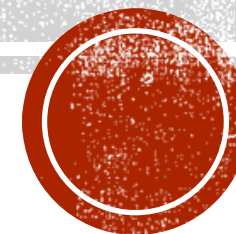


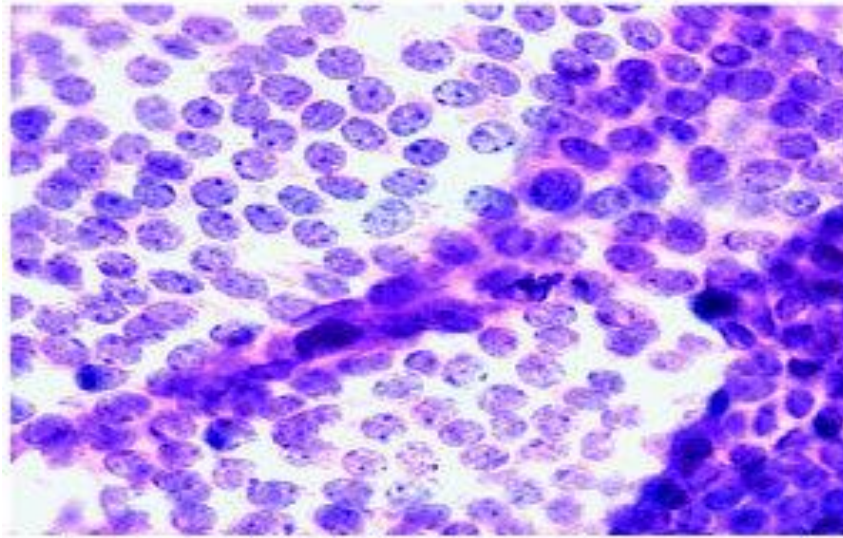
ANÁLISE CÂNCER MAMA WISCONSIN



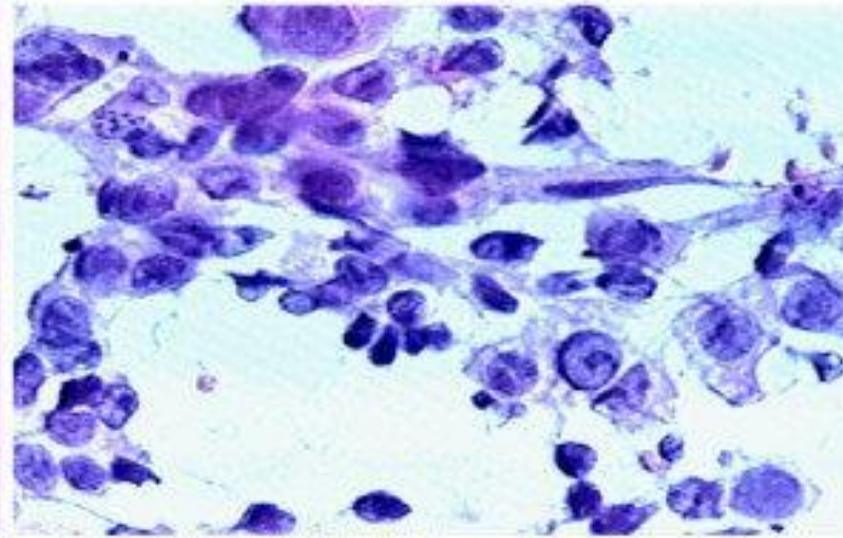
Machine learning e estatística não-paramétrica

SOBRE OS DADOS

- Consiste em um grupo de 569 mulheres, onde 357 possuem câncer benigno e 212 câncer maligno
- As características são calculadas a partir de uma imagem digitalizada de uma Aspirativa por agulha fina de uma massa mamária. Eles descrevem características dos núcleos celulares presentes na imagem.



BENIGNO

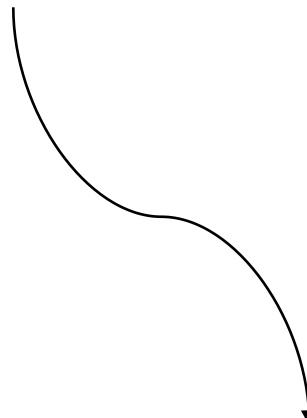


MALIGNO



SOBRE OS DADOS

- Dessa imagem são calculadas as seguintes métricas:
 - ❖ Raio das células
 - ❖ Textura (desvio padrão do nível de cinza)
 - ❖ Perímetro
 - ❖ Área
 - ❖ Suavidade (variação local do comprimento do raio)
 - ❖ Compacidade ($\text{perímetro}^2 / \text{área} - 1.0$)
 - ❖ Concavidade
 - ❖ Pontos côncavos (numero de pontos côncavos)
 - ❖ Simetria
 - ❖ Dimensão fractal (distância da fronteira entre as células)



