CIÊNCIA DE DADOS E BIG DATA

19 de Novembro, 2025

Diagnóstico de Câncer de Mama com Ciência de Dados

Classificação de Tumores Benignos vs Malignos

Taimisson C. Schardosim

UNISINOS

Nossa sala de aula é o mundo.

UNISINOS

Contexto

O Câncer de Mama

é o tipo de câncer que **mais mata mulheres** no Brasil

1 a cada 8 mulheres, serão diagnosticadas com câncer de mama durante sua vida

95% dos casos **têm cura** quando diagnosticado cedo

80%

dos tumores são descobertos pela própria mulher ao **palpar suas mamas.**

a cada ano, mais de

600 MIL

mulheres **perdem a vida** em todo o mundo por causa dessa doença.

19 de Novembro, 2025

fonte: https://unintese.com.br/outubro-rosa

UNISINOS

Conjunto de Dados
Wisconsin Diagnostic Breast Cancer

- Origem: Repositório público da University of California, Irvine
- 569 amostras com 32 colunas (1 ID, 1 Diagnóstico e 30 atributos numéricos)
- Diagnóstico: Benigno (B) x Maligno (M).
- Dados obtidos de imagens de aspirado por agulha fina de massas mamárias

UNIVERSITY of CALIFORNIA
IRVINE

Dimensões do dataset: (569, 32)

	ID	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave_poin
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	

5 rows × 32 columns

https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

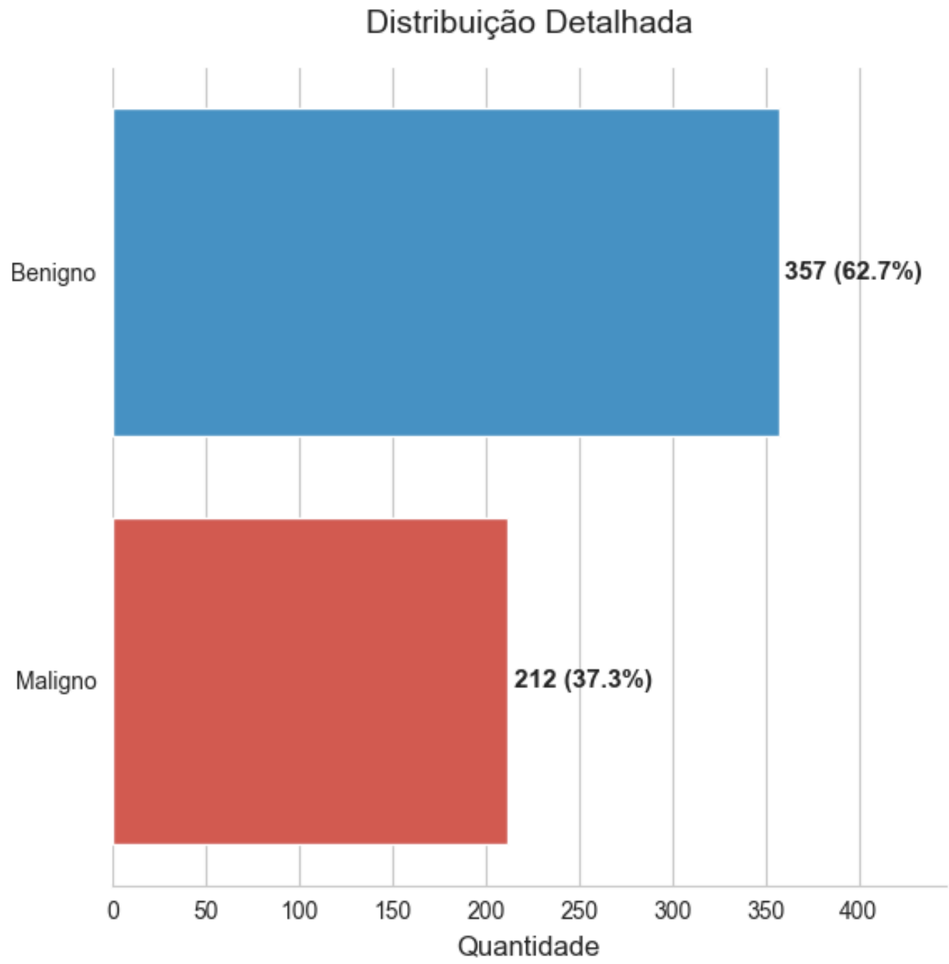
Estrutura das Variáveis e Estatísticas Descritivas

- Resumo da estatística descritiva (describe())
- 357 tumores benignos e 212 malignos

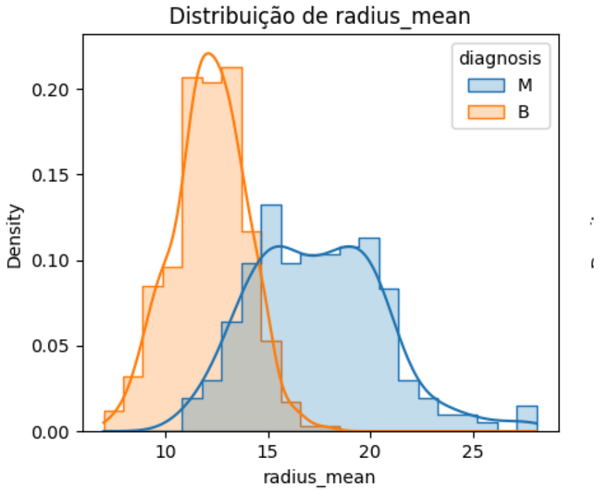
radius_mean float64
texture_mean float64
perimeter_mean float64
area_mean float64
smoothness_mean float64
dtype: object

