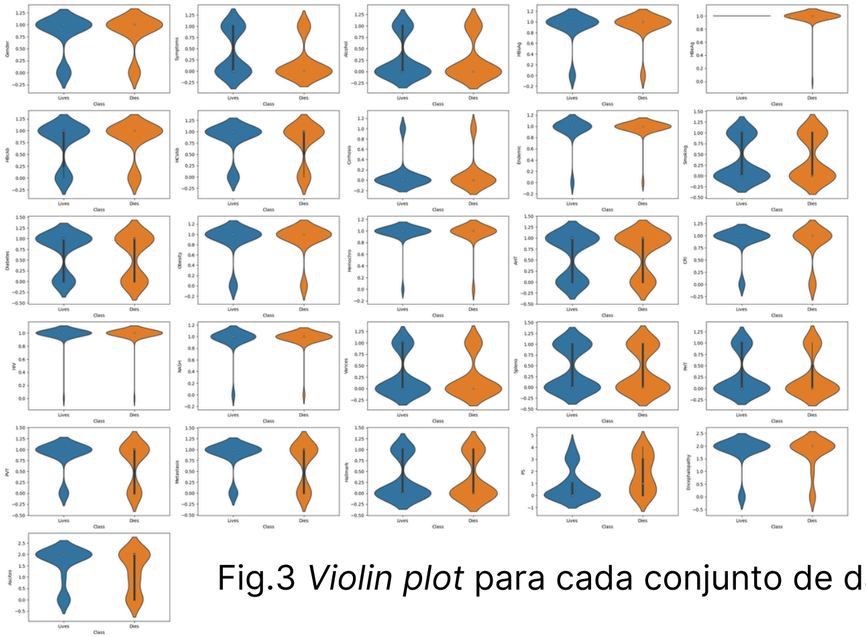


Fig.3 *Violin plot* para cada conjunto de dados categóricos

- Dados numéricos distribuídos num ‘Heatmap’.

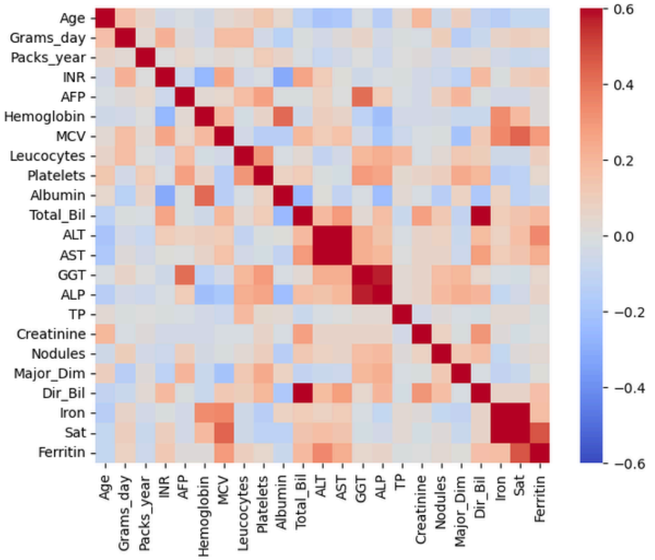


Fig.4 *Heatmap* para cada conjunto de dados numéricos

DATA PREPROCESSING

ALTERAÇÕES INICIAIS DOS DADOS

- Remover espaços entre os nomes das colunas;
- Alterar as respostas ‘NaN’ para 'Not affected';
- No ficheiro os dados em falta são identificados por ‘?’:
  - Alterar os dados categóricos em falta pela moda;
  - Alterar os dados numéricos em falta pela média dos que estão presentes.

Consideramos que eliminar os dados em falta nos iria retirar a maioria dos dados, concluindo que não o poderíamos fazer.

ORGANIZAÇÃO DOS DADOS

- Alterar os dados ‘Yes’ ou ’No’ para 0 e 1, respetivamente;
- Alterar os restantes dados categóricos utilizando o ‘LabelEnconder’.

DATA MODELING

ALGORÍTMOS DE MACHINE LEARNING APLICADOS

01 Decision Tree

Modelo de *Machine Learning* que utiliza uma estrutura em forma de árvore para tomar decisões, utilizando condições baseadas em variáveis numéricas e categóricas. Cada nó interno representa uma condição num atributo, cada ramo representa o resultado da condição e cada nó 'folha' representa uma classe ou valor de saída.

