

1. Construya el árbol de decisión de acuerdo al algoritmo ID3 para los siguientes datos donde se concluye dar o no un crédito respecto a un cliente bancario de acuerdo a ciertos atributos (nivel de ingresos, tipo de contrato, etc.).

Cliente	Moroso	Antigüedad	Ingresos	Trabajo Fijo	Conceder
1	si	> 5	600 – 1200	si	no
2	no	< 1	600 – 1200	si	si
3	si	1 – 5	> 1200	si	no
4	no	> 5	> 1200	no	si
5	no	< 1	> 1200	si	si
6	si	1 – 5	600 – 1200	si	no
7	no	1 – 5	> 1200	si	si
8	no	< 1	< 600	si	no
9	no	> 5	600 – 1200	no	no
10	si	1 – 5	< 600	no	no

### Solución

Primero analizamos el nodo raíz. Dado que hay ejemplos pertenecientes clases, no es nodo terminal y hay que encontrar el mejor atributo.

Para ello calculamos primero la entropía global sobre todos los ejemplos de este nodo (los 10 de la tabla). Hay 4 "si" y 6 "no". Por tanto, la entropía sería:

$$E_{global} = (-6/10 * \log_2 6/10) + (-4/10 * \log_2 4/10) = 0,97$$

Ahora analizamos los diferentes atributos disponibles:

Moroso:

Moroso=si: hay 4 casos de los cuales 4 son "no"

$$E_{si} = (-1 * \log_2 1) + (-0 * \log_2 0) = 0$$

Moroso=no: hay 6 casos de los cuales 4 son "si" y 2 son "no"

$$E_{no} = (-2/3 * \log_2 2/3) + (-1/3 * \log_2 1/3) \approx 0,92$$

Por tanto la entropía para Moroso ( la media ponderada de las dos) es:

$$E_{Moroso} = 4/10 * 0 + 6/10 * 0,92 \approx 0,54$$

Con ello se obtiene la ganancia para el atributo moroso:  $G_{moroso} = E_{global} - E_{moroso} \approx 0,43$

Antigüedad:

Antigüedad=< 1: 3 casos de los cuales 2 son "si" y 1 "no"

$$E_{<1} = (-2/3 * \log_2 2/3) + (-1/3 * \log_2 1/3) \approx 0,92$$

Antigüedad=1-5: hay 4 casos de los cuales 3 son "no" y 1 "si"

$$E_{1-5} = (-3/4 * \log_2 3/4) + (-1/4 * \log_2 1/4) \approx 0,81$$

Antigüedad=> 5: hay 3 casos de los cuales 2 son "no" y 1 "si"

$$E_{>5} = (-2/3 * \log_2 2/3) + (-1/3 * \log_2 1/3) \approx 0,92$$

Por tanto la entropía para Antigüedad ( la media ponderada de las tres) es:

$$E_{Antigüedad} = 3/10 * 0,92 + 4/10 * 0,81 + 3/10 * 0,92 \approx 0,87$$

Con ello se obtiene la ganancia para el atributo Antigüedad:  $G_{Antigüedad} = E_{global} - E_{Antigüedad} \approx 0,1$

Ingresos:

Ingresos=< 600: 2 casos de los cuales 2 son "no"

$$E_{<600} = (-1 * \log_2 1) + (-0 * \log_2 0) = 0$$

Ingresos=600-1200: hay 4 casos de los cuales 3 son "no" y 1 "si"

$$E_{600-1200} = (-3/4 * \log_2 3/4) + (-1/4 * \log_2 1/4) \approx 0,81$$

Ingresos=> 1200: hay 4 casos de los cuales 3 son "si" y 1 "no"

$$E_{>1200} = (-3/4 * \log_2 3/4) + (-1/4 * \log_2 1/4) \approx 0,81$$

Por tanto la entropía para Ingresos ( la media ponderada de las tres) es:

$$E_{Ingresos} = 2/10 * 0 + 8/10 * 0,81 \approx 0,64$$

Con ello se obtiene la ganancia para el atributo Antigüedad:  $G_{Ingresos} = E_{global} - E_{Ingresos} \approx 0,33$

TrabajoFijo:

TrabajoFijo=si: 7 casos de los cuales 3 son "si" y 4 "no"

$$E_{si} = (-3/7 * \log_2 3/7) + (-4/7 * \log_2 4/7) \approx 0,98$$

TrabajoFijo=no: hay 3 casos de los cuales 2 son "no" y 1 "si"

$$E_{no} = (-2/3 * \log_2 2/3) + (-1/3 * \log_2 1/3) \approx 0,92$$

Por tanto la entropía para TrabajoFijo ( la media ponderada de las dos) es:

