CADERNO DE EXERCÍCIOS – PARTE 02

4 CAPÍTULO 04: AGRUPAMENTO DE DADOS

4.1 EXERCÍCIOS CONCEITUAIS

4.1.1 Qual é a definição e o objetivo da tarefa de agrupamento de dados?

Resposta: A tarefa de agrupamento de dados tem por objetivo descobrir grupos homogêneos de objetos utilizando métodos numéricos de análise de dados multivariados.
Ela pode ser definida como a organização de um conjunto de objetos, normalmente representados por pontos em um espaço multidimensional em grupos baseada na similaridade entre eles.

4.1.2 A avaliação da saída de um algoritmo de agrupamento de dados pode determinar a qualidade do agrupamento resultante. Para isso utilizamos medidas de avaliação de desempenho que são responsáveis por aferir quantitativamente o agrupamento resultante. As medidas podem ser categorizadas em dois tipos. Quais são os tipos de medidas de avaliação e como eles funcionam?

Resposta: Os dois tipos de medidas são: internas e externas.

- **Internas:** são medidas que utilizam apenas informações intrínsecas aos objetos do agrupamento baseando-se em medidas de similaridade e avaliando as distâncias intragrupos e/ou intergrupos.
- **Externas:** são medidas que avaliam quão correto está um agrupamento dado um agrupamento ideal que se deseja alcançar. O cálculo dessas medidas requer o conhecimento prévio do grupo ao qual cada objeto pertence.

4.1.3 Existem diversos algoritmos de agrupamento disponíveis, mas de forma abrangente estes algoritmos podem ser divididos em três categorias. Quais são estas categorias e suas características?

Resposta: As três categorias são: hierárquicos, particionais e não exclusivos. Abaixo estão as características de cada um.

- **Hierárquicos:** os métodos hierárquicos criam uma decomposição hierárquica dos dados. Esses métodos podem ser aglomerativos ou divisivos, baseados em como o processo de decomposição é efetuado.
- **Particionais:** a partir de um número n de partições, o método constrói k partições dos dados, sendo cada partição representa um cluster onde $k \leq n$.
- **Não Exclusivos:** esse método permite que um objeto pertença completamente ou parcialmente a mais de um grupo ao mesmo tempo.

4.2 EXERCÍCIOS NUMÉRICOS

4.2.1 Para a base de dados apresentada na tabela abaixo execute dois passos do algoritmo k-Médias, com $k = 2$ e distância Euclidiana. Considere os objetos 1 e 4 como centroides iniciais.

Objeto	Atributo A	Atributo B	Atributo C	Atributo D
1	5,4	3,9	1,7	0,4
2	5,0	3,4	1,5	0,2
3	6,6	2,9	4,6	1,3
4	5,2	2,7	3,9	1,4

Resposta:
Centroides

Cluster	Objeto	Atributo A	Atributo B	Atributo C	Atributo D	Centroides
C1	1	5.4	3.9	1.7	0.4	(5.4, 3.9, 1.7, 0.4)
C2	4	5.2	2.7	3.9	1.4	(5.2, 2.7, 3.9, 1.4)

Cluster	Centroides
C1	(5.4, 3.9, 1.7, 0.4)
C2	(5.2, 2.7, 3.9, 1.4)

Iteração 1

Objeto	Atributo A	Atributo B	Atributo C	Atributo D
2	5.0	3.4	1.5	0.2

$$d(x_2, c_1) = \sqrt{(5.0 - 5.4)^2 + (3.4 - 3.9)^2 + (1.5 - 1.7)^2 + (0.2 - 0.4)^2}$$
$$d(x_2, c_1) = \sqrt{0.16 + 0.25 + 0.04 + 0.16}$$
$$d(x_2, c_1) = \sqrt{0.61}$$
$$d(x_2, c_1) = \mathbf{0.78}$$

$$d(x_2, c_2) = \sqrt{(5.0 - 5.2)^2 + (3.4 - 2.7)^2 + (1.5 - 3.9)^2 + (0.2 - 1.4)^2}$$
$$d(x_2, c_2) = \sqrt{0.4 + 0.49 + 5.76 + 1.44}\ d(x_2, c_1) = \sqrt{8.09}$$
$$d(x_2, c_1) = 2.84$$

Objeto	Cluster
1	C1
2	C1
3	
4	C2

Recalculando Centroide C1
Atual

Cluster	Objeto	Atributo A	Atributo B	Atributo C	Atributo D	Centroides
C1	1	5.4	3.9	1.7	0.4	(5.4, 3.9, 1.7, 0.4)

Novo

$$C_1 = (\frac{5.4 + 5.0}{2}, \frac{3.9 + 3.4}{2}, \frac{1.7 + 1.5}{2}, \frac{0.4 + 0.2}{2})$$
$$C_1 = (5.2, 3.65, 1.6, 0.3)$$

Cluster	Centroides
C1	(5.2,3.65,1.6,0.3)
C2	(5.2, 2.7, 3.9, 1.4)

