EJERCICIOS DE ÁRBOLES DE DECISIÓN

Ejercicio 1:

Supongamos que queremos clasificar si un globo en particular, está inflado en función de los siguientes cuatro atributos: color del globo(*color*), tamaño del globo(*size*), 'acto' en el que está involucrada la persona que sostiene el globo(*Act*) y edad de la persona que sostiene el globo(*Age*).
Según los datos de entrenamiento facilitados en la tabla, usar el algoritmo ID3 para obtener un árbol de decisión.
¿Cuál es la clase que se puede predecir para la muestra de prueba (*Purple, Large, Dip, Child*)?

Example	Color	Size	Act	Age	Inflated?
1	Yellow	Small	Stretch	Adult	T
2	Yellow	Small	Stretch	Child	T
3	Yellow	Small	Dip	Adult	T
4	Yellow	Small	Dip	Child	F
5	Yellow	Small	Dip	Child	F
6	Yellow	Large	Stretch	Adult	T
7	Yellow	Large	Stretch	Child	T
8	Yellow	Large	Dip	Adult	T
9	Yellow	Large	Dip	Child	F
10	Yellow	Large	Dip	Child	F
11	Purple	Small	Stretch	Adult	T
12	Purple	Small	Stretch	Child	T
13	Purple	Small	Dip	Adult	T
14	Purple	Small	Dip	Child	F
15	Purple	Small	Dip	Child	F
16	Purple	Large	Stretch	Adult	T
17	Purple	Large	Stretch	Child	T
18	Purple	Large	Dip	Adult	T
19	Purple	Large	Dip	Child	F
20	Purple	Large	Dip	Child	F

Ejercicio 2:

Consideremos los datos de capacitación que se proporcionan a continuación para una tarea de evaluación del riesgo crediticio (*Risk* es el atributo objetivo). Utilice el algoritmo ID3 para obtener un árbol de decisión. ¿Cuál es la clase prevista para la muestra de prueba (*bad, low, adequate, over $35K*)?

Example	Risk	History	Debt	Collateral	Income
1	high	bad	high	none	$0-15K
2	high	unk	high	none	$15-35K
3	mod	unk	low	none	$15-35K
4	high	unk	low	none	$0-15K
5	low	unk	low	none	over $35K
6	low	unk	low	adequate	over $35K
7	high	bad	low	none	$0-15K
8	mod	bad	low	adequate	over $35K
9	low	good	low	none	over $35K
10	low	good	high	adequate	over $35K
11	high	good	high	none	$0-15K
12	mod	good	high	none	$15-35K
13	low	good	high	none	over $35K
14	high	bad	high	none	$15-35K

Ejercicio 3:

Consideremos los datos de entrenamiento que se dan a continuación para predecir el sexo de una persona (*Sexo* es el atributo objetivo). Utilice el algoritmo ID3 para obtener un árbol de decisión. ¿Cuál es la clase predicha para la muestra de prueba (*Long, High, Young*)?

Person	Hair length	Weight	Age	Class
1	Short	High	Young	Male
2	Long	Low	Young	Female
3	Short	Low	Young	Male
4	Long	Low	Young	Female
5	Short	Low	Young	Female
6	Short	High	Old	Male
7	Long	Low	Old	Female
8	Long	High	Young	Male
9	Long	High	Old	Male

Ejercicio 4:

Consideremos los datos de entrenamiento que se dan a continuación para predecir si una persona se quema con el sol (*Result* es el atributo objetivo). Utilice el algoritmo ID3 para obtener un árbol de decisión. ¿Cuál es la clase predicha para la muestra de prueba (*red, tall, average, yes*)?

Person	Hair	Height	Weight	Lotion	Result
1	blonde	average	light	no	sunburned
2	blonde	tall	average	yes	none
3	brown	short	average	yes	none
4	blonde	short	average	no	sunburned
5	red	average	heavy	no	sunburned
6	brown	tall	heavy	no	none
7	brown	average	heavy	no	none
8	blonde	short	light	yes	none

Ejercicio 5:

Dado que ID3 es un procedimiento de búsqueda local en el espacio de los árboles de decisión, es posible que no encuentre el árbol óptimo en términos de longitud y precisión en el conjunto de entrenamiento. Dé un ejemplo de un conjunto de datos para el cual ID3 no devuelve el árbol más corto que clasifica correctamente todos los datos de entrenamiento.