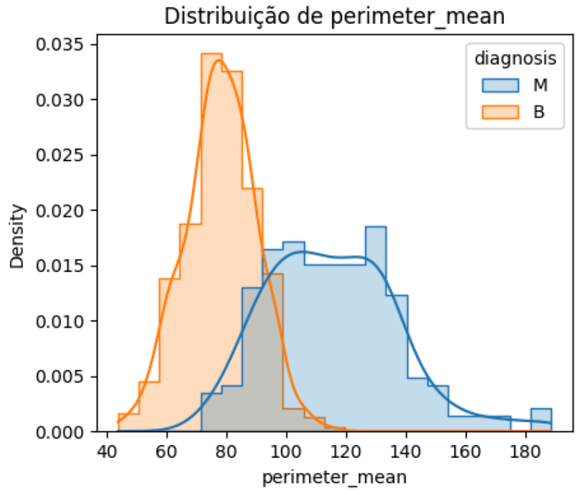
	count	mean	std	min	25%	50%
radius_mean	569.0	14.127292	3.524049	6.981000	11.700000	13.370000
texture_mean	569.0	19.289649	4.301036	9.710000	16.170000	18.840000
perimeter_mean	569.0	91.969033	24.298981	43.790000	75.170000	86.240000
area_mean	569.0	654.889104	351.914129	143.500000	420.300000	551.100000
smoothness_mean	569.0	0.096360	0.014064	0.052630	0.086370	0.095870
compactness_mean	569.0	0.104341	0.052813	0.019380	0.064920	0.092630

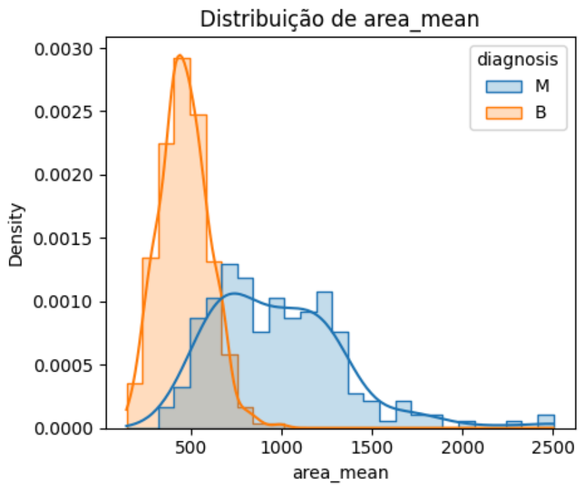
O dataset é levemente desbalanceado, mas suficiente para treinamento sem técnicas artificiais de balanceamento.