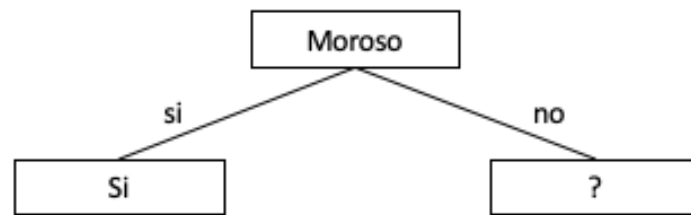
$$E_{TrabajoFijo} = 7/10 * 0,98 + 3/10 * 0,92 \approx 0,97$$

Con ello se obtiene la ganancia para el atributo TrabajoFijo:  $G_{TrabajoFijo} = E_{global} - E_{TrabajoFijo} \approx 0,01$

Dado este resultado, el mejor atributo (el de mayor ganancia) es el atributo Moroso y por ello se elige como atributo para el nodo raíz del árbol. Respecto al nodo de los ejemplos con Moroso=Si, este node es etiquetado con "NO" ya que todos estos ejemplos son del tipo "no". Con ello, el árbol queda en este momento de la siguiente forma:

Para el nodo de los ejemplos con Moroso=no, se determina nuevamente el mejor atributo. En este caso sólo se tienen en cuenta los 6 ejemplos que llegan a este nodo, es decir, los que tienen el valor Moros=no.

Nivel 2:



Calculamos de nuevo la entropía global, ahora sobre los 6 ejemplos no morosos de los cuales 4 son "si" y 2 son "no":

$$E_{global} = -2/3 * \log_2 2/3 + (-1/3 * \log_2 1/3) \approx 0,92$$

Ahora analizamos los diferentes atributos disponibles:

Antigüedad:

Antigüedad=< 1: 3 casos de los cuales 2 son "si" y 1 "no"

$$E_{<1} = (-2/3 * \log_2 2/3) + (-1/3 * \log_2 1/3) \approx 0,92$$

Antigüedad=1-5: 1 caso que es un "si"

$$E_{1-5} = 0$$

Antigüedad=> 5: hay 2 casos un "no" y un "si"

$$E_{>5} = (-1/2 * \log_2 1/2) + (-1/2 * \log_2 1/2) = 1$$

Por tanto la entropía para Antigüedad ( la media ponderada de las tres) es:

$$E_{Antigüedad} = 3/6 * 0,92 + 1/6 * 0 + 2/6 * 1 \approx 0,79$$

Con ello se obtiene la ganancia para el atributo Antigüedad:  $G_{Antigüedad} = E_{global} - E_{Antigüedad} \approx 0,13$

Ingresos:

Ingresos=< 600: 1 casos que es un "no"

$$E_{<600} = 0$$

Ingresos=600-1200: hay 2 casos un "no" y un "si"

$$E_{600-1200} = (-1/2 * \log_2 1/2) + (-1/2 * \log_2 1/2) = 1$$

Ingresos=> 1200: hay 3 casos de los cuales 3 son "si"

$$E_{>1200} = 0$$

Por tanto la entropía para Ingresos ( la media ponderada de las tres) es:

$$E_{Ingresos} = 4/6 * 0 + 2/6 * 1 \approx 0,33$$

Con ello se obtiene la ganancia para el atributo Antigüedad:  $G_{Ingresos} = E_{global} - E_{Ingresos} \approx 0,59$

TrabajoFijo:

TrabajoFijo=si: 4 casos de los cuales 3 son "si" y 1 "no"

$$E_{si} = (-3/4 * \log_2 3/4) + (-1/4 * \log_2 1/4) \approx 0,81$$

TrabajoFijo=no: hay 2 casos uno es "no" y uno es "si"

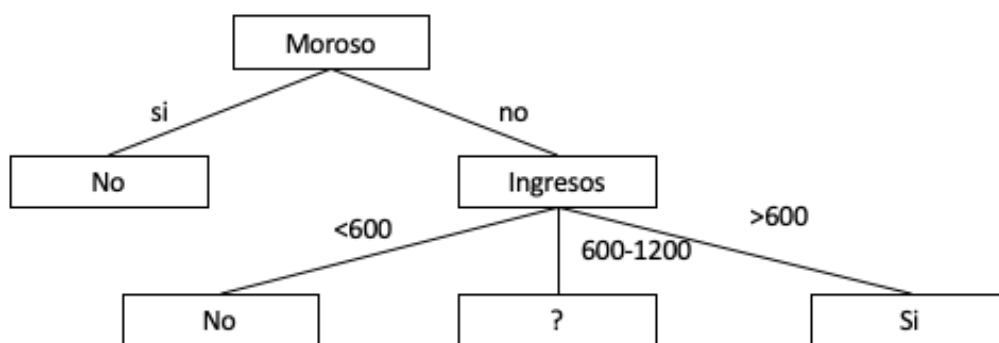
$$E_{no} = (-1/2 * \log_2 1/2) + (-1/2 * \log_2 1/2) = 1$$