DESSAS MÉTRICAS SÃO
CALCULADAS SUAS MÉDIAS,
ERROS PADRÃO E MÉDIA DOS
3 MAIORES VALORES



MÉTODOS USADOS

- **Divisão de treino e teste**

70% dos dados para treino

30% dos dados para teste



MÉTODOS USADOS

PARA REDUÇÃO DE DIMENSIONALIDADE DOS DADOS:

- **Teste t para variáveis normais:**

H0: As médias são iguais

H1: As médias são diferentes

- **Teste de Wilcoxon e Mannwhitney para variáveis não normais:**

H0: As medianas são iguais

H1: As medianas são diferentes

- **Correlação entre as covariáveis (0.95)**

ALGORITMOS USADOS:

▪ **Árvore de decisão:**

Assim como um fluxograma, a árvore de decisão estabelece **nós** (decision nodes) que se relacionam entre si por uma hierarquia.

▪ **KNN:**

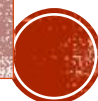
o KNN tenta classificar cada amostra de um conjunto de dados avaliando sua distância em relação aos vizinhos mais próximos.

▪ **XGBOOST:**

Se favorece de múltiplas árvores de decisão que são melhoradas a cada interação

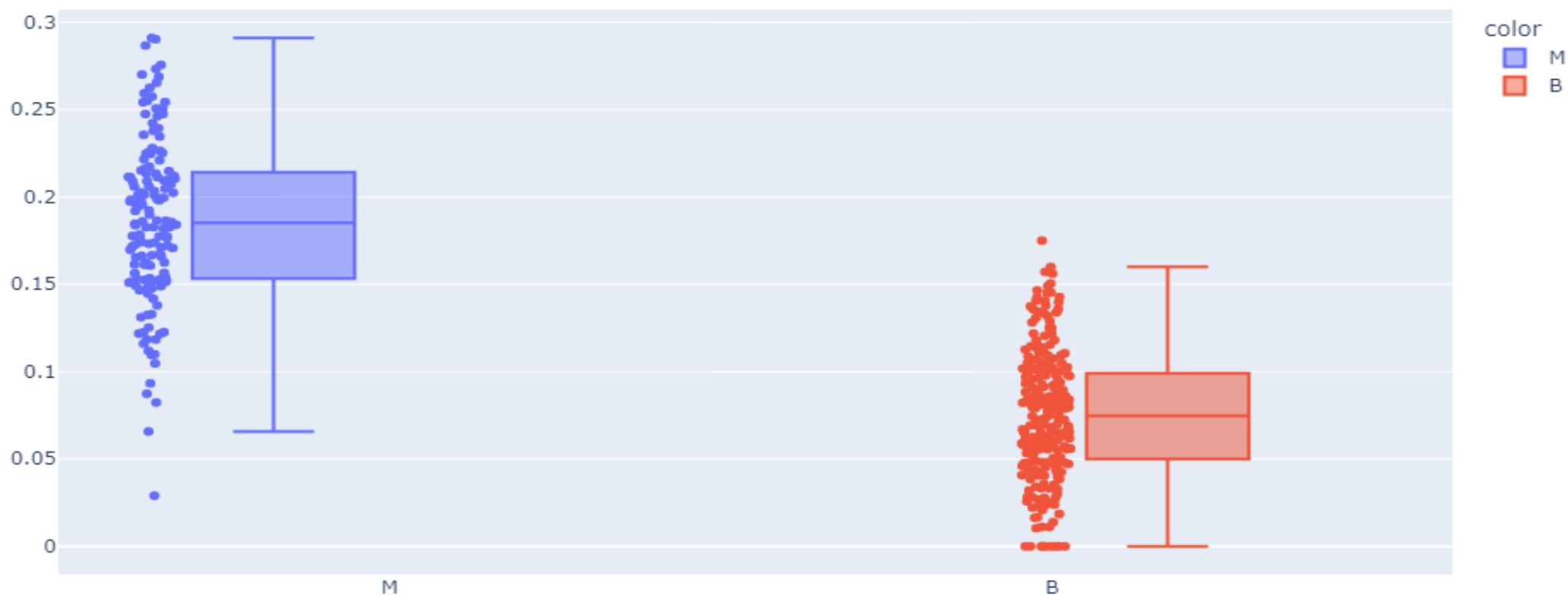
▪ **Regressão logística:**

A regressão logística é um modelo estatístico usado para determinar a probabilidade de um evento acontecer.



