Classificação de Tumores de Mama:
Comparação de Modelos de Machine Learning

Adinan Alves de Brito Filho
Lucas Jean Emmanuel Gauvain
João Pedro Sousa Santos

Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas
Sistemas Inteligentes - Prof. Dr. Luneque Silva Junior

17 de dezembro de 2024

UFABC

Visão geral

O câncer de mama é um dos tipos mais **comuns** no mundo.

1. No mundo, o câncer de mama é o terceiro tipo de câncer mais incidente, juntamente com o de pulmão e o colorretal. [6]
2. No Brasil, o câncer de mama é o segundo tipo de câncer mais comum em mulheres. [1]
3. Em 2022, foram estimados cerca de 2,3 milhões de novos casos de câncer de mama no mundo. [3]
4. No Brasil, são estimados cerca de 73.610 novos casos de câncer de mama em 2024. [2]

UFABC

Objetivo e impacto

Objetivo

- Criar modelos de Machine Learning para classificar tumores como malignos ou benignos usando o dataset `load_breast_cancer` do Scikit-learn.

Impacto

1. Detecção precoce da doença
2. Decisões clínicas mais rápidas
3. Redução de custos e recursos médicos

UFABC

Informações do dataset

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst texture	worst perimeter	worst area	worst smoothness	worst compactness	worst concavity	worst concave points	worst symmetry	worst fractal dimension	target
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	0.07871	...	17.33	184.60	2019.0	0.1622	0.6656	0.7119	0.2654	0.4601	0.11890	0
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	...	23.41	158.80	1956.0	0.1238	0.1866	0.2416	0.1860	0.2750	0.08902	0
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	0.05999	...	25.53	152.50	1709.0	0.1444	0.4245	0.4504	0.2430	0.3613	0.08758	0
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597	0.09744	...	26.50	98.87	567.7	0.2098	0.8663	0.6869	0.2575	0.6638	0.17300	0
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809	0.05883	...	16.67	152.20	1575.0	0.1374	0.2050	0.4000	0.1625	0.2364	0.07678	0

5 rows × 31 columns

- **Dataset:** `load_breast_cancer` do Scikit-learn, confeccionado por W. Street, W. Wolberg, O. Mangasarian, 1993 [8].
- **Descrição:** Contém informações sobre características de núcleos celulares extraídas de imagens digitalizadas de punções aspirativas por agulha fina (PAAF) de massas mamárias.
- **Tamanho:** 569 amostras.
- **Classes:** 212 malignos (0), 357 benignos (1).
- **Features:** 30 características numéricas.

