Universidade Federal de Uberlândia

Relatório Comparativo de Classificadores para Dados de Coluna Vertebral

Waldemar Patrique Flores Silva

November 1, 2024

1 Introdução

Este relatório apresenta uma análise comparativa de três classificadores – Árvore de Decisão, Naive Bayes e Support Vector Machine (SVM) – aplicados a dois conjuntos de dados distintos relacionados a problemas na coluna vertebral: um conjunto de dados 2C (duas classes: Normal e Anormal) e um conjunto de dados 3C (três classes: Normal, Hérnia e Espondilolistese). A avaliação do desempenho dos classificadores foi realizada utilizando as métricas de acurácia e matrizes de confusão.

2 Resultados

2.1 Conjunto de Dados 2C

Table 1: Acurácia dos classificadores no conjunto de dados 2C.

Classificador	Acurácia
Árvore de Decisão	0.77
Naive Bayes	0.79
SVM	0.85

A SVM apresentou a maior acurácia (85%) no conjunto de dados 2C, seguida por Naive Bayes (79%) e Árvore de Decisão (77%). As matrizes de confusão para cada classificador (apresentadas nas figuras 1) permitem uma análise mais detalhada dos erros e acertos. Por exemplo, podemos observar que a SVM teve um número menor de falsos negativos (classificar um paciente "Anormal" como "Normal") em comparação com os outros dois classificadores, o que pode ser crucial em um contexto médico.

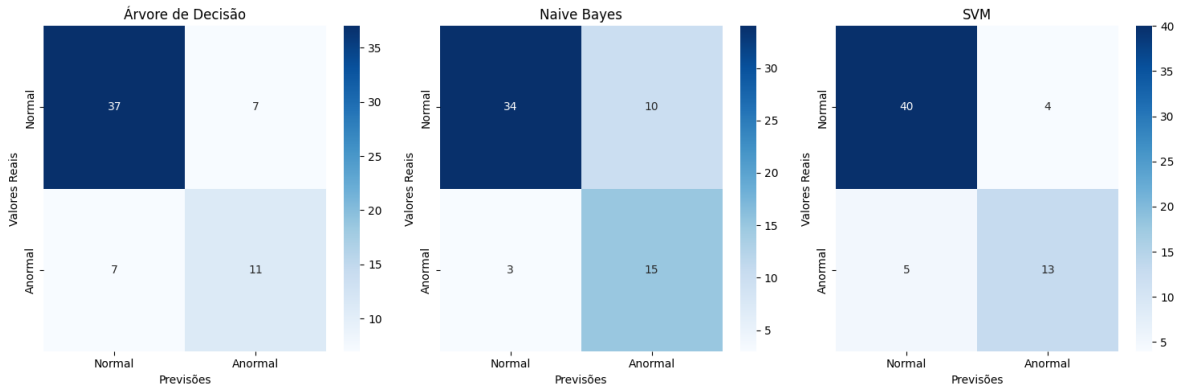

Figure 1: Matrizes de confusão para o conjunto de dados 2C.

2.2 Conjunto de Dados 3C

Table 2: Acurácia dos classificadores no conjunto de dados 3C.

Classificador	Acurácia
Árvore de Decisão	0.76
Naive Bayes	0.87
SVM	0.85

No conjunto de dados 3C, Naive Bayes alcançou a maior acurácia (87%), seguido de perto pela SVM (85%) e, por último, a Árvore de Decisão (76%). Observando as matrizes de confusão (apresentadas nas figuras 2), podemos verificar o desempenho dos classificadores para cada classe. Naive Bayes, por exemplo, demonstra uma boa capacidade de distinguir entre as três classes, com poucos falsos positivos e falsos negativos, especialmente para a classe "Espondilolistese".

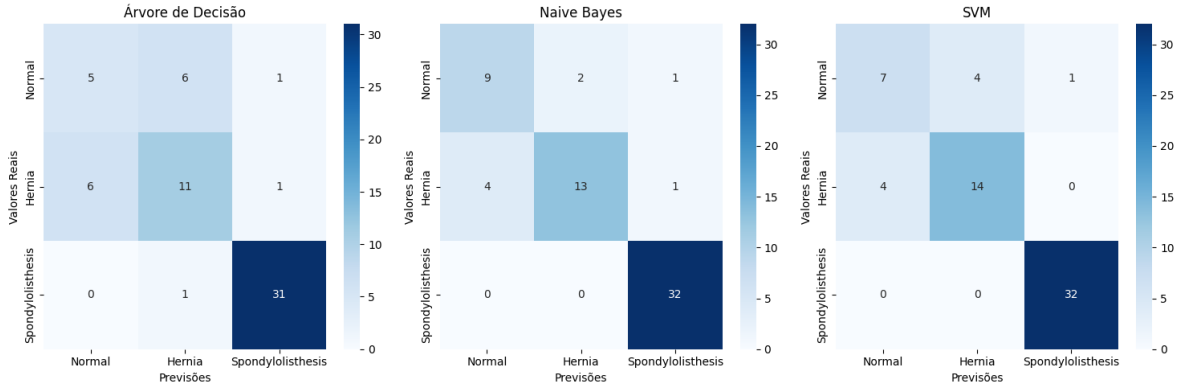

Figure 2: Matrizes de confusão para o conjunto de dados 3C.

3 Conclusão

Em ambos os conjuntos de dados, Naive Bayes e SVM demonstraram desempenho superior à Árvore de Decisão. Para o problema de duas classes (2C), a SVM apresentou uma ligeira vantagem em termos de acurácia. Já para o problema de três classes (3C), Naive Bayes obteve a melhor acurácia. A escolha do melhor classificador depende do contexto da aplicação e da importância relativa de cada tipo de erro (falsos positivos vs. falsos negativos). Em cenários médicos, onde a precisão do diagnóstico é crucial, a análise da matriz de confusão é fundamental para escolher o classificador mais adequado.

4 Observações

- A metodologia utilizada nos scripts Python incluiu pré-processamento dos dados (remoção de valores nulos e codificação de rótulos), divisão dos dados em conjuntos de treino e teste, treinamento dos classificadores e avaliação do desempenho usando acurácia e matriz de confusão.
- O parâmetro `random_state` foi fixado para garantir a reprodutibilidade dos resultados.
- As matrizes de confusão fornecem uma visão mais granular do desempenho, permitindo identificar os tipos de erros cometidos por cada classificador.