Predição do Diagnóstico de Câncer de Mama Utilizando Classificador Ingênuo de Bayes

ET586 - Estatística e Probabilidade

Centro de Informática - UFPE

Equipe

Alexandre Burle [aqb]

Lucas Morais [lma6]

Marco Aurélio [mafs3]

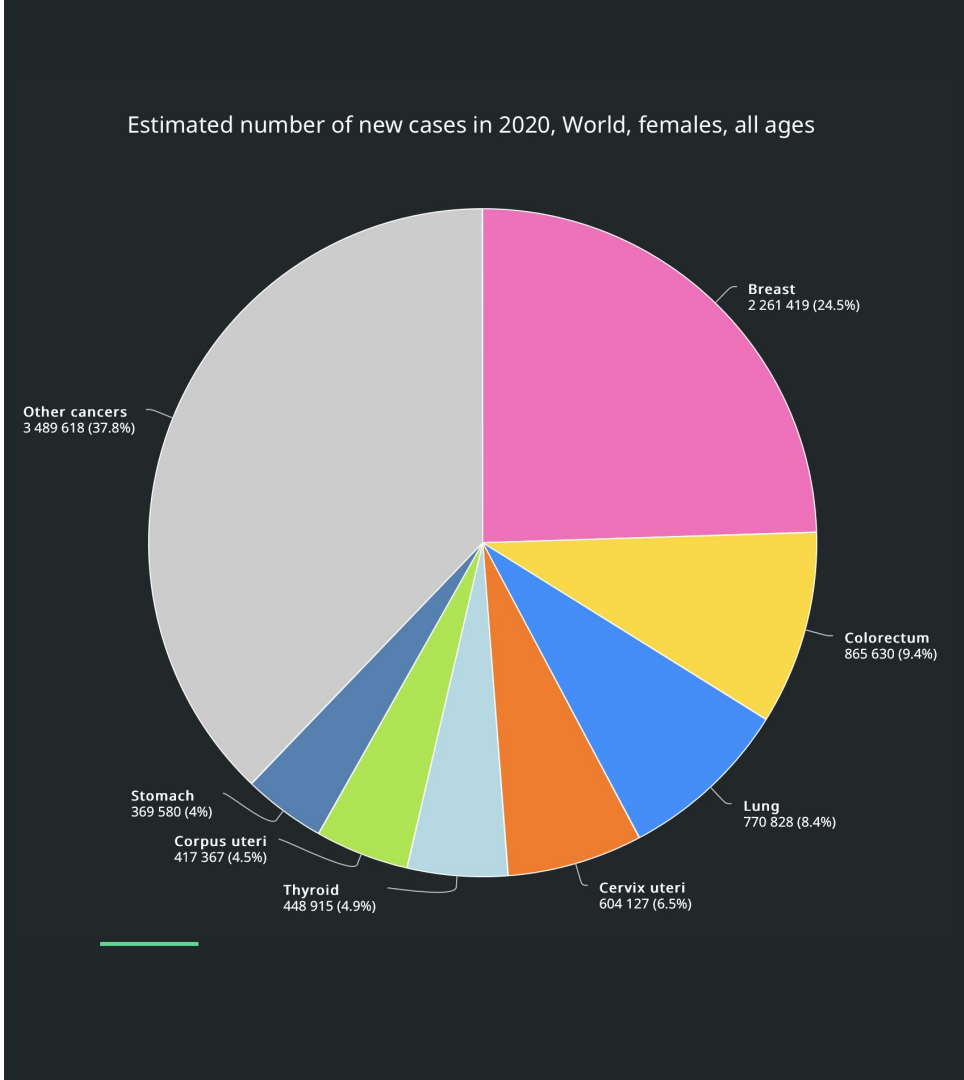
Matheus Andrade [mvtna]

