Análise de Dados e Projeto de um Classificador

Objetivo Principal: Desenvolver classificadores baseados em distância e o k-NN para o conjunto de dados "*Dry Bean*" da UCI. https://archive.ics.uci.edu/dataset/602/dry+bean+dataset

Inicialmente deve-se analisar os dados, realizar o pré-processamento, para posteriormente projetar e analisar os classificadores.

Pré-processamento:

A. Trabalhar com o método do *holdout*, deixando 30% dos dados para o conjunto de teste.
B. Normalizar os dados, lembrando que a normalização é aprendida no conjunto de treino e aplicado aos conjuntos de treinamento e teste.
C. Esse conjunto de dados não possui dados faltantes ou dados inconsistentes, mas possui desbalanceamento entre as classes. Utilizar o método SMOTE ("*Synthetic Minority Oversampling TEchnique*") para balancear as classes apenas no conjunto de treinamento.
D. Transformar as variáveis categóricas da saída em valores numéricos (pode usar *Preprocessing.LabelEncoder do sklearn*) para poder utilizar informação mútua e correlação para seleção de parâmetros.
E. Obter a matriz de correlação das entradas e da saída (plotar) e verificar quais os parâmetros de entrada mais correlacionadas e quais os 5 parâmetros de entrada que possuem maior correlação com a saída.
F. Obter a informação mútua entre os parâmetros de entrada e saída e identificar os 5 parâmetros com maior valor.
G. Obter a razão discriminante de Fisher para os parâmetros de entrada e identificar os 5 parâmetros com maior valor.
H. Aplicar a transformação PCA aos parâmetros de entrada e utilizá-la para selecionar os 5 parâmetros de maior energia.

Projeto dos classificadores:

1. Projetar um classificar baseado na distância de Mahalanobis, utilizando o conjunto de treinamento, utilizando os 16 parâmetros de entrada (C). Avaliar o desempenho no conjunto de teste, através da acurácia e da matriz de confusão.
2. Projetar um classificar baseado na distância de Mahalanobis, utilizando o conjunto de treinamento, utilizando os 5 parâmetros de maior correlação com a saída (E). Avaliar o desempenho no conjunto de teste, através da acurácia e da matriz de confusão.
3. Projetar um classificar baseado na distância de Mahalanobis, utilizando o conjunto de treinamento, utilizando os 5 parâmetros de maior informação mútua com a saída (F). Avaliar o desempenho no conjunto de teste, através da acurácia e da matriz de confusão.
4. Projetar um classificar baseado na distância de Mahalanobis, utilizando o conjunto de treinamento, utilizando os 5 parâmetros de maior discriminante de Fisher (G). Avaliar o desempenho no conjunto de teste, através da acurácia e da matriz de confusão.
5. Projetar um classificar baseado na distância de Mahalanobis, utilizando o conjunto de treinamento, utilizando os 5 parâmetros da PCA (H). Avaliar o desempenho no conjunto de teste, através da acurácia e da matriz de confusão.
6. Repetir os itens de 1 a 5 para o classificador k-NN, utilizando k=10.
7. Repetir os itens de 1 a 5 para o classificador linear (*LinearDiscriminantAnalysis do sklearn*).
8. Comparar os resultados obtidos pelos diferentes classificadores e indicar qual seria o melhor projeto.