Universidade de Brasília
Departamento de Ciência da Computação
Projeto 1, Fundamentos de Sistemas Inteligentes, Turma A, 2022/1
Prof. Díbio

Técnicas de Aprendizagem de Máquinas como Árvores de Decisão e Florestas Randômicas têm atingido excelentes resultados na predição/classificação diagnóstica de várias doenças.

O projeto visa aplicar algoritmos de árvores de decisão e Florestas Randômicas para predição de hepatite nesse banco de dados (ps. Já separado em treinamento e teste):

https://github.com/zahangirbd/medical_data_for_classification/tree/master/data/Hepatitis

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=19), 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 variaúeis preditoras (i.e. m=4), 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.**