02 KNN

Algoritmo K-Nearest Neighbors é um método de *Machine Learning* que classifica e faz previsões com base na proximidade de um determinado dado aos seus dados vizinhos mais próximos no espaço de características.

03 Random Forest

Algorítmo de *Machine Learning* que combina múltiplas árvores de decisão para melhorar a precisão e reduzir o risco de overfitting. Cada árvore é treinada numa amostra aleatória de dados e faz uma previsão. As previsões de todas as árvores são agregadas de forma a produzir a previsão final.

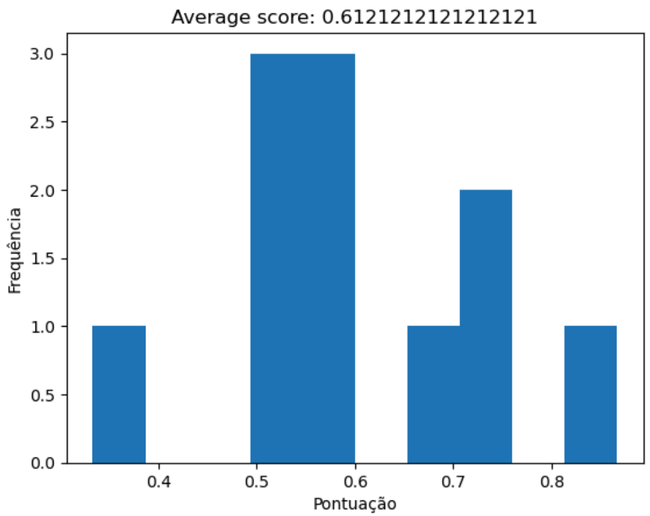
De seguida são apresentados os resultados obtidos com cada um destes modelos

PRIMEIRA UTILIZAÇÃO DOS MODELOS

GRÁFICOS UTILIZANDO O MÉTODO DE PARTIÇÃO CROSS-VALIDATION

Aplicamos os 3 modelos, abaixo apresentados, com todos os dados incluídos, apenas com o tratamento de dados anteriormente referido.

Decision Tree

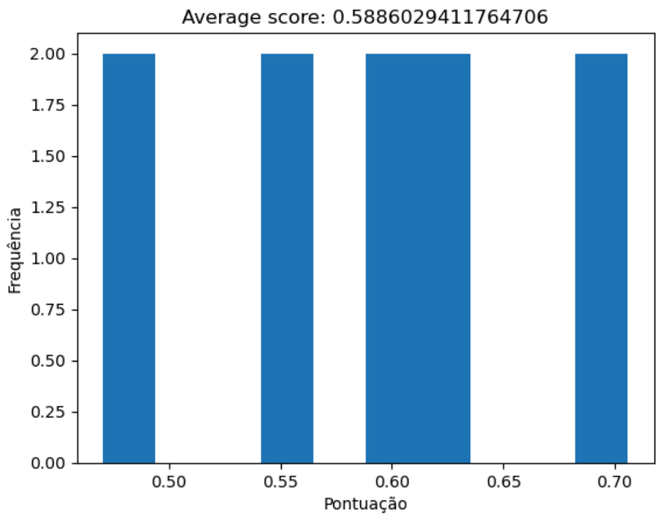


Decision Tree Accuracy: 0.6190

[[10 7]

[9 16]]

K Nearest Neighbours

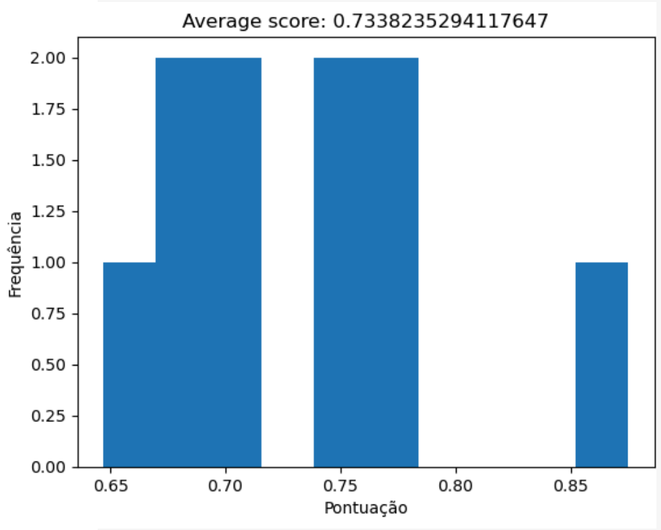


KNN Accuracy: 0.5714

[[6 11]

