Short-Paper de Classificadores
Sistemas Inteligentes

Hector J. R. Salgueiros[1], **Willians S. Santos**[1]

[1]Sistemas de Informação – Universidade Federal do Piauí (UFPI)
64.607-670 – Picos – PI – Brasil

{hectorsalg,willianssilva}@ufpi.edu.br

1. Introdução

O câncer de mama representa um dos mais significativos desafios de saúde pública globalmente, sendo a neoplasia maligna mais frequente entre mulheres e a principal causa de morte por câncer em muitas regiões [World Health Organization and International Agency for Research on Cancer 2022].

Identificar de forma acurada quais pacientes apresentam maior risco de recorrência é crucial para otimizar as estratégias terapêuticas, permitindo intervenções mais agressivas e personalizadas para aqueles em maior risco, enquanto se evita o excesso de tratamento para outros.

2. Métricas & Métodos

2.1. Resolução

A abordagem para solucionar o problema de classificação neste estudo envolveu a implementação e comparação de quatro algoritmos distintos: K-Nearest Neighbors (KNN) e Naive Bayes desenvolvidos manualmente, juntamente com Regressão Logística e Random Forest do scikit-learn. Cada classificador foi avaliado em cinco execuções independentes, utilizando uma divisão estratificada dos dados de treino e teste (70% e 30%, respectivamente) para garantir a representatividade das classes em cada fold. As métricas de desempenho empregadas foram acurácia, precisão e recall, a partir de uma média de 5 testes.

2.2. Base de dados

A base de dados ”Breast Cancer” do UCI Machine Learning Repository [Zwitter and Soklic 1988], possui 286 instâncias, cada uma descrita por nove atributos categóricos.

- **Age (Idade):** Categoria de idade da paciente no momento do diagnóstico. Valores: ’10-19’, ’20-29’, ’30-39’, ’40-49’, ’50-59’, ’60-69’, ’70-79’, ’80-89’, ’90-99’.
- **Menopause (Estágio da Menopausa):** Status da menopausa da paciente. Valores: ’premeno’ (pré-menopausa), ’meno’ (menopausa), ’lt40’ (idade inferior a 40 anos, com pré-menopausa).
- **Tumor-size (Tamanho do Tumor):** Categoria do tamanho do tumor em milímetros. Valores: ’0-4’, ’5-9’, ’10-14’, ’15-19’, ’20-24’, ’25-29’, ’30-34’, ’35-39’, ’40-44’, ’45-49’, ’50-54’, ’55-59’.

- **Inv-nodes (Linfonodos Axilares Positivos):** Número de linfonodos axilares com presença de células cancerígenas. Valores: '0-2', '3-5', '6-8', '9-11', '12-14', '15-17', '18-20', '21-23', '24-26', '27-29', '30-32', '33-35', '36-39'.
- **Node-caps (Cápsula Linfonodal):** Presença de metástase na cápsula linfonodal. Valores: 'yes' (sim), 'no' (não).
- **Deg-malig (Grau de Malignidade):** Grau histológico de malignidade do tumor. Valores: '1', '2', '3' (onde 1 é o grau mais baixo de malignidade e 3 o mais alto).
- **Breast (Localização do Tumor na Mama):** Quadrante da mama onde o tumor está localizado. Valores: 'left' (mama esquerda), 'right' (mama direita).
- **Breast-quad (Quadrante da Mama):** Quadrante específico da mama onde o tumor foi detectado. Valores: 'left-up' (superior esquerdo), 'left-low' (inferior esquerdo), 'right-up' (superior direito), 'right-low' (inferior direito), 'central'.
- **Irradiat (Radioterapia):** Indicação se a paciente recebeu radioterapia. Valores: 'yes' (sim), 'no' (não).
- **Class (Classe):** É a variável alvo deste estudo. Esta coluna indica se houve recorrência da doença dentro de um período específico ou se a paciente permaneceu sem evidências da doença. Valores: 'no-recurrence-events' (sem eventos de recorrência) e 'recurrence-events' (eventos de recorrência).

2.3. Classificação

Para abordar o problema da previsão da recorrência do câncer de mama, este estudo explorará quatro algoritmos de classificação distintos. KNN e Naive Bayes feitos manualmente, Regressão Logística e Random Forest feitos pelo Sklearn.

