

## Bootcamp: Arquiteto(a) de Big Data

### Desafio Prático

#### Módulo 1: Fundamentos de Big Data

#### Objetivos de Ensino

Exercitar os seguintes conceitos trabalhados no Módulo:

- ✓ Coleta de dados.
- ✓ Manipulação e visualização de dados.
- ✓ Tratamento de dados.
- ✓ Aplicação de algoritmo de machine learning.
- ✓ Análise de dados gerados.
- ✓ Conhecimento teórico ministrado nas videoaulas.

#### Enunciado

Uma operadora de seguro de saúde identificou na sua base de dados de clientes a relação entre os dados de colesterol e peso, com a incidência de problemas que influenciam no desenvolvimento de doenças cardíacas. Pensando no bem-estar dos seus clientes e ao mesmo tempo pensando em diminuir problemas de internação e tratamento para esses tipos de causa, a operadora quer realizar um estudo para identificar o perfil de pessoas que se encaixam nos grupos de risco e assim realizar medidas preventivas e palestras de cuidados médicos. Para isso, a operadora conta com a equipe de arquiteto de big data para ajuda-los a encontrar o grupo de risco dentre essas pessoas.

Através da análise dos dados, os analistas da operadora de seguro de saúde identificaram quatro grandes grupos:

1. Alto Risco.
2. Risco Moderado alto.
3. Risco Moderado baixo.
4. Baixo risco.

## Atividades

Os alunos deverão desempenhar as seguintes atividades:

1. Criar um projeto no Google Drive;
2. Coletar e analisar os dados dos seguintes datasets:
  - a. dados\_clientes.
  - b. estados\_brasileiros.
  - c. idade\_clientes.
3. Manipular dados e corrigir erros se necessário;
4. Implementar algoritmo não supervisionado kmeans;
5. Criar agrupamento para quatro grupos distintos;
6. Responder as questões teóricas e práticas do trabalho.

### Dicas do professor:

1. Para o cálculo do WCSS, vocês devem escolher apenas os atributos de peso e colesterol.
2. Utilize se necessário a estratégia de exclusão de dados ausentes nos datasets.
3. Atenção para os indicadores dos agrupamentos realizados.
4. Analisem os dados com cuidados, visualizem os dados nos gráficos.
5. Terceiro cluster e diferente de cluster = 3.
6. Ao analisar os grupos de riscos, observem cuidadosamente a coloração para responder as questões de centroides.
7. As bases de dados utilizadas no trabalho podem ser obtidas também no link:
  - <https://github.com/ProfLeandroLessa/classroom-datasets/tree/master/FDA/Desafio>

**Bom desafio a todos!**