UFABC

Informações do dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 31 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   mean radius              569 non-null    float64
 1   mean texture             569 non-null    float64
 2   mean perimeter           569 non-null    float64
 3   mean area                569 non-null    float64
 4   mean smoothness          569 non-null    float64
 5   mean compactness         569 non-null    float64
 6   mean concavity           569 non-null    float64
 7   mean concave points      569 non-null    float64
 8   mean symmetry            569 non-null    float64
 9   mean fractal dimension   569 non-null    float64
 10  radius error             569 non-null    float64
 11  texture error            569 non-null    float64
 12  perimeter error          569 non-null    float64
 13  area error               569 non-null    float64
 14  smoothness error         569 non-null    float64
 15  compactness error        569 non-null    float64
 16  concavity error          569 non-null    float64
 17  concave points error     569 non-null    float64
 18  symmetry error           569 non-null    float64
 19  fractal dimension error  569 non-null    float64
 20  worst radius             569 non-null    float64
 21  worst texture            569 non-null    float64
 22  worst perimeter          569 non-null    float64
 23  worst area               569 non-null    float64
 24  worst smoothness         569 non-null    float64
 25  worst compactness        569 non-null    float64
 26  worst concavity          569 non-null    float64
 27  worst concave points     569 non-null    float64
 28  worst symmetry           569 non-null    float64
 29  worst fractal dimension  569 non-null    float64
 30  target                   569 non-null    int64
dtypes: float64(30), int64(1)
memory usage: 137.9 KB
```

Figura: Features do dataset.

UFABC

Informações do dataset

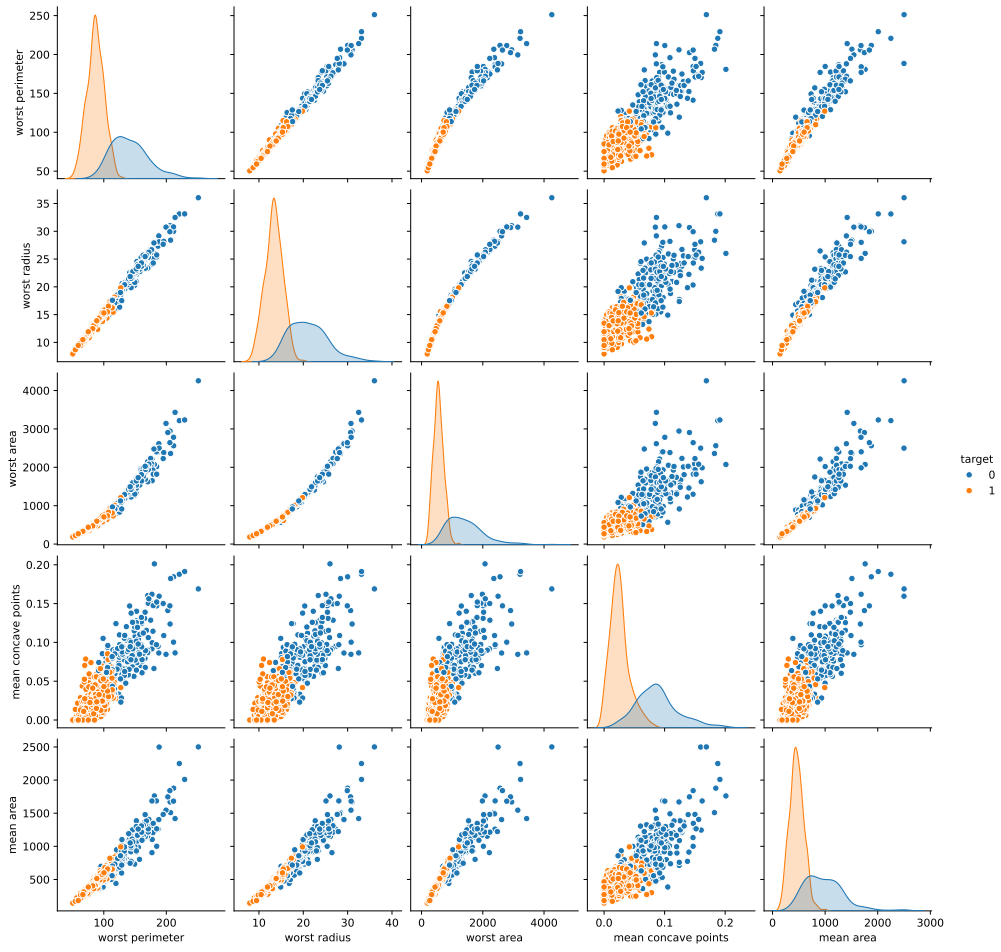

Figura: Pairplot - Correlação entre algumas variáveis do dataset.

UFABC

KFold

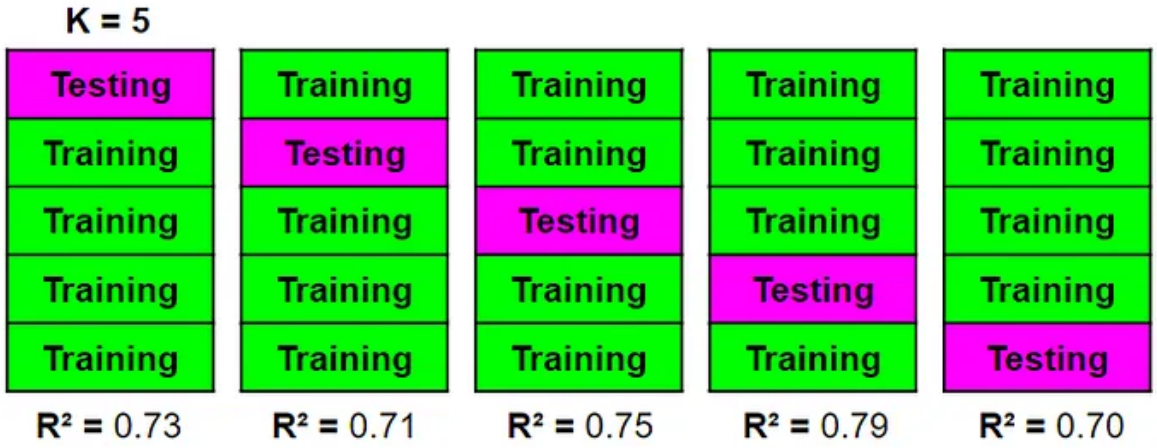

