Programa de Pós-Graduação em Engenharia de Computação e Sistemas - PECS/UEMA

Professor: Omar Andres Carmona Cortes
Aluno: Adrielson Ferreira Justino
Atividade I: Implementação de MLP

Descrição: Envie um relatório em PDF contendo a descrição da base de dados, as arquiteturas criadas e a comparação entre MLP e dois outros algoritmos de AM.

1. DESCRIÇÃO DA BASE DE DADOS

Para esta atividade foi escolhida a base de dados **Breast Cancer Wisconsin (Diagnostic)**[1] que consiste em um conjunto de dados utilizado em tarefas de diagnóstico de câncer de mama. Ela contém informações sobre características extraídas de imagens de células tumorais obtidas de biópsias de tecido mamário, sendo ideal para tarefas de classificação binária, no treinamento de modelos de aprendizado de máquina para prever se um tumor é maligno ou benigno com base nas características descritas.

As características da base foram computadas a partir de imagens de aspiração com agulha fina (FNA) de diferentes massas de mama. A base contém um total de 569 amostras e 31 atributos, conforme descrito a seguir:

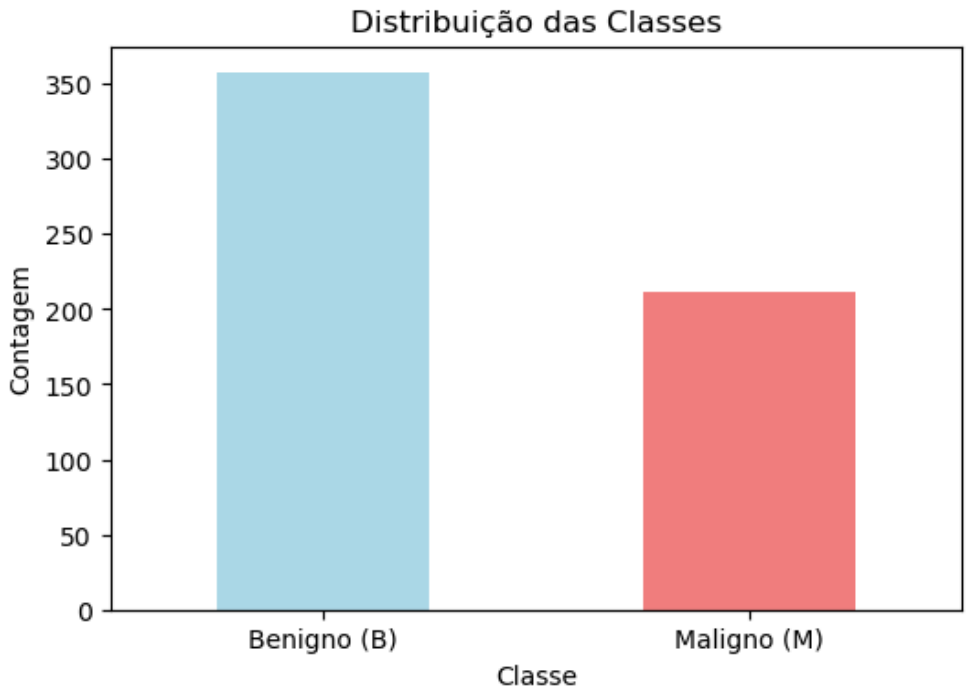
- ID number: Cada entrada no conjunto de dados é identificada por um número de ID único.
- Diagnosis (M = maligno, B = benigno): Esta é a variável de destino que indica se o tumor é maligno (câncer de mama) ou benigno (não canceroso).
- Dez características de valor real são calculadas para cada núcleo celular:
    - *radius* (média das distâncias do centro aos pontos na circunferência)
    - *texture* (desvio padrão dos valores em escala de cinza)
    - *perimeter*
    - *area*
    - *smoothness* (variação local nos comprimentos dos raios)
    - *compactness* (perímetro^2 / área - 1.0)
    - *concavity* (gravidade das porções côncavas do contorno)
    - *concave points* (número de porções côncavas do contorno)
    - *symmetry*
    - *fractal dimension* (aproximação à "linha costeira" - 1)

[1] https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

- Para cada imagem, foram calculados os valores médios, erros padrão e "pior" (maior valor das três maiores) dessas características, resultando em um total de 30 características. Por exemplo, o campo 3 é o Raio Médio, o campo 13 é o Erro Padrão do Raio e o campo 23 é o Pior Raio.
- Não há valores ausentes neste conjunto de dados.

