Comparação de modelos de ML para diagnóstico do câncer de Mama

Modelos: Regressão Logística, Máquinas de vetores de suporte - SVM e Arvore de decisão.

Vivian Kailany

Introdução - Câncer de Mama

O câncer de mama é uma doença que atinge milhões de pessoas, a estimativa global é que cerca de **2,3 milhões de mulheres** foram diagnosticadas com câncer de mama em 2020, de acordo com a Organização Mundial da Saúde (OMS).

Modelos utilizados

Regressão Logística

Modelo probabilístico que estima a probabilidade de uma amostra pertencer a uma classe.

Máquina de Vetores de Suporte (SVM)

Algoritmo que separa as classes através de um hiperplano de margem máxima.

Árvore de decisão

Modelo baseado em árvores de decisão que divide os dados em subconjuntos baseados nos atributos.

Conjunto de Dados

1

Breast Cancer Wisconsin Dataset

O conjunto de dados é composto por 569 amostras, sendo 212 malignas e 357 benignas. Cada amostra possui 30 atributos.

2

Atributos

Os atributos são valores que representam medidas como tamanho, forma e textura das células.

3

Objetivo

O objetivo é classificar os tumores como malignos ou benignos, utilizando os 29 atributos para treinar e avaliar modelos de Machine Learning.

Otimização de Hiperparâmetros

Regressão Logística

A Regressão Logística utilizará **a** e **epocas** como hiperparâmetros. A taxa de aprendizado **a** varia de 0,001 a 0,1, e as **épocas** variam de 100 a 300.

{'a': 0.1, 'epocas': 200}

Máquina de Vetores de Suporte (SVM)

O modelo SVM utiliza os hiperparâmetros **C**, **gamma** e **kernel** para definir a complexidade do modelo. **C** varia de 2^-25 a 2^25, e o **gamma** varia de 2^-25 a 2^23, **kernel**: ['linear', 'rbf'].

{'C': 8192, 'gamma': 0.0001220703125, 'kernel': 'rbf'}

Árvore de Decisão

A Árvore de Decisão utiliza **max_depth** e **min_samples_leaf**. O **max_depth** varia de 1 a 20, definindo a profundidade da árvore, e o **min_samples_leaf** varia de 1 a 19, definindo o número mínimo de amostras em cada folha.

{'max_depth': 7, 'min_samples_leaf': 2}

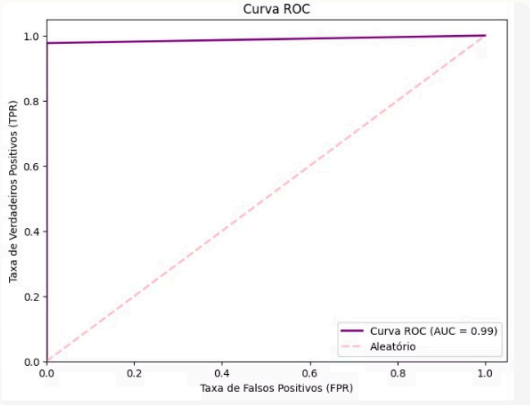
Metodologia

- **Validação Cruzada de 10 folds**: Garante que os modelos sejam avaliados de forma robusta, dividindo o conjunto de dados em 10 partes e utilizando 9 partes para treinamento e 1 parte para teste, repetindo o processo 10 vezes.
- **Métricas**: Acurácia, Precisão, Recall e F1-score serão utilizadas para avaliar o desempenho dos modelos, fornecendo uma visão abrangente da capacidade de cada modelo em classificar corretamente tumores malignos e benignos.
- **Ferramentas**: O Scikit-learn será utilizado para implementar os modelos de Machine Learning, enquanto o Grid Search CV será empregado para otimizar os hiperparâmetros de cada modelo, buscando a melhor configuração para alcançar o desempenho desejado.

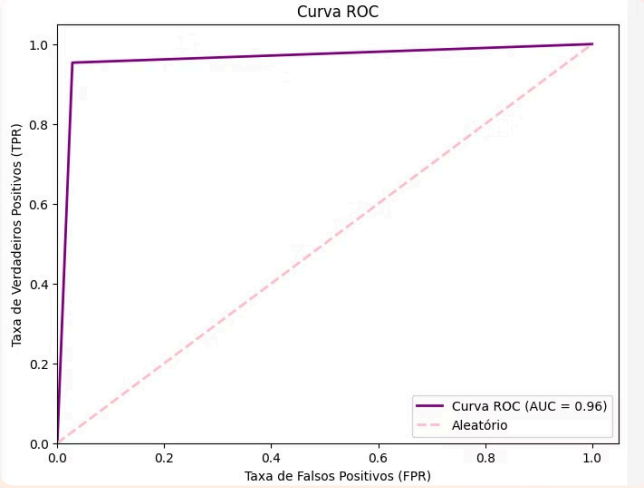
Métricas Obtidas

Modelo	Acurácia	Precisão	Revocação	F1
Regressão Logística	99%	100%	97%	98%
SVM	96%	95%	95%	95%
Árvore de Decisão	94%	95%	90%	92%

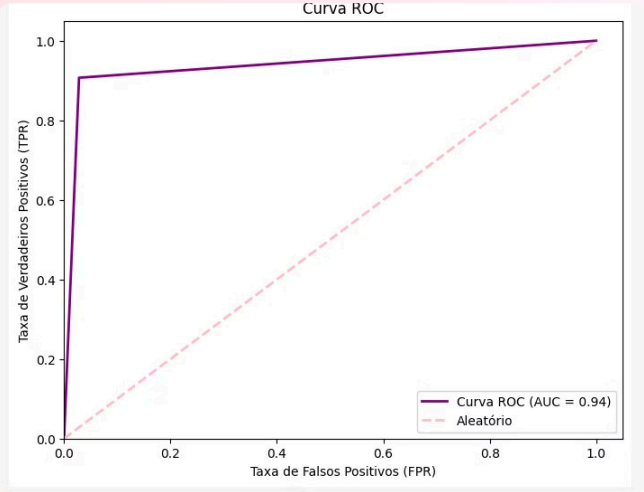
Curva ROC



Regressão Logística 99%



SVM 96%



Árvore de decisão 94%

Conclusão e Próximos Passos

Embora a regressão logística tenha se saído bem em todas as métricas com esse conjunto de dados, em casos clínicos como este, a **revocação** é um fator crucial que não pode ser negligenciado.

Novas abordagens!

Referências

UCI Machine Learning Repository: https://archive.ics.uci.edu/

Carga global de câncer aumenta em meio à crescente necessidade de serviços

Câncer de mama