Sumário

Introdução

Casos de Câncer de Mama

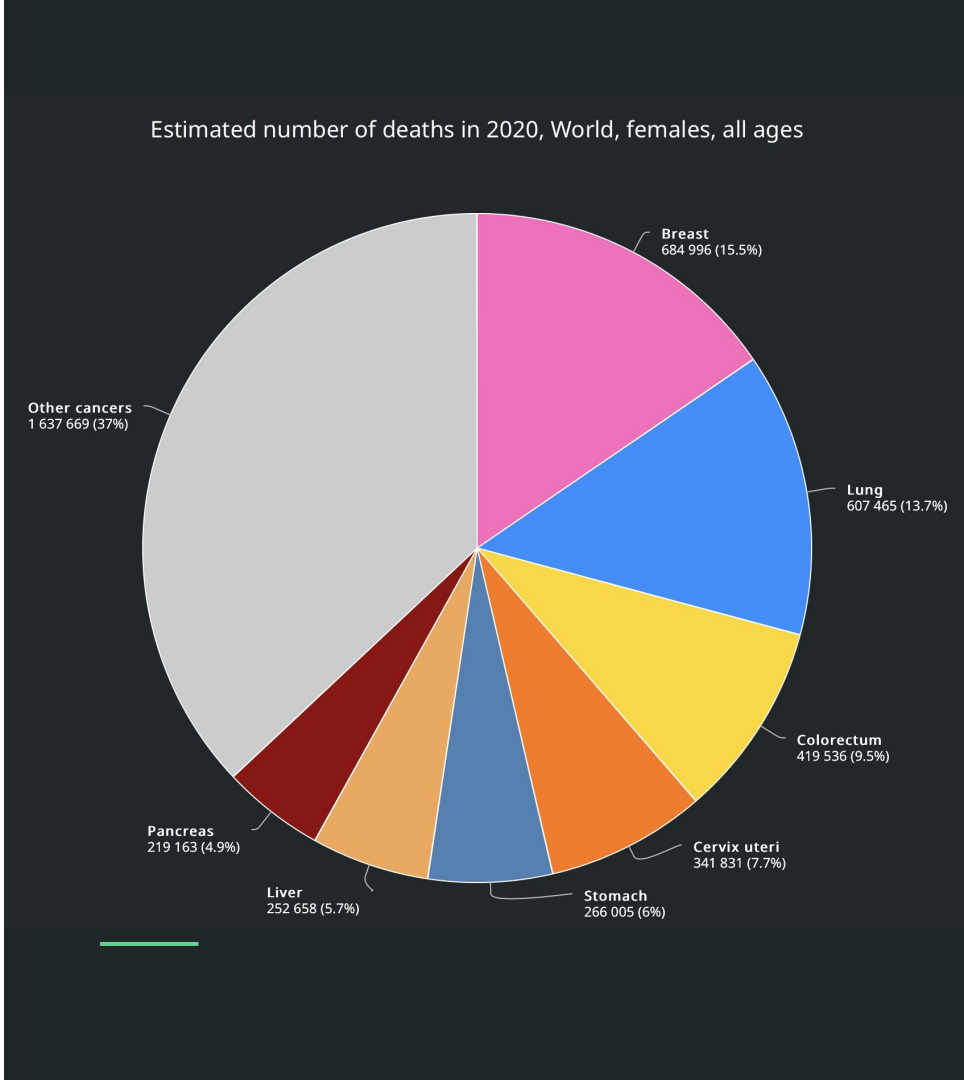
Em 2020, foram registrados cerca de 2,2 milhões de novos casos de câncer de mama ao redor do mundo para a população feminina.



Fonte: [Cancer Today - ONU]

Mortes por Câncer

A estimativa de mortes por essa doença atingiu a marca de **684.996** mulheres.



Fonte: [Cancer Today - ONU]

Sintomas

Pontos para ficar atento!