Por tanto la entropía para TrabajoFijo ( la media ponderada de las dos) es:

$$E_{TrabajoFijo} = 4/6 * 0,81 + 2/6 * 1 \approx 0,87$$

Con ello se obtiene la ganancia para el atributo TrabajoFijo:  $G_{TrabajoFijo} = E_{global} - E_{TrabajoFijo} \approx 0,05$

Se observa que el mejor atributo es Ingresos y se elige este atributo. Dos de las nuevas ramas del árbol se resuelven (terminan) porque todos sus ejemplos son de la misma clase y el árbol queda de la siguiente forma:



Ahora queda encontrar el mejor atributo para el nodo que todavía no tiene etiqueta. A éste nodo sólo llegan dos ejemplo: el 2 y el 9.

Nivel 3.

Calculamos de nuevo la entropía global, ahora sobre los 2 ejemplos un "si" y un "no":

$$E_{global} = 1$$

Ahora analizamos los diferentes atributos disponibles:

TrabajoFijo:

TrabajoFijo=si: 1 casos que es "si"

$$E_{si} = 0$$

TrabajoFijo=no: 1 caso que es un "no"

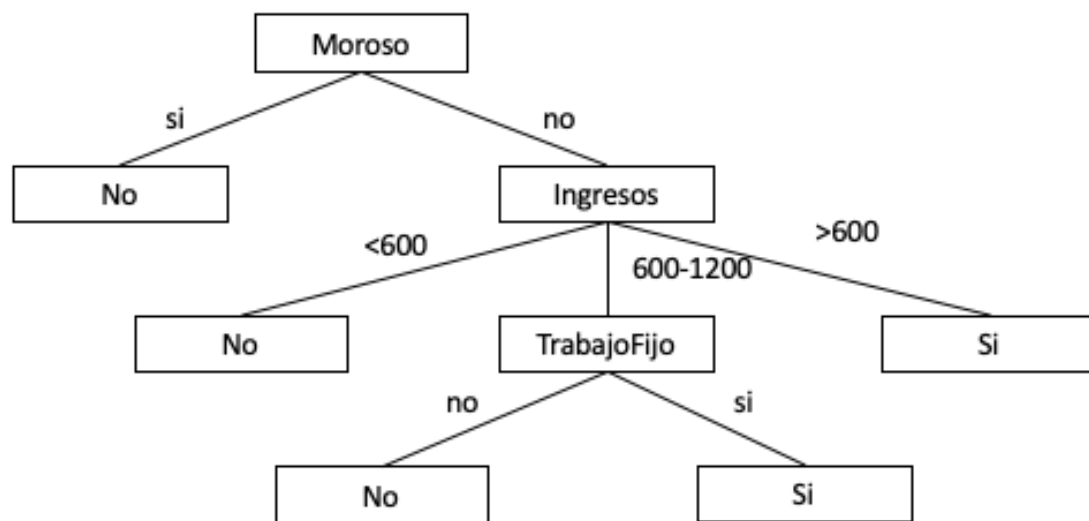
$$E_{no} = 0$$

Por tanto la entropía para TrabajoFijo ( la media ponderada de las dos) es:

$$E_{TrabajoFijo} = 0$$

Con ello se obtiene la ganancia para el atributo TrabajoFijo:  $G_{TrabajoFijo} = E_{global} - E_{TrabajoFijo} = 1$

Dado que esta ganancia es máxima, no es necesario realizar el cálculo para el atributo Antigüedad (no podría tener mayor ganancia) y se puede elegir el atributo TrabajoFijo. Con ello, y determinando las clases de los nuevos nodos el resultado final quedaría de la siguiente forma:



