1. Dado el siguiente conjunto de datos (entradas p_1 y p_2, salida y):

p_1	p_2	y
1	0	1
0	1	0
0	0	1
1	1	0

   a) ¿Cuál es la entropía del conjunto de dado?

   b) Calcule la ganancia sobre el atributo p_2.

   c) Grafique el árbol que resuelve el problema mediante el algoritmo ID3, puede resolverlo intuitivamente.

Soluciones

a) $-\frac{2}{4} \cdot \log_2 \left(\frac{2}{4}\right) - \frac{2}{4} \cdot \log_2 \left(\frac{2}{4}\right) = 1.$

b) COMPLETAR.

c) COMPLETAR.

2. Considerando el siguiente ejemplo de los Simpsons

Personaje	Longitud Pelo	Peso	Edad	Género
Homero	0	250	36	H
Bart	2	90	10	H
Abe	1	170	70	H
Otto	10	180	38	H
Kruty	6	200	45	H
Marge	10	150	34	M
Lisa	6	78	8	M
Maggie	4	20	1	M
Selma	8	160	41	M
Comic	8	290	38	?

   a) ¿Puede desarrollar un árbol de decisión que utilice sólo dos variables para determinar el género de un personaje en ese contexto? ¿Que valores de corte propondría para esas dos variables?

b) Resolver en forma intuitiva primero y luego fundamentar con ganancia de información.

Soluciones

a) Podemos considerar el peso y luego el peso. Observemos que si el peso mayor a 160 todos los personajes son hombres, luego basta separar al individuo restante por ejemplo si la longitud del pelo es menor que 4.

b) Observemos que:

- $entropia\,(S) = -\frac{5}{9}\cdot\log_2\left(\frac{5}{9}\right) - \frac{4}{9}\cdot\log_2\left(\frac{4}{9}\right) \approx 0,99.$
- $entropia\,(peso > 160) = -\frac{4}{4}\cdot\log_2\left(\frac{4}{4}\right) - 0 = 0.$
- $entropia\,(peso \leq 160) = -\frac{1}{5}\cdot\log_2\left(\frac{1}{5}\right) - \frac{4}{5}\cdot\log_2\left(\frac{4}{5}\right) \approx 0,722.$
- $entropia\,(peso \leq 160 \wedge pelo < 4) = -\frac{1}{1}\cdot\log_2\left(\frac{1}{1}\right) - 0 = 0.$
- $entropia\,(peso \leq 160 \wedge pelo \geq 4) = -\frac{4}{4}\cdot\log_2\left(\frac{4}{4}\right) - 0 = 0.$