# Bootcamp: Arquiteto(a) de Big Data

## Trabalho 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

Previsão do Nível de Colesterol com Base no Peso usando Regressão Linear

Uma empresa de saúde coletou dados clínicos de diversos pacientes, organizados por gênero, e deseja prever os níveis de colesterol dos pacientes com base em seus pesos. Esta atividade visa explorar a relação entre o peso e o nível de colesterol, utilizando técnicas de regressão linear para desenvolver um modelo preditivo.

Como arquiteto de dados, seu objetivo é desenvolver um modelo de regressão linear que preveja os níveis de colesterol dos pacientes com base em seus pesos, levando em consideração as diferenças de gênero. Você utilizará os dados clínicos fornecidos pela empresa de saúde para realizar essa análise.

Importância da Atividade

A regressão linear é uma ferramenta poderosa para modelar e entender a relação entre variáveis dependentes e independentes. Neste caso, a capacidade de prever os níveis de colesterol com base no peso pode fornecer insights valiosos para a saúde pública. Essa previsão pode identificar indivíduos com maior risco de problemas cardiovasculares, permitindo intervenções preventivas e personalizadas.

Riscos de Problemas Cardíacos

Os níveis elevados de colesterol estão intimamente relacionados ao aumento do risco de doenças cardiovasculares, incluindo ataques cardíacos e acidentes vasculares cerebrais. Identificar e monitorar indivíduos em risco pode ajudar a prevenir o desenvolvimento dessas condições graves. Portanto, o uso da regressão linear para prever os níveis de colesterol pode ter um impacto significativo na promoção da saúde cardiovascular e na redução da morbidade e mortalidade associadas a doenças cardíacas.

ATENÇÃO PARA TRATAMENTO DE DADOS

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

Instruções para correção de dados ausentes

1. Média arredondada para 2 casas decimais para as variáveis do tipo numéricas.

2. Moda para as variáveis categóricas.

## Atividades

Para esta atividade, os alunos deverão criar um algoritmo de regressão linear para prever o nível de colesterol com base no peso das amostras.

1. Criar um projeto no Google Drive.
2. Coletar e inserir o arquivo dado_clinicos.zip.
    a. O arquivo se encontra em: https://leandrolessa.com.br/datasets/
3. Analisar os dados coletados.
4. Avaliar a relação entre as variáveis.
5. Criar algoritmo de regressão linear.
6. Responder às questões teóricas e práticas do trabalho.

## Dicas do professor:

1. Antes de prosseguir com qualquer análise ou desenvolvimento de algoritmos de machine learning, é essencial compreender profundamente a base de dados e realizar uma verificação minuciosa dos dados disponíveis.
2. Realize o tratamento de dados antes de responder às questões.
3. Analise com cuidado os dados através da representação gráfica.
4. Elimine os dados duplicados caso necessário.
5. Corrija dados ausentes caso necessário.
6. Analise bem o gráfico gerado e a disponibilização dos dados.
7. Antes de enviar as respostas, verifique se o gabarito está correto.

8. Tenha atenção no que pede cada questão.

9. Para obter mais orientações sobre como criar um gráfico de dispersão solicitado em uma questão, siga estas dicas adicionais:

   https://leandrolessa.com.br/tutoriais/grafico-de-dispersao-como-criar-e-analisar-na-pratica/