Ejercicio 6:

Dé un ejemplo de un conjunto de datos para el que no hay un árbol de decisión que pueda clasificar correctamente todos los datos de entrenamiento.

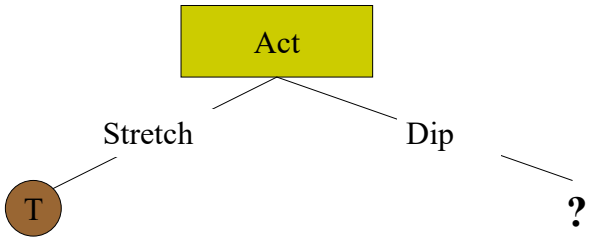
POSIBLES SOLUCIONES:

Ejercicio 1:

La entropía inicial es $H(\{\mathbf{x}_1,\ldots,\mathbf{x}_{20}\})=0.971$. Calculamos la ganancia de información para dividir cada atributo en el nodo raíz:

$$IG(\{\mathbf{x}_1,\ldots,\mathbf{x}_{20}\},Color)=0.971-(10/20)*0.971-(10/20)*0.971=0$$
$$IG(\{\mathbf{x}_1,\ldots,\mathbf{x}_{20}\},Size)=0.971-(10/20)*0.971-(10/20)*0.971=0$$
$$IG(\{\mathbf{x}_1,\ldots,\mathbf{x}_{20}\},Act)=0.971-(8/20)*0-(12/20)*0.918=0.42$$
$$IG(\{\mathbf{x}_1,\ldots,\mathbf{x}_{20}\},Age)=0.971-(8/20)*0-(12/20)*0.918=0.42$$

En consecuencia, podemos elegir Act o Age para dividir en el nodo raíz. Si elegimos Act, obtenemos:



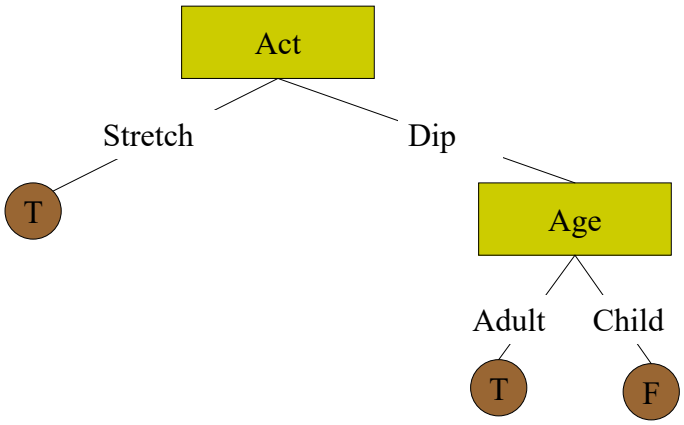
Ahora debemos elegir un atributo para dividir en el nodo *Act=Dip,* donde $S_{Dip}=\{\ \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_8, \mathbf{x}_9, \mathbf{x}_{10}, \mathbf{x}_{13}, \mathbf{x}_{14}, \mathbf{x}_{15}, \mathbf{x}_{18}, \mathbf{x}_{19}, \mathbf{x}_{20}\ \}$:

Example	Color	Size	Age	Inflated?
3	Yellow	Small	Adult	T
4	Yellow	Small	Child	F
5	Yellow	Small	Child	F
8	Yellow	Large	Adult	T
9	Yellow	Large	Child	F
10	Yellow	Large	Child	F
13	Purple	Small	Adult	T
14	Purple	Small	Child	F
15	Purple	Small	Child	F
18	Purple	Large	Adult	T
19	Purple	Large	Child	F
20	Purple	Large	Child	F

La entropía inicial es $H(S_{Dip})=0.918$.

$$IG(S_{Dip},Color)=0.918-(6/12)*0.918-(6/12)*0.918=0$$
$$IG(S_{Dip},Size)=0.918-(6/12)*0.918-(6/12)*0.918=0$$
$$IG(S_{Dip},Age)=0.918-(4/12)*0-(8/12)*0=0.918$$

Consecuentemente debemos elegir *Age* para dividir en el nodo *Act=Dip*, y así obtenemos el árbol final:



La clase predicha para la muestra de prueba (*Purple*, *Large*, *Dip*, *Child*) es F (not inflated), ya que *Act=Dip* y *Age=Child*.

