CENTRO UNIVERSITÁRIO DO PARÁ - CESUPA
ESCOLA DE NEGÓCIOS, TECNOLOGIA E INOVAÇÃO - ARGO
CURSO DE ENGENHARIA DA COMPUTAÇÃO

VINÍCIUS FIGUEIREDO DA SILVEIRA FARIAS
JEFFERSON BITTENCOURT AFONSO

BELÉM
2022

1. INTRODUÇÃO

O estudo da estatística é base fundamental para o desenvolvimento de pesquisas científicas. Segundo Barbetta *et al* (2010), no contexto da informática, a estatística procura obter informações pertinentes a partir da massa de dados coletados, sejam tabelas, imagens, e outros. Esse artigo vai tratar da aplicação da linguagem de programação Python na análise de dados coletados sobre o câncer de mama e tirando conclusões a partir dos resultados obtidos.

2. METODOLOGIA

Para esse artigo, foram utilizados os dados coletados por Wolberg *et al* (1995), referentes a amostras de tumores de câncer de mama coletados. A pesquisa de Wolberg *et al* (1995) analisa a eficácia de um algoritmo desenvolvido na identificação de tumores malignos e benignos. Neste artigo, será utilizado um algoritmo desenvolvido com a linguagem de programação Python para fazer a análise e tratamento dos dados disponibilizados

3. CONCEITOS FUNDAMENTAIS

Para uma melhor compreensão dos resultados que serão obtidos e analisados, deve-se, primariamente, ter em mente os seguintes conceitos da estatística, de acordo com Barbosa *et al* (2017).

A média é o valor que se obtém ao somar os valores da amostra, e dividindo pelo total desses valores. A média traz um valor de equilíbrio dentro de uma amostra, entretanto, ela é facilmente influenciada por valores muito discrepantes, por isso, a média, exclusivamente, não é confiável para se ter uma ideia geral de uma amostra, sendo necessário o acompanhamento de outros valores, como a mediana, ou então gráficos.

A mediana, diferente da média, ao invés de trazer um valor de equilíbrio, traz um valor que divide uma amostra no meio (Considerando uma amostra: 1, 2, 3, 4, 5. A mediana é o valor 3). Diferente da média aritmética, o cálculo da mediana não sofre influência por valores muito discrepantes.

Outro valor importante é o da variância, que se obtém ao fazer a diferença entre a média e a mediana, com o resultado elevado ao quadrado, e ao se tirar a raiz quadrada da variância, obtém-se o desvio padrão. O desvio padrão é um valor importante, visto que ele representa o grau de desvio de um valor da amostra em relação à média.

4. DADOS COLETADOS

Primariamente, deve-se ter noção dos valores com os quais serão trabalhados nesse artigo. A imagem abaixo mostra parte da tabela com os dados fornecidos por Wolberg *et al* (1995):

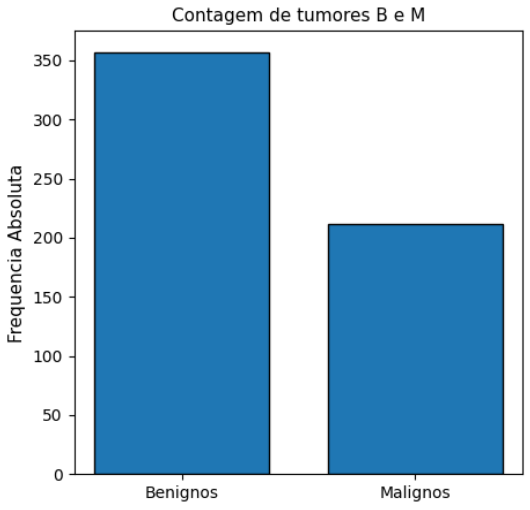
imagem 01 - tabela com valores extraídos da coleta

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.30010	0.14710	...
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690	0.07017	...
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.19740	0.12790	...
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.24140	0.10520	...
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800	0.10430	...
...	...	...	...	...	...	...	...	...	...	...	...
564	926424	M	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390	0.13890	...
565	926682	M	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791	...
566	926954	M	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302	...
567	927241	M	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140	0.15200	...
568	92751	B	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.00000	...

569 rows × 33 columns

Dos campos da tabela, será inicialmente analisado o campo *diagnosis*, que classifica um tumor como benigno ou maligno. Fazendo a filtragem, verificou-se que, das 569 amostras, 357 eram de tumores benignos e 212 de tumores malignos, plotando-se o seguinte gráfico de frequência:

imagem 02 - histograma de frequência de tumores



Além do gráfico, também foi coletada a média, mediana e desvio padrão do valor *fractal_mean* da tabela, tanto para tumores malignos, como benignos. A importância do fractal será discutida no tópico de resultados.