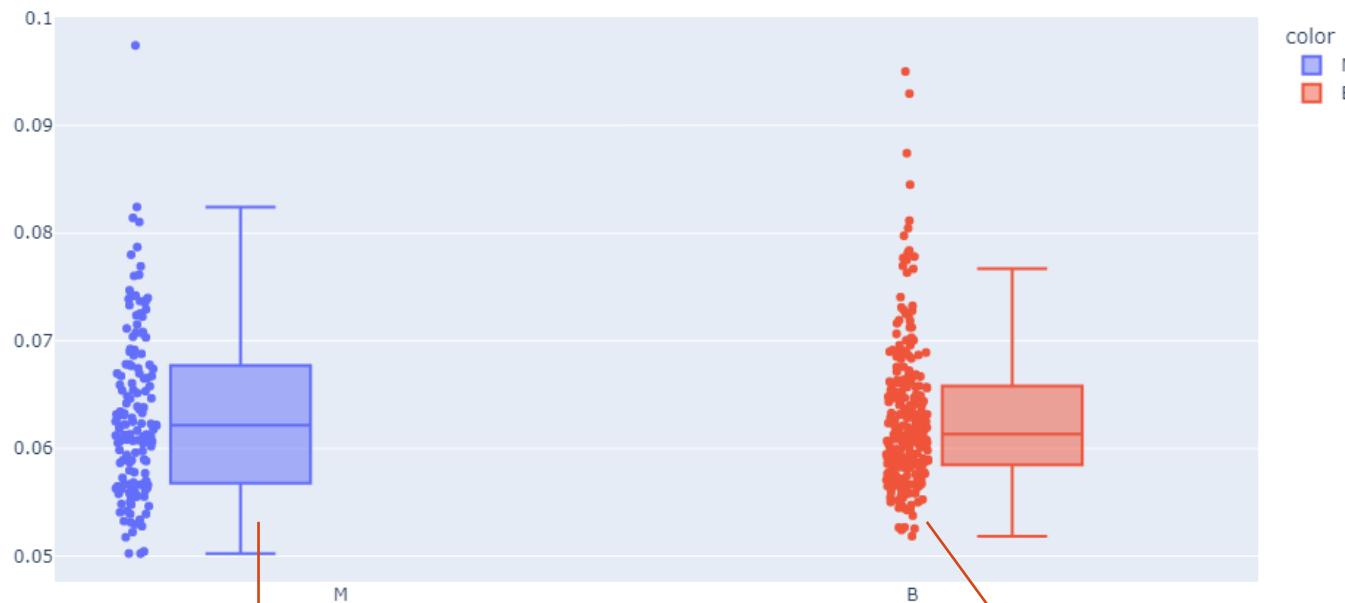
RESULTADOS

- Do par de variáveis a única que apresentou normalidade foi “piores pontos côncavos”;
- O teste t indicou que há diferença entre as médias do diagnóstico Benigno e Maligno para essa variável ($p\text{-valor} < 0,05$)