Ejercicio 2:

La entropía inicial es $H(\{\mathbf{x}_1,\ldots,\mathbf{x}_{14}\})=1.531$. Calculamos la ganancia de información para dividir cada atributo en el nodo raíz:

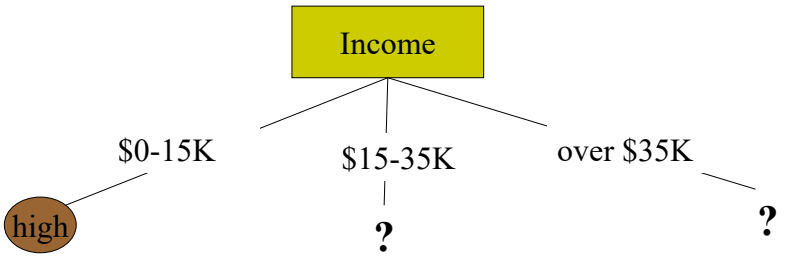
$IG(\{\mathbf{x}_1,\ldots,\mathbf{x}_{14}\},History)= 1.531–(4/14)*0.811–(5/14)*1.371–(5/14)*1.522=0.266$

$IG(\{\mathbf{x}_1,\ldots,\mathbf{x}_{14}\},Debt)= 1.531–(7/14)*1.379–(7/14)*1.557=0.063$

$IG(\{\mathbf{x}_1,\ldots,\mathbf{x}_{14}\},Collateral)= 1.531–(3/14)*0.918–(11/14)*1.435=0.207$

$IG(\{\mathbf{x}_1,\ldots,\mathbf{x}_{14}\},Income)=1.531–(4/14)*0–(4/14)*1–(6/14)*0.650=0.967$

Consecuentemente debemos elegir *Income* para dividir en el nodo raiz:



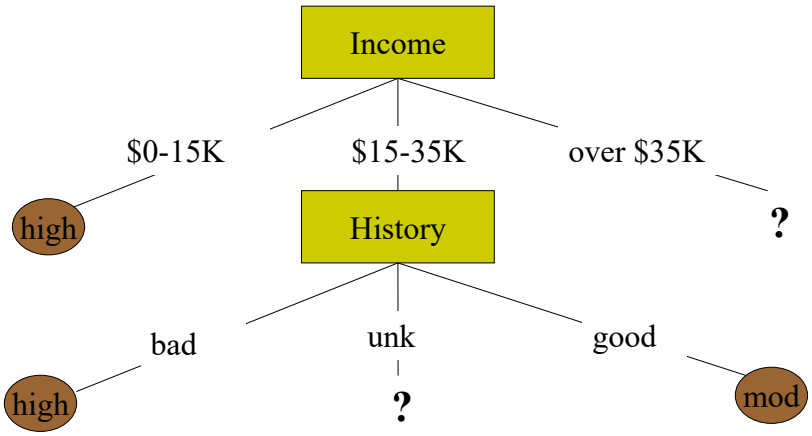
Ahora debemos elegir un atributo para dividir en el nodo *Income=$15-35K*, donde $S_{\$15\text{-}35K}=\{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_{12}, \mathbf{x}_{14}\}$. La entropía inicial es $H(S_{\$15\text{-}35K})=1$.

$IG(S_{\$15\text{-}35K}, History)=1–(1/4)*0–(1/4)*0–(2/4)*1=0.5$

$IG(S_{\$15\text{-}35K}, Debt)=1–(3/4)*0.918–(1/4)*0= 0.311$

$IG(S_{\$15\text{-}35K}, Collateral)=1–(4/4)*1=0$

Consecuentemente debemos elegir *History* para dividir en el nodo *Income=$15-35K*:

