INSTITUTO FEDERAL
ESPÍRITO SANTO

Bacharelado em Sistemas de Informação

Trabalho sobre Mineração de Dados

Disciplina: Inteligência de Negócios - 2022/1 - Profa Kelly

Classificação de Lesões no Tecido Mamário

1.0 Introdução

De acordo com o site do Instituo Nacional do Cancer [1]:

> "o câncer de mama é uma doença causada pela multiplicação desordenada de células anormais da mama, que forma um tumor com potencial de invadir outros órgãos. Há vários tipos de câncer de mama. Alguns têm desenvolvimento rápido, enquanto outros crescem lentamente. A maioria dos casos, quando tratados adequadamente e em tempo oportuno, apresentam bom prognóstico".

A espectroscopia de impedância elétrica é a principal ferramenta para triagem de classificação de lesões no tecido mamário, pois permite a identificação de um tumor antes de que o mesmo seja palpável [2]. Em um estudo feito por [3], verificou-se que, das lesões previamente diagnosticadas como suspeitas e encaminhadas para biópsia, aproximadamente, 25% eram de fato lesões malignas, ou seja, aproximadamente, 75% das lesões foram diagnosticadas como benignas. Essa alta taxa de falsos positivos evidencia a grande dificuldade de se obter um diagnóstico preciso [4]. Nesse contexto, a análise computadorizada de imagens se apresenta como uma importante ferramenta para melhorar o diagnóstico inicial.

O conjunto de dados utilizado neste trabalho, disponível em [5], foi obtido a partir de operações de espectroscopia de impedância elétrica e apresenta 10 atributos. Sendo o atributo alvo a classificação do tecido, podendo ser: car(carcicoma), fad (fibro-adenoma), mas (mastopathy), gla (glandular), con (connective), adi (adipose).

2.0 Objetivo do Trabalho

O objetivo do trabalho é realizar um estudo comparativo das técnicas de classificação KNN e SVM (*Support Vector Machine*) na classificação de lesões no tecido mamário, utilizando métricas e técnicas de validação de modelos.

Para esse fim, os seguintes passos devem ser adotados.

Para avaliação dos modelos de classificação, deve-se utilizar a técnica de validação cruzada k-fold, com k=5. Note que os mesmos subconjuntos de dados deverão ser utilizados em todos os cenários de implementados.

I - Implementação dos cenários:

KNN:

- Para cada divisão treino/teste gerada:
  - Realizar o treinamento do algoritmo.
  - Executar os testes
- criar a matriz de confusão com os resultados acumulados (soma de todas as matrizes de confusão geradas para cada divisão treino/teste)
- calcular a média das acurácias obtidas nos testes executados no passo anterior.

SVM:

- Kernel Linear
  - Para cada divisão de treino/teste gerada:
    - Treinar o modelo utilizando o kernel linear
    - Executar os testes.
  - criar a matriz de confusão com os resultados acumulados (soma de todas as matrizes de confusão geradas a cada divisão treino/teste)
  - calcular a média das acurácias obtidas nos testes executados no passo anterior.

II - Elaboração de um artigo científico apresentando os resultados obtidos.

3.0 Implementação

O código deve ser implementado em Python e será executado no Google Colab.

O programa gerado não deverá solicitar nenhuma entrada ao usuário. Vale dizer que os códigos que não executarem serão desconsiderados.

4.0 Artigo

O artigo deverá conter as seguintes seções:

- Resumo: onde o conteúdo do artigo deve ser descrito de forma sucinta.
- Introdução: contextualizando o trabalho.

- Referencial teórico: onde deve ser apresentada a teoria acerca das técnicas de classificação KNN e SVM.
- Metodologia: onde devem ser descritos os testes realizados e a base utilizada (não precisa descrever cada atributo). Além disso, deve ser abordada a teoria sobre matriz de confusão, validação cruzada e métricas (acurácia, precisão e revocação)
- Resultados: onde serão expostos os resultados alcançados.
- Conclusão: que deve estar relacionada ao objetivo aqui proposto para o trabalho.

O documento deverá ser produzido seguindo o modelo do IEEE transaction, a ser disponibilizado no AVA, ser entregue no formato PDF, e ter no **máximo 6 páginas** (incluindo imagens, tabelas e referências).

Observações:

1. Trabalhos que excederem o número de páginas serão penalizados.
2. Consulte as fontes, referencie-as, mas escreva o texto com suas próprias palavras. Caso seja detectado o plágio, o trabalho será avaliado com nota zero.
3. Documentos desacompanhados do código fonte serão avaliados com nota zero.

5.0 Produção e entrega

- O artigo e os códigos produzidos devem ser entregues em um arquivo compactado por meio do AVA até as 23h do dia 08/07/22.
- Trabalho em grupo de até 3 pessoas
- Nota máxima: 40 pontos
- É possível que os autores sejam convocados para responder a questionamentos relativos ao trabalho desenvolvido.
- Caso seja detectada algum tipo de fraude, o trabalho será avaliado com nota zero.
- Envios sem a codificação ou sem o artigo serão avaliados com nota zero.
- Além do conteúdo técnico, será avaliada a qualidade do texto produzido, que deve ser claro, objetivo e ter as informações apresentadas de forma organizada.

6.0 Referências:

[1] Site do Instituto Nacional do Câncer: https://www.inca.gov.br/tipos-de-cancer/cancer-de-mama. Acesso em 01/06/22.

[2] Abdulkader Helwan, John Bush Idoko, Rahib H. Abiyev. Machine learning techniques for classification of breast tissue. Procedia Computer Science, Volume 120, 2017, Pages 402-410.

[3] P Vacek, B Geller, D Weaver, R Foster. Increased mammography use and its impact on earlier breast cancer detection in vermont Cancer;, 94 (2002), pp. 2160-2168.

[4] Basset L, and Gold R., 1987. Breast Cancer Detection: Mammograms and Other Methods in Breast Imaging. Grune & Stratton, New York.

[5] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository Irvine, CA: University of California, School of Information and Computer Science. Disponível em https://archive.ics.uci.edu/ml/datasets/Breast+Tissue. Acesso em 01/06/22