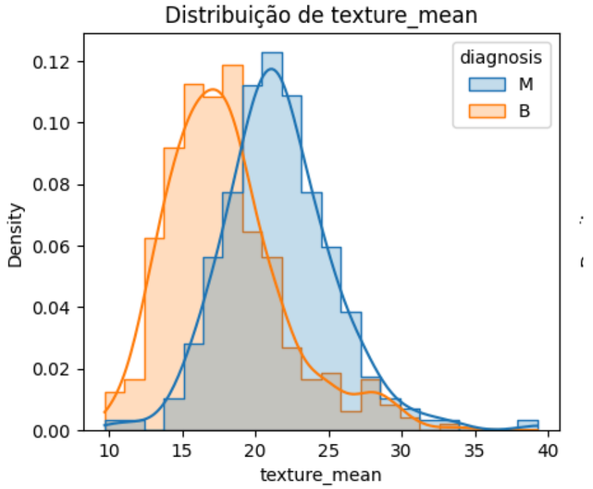


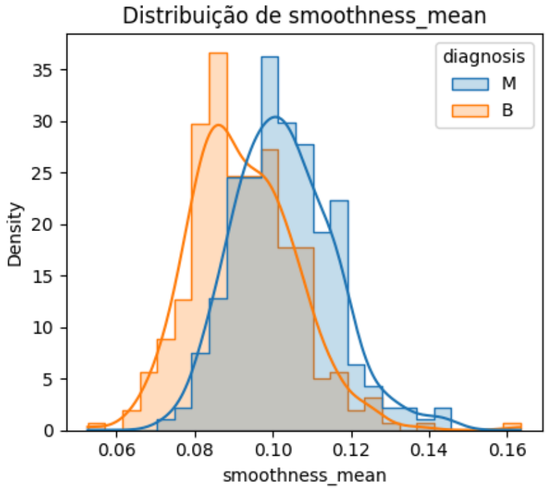
Visualização 1: Histogramas por Classe







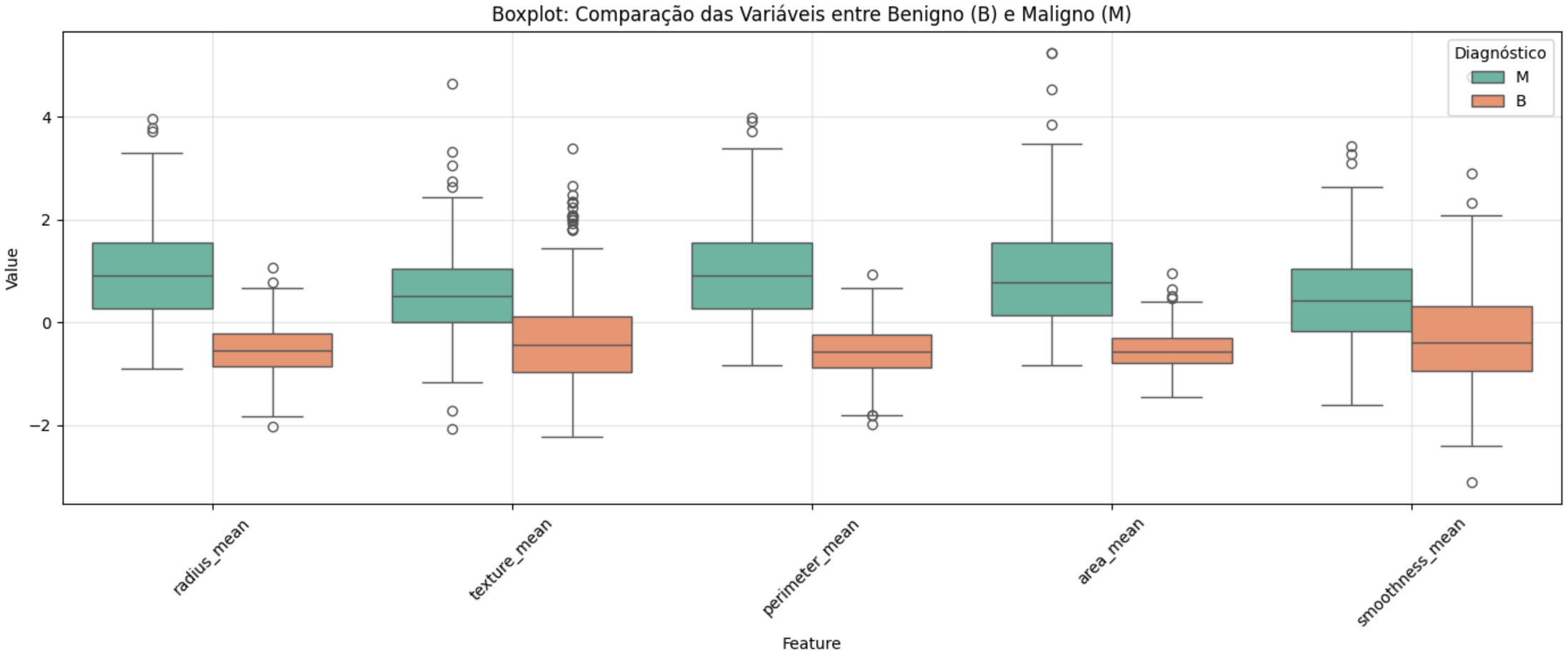