RESULTADOS

- Variáveis com medianas iguais de acordo com o teste de Wilcoxon



Media dimensões factrais

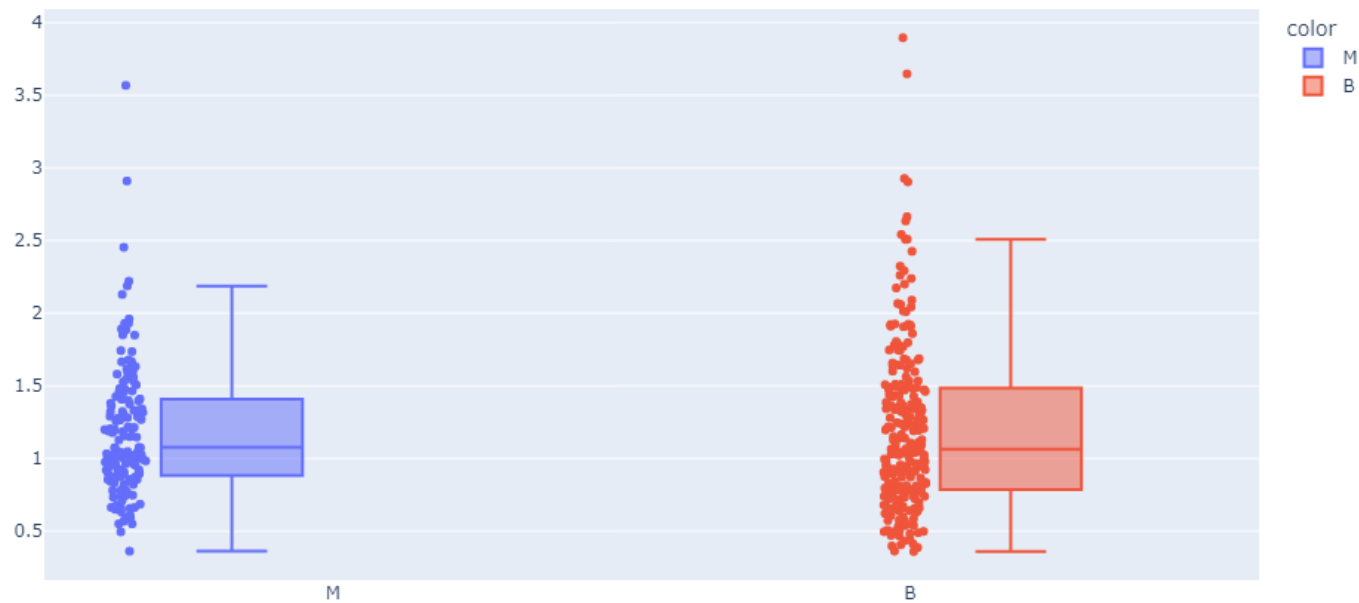
Mediana = 0.062

Mediana = 0.061



RESULTADOS

- Variáveis com medianas iguais de acordo com o teste de Wilcoxon



Erro padrão da textura

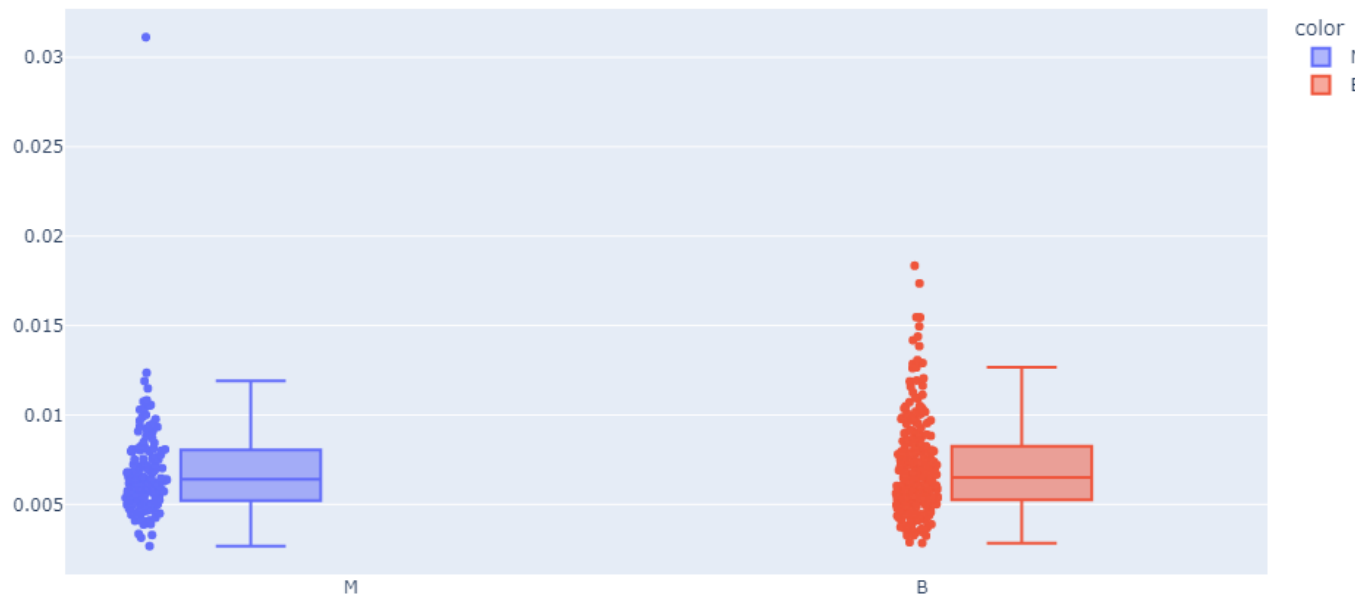
Mediana = 1.07

Mediana = 1.06



RESULTADOS

- Variáveis com medianas iguais de acordo com o teste de Wilcoxon



Erro padrão da suavidade

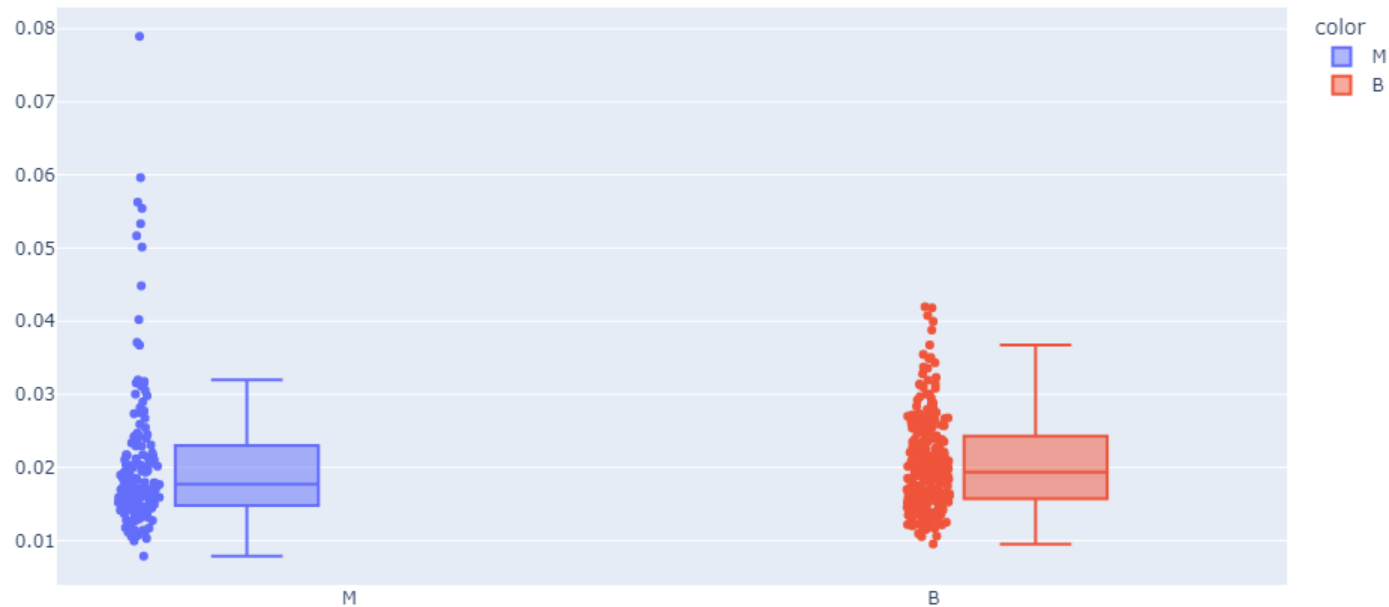
Mediana = 0.0063

Mediana = 0.0065



RESULTADOS

- Variáveis com medianas iguais de acordo com o teste de Wilcoxon



Erro padrão da simetria

Mediana = 0.017

Mediana = 0.019



RESULTADOS

- Variáveis altamente correlacionadas:

		media_raio	media_perimetro	media_area	pior_raio	pior_perimetro	pior_area	errP_raio	errP_perimetro
●	media_raio	1.00	1.00	0.99	0.97	0.97	0.94	0.67	0.66
✗	media_perimetro	1.00	1.00	0.99	0.97	0.97	0.94	0.68	0.69
✗	media_area	0.99	0.99	1.00	0.97	0.96	0.96	0.71	0.70
✗	pior_raio	0.97	0.97	0.97	1.00	0.99	0.98	0.73	0.71
✗	pior_perimetro	0.97	0.97	0.96	0.99	1.00	0.98	0.74	0.74
✗	pior_area	0.94	0.94	0.96	0.98	0.98	1.00	0.76	0.74
✗	errP_raio	0.67	0.68	0.71	0.73	0.74	0.76	1.00	0.97
●	errP_perimetro	0.66	0.69	0.70	0.71	0.74	0.74	0.97	1.00



RESULTADOS

AGORA VAMOS APLICAR OS MODELOS NO PYTHON...



RESULTADOS FINAL

- **O xgboost foi o melhor modelo para prever câncer de mama, acurácia de (0,95);**
- **A profundidade máxima de suas árvores de decisão foi 4;**
- **O knn foi o com a menor acurácia (0.91);**
- **Apesar disso, todos os 4 modelos obtiveram bons resultados;**