Na distribuição das classes no conjunto de dados estão dispostas 357 amostras são benignas (B) e 212 amostras são malignas (M), conforme a Figura 1.

Figura 1. Distribuição das Classes



2. PRÉ-PROCESSAMENTO

Inicialmente, os dados foram preparados para o treinamento do modelo. Para isso foram removidas duas colunas que não são necessárias para a análise: a coluna "id" e a coluna "Unnamed: 32". Em seguida, foi realizada uma codificação dos rótulos das classes. Ele mapeia a classe "M" (Maligno) para o valor 1 e a classe "B" (Benigno) para o valor 0, tornando os rótulos binários. As características foram selecionadas utilizando a indexação de colunas do DataFrame. As características (atributos preditivos) foram armazenadas em uma variável X, enquanto os rótulos (diagnósticos malignos ou benignos) foram armazenados em uma variável y. As características foram selecionadas a partir da coluna 1 até o final do DataFrame, excluindo a primeira coluna que continha o diagnóstico. Finalmente, os dados são divididos em conjuntos de treinamento (X_train, y_train) e teste (X_test, y_test), 30% dos dados foram reservados para o conjunto de teste, enquanto 70% foram usados para o conjunto de treinamento.

3. ARQUITETURA DOS MODELOS

Inicialmente, os dados foram preparados para o treinamento dos modelos. As características foram selecionadas utilizando a indexação de colunas do DataFrame. As características (atributos preditivos) foram armazenadas em uma variável `X`, enquanto os rótulos (diagnósticos malignos ou benignos) foram armazenados em uma variável `y`. As características foram selecionadas a partir da segunda coluna até o final do DataFrame, excluindo a primeira coluna que continha o diagnóstico. Em sequência, 30% dos dados foram reservados para o conjunto de teste, enquanto 70% foram utilizados para o conjunto de treinamento.

O principal objetivo desta atividade foi identificar algoritmos eficazes para a detecção do câncer de mama. Para isso, utilizamos três modelos de aprendizado de máquina: *Decision Tree* (Árvore de Decisão), *Random Forest* e uma Rede Neural *Multilayer Perceptron* (MLP).

3.1 Árvore de Decisão

A Árvore de Decisão é um modelo simples, mas poderoso, que divide os dados em base nas características mais relevantes para a classificação. O modelo foi configurado com o parâmetro `random_state=42` para garantir a reprodutibilidade dos resultados. O treinamento foi realizado utilizando o conjunto de dados de treinamento, e o modelo foi ajustado para maximizar a acurácia.

3.2 *Random Forest*

O modelo de *Random Forest* é uma técnica de ensemble que combina múltiplas árvores de decisão para melhorar a precisão e reduzir o overfitting. Para este experimento, o modelo foi configurado com 100 estimadores (`n_estimators=100`) e `random_state=42`, também para garantir a reprodutibilidade. O *Random Forest* é capaz de capturar interações mais complexas entre as características e é conhecido por sua robustez em problemas de classificação.

3.3 MLP

A Rede Neural MLP foi configurada com duas camadas ocultas. A primeira camada oculta possui 64 neurônios e a segunda possui 32 neurônios, ambas utilizando a função de ativação ReLU. A camada de saída foi configurada para classificação binária com a função de ativação sigmoid. O modelo foi compilado com o otimizador Adam e a função de perda binary_crossentropy. Para evitar overfitting, foi utilizado o método de early stopping, que interrompe o treinamento quando o erro de validação não melhora por um determinado número de épocas. O valor de random_state foi definido para 42 para garantir a reprodutibilidade dos resultados. O MLP é

adequado para capturar relações não lineares entre as características, tornando-o útil para problemas de classificação complexos.