UNISINOS

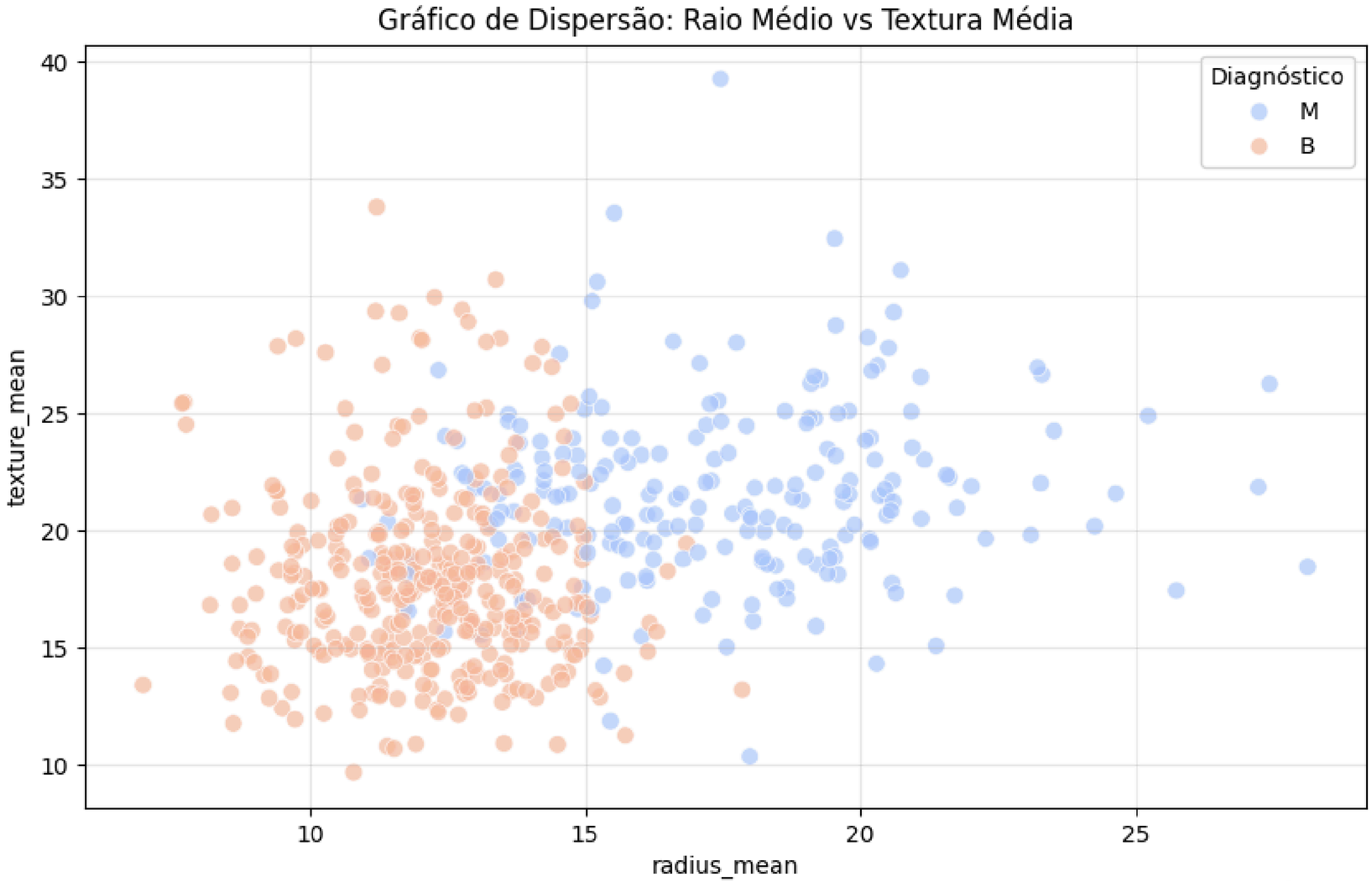
19 de Novembro, 2025

Visualização 2: Boxplots



UNISINOS

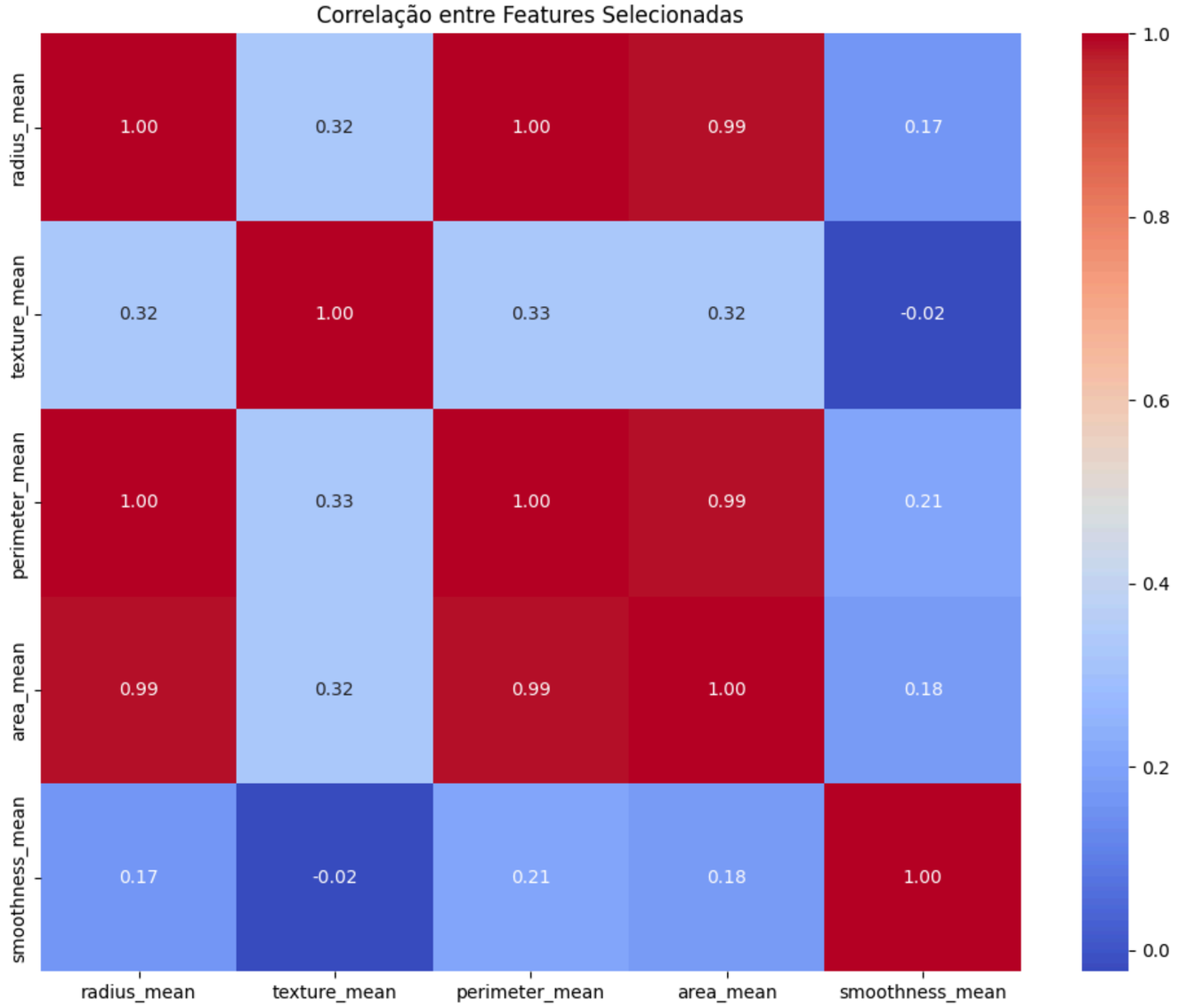
