IDENTIFICAÇÃO DO CÂNCER DE MAMA UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

Adriel Wippel, Alexandre Rosa Liermann, Leonardo Gian Pottmayer e Rael dos Santos Nehring

Curso de Ciência da Computação

Departamento de Sistemas e Computação

Universidade Regional de Blumenau (FURB) – Blumenau, SC – Brasil

[adrielw@furb.br](mailto:adrielw@furb.br), [aleliermann@fur.br](mailto:aleliermann@fur.br), [lgpottmayer@furb.br](mailto:lgpottmayer@furb.br), [rsnehring@furb.br](mailto:rsnehring@furb.br), [asartori@furb.br](mailto:asartori@furb.br)

**Resumo.** O câncer de mama é uma das principais causas de mortalidade feminina em todo o mundo. A detecção precoce é crucial para o tratamento eficaz e aumento das chances de cura. Este estudo utiliza técnicas de aprendizado de máquina para classificar tumores mamários como malignos ou benignos, com base no conjunto de dados “Breast Cancer Wisconsin (Diagnostic) Dataset”. Foram implementados três algoritmos de classificação: Regressão Logística, K-Nearest Neighbors (KNN) e Support Vector Classifier (SVC). O desempenho dos modelos foi avaliado com métricas de acurácia, precisão, recall e F1-Score, atingindo até 98,24% de acurácia. Os resultados destacaram a eficácia dos algoritmos na detecção do câncer de mama.

**Palavras-chave**: Câncer de Mama. Aprendizado de Máquina. Regressão Logística. K-Nearest Neighbors. Support Vector Classifier.

# INTRODUÇÃO

O câncer de mama é um problema de saúde pública mundial, sendo responsável por um número significativo de óbitos a cada ano. Identificar tumores de maneira eficiente é essencial para o diagnóstico precoce e intervenção clínica. Técnicas de aprendizado de máquina têm sido amplamente utilizadas na análise de dados médicos, contribuindo significativamente para a melhoria dos diagnósticos e decisões clínicas.

Este trabalho utiliza o conjunto de Dados “*Breast Cancer Wisconsin (Diagnostic) Dataset*”, amplamente reconhecido para estudos em aprendizado de máquina, aplicando os algoritmos Regressão Logística, KNN e SVC. O objetivo desse trabalho é classificar tumores mamários como malignos ou benignos, com base em características extraídas de exames, avaliando o desempenho de diferentes modelos.

# MATERIAIS E MÉTODOS

## ANÁLISE DO DATASET

O conjunto de dados utilizado contém 569 amostras com 30 características numéricas, descritas como propriedades dos tumores, como raio, textura e simetria. A variável *diagnosis* é binária, representando tumores malignos (M) ou benignos (B).

Inicialmente, foi utilizada a biblioteca “missingno” com o intuito de identificar se o Dataset possui alguma amostra sem informação em alguma das colunas, onde constatou-se que não haviam valores ausentes. Em seguida, foi realizada a análise exploratória das variáveis, com o objetivo de compreender as distribuições dos valores e identificar possíveis padrões ou anomalias que poderiam impactar no desempenho do modelo.

Figura 1 – Distribuição das variáveis.

Forma, Seta

Descrição gerada automaticamente

Fonte: elaborado pelos autores.

A representação gráfica indica que as variáveis variam muito em suas escalas, radius\_mean e area\_mean têm valores na casa das dezenas e milhares, nessa sequência. Já smoothnes\_mean e fractal\_dimension, possuem valores muito pequenos (< 1). Isso indica a necessidade de normalização dos dados. Além disso, há indícios de outliers em variáveis como area\_mean e radius\_mean, destacados pelas caudas longas e desvios dos valores médios.

Para facilitar a interpretação do relacionamento entre as variáveis, foi construída a matriz de correlação, conforme a imagem abaixo.

Figura 2 –Matriz de Correlação.

