# Bootcamp: Arquiteto(a) de Big Data

## Desafio Prático

## Módulo 1: Fundamentos de Big Data

### Objetivos de Ensino

Exercitar os seguintes conceitos trabalhados no Módulo:

1. Coleta de dados.
2. Analisar e realizar tratamento de dados.
3. Criar visualização de dados.
4. Implementar algoritmo de Machine Learning.
5. Analisar resultados obtidos.
6. Conhecimento teórico ministrado nas videoaulas.

### Enunciado

Como Arquiteto de Big Data em uma equipe de saúde, você faz parte de um projeto com o objetivo de identificar padrões e riscos de saúde em pacientes, com base em suas informações demográficas e biomarcadores específicos. Nesse estudo, nosso foco é analisar as variáveis de gênero, idade, peso e colesterol para avaliar o risco de desenvolvimento de problemas cardiovasculares.

Para alcançar esse objetivo, a equipe decidiu utilizar o algoritmo de agrupamento k-means para segmentar os pacientes em três grupos distintos, com base nessas variáveis selecionadas. O intuito é identificar padrões nos dados que possam indicar o risco de um paciente desenvolver doenças cardiovasculares. A

identificação precoce desses riscos é essencial para proporcionar intervenções adequadas e tratamentos personalizados, a fim de prevenir complicações graves.

Os três clusters identificados neste estudo são:

Baixo Risco: esse grupo inclui pacientes com características demográficas e biomarcadores associados a um risco relativamente baixo de desenvolvimento de problemas cardiovasculares. Eles podem ter um perfil mais jovem, peso saudável e níveis de colesterol dentro da faixa normal.

Risco Moderado: nesse grupo, encontramos pacientes que possuem algumas características que indicam um risco moderado de problemas cardiovasculares. Eles podem apresentar uma combinação de fatores de risco, como idade mais avançada, peso um pouco acima do ideal e níveis de colesterol elevados, mas ainda dentro de limites considerados moderados.

Risco Alto: este grupo contém pacientes com características que indicam um risco significativamente elevado de desenvolvimento de problemas cardiovasculares. Esses pacientes podem ser mais velhos, apresentar excesso de peso ou obesidade e ter níveis de colesterol muito acima dos limites recomendados.

Através dessa análise, nosso objetivo é fornecer informações valiosas para a equipe de saúde, possibilitando a tomada de decisões mais assertivas em relação aos pacientes e a implementação de medidas preventivas e personalizadas para garantir uma melhor qualidade de vida e redução dos riscos cardiovasculares.

## ATENÇÃO PARA TRATAMENTO DE DADOS

Avaliem se será necessário realizar tratamento de dados ausentes nos datasets disponibilizados.

Instruções para correção de dados ausentes

1. Para dados de estados utilize a estratégia de exclusão dos dados.
2. Para os dados de clientes utilize:

a. Mediana arredondada para duas casas decimais para as variáveis do tipo numéricas.

b. Moda para as variáveis categóricas.

## Atividades

Para essa atividade, os alunos deverão criar um algoritmo de K-means para criar um agrupamento de pessoas baseados nas suas características.

Criar um projeto no Google Drive.

1. Coletar e inserir na plataforma os arquivos:
    a. dados_clientes.xlsx
    b. estados_brasileiros.csv
    c. idade_clientes.csv
2. Analisar os dados coletados.
3. Avaliar a correlação entre as variáveis.
4. Criar algoritmo de k-means com as configurações:
    a. random_state=0, init='k-means++'
5. Responder as questões teóricas e práticas do trabalho.
6. Realizar a segmentação dos clientes.
7. Análise dos dados de acordo com clusters criados.

## Dicas do professor:

1. Leia atentamente todas as instruções da atividade.
2. Analisem com cuidado os dados através da representação gráfica.

3. Crie um novo dataframe (pacientes) com os dados obtidos entre a relação das tabelas disponibilizadas.
4. Para criação do algoritmo de clusterização, utilize as variáveis de idade e colesterol.
5. Analisem bem o gráfico gerado e a disponibilização dos dados do agrupamento.
    a. Atribua o grupo de risco baseado nas cores apresentadas no gráfico.
    b. Faz parte da atividade o aluno analisar os dados no gráfico e escolher qual o indicador para cada grupo.
    c. Utilizem o bom senso no momento de analisar os dados do gráfico e atribuir os grupos de risco. Avaliem a disposição.
6. Antes de enviar as respostas, verifiquem se o gabarito está correto.
7. Tenham atenção no que pede cada questão.
8. Os dados disponibilizados no dataset são fictícios. Ou seja, não tem relação com o mundo real.
9. Sigam exatamente o que o enunciado do trabalho solicita.
    a. Obs.: não existe pegadinhas nas alternativas.
10. Os datasets utilizados no trabalho podem ser obtidos no link:
    a. https://github.com/ProfLeandroLessa/classroom-datasets/tree/master/FDA/Desafio

## Biblioteca utilizadas

```
pandas: 1.5.2
sklearn: 1.2.0
plotly: 5.11.0
```