Figura: Método KFold para Validação Cruzada - Em cada iteração, o conjunto de dados é particionado em diferentes subconjuntos de treino e teste. Isso permite obter resultados variados a cada divisão, possibilitando a avaliação do desempenho do modelo em diferentes partes dos dados. Dessa forma, é possível escolher o modelo que melhor generaliza o comportamento do dataset completo [4].

UFABC

Random Forest

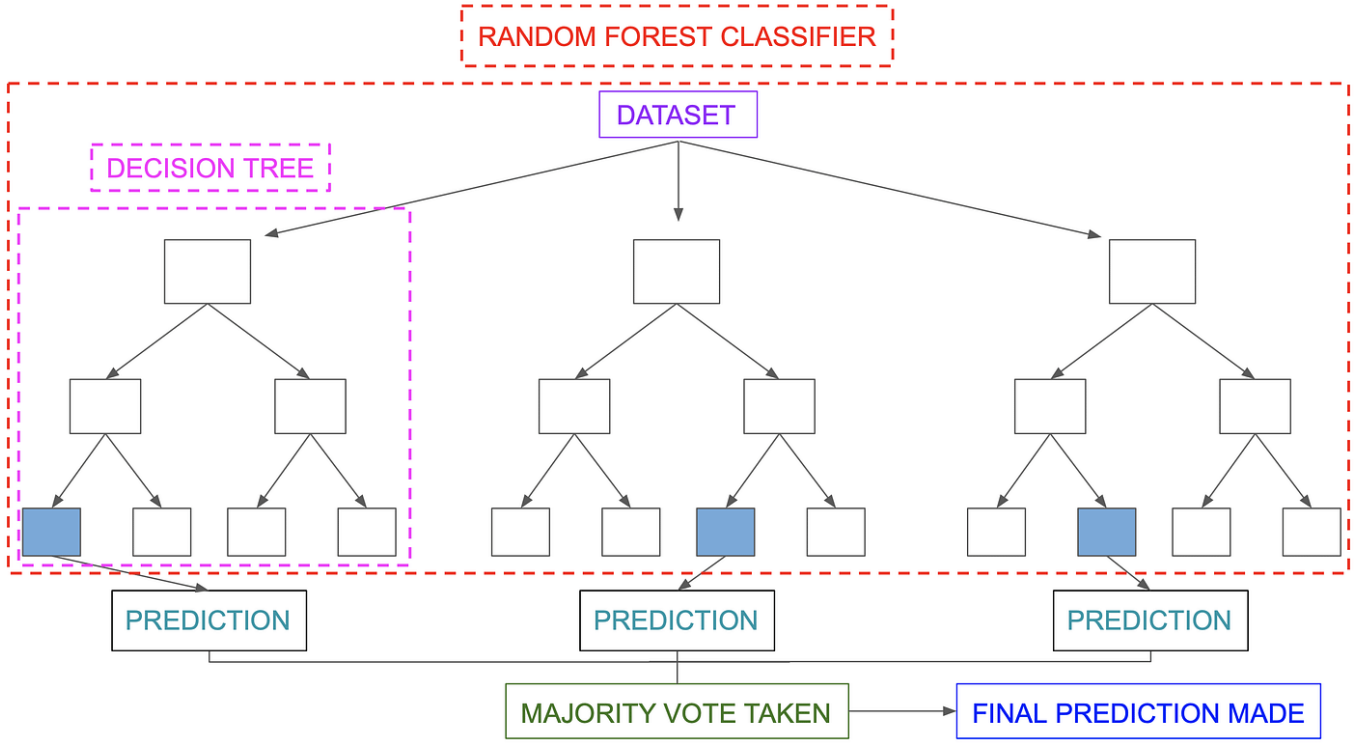

Figura: Um classificador *Random Forest* é composto por várias árvores de decisão para uma classificação aprimorada [7].

UFABC

Linear Support Vector Classification

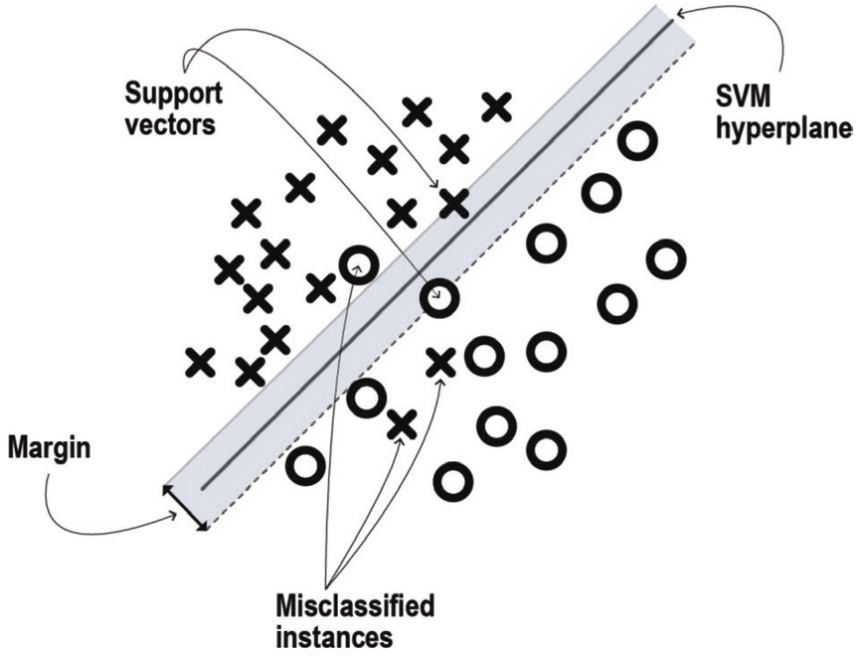

Figura: O LinearSVC é uma implementação do algoritmo Support Vector Machine (SVM), projetada especificamente para lidar com dados linearmente separáveis. O classificador funciona tentando encontrar um hiperplano em um espaço N-dimensional que classifique distintamente os pontos de dados [5].

UFABC

Métricas de avaliação

- **Acurácia**: Valor numérico que define a proporção de predições corretas feitas pelo modelo em relação ao total de predições realizadas.
