

# Projeto da Disciplina

Olá Pessoal, o projeto da disciplina será uma competição entre vocês para que escolher o melhor classificador com o *dataset* presente no link <https://bit.ly/2Ehg4Iu>.

## Dataset


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

## Framingham Heart Study

The Framingham Heart Study is a project of Boston University & the National Heart, Lung, & Blood Institute.

ABOUT ▼ PARTICIPANTS ▼ INVESTIGATORS ▼

RISK FUNCTIONS ▼ BIBLIOGRAPHY ▼ FOR RESEARCHERS ▼



The Framingham Heart Study has been renewed for an additional six-years, including the next research examination of the Offspring and Omni 1 Cohorts.

Thank you to our participants and investigators for their continued dedication.

Thank you to the National Heart, Lung, and Blood Institute (NHLBI) for its continued support.

*Additional details here.*

O *dataset* possui 4240 entradas e os seguintes campos:

```
import pandas as pd

df = pd.read_csv('https://bit.ly/2Ehg4Iu')
df.head()
```

age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
39	4.0	0	0.0	0.0	0	0	0	195.0	106.0	70.0	26.97	80.0	77.0	0
46	2.0	0	0.0	0.0	0	0	0	250.0	121.0	81.0	28.73	95.0	76.0	0
48	1.0	1	20.0	0.0	0	0	0	245.0	127.5	80.0	25.34	75.0	70.0	0
61	3.0	1	30.0	0.0	0	1	0	225.0	150.0	95.0	28.58	65.0	103.0	1
46	3.0	1	23.0	0.0	0	0	0	285.0	130.0	84.0	23.10	85.0	85.0	0

## Features:

- **Age:** Idade do indivíduo
- **Education:** Nível de educação (quanto mais alto maior o nível de educação)
- **currentSmoker:** Se fumante (1) ou não fumante (0)

- **cigsPerDay:** Número de cigarros por dia
- **BPMeds:** Número de batimentos por dia
- **prevalentStroke:** Se paciente já infartou alguma vez
- **prevalentHyp:** Se paciente já foi hipertenso
- **diabetes:** Se paciente já foi diabético
- **totChol:** Quantidade de colesterol total do paciente
- **sysBP:** Pressão sistólica
- **diaBP:** Pressão diastólica
- **BMI:** Índice de massa corporal BMI = Peso/altura^2
- **heartRate:** Frequência cardíaca
- **glucose:** Taxa de açúcar no sangue

Label:

- **TenYearCHD (Label):** Risco em 10 anos de doença coronária

## Objetivo do Projeto

Suponha que eu seja o seu cliente e você já extraiu os dados presentes no dataset (<https://bit.ly/2Ehg4lu>). Eu preciso que você me traga:

1. O melhor classificador que vocês consigam obter para o label **TenYearCHD**.
2. Um relatório com descobertas interessantes que se possa obter com esses dados. Esse relatório pode ser escrito no próprio notebook ou um arquivo word / pdf com as conclusões que vocês podem tirar dos dados (Sejam criativos).
3. **IMPORTANTE: Implementem o algoritmo de regressão logística como uma rede neural assim como mostrado na aula do dia 18/05. O algoritmo é apresentado na imagem a seguir.**

**Lembre-se da função de Erro  $L(a, y)$  e a função  $y = a \rightarrow$  função sigmoide que aprendemos na nossas aulas.**

$$z = w^T x + b$$

$$\hat{y} = a = \sigma(z)$$

$$\mathcal{L}(a, y) = -(y \log(a) + (1 - y) \log(1 - a))$$

- $y = a$  = Função sigmoide de  $z$ . Basicamente a função de regressão logística que aprendemos nas últimas aulas.
- $\mathcal{L}(a, y)$  = Função de custo. Essa função determina o erro na previsão

O algoritmo para treinamento da rede com um único neurônio é dado a seguir:

$$\begin{aligned}
 &J=0; \underline{dw}_1=0; \underline{dw}_2=0; \underline{db}=0 \\
 &\text{For } i=1 \text{ to } m \\
 &\quad z^{(i)} = w^T x^{(i)} + b \\
 &\quad a^{(i)} = \sigma(z^{(i)}) \\
 &\quad J += -[y^{(i)} \log a^{(i)} + (1-y^{(i)}) \log(1-a^{(i)})] \\
 &\quad \underline{dz}^{(i)} = a^{(i)} - y^{(i)} \\
 &\quad dw_1 += x_1^{(i)} dz^{(i)} \\
 &\quad dw_2 += x_2^{(i)} dz^{(i)} \\
 &\quad db += dz^{(i)} \\
 &J /= m \leftarrow \\
 &\underline{dw}_1 /= m; \underline{dw}_2 /= m; \underline{db} /= m. \leftarrow
 \end{aligned}$$

$dw_1 = \frac{\partial J}{\partial w_1}$   
  
 $w_1 := w_1 - \alpha \underline{dw}_1$   
 $w_2 := w_2 - \alpha \underline{dw}_2$   
 $b := b - \alpha \underline{db}$

Onde cada variável é definida como:

- J, Função de Erro
- dw, derivada de J em função de w
- db, derivada de J em função de b
- m, número de elementos no conjunto de treinamento
- a, previsão da regressão logística
- w, pesos da regressão

4.

Informações Importantes:

- Você pode manipular o dataset como quiser, inclusive removendo e inserindo novas colunas se for necessário.
- Reporte também quais variáveis são mais importantes utilizando gráficos ou estatísticas. (Lembre-se – O máximo de informação relevante vai ajudar no entendimento do problema)
- Traga insights interessantes como percentuais de pessoas que já enfartaram ou não dependendo se ela fuma ou não. (Dica – utilize probabilidades condicionais)
- Analise dependências entre as variáveis e apresente insights sobre isso.
- Apresente comentários no notebook para que seja fácil de ser entendido por você e por mim.
- Na dúvida, pergunte no Edmodo.

Prazo

O projeto deverá ser enviado até o dia 31 de maio. Atrasos não serão tolerados. Portanto, se você não terminou o trabalho, mande do modo que está que considerarei o que foi feito.