Iteração 2

Objeto	Atributo A	Atributo B	Atributo C	Atributo D
3	6.6	2.9	4.6	1.3

$$d(x_3, c_1) = \sqrt{(6.6-5.2)^2 + (2.9-3.65)^2 + (4.6-1.6)^2 + (1.3-0.3)^2}$$
$$d(x_3, c_1) = \sqrt{1.96 + 0.56 + 9 + 1}$$
$$d(x_3, c_1) = \sqrt{12.52}$$
$$d(x_3, c_1) = 3.54$$

$$d(x_3, c_2) = \sqrt{(6.6-5.2)^2 + (2.9-2.7)^2 + (4.6-3.9)^2 + (1.3-1.4)^2}$$
$$d(x_3, c_2) = \sqrt{1.96 + 0.04 + 0.49 + 0.01}$$
$$d(x_3, c_2) = \sqrt{2.5}$$
$$d(x_3, c_2) = \mathbf{1.58}$$

Objeto	Cluster
1	C1
2	C1
3	C2
4	C2

Recalculando Centroide C2 Atual

Cluster	Objeto	Atributo A	Atributo B	Atributo C	Atributo D	Centroides
C2	4	5.2	2.7	3.9	1.4	(5.2, 2.7, 3.9, 1.4)

Novo

$$C_2 = (\frac{5.2+6.6}{2}, \frac{2.7+2.9}{2}, \frac{3.9+4.6}{2}, \frac{1.4+1.3}{2})$$
$$C_2 = (5.9, 2.8, 4.25, 1.35)$$

Cluster	Centroides
C1	(5.2,3.65,1.6,0.3)
C2	(5.9, 2.8, 4.25, 1.35)

Conclusão
Cluster com $k = 2$

Objeto	Cluster
1	C1
2	C1
3	C2
4	C2

Centroides finais

Cluster	Centroides
C1	(5.2,3.65,1.6,0.3)
C2	(5.9, 2.8, 4.25, 1.35)

4.2.2 Para o agrupamento resultante do exercício anterior, determine o valor do índice de Dunn. Utilize a distância Euclidiana para identificar os objetos mais similares. Mostre passo-a-passo a realização do cálculo do índice.

Resposta:

Clusters

Cluster	Centroides
C1	(5.2, 3.65, 1.6, 0.3)
C2	(5.9, 2.8, 4.25, 1.35)

Cáculo da medida intragrupo: $Intra(g_i) = max_{x,y \in gi}\{d(x,y)\}$

$$x = max(C1) = 5.2$$
$$y = max(C2) = 5.9$$
$$d(x,y) = \sqrt{(5.2 - 5.9)^2}$$
$$d(x,y) = 0.7$$
$$Intra(gi) = \mathbf{0.7}$$

Cálculo da medida intergrupo: $Inter(g_i, g_j) = \dfrac{1}{\mid g_i \mid \times \mid g_j \mid} \sum d(x,y) \mid x \in g_i, y \in g_j$

$$d(x,y) = \sqrt{(5.2 - 5.9)^2 + (3.65 - 2.8)^2 + (1.6 - 4.25)^2 + (0.3 - 1.5)^2}$$
$$d(x,y) = \sqrt{0.49 + 0.72 + 7.02 + 1.44}\ d(x,y) = 3.11\ Inter(g_i, g_j) = \frac{1}{4 \times 4} \times 3.11$$
$$Inter(g_i, g_j) = \mathbf{0.38}$$

Cálculo do Índice de Dunn: $DU(g) = \min_{i=1,\ldots,k} \left\{ \min_{j=1,\ldots,k; i \neq i} \left\{ \dfrac{Inter(g_i, g_j)}{\max_{l=1,\ldots,k} \{Intra(g_l)\}} \right\} \right.$

$$DU(g) = \frac{0.38}{0.7}$$
$$DU(g) = \mathbf{0.54}$$

4.2.3 Aplique o método de agrupamento single-link e desenhe o dendrograma da matriz de distâncias abaixo.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
P1										
P2	14									
P3	10	16								
P4	25	27	15							
P5	28	26	18	7						
P6	24	22	18	17	18					
P7	11	15	7	16	21	15				
P8	24	20	14	21	22	10	19			
P9	20	30	14	15	14	20	15	20		
P10	22	30	16	23	26	20	21	12	14	

Resposta: **Cluster Inicial:** $[P4, P5] = 7$

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
P1										
P2	14									
P3	10	16								
P4	25	27	15							
P5	28	26	18	7						
P6	24	22	18	17	18					
P7	11	15	7	16	21	15				
P8	24	20	14	21	22	10	19			
P9	20	30	14	15	14	20	15	20		
P10	22	30	16	23	26	20	21	12	14	

$$d(P1, [P4P5])$$
$$min(d(P1, P4), d(P1, P5))$$
$$min(d(25, 28)) \Rightarrow 25$$

$$d(P2, [P4P5])$$
$$min(d(P2, P4), d(P2, P5))$$
$$min(d(27, 26)) \Rightarrow 26$$

$$d(P3, [P4P5])$$
$$min(d(P3, P4), d(P3, P5))$$
$$min(d(15, 18)) \Rightarrow 15$$

$$d([P4P5], P6)$$
$$min(d(P4, P6), d(P5, P6))$$
$$min(d(17, 18)) \Rightarrow 17$$

$$d([P4P5], P7)$$
$$min(d(P4, P7), d(P5, P7))$$
$$min(d(16, 21)) \Rightarrow 16$$

$$d([P4P5], P8)$$
$$min(d(P4, P8), d(P5, P8))$$
$$min(d(21, 22)) \Rightarrow 21$$

$$d([P4P5], P9)$$
$$min(d(P4, P9), d(P5, P9))$$
$$min(d(15, 14)) \Rightarrow 14$$

$$d([P4P5], P10)$$
$$min(d(P4, P10), d(P5, P10))$$
$$min(d(23, 26)) \Rightarrow 2317$$

