# LABORATÓRIO DE R – 03

**Classificação:**

1. Considere o conjunto de dados inibina (disponível aqui) obtido de um estudo cuja finalidade era avaliar a utilização da inibina B como marcador da reserva ovariana de pacientes submetidas à fertilização in vitro. A variável explicativa é a diferença entre a concentração sérica de inibia B após estímulo com o hormônio FSH (inibpos) e sua concentração sérica pré estímulo (inibpre) e a variável resposta é a classificação das pacientes como boas ou más respondedoras com base na quantidade de oócisto recuperados.
    a. Faça uma breve análise descritiva dos dados.
    b. Como você modelaria esse conjunto de dados de acordo com o objetivo do estudo?
    c. Ajuste um modelo de regressão logística aos dados. Qual é a acurácia do modelo em fazer classificação?
    d. Use o classificador linear de Fisher para classificar a variável resposta de acordo com a variável preditora. Qual é a acurácia do classificador?
    e. Use o classificador linear de Bayes para classificar a variável resposta de acordo com as variáveis explicativas. Utilize priori 0,65 e 0,35 para resposta negativa e positiva, respectivamente. Qual é a acurácia do classificador?
    f. Use o classificador knn para classificar a variável resposta de acordo com as variáveis preditoras. Utilize k = 1, 3, 5.
    g. Use naive Bayes para para classificar a variável resposta de acordo com as variáveis preditoras.
    h. Use uma arvore de decisão para classificar a variável resposta de acordo com as variáveis preditoras.
    i. Use svm para fazer para classificar a variável resposta de acordo com as variáveis preditoras.
    j. Use uma rede neural para classificar a variável resposta de acordo com as variáveis preditoras.
    k. Refaça os itens *c* à *j* usando o método de validação cruzada *leave-one-out*. Dica: use a função train do pacote caret.
    l. Qual classificador você escolheria?

2. Os dados disponíveis aqui foram extraídos de um estudo realizado no Hospital Universitário da Universidade de São Paulo com o objetivo de avaliar se algumas medidas obtidas ultrassonograficamente poderiam ser utilizadas como substitutas de medidas obtidas por métodos de ressonância magnética, considerada como padrão ouro para avaliação do deslocamento do disco da articulação temporomandibular (referido simplesmente como disco). Distâncias cápsula-côndilo (em mm) com boca aberta ou fechada (referidas, respectivamente, como distância aberta ou fechada) foram obtidas ultrassonograficamente de 104 articulações e o disco correspondente foi classificado como deslocado (1) ou não (0) segundo a avaliação por ressonância magnética.
    a. Qual é a variável resposta e quais são as explicativas? Faça uma breve análise descritiva dos dados.
    b. Como você modelaria esse conjunto de dados?
    c. Separe o conjunto de dados em conjunto de treinamento (70% doas dados) e conjunto de teste (30%).
    d. Ajuste um modelo de regressão logística conjunto de treinamento. Qual é a acurácia do modelo no conjunto de teste?

e. Use o classificador de Fisher para classificar a variável resposta de acordo com a variável preditora. Qual é a acurácia do classificador? Ajuste o modelo no conjunto de treino e faça a predição no conjunto de teste.
f. Use o classificador de Bayes para classificar a variável resposta de acordo com a variável preditora. Utilize priori 0,65 e 0,35 para 0 e 1, respectivamente. Qual é a acurácia do classificador? Ajuste o modelo no conjunto de treino e faça a predição no conjunto de teste.
g. Use o classificador knn para classificar a variável resposta de acordo com a variável preditora. Utilize k = 1, 3, 5. Ajuste o modelo no conjunto de treino e faça a predição no conjunto de teste.
h. Use naive Bayes para classificar a variável resposta de acordo com as variáveis preditoras. Ajuste o modelo no conjunto de treino e faça a predição no conjunto de teste.
i. Use uma arvore de decisão para classificar a variável resposta de acordo com as variáveis preditoras. Ajuste o modelo no conjunto de treino e faça a predição no conjunto de teste.
j. Use svm para fazer para classificar a variável resposta de acordo com as variáveis preditoras. Ajuste o modelo no conjunto de treino e faça a predição no conjunto de teste.
k. Use uma rede neural para classificar a variável resposta de acordo com as variáveis preditoras. Ajuste o modelo no conjunto de treino e faça a predição no conjunto de teste.
l. Qual classificador você escolheria?