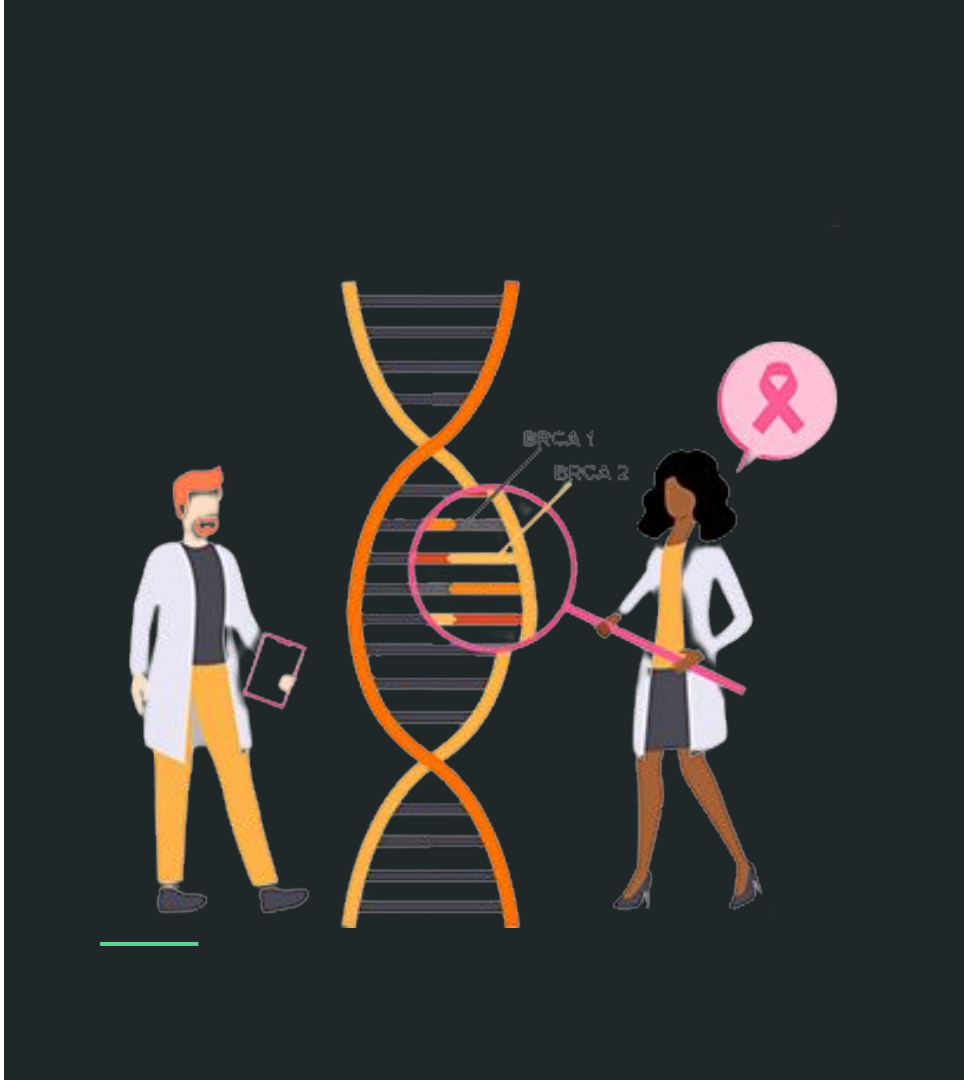
Objetivo

Prevenção

Diagnóstico Precoce e Rastreamento.

