

# Data exploration and enrichment for supervised classification

Elementos de Inteligência Artificial e Ciência de Dados

Beatriz Pereira, 202304769

Carolina Leite, 202307856

Inês Santos, 202305589

# Especificação do trabalho a realizar

## Formulação do problema

A análise exploratória de dados e a aplicação de modelos de aprendizagem supervisionados desempenham um papel crucial na ciência de dados, permitindo a extração de insights valiosos.



### Objetivo

Desenvolver um pipeline capaz de determinar a capacidade de sobrevivência dos doentes 1 ano após o diagnóstico (“vive” ou “morre”).



### DataSet

Base de dados com um total de 165 inputs, constituída por 50 colunas com diferentes parâmetros a serem estudados, e 165 pacientes, classificados conforme o género.



### Etapas

Análise de dados, pré-processamento de dados, análise exploratória, classificação, comparação e análise de resultados.

# Descrição das ferramentas e dos algoritmos a utilizar no trabalho

## PYTHON

Linguagem de programação amplamente utilizada na ciência de dados devido à sua fácil utilização e à grande quantidade de bibliotecas especializadas.

### BIBLIOTECAS PYTHON

#### PANDAS

Utilizada para manipulação e análise de dados de forma eficiente.

#### NUMPY

Utilizada para programação numérica sobre arrays.

#### SCIKIT-LEARN

Fornece implementações eficientes de algoritmos de aprendizagem supervisionados e não supervisionados.

#### MATPLOTLIB E SEABORN

Utilizada para visualização de dados.

#### JUPYTER NOTEBOOK

Ambiente de desenvolvimento interativo para criar e apresentar projetos de ciência de dados

### ALGORITMOS

#### KNN (K-Nearest Neighbors)

Classifica novos dados com base na proximidade com os dados já rotulados

#### DECISION TREE

Modelo de decisão hierárquico, semelhante a uma árvore

# Classificação

## DEFINIR O TARGET

```
target_c = df_filled['Class']
```

## CONFIGURAÇÃO

Todas as variáveis, incluindo 'Class'

## DIVISÃO

Preparação para treinar os modelos

## IMPORTÂNCIA DAS VARIÁVEIS

'Class' com importância de 1.0

## DECISION TREE

1.00

## KNN

0.56



# Classificação

## Testes realizados

2

Utilizamos todos os dados disponíveis com exceção das 6 variáveis menos correlacionadas: AFP, ALT, HIV, Hemochro, Spleno, Obesity, Class. Houve uma alteração muito significativa nos resultados.

4

Removemos outras 6 variáveis menos correlacionadas: Alcohol, HBsAg, Packs\_year, NASH, AHT, HBcAb. A alteração neste teste já foi menor, não apresentou resultados muito diferentes.

6

Removemos outras 6 variáveis menos correlacionadas: Nodules, Endemic, HBeAg, Diabetes, Major\_Dim, Grams\_day. Não houve alteração dos dados de um modo relevante.

3

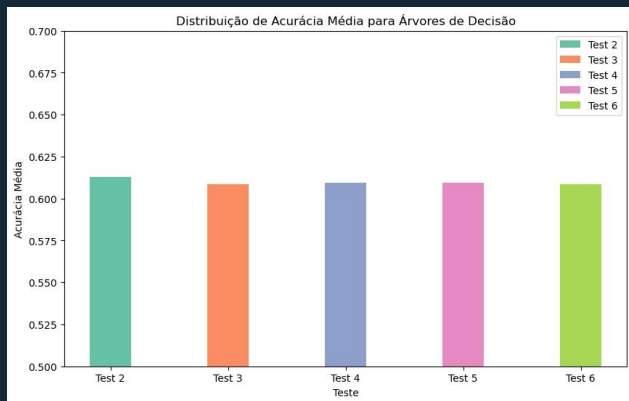
Removemos outras 6 variáveis menos correlacionadas: TP, Hallmark, Varices, MCV, Cirrhosis, Gender. Verificou-se uma diferença significativa nos resultados.

5

Removemos outras 6 variáveis menos correlacionadas: Smoking, PHT, Sat, Leucocytes, CRI, Creatinine. Os resultados obtidos foram semelhantes.

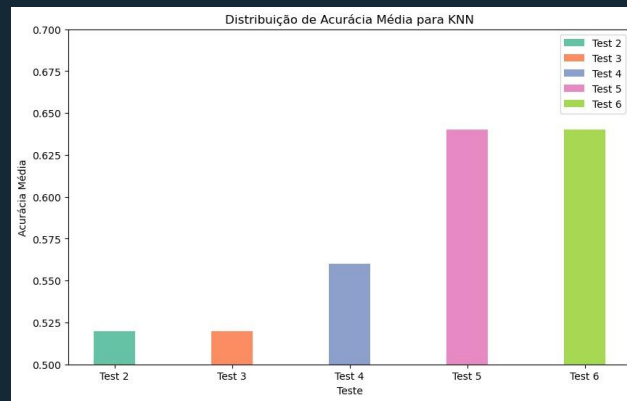
# Comparação de resultados

## DECISION TREE



O teste que apresentou melhores resultados foi o 2 (~0.61). Após isso, foi realizado um 'Parameter Tuning' onde foi possível alcançar um accuracy de 0.96.