- **R^2 score**: Mede a proporção da variabilidade total dos dados que é explicada pelo modelo. Em outras palavras, o R^2 indica quão bem o modelo se ajusta aos dados.
- **Matriz de Confusão**: É uma tabela que descreve o desempenho de um modelo de classificação, mostrando os números de verdadeiros positivos (TP), falsos positivos (FP), verdadeiros negativos (TN) e falsos negativos (FN).
- **AUC-ROC (Área sob a Curva ROC)**: A curva ROC *(Receiver Operating Characteristic)* é um gráfico que mostra a taxa de verdadeiros positivos contra a taxa de falsos positivos.

UFABC

Fluxo de trabalho

1. **Importação e preparação dos dados**
2. **Configuração do KFold**
   - Divisão dos dados em **5 folds** para validação cruzada
3. **Inicialização de variáveis para armazenar os resultados**
4. **Loop principal (Validação cruzada com KFold)**
   Para cada fold:
   - Separação dos dados de treino e teste
   1. **Subloop: Iteração de 0 a 10 para treinamento de modelos aleatórios**
      Para cada random_state (0-9):
      - Definição dos modelos *(Random Forest e SVM)*
      - Treinamento dos modelos
      - Cálculo das métricas *(acurácia, MSE, R^2 score)*
      - Comparação dos modelos e atualização do melhor modelo, caso positivo

UFABC

Resultados gerais

Model	Fold	Model Random State	Accuracy	R2 Score
Random Forest	2	3	0.991228	0.959986
Support Vector Machine	4	0	0.982456	0.925319

Tabela: Melhores resultados de treinamento dos modelos.

Pode-se observar que os dois modelos testados apresentaram excelente desempenho, com acurácias altíssimas.
Contudo, dentre os modelos avaliados, o Random Forest se destacou como o melhor modelo.

UFABC

Resultados gerais

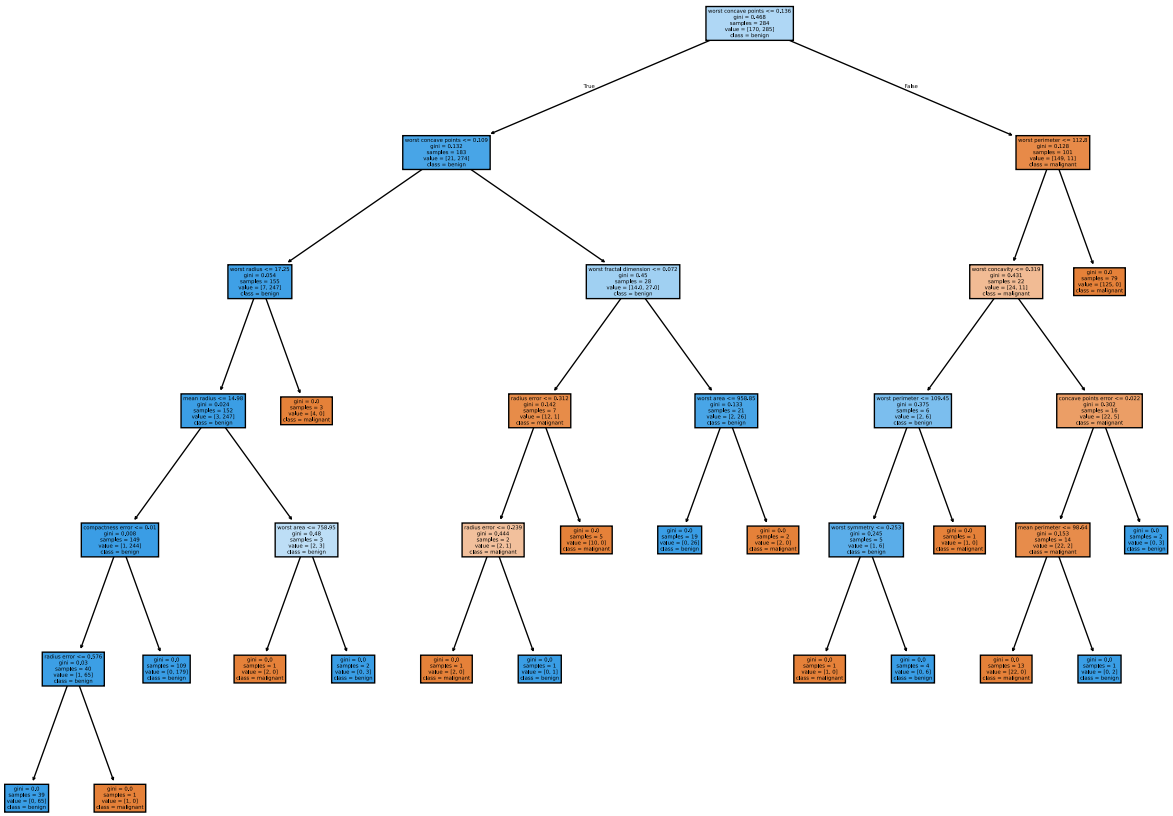

Figura: Primeira árvore de decisão do modelo *Random Forest*.

UFABC

Matriz de Confusão

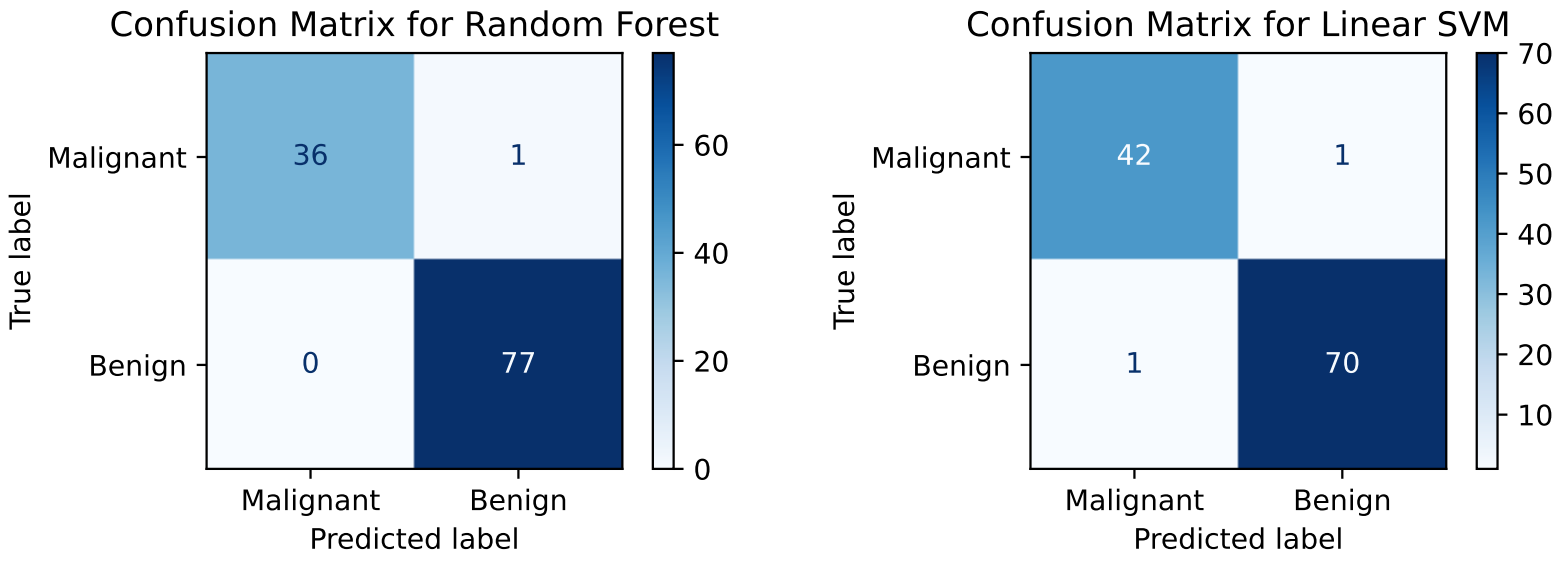

Figura: Matriz de confusão dos modelos *Random Forest* e *LSVC*.

Nota-se que o *Random Forest* é muito preciso para detectar tumores benignos (zero FP) e tem ótimo desempenho para tumores malignos.

UFABC

Curva ROC

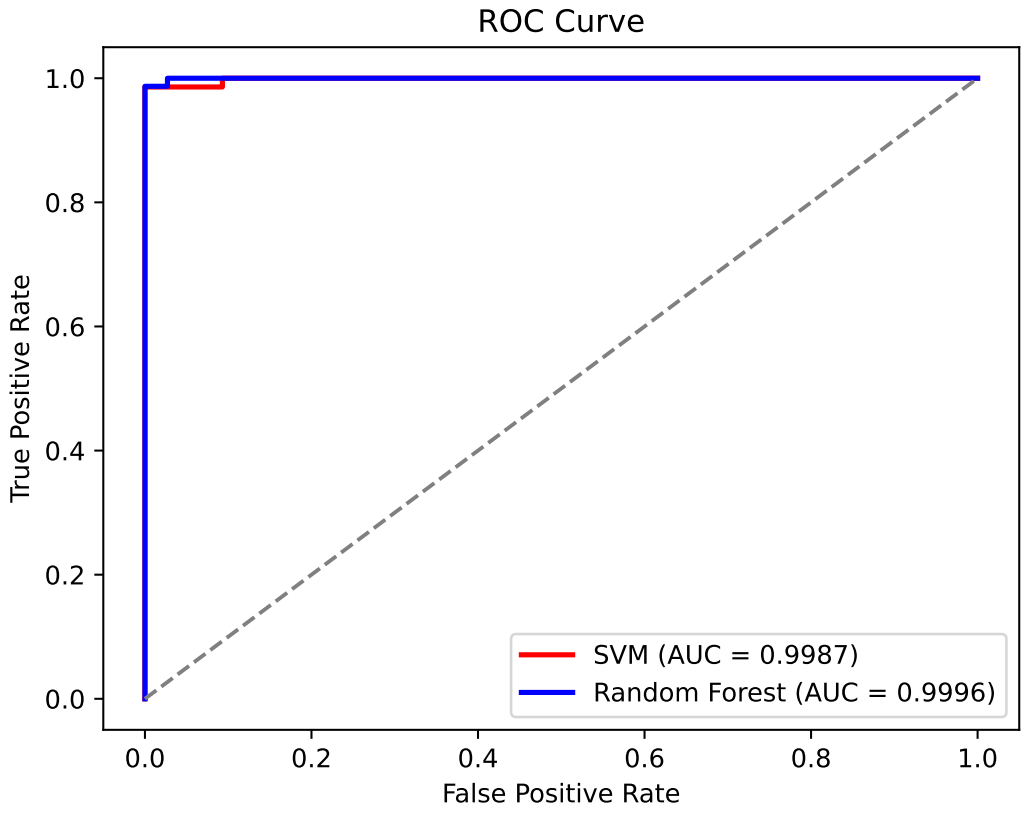

Figura: A curva ROC para o modelo *Random Forest* demonstra desempenho quase perfeito, com uma AUC de 0,9996. Isso indica que o modelo pode distinguir muito bem as classes positiva e negativa.