[7 18]]

Random Forest



Random Forest Accuracy: 0.7381

[[11 6]

[5 20]]

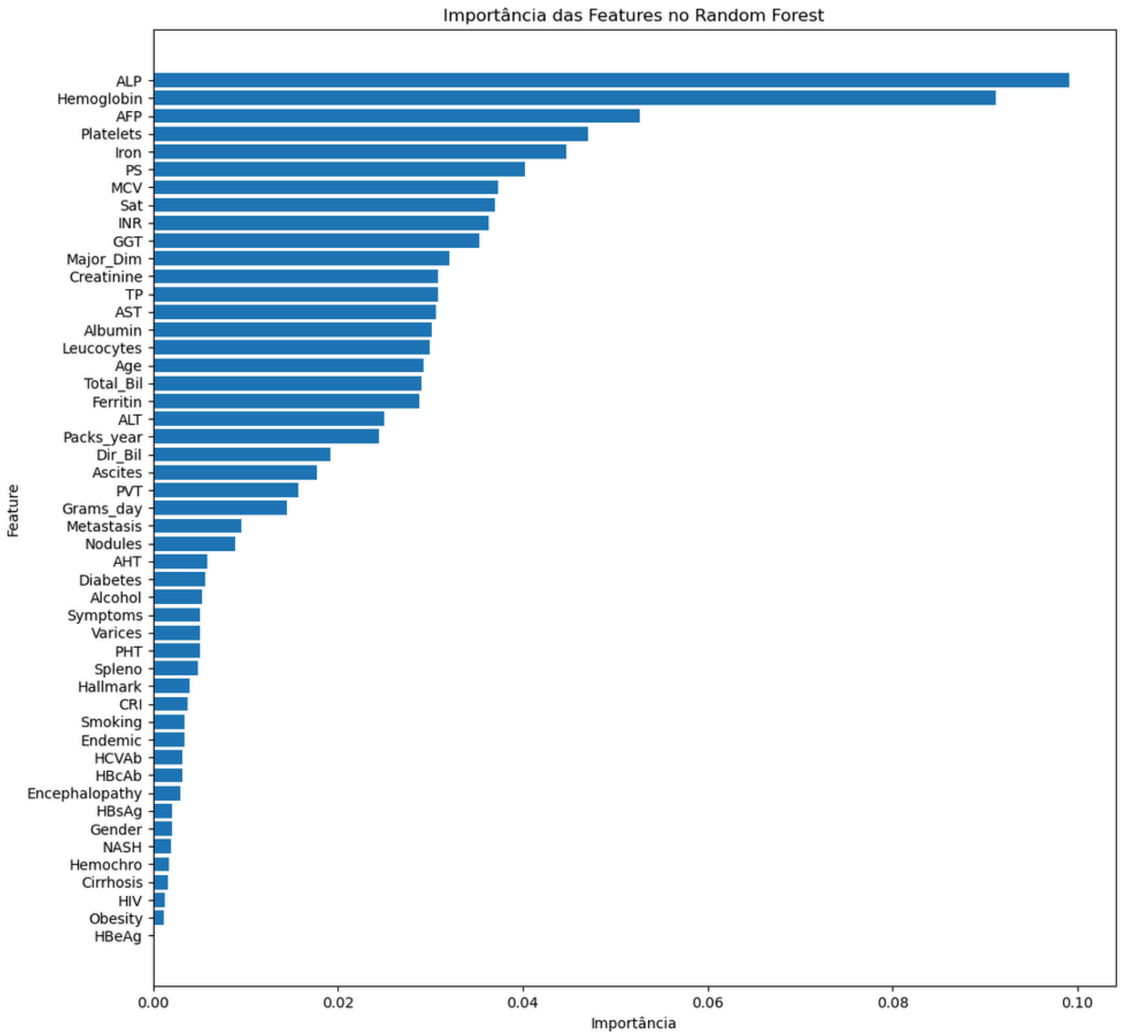
RESULTADOS UTILIZANDO O MÉTODO DE PARTIÇÃO LEAVE-ONE-OUT:

Decision Tree Accuracy: 0.6848

KNN Accuracy: 0.5758

Random Forest Accuracy: 0.6909

SIMPLIFICAÇÃO DOS DADOS



Importância de cada feature

Da comparação dos resultados dos três algoritmos, concluímos que o modelo *Random Forest* é o melhor entre os três, para este conjunto de dados. Deste modo, plotamos o gráfico ao lado que nos dá a importância de cada feature.