Detecção Eficiente

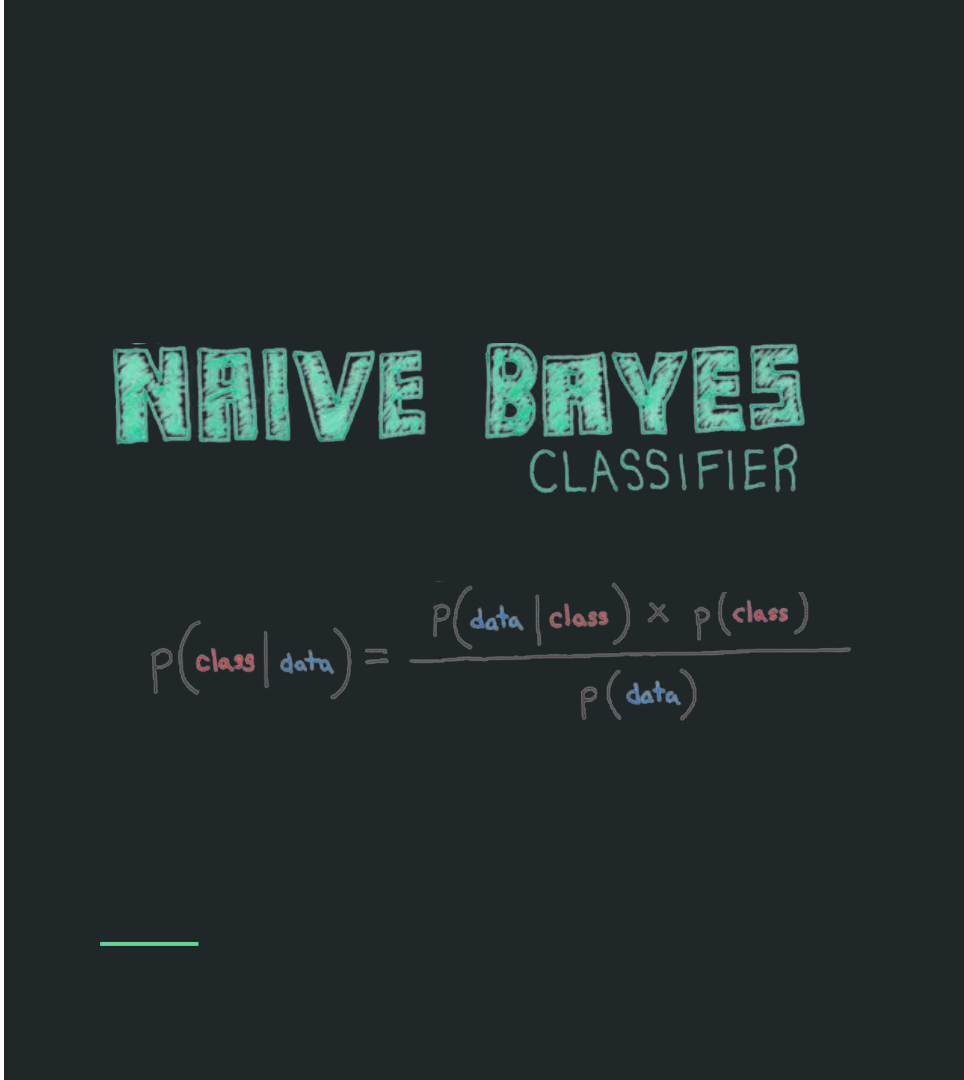
A detecção precoce do câncer de mama é crucial para o sucesso do tratamento.



Fonte: Cancer ONU Control

Objetivo

Analisar um sistema de previsão de diagnóstico de câncer de mama baseado no classificador Ingênuo de Bayes no dataset Breast Cancer Wisconsin (Diagnostic)



Metodologia

Classificador Ingênuo de Bayes

- Algoritmo de aprendizagem de máquina que utiliza a probabilidade para classificar dados em diferentes categorias;
- O Naive Bayes é um dos modelos mais conhecidos a aplicar o conceito de probabilidade e se baseia no Teorema de Bayes.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Dataset

- *Breast Cancer Wisconsin (Diagnostic) Data Set;*
- Criado por pesquisadores da Universidade de Wisconsin;
- Distribuído no repositório de aprendizagem de máquina UCI;
- O dataset é formado pelas características de imagens digitalizadas de amostras de tecido da mama obtidos por meio de aspiração por agulha fina (PAAF)
- Uso de uma árvore de decisão *Multisurface Method-Tree (MSM-T*) e programas computacionais para separação da célula em planos em um espaço tridimensional.