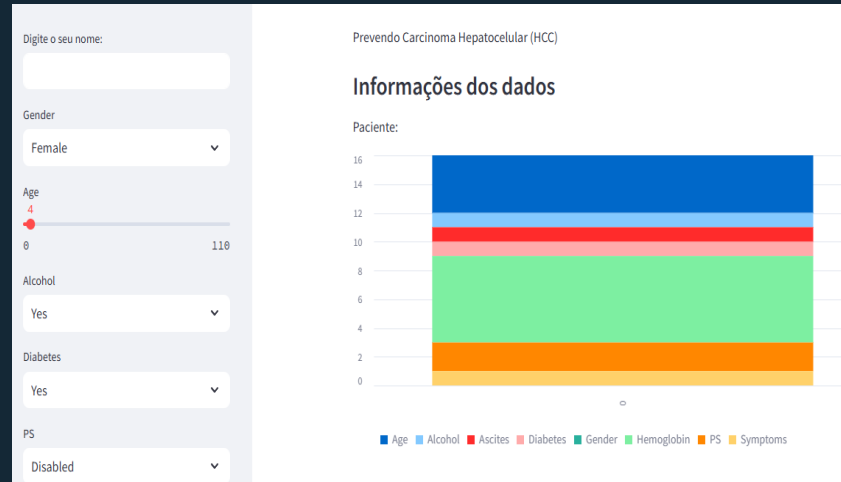
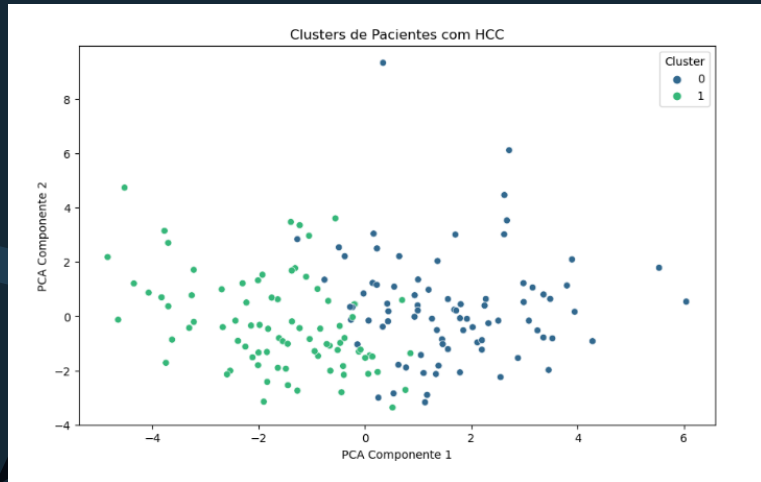
## K-NN



O teste que apresentou melhores resultados foi o 5 (~0.64). Após isso, foi realizado um 'Parameter Tuning' onde foi possível alcançar um accuracy de 0.72.

# Elementos extras

Para melhorar ainda mais o nosso trabalho, decidimos explorar um pouco mais outras funcionalidades e métodos para analisar os nossos dados. Para isso, utilizamos um método de aprendizagem não supervisionada (*clustering*) e o SMOTE (para atenuar certos desequilíbrios que pudesse existir no nosso target). Além disso, implementámos uma aplicação de previsão.



# Conclusão

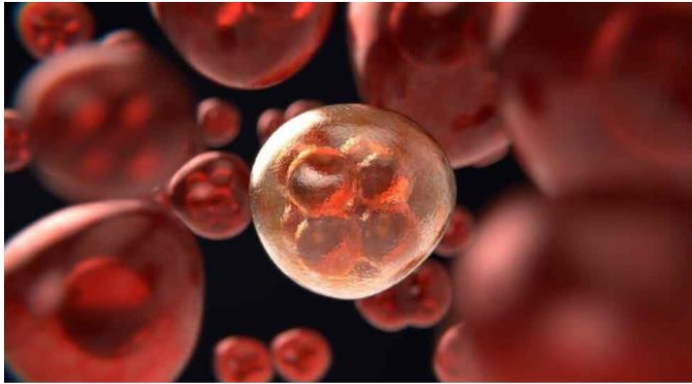
- Na análise de dados, concluímos que afinal nem todos os dados fornecidos são estritamente necessários para a realização do trabalho e que existiam valores com entradas nulas.
- Na classificação, realizamos 6 testes, onde tentamos responder a diversas perguntas, como a importância dos dados e o impacto desses dados nos resultados. Tendo sido feito o cross-validation para todos os testes exceto o primeiro, de modo a tornar o parameter tuning mais fácil de calcular. Neste foi apenas escolhido o teste com maior precisão para cada caso.
- Ao comparar os testes constatamos que estes não apresentam uma discrepância de valores, sendo todos muito próximos. De um modo geral concluímos que neste caso o melhor método foi o KNN, já que atingiu valores mais elevados.



# Link para o GitHub

HCC

*PREVISÃO ACERCA DA SOBREVIVÊNCIA DE PACIENTES COM CARCINOMA HEPATOCELULAR*



Para consultar o trabalho:

<https://github.com/caroleite05/HCC>



# Pesquisa bibliográfica

[https://github.com/Dr-Salcedo/hepatocellular\\_carcinoma\\_one\\_year\\_survival](https://github.com/Dr-Salcedo/hepatocellular_carcinoma_one_year_survival)