Universidade Federal da Paraíba
Centro de Informática

Prova 02 | Paradigmas de Aprendizagem de Máquina
Professoras Daniela Coelho e Thaís Gaudencio
22 de abril de 2025

Instruções

- Utilize validação cruzada estratificada 10-fold e fixe o `random_state` para garantir reprodutibilidade nos resultados.
- O código deve ser disponibilizado no Google Colab ou GitHub e conter a documentação adequada para sua execução.
- Justifique suas escolhas e interprete os resultados obtidos em cada etapa da análise.

Contextualização sobre Features Radiômicas

Os dados utilizados nesta prova foram extraídos por meio da biblioteca `PyRadiomics`, que permite a obtenção de features radiômicas a partir de imagens médicas. Essas features fornecem informações quantitativas sobre a textura, forma e intensidade das imagens. As features radiômicas têm sido aplicadas com sucesso em problemas de aprendizado de máquina para a detecção e classificação de lesões.

Algumas categorias principais dessas features incluem:

- **First-order statistics:** descrevem a distribuição dos valores de pixel (ex.: média, desvio padrão, entropia).
- **Shape-based features:** caracterizam a forma e o tamanho da região segmentada (ex.: esfericidade, volume).
- **Texture features:** analisam padrões texturais da imagem (ex.: Matriz de Coocorrência de Níveis de Cinza - GLCM, Matriz de Dependência de Níveis de Cinza - GLDM).

QUESTÕES

Questão 01

A base de dados está disponível em: Radiomic Data.

O repositório no *GitHub* acima disponibiliza o arquivo *radiomic_data_binary.csv*, o qual contém features radiômicas extraídos de regiões de interesse em mamografias, utilizando a biblioteca *PyRadiomics*. A base contempla duas classes: *BENIGN* e *MALIGNANT*, totalizando 2018 amostras e 116 características. A partir dessa base, realize as seguintes tarefas:

a) Efetue o pré-processamento da base de dados, considerando:
- Tratamento de valores ausentes (se houver);
- Redução de dimensionalidade via seleção de atributos ou técnicas como PCA;
- Avalie o impacto da redução de dimensionalidade nos modelos.
  - **Obs.: Você tem total liberdade para explorar diferentes abordagens de pré-processamento, experimentar novas técnicas e analisar seus impactos nos resultados finais.**

b) Treine e compare o desempenho dos algoritmos:
- Árvore de Decisão
- MLP (Perceptron Multicamadas)
  - **Obs.: Caso prefira, você pode utilizar outro algoritmo, desde que justifique sua escolha e explique como ele se compara aos modelos propostos.**

c) Avalie os modelos utilizando F1-score, matriz de confusão e curva ROC/AUC. Compare os resultados obtidos e discuta qual algoritmo apresentou o melhor desempenho e por quê.

Questão 02

Com base nos dados fornecidos anteriormente (removendo os rótulos das classes), realize uma análise de agrupamento (*clustering*) para identificar padrões sem supervisão:

a) Execute os algoritmos de **K-means** e **Hierárquico**.
b) Determine o valor ideal de K para o K-means utilizando o **método do cotovelo**.
c) Com o valor de K obtido, compare os resultados do K-means com os do Hierárquico.
d) No algoritmo Hierárquico, teste pelo menos **dois** métodos diferentes de *linkage*.
e) Compare os resultados das técnicas de *clustering* utilizando métricas adequadas, justificando suas escolhas.