Exemplo de Exame (Parte A)

Parte A:

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

Tabela 1: Dados relativos a 11 doentes.

Hipertensão	S	S	S	S	S	N	S	S	N	N	S
Edema	S	S	N	N	S	N	S	S	S	S	S
Ritmo Cardíaco (BPM)	150	120	80	67	88	56	67	90	96	78	130
Glicémia (mg/dL)	150	190	125	90	66	199	121	321	210	170	199
HF Descompensado	S	S	S	S	N	N	N	S	S	S	S

Assumindo que S=”Sim”, N=”Não”, considere que lhe é pedido que projete um preditor do risco de descompensação usando uma variável categórica (Hipertensão e Edema) e uma variável numérica (Ritmo cardíaco e Glicémia). Nesse contexto, responda às seguintes questões:

a) Determine o valor Kruskal Lambda para as variáveis independentes Hipertensão e Edema relativamente à variável dependente HF Descompensação. Apresente todos os cálculos; não use implementações computacionais.
b) Se tivesse de escolher uma das variáveis independentes categóricas como feature num classificador, qual usaria? Fundamente.
c) Determine a relevância usando o score de Fisher das variáveis Ritmo Cardíaco e Glicémia. Apresente todos os cálculos (não deverá usar bibliotecas já definidas).

2 – Considere a série temporal presente na figura e tabla seguinte:

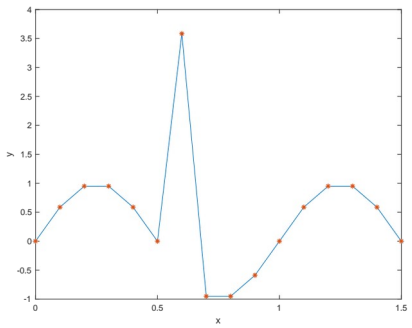

Tabela 2: Pontos (x,y) e da variável dependente “output”.

x	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5
y	0.	0.587	0.951	0.951	0.587	0	3.585	-0.9511	-0.9511	-0.5878	0	0.587	0.951	0.951	0.587	0.0

	0	8	1	1	8		5					8	1	1	8	
output	A	A	A	A	A	A	B	B	B	B	A	A	A	A	A	A

a) Assumindo que os pontos são distribuídos segundo uma distribuição normal, determine os outliers da série; não use implementações computacionais.
b) Use um processo autorregressivo (regressão linear) de dimensão 3 para imputar os outliers determinados em a). Apresente todos os cálculos e matrizes relevantes; não use implementações computacionais.

Parte B

3 – Pretende-se criar um modelo computacional para previsão do risco de insuficiência cardíaca (risco: alto, médio, baixo). Para isso, foi criada uma pequena base de dados envolvendo 200 sujeitos, onde foram recolhidos dados para os seguintes atributos: paciente tem doença arterial coronariana (CAD: sim, não), faixa etária (G1: < 45 anos; G2: ≥ 45 e < 65; G3: ≥ 65) e nível de pressão arterial do paciente (BPL: baixa, media, alta).

a) A tabela seguinte representa uma amostra do conjunto de treino, contendo os dados de apenas 12 sujeitos. Na construção de um classificador OneR, que atributo seria selecionado, que regras seriam obtidas e que taxa de erro seria alcançada? Apresente os cálculos necessários.

CAD	Age Group	BPL	Heart Failure
N	G1	B	B
N	G1	A	B
S	G1	B	B
N	G2	M	B
N	G2	A	M
S	G1	A	M
S	G2	M	M
N	G3	B	B
N	G3	A	A
S	G2	A	A
S	G3	M	A
S	G3	A	A

b) Agora imagine que um classificador Naive Bayes foi utilizado no mesmo conjunto de treino. Como é que esse modelo classificaria uma nova amostra (CAD = N; Faixa Etária = G1; BPL = M)? Apresente os cálculos necessários.

c) No cenário descrito na alínea b), quão confiante poderia estar em relação à classificação proposta? Justifique a sua resposta quantitativamente.

d) Se utilizasse um Multi-Layer Perceptron, como definiria a camada de saída (número de neurónios, tipo de funções de ativação e representação da saída alvo – target output)? Justifique as suas decisões.

4 – Um dos principais hiperparâmetros das máquinas vetoriais de suporte (SVM) é o parâmetro Cost. Explique como funciona (em termos gerais) e o seu impacto em termos de enviesamento (bias) e variância.