- Tumores malignos:
  - Média = 0.062680
  - Mediana = 0.06157
  - Variância = 0.007573
  - Desvio padrão = 0.007573
  - Valor máximo = 0.09744
  - Valor mínimo = 0.04996
- Tumores benignos:
  - Média = 0.062867
  - Mediana = 0.061540
  - Variância = 0.006747
  - Desvio padrão = 0.006747
  - Valor máximo = 0.09575
  - Valor mínimo = 0.05185

Além da coleta desses valores, foi feita, também, a plotagem dos gráficos de valor de fractal para os tumores usando, respectivamente, todos os valores de fractal da tabela, apenas os de tumores benignos, e apenas os de tumores malignos:

imagem 03 - histograma de todos os valores de fractal

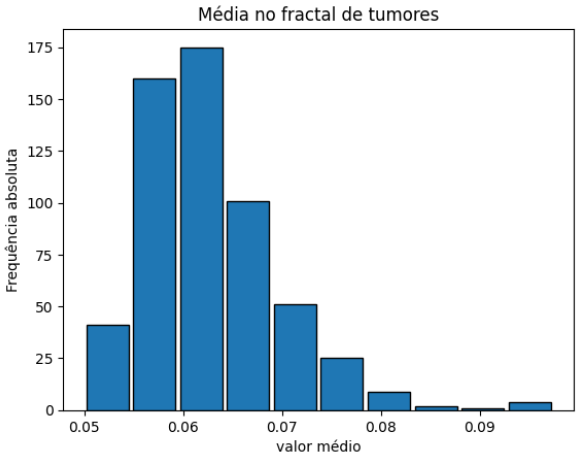


imagem 04 - diagrama de caixa mostrando a média com todos os valores de fractal

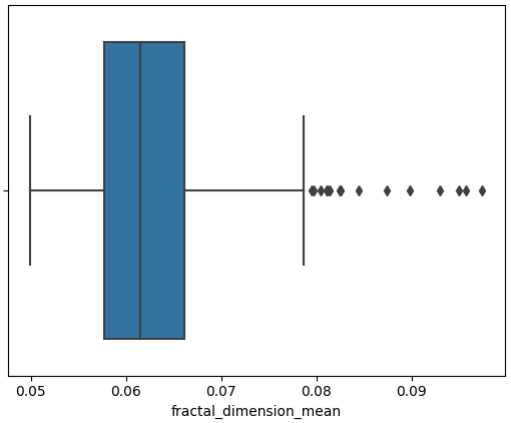


imagem 05 - histograma de frequência para todos os valores de fractal

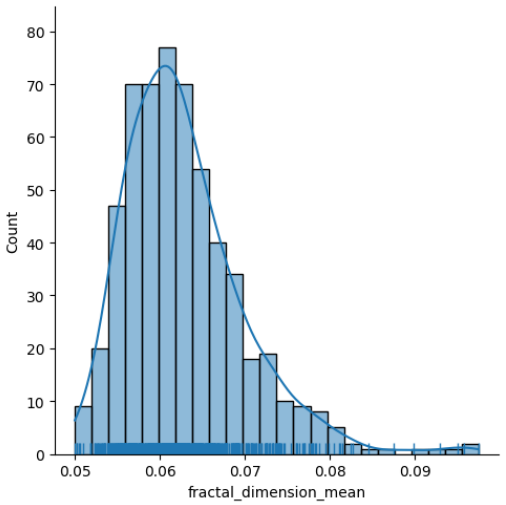


imagem 06 - histograma dos valores de fractal para tumores benignos

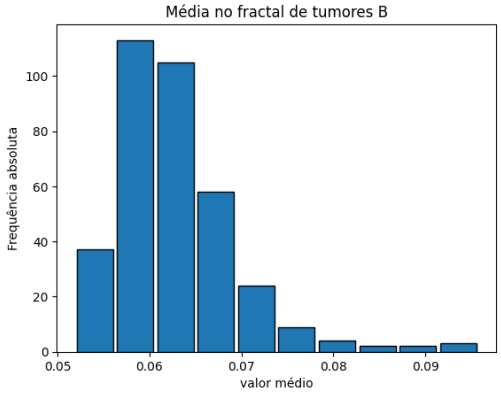


imagem 07 - diagrama de caixa mostrando a média para fractal de tumores benignos

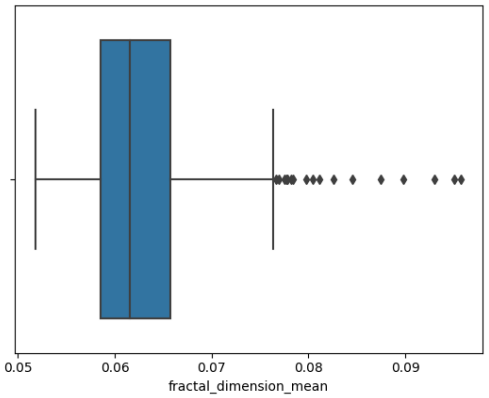


imagem 08 - histograma de frequência para os valores de fractal de tumores benignos