	P1	P2	P3	[P4 P5]	P6	P7	P8	P9	P10
P1									
P2	14								
P3	10	16							
[P4 P5]	25	26	15						
P6	24	22	18	17					
P7	11	15	7	16	15				
P8	24	20	14	21	10	19			
P9	20	30	14	14	20	15	20		
P10	22	30	16	23	20	21	12	14	

Próximo cluster: $[P3P7] = 7$

	P1	P2	P3	[P4 P5]	P6	P7	P8	P9	P10
P1									
P2	14								
P3	10	16							
[P4 P5]	25	26	15						
P6	24	22	18	17					
P7	11	15	7	16	15				
P8	24	20	14	21	10	19			
P9	20	30	14	14	20	15	20		
P10	22	30	16	23	20	21	12	14	

$$d(P1, [P3P7])$$
$$min(d(P1, P3), d(P1, P7))$$
$$min(d(10, 11)) \Rightarrow 10$$

$$d(P2, [P3P7])$$
$$min(d(P2, P3), d(P2, P7))$$
$$min(d(16, 15)) \Rightarrow 15$$

$$d([P3P7], [P4P5])$$
$$min(d(P3, [P4P5]), d(P7, [P4P5]))$$
$$min(d(15, 16)) \Rightarrow 15$$

$$d([P3P7], P6)$$
$$min(d(P3, P6), d(P7, P6))$$
$$min(d(18, 15)) \Rightarrow 15$$

$$d([P3P7], P8)$$
$$min(d(P3, P8), d(P7, P8))$$
$$min(d(14, 19)) \Rightarrow 14$$

$$d([P3P7], P9)$$
$$min(d(P3, P9), d(P7, P9))$$
$$min(d(14, 15)) \Rightarrow 14$$

$$d([P3P7], P10)$$
$$min(d(P3, P10), d(P7, P10))$$
$$min(d(16, 21)) \Rightarrow 16$$

	P1	P2	[P3 P7]	[P4 P5]	P6	P8	P9	P10
P1								
P2	14							
[P3 P7]	10	15						
[P4 P5]	25	26	15					
P6	24	22	15	17				
P8	24	20	14	21	10			
P9	20	30	14	14	20	20		
P10	22	30	16	23	20	12	14	

Próximo cluster: $[P1[P3P7]] = 10$

	P1	P2	[P3 P7]	[P4 P5]	P6	P8	P9	P10
P1								
P2	14							
[P3 P7]	10	15						
[P4 P5]	25	26	15					
P6	24	22	15	17				
P8	24	20	14	21	10			
P9	20	30	14	14	20	20		
P10	22	30	16	23	20	12	14	

$$d([P1[P3P7]], P2)$$
$$min(d(P1, P2), d(P2, [P3P7]))$$
$$min(d(14, 15)) \Rightarrow 14$$

$$d([P1[P3P7]], [P4P5])$$
$$min(d(P1, [P4, P5]), d([P3P7], [P4P5]))$$
$$min(d(25, 15)) \Rightarrow 15$$

$$d([P1[P3P7]], P6)$$
$$min(d(P1, P6), d([P3P7], P6))$$
$$min(d(24, 15)) \Rightarrow 15$$

$$d([P1[P3P7]], P8)$$
$$min(d(P1, P8), d([P3P7], P8))$$
$$min(d(24, 14)) \Rightarrow 14$$

$$d([P1[P3P7]], P9)$$
$$min(d(P1, P9), d([P3P7], P9))$$
$$min(d(20, 14)) \Rightarrow 14$$

$$d([P1[P3P7]], P10)$$
$$min(d(P1, P10), d([P3P7], P10))$$
$$min(d(22, 16)) \Rightarrow 16$$

	[P1 P3 P7]	P2	[P4 P5]	P6	P8	P9	P10
[P1 P3 P7]							
P2	14						
[P4 P5]	15	26					
P6	15	22	17				
P8	14	20	21	10			
P9	14	30	14	20	20		
P10	16	30	23	20	12	14	

Próximo cluster: $[P6P8] = 10$

	[P1 P3 P7]	P2	[P4 P5]	P6	P8	P9	P10
[P1 P3 P7]							
P2	14						
[P4 P5]	15	26					
P6	15	22	17				
P8	14	20	21	10			
P9	14	30	14	20	20		
P10	16	30	23	20	12	14	

$$d([P1P3P7], [P6P8])$$
$$min(d([P1P3P7], P6), d([P1P3P7], P8))$$
$$min(d(15, 14)) \Rightarrow 14$$

$$d(P2, [P6P8])$$
$$min(d(P2, P6), d(P2, P8))$$
$$min(d(22, 20)) \Rightarrow 20$$

$$d([P4P5], [P6P8])$$
$$min(d([P4P5], P6), d([P4P5], P8))$$
$$min(d(17, 21)) \Rightarrow 17$$

$$d([P6P8], P9)$$
$$min(d(P6, P9), d(P8, P9))$$
$$min(d(20, 20)) \Rightarrow 20$$

$$d([P6P8], P10)$$
$$min(d(P6, P10), d(P8, P10))$$
$$min(d(20, 12)) \Rightarrow 12$$