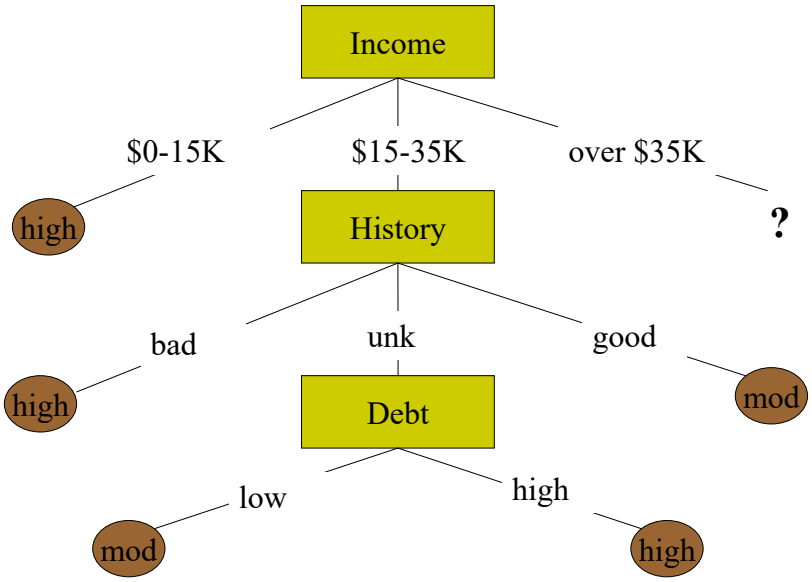


A continuación, debemos elegir un atributo para dividir en el nodo *History=unk*, donde S_{unk}={ $\mathbf{x}_2$, $\mathbf{x}_3$ }. La entropía inicial es $H(S_{unk})$=1.

$IG(S_{unk}, Debt)$=1–(1/2)*0–(1/2)*0=1

$IG(S_{unk}, Collateral)$=1–(2/2)*1=0

Consecuentemente debemos elegir *Debt* para dividir en el nodo *History=unk*:



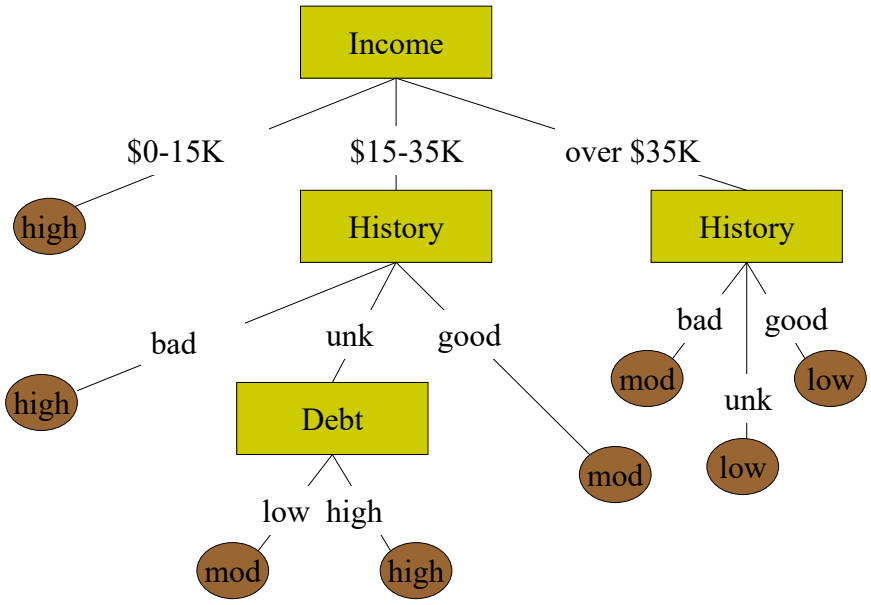
Ahora debemos elegir un atributo para dividir en el nodo *Income=over $35K*, donde $S_{over\ \$35K}$={ $\mathbf{x}_5$, $\mathbf{x}_6$, $\mathbf{x}_8$, $\mathbf{x}_9$, $\mathbf{x}_{10}$, $\mathbf{x}_{13}$ }. La entropía inicial es $H(S_{over\ \$35K})$=0.65.

$IG(S_{over\ \$35K}, History)$=0.65–(2/6)*0–(3/6)*0–(1/6)*0=0.65

$IG(S_{over\ \$35K}, Debt)$= 0.65–(4/6)*0.811–(2/6)*0=0.109

$IG(S_{over\ \$35K}, Collateral)$= 0.65–(3/6)*0–(3/6)*0.918=0.191

Consecuentemente debemos elegir *History* para dividir en el nodo *Income=over $35K*. El árbol final queda de la siguiente manera:



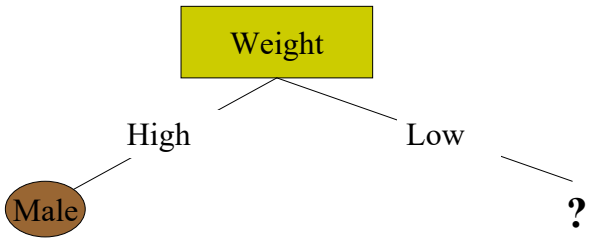
La clase predicha para la muestra de prueba (*bad, low, adequate, over $35K*) es *mod*.

