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:

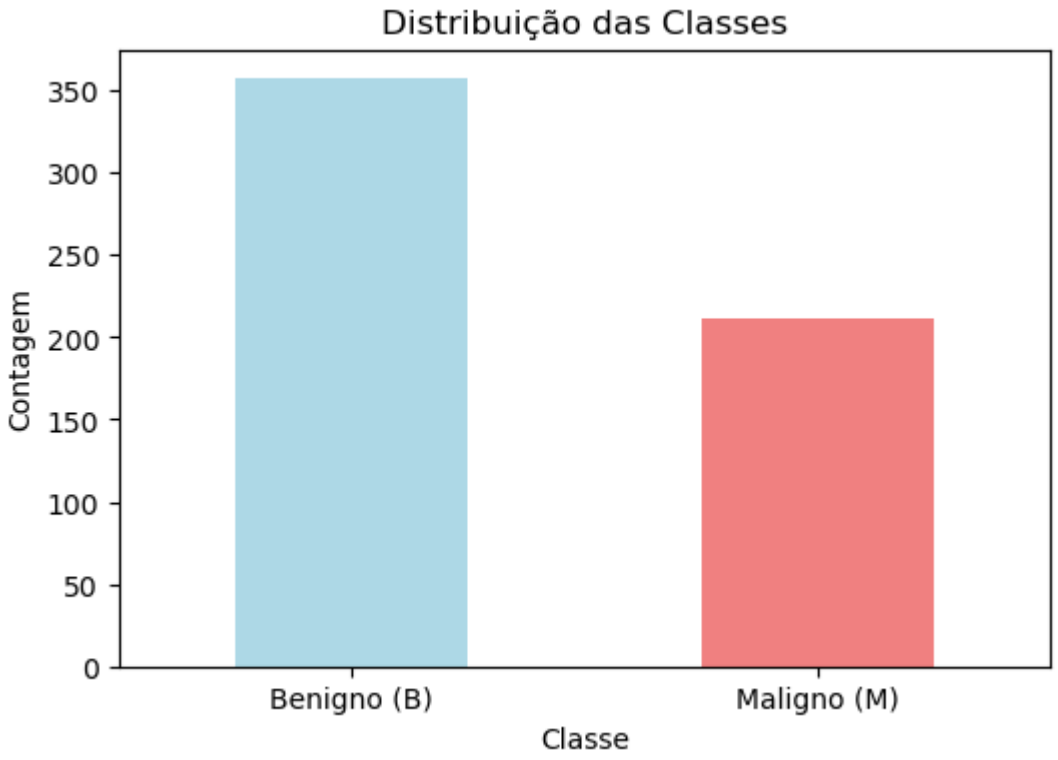
- ID number: Cada entrada no conjunto de dados é identificada por um número de ID único.
- Diagnosis (M = maligno, B = benigno): Esta é a variável de destino que indica se o tumor é maligno (câncer de mama) ou benigno (não canceroso).
- Dez características de valor real são calculadas para cada núcleo 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. ARQUITETURA DOS MODELOS

Inicialmente, os dados foram preparados para o treinamento do modelo. 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. Em sequência, 30% dos dados foram reservados para o conjunto de teste, enquanto 70% foram usados para o conjunto de treinamento.

O principal objetivo dessa atividade foi identificar um algoritmo eficaz e preditivo para a detecção do câncer de mama, portanto aplicamos classificadores de aprendizado de máquina *Support Vector Machine* (SVM), *k-Nearest Neighbors* (KNN) e uma Rede Neural *Multilayer Perceptron* (MLP).

O SVM é um algoritmo de aprendizado supervisionado que visa encontrar um hiperplano de separação ótimo entre as classes, maximizando a margem entre os pontos de dados mais próximos de diferentes classes. Nesta atividade o modelo foi configurado para considerar dois tipos de kernels: 'linear' e 'rbf'. Além disso, ajustamos o hiperparâmetro C, testando valores no intervalo [6, 7, 8, 9, 10, 11, 12].