Visualização 3: Dispersão



UNISINOS

19 de Novembro, 2025

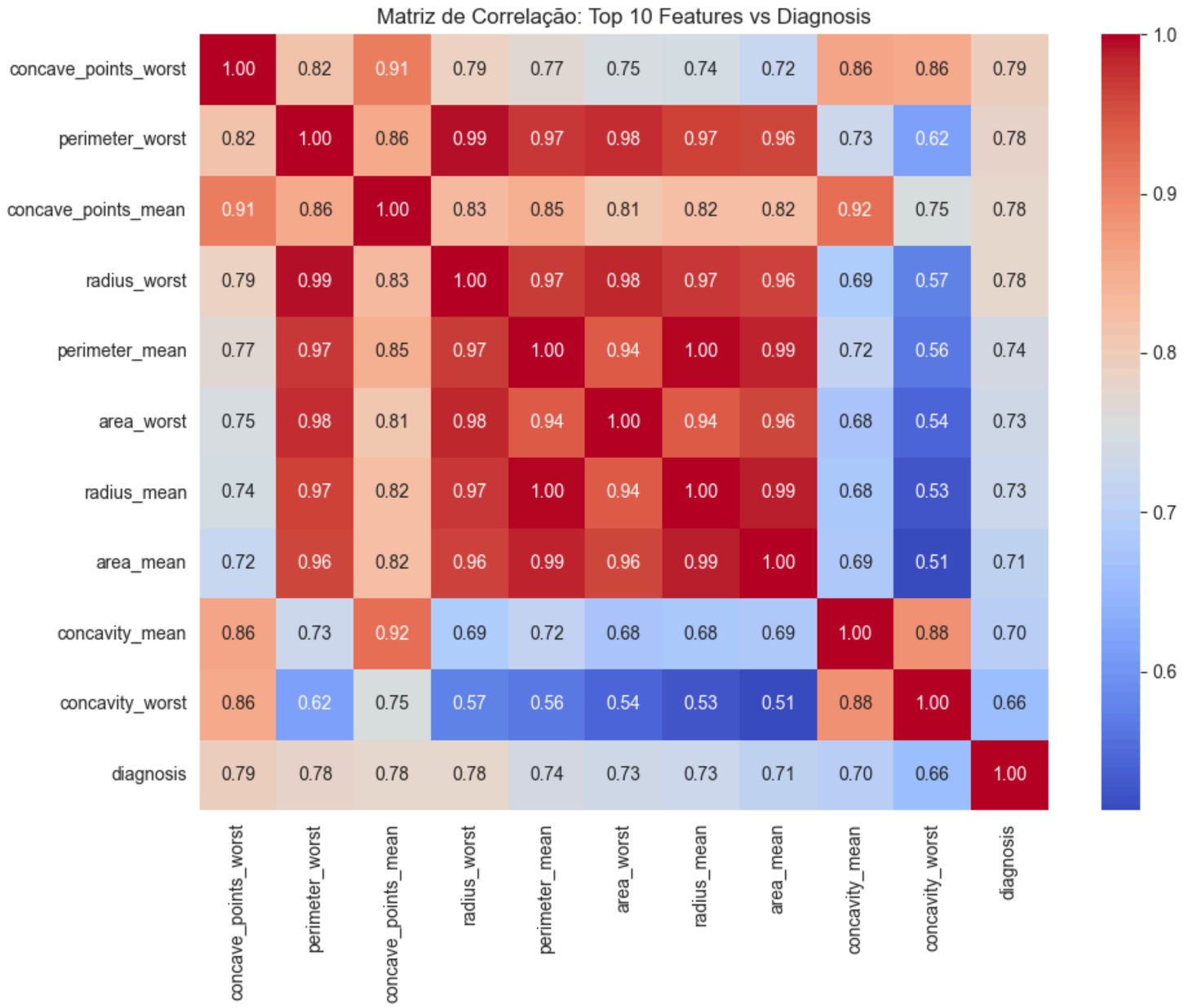
Correlação, Seleção de Atributos e Classificação das Variáveis



