SMARKIO - Teste Prático Ciência de Dados

Isabela Corsi

1.

Estatística descritiva

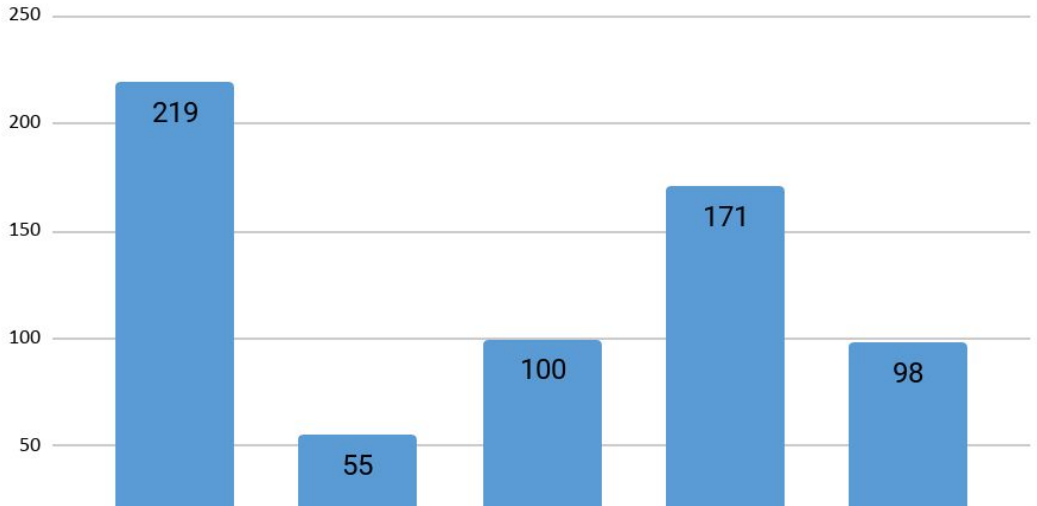
obs: todos os dados estatisticos retirados pelo Excel, no arquivo “Analise_ml_estatistica”, que está na pasta “Arquivos de Dados Excel”.

	Pred_class	probabilidade	True_class
MÁXIMO	118	1	118
MÍNIMO	2	0,04	0
AMPLITUDE	116	0,96	118
LARGURA DE CLASSE	24	0,19	23,6
MÉDIA	52,71228616	0,62	48,251944
MODA	3	1,00	74
MEDIANA	59	0,62	55
DESVIO PADRÃO	37,57281728	0,27	38,54226
TAM (n)	643	643	643
25%	12	0,41	3
50%	59	0,62	55
75%	81	0,04	77

A seguir, segue a tabela de frequência de **Pred_class**, assim como seu gráfico,

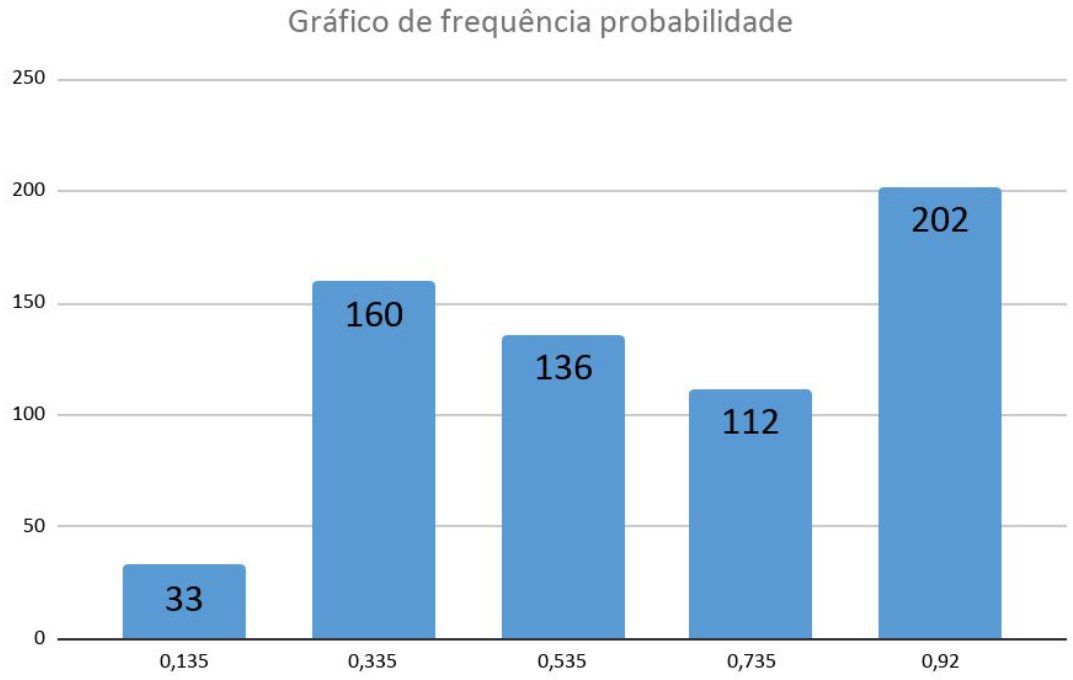
Tabela de Frequência Pred_class							
	Classes			Frequência	Ponto Médio	Frequência Relativa	Frequência Acumulada
1	2	-	25	219	13,5	0,34	219
2	26	-	49	55	37,5	0,09	274
3	50	-	73	100	61,5	0,16	155
4	74	-	97	171	85,5	0,27	271
5	98	-	121	98	109,5	0,15	269