4. COMPARAÇÃO ENTRE MLP, ÁRVORE DE DECISÃO e *RANDON FOREST*

As métricas utilizadas para fins de comparação entre os modelos foram acurácia, precisão, recall e F1-score. A acurácia mede a proporção de previsões corretas feitas pelo modelo em relação ao total de previsões, indicando a taxa geral de acerto do modelo. A precisão mede a proporção de verdadeiros positivos (tumores malignos corretamente identificados) em relação ao total de casos positivos previstos pelo modelo (tumores malignos previstos), sendo útil para avaliar a capacidade do modelo de evitar falsos positivos. O recall mede a proporção de verdadeiros positivos em relação ao total de casos positivos reais (tumores malignos reais), avaliando a capacidade do modelo de identificar todos os casos positivos. Por fim, o F1-Score combina precisão e recall em uma única métrica, sendo a média harmônica entre as duas, útil quando se deseja equilibrar ambas as métricas, especialmente em situações com classes desbalanceadas.

Tabela 1. Pontuações por modelo

Conforme a Tabela 1, o modelo MLP apresentou uma acurácia de 87,13%, indicando que a maioria das previsões foi correta. A precisão foi de 90,20%, o que mostra que a maioria dos casos classificados como malignos realmente são malignos. O recall de 73,02% indica que o modelo teve um desempenho um pouco inferior na identificação de todos os casos malignos, deixando de detectar alguns casos. Entretanto, o F1-Score de 80,70% sugere que o modelo conseguiu manter um equilíbrio razoável entre precisão e recall, apesar de não ser o modelo com o melhor desempenho geral.

O modelo Árvore de Decisão obteve uma acurácia superior, de 92,98%, o que indica um bom desempenho geral na classificação. A precisão de 88,06% sugere que o modelo conseguiu evitar a maioria dos falsos positivos, enquanto o recall de 93,65% mostra que ele foi bastante eficaz em identificar a maioria dos casos malignos. Com um F1-Score de 90,77%, o Árvore de Decisão equilibrou bem a precisão e o recall, mostrando-se um modelo robusto e eficaz para a classificação de tumores de mama.

O modelo *Random Forest* foi o melhor entre os três, com uma acurácia de 96,49%, o que indica que ele acertou a grande maioria das previsões. A precisão extremamente alta, de 98,31%, mostra que quase todos os casos que o modelo classificou como malignos realmente são malignos, minimizando falsos positivos. O recall de 92,06% indica que o modelo conseguiu identificar a maioria dos casos

malignos, embora tenha perdido alguns poucos casos. No entanto, com um F1-Score de 95,08%, o *Random Forest* demonstra um excelente equilíbrio entre precisão e recall, sendo o modelo mais eficaz para a tarefa.

5. CONCLUSÃO

Entre os modelos comparados, o *Random Forest* apresentou o melhor desempenho, com a maior acurácia e F1-Score, destacando-se em termos de precisão e recall. A Árvore de Decisão também obteve bons resultados, com alta acurácia e um bom equilíbrio entre precisão e recall, sendo uma opção válida dependendo do cenário. O MLP teve o desempenho mais modesto, especialmente em termos de recall, indicando que ele deixou de detectar alguns casos malignos. No entanto, o F1-Score do MLP ainda mostra um equilíbrio razoável entre precisão e recall.

A escolha do modelo ideal pode depender de fatores como interpretabilidade e tempo de treinamento. Enquanto o *Random Forest* se destaca em termos de precisão, seu tempo de treinamento pode ser mais longo que o do Árvore de Decisão. A MLP, por outro lado, pode ser útil para capturar relações não lineares entre as características, embora não tenha superado os outros modelos neste experimento. Assim, é importante considerar essas variáveis ao escolher o modelo mais adequado para a tarefa de classificação de tumores de mama.