UNISINOS

19 de Novembro, 2025

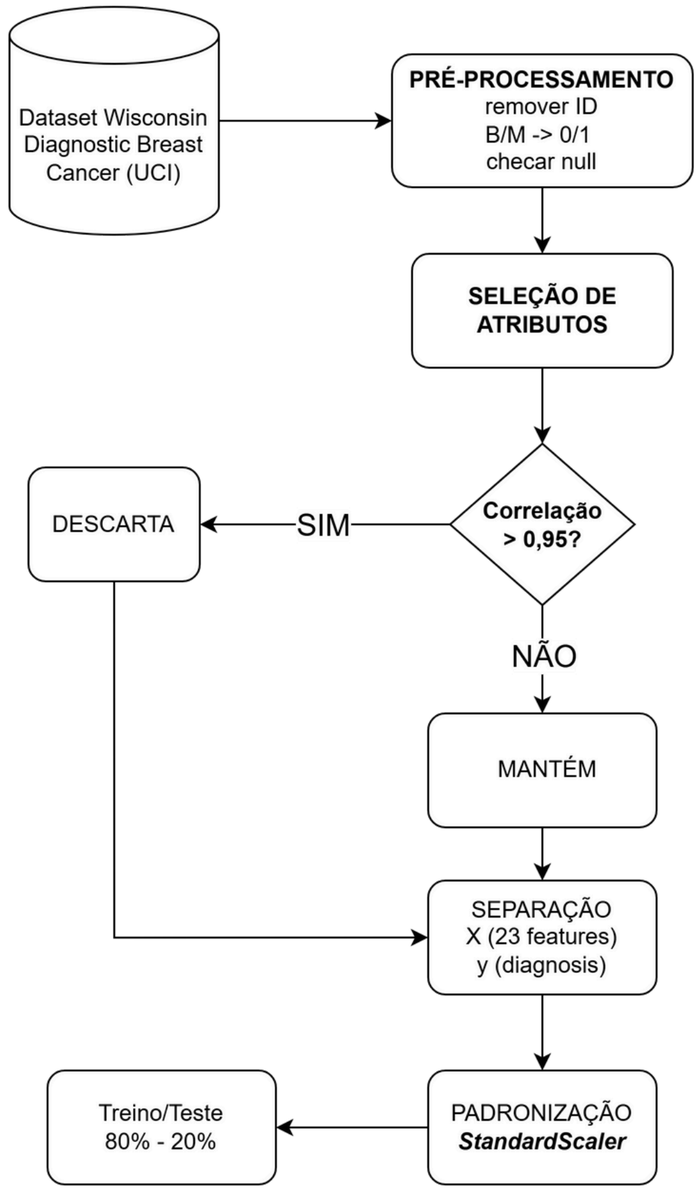
Correlação, Seleção de Atributos e Classificação das Variáveis



UNISINOS

19 de Novembro, 2025

Preparação dos Dados e Pipeline de Modelagem




```python
# 10. Separando X (preditores) e y (alvo)
print('\n=== ETAPA 10: Separação entre Atributos Preditores (X) e Variável-Alvo (y) ===')
X = df.drop(columns=['diagnosis'])
y = df['diagnosis']
print(f"Shape de X (features): {X.shape}")
print(f"Shape de y (target): {y.shape}")
print(f"Distribuição da variável alvo:\n{y.value_counts()}")

# 8. Seleção de atributos via correlação
print('\n=== ETAPA 8: Seleção de Atributos Relevantes ===')
# Calculando correlação absoluta
corr_matrix = X.corr().abs()
upper = corr_matrix.where(np.triu(np.ones(corr_matrix.shape), k=1).astype(bool))

# Selecionando features com correlação acima de um limiar (por exemplo, 0.95)
threshold = 0.95
to_drop = [column for column in upper.columns if any(upper[column] > threshold)]

print(f"Atributos altamente correlacionados a serem removidos (threshold={threshold}):")
print(to_drop if to_drop else "Nenhum atributo removido com este threshold.")

# Removendo features correlacionadas
X_reduced = X.drop(columns=to_drop)

# Exibindo dimensões após redução
print(f'\nDimensão original de X: {X.shape}')
print(f'Dimensão reduzida de X: {X_reduced.shape}')
print(f'Features removidas: {len(to_drop)}')
```

Modelos Supervisionados e Métricas Globais

Regressão Logística
- Modelo linear para classificação binária
- Bom baseline e fácil de interpretar

Random Forest
- Conjunto de várias árvores de decisão
- Captura relações não lineares e interações entre features

Modelo	Acurácia	Precisão	Recall	F1-score
Regressão Logística	0,9649	0,975	0,9286	0,9512
Floresta Aleatória	0,9474	0,9737	0,881	0,925