Dataset

Características do Conjunto de Dados	Multivariado
Quantidade de Amostras	569
Atributos (por amostra)	32
Características dos Atributos	Real e Categórico

Informações dos Atributos

Índice	Nome	Descrição	Tipo
a)	Número de identificação	ID	Inteiro
b)	Diagnóstico	Indica o resultado da doença Valores possíveis	Categórica
c)	Raio	Média das distâncias do centro aos pontos do perímetro	Número Real
d)	Textura	Desvio padrão dos valores da escala de cinza	Número Real
e)	Perímetro	Perímetro do núcleo celular	Número Real
f)	Área	Área do núcleo celular	Número Real
g)	Suavidade	Variação local em comprimentos de raio	Número Real
h)	Compacidade	$\left(\frac{perimetro^2}{area - 1.0}\right)$	Número Real
i)	Concavidade	Gravidade das porções côncavas do contorno	Número Real
j)	Pontos Côncavos	Número de porções côncavas do contorno	Número Real
k)	Simetria	Simetria do corpúsculo	Número Real
l)	Dimensão fractal	(”coastline approximation” - 1)	Número Real

Informações dos Atributos

Índice	Nome	Descrição	Tipo
a)	Número de identificação	ID	Inteiro
b)	Diagnóstico	Indica o resultado da doença Valores possíveis	Categórica
c)	Raio	Média das distâncias do centro aos pontos do perímetro	Número Real
d)	Textura	Desvio padrão dos valores da escala de cinza	Número Real
e)	Perímetro	Perímetro do núcleo celular	Número Real
f)	Área	Área do núcleo celular	Número Real
g)	Suavidade	Variação local em comprimentos de raio	Número Real
h)	Compacidade	$\left(\frac{perimetro^2}{area - 1.0}\right)$	Número Real
i)	Concavidade	Gravidade das porções côncavas do contorno	Número Real
j)	Pontos Côncavos	Número de porções côncavas do contorno	Número Real
k)	Simetria	Simetria do corpúsculo	Número Real
l)	Dimensão fractal	(”coastline approximation” - 1)	Número Real

São repetidos para cada plano do corpúsculo. Assim, temos 3 valores para cada um destes atributos.

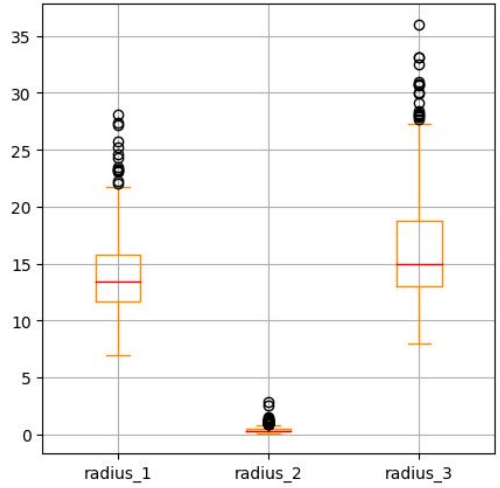
Valores Ausentes

Quantidade de Amostras	569
Benignas	357
Malígnas	212

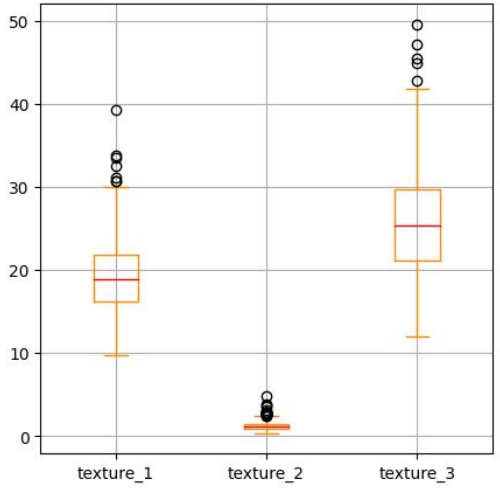
Valores Ausentes: 0

Dados transformados para float

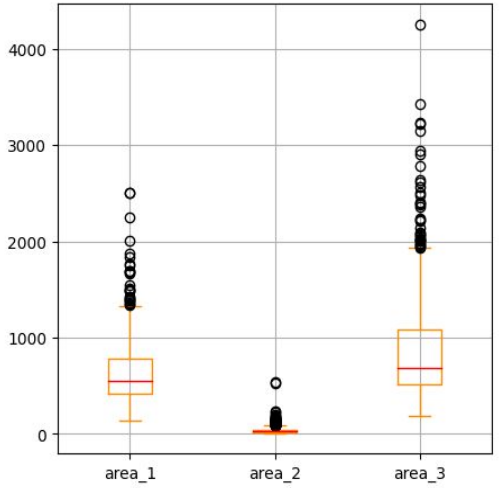
Outliers



Visualização de outliers quanto aos atributos de raio.