UFABC

Feature Importance

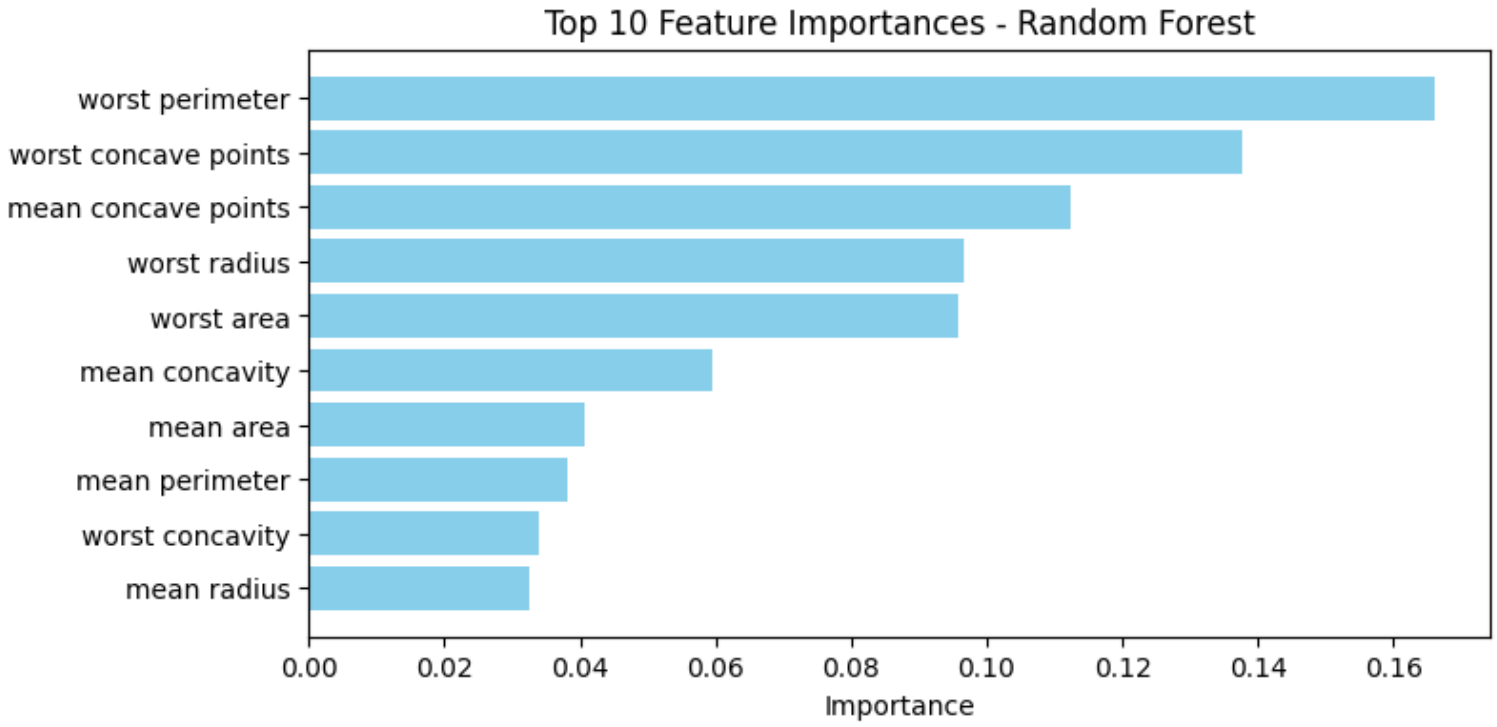

Figura: Gráfico exibindo as características mais importantes para a predição de tumores.

UFABC

Conclusão

Os resultados obtidos demonstraram um excelente desempenho dos modelos, com acurácia próxima a **99%** para o **Random Forest**, que se destacou como o modelo mais preciso.

O sucesso do **Random Forest** pode ser atribuído à sua **robustez** em lidar com dados complexos e à capacidade de corrigir o *overfitting* característico de árvores de decisão individuais.

Por outro lado, o **Linear SVC** mostrou-se eficiente para dados linearmente separáveis, porém apresentou desempenho ligeiramente inferior em comparação ao Random Forest.

Bibliografia

Ministério da Saúde.

Câncer de mama.

https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/c/cancer-de-mama.

[Accessado em 14-12-2024].

Instituto Nacional de Câncer (Brasil).

Controle do câncer de mama no Brasil: dados e números 2024.

Ministério da Saúde, 2024.

Instituto Nacional de Câncer INCA.

Controle do câncer de mama - conceito e magnitude.

