Assignment No.2:

"Data exploration and enrichment for supervised classification " -

The Hepatocellular Carcinoma Dataset

Ferramentas:
Pyzo or Pycharm
Streamlit
Google Collab

EAICD (1°ano, 2° semestre)

BioInformática

FCUP

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;

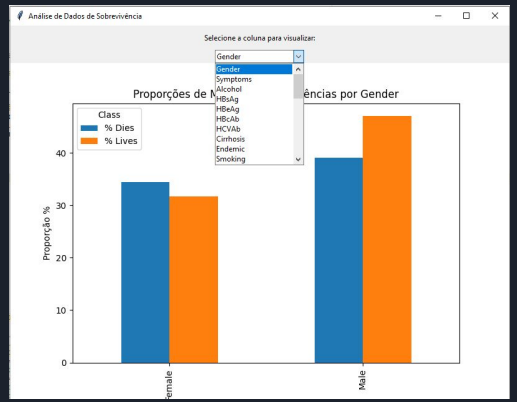
Análise de Dados

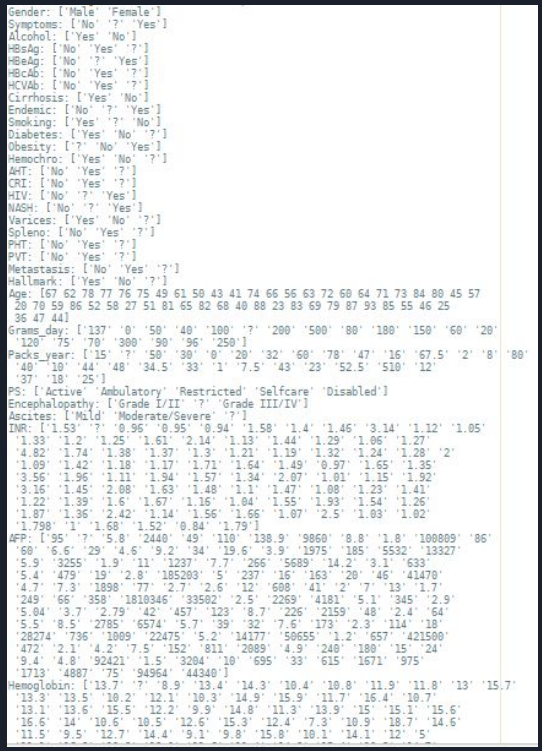
	Gender	Symptoms	Alcohol	HBsAg	HBeAg	HBcAb	HCVAb	Cirrhosis	Endemic	Smoking	\
0	Male	No	Yes	No	No	No	No	Yes	No	Yes	
1	Female	?	No	No	No	No	Yes	Yes	?	?	
2	Male	No	Yes	Yes	No	Yes	No	Yes	No	Yes	
3	Male	Yes	Yes	No	No	No	No	Yes	No	Yes	
4	Male	Yes	Yes	Yes	No	Yes	No	Yes	No	Yes	

	...	ALP	TP	Creatinine	Nodules	Major_Dim	Dir_Bil	Iron	Sat	Ferritin	Class
0	...	150	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	Yes	No	Yes	
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	...	150	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





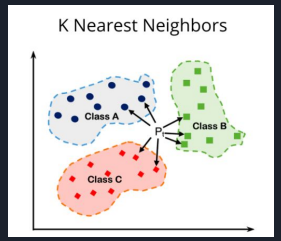
- Como vamos interpretar os dados?
  Será que estão corretos e completos?
- O que vamos fazer com os dados problemáticos?
  Para preencher os dados que faltavam.
  Troca de os termos desconhecidos “?” pela média para colunas numéricas, e moda para as colunas categóricas.
- Quais são os dados mais importantes?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

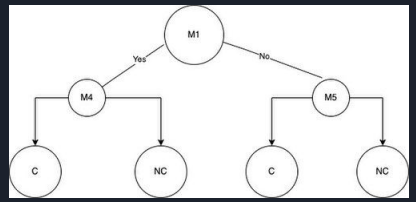
Algoritmos: Decision Tree e KNN

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.

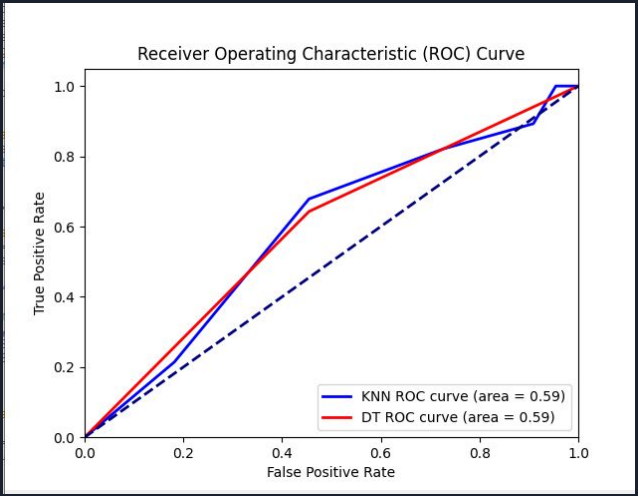
Desvantagens: Propenso a overfitting, especialmente com árvores profundas.