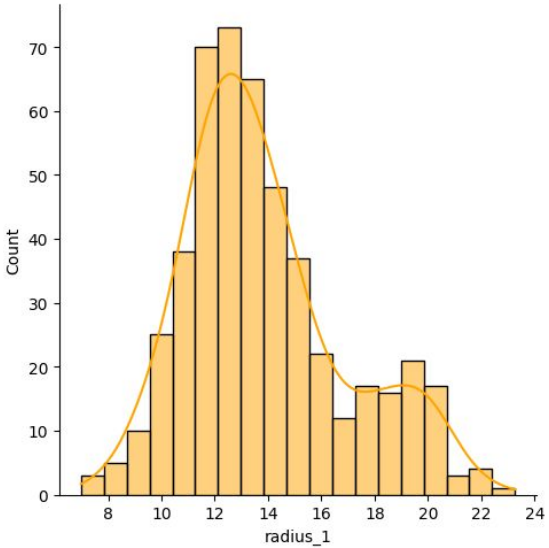


Visualização de outliers quanto aos atributos de textura.

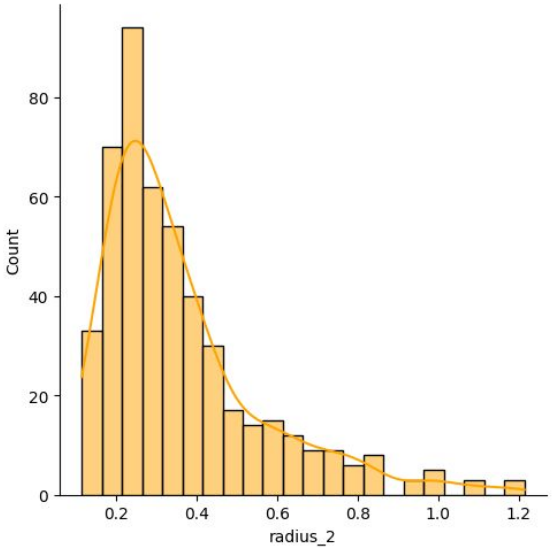


Visualização de outliers quanto aos atributos de área.

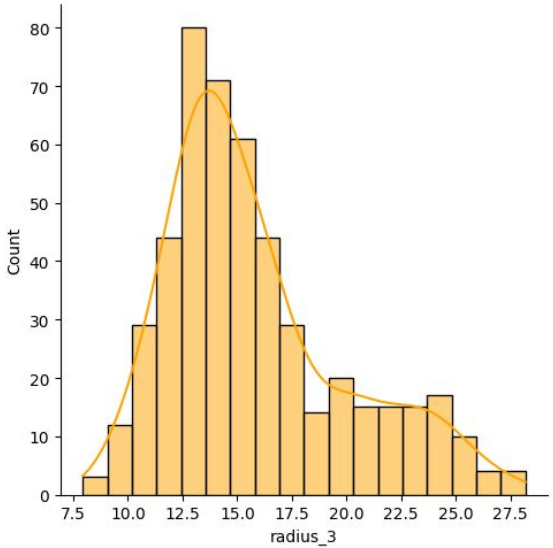
Representação Gráfica dos Dados



Visualização gráfica do atributo raio do primeiro núcleo.