Gráfico, Gráfico de mapa de árvore

Descrição gerada automaticamente

Fonte: elaborado pelos autores.

Como indicado pela matriz, algumas variáveis como radius\_mean, perimeter\_mean e area\_mean têm correlações muito fortes, indicando uma redundância entre elas.

## construção do modelo

Com base na análise do modelo, foram realizados vários testes para determinar quais variáveis poderiam ser removidas, a fim de tornar o modelo mais eficiente e reduzir o risco de overfitting. Ao final dos testes, foi identificado que a melhor solução seria remover as variáveis com limiar de correlação superior a 0.92, resultando na remoção de 8 variáveis. Após a remoção das variáveis, foi aplicada a normalização dos dados.

Para o treinamento, foram utilizadas 455 amostras e 22 variáveis, representando 80% do dataset. As demais amostras foram separadas para o grupo de teste.

## aplicando OS ALGORITMOS

Os algoritmos aplicados foram Regressão Logística, KNN e SVC. A Regressão Logística é um modelo linear que estima a probabilidade de uma amostra pertencer a uma classe específica. O K-Nearest Neighbors (KNN) classifica uma instância com base na classe mais comum entre seus *k* vizinhos mais próximos, definidos por uma métrica de distância. Já o Support Vector Classifier (SVC), uma implementação específica do SVM, busca o hiperplano que melhor separa as classes, maximizando a margem entre elas, sendo eficaz tanto para problemas lineares quanto não lineares.

Para implementar os algoritmos, cada modelo foi configurado e ajustado com parâmetros adequados para maximizar seu desempenho. A Regressão Logística foi utilizada com o parâmetro class\_weight='balanced', que corrige eventuais desequilíbrios entre as classes malignas e benignas, assegurando que ambas contribuam igualmente para o aprendizado. No caso do K-Nearest Neighbors (KNN), foi configurado com n\_neighbors=5, considerando os cinco vizinhos mais próximos para classificar uma amostra. O Support Vector Classifier (SVC) foi otimizado por meio de uma busca em grade (GridSearchCV), ajustando os hiperparâmetros C e gamma para maximizar a acurácia com validação cruzada de 5 folds.

# RESULTADOS

Para avaliar o desempenho dos modelos, foram utilizadas métricas como acurácia, precisão, recall e F1-Score. Também foi analisada a Curva ROC para medir o desempenho do modelo para diferentes limiares de decisão. Neste caso, quando mais próximo de 1 o AUC estiver, melhor o modelo, á um AUC próxima de 0.5 indica um modelo aleatório.

Figura 3 –Matriz de Correlação. Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Fonte: elaborado pelos autores.

# CONCLUSÕES

Os modelos utilizados demonstraram um ótimo desempenho, com destaque para o SVC, que apresentou os melhores resultados gerais, incluindo uma precisão de 100%, acurácia de 98,24% e AUC de 0,999. A Regressão Logística também se mostrou eficiente, resultando em uma precisão 97,83% e recall 95,74%, com uma AUC de 0,992. Já o KNN, embora tendo obtido uma precisão de 97,73%, teve menor recall (91,49%) e AUC (0,968).

O SVC mostrou-se mais adequado para o cenário estudado, combinando alta acurácia e excelente equilíbrio entre as métricas avaliadas.

# Referências

ALASWAD, Nancy. Breast Cancer Dataset [dataset]. Disponível em: <https://www.kaggle.com/datasets/nancyalaswad90/breast-cancer-dataset>. Acesso em: 10 out. 2024.

Noreen Fatima, Li Liu, Sha Hong, and Haroon Ahmed. "Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis." *IEEE Access*, vol. 8, 2020, pp. 150360-150375. doi:10.1109/ACCESS.2020.3016715.

Wenbin Yue, Zidong Wang, Hongwei Chen, Annette Payne, and Xiaohui Liu. "Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis." *Designs*, vol. 2, no. 2, 2018, Article 13. doi:10.3390/designs2020013.