



INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL CLASE 3 PARTE 2

AGENDA

- ARBOLES DE DECISIÓN (APRENDIZAJE SUPERVISADO)
 - APLICACIÓN
 - EJEMPLO
 - ALGORITMO BÁSICO
 - MÉTODOS DE SELECCIÓN DE ATRIBUTOS.

DATOS DE ENTRENAMIENTO Y DE TESTEO

El conjunto de datos original se dividirá en dos partes

■ Conjunto de datos de entrenamiento

- Se utilizarán para construir el modelo. Como el aprendizaje es supervisado el método buscará ajustar su respuesta a lo indicado en estos ejemplos.

■ Conjunto de datos de testeo

- Una vez construido el modelo será utilizado para medir su calidad.
- Se espera que la respuesta del modelo coincida lo más posible con lo indicado en estos ejemplos.

ARBOL DE DECISIÓN

- Es un **modelo de predicción muy utilizado.**
 - Por su **forma jerárquica, permite visualizar la organización de los atributos.**
 - **Se construye a partir de la identificación sucesiva de los atributos más relevantes.**
-

ARBOL DE DECISIÓN – APLICACIONES

■ Predicción

- Recorriendo sus ramas **se obtienen reglas que permiten tomar decisiones.**
- **Si todas las hojas se refieren al mismo atributo y es discreto** es un árbol de **clasificación.**

■ Descripción

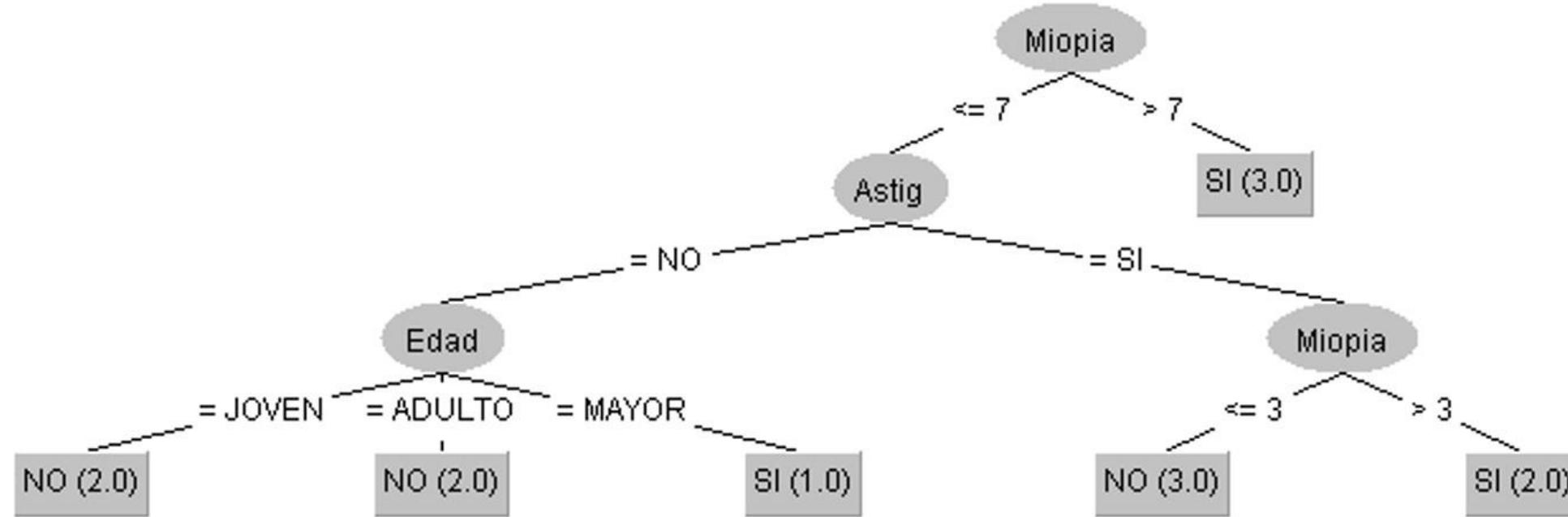
- Su **estructura jerárquica les permite mostrar cómo está organizada la información disponible.**

ARBOL DE DECISIÓN. EJEMPLO

- Suponga que se dispone de la **siguiente información de pacientes tratados previamente por problemas visuales**
 - Edad
 - Astigmatismo (si o no)
 - Grado de miopía
 - Recomendación de operarse (si o no)