Ejercicio 3:

La entropía inicial es $H(\{\mathbf{x}_1,\ldots,\mathbf{x}_9\})=0.991$. Calculamos la ganancia de información para dividir cada atributo en el nodo raíz:

$$IG(\{\mathbf{x}_1,\ldots,\mathbf{x}_9\},Hair)= 0.991–(5/9)*0.971–(4/9)*0.811=0.091$$
$$IG(\{\mathbf{x}_1,\ldots,\mathbf{x}_9\},Weight)= 0.991–(4/9)*0–(5/9)*0.722=0.590$$
$$IG(\{\mathbf{x}_1,\ldots,\mathbf{x}_9\},Age)= 0.991–(3/9)*0.918–(6/9)*1=0.018$$

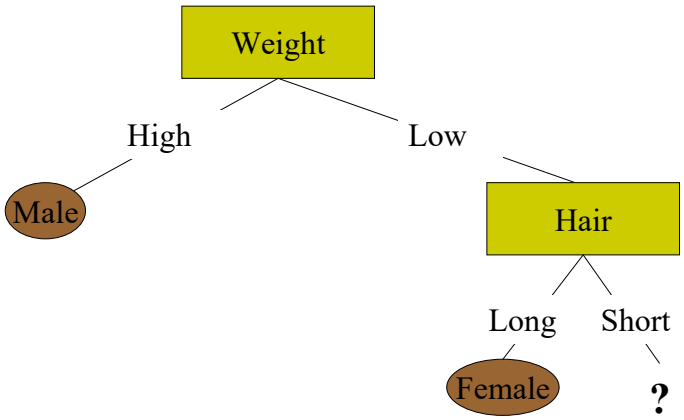
En consecuencia, podemos elegir *Weight* para dividir en el nodo raíz:



Ahora debemos elegir un atributo para dividir en el nodo *Weight=Low*, donde $S_{Low}=\{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_7\}$. La entropía inicial es $H(S_{Low})=0.722$.

$$IG(S_{Low}, Hair)= 0.722–(3/5)*0–(2/5)*1=0.322$$
$$IG(S_{Low}, Age)= 0.722–(1/5)*0–(4/5)*0.811=0.073$$

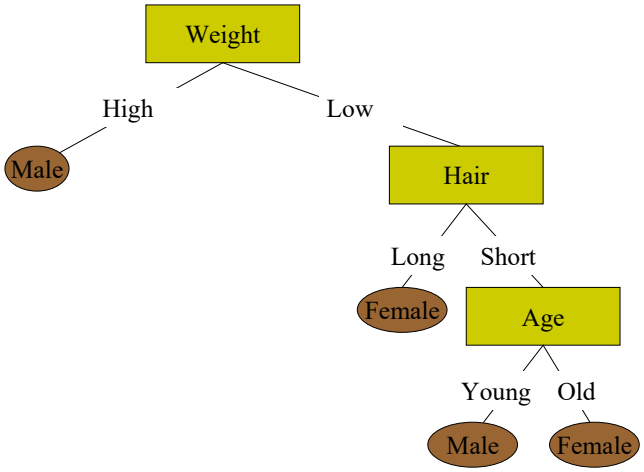
Consecuentemente debemos elegir *Hair* para dividir en el nodo *Weight=Low*:



Ahora, el único atributo possible para dividir en el nodo *Hair=Long* es *Age*, donde S_{Long}={ $\mathbf{x}_3$, $\mathbf{x}_5$ }. Pero encontramos que:

1) no hay ejemplos para *Age=Old*;
2) no hay un valor objetivo más frecuente (hay un ejemplo de cada clase).

Una posible solución para romper este empate es la siguiente:



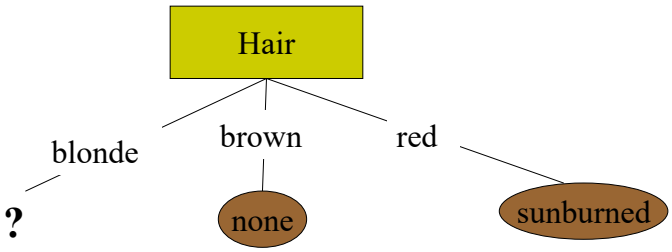
La clase predicha para la muestra de prueba (*Long*, *High*, *Young*) es *Male*.