Aplicação dos métodos em partição

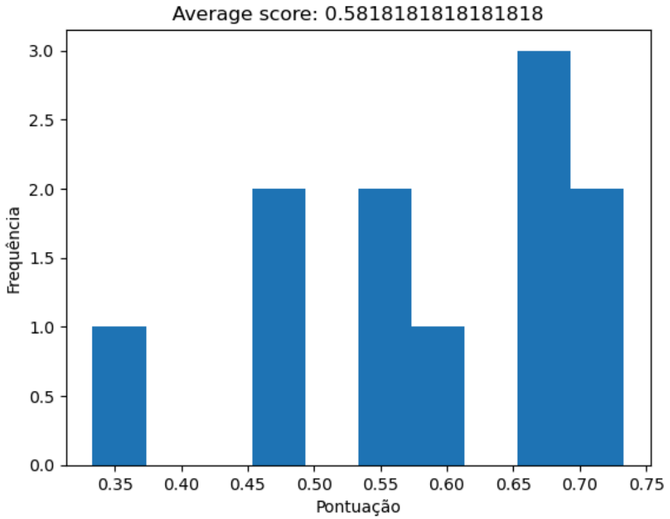
Para otimizar os resultados, realizamos 6 tentativas, eliminando a cada nova tentativa mais features, começando pelas de menor importância e indo aumentando, até ficarem apenas com as de maior importância. Em cada tentativa, avaliamos os resultados de cada um dos três modelos aplicados e escolhemos a combinação de features que apresentava o melhor desempenho. Esse processo garantiu que o modelo final não apenas fosse o mais preciso, mas também eficiente, utilizando apenas as features mais relevantes para a precisão.

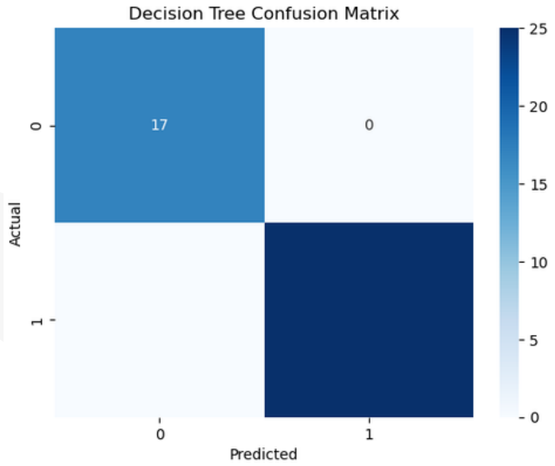
DATA EVALUATION

RESULTADOS FINAIS

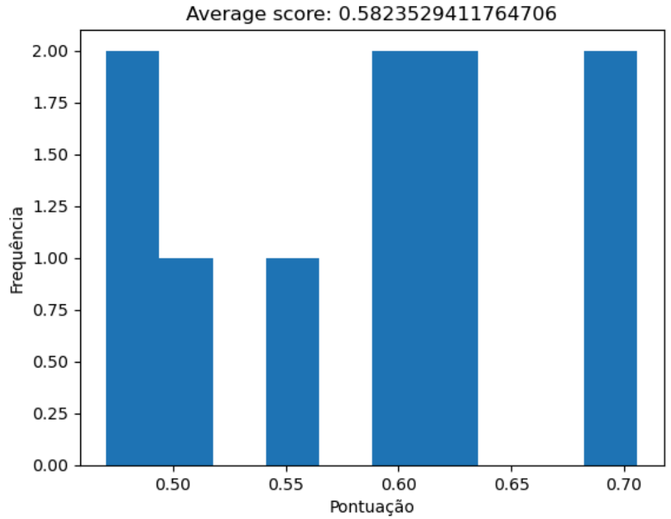
O modelo *Random Forest* é o que apresenta melhores valores de precisão, recall e F1-score e o menor número de falsos positivos e falsos negativos.