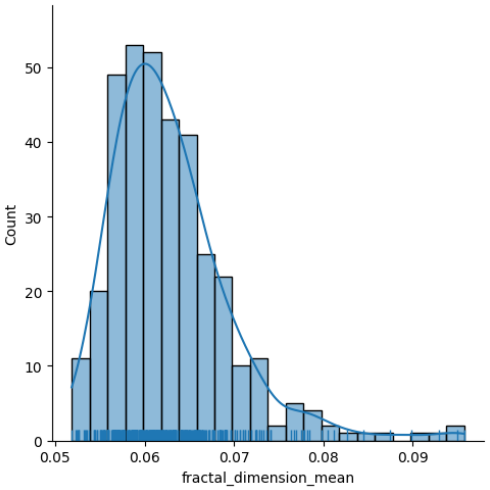


imagem 09 - histograma dos valores de fractal para tumores malignos

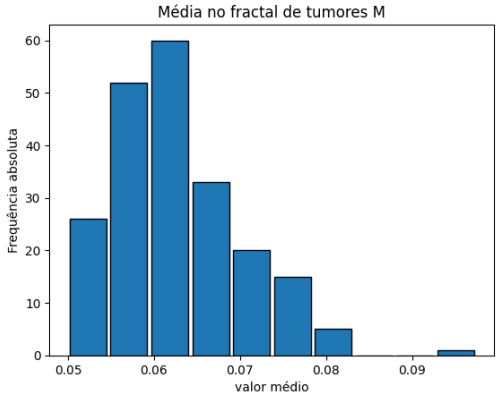


imagem 10 - diagrama de caixa mostrando a média para fractal de tumores malignos

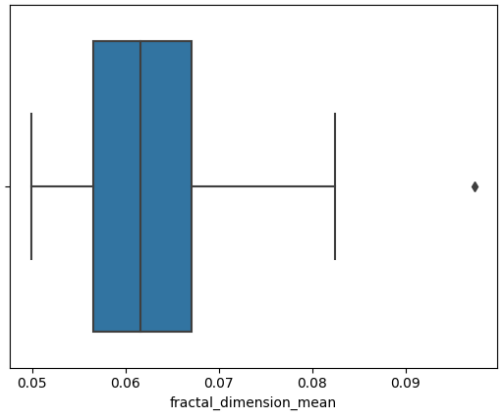
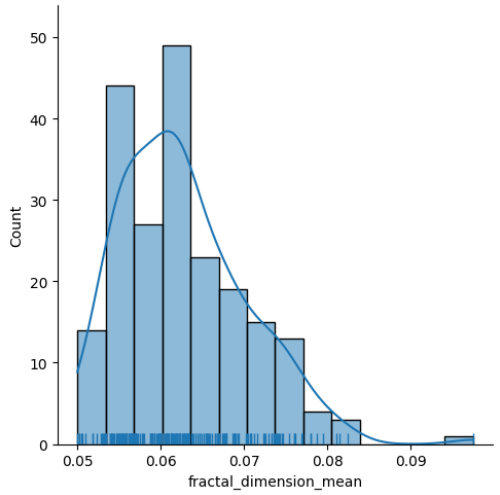


imagem 11 - histograma de frequência para os valores de fractal de tumores malignos