Ejercicio 4:

La entropía inicial es $H(\{\mathbf{x}_1,\ldots,\mathbf{x}_8\})$=0.954. Calculamos la ganancia de información para dividir cada atributo en el nodo raíz:

$$IG(\{\mathbf{x}_1,\ldots,\mathbf{x}_8\},Hair)= 0.954–(4/8)*1–(3/8)*0–(1/8)*0=0.454$$
$$IG(\{\mathbf{x}_1,\ldots,\mathbf{x}_8\},Height)= 0.954–(3/8)*0.918–(3/8)*0.918–(2/8)*0=0.265$$
$$IG(\{\mathbf{x}_1,\ldots,\mathbf{x}_8\},Weight)= 0.954–(3/8)*0.918–(3/8)*0.918–(2/8)*1=0.015$$
$$IG(\{\mathbf{x}_1,\ldots,\mathbf{x}_8\},Lotion)= 0.954–(5/8)*0.971–(3/8)*0=0.347$$

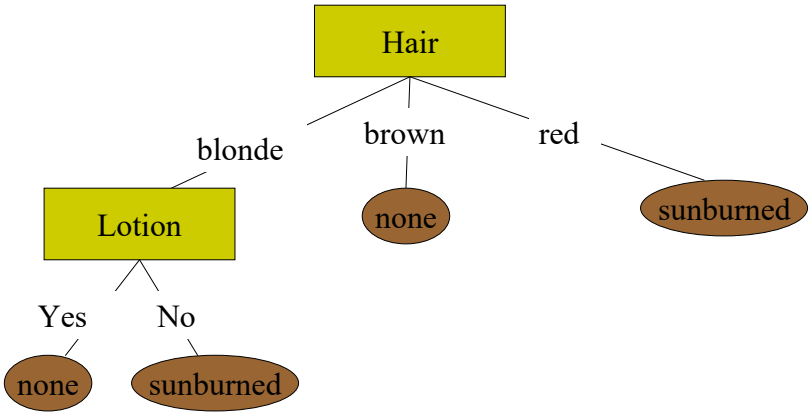
Consecuentemente debemos elegir *Hair* para dividir en el nodo raíz:



Ahora debemos elegir un atributo para dividir en el nodo *Hair=blonde* node, donde S_{blonde}={ $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_4$, $\mathbf{x}_8$ }. La entropía inicial es $H(S_{blonde})=1$.

$$IG(S_{blonde},Height)= 1–(1/4)*0–(2/4)*1–(1/4)*0=0.5$$
$$IG(S_{blonde},Weight)= 1–(2/4)*1–(2/4)*1=0$$
$$IG(S_{blonde},Lotion)= 1–(2/4)*0–(2/4)*0=1$$

Consecuentemente debemos elegir *Lotion* para dividir en el nodo *Hair=blonde*. El árbol final queda de la siguiente manera:



La clase predicha para la muestra de prueba (*red*, *tall*, *average*, *yes*) es *sunburned*.

Ejercicio 5:

A continuación se proporciona un posible conjunto de datos. El árbol de decisión más pequeño tiene una profundidad dos y se logra ramificando primero en A1 y luego en A2. Sin embargo, ID3 elige A3 y produce un árbol de decisión de profundidad tres. Los detalles se dejan al lector.

Example	A1	A2	A3	Class
1	T	T	T	+
2	T	T	T	+
3	T	F	F	-
4	T	F	T	-
5	F	F	T	+
6	F	F	F	+
7	F	T	F	-
8	F	T	T	-

Ejercicio 6:

A continuación se proporciona un posible conjunto de datos. Tenga en cuenta que los ejemplos 1 y 2 tienen los mismos valores para el vector de entrada (A1, A2, A3), mientras que los valores del objetivo son diferentes.

Example	A1	A2	A3	Class
1	T	F	T	+
2	T	F	T	-
3	T	F	F	-
4	T	T	T	-
5	F	F	T	+
6	F	F	F	+
7	F	T	F	-
8	F	T	T	-