Classificação de Tumores de Mama usando Aprendizado de Máquina

Daniele Carnaúba, Felipe Barros, João Samuel e Leyberson Assunção

29 de Agosto de 2024

1 Introdução

O câncer de mama é um dos tipos de câncer mais comuns entre as mulheres em todo o mundo. A detecção precoce e precisa é crucial para um tratamento eficaz e para melhorar as taxas de sobrevivência. Este estudo visa classificar tumores mamários como malignos ou benignos usando um conjunto de dados contendo várias características extraídas de imagens digitalizadas de aspirados por agulha fina de massas mamárias. O objetivo é desenvolver e avaliar modelos de aprendizado de máquina que possam diferenciar efetivamente tumores malignos e benignos.

2 Descrição do Conjunto de Dados

O conjunto de dados a ser utilizado neste estudo é um conjunto de dados de câncer de mama disponível no Kaggle. Ele contém 569 instâncias, cada uma rotulada como maligna (M) ou benigna (B). O conjunto de dados inclui 30 características numéricas que descrevem características dos núcleos celulares presentes nas imagens, como raio, textura, perímetro, área, suavidade, compacidade, concavidade e simetria.

2.1 Rótulos e Variáveis

O rótulo alvo para este conjunto de dados é o *diagnóstico*:

- **M**: Maligno (tumor canceroso)
- **B**: Benigno (tumor não canceroso)

As variáveis independentes incluem várias medições numéricas, agrupadas da seguinte forma:

- **Características Médias:** radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave points_mean, symmetry_mean, fractal_dimension_mean.
- **Piores Características (Valores Máximos):** radius_worst, texture_worst, perimeter_worst, area_worst, smoothness_worst, compactness_worst, concavity_worst, concave points_worst, symmetry_worst, fractal_dimension_worst.
- **Erro Padrão das Características:** radius_se, texture_se, perimeter_se, area_se, smoothness_se, compactness_se, concavity_se, concave points_se, symmetry_se, fractal_dimension_se.

3 Questão de Pesquisa

A principal questão deste estudo é: **"Podemos classificar com precisão os tumores mamários como malignos ou benignos com base nas características extraídas de imagens digitalizadas de aspirados por agulha fina?"**

4 Metodologia

4.1 Preparação e Exploração de Dados

- **Inspecionar os Dados:** Identificar e lidar com quaisquer valores ausentes, *outliers* ou inconsistências.
- **Visualizar os Dados:** Usar gráficos e medidas estatísticas para entender a distribuição e as relações entre as características.
- **Normalizar/Padronizar Características:** Padronizar as escalas das características para garantir uma comparação justa e desempenho do modelo.

4.2 Seleção e Treinamento de Modelos

Vários modelos de aprendizado de máquina são considerados para classificar os tumores, incluindo:

- **Regressão Logística**
- **Árvores de Decisão**
- **Random Forest**
- **Máquinas de Vetores de Suporte (SVM)**
- **k-Nearest Neighbors (k-NN)**
- **Naive Bayes**
- **Redes Neurais Artificiais (ANN)**
- **Máquinas de Gradiente Boosting (GBM) / XGBoost**
- **LightGBM ou CatBoost**
- **Métodos de Conjunto (e.g., Stacking, Bagging)**

4.3 Avaliação do Modelo

- Utilizar métricas como acurácia, precisão, revocação, F1-score, AUROC e AUCPR.
- Utilizar matrizes de confusão para analisar falsos positivos e negativos.

4.4 Tratamento de Dados Ausentes

- **Remover Colunas Irrelevantes:** Excluir colunas sem informações úteis (por exemplo, *Unnamed: 32*).
- **Imputação de Valores Ausentes:** Usar métodos como média, mediana e moda.
- **Avaliar Impacto:** Comparar o desempenho do modelo antes e depois da imputação.

5 Conclusão

Ao aplicar vários modelos de aprendizado de máquina e lidar com dados ausentes de forma apropriada, esperamos desenvolver um modelo robusto e preciso para a classificação de tumores mamários. O sucesso do modelo será avaliado com base em sua capacidade de generalizar bem para novos dados e em seu desempenho em várias métricas, apoiando, assim, a tomada de decisões clínicas no diagnóstico do câncer de mama.