Gráfico das frequências Pred_class



A seguir, segue a tabela de frequência de **probabilidade**, assim como seu gráfico,

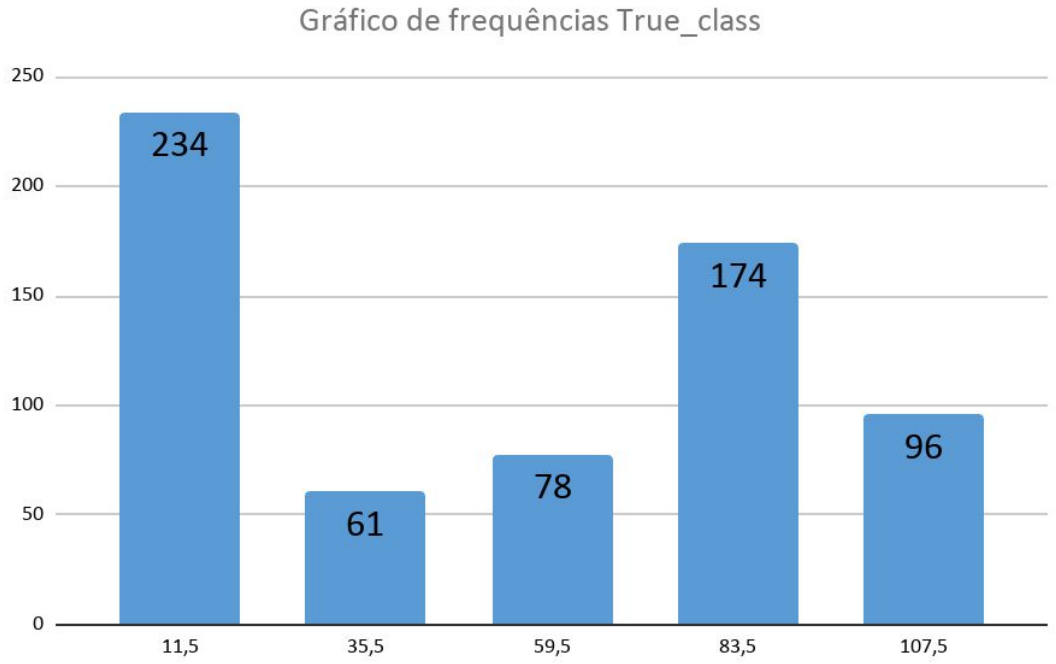
Tabela de Frequência probabilidade							
	Classes			Frequência	Ponto Médio	Frequência Relativa	Frequência Acumualda
1	0	-	23	234	11,5	0,36	234
2	24	-	47	61	35,5	0,09	295
3	48	-	71	78	59,5	0,12	373
4	72	-	95	174	83,5	0,27	547
5	96	-	119	96	107,5	0,15	643



A seguir, segue a tabela de frequência de **true_class**, assim como seu gráfico,

Tabela de Frequência true_class							
	Classe			frequência	ponto médio	Frequência Relativa	Frequência Absoluta
1	0,04	-	0,23	33	0,135	0,051	33
2	0,24	-	0,43	160	0,335	0,249	193
3	0,44	-	0,63	136	0,535	0,212	329

4	0,64	-	0,83	112	0,735	0,174	441
5	0,84	-	1	202	0,92	0,314	643



Estatística Inferencial - Teste de Hipótese

Selecionada uma amostra de 186 elementos, separou-se os dados de modo que quando os elementos na coluna pred_class coincidissem com aqueles da coluna true_class, tal acerto era contabilizado em uma variável sucesso. Obteve-se, dessa amostra, 133 sucessos, ou seja, 71,5% de acertos.

