Relato Técnico e Reflexivo — Análise Framingham Heart Study

Raciocínio por trás da construção do projeto

Ao receber o dataset do Framingham Heart Study, a primeira coisa que considerei foi: este é um dataset epidemiológico clássico, utilizado historicamente para estudos de risco cardiovascular. Portanto, ele não é um dataset construído especificamente para modelos de previsão, mas sim para análise de fatores de risco.

Diante disso, minha proposta foi transformar esse problema em uma análise quantitativa, com foco em explicar como variáveis clínicas e comportamentais impactam os níveis de **colesterol total (totChol)**.

A escolha do colesterol como variável-alvo foi baseada no fato de que:

- É um dos principais marcadores de risco cardiovascular.
- Tem dependência multifatorial clara — envelhecimento, obesidade, pressão arterial, glicemia e tabagismo.
- É uma variável clínica de fácil interpretação.

Estrutura do trabalho

1. **Entendimento dos dados:**
   - Realizei uma análise exploratória completa (EDA) para entender os padrões, outliers e dados ausentes.
   - Optei por preencher valores ausentes com a mediana, devido à robustez dessa medida frente a dados assimétricos comuns em saúde.
2. **Seleção de variáveis:**
   - Foco nas variáveis clinicamente justificáveis:
     - Idade
     - IMC
     - Pressão Sistólica (sysBP)
     - Pressão Diastólica (diaBP)
     - Glicose
     - Cigarros por dia
3. **Modelagem:**
   - Construção da regressão linear **manual**, utilizando a fórmula dos mínimos quadrados.
   - Esse caminho foi proposital para reforçar a compreensão dos fundamentos matemáticos da regressão.
4. **Avaliação:**
   - Métricas como MSE, RMSE e R^2.
   - Visualização Real vs. Predito para validação gráfica da aderência do modelo.

O que aprendi no processo

- A regressão linear vai muito além de simplesmente rodar um método pronto como .fit(). Entender o que acontece por trás, especialmente a construção da equação de mínimos quadrados, fortalece muito a segurança na análise.
- Dados clínicos exigem cuidado extremo com tratamento de outliers, valores faltantes e interpretação.
- A relação entre as variáveis é multifatorial. Nenhuma variável sozinha explica o colesterol, mas o conjunto tem um poder preditivo relevante.
- A parte mais rica do processo foi **interpretar os coeficientes**, entendendo que:
  - Idade, IMC, pressão, glicose e tabagismo impactam diretamente o colesterol.
  - Esse impacto pode ser quantificado e comunicado de forma objetiva.

Reflexão final

- Esse trabalho foi mais do que uma análise de dados; foi um exercício completo de pensamento crítico e construção analítica.
- Cada etapa, desde o entendimento dos dados até a avaliação, foi guiada por critérios técnicos e também clínicos.
- O maior ganho não foi gerar um modelo matematicamente ajustado, mas sim gerar uma análise **clinicamente interpretável, estatisticamente robusta e aplicável ao mundo real da saúde.**

Conclusão

O desenvolvimento desse projeto me proporcionou uma visão madura de como a estatística, a ciência de dados e o conhecimento de domínio (neste caso, saúde) se integram para gerar valor, entendimento e soluções que fazem sentido na prática.

Reflexão final

- Esse trabalho foi mais do que uma análise de dados; foi um exercício completo de pensamento crítico e construção analítica.
- Cada etapa, desde o entendimento dos dados até a avaliação, foi guiada por critérios técnicos e também clínicos.
- O maior ganho não foi gerar um modelo matematicamente ajustado, mas sim gerar uma análise **clinicamente interpretável, estatisticamente robusta e aplicável ao mundo real da saúde.**