UNIVERSIDADE FEDERAL DA PARAÍBA - UFPB
CENTRO DE INFORMÁTICA - CI

PROVA 02 - INTRODUÇÃO A INTELIGÊNCIA ARTIFICIAL

Utilize validação cruzada estratificada 10-fold e fixe o random_state para responder. Os código devem ser escritos em um notebook, enviando o link ou o código por e-mail.

1. Utilizando a base de dados de https://www.kaggle.com/datasets/csafrit2/higher-education-students-performance-evaluation, elabore uma solução para identificar o OUTPUT Grade do estudante. Lembre-se de comentar seu código no notebook detalhadamente, explicando cada passo.
    A. Faça o pré-processamento dos dados (limpeza, engenharia de variáveis, etc) e deixe os seus dados preparados para aplicar o modelo.
        a. **OBS: Utilize pelo menos uma forma de redimensionamento de atributos (selecionando ou agregando) e avalie o resultado de utilizar todas eles e essa amostra.**
    B. Faça uma análise exploratória dos dados de saída, utilizando box plot, mostrando a aplicação de técnicas de under ou oversampling para que as classes tenham o mesmo tamanho.
    C. Utilize a biblioteca AUTOML para fazer a previsão. Para o melhor algoritmo, teste 3 variações de um dos seus hiperparâmetros.
    D. Para avaliar os resultados, utilize e explique a matriz de confusão. Além disso, escolha 2 métricas de sua preferência e o que o seu resultado significa.
2. Utilize a mesma base de dados (lembrem de tirar o rótulo, obviamente) da questão anterior de forma que:
    A. Execute o K-means e Hierárquico.
    B. Teste o K igual à 5 e 7.
    C. Na execução do Hierárquico, varie 2 métodos do linkage; **OBS.: utilize os mesmos valores de clusters escolhidos na questão anterior.**
    D. Por fim, faça uma comparação entre os 2 resultados das execuções anteriores e adote uma medida de avaliação própria para clusterização.

FASE BÔNUS: Diga vantagens e desvantagens do uso do AUTOML e como você faria para usá-lo, garantindo os melhores resultados possíveis.