Pretende-se testar, dessa forma, se na população de 643 elementos é aceitável esperar uma taxa de sucesso de pelo menos 75%, a um nível de significância de 5%. É feita a aproximação de que os dados amostrais se comportam como uma distribuição normal. Para o teste, foram definidas as seguintes hipóteses:

$H_0 : p_0 = 0,75$ (Hipótese nula)

$H_1 : p_0 \neq 0.75$ (HIpótese alternativa)

A equação (1) refere-se à estatística teste utilizada, considerando-se n = 186, p = 0,715 e p_0 = 0,75.

$$Z_0 = (p - p_0)/\sqrt{p_0(1-p_0)/n} \qquad (1)$$

Para um nível de significância de 5%, obtém-se da tabela padronizada $Z_{0,025} = 1,96,$ e a estatística teste resulta em Z_0 = -1,09. Como $Z_0 > -Z_{0,025}$, conclui-se que a hipótese nula é aceita, pois a estatística teste se encontra fora da região crítica do teste de hipóteses. Sendo assim, uma taxa de acertos de 75% para a população é uma estimativa realista.

.

---

2. O Desempenho do classificador foi avaliado conforme as seguintes métricas:

- ***Accuracy* :** calcula a precisão do subconjunto, ou seja, o conjunto de rótulos previsto para uma amostra deve corresponder exatamente ao conjunto correspondente de rótulos reais.

  Resultado : `71.85%`

- ***Precision* :** calcula a capacidade do classificador de não rotular como positivo uma amostra que na verdade é negativa, por meio da fórmula:

$$\frac{Verdadeiro\ Positivo}{Verdadeiro\ Positivo + falso\ positivo}$$

  Além disso, foi avaliado o precision com average micro e macro. O primeiro calcula métrica global contando todos os Verdadeiros Positivos, Falsos Negativos e Falsos Positivos. E o macro calcula a métrica para cada classe, e depois, obtém a média.

  Resultado com average = 'micro': `71.85%`
  Resultado com average = 'macro': `70%`

- **Recall:** calcula capacidade do classificador de encontrar as amostras positivas, por meio da fórmula:

$$\frac{Verdadeiro\ Positivo}{Verdadeiro\ Positivo + falso\ negativo}$$

  Resultado com average = 'micro': `71.85%`
  Resultado com average = 'macro': `63%`

- **Classification_report:** mostra as principais as métricas de cada classe, conforme a figura abaixo. Sendo que F1_score é dada pela fórmula:

$$\frac{2 * (precision * recall)}{(precision + recall)}$$

E o Support é o número de ocorrências de determinada classe no teste.

	precision	recall	f1-score	support
0.0	0.00	0.00	0.00	0
2.0	0.77	0.77	0.77	61
3.0	0.83	0.79	0.81	63
4.0	0.86	0.78	0.82	23
11.0	1.00	0.44	0.62	9
12.0	0.71	0.83	0.77	6
15.0	0.67	0.67	0.67	3
17.0	0.75	0.86	0.80	7
19.0	0.40	0.40	0.40	5
21.0	0.00	0.00	0.00	1
22.0	0.91	0.67	0.77	15
24.0	0.62	0.71	0.67	14
25.0	1.00	0.83	0.91	12
26.0	0.50	0.33	0.40	3
28.0	1.00	0.50	0.67	2
29.0	1.00	1.00	1.00	7
30.0	1.00	0.60	0.75	5
31.0	0.00	0.00	0.00	2
32.0	0.50	0.25	0.33	8
33.0	0.00	0.00	0.00	3
36.0	1.00	1.00	1.00	1
39.0	1.00	0.67	0.80	6
40.0	0.78	1.00	0.88	7
43.0	1.00	0.50	0.67	6
46.0	1.00	1.00	1.00	1
48.0	0.25	0.50	0.33	2
49.0	0.00	0.00	0.00	2
50.0	0.00	0.00	0.00	2
52.0	1.00	0.30	0.46	20
54.0	1.00	1.00	1.00	2
55.0	0.93	0.82	0.87	17
56.0	1.00	1.00	1.00	3
58.0	0.50	1.00	0.67	1
59.0	1.00	0.25	0.40	4
60.0	0.89	0.81	0.85	31
62.0	0.60	0.75	0.67	4
63.0	1.00	1.00	1.00	2
64.0	0.00	0.00	0.00	1
65.0	0.50	0.33	0.40	3
66.0	0.00	0.00	0.00	1
68.0	1.00	0.67	0.80	3
69.0	1.00	1.00	1.00	1