O KNN é um método não-paramétrico e baseado em instâncias, que classifica um novo ponto de dados com base na classe majoritária de seus K vizinhos mais próximos no espaço de características.

O classificador Naive Bayes é um algoritmo probabilístico baseado no Teorema de Bayes, que pressupõe independência entre as características, dado a classe.

A Regressão Logística é um algoritmo fundamental para classificação binária. Ela modela a probabilidade de uma instância pertencer a uma determinada classe usando a função logística, que mapeia qualquer valor real para um valor entre 0 e 1.

Random Forest é um algoritmo de aprendizado em conjunto que constrói múltiplas árvores de decisão durante a fase de treinamento e produz a classe que é a moda das classes ou a média das previsões das árvores individuais.

2.4. Pré-Processamento

Para preparar o conjunto de dados Breast Cancer para a modelagem. Inicialmente, a coluna 'Sample_code_number' foi descartada por não conter informações preditivas relevantes. Em seguida, todas as características de tipo 'object' foram submetidas ao Label Encoding, transformando seus valores textuais em representações numéricas inteiras. Após essa conversão, todas as colunas foram explicitamente convertidas para o tipo numérico, e quaisquer valores ausentes (NaNs) resultantes foram imputados utilizando a mediana da respectiva coluna. Por fim, os índices das features contínuas e categóricas foram identificados e separados, permitindo que os algoritmos de classificação, especialmente o Naive Bayes misto, pudessem tratar cada tipo de feature de forma apropriada.

3. Avaliações dos Resultados

A avaliação do desempenho de modelos de classificação é uma etapa crucial para determinar sua eficácia e robustez. Para este estudo, a qualidade dos modelos será medida através de métricas Acurácia, Precisão (Precision) e Revocação (Recall). A Acurácia mede a proporção de previsões corretas (tanto positivas quanto negativas) em relação ao número total de previsões. A Precisão foca na qualidade das previsões positivas do modelo. O Recall foca na capacidade do modelo de identificar corretamente todas as instâncias positivas.

4. Resultados

A Tabela 1 apresenta o desempenho dos quatro modelos de classificação avaliados – KNN Manual, Naive Bayes Manual, Regressão Logística (Scikit-learn) e Random Forest (Scikit-learn) – utilizando a média das métricas de Acurácia, Precisão e Recall, acompanhadas de seus respectivos desvios padrão obtidos a partir das validações. A média é realizada em 5 teste de cada algoritmo.

Table 1. Resultados do Desempenho dos Modelos de Classificação (Média $\pm$ Desvio Padrão)

Modelo	Acurácia	Precisão	Recall
KNN Manual	0.6837 ± 0.0114	0.4664 ± 0.0277	0.2692 ± 0.0544
Naive Bayes Manual	0.7093 ± 0.0329	0.5222 ± 0.0584	0.4538 ± 0.1125
Sklearn Regressão Logística	0.7093 ± 0.0164	0.5533 ± 0.0728	0.2231 ± 0.0377
Sklearn Random Forest	0.7000 ± 0.0412	0.5007 ± 0.1030	0.3538 ± 0.0923

5. Conclusão

A análise dos resultados obtidos com os quatro modelos de classificação no dataset permitiu uma avaliação comparativa de suas capacidades preditivas. As acurácias dos modelos apresentaram-se em um intervalo similar, aproximadamente entre 68% e 70%, indicando uma consistência geral na capacidade de acerto global.

Ao examinar as métricas de Precisão e Recall, observam-se distinções mais notáveis. A Regressão Logística demonstrou a maior precisão, sugerindo uma menor ocorrência de falsos positivos. Por outro lado, o modelo Naive Bayes Manual destacou-se pelo maior recall, indicando uma superioridade na identificação dos casos de recorrência real, minimizando os falsos negativos. O Random Forest e o KNN apresentaram desempenhos intermediários nessas métricas.

6. References

References

World Health Organization and International Agency for Research on Cancer (2022). Global Cancer Observatory (GLOBOCAN): Breast Cancer Fact Sheet. `https://gco.iarc.fr/today/data/factsheets/cancers/20-Breast-fact-sheet.pdf`. Acessado em 26 de maio de 2025.

Zwitter, M. and Soklic, M. (1988). Breast Cancer. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C51P4M.