# Bootcamp Data Analytics 2024

## Challenge:  Modelos de Classificação

# Classificação de Doenças Cardíacas

**Contexto**

O **conjunto de dados utilizado é o Statlog Heart Disease, disponível no repositório UCI. Ele contém informações de 270 indivíduos e é composto por 14 colunas, selecionadas a partir de um conjunto maior que continha 75 colunas. Nesse conjunto, não há valores ausentes.**

**O objetivo desse conjunto de dados é realizar uma tarefa de classificação, onde se deve prever se uma pessoa tem ou não tem doença cardíaca. A variável de saída (o que queremos prever) é binária, sendo:**

**0: a pessoa não sofre de doença cardíaca.**

**1: a pessoa sofre de doença cardíaca.**

**Este é um estudo importante porque a saúde é um ponto vital de pesquisa para melhor ajudar os pacientes com certas condições. Além disso, a pressão arterial geralmente não apresenta sintomas e, no entanto, se a pressão alta não for tratada, pode ser um grande contribuinte para condições de saúde mais graves, como um derrame ou ataque cardíaco [2].**

**Dicionário de dados**

Este banco de dados contém 13 atributos e uma variável de destino. Possui 8 valores nominais e 5 valores numéricos. A descrição detalhada de todos esses recursos é a seguinte:

- **Age:** idade dos pacientes em anos
- **Sex**: (Masculino: 1; Feminino: 0)
- **cp**: Tipo de dor torácica sentida pelo paciente. Este termo é categorizado em 4 categorias.
- 0 angina típica,
- 1 angina atípica,
- 2 dor não anginosa,
- 3 assintomática
- **trestbps**: nível de pressão arterial do paciente no modo de repouso em mm/HG
- **chol**: colesterol sérico em mg/dl
- **fbs**: Níveis de açúcar no sangue em jejum > 120 mg/dl representa 1 em caso de verdadeiro e 0 como falso (Nominal) ●
  **restecg**: O resultado do eletrocardiograma em repouso é representado em 3 valores distintos
  0: Normal
- 1: com anormalidade da onda ST-T (inversões da onda T e/ou elevação ou depressão do ST > 0,05 mV)
- 2: mostrando provável ou definitiva hipertrofia ventricular esquerda por Critérios de Estes ●
  **thalach:** frequência cardíaca máxima alcançada
- **exang**: Angina induzida pelo exercício
- 0 retratando **Não**
- 1 retratando **Sim**
- **oldpeak:** Depressão do ST induzida pelo exercício em relação ao estado de repouso

- **slope:** segmento ST medido em termos de inclinação durante o pico do exercício
0: inclinação ascendente; ●
  1: plano;
- 2: inclinação descendente
- **ca:** O número de vasos principais (0–3) (nominal)
- **thal:** Um distúrbio sanguíneo chamado talassemia
  0: NULO
- 1: fluxo sanguíneo normal
- 2: defeito fixo (sem fluxo sanguíneo em alguma parte do coração)
- 3: defeito reversível (um fluxo sanguíneo é observado, mas não é normal (nominal)
- **target**: É a variável alvo que temos que prever 1 significa que o paciente sofre de doença cardíaca e 0 significa que o paciente é normal.

**Perguntas:**

1- Faça uma análise exploratória dos dados, observando as principais variáveis e sua relação com a variável target.

2 - Construa um modelo de regressão logística para classificar se o indivíduo sofre de doença cardíaca ou não.

3 - Analise o resultado da regressão logística e plote a matriz de confusão.