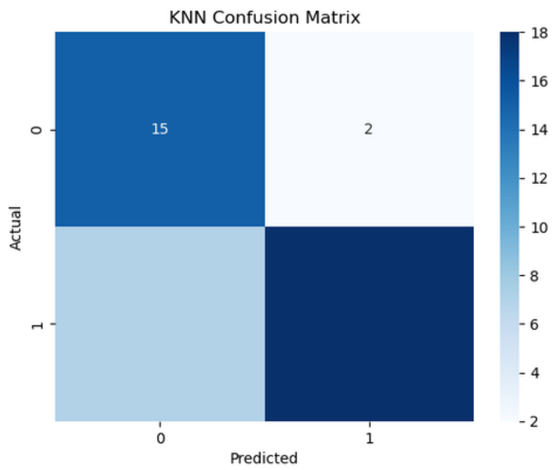
Decision Tree



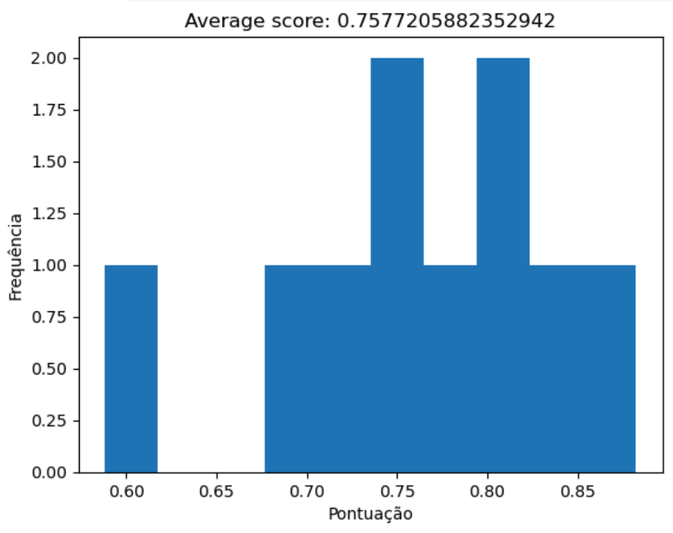


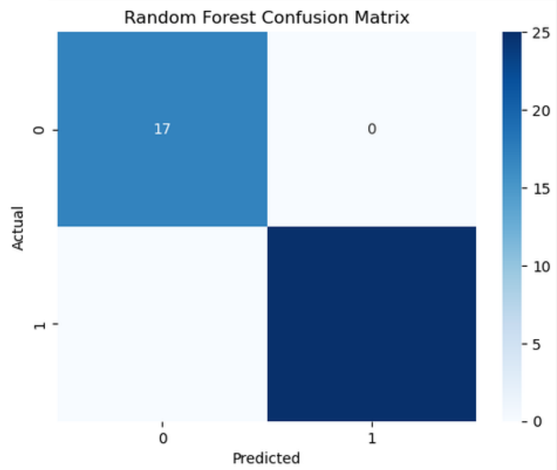
K Nearest Neighbours





Random Forest





ALGORÍTMOS E FERRAMENTAS UTILIZADAS

Linguagem de programação

PYTHON

Bibliotecas utilizadas

PANDAS
NUMPY/SCIPY
SCIKIT-LEARN
MATPLOTLIB
SEABORN
IMBLEARN-LEARN

Algoritmos utilizados

DECISION TREES
KNN (SCIKIT-LEARN)
RANDOM FOREST

TRABALHOS RELACIONADOS

DURANTE O DESENVOLVIMENTO DO PROJETO, EXECUTAMOS CONSULTA BIBLIOGRÁFICA PARA FUNDAMENTAR A CONCRETIZAÇÃO DO TRABALHO.

Artígos científicos

Retirados do livro *Hepatocellular Carcinoma (https://www.ncbi.nlm.nih.gov/books/NBK553759/?term=survival%20of%20patients%20with%20HCC), foram* consultados para nos darem informações relacionadas sobre o diagnóstico de carcinoma hepatocelular (HCC).

https://www.ncbi.nlm.nih.gov/books/NBK553754/
https://www.ncbi.nlm.nih.gov/books/NBK553748/

Consulta para utilização de bases de dados

https://www.datacamp.com/cheat-sheet/pandas-cheat-sheet-for-data-science-in-python
https://www.freecodecamp.org/portuguese/news/como-criar-e-manipular-bancos-de-dados-sql-com-python/
https://matplotlib.org/
https://www.datacamp.com/cheat-sheet/matplotlib-cheat-sheet-plotting-in-python