



Observe que:

Deverá realizar as partes A e B em folhas separadas.

Exame com consulta

Qualquer tentativa de fraude conduzirá à anulação da prova para todos os intervenientes.

Parte A:

1 – Considere o seguinte conjunto de variáveis categóricas adquiridas de um conjunto de 11 doentes:

Fumador	S	S	S	S	N	N	N	S	S	S	N
Creatinina Alta	S	S	S	S	S	N	N	N	N	N	N
Risco Cardiovascular	B	B	A	A	B	B	B	B	A	A	A

Assumindo que S="Sim", N="Não" e que A="Alto", B="Baixo", considere que lhe é pedido que projete um preditor do risco cardiovascular usando o atributo "Antecedentes Familiares". Nesse contexto, apresente o ranking de atributos usando a metodologia baseada na Informação mútua. Deverá apresentar todos os cálculos necessários.

2 – Considere o conjunto de pontos (x,y) das classes C1 e C2, respetivamente:

Tabela 1: Pontos das classes C1 e C2.

Classe C1 (x,y)	Classe C2 (x,y)
P1=(4, 5)	P26=(6, 2)
P2=(3, 6)	P27=(13, 14)
P3=(4, 8)	P28=(25, 31)
P4=(8, 10)	P29=(26, 28)
P5=(6, 1)	P30=(7, 2)
P6=(7, 5)	
P7=(8, 16)	
P8=(3, 0)	
P9=(18, 18)	
P10=(19, 14)	
P11=(8, 9)	
P12=(2, 11)	
P13=(5, 4)	
P14=(8, 9)	
P15=(11, 19)	
P16=(5, 10)	
P17=(12, 10)	
P18=(14, 4)	
P19=(4, 9)	
P20=(2, 12)	
P21=(28, 32)	
P22=(22, 33)	
P23=(22, 30)	

P24=(34, 16)	
P25=(15, 20)	

- a) Usando o DBScan com ϵ = 10 e minPts = 4 indique, fundamentando, o resultado da classificação do ponto P24.
- b) Considere o ponto P31 = (15, 15) cuja classe não é conhecida. Sugira uma estratégia de imputação para este ponto e apresente o respetivo resultado.
- c) Determine a projeção que maximiza a variância explicada dos dados da tabela 1. Apresente todas as matrizes e indique todos os cálculos necessários.

Parte B:

No seguinte dataset, o estado de saúde de pessoas adultas é classificado com valores entre 1 (pessoas mais saudáveis) e 4 (pessoas com maiores riscos de saúde). As características (features) incluídas no dataset são o índice de massa corporal (IMC, qualitativamente indicado como Baixo, Normal ou Alto), a idade (quantitativamente indicado em anos), e se realiza regularmente atividade física (Sim ou Não).

Tabela B1				
Classes verdadeiras	IMC	Idade	Atividade Física	Predição
2	Normal	25	Sim	2
2	Alto	40	Sim	3
4	Normal	63	Não	4
3	Alto	18	Não	3
4	Alto	81	Não	1
1	Normal	33	Sim	1
2	Normal	38	Sim	2
1	Normal	64	Sim	3
3	Baixo	22	Sim	1
2	Normal	45	Sim	2
4	Normal	90	Não	4
2	Alto	57	Sim	2
4	Normal	70	Não	4
4	Alto	68	Não	4
2	Normal	54	Sim	4
3	Baixo	20	Sim	2

1 – Considerando o resultado de uma classificação obtida por um modelo desconhecido (Tabela B1, coluna 5), implemente um método apropriado para avaliação da exatidão deste modelo, tendo em consideração as características do dataset.

2 – Pretende-se construir um novo classificador baseado em *feed-forward neural networks*:

- a) Indicar as entradas e as saídas da NN (assim como os possíveis valores que podem adquirir)
- b) Indicar os parâmetros que devem ser ajustados durante a fase de treino.
- c) Explicar brevemente (com suas próprias palavras, sem o uso de equações) em que consiste o método de *backpropagation* utilizado no processo de treino de uma NN.

3 – Considerando as características do dataset dado na Tabela 1B, indicar dois (só dois) estratégias que considere as mais apropriadas para a divisão deste dataset, que permita uma correta avaliação do desempenho do modelo. Justificar a escolha.

4 – Considerar que se pretende implementar um modelo de classificação binária baseado em SVM. Para isto devem ser definidas duas classes, sendo que a primeira classe (pessoas mais saudáveis)

incluir as classes 1 e 2 da Tabela 1B, e a segunda classe (pessoas com maior risco de saúde) inclui as classes 3 e 4.

Considerando só a variável Idade, diga que estratégia pode ser aplicada para melhorar a separabilidade entre as duas classes definidas. Justificar a resposta.

5 – Selecionar a melhor regra (rule) para o problema de classificação binária apresentado no ponto 4, considerando as três variáveis disponíveis. Expresse a regra escolhida.

Nota: para isto, deverá discretizar a variável Idade, sendo que o primeiro grupo deve abranger idades entre 18 e 59 anos, e o segundo grupo as idades iguais ou superiores a 60 anos.