6ª Lista de exercícios utilizando o R

Neste exercício, um banco de dados público (obtido do site Machine Learning Repository (Dua e Graff, 2019) e disponível em https://archive.ics.uci.edu/ml/datasets.php), denominado ***Adult*** (**adult.data** e **adult.names**), inclui dados extraídos do Census dos EUA em 1994 para prever se a renda individual excede U$ 50.000/ano, com base em 15 variáveis preditoras. No total, 32.561 observações estão disponíveis. Mais detalhes sobre o banco de dados são encontrados em Kohavi (1996) e em outras publicações.

a) Para a base de dados **Adult**, faça uma análise exploratória procurando identificar o impacto de cada potencial variável preditora na ocorrência (ou não) da renda individual excedendo U$ 50.000/ano.

b) Faça um ajuste de um modelo de regressão logístico realizando, quando pertinente a seleção de variáveis que tenham significância estatística para a previsão da resposta.

c) Faça uma análise da curva ROC do modelo.

d) Utilizando o procedimento de validação cruzada, avalie a nova estimativa da curva ROC.

e) Repita os itens (c) e (d) utilizando o método de Random Forests (Florestas Aleatórias)

Inclua em suas análises comentários sobre o ajuste dos modelos e suas respectivas capacidades preditivas. Faça um relatório sucinto e apresente discussões dos resultados obtidos.

Elabore o seu relatório utilizando o **Rmarkdown** e envie o documento em formato **PDF** para avaliação pelo sistema minha.ufmg. O tamanho máximo do relatório é de 5 páginas.