Universidade Federal do Pará - UFPA
Instituto de Ciências Biológicas – ICB
Programa de Pós-Graduação em Genética e Biologia Molecular

Disciplina/módulo: Machine Learning para Bioinformática
Professor: Gilderlanio Santana de Araújo

Exercício II

Contexto: Você foi contratado para análises de um conjunto de dados de expressão gênica de pacientes com câncer. O conjunto de dados contém informações de um painel de genes em cada paciente. No entanto, alguns valores de expressão gênica estão ausentes devido a erros experimentais. Suas tarefas são:

TF1: Imputação por Média

A imputação por média é um método simples e rápido para lidar com dados ausentes em conjuntos de dados. Ela funciona substituindo os valores ausentes pela média dos valores presentes na mesma coluna.

TF2: Imputação por MICE

O MICE (Multiple Imputation by Chained Equations) é um método estatístico avançado para lidar com dados ausentes em conjuntos de dados multivariados. Ele funciona através da criação de múltiplas imputações dos valores ausentes, levando em consideração as relações entre as variáveis do conjunto de dados.

Referência: https://www.jstatsoft.org/article/view/v045i03/

Passo a passo:

1. Carregue o conjunto de dados de expressão gênica, disponível no GitHub:
    1. Dados se encontram na pasta:
    https://github.com/Gilderlanio/notasdeaulas/tree/main/machine_learning/datasets/gene_expression
2. Identifique a proporção de dados ausentes em cada coluna.
3. Impute os valores ausentes em cada coluna utilizando a média dos valores presentes e a estratégia MICE.
4. Compare a distribuição dos dados originais com a distribuição dos dados imputados em cada estratégia.
5. Calcule a correlação entre as variáveis antes e depois da imputação.
6. Discuta o impacto da imputação por média e por MICE na distribuição dos dados e na correlação entre as variáveis.