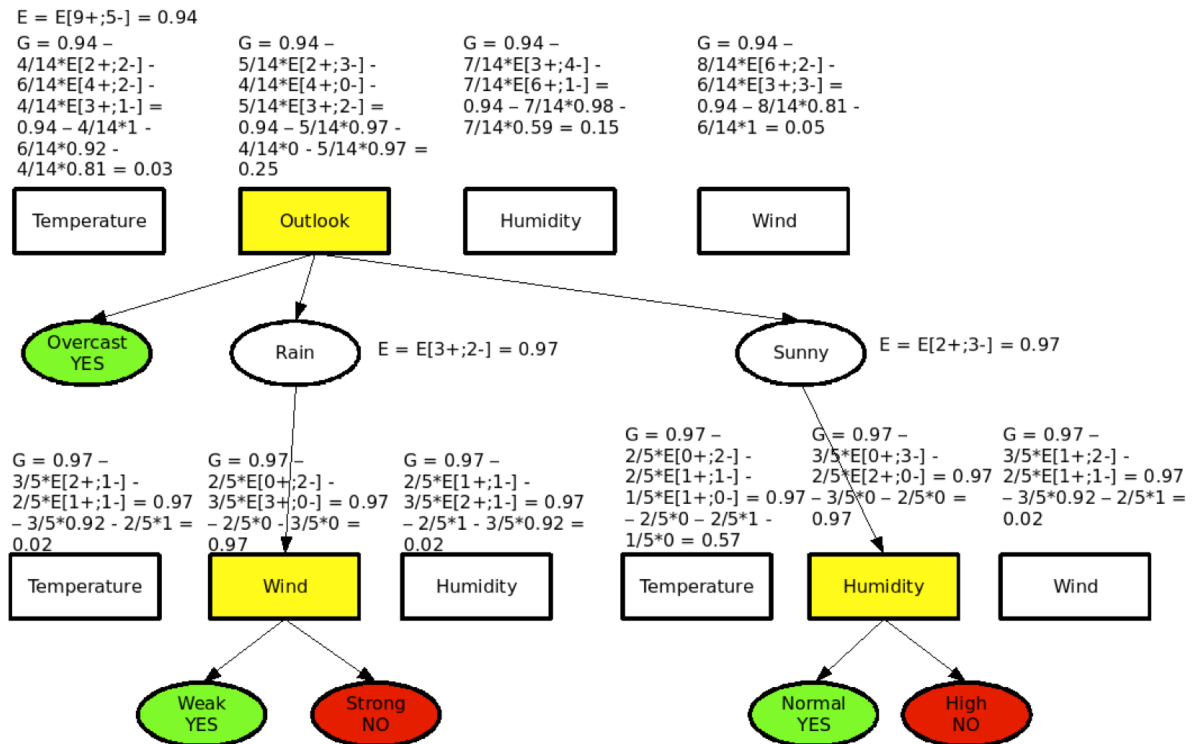
Finalmente cabe observar que en el último paso, si se hubiera analizado el atributo Antigüedad, se hubiera obtenido la misma ganancia. En este caso, ambos atributos tienen la misma importancia, por lo que da igual cual de los dos elegir. Sin embargo, con el atributo Antigüedad, se hubiera tenido un nodo para el cual no existen ejemplos. En este caso se debería decidir que clase se devolvería en este nodo.

2. Construya el árbol de decisión de acuerdo al algoritmo ID3 para los siguientes datos donde se concluye si jugar o no al tenis de acuerdo a las condiciones del tiempo.

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

## Solución

El proceso de creación del árbol ID3 se resume en la siguiente figura:



3. En una tienda de electrodomésticos quieren evaluar la calidad de los productos que venden. Para ello han hecho una entrevista a los clientes de la tienda cuyos resultados se detallan en la siguiente tabla:

Cliente	Marca	Material	Calidad
1	UFESA	Fibra	Media
2	LG	Plástico	Baja
3	Siemens	Metal	Alta
4	LG	Metal	Alta
5	UFESA	Plástico	Baja
6	Siemens	Fibra	Media
7	Siemens	Metal	Media
8	UFESA	Metal	Media

Usando estos ejemplos, desarrolle un árbol de decisión que permite predecir la calidad de los electrodomésticos. Utiliza el algoritmo presentado en clase, indicando en cada paso la expresión y el valor de la ganancia de cada atributo.

Se puede emplear un “cálculo aproximado” basado en las siguientes igualdades:

$\log_2 1 = 0$	$\log_2 1/6 = -2,6$	$\log_2 5/8 = -0,7$
$\log_2 1/2 = -1$	$\log_2 5/6 = -0,25$	$\log_2 7/8 = -0,2$
$\log_2 1/3 = -1,6$	$\log_2 1/7 = -2,8$	
$\log_2 2/3 = -0,6$	$\log_2 2/7 = -1,8$	
$\log_2 1/4 = -2$	$\log_2 3/7 = -1,2$	
$\log_2 3/4 = -0,4$	$\log_2 4/7 = -0,8$	
$\log_2 1/5 = -2,3$	$\log_2 5/7 = -0,5$	
$\log_2 2/5 = -1,3$	$\log_2 6/7 = -0,2$	
$\log_2 3/5 = -0,7$	$\log_2 1/8 = -3$	
$\log_2 4/5 = -0,3$	$\log_2 3/8 = -1,4$	