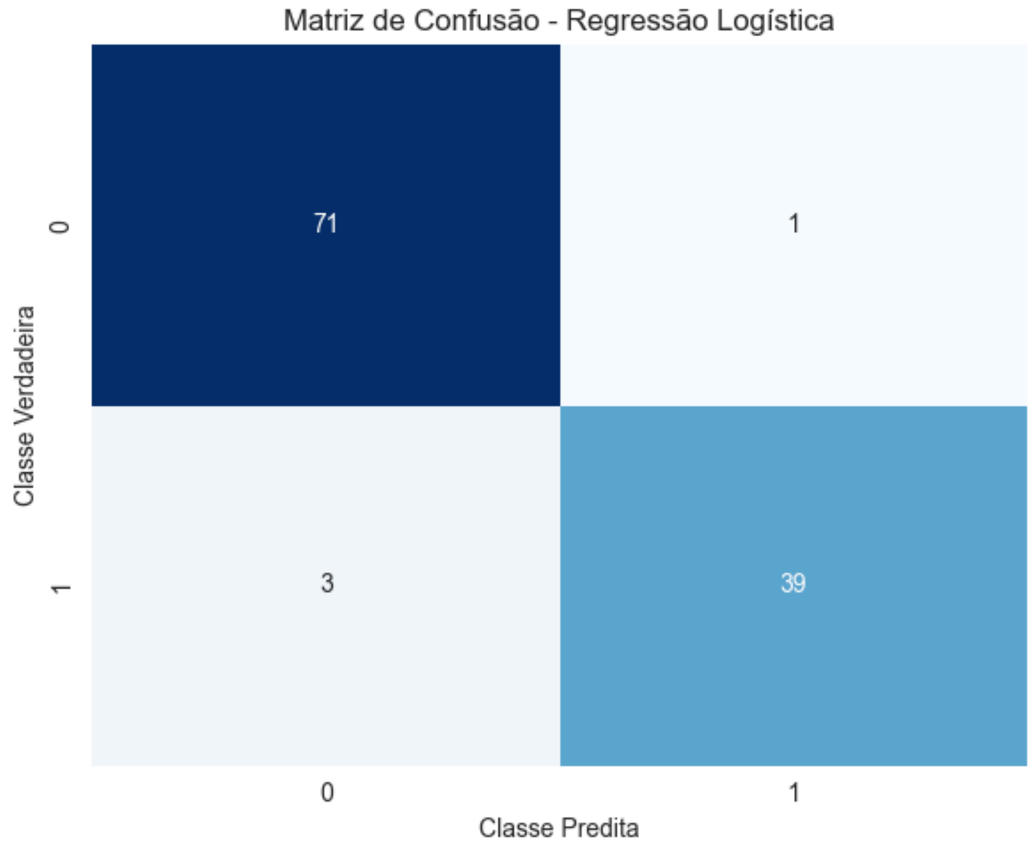
Matrizes de Confusão e Erros Críticos

0 = benigno, 1 = maligno

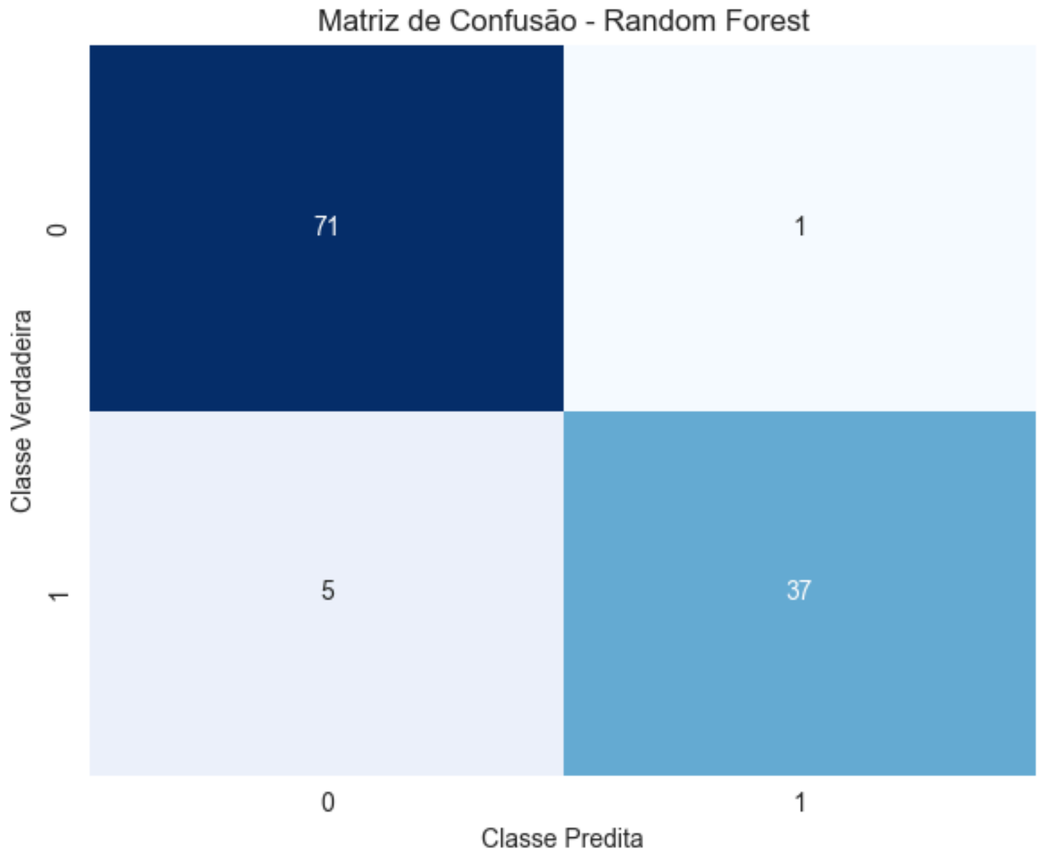
Reg. Logística: 3 malignos classificados como benignos