	[P1 P3 P7]	P2	[P4 P5]	[P6 P8]	P9	P10
[P1 P3 P7]						
P2	14					
[P4 P5]	15	26				
[P6 P8]	14	20	17			
P9	14	30	14	20		
P10	16	30	23	20	12	

Próximo cluster: $[P9P10] = 12$

	[P1 P3 P7]	P2	[P4 P5]	[P6 P8]	P9	P10
[P1 P3 P7]						
P2	14					
[P4 P5]	15	26				
[P6 P8]	14	20	17			
P9	14	30	14	20		
P10	16	30	23	20	12	

$$d([P1P3P7], [P9P10])$$
$$min(d([P1P3P7], P9), d([P1P3P7], P10))$$
$$min(d(14, 16)) \Rightarrow 14$$

$$d(P2, [P9P10])$$
$$min(d(P2, P9), d(P2, P10))$$
$$min(d(30, 30)) \Rightarrow 30$$

$$d([P4P5], [P9P10])$$
$$min(d([P4P5], P9), d([P4P5], P10))$$
$$min(d(14, 23)) \Rightarrow 14$$

$$d([P6P8], [P9P10])$$
$$min(d([P6P8], P9), d([P6P8], P10))$$
$$min(d(20, 20)) \Rightarrow 20$$

	[P1 P3 P7]	P2	[P4 P5]	[P6 P8]	[P9 P10]
[P1 P3 P7]					
P2	14				
[P4 P5]	15	26			
[P6 P8]	14	20	17		
[P9 P10]	14	30	14	20	

Próximo cluster: $[[P1P3P7]P2] = 14$

	[P1 P3 P7]	P2	[P4 P5]	[P6 P8]	[P9 P10]
[P1 P3 P7]					
P2	14				
[P4 P5]	15	26			
[P6 P8]	14	20	17		
[P9 P10]	14	30	14	20	

$$d([[P1P3P7]P2], [P4P5])$$
$$min(d([P1P3P7], [P4P5]), d(P2, [P4P5]))$$
$$min(d(15, 26)) \Rightarrow 15$$

$$d([[P1P3P7]P2], [P6P8])$$
$$min(d([P1P3P7], [P6P8]), d(P2, [P6P8]))$$
$$min(d(14, 20)) \Rightarrow 14$$

$$d([[P1P3P7]P2], [P9P10])$$
$$min(d([P1P3P7], [P9P10]), d(P2, [P9P10]))$$
$$min(d(14, 30)) \Rightarrow 14$$

	[P1 P2 P3 P7]	[P4 P5]	[P6 P8]	[P9 P10]
[P1 P2 P3 P7]				
[P4 P5]	15			
[P6 P8]	14	14		
[P9 P10]	14	14	20	

Próximo cluster: $[P1P2P3P7][P6P8] = 14$

	[P1 P2 P3 P7]	[P4 P5]	[P6 P8]	[P9 P10]
[P1 P2 P3 P7]				
[P4 P5]	15			
[P6 P8]	14	14		
[P9 P10]	14	14	20	

$$d([[P1P2P3P7][P6P8]], [P4P5])$$
$$min(d([P1P2P3P7], [P4P5]), d([P4P5], [P6P8]))$$
$$min(d(15, 14)) \Rightarrow 14$$

$$d([[P1P2P3P7][P6P8]], [P9P10])$$
$$min(d([P1P2P3P7], [P9P10]), d([P6P8], [P9P10]))$$
$$min(d(14, 20)) \Rightarrow 14$$

	[P1 P2 P3 P6 P7 P8]	[P4 P5]	[P9 P10]
[P1 P2 P3 P6 P7 P8]			
[P4 P5]	14		
[P9 P10]	14	14	

Próximo cluster: $[[P1P2P3P4P6P7P8][P4P5]] = 14$

$$d([[P1P2P3P6P7P8][P4P5]], [P9P10])$$
$$min(d([P1P2P3P6P7P8], [P9P10]), d([P4P5], [P9P10]))$$
$$min(d(14, 14)) \Rightarrow 14$$