5. RESULTADOS

Para fazer uma análise dos resultados, tem que se ter em mente o significado do valor do fractal. Segundo Wolberg *et al* (1995), o valor do fractal determina o nível de deformação da célula cancerígena, quanto maior o fractal, maior esse nível de deformação. Analisando a imagem 07, percebe-se que para os tumores benignos, há uma quantidade considerável de valores discrepantes afetando a média. Considerando a definição do fractal para esse estudo, esses valores muito fora da curva poderiam significar células cancerígenas com uma dificuldade de tratamento maior, todavia, considerando a distribuição das frequências no gráfico 08, percebe-se que maioria dos tumores benignos estão dentro da média, sendo poucos os casos que o fractal tende a um valor mais extremo. Quando essa análise passa para os tumores malignos, percebe-se uma concentração maior dos valores nas faixas de 0.053 à 0.056 e 0.06 à 0.063, com apenas um valor fora da curva. Os valores de fractal para os tumores malignos estão mais bem concentrados do que os de tumores benignos, com média e mediana mais próximas, entretanto, apesar de eles não possuírem tanta variação no fractal como os benignos, podem haver outros fatores de ordem médica e biológica que, associados ao valor do fractal, possam facilitar na identificação desses tumores. Com isso, verifica-se a possibilidade da aplicação da linguagem de programação Python em trabalhos que envolvam análise de dados, tornando-a uma ferramenta essencial para profissionais que desejem trabalhar com coleta e tratamento de dados, permitindo facilidade de visualização e interpretação das informações processadas.

REFERÊNCIAS BIBLIOGRÁFICAS

BARBETTA, Pedro Alberto; REIS, Marcelo Menezes; BORNIA, Antonio Cezar. **Estatística para cursos de engenharia e informática**. 3. ed. São Paulo: Atlas, 2010.

SPIEGEL, Murray R.; STEPHENS, Larry J.. **Estatística**. 4. ed. Porto Alegre: Bookman, 2009.

STREET, W. Nick; WOLBERG, William H.; MANGASARIAN, O. L.. Nuclear Feature Extraction For Breast Tumor Diagnosis. **International Symposium On Electronic Imaging: Science And Technology**, San Jose, v. 1905, p. 861-870, 28 dez. 1992. Disponível em: http://rexa.info/paper/b98475235164960529ad2ff9fda3816e9335cf8a. Acesso em: 17 set. 2022.

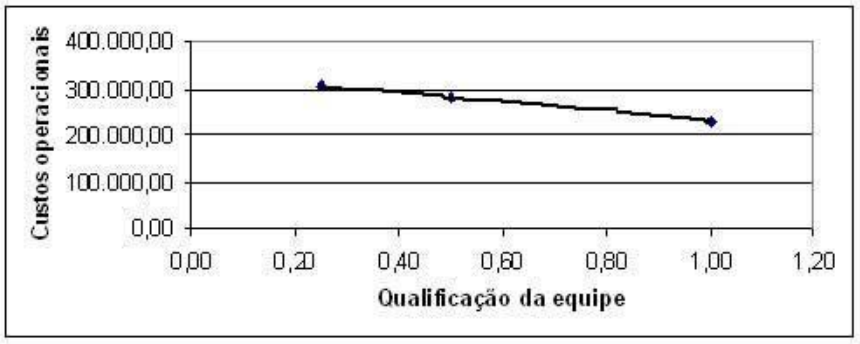
ELEMENTOS GRÁFICOS

Abaixo são apresentados exemplos de como figuras, gráficos e tabelas deverão ser inseridos ao longo do texto do **TRABALHO DE CURSO**. As figuras devem ser formatadas com espaçamento simples e texto e figura centralizados na página. A legenda deve ser inserida acima da figura e a fonte abaixo, ambas com fonte tamanho 10.

FIGURAS

As figuras devem sempre ser referenciadas no texto, de preferências, antes de serem apresentadas, e o mesmo vale para quadros e tabelas. Nunca se deve começar ou terminar um capítulo, seção ou subseção com figuras/tabelas/quadro.

Figura 1- Exemplo de figura



Fonte: Adaptado de Mays *apud* Greenhalg (1997)

TABELAS

Para formatação de tabelas sempre utilize fonte tamanho 10, espaçamento simples e texto e tabelas centralizados na página. Insira a legenda acima da tabela e a fonte abaixo, ambas com fonte tamanho 10. A tabela não pode ser fechada nas laterais e deve conter números.

Tabela 1 - Exemplo de tabela

Item	Quantidade	Percentual
Teoria social	22	7,9%
Método	34	12,3%
Questão	54	19,5%
Raciocínio	124	44,8%
Método de amostragem	33	11,9%
Força	10	3,6%

Fonte: Adaptado de Mays *apud* Greenhalg (1997).

QUADROS

Os quadros devem ser formatados utilizando fonte tamanho 10, espaçamento simples, e centralizadas, assim como o texto. Inserir legenda acima do quadro e a fonte abaixo, ambas com fonte tamanho 10. O quadro deve ser fechado nas laterais e deve conter texto.

Quadro 1 - Exemplo de quadro

Tema	**Autor**
Teoria social	Fulano
Método	Ciclano
Questão	Beltrano
Raciocínio	João

Fonte: Simons (2007).