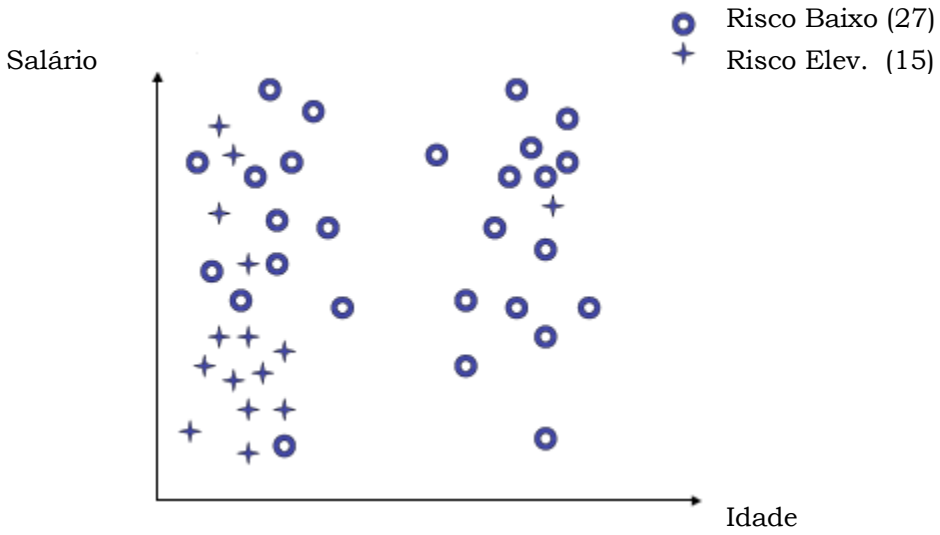
Departamento de Engenharia Informática

4 valores

1. Considere o seguinte conjunto de dados:



a) Encontre a melhor partição, e o correspondente ganho de informação. Considere pelo menos duas alternativas, uma por eixo, e compare o ganho de informação de cada uma.

b) Múltiplas partições de um atributo resultam sempre numa maior redução de entropia do que partições binárias. Diga se concorda com a afirmação, justifique.

6 valores

2. Considere o seguinte conjunto de dados relativo a cogumelos não-venenosos/venenosos

Cogumelo	Pesado	Mau-cheiro	Manchado	Venenoso
A	0	0	0	0
B	0	1	1	1
C	1	1	0	0
D	1	0	0	1
E	1	1	1	?
F	0	0	1	?
G	0	1	0	?
H	1	0	1	?

a) Sem recorrer a cálculos, desenhe a árvore de decisão apenas com o atributo raiz, relativa a este conjunto de dados. Justifique.

b) Apresente a matriz de confusão. Qual a taxa de Precisão do modelo?

c) O modelo apresentado deverá apresentar diferentes custos relativamente às falsas previsões? Justifique.

d) Qual a medida mais adequada para avaliar o desempenho deste modelo? Apresente o seu valor e compare face à Precisão do modelo.

e) Usando o classificador Naive Bayes sobre este conjunto de dados qual seria a previsão para a instância E? Apresente os cálculos. Porque razão o classificador Naive Bayes é designado Naive (ingénuo)?

f) Construa um classificador K-NN baseado na distância Euclediana para prever a toxidade da instância H. Use K = 1.

4 valores

3. Suponha que o algoritmo Apriori produziu os seguintes itemsets frequentes e respectivos suportes:

L_1	Sup
{A}	0,4
{C}	0,3
{B}	0,6
{E}	0,5

L_2	Sup
{A, C}	0,2
{A,E}	0,2
{C,E}	0,2
{B,E}	0,4

L_3	Sup
{A, C, E}	0,1

$L_4 = \varnothing$

a) Seja β=0,5 o nível mínimo de confiança, apresente todas as regras de associação com confiança maior ou igual a β.

b) Explique porque o algoritmo Apriori (que gera todos os itemsets frequentes) sempre pára, isto é, porque razão existe sempre um K tal que $L_k = \varnothing$.

c) O que são itemsets negativamente correlacionados ? O que significa isto em termos práticos?

d) Dê um exemplo de uma regra de associação A $\Rightarrow$ B que tenha suporte e confiança acima dos mínimos exigidos mas onde os conjuntos de itens A e B são negativamente correlacionados.

3 valores

4. Considere o seguinte algoritmo de clustering designado por *Leader Clustering*. Este algoritmo recebe dois parâmetros: um valor inteiro k e um valor real t. Funciona analogamente ao algoritmo K-means, ou seja, começa por seleccionar k instâncias – Leaders e em seguida atribui cada instância de treino ao leader mais próximo, com excepção se a distância da instância ao leader for superior ao parâmetro t, então essa instância passa a ser um novo *leader*. Após o processamento de todas as instâncias de treino, são calculados os centros de cada cluster, sendo estes centros os novos *leaders*, e o processo é repetido até ser encontrada uma partição estável.

a) Dado um conjunto de dados, e os valores dos parâmetros k e t, será a partição produzida pelo algoritmo Leader Clustering igual à do algoritmo K-means? Assuma que as k instâncias iniciais seleccionadas como leaders/centróides para ambos os algoritmos são as mesmas. Considere diferentes casos dependentes do parâmetro t, justifique a sua resposta.

b) Qual dos dois métodos será melhor a lidar com valores isolados (outliers). Explique a sua resposta.

3 valores

5. Considere a seguinte matriz de documentos caracterizados pelas seguintes Expressões Relevantes (ERs) como atributos:

	ER1	ER2	ER3	ER4	ER5	ER6
Doc1	2	0	0	0	1	10
Doc2	2	0	14	0	1	10
Doc3	2	0	22	3	1	13
Doc4	3	0	9	3	1	13
Doc5	2	0	81	0	1	10
Doc6	2	1	30	0	1	10
Doc7	2	0	200	0	1	10
Doc8	1	1	100	0	1	10
Doc9	2	0	1	0	1	10

INSTITUTO POLITÉCNICO DO PORTO

a) Dos atributos acima apresentados, há algum que seja irrelevante? Porquê?

b) Qual o atributo que mais contribuiria para o agrupamento (clustering) de documentos considerando o conteúdo neles residente? Porquê?

c) Entre os atributos ER4 e ER6, qual deles acha mais informativo? Porquê?

d) Com vista ao cálculo de uma matriz de semelhanças entre documentos, pesaria de forma diferente a influência de cada ER no cálculo dessas semelhanças? Se sim, qual o princípio básico de um tal critério?

FIM