**Solución**

$$E_{global} = 2(-2/8 * \log_2 2/8) + (-4/8 * \log_2 4/8) = 1,5$$

Marca:

$$E_{Ufesa} = (-2/3 * \log_2 2/3) + (-1/3 * \log_2 1/3) + (-0 * \log_2 0) = (2/3 * 0,6) + (1/3 * 1,6) \approx 0,4 + 0,5 = 0,9$$

$$E_{Siemens} = (-2/3 * \log_2 2/3) + (-1/3 * \log_2 1/3) + (-0 * \log_2 0) = (2/3 * 0,6) + (1/3 * 1,6) \approx 0,4 + 0,5 = 0,9$$

$$E_{LG} = 2(-1/2 * \log_2 1/2) + (-0 * \log_2 0) = 1$$

$$E_{marca} = 3/8 * 0,9 + 3/8 * 0,9 + 2/8 * 1 \approx 0,93$$

ganancia 0,57

Material:

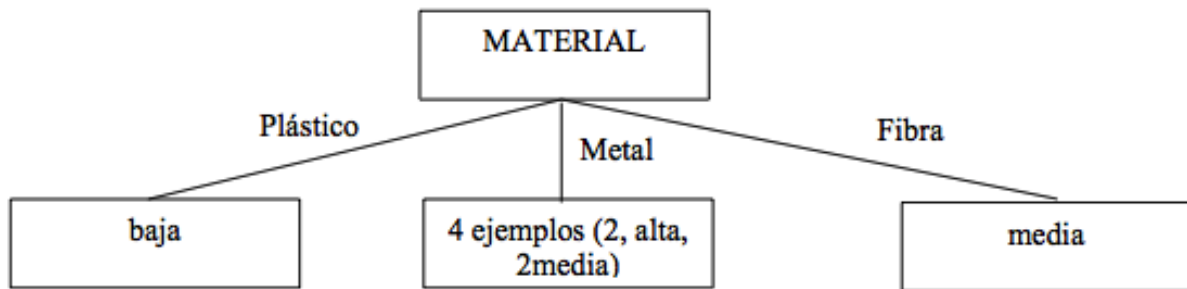
$$E_{Plastico} = (-1 * \log_2 1) + 2(-0 * \log_2 0) = 0$$

$$E_{Metal} = 2(-2/4 * \log_2 2/4) = 1$$

$$E_{Fibra} = (-1 * \log_2 1) + 2(-0 * \log_2 0) = 0$$

$$E_{material} = 4/8 * 0 + 4/8 * 1 = 0,5$$

ganancia=1 (mayor ganancia)



A partir de este punto, aunque sólo queda un único atributo (marca), se debe calcular de nuevo la ganancia que se obtendría al clasificar por este atributo. Si la ganancia fuese 0, no se debería abrir más el árbol por la rama “metal”.

Calculamos la ganancia que se obtendría al clasificar la rama metal por el atributo marca:

$$E_{global} = 1$$

Marca:

$$E_{LG} = 0$$

$$E_{Siemens} = 1$$

$$E_{Ufesa} = 0$$

$$E_{marca} = 0,5$$

ganancia=0,5

Como hay una ganancia positiva, se debe seguir desarrollando el árbol.

En el nodo metal/siemens, no hay un resultado único, pero sí tenemos que terminar el árbol en este nodo, ya que no hay más atributos para diferenciar. Como se ha comentado en la clase de teoría, en este caso se podría aplicar el criterio de devolver “Alta” o “Media” de forma aleatoria con una probabilidad de 0,5 cada uno, pues esta probabilidad corresponde a los ejemplos que llegan a este nodo. Sin embargo, si consideramos el contexto de este ejercicio, no parece muy razonable esta idea. En este ejercicio se pretende dar información a una tienda sobre cómo evalúa la gente sus productos. Lógicamente estas evaluaciones son subjetivos y la tienda estará interesada en saber estas evaluaciones al máximo. Por tanto, en este caso, parece más lógico asignar a este nodo incluso un nuevo concepto, tipo “media-alta”, ya que eso reflejaría mejor el resultado. La situación del ejercicio es muy diferente a, por ejemplo, un caso, donde un sistema debe tomar decisiones de forma autónoma, por ejemplo apagar o encender una máquina. En este último caso, sí puede ser necesario que el sistema decida claramente entre una u otra acción, y por tanto, el criterio de decidir basado en probabilidades parece razonable.

