

Técnicas de Aprendizagem de Máquinas como Árvores de Decisão e Florestas Randômicas têm atingido excelentes resultados na predição/classificação diagnóstica de várias doenças. Um banco de dados foi coletado na África do Sul sobre doenças cardíacas. São 462 amostras com 10 variáveis. A variável **chd** (*coronary heart disease, yes=1, no=0*) é a classificação **y**.

O projeto visa aplicar algoritmos de árvores de decisão e Florestas Randômicas para predição de diabetes com base nesses dados. Sua solução deverá incluir:

1. Faça uma análise estatística inicial dos dados, plotando as quantidades médias, desvios padrões de todas as variáveis dos dados; (1,0 ponto)
2. Construa um modelo de árvore de decisão (ID3, C4.5 ou CART), separando aleatoriamente sempre 10% dos dados para teste, em validação cruzada (com 10 rodadas), e mostre o resultado final em termos de: curva ROC, curva AUC ROC, e matriz de confusão. (2,0 pontos)
3. Construa um modelo de “floresta randômica”, com 100 árvores, usando todas as variáveis preditoras (i.e.  $m=9$ ), separando aleatoriamente sempre 10% dos dados para teste, em validação cruzada (com 10 rodadas), e mostre o resultado final em termos de: curva ROC, curva AUC ROC, e matriz de confusão. (2,0 pontos)
4. Construa um modelo de “floresta randômica”, com 100 árvores, usando a raiz quadrada das variáveis preditoras (i.e.  $m=3$ ), separando aleatoriamente sempre 10% dos dados para teste, em validação cruzada (com 10 rodadas), e mostre o resultado final em termos de: curva ROC, curva AUC ROC, e matriz de confusão. (2,0 pontos)
5. Mostre, para o caso do melhor resultado, quais as 2 mais importantes/relevantes variáveis preditoras. (1,0 ponto)
6. Gere, ou nos comentários do código, ou em um texto à parte as saídas e explicações pedidas no projeto. (2,0 pontos)

O código deve ser bem documentado, escrito em Python, por um (1) estudante individualmente do curso, e entregue somente via sistema <http://aprender3.unb.br> do curso, no prazo estipulado. **O estudante deve indicar no código se, e de onde, estão usando fontes públicas de outros, e realizar suas próprias alterações para entendimento. Códigos iguais, ou tendo indicativo de plágios, ou feitos por outros, poderão receber nota zero.**