O KNN é uma técnica simples e eficaz para classificação baseada na similaridade entre os dados. Este modelo considera a proximidade dos vizinhos mais próximos para fazer previsões. Para esta tarefa foi configurado com vários hiperparâmetros, incluindo *n_neighbors* em que foram testados vários valores, incluindo 5, 6, 7, 8, 9 e 10. Este hiperparâmetro define quantos vizinhos mais próximos serão considerados para determinar a classe de uma observação. Para o *leaf_size* foi testado diferentes valores, como 1, 2, 3 e 5. O *leaf_size* afeta a eficiência do algoritmo e a rapidez com que a pesquisa pelos vizinhos mais próximos é realizada. Quanto ao hiperparâmetro *weights* foram considerados dois tipos de pesos: *'uniform'* e *'distance'*. O peso *'uniform'* atribui a mesma importância a todos os vizinhos mais próximos, enquanto o peso *'distance'* atribui maior importância aos vizinhos mais próximos, ponderando-os de acordo com a distância. Por fim em *algorithm* foram testados diferentes algoritmos de pesquisa como *'auto', 'ball_tree', 'kd_tree'* e *'brute'*.

As configurações para MLP para resolver o problema de classificação consistiu em duas camadas ocultas com 64 e 32 neurônios, respectivamente, e utilizou a função de ativação *'relu'* nas camadas ocultas. O treinamento da MLP foi realizado com um máximo de 1000 iterações, e o valor de *random_state* foi fixado em 42 para garantir a reprodutibilidade dos resultados. O MLP é uma abordagem mais complexa que pode capturar relações não lineares entre os recursos, tornando-o adequado para problemas de classificação mais complexos.

3. COMPARAÇÃO ENTRE SVM, KNN E MLP

As métricas utilizadas para fins de comparação entre os modelos foram acurácia, precisão, *recall* e *f1-score*. No qual 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. Precisão que mede a proporção de verdadeiros positivos (tumores malignos corretamente identificados) em relação ao total de casos positivos previstos pelo modelo (tumores malignos previstos). É uma métrica que avalia 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). É uma métrica que avalia a capacidade do modelo de identificar todos os casos positivos. E por fim *F1-Score* que combina precisão e recall em uma única pontuação. Ele é calculado como a média harmônica entre precisão e recall e é útil quando se deseja equilibrar ambas as métricas. É especialmente útil quando as classes estão desbalanceadas.

A **Tabela 1** apresenta o resumo das pontuações por modelo.

Tabela 1. Pontuações por modelo

	SVM	KNN	MLP
Acurácia	0.959064	0.929825	0.935673
Precisão	0.963406	0.942584	0.943146
Recall	0.959064	0.929825	0.935673
F1-Score	0.959550	0.931437	0.936704

De acordo com a Tabela 1 o modelo SVM obteve uma alta acurácia de 95,91%, o que indica que a maioria das previsões está correta. Além disso, apresenta uma alta precisão de 96,34%, o que significa que a maioria dos casos classificados como malignos são realmente malignos. O recall é igual à acurácia, o que indica que o modelo identificou a maioria dos casos malignos. O F1 Score é alto, indicando um bom equilíbrio entre precisão e recall.

O modelo KNN também obteve um bom desempenho com uma acurácia de 92,98%. A precisão é um pouco menor que a do SVM, mas ainda é alta, indicando que o modelo minimiza falsos positivos. O recall é igual à acurácia, o que significa que o modelo identificou a maioria dos casos malignos. O F1 Score é ligeiramente inferior ao do SVM.

O modelo MLP apresenta uma acurácia de 93,57%, que é comparável à do KNN. A precisão e o recall também são semelhantes aos do KNN. O F1 Score é ligeiramente superior ao do KNN, indicando um melhor equilíbrio entre precisão e recall.

4. CONCLUSÃO

Os modelos SVM, KNN e MLP apresentam resultados bastante similares em termos de acurácia, precisão e recall na tarefa de classificação de tumores de mama. No entanto, a escolha do modelo ideal pode depender de outros fatores, como a interpretabilidade do modelo e o tempo de treinamento. É importante avaliar esses fatores em conjunto com as métricas de avaliação ao escolher o modelo mais adequado para a tarefa.