3. Os dados do arquivo tipofacial (disponível aqui) foram extraídos de um estudo odontológico realizado pelo Dr. Flávio Cotrim Vellini. Um dos objetivos era utilizar medidas entre diferentes pontos do crânio para caracterizar indivíduos com diferentes tipos faciais, a saber, braquicéfalos, mesocéfalos e dolicocéfalos. O conjunto de dados contém observações de 11 variáveis em 101 pacientes. Para efeitos didáticos, considere apenas a altura facial (altfac) e a profundidade facial (proffac) como variáveis preditoras.
    a. Faça uma breve análise descritiva. Desenhe um gráfico das variáveis altura facial X profundidade facial e marque cada observação de acordo com o tipo facial do indivíduo.
    b. Separe o conjunto de dados em conjunto de treinamento (70% dos dados) e conjunto de teste (30%). Faça isso de 5 formas diferentes.
    c. É possível ajustar um modelo de regressão logística nesse conjunto de dados? Por quê?
    d. Use o classificador de Fisher para classificar a variável resposta de acordo com variáveis preditoras. Qual é a acurácia do classificador? Ajuste o modelo no conjunto de treino e faça a predição no conjunto de teste.
    e. Use o classificador de Bayes para classificar a variável resposta de acordo com as variáveis preditoras. Utilize como priori a proporção amostral. Qual é a acurácia do classificador? Ajuste o modelo no conjunto de treino e faça a predição no conjunto de teste.
    f. Use o classificador knn para classificar a variável resposta de acordo com a variável preditora. Utilize k = 1,3,5. Ajuste o modelo no conjunto de treino e faça a predição no conjunto de teste.
    g. Use naive Bayes para classificar a variável resposta de acordo com as variáveis preditoras. Ajuste o modelo no conjunto de treino e faça a predição no conjunto de teste.
    h. Use uma arvore de decisão para classificar a variável resposta de acordo com as variáveis preditoras. Ajuste o modelo no conjunto de treino e faça a predição no conjunto de teste.
    i. Use svm para fazer para classificar a variável resposta de acordo com as variáveis preditoras. Ajuste o modelo no conjunto de treino e faça a predição no conjunto de teste.
    j. Use uma rede neural para classificar a variável resposta de acordo com as variáveis preditoras. Ajuste o modelo no conjunto de treino e faça a predição no conjunto de teste.
    k. Qual classificador você escolheria?

4. Refaça o exercício 3 usando outras variáveis como preditoras (a sua escolha) e usando outro método de validação cruzada no a partido do item b.

**Agrupamento**:

5. Considere os dados a seguir do consumo alimentar médio de diferentes tipos de alimentos para famílias classificadas de acordo com o número de filhos (2, 3, 4 ou 5) e principal área de trabalho (MA: Setor de Trabalho Manual, EM: Empregados do Setor Público ou CA: Cargos Administrativos):

| | | Pães | Vegetais | Frutas | Carnes | Aves | Leite | Alcóolicos |
|---|---|---|---|---|---|---|---|---|
| 1 | MA2 | 332 | 428 | 354 | 1437 | 526 | 247 | 427 |
| 2 | EM2 | 293 | 559 | 388 | 1527 | 567 | 239 | 258 |
| 3 | CA2 | 372 | 767 | 562 | 1948 | 927 | 235 | 433 |
| 4 | MA3 | 406 | 563 | 341 | 1507 | 544 | 324 | 407 |
| 5 | EM3 | 386 | 608 | 396 | 1501 | ——— | 319 | 363 |
| 6 | CA3 | 438 | 843 | 689 | 2345 | 1148 | 243 | 341 |
| 7 | MA4 | 534 | 660 | 367 | 1620 | 638 | 414 | 407 |
| 8 | EM4 | 460 | 699 | 484 | 1856 | 762 | 400 | 416 |
| 9 | CA4 | 385 | 789 | 621 | 2366 | 1149 | 304 | 282 |
| 10 | MA5 | 655 | 776 | 423 | 1848 | 759 | 495 | 486 |
| 11 | EM5 | 584 | 995 | 548 | 2056 | 893 | 518 | 319 |
| 12 | CA5 | 515 | 1097 | 887 | 2630 | 1167 | 561 | 284 |

Um código no R para o conjunto de dados é dado a seguir:

```
library(tibble)
dados = tibble(AreaTrabalho = as.factor(rep(c("MA", "EM", "CA"), 4)),
               Filhos = as.factor(rep(2:5, each = 3)),
               Paes =       c(332, 293, 372, 406, 386, 438, 534, 460, 385, 655, 584, 515),
               Vegetais =   c(428, 559, 767, 563, 608, 843, 660, 699, 789, 776, 995, 1097),
               Frutas =     c(354, 388, 562, 341, 396, 689, 367, 484, 621, 423, 548, 887),
               Carnes =     c(1437,1527,1948,1507,1501,2345,1620,1856,2366,1848,2056,2630),
               Aves =       c(526, 567, 927, 544, NA,  1148,638, 762, 1149,759, 893, 1167),
               Leite =      c(247, 239, 235, 324, 319, 243, 414, 400, 304, 495, 518, 561),
               Alcoolicos = c(427, 258, 433, 407, 363, 341, 407, 416, 282, 486, 319, 284))
```

a. Utilize regressão linear para predizer o dado faltante em Aves.
b. Faça uma análise de agrupamento com as variáveis numéricas. Compare vários métodos hierárquicos, combinando com os tipos de distâncias. Compare também com o método k-médias.

6. Faça uma análise de agrupamento para os dados USArrests. Compare vários métodos hierárquicos, combinando com os tipos de distâncias. Compare também com o método k-médias.