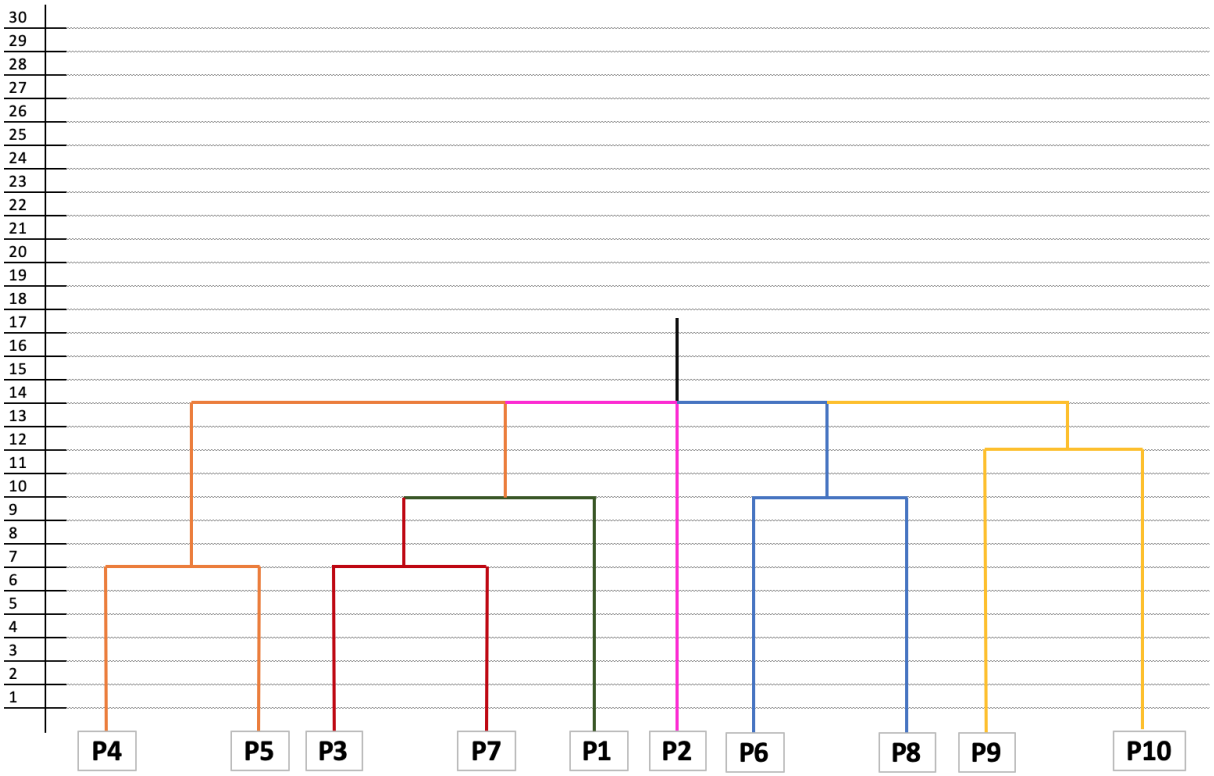
	[P1 P2 P3 P6 P7 P8]	[P9 P10]
[P1 P2 P3 P4 P5 P6 P7 P8]		
[P9 P10]	14	

Próximo cluster: $[P1P2P3P4P5P6P7P8][P9P10] = 14$

	[P1 P2 P3 P4 P5 P6 P7 P8]	[P9 P10]
[P1 P2 P3 P4 P5 P6 P7 P8]		
[P9 P10]	14	

	[P1 P2 P3 P4 P5 P6 P7 P8 P9 P10]
[P1 P2 P3 P4 P5 P6 P7 P8 P9 P10]	

Dendrograma



5 CAPÍTULO 05: CLASSIFICAÇÃO DE DADOS

5.1 EXERCÍCIOS CONCEITUAIS

5.1.1 Durante o desenvolvimento de um modelo de classificação a base de dados é dividida em dois conjuntos. Quais são estes conjuntos e qual é o propósito desta separação

Resposta: Os dois conjuntos são: treinamento e testes. Seu propósito é gerar modelos preditivos capazes de identificar classes ou valores de registros não rotulados com um resultado satisfatório.

5.1.2 Discuta por quê para bases de dados reais na maioria das vezes treinar um sistema preditivo até que o erro para os dados de treinamento seja zero não é aconselhável.

Resposta: Durante o processo de treinamento de modelos, mesmo com o desempenho cada vez melhor e a quantidade de erros possa estar decaindo, o desempenho de generalização pode começar a se deteriorar após determinado número de iterações.
O ideal é utilizar uma validação cruzada para interromper o treinamento para conseguir obter um melhor resultado.

5.1.3 O que é a validação cruzada em k-pastas? Qual é a sua finalidade?

Resposta: A validação cruzada em k-pastas consiste em dividir a base de dados em k subconjuntos, sendo k –1 pastas para treinamento e 1 pasta para teste. Esse processo de treinamento e teste é repetido com todos os k subconjuntos, e a média dos desempenhos para as bases de treinamento e as bases de teste é adotada como indicador de qualidade do modelo.
A validação cruzada em k-pastas é bastante usual para se estimar o erro de generalização de preditores.

5.1.4 O que é a matriz de confusão de um problema de classificação binária? Explique os quatro valores apresentados pela matriz.

Resposta: Matriz de confusão é uma forma de apresentar integralmente o desempenho de um algoritmo de classificação binária utilizando uma matriz que relaciona as classes desejadas com as classes preditas.
Essa matriz também é conhecida como *matriz de contingência* ou *matriz de erro* e na sua composição ela tem nas linhas os objetos nas classes originais e nas colunas os objetos nas classes preditas.

		Classe predita	
		Positiva	Negativa
Classe original	Positiva	VP	FN
	Negativa	FP	VN

Abaixo estão as explicações dos quatro valores apresentados na tabela.

- **VP (verdadeiro positivo):** você previu positivo e é verdadeiro
  Ex.: Você previu que uma mulher está grávida e ela realmente está.
- **VN (verdadeiro negativo):** você previu negativo e é verdadeiro
  Ex.: Você previu que um homem não está grávido e ele realmente não está.
- **FP (falso positivo):** você previu positivo e é falso
  Ex.: Você previu que um homem está grávido, mas na verdade ele não está.
- **FN (falso negativo):** você previu negativo e é falso
  Ex.: Você previu que uma mulher não está grávida, mas na verdade ela está.

5.2 EXERCÍCIOS NUMÉRICOS

5.2.1 Utilizando a árvore de decisão para o conjunto de treinamento da base de dados Cogumelos, apresentada na Figura 5.19, extraia todas as regras de classificação para cogumelos venenosos.