https://www.gov.br/inca/pt-br/assuntos/gestor-e-profissional-de-saude/controle-do-cancer-de-mama/conceito-e-magnitude, 2024.

[Accessado em 14-12-2024].

UFABC

Bibliografia

Rodrigo Leite.

Introdução a validação-cruzada: K-fold.

https://rodrigols89.medium.com/introdu%C3%A7%C3%A3o-a-valida%C3%A7%C3%A3o-cruzada-k-fold-2a6bced32a90, 2020.

[Accessado em 14-12-2024].

Juan Manuel Núñez, Sandra Medina-Fernández, F. Gerardo Ávila, and Jorge Montejano.

High-Resolution Satellite Imagery Classification for Urban Form Detection, pages 1–9.

IntechOpen, 02 2019.

Grupo Oncoclínicas.

Tudo sobre o câncer de mama — oncoclínicas.

https://grupooncoclinicas.com/tudo-sobre-o-cancer/tipos-de-cancer/cancer-de-mama.

[Accessado em 14-12-2024].

UFABC

Bibliografia

Bhushan Talekar.

A detailed review on decision tree and random forest.

Bioscience Biotechnology Research Communications, 13:245–248, 12 2020.

W. Wolberg, O. Mangasarian, and N. Street.

Breast cancer wisconsin (diagnostic).

UCI Machine Learning Repository, 1993.

DOI: https://doi.org/10.24432/C5DW2B.

UFABC