Universidade de São Paulo
ICMC/USP
Bacharelado em Ciências de Dados

Professor: Dr. Fabiano Rodrigues Coelho
1º Semestre de 2025
Disciplina: 5500004

Instruções: Entregar via e-mail em Google Colab ou similares. Não se esqueça de abrir uma caixa de texto no começo do arquivo e inserir um um link do seu trabalho

---

Trabalho 6: Data de Entrega 11/06/2025 até às 23h59min

---

1. A base de dados `insurance` contém informações sobre clientes de um plano de saúde nos Estados Unidos. Cada linha representa um cliente, com variáveis como idade, sexo, IMC (índice de massa corporal), número de filhos, se a pessoa é fumante ou não e a região de residência. A variável `charges` representa o valor cobrado pelo seguro.

   Com base nessas informações, responda aos itens a seguir:

   (a) Faça a leitura da base de dados e imprima as cinco primeiras linhas.

   ```
   import pandas as pd
   url = "https://raw.githubusercontent.com/stedy/Machine-Learning-with-R-datasets/master/insurance.csv"
   df = pd.read_csv(url)
   df.head()
   ```

   (b) Faça uma análise descritiva dos dados.

   (c) Considerando que o objetivo é prever o valor cobrado pelo seguro, treine um modelo de aprendizado de máquina adequado para essa tarefa.

   (d) Descreva o processo de aprimoramento do modelo treinado, considerando estratégias como transformação de variáveis, inclusão de variáveis dummies, regularização ou seleção de variáveis, investigação de multicolinearidade e diagnóstico de influência.

   (e) Avalie a capacidade preditiva do modelo por meio de métricas apropriadas (como R^2, RMSE ou MAE), interpretando os resultados obtidos.