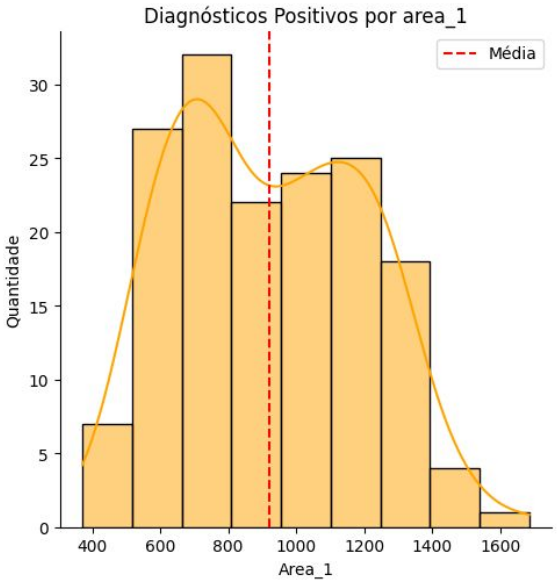


Visualização gráfica do atributo raio do segundo núcleo.

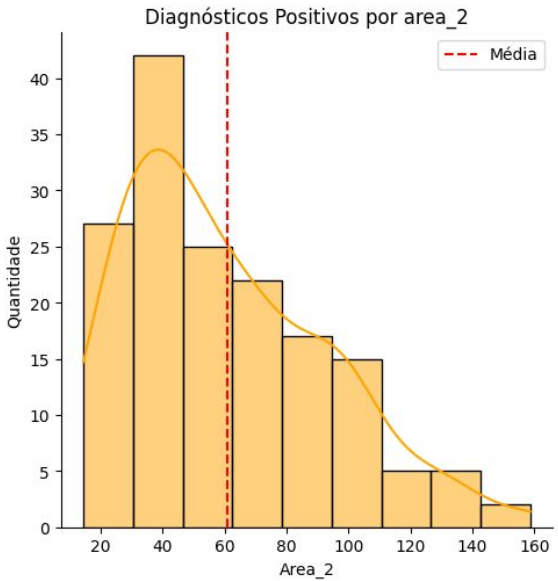


Visualização gráfica do atributo raio do terceiro núcleo.

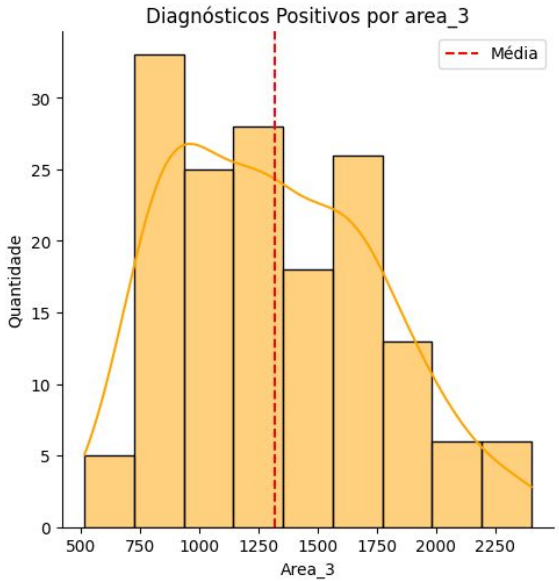
Análise Estatística



Análise estatística do atributo área do núcleo 1.



Análise estatística do atributo área do núcleo 2.



Análise estatística do atributo área do núcleo 3.

Treinamento do Modelo

- Para o treinamento do modelo, foi utilizada a biblioteca ***scikit-learn***, na linguagem ***Python***;
    - Gaussian Naive Bayes.
- Dois métodos foram aplicados durante o treinamento:
    - Estratificação: a coluna ***diagnostic*** (predição do modelo) foi escolhida para ser estratificada em alguns casos de teste;
    - Validação cruzada: 5 iterações, utilizada em todos os casos teste.

Métricas

As métricas que fornecem informações para a análise da performance do modelo serão baseadas na Matriz de Confusão do sistema.