70.0	0.75	1.00	0.86	3
73.0	0.50	0.50	0.50	2
74.0	0.72	0.95	0.82	59
76.0	0.80	1.00	0.89	8
77.0	0.83	0.77	0.80	31
78.0	1.00	1.00	1.00	3
79.0	0.43	1.00	0.60	3
81.0	0.33	0.60	0.43	5
82.0	1.00	1.00	1.00	5
84.0	1.00	1.00	1.00	1
85.0	0.60	0.43	0.50	14
86.0	0.20	0.33	0.25	3
87.0	0.67	1.00	0.80	4
88.0	1.00	1.00	1.00	3
90.0	1.00	0.67	0.80	3
92.0	0.50	0.20	0.29	5
93.0	1.00	1.00	1.00	1
94.0	1.00	1.00	1.00	1
95.0	0.00	0.00	0.00	1
96.0	0.90	0.90	0.90	21
98.0	0.71	1.00	0.83	5
99.0	0.80	0.57	0.67	14
100.0	1.00	1.00	1.00	1
102.0	0.62	1.00	0.77	5
103.0	1.00	0.67	0.80	9
104.0	1.00	0.25	0.40	4
105.0	0.00	0.00	0.00	1
106.0	1.00	0.25	0.40	4
107.0	1.00	1.00	1.00	1
108.0	0.90	0.69	0.78	13
109.0	0.00	0.00	0.00	4
110.0	1.00	0.69	0.81	16
111.0	1.00	0.33	0.50	3
112.0	0.67	0.50	0.57	4
113.0	1.00	1.00	1.00	1
114.0	1.00	1.00	1.00	1
115.0	0.80	0.80	0.80	5
116.0	0.67	1.00	0.80	2
117.0	0.00	0.00	0.00	0
118.0	1.00	0.40	0.57	5
accuracy			0.72	643
macro avg	0.70	0.63	0.64	643
weighted avg	0.80	0.72	0.73	643

4.O Desempenho do classificador do item 3 foi avaliado conforme as seguintes métricas:
O dados de teste se deram da seguinte forma,comparando o previsto com o real:

Valor real	Valor previsto
4	4
43	43
2	2
25	25
84	86
96	96
3	3
113	114
55	55
4	4
3	3
39	39
77	77

- ***Accuracy:*** 84.62%
- ***Precision:***
  com average = ‘micro’ : 84.6154%
  com average = ‘macro’: 69%
- ***Recall:***
  com average = ‘micro’ : 84.6154%
  com average = ‘macro’: 69%
- ***F1_Score:***
  com average = ‘micro’ : 84.6154%
  com average = ‘macro’: 69%

- ***Classification_report:***

	precision	recall	f1-score	support
2.0	1.00	1.00	1.00	1
3.0	1.00	1.00	1.00	2
4.0	1.00	1.00	1.00	2
25.0	1.00	1.00	1.00	1
39.0	1.00	1.00	1.00	1
43.0	1.00	1.00	1.00	1
55.0	1.00	1.00	1.00	1
77.0	1.00	1.00	1.00	1
84.0	0.00	0.00	0.00	0
86.0	0.00	0.00	0.00	1
96.0	1.00	1.00	1.00	1
113.0	0.00	0.00	0.00	0
114.0	0.00	0.00	0.00	1
accuracy			0.85	13
macro avg	0.69	0.69	0.69	13

- **Matriz de confusão:**