

## Quarta Lista de Exercícios: Agrupamento Hierárquico, Árvores de Decisão e Florestas Aleatórias

**Exercício 1.** Considere a seguinte matriz de distâncias:

$$\begin{bmatrix} 0 & 9 & 3 & 6 & 11 \\ 9 & 0 & 7 & 5 & 10 \\ 3 & 7 & 0 & 9 & 2 \\ 6 & 5 & 9 & 0 & 8 \\ 11 & 10 & 2 & 8 & 0 \end{bmatrix}$$

- (a) Com base na matriz de distâncias acima, esboce o dendograma que resulta do processo de aglomeração hierárquica dessas 5 observações usando o método **complete** como a distância entre dois aglomerados.
- (b) Repita o exercício (a) utilizando o método **single** como a distância entre dois aglomerados.
- (c) Suponha que um corte seja feito no dendograma encontrado em (a) de forma a deixar dois aglomerados. Quais observações estão em cada aglomerado?
- (d) Suponha que um corte seja feito no dendograma encontrado em (b) de forma a deixar dois aglomerados. Quais observações estão em cada aglomerado?

**Exercício 2.** O objetivo desse exercício é construir manualmente uma árvore de decisão para ajudar a prever quando um paciente poderá ter um ataque cardíaco. Os dados de treinamento par esse problema estão no arquivo `heart.txt`.

- (a) Usando a impureza de Gini construa uma árvore de decisão que irá prever quando um paciente terá ou não um ataque cardíaco. Deixe indicado todos os passos da construção.
- (b) Traduza a árvore construída acima em um código. O código deve ser uma função cuja entrada é um vetor de tamanho 4 referente às variáveis explanatórias do modelo (cada entrada sendo yes ou no) e cuja saída é **yes** (é provável que o paciente tenha um ataque cardíaco) ou **no** (é provável que o paciente não tenha um ataque cardíaco).

**Exercício 3.** O conjunto de dados `SBI.csv` contém informações de mais de 2.348 crianças que compareceram aos serviços de emergência de um hospital com febre e que foram submetidas a um teste para detecção de infecção bacteriana grave. A variável `sbi` possui 4 categorias: Not Applicable, UTI, Pneum e Bact. Not Applicable significa que o teste deu negativo; já as outras categorias indicam a existência de infecção bacteriana grave.

- (a) Acrescente ao conjunto de dados uma nova coluna chamada **infection**. Essa variável será **yes** se a criança foi diagnosticada com infecção grave e **no** caso contrário. Lembre-se que essa variável deve ser do tipo categórica (factor).
- (b) Retire do conjunto de dados as variáveis **X**, **id** e **sbi**.
- (c) Separe o conjunto de dados em dois novos conjuntos, um para treino e um para teste. O conjunto para treino deverá ter 80% dos dados iniciais.
- (d) Crie um modelo de árvore de decisão para classificar a variável **infection** a partir das outras variáveis do conjunto de treinamento. Plote a representação gráfica da árvore resultante. A partir do conjunto de teste e da função **predict**, verifique a acurácia do modelo. Por fim, crie uma matriz de confusão da previsão versus respostas verdadeiras.
- (e) Repita o item acima (exceto a parte referente à representação gráfica) para um modelo de floresta aleatória.