Projeto da Disciplina

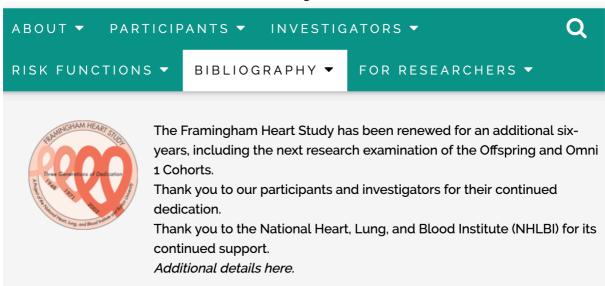
Olá Pessoal, o projeto da disciplina será uma competição entre vocês para que escolher o melhor classificador com o *dataset* presente no link https://bit.ly/2Ehg4lu.

Dataset

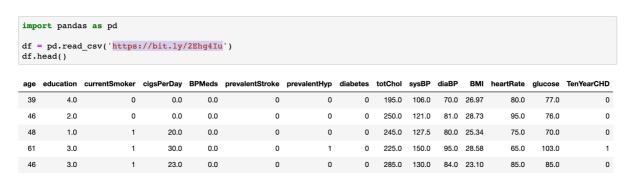
https://www.framinghamheartstudy.org/

Framingham Heart Study

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



O dataset possui 4240 entradas e os seguintes campos:



Features:

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

- <u>cigsPerDay:</u> Número de cigarros por dia
- BPMeds: Número de batimentos por dia
- prevalentStroke: Se paciente já infartou alguma vez
- prevalentHyp: Se pacience 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 cardiáca
- 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. <u>IMPORTANTE: Implementem o algoritmo de regressão logística como uma rede</u> 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 -> 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.
- 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:

$$J=0; d\omega_{1}=0; d\omega_{2}=0; db=0$$

$$Eor i=1 to m$$

$$E(i)=\omega^{T}x^{(i)}tb$$

$$E(i)=\omega^{T}x^{(i)}$$

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.