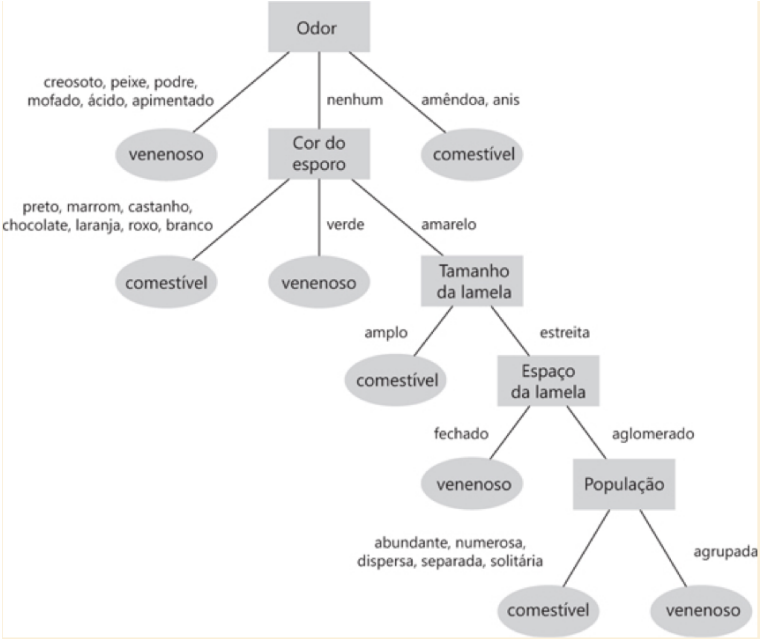


Figura 5.19

Resposta:

Atributo	Valor	Classe
odor	creosoto	venenoso
odor	peixe	venenoso
odor	podre	venenoso
odor	mofado	venenoso
odor	ácido	venenoso
odor	apimentado	venenoso
cor do esporo	verde	venenoso
espaço da lamela	fechado	venenoso
população	agrupada	venenoso

5.2.2 Considere o seguinte problema: Um determinado algoritmo é responsável por indicar se um paciente está doente ou não. Após a realização de um experimento com 150 pacientes, onde 30 estavam doentes, o algoritmo informou a existência de 28 doentes sendo que 8 destes estavam saudáveis. Calcule a taxa de verdadeiro positivo *(TPR)*, a taxa de falso positivo *(FPR)*, a acurácia *(ACC)*, e a precisão *(Pr)*.

Resposta:

		Classe predita	
		Positiva	Negativa
Classe original	Positiva	20	10
	Negativa	8	112

Taxa de verdadeiro positivo (TVP):

$$TVP = \frac{VP}{VP + FN} = \frac{20}{30} = 0.6666 \cdot 100 = 66.66\%$$

Taxa de falso positivo (TFP):

$$TFP = \frac{FP}{FP + VN} = \frac{8}{120} = 0.0666 \cdot 100 = 6,67\%$$

Acurácia (ACC):

$$ACC = \frac{VP + VN}{VP + FP + VN + FN} = \frac{132}{150} = 0.88 \cdot 100 = 88\%$$

Precisão (PR):

$$Pr = \frac{VP}{FP + VP} = \frac{20}{28} = 0.7143 \cdot 100 = 71.43\%$$

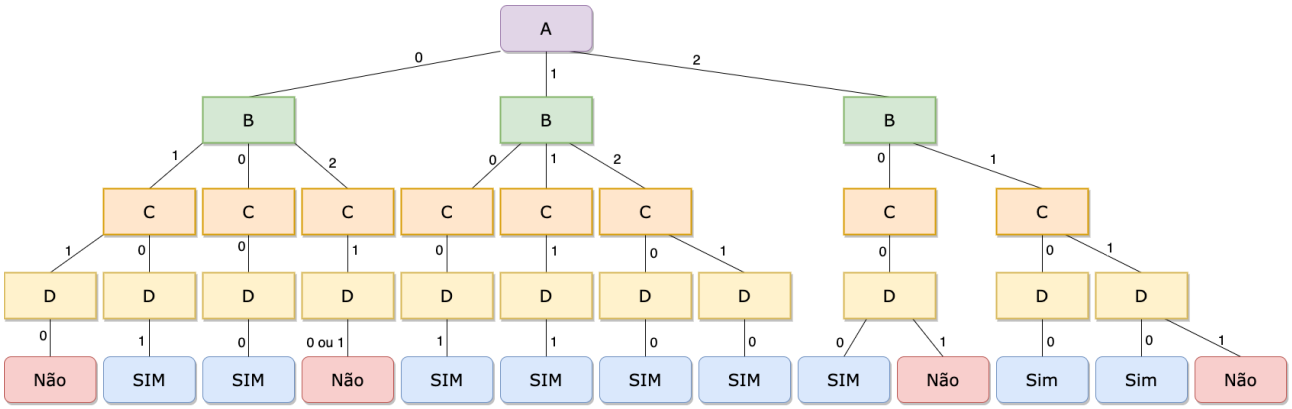
Conclusão:

$$TVP = 66.66\%$$
$$TFP = 6.67\%$$
$$ACC = 88\%$$
$$Pr = 71.43\%$$

5.2.3 Construa uma árvore de decisão para a base de dados abaixo.

A	B	C	D	Classe
2	1	1	0	Sim
2	0	0	0	Sim
2	0	0	1	Não
2	1	0	0	Sim
2	1	1	1	Não
0	2	1	0	Não
0	2	1	1	Não
0	1	1	0	Não
0	0	0	0	Sim
0	1	0	1	Sim
1	2	1	0	Sim
1	0	0	1	Sim
1	1	1	1	Sim
1	2	0	0	Sim