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

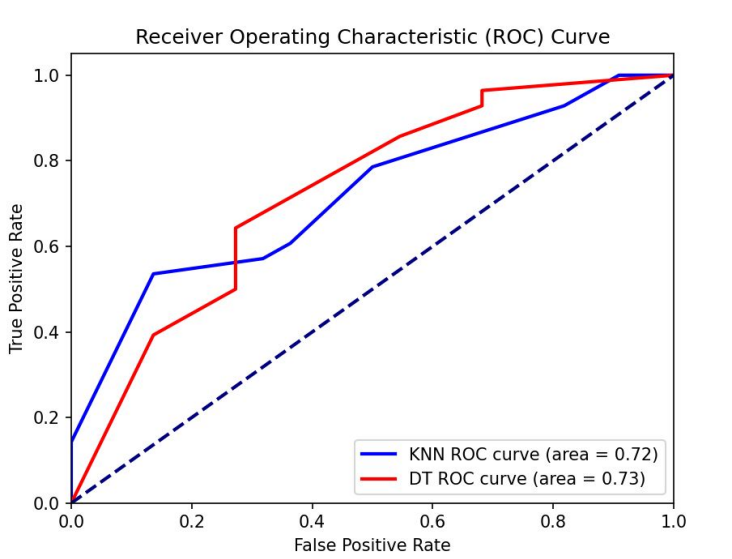
Envolve ajustar os parâmetros do modelo para melhorar seu desempenho. Encontramos que isso podia ser feito usando técnicas como “GridSearchcv “ ou “RandomizedSearchcv” da biblioteca scikit-learn.

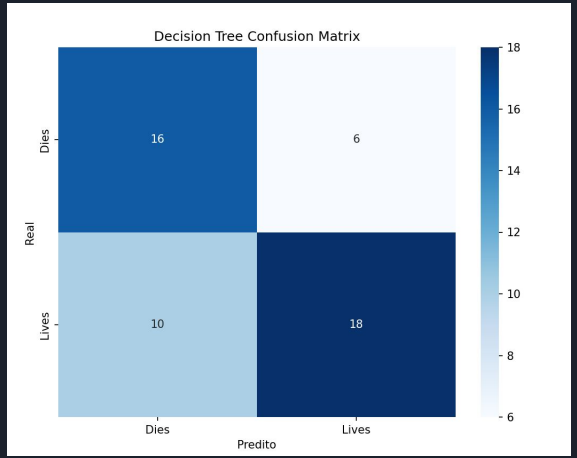
1. Grid Search: Testa todas as combinações de hiperparâmetros especificados - demorado para muita pesquisa de informação.
2. Randomized Search: Estuda aleatoriamente um conjunto dos hiperparâmetros, sendo mais eficiente caso haja muita informação, embora não garante a melhor solução.

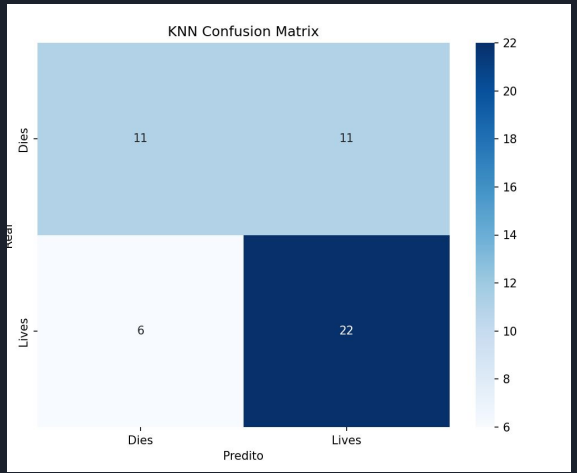
Técnicas Avançadas de Pré-Processamento de Dados

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 InterativeImputer.
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





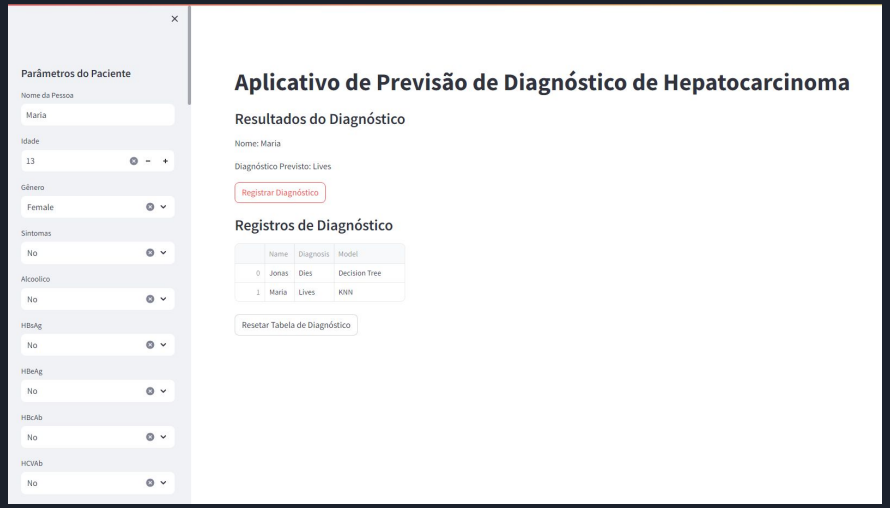


Interpretação dos Resultados

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;

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