Cuestiones

Responde brevemente a las siguientes cuestiones justificando las respuestas:

1. **(2)** Indique dos formas de gestionar los valores perdidos (*missing values)* en un conjunto de datos. Indique también qué ventajas e inconvenientes ve en cada una de ellas.

   RESPUESTA: Hay dos formas fundamentales, ignorar los valores perdidos o estimarlos. Si se ignoran se puede hacer a nivel de instancia, ignorando todas las instancias con al menos un valor perdido, o a nivel de variable, ignorando aquellas variables que tienen al menos un valor perdido en una instancia. El problema de esta aproximación es que podemos perder mucha información.

   En el segundo enfoque se estiman los valores perdidos, usando modas, medianas, medias, estimaciones estadísticas o vecinos más cercanos. En cualquier métodos que usemos tenemos el problema de que introducimos ruido en la muestra.

2. **(2)** Considere el box plot de las cuatro variables del problema iris mostrado en la figura 1. ¿Qué información puede obtener de esa representación respecto al comportamiento de las variables?

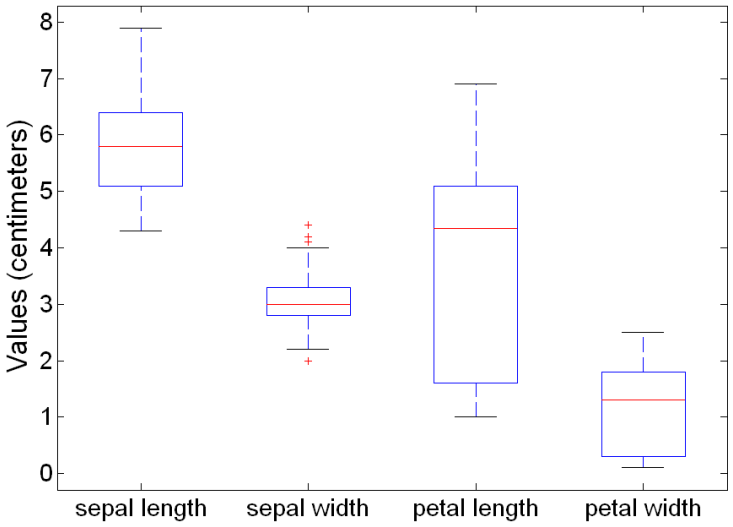

Figura 1: Box plot de las cuatro variables del problema de tres clases *iris*.

   RESPUESTA: Según el gráfico podemos observar que la variable *petal lenght* tiene una gran dispersión al igual que, en menor medida, las variables *sepal length* y *pepal width*. Por el contrario, *sepal width* tiene valores muy homogéneos entre los diferentes patrones. Respecto a la importancia de estas variables en la clasificación, de un gráfico tipo box plot **NO** podemos deducir nada, ya que una variable más homogénea puede ser más discriminante que una variable con mayor dispersión.

3. **(2)** Indique un aspecto positivo y otro negativo de cada uno de los tres siguientes métodos de clasificación: un árbol de decisión, una máquina de vectores soporte (SVM) y el método del vecino más cercano (*nearest neighbor*).

   RESPUESTA: Árboles de decisión:

Positivo: Capaces de tratar con problemas muy grandes, muy rápidos en la clasificación, interpretables cuando son pequeños. buena relación coste/rendimiento, inestables.

Negativos: Menor rendimiento que otros métodos, inestables.

Máquinas de vectores soporte (SVM):

Positivo: Pueden ser muy eficientes con conjuntos de datos con miles de variables, muy buen rendimiento, robustos ante la presencia de ruido, estables.

Negativo: Son muy costosos computacionalmente, muy sensibles a los parámetros de entrenamiento, estables.

Vecino más cercano:

Positivo: No necesitan entrenamiento, buen rendimiento, estables antes variaciones en el conjunto de instancias, inestables ante variaciones en el conjunto de variables.

Negativo: Necesitan almacenar el conjunto de entrenamiento completo por lo que tienen problemas de escalabilidad, estables antes variaciones en el conjunto de instancias, inestables ante variaciones en el conjunto de variables.

4. **(2)** En la construcción de reglas de asociación, ¿qué efecto tiene el uso de un soporte mínimo variable según los ítems en un itemset sobre el algoritmo Apriori?

   RESPUESTA: El soporte pierde la propiedad de anti-monotonía y por lo tanto el algoritmo Apriori deja de ser aplicable porque está basado en dicha propiedad. Existen diferentes modificaciones del algoritmo para poder seguir siendo aplicado aunque con menor efectividad.

5. **(2)** Indique dos puntos fuertes del agrupamiento jerárquico con respecto al particional.

   RESPUESTA: Un punto fuerte es que no asume un número determinado de clústers en el conjunto de datos. Otro punto fuerte es que el resultado es una taxonomía de las instancias que puede ser de mucha utilidad en muchas áreas de conocimiento.

Problemas

1. **(3)** La siguiente tabla representa la matriz de similaridad entre los elementos de un cierto conjunto de datos compuesto por 6 puntos:

	p1	**p2**	**p3**	**p4**	**p5**	**p6**
p1	1.00	—	—	—	—	—
p2	0.72	1.00	—	—	—	—
p3	0.41	0.73	1.00	—	—	—
p4	0.18	0.59	0.65	1.00	—	—
p5	0.97	0.21	0.38	0.35	1.00	—
p6	0.76	0.11	0.87	0.49	0.33	1.00

   Realiza el dendograma correspondiente al clustering jerárquico mediante los métodos de *complete link* y *single link* y comente los resultados.

2. **(3)** Considera la siguiente tabla que incluye la información de 10 transacciones realizadas en un cierto establecimiento:

Transacción	**Ítems**
1	{c, d, e}
2	{b, d, e}
3	{b, c, d}
4	{a, c, e}
5	{a, c, e, d}
6	{a, b, c, e}
7	{b, c, d, e}
8	{c, d}
9	{a, b, c, e}
10	{a, d, e}

Considera una soporte mínimo del 25%. Construye la rejilla correspondiente a todos los posibles conjuntos de ítems. Marca en la rejilla cada nodo con una F si es frecuente, una I si es infrecuente y una N si es podado por el algoritmo *Apriori*. Adicionalmente marca los nodos maximalmente frecuentes con una M y los nodos cerrados y frecuentes con una C.

Obtén la confianza de las reglas siguientes:

$$a, d \rightarrow b, c$$
$$b, c \rightarrow a$$
$$a, c \rightarrow b, d$$
$$\emptyset \rightarrow a, b, c$$
$$a \rightarrow c, e$$
$$a, c \rightarrow \emptyset$$

3. **(4)** Considera el conjunto de datos dónde cada instancia tiene 3 atributos, dos de ellos de tipo lógico y un tercero nominal y dos clases, c_1 y c_2. La siguiente tabla indica el número de instancias para cada una de las tres clases en función de los valores posibles de los atributos:

Atributos			**Número de instancias**	
x_1	x_2	x_3	c_1	c_2
V	V	a	36	7
V	V	b	10	44
V	V	c	7	1
V	F	a	7	16
V	F	b	6	8
V	F	c	6	34
F	V	a	28	10
F	V	b	11	8
F	V	c	5	4
F	F	a	3	18
F	F	b	34	3
F	F	c	15	0

Construye un árbol de decisión **binario** utilizando una estrategia voraz y cómo criterio de división de cada nodo el error de clasificación. Para crear una nueva división es necesario que el nuevo subárbol mejore al nodo padre. Calcula el error de entrenamiento del árbol completo.

Tiempo de realización: **6** horas. Calificación de cada ejercicio entre paréntesis.