Resposta:



5.2.4 Aplique o classificador *Naïve Bayes* na base de dados abaixo e determine o valor para classe Jogar dos seguintes objetos:

a X = (Ensolarado, Branda, Alta, Não)

b Y = (Chuvoso, Fria, Alta, Sim)

c Z = (Ensolarado, Branda, Normal, Não)

d W = (Fechado , Fria, Normal, Sim)

Tempo	Tempera-tura	Umidade	Vento	Jogar
Ensolarado	Quente	Alta	Não	**Não**
Ensolarado	Quente	Alta	Sim	**Não**
Fechado	Quente	Alta	Não	**Sim**
Chuvoso	Branda	Alta	Não	**Sim**
Chuvoso	Fria	Normal	Não	**Sim**
Chuvoso	Fria	Normal	Sim	**Não**
Fechado	Fria	Normal	Sim	**Sim**
Ensolarado	Branda	Alta	Não	**Não**
Ensolarado	Fria	Normal	Não	**Sim**
Chuvoso	Branda	Normal	Não	**Sim**
Ensolarado	Branda	Normal	Sim	**Sim**
Fechado	Branda	Alta	Sim	**Sim**
Fechado	Quente	Normal	Não	**Sim**
Chuvoso	Branda	Alta	Sim	**Não**

Resposta: **Teorema de Bayes:** $P(C_i \mid x) = \dfrac{P(x \mid C_i)P(C_i)}{P(x)}$

Probabilidade por atributo:

Tempo	Jogar = Sim	Jogar = Não	Total
Ensolarado	$2/9 = 0.22$	$3/5 = 0.6$	$5/14 = 0.36$
Fechado	$4/9 = 0.44$	$0/5 = 0$	$4/14 = 0.28$
Chuvoso	$3/9 = 0.33$	$2/5 = 0.55$	$5/14 = 0.36$

Temperatura	Jogar = Sim	Jogar = Não	Total
Quente	$2/9 = 0.22$	$2/5 = 0.4$	$4/14 = 0.28$
Branda	$4/9 = 0.44$	$2/5 = 0.4$	$6/14 = 0.43$
Fria	$3/9 = 0.33$	$1/5 = 0.2$	$4/14 = 0.28$

Umidade	Jogar = Sim	Jogar = Não	Total
Alta	$3/9 = 0.33$	$4/5 = 0.8$	$7/14 = 0.5$
Normal	$6/9 = 0.67$	$1/5 = 0.2$	$7/15 = 0.5$

Vento	Joga = Sim	Jogar = Não	Total
Sim	$3/9 = 0.33$	$3/5 = 0.6$	$6/14 = 0.43$
Não	$6/9 = 0.67$	$2/5 = 0.4$	$8/14 = 0.57$

Probabilidade de Jogar (Sim/Não) $P(C)$**:** $P(Jogar = SIM) = 9/14 = 0.64$
$P(Jogar = NAO) = 5/14 = 0.36$

Problema: a) - X = (Ensolarado, Branda, Alta, Não):

Probabilidade de Jogar=Sim por atributo:
$P(Tempo = Ensolarado \mid Jogar = Sim) = 0.22$
$P(Temperatura = Branda \mid Jogar = Sim) = 0.44$
$P(Umidade = Alta \mid Jogar = Sim) = 0.33$
$P(Vento = Nao \mid Jogar = Sim) = 0.67$
$P(Jogar = Sim) = 0.64$
$P(X \mid Jogar = Sim)P(Jogar = Sim) = 0.22 \cdot 0.44 \cdot 0.33 \cdot 0.67 \cdot 0.64 = 0.0137$

Probabilidade de Jogar=Nao por atributo:
$P(Tempo = Ensolarado \mid Jogar = Nao) = 0.6$
$P(Temperatura = Branda \mid Jogar = Nao) = 0.4$
$P(Umidade = Alta \mid Jogar = Nao) = 0.8$
$P(Vento = Nao \mid Jogar = Nao) = 0.4$
$P(Jogar = Nao) = 0.36$
$P(X \mid Jogar = Nao)P(Jogar = Nao) = 0.6 \cdot 0.4 \cdot 0.8 \cdot 0.4 \cdot 0.36 = 0.0276$

Probabilidade do total dos atributos:
$P(X) = P(Tempo = Ensolarado) \cdot P(Temperatura = Branda) \cdot P(Umidade = Alta) \cdot P(Vento = Nao)$
$P(X) = 0.36 \cdot 0.43 \cdot 0.5 \cdot 0.57$
$P(X) = 0.0441$

Probabilidade de Jogar:
$P(Jogar = Sim) \mid X) = \frac{0.0137}{0.0441} = 0.3106$
$P(Jogar = Nao) \mid X) = \frac{0.0276}{0.0441} = \mathbf{0.6258}$

Conclusão de X: Probabilidade de Jogar = **NÃO**

Problema: b) - Y = (Chuvoso, Fria, Alta, Sim):

Probabilidade de Jogar=Sim por atributo:
$P(Tempo = Chuvoso \mid Jogar = Sim) = 0.33$
$P(Temperatura = Fria \mid Jogar = Sim) = 0.33$
$P(Umidade = Alta \mid Jogar = Sim) = 0.33$
$P(Vento = Sim \mid Jogar = Sim) = 0.33$
$P(Jogar = Sim) = 0.64$
$P(Y \mid Jogar = Sim)P(Jogar = Sim) = 0.33 \cdot 0.33 \cdot 0.33 \cdot 0.33 \cdot 0.64 = 0.0076$