MATRIZ DE CONFUSÃO

	Classe Prevista	
	Positivo	Negativo
Classe Real	TP	FN
	FP	TN

Sendo:
TP = Verdadeiro positivo
FP = Falso positivo
FN = Falso negativo
TN = Verdadeiro negativo

Métricas - Acurácia

- Fornece uma informação do desempenho geral do modelo;
- Mapeia a performance do sistema a partir da razão entre as predições corretas do sistema e as predições totais.

$$Acurácia = \frac{TP + \mathrm{TN}}{TP + \mathrm{TN} + FP + FN}$$

Métricas - Precisão

- Fornece uma informação referente a quantidade de predições positivas que são, de fato, positivas.

$$Precisão = \frac{TP}{TP + FP}$$

Métricas - Sensibilidade

- Fornece uma informação referente a capacidade do sistema em detectar com sucesso resultados classificados como positivos.

$$Sensibilidade = \frac{TP}{TP + FN}$$

Métricas - *f1-score*

- Fornece uma informação referente a uma média harmônica entre as métricas de sensibilidade e precisão.

$$F1 - Score = \frac{2TP}{2TP + FP + FN}$$

Métricas - *f1-score*

- Fornece uma informação referente a uma média harmônica entre as métricas de sensibilidade e precisão.

$$F1 - Score = 2 * \frac{precisão * sensibilidade}{precisão + sensibilidade}$$

Experimentos

Cenários de Teste

Quatro cenários diferentes foram testados em relação ao tratamento dos dados utilizados no treinamento:

1. Dataset contendo dados brutos;
2. Remoção do atributo ***id*** que não carrega uma informação relevante para o modelo;
3. Remoção dos atributos ***id*** e ***compacity***, o qual é obtido pelos atributos área e perímetro, o que vai no caminho inverso do Naive Bayes, o qual se baseia no fato de que os parâmetros são independentes;
4. Seleção de *features* através do algoritmo Random Forest.

Para **cada uma** das abordagens acima, foram testadas as estratégias de estratificação dos dados e balanceamento do *dataset* utilizando o algoritmo de Random Oversampling.

Resultados

Resultados com Dados Não Balanceados

Cenário	Acurácia	Precisão	F1-Score	Recall
Dados brutos não-estratificados	0.71	0.81	0.60	0.62
Dados brutos estratificados	0.78	0.88	0.67	0.66
Dados removendo *id* não estratificado	0.94	0.93	0.94	0.94
Dados removendo *id* estratificado	0.94	0.94	0.93	0.96
Dados removendo *id* e *compacity* não-estratificado	0.96	0.96	0.96	0.96
Dados removendo *id* e *compacity* estratificado	0.95	0.95	0.94	0.93
Dados balanceados, não estratificados e com seleção de features	0.95	0.94	0.94	0.95
Dados balanceados, estratificados e com seleção de features	0.94	0.94	0.93	0.92

Balanceamento

- O total de diagnósticos positivos representam menos da metade do total de diagnósticos negativos;
- Desvantagens: Pode causar overfitting e criar ruídos na classificação.
- Vantagens: Pode melhorar a precisão com classes desbalanceadas;
- Para aumentar a nossa amostra, utilizamos a classe `RandomOverSampler` da biblioteca *imbalanced-learn* fornecida pela *scikit-learn*.

Resultados com Dados Balanceados

Cenário	Acurácia	Precisão	F1-Score	Recall
Dados brutos não-estratificados	0.79	0.85	0.79	0.78
Dados brutos estratificados	0.82	0.85	0.82	0.82
Dados removendo *id* não estratificado	0.92	0.93	0.93	0.92
Dados removendo *id* estratificado	0.94	0.94	0.94	0.94
Dados removendo *id* e *compacity* não-estratificado	0.94	0.94	0.94	0.94
Dados removendo *id* e *compacity* estratificado	0.92	0.93	0.92	0.92
Dados balanceados, não estratificados e com seleção de features	0.79	0.85	0.79	0.78
Dados balanceados, estratificados e com seleção de features	0.82	0.85	0.82	0.82

Conclusão

Melhores Resultados e Performance

Considerando todos os experimentos realizados, o melhor resultado se deu com os seguintes dados:

- Dados originais (não balanceados);
- Remoção dos atributos *Id* e *Compacidade;*
- Não-estratificados.

Métrica	Acurácia Média	Desvio Padrão	Acurácia	Precisão	Recall	F1-score
Resultado	0.96	0.02	0.96	0.96	0.96	0.96

Métricas obtidas com dados originais não-estratificados sem *id* e sem *compacidade*.

Obrigado!