Em diagnósticos, falsos negativos são os erros mais críticos

Random Forest: 5 malignos classificados como benignos

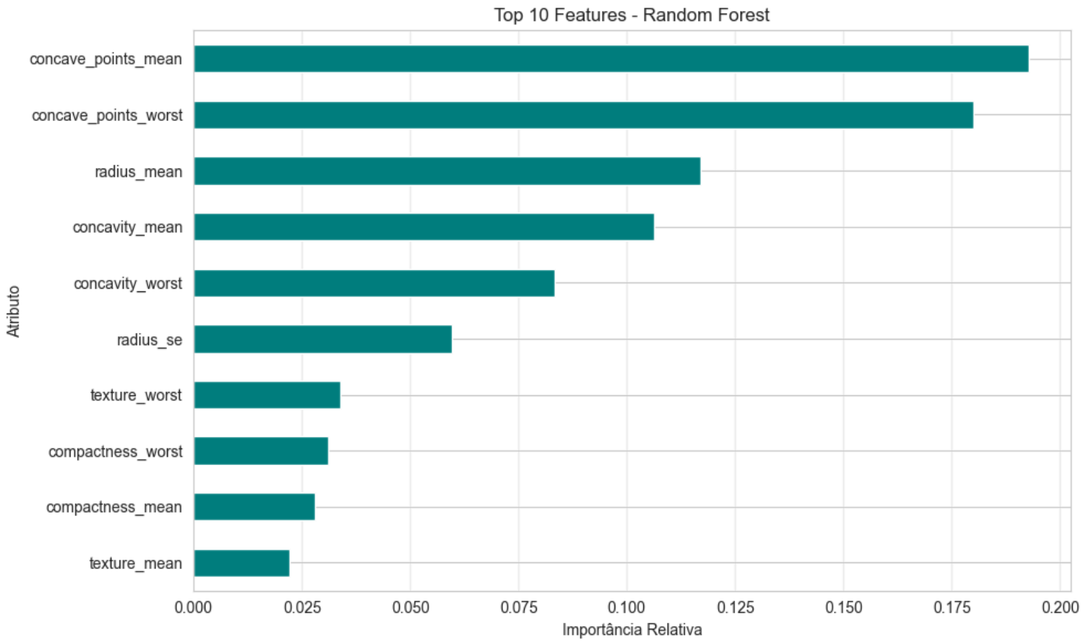


Logistic Regression: 3 falsos negativos, 1 falso positivo



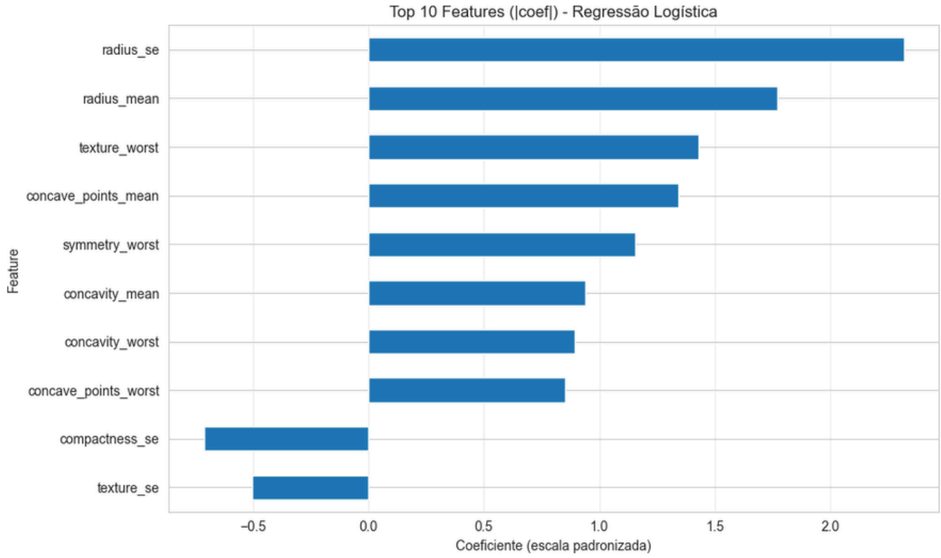
Random Forest: 5 falsos negativos, 1 falso positivo

Importância das Variáveis nos Dois Modelos



- Floresta Aleatória mostra a importância em termos de ganho de informação.
- Regressão Logística mostra o peso de cada variável na probabilidade de ser maligno (coeficientes).

- Ambos os modelos destacam: raio, concavidade e pontos côncavos como variáveis importantes.
- Em comum: tumores maiores e mais irregulares tendem a ser classificados como malignos.



UNISINOS

Conclusão

19 de Novembro, 2025

UNISINOS

Obrigado pela atenção!

UNISINOS

Nossa sala de aula é o mundo.