Probabilidade de Jogar=Nao por atributo:
$P(Tempo = Chuvoso \mid Jogar = Nao) = 0.6$
$P(Temperatura = Fria \mid Jogar = Nao) = 0.2$
$P(Umidade = Alta \mid Jogar = Nao) = 0.8$
$P(Vento = Sim \mid Jogar = Nao) = 0.6$
$P(Jogar = Nao) = 0.36$
$P(Y \mid Jogar = Nao)P(Jogar = Nao) = 0.6 \cdot 0.2 \cdot 0.8 \cdot 0.6 \cdot 0.36 = 0.0207$

Probabilidade do total dos atributos:
$P(Y) = P(Tempo = Chuvoso) \cdot P(Temperatura = Fria) \cdot P(Umidade = Alta) \cdot P(Vento = Sim)$
$P(Y) = 0.36 \cdot 0.28 \cdot 0.5 \cdot 0.43$
$P(Y) = 0.0217$

Probabilidade de Jogar:
$P(Jogar = Sim) \mid Y) = \frac{0.0076}{0.0217} = 0.3502$
$P(Jogar = Nao) \mid Y) = \frac{0.0207}{0.0217} = \mathbf{0.9539}$

Conclusão de Y: Probabilidade de Jogar = **NÃO**

Problema: c) - Z = (Ensolarado, Branda, Normal, Não):

Probabilidade de Jogar=Sim por atributo:
$P(Tempo = Ensolarado \mid Jogar = Sim) = 0.22$
$P(Temperatura = Branda \mid Jogar = Sim) = 0.44$
$P(Umidade = Normal \mid Jogar = Sim) = 0.67$
$P(Vento = Nao \mid Jogar = Sim) = 0.67$
$P(Jogar = Sim) = 0.64$
$P(Z \mid Jogar = Sim)P(Jogar = Sim) = 0.22 \cdot 0.44 \cdot 0.67 \cdot 0.67 \cdot 0.64 = 0.0278$

Probabilidade de Jogar=Nao por atributo:
$P(Tempo = Ensolarado \mid Jogar = Nao) = 0.6$
$P(Temperatura = Branda \mid Jogar = Nao) = 0.4$
$P(Umidade = Normal \mid Jogar = Nao) = 0.2$
$P(Vento = Nao \mid Jogar = Nao) = 0.4$
$P(Jogar = Nao) = 0.36$
$P(Z \mid Jogar = Nao)P(Jogar = Nao) = 0.6 \cdot 0.4 \cdot 0.2 \cdot 0.4 \cdot 0.36 = 0.0069$

Probabilidade do total dos atributos:
$P(Z) = P(Tempo = Ensolarado) \cdot P(Temperatura = Branda) \cdot P(Umidade = Normal) \cdot P(Vento = Nao)$
$P(Z) = 0.36 \cdot 0.43 \cdot 0.5 \cdot 0.57$
$P(Z) = 0.0441$

Probabilidade de Jogar:
$P(Jogar = Sim) \mid Z) = \dfrac{0.0278}{0.0441} = \mathbf{0.6303}$
$P(Jogar = Nao) \mid Z) = \dfrac{0.0069}{0.0441} = 0.1564$

Conclusão de Z: Probabilidade de Jogar = **SIM**

Problema: d) - W = (Fechado, Fria, Normal, Sim):

Probabilidade de Jogar=Sim por atributo:
$P(Tempo = Fechado \mid Jogar = Sim) = 0.44$
$P(Temperatura = Fria \mid Jogar = Sim) = 0.33$
$P(Umidade = Normal \mid Jogar = Sim) = 0.67$
$P(Vento = Sim \mid Jogar = Sim) = 0.33$
$P(Jogar = Sim) = 0.64$
$P(W \mid Jogar = Sim)P(Jogar = Sim) = 0.44 \cdot 0.33 \cdot 0.67 \cdot 0.33 \cdot 0.64 = 0.0109$

Probabilidade de Jogar=Nao por atributo:
$P(Tempo = Fechado \mid Jogar = Nao) = 0$
$P(Temperatura = Fria \mid Jogar = Nao) = 0.33$
$P(Umidade = Normal \mid Jogar = Nao) = 0.2$
$P(Vento = Sim \mid Jogar = Nao) = 0.6$
$P(Jogar = Nao) = 0.36$
$P(W \mid Jogar = Nao)P(Jogar = Nao) = 0 \cdot 0.33 \cdot 0.2 \cdot 0.6 \cdot 0.36 = 0$

Probabilidade do total dos atributos:
$P(W) = P(Tempo = Fechado) \cdot P(Temperatura = Fria) \cdot P(Umidade = Normal) \cdot P(Vento = Sim)$
$P(W) = 0.28 \cdot 0.28 \cdot 0.5 \cdot 0.43$
$P(W) = 0.0168$

Probabilidade de Jogar:
$P(Jogar = Sim) \mid W) = \dfrac{0.0109}{0.0168} = \mathbf{0.6488}$
$P(Jogar = Nao) \mid W) = \dfrac{0}{0.0168} = 0$

Conclusão de W: Probabilidade de Jogar = **SIM**