

**Residência Tecnológica - Machine Learning - Aula Assíncrona**

**Questão 1)** Nesta atividade, utilize o *k-Nearest Neighbor* para verificar o desempenho do algoritmo de acordo com a engenharia de *features* empregada. Para tal fim, siga os seguintes passos:

1. Utilize o *dataset ionosphere* presente na pasta *datasets* para realizar uma tarefa de **classificação binária**.
2. Separe os dados de entrada e saída considerando que a coluna de índice 34 corresponde ao *label* dos dados de cada linha.
3. Separe os dados em dados de treinamento e teste utilizando a proporção 70% e 30%, respectivamente, e *random state* = 0.
4. Estabeleça arbitrariamente o número de vizinhos  $k = 5$ , inicialmente.
5. Realize a classificação e obtenha os seguintes resultados:
  - (a) Acurácia.
  - (b) Precisão.
  - (c) *Recall*.
  - (d) F1-Score.
  - (e) Plote a matriz de confusão obtida.
6. Repita todo o processo dos itens 1 até 5 utilizando o método de Seleção de *Features* por limiar de variância, considerando o limiar de 35%.
7. Repita todo o processo itens 1 até 5 utilizando o método de Extração de *Features* por Análise de Componentes Principais (PCA), considerando a porcentagem de informação a ser mantida no problema de 85%.
8. Considerando os dados transformados através do uso da PCA, repita todo o processo dos itens 1 a 5, mas aplique a técnica de Validação Cruzada por ***k-fold*** para encontrar o número de vizinhos mais próximos ideal para o problema. Considere variar  $k$  de 1 até 20, e considere o número de pastas do *k-fold* (o parâmetro *cv* da função *cross\_val\_score*) igual a 5.

**Questão 2)** Compare os resultados obtidos em todos os processos do item anterior e justifique as diferenças obtidas através dos conceitos de engenharia de *features* estudados durante as aulas síncronas.