Avaliação Somativa 1

Aluno: _____

Suponha que você foi contrato por uma empresa que precisa de um modelo preditivo para um problema de classificação de tipos de vidros, problema denominado **Glass**.

A empresa preparou uma base de dados contendo 214 instâncias cada uma com 11 atributos numéricos. O primeiro atributo deve ser ignorado pois é um ID# da instância, o qual é seguido por 9 atributos de entrada numéricos e a classe.

Cada atributo de entrada diz respeito à composição do vidro (índice de refração e as quantidades de sódio, magnésio, alumínio, sílica, potássio, cálcio, bário e ferro).

Há 7 classes de vidros, sendo: 1,2,3,4,5,6,7. Contudo, não há exemplos da classe 4 na base. Isto não impede a construção de classificadores para o problema.

Apenas para seu conhecimento, segue o nome de cada classe:

1 building_windows_float_processed
2 building_windows_non_float_processed
3 vehicle_windows_float_processed
4 vehicle_windows_non_float_processed (sem exemplos na base)
5 containers
6 tableware
7 headlamps

Leitura da base:

```
# Ler direto da URL
import urllib
import urllib.request as request
import numpy as np

url = "https://archive.ics.uci.edu/ml/machine-learning-databases/glass/glass.data"
raw_data = urllib.request.urlopen(url)

# Carrega arquivo como uma matriz
dataset = np.loadtxt(raw_data, delimiter=",")

# Separa atributos de entrada em X e as classes em y
# Já ignora o ID da instância
X = dataset[:,1:10]
y = dataset[:,10]
```

A) CONSTRUÇÃO CLASSIFICADOR: Encontrar a melhor solução para este problema através da avaliação de soluções monolíticas (uso de um único classificador). Para tal, avalie as três técnicas estudas em sala (KNN, Naive Bayes e Árvores de Decisão). Anote na tabela abaixo o melhor resultado encontrado para cada uma em termos de taxa de acerto e f1_score. Utilize validação cruzada considerando 5 folds.

Tabela de Resultados (Validação cruzada = 5 folds)

Classificador	Taxa de Acerto (%)	F1_score
KNN		
Naive Bayes		
Árvores de Decisão		

B) CONSIDERANDO O SEU MELHOR RESULTADO, APRESENTE:

B.1) Os parâmetros utilizados para que o professor possa reproduzir seus resultados.

B.2) A matriz de confusão.

B.3) A taxa de acerto da classe com mais erros?

B.4) Para este problema você recomenda o uso de taxa de acerto ou f1_score como métrica mais adequada. Justifique a sua resposta.

C) DESAFIO

Altere o script desenvolvido em sala para considerar uma outra técnica também disponível no SKLEARN denominada LDA (Linear Discriminant Analysis). Execute o LDA para o problema dado e anote abaixo os resultados e parâmetros.

Tabela de Resultados (Validação cruzada = 5 folds)

Classificador	Taxa de Acerto (%)	F1_score
LDA		

Parâmetros utilizados no LDA:

_____