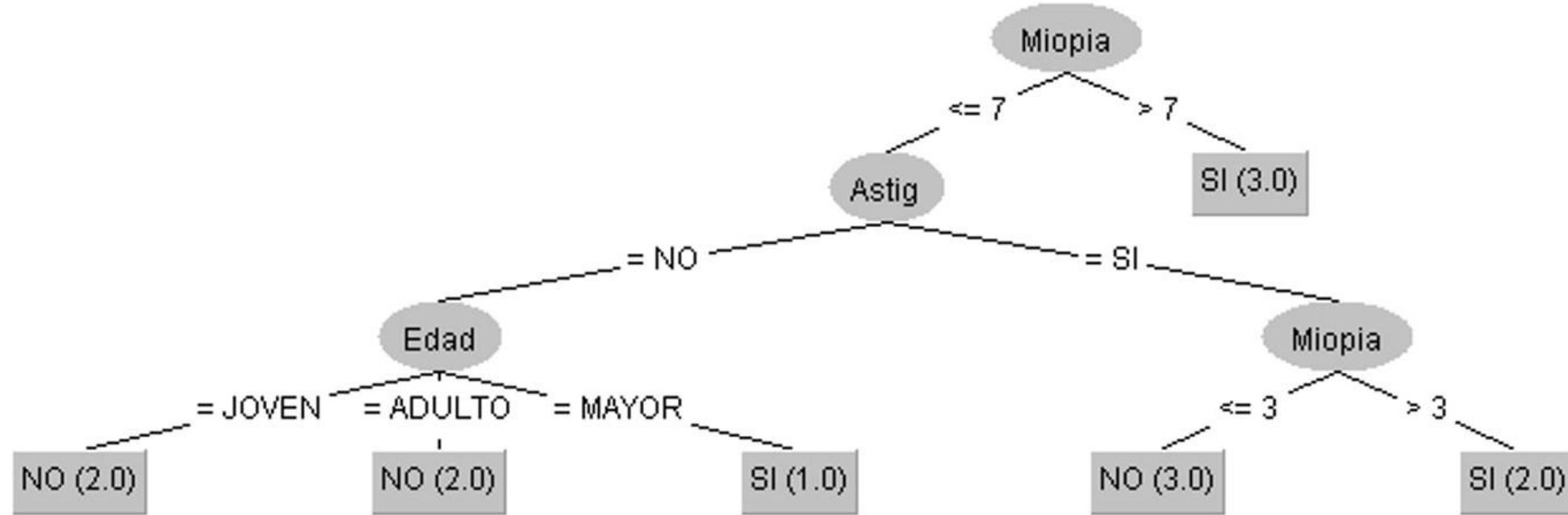
A partir de esta información puede obtenerse un **modelo** en forma de **árbol** que resuma el criterio seguido para recomendar si debe operarse o no.

ARBOL DE DECISIÓN. EJEMPLO



Note que las opciones son excluyentes

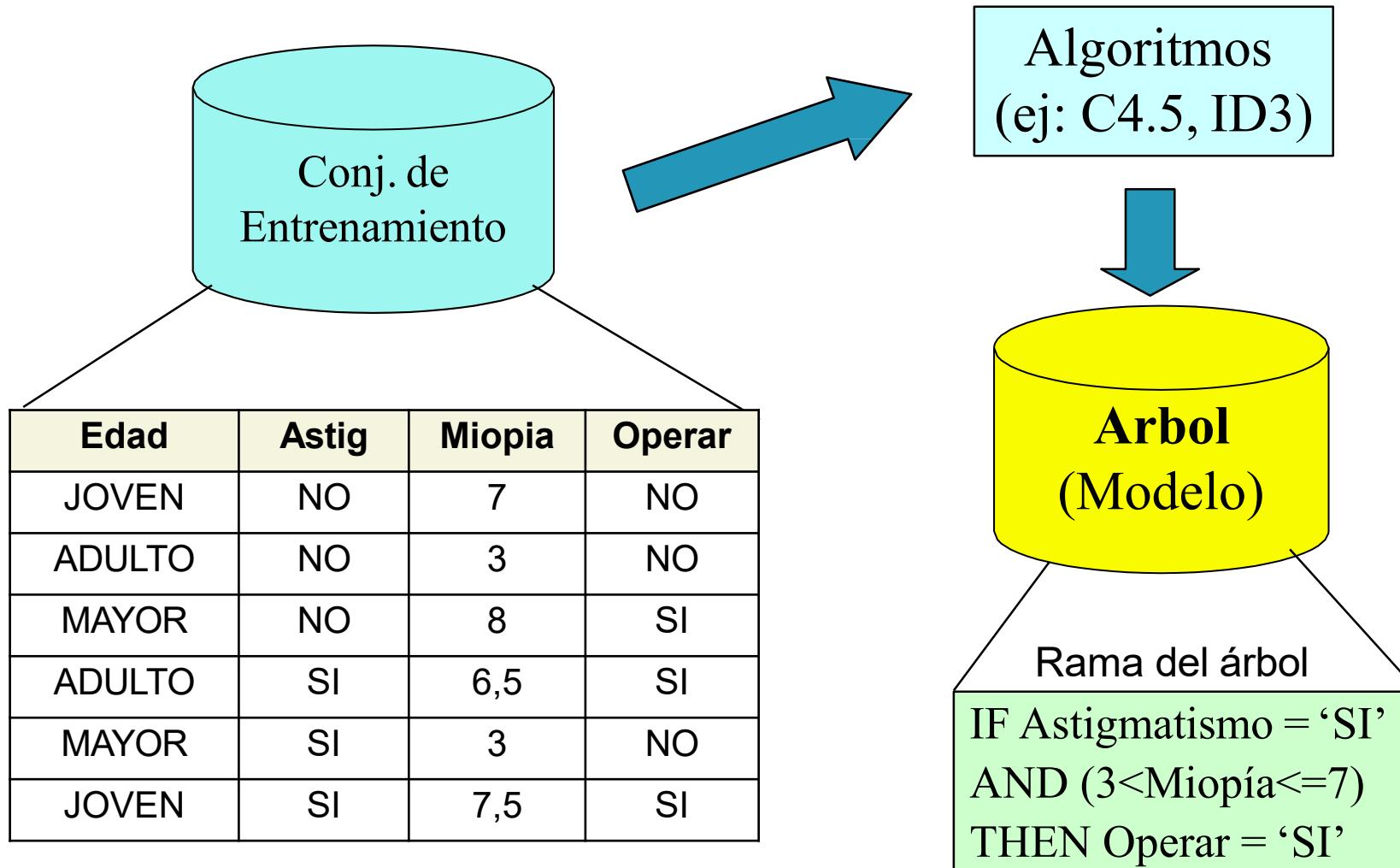
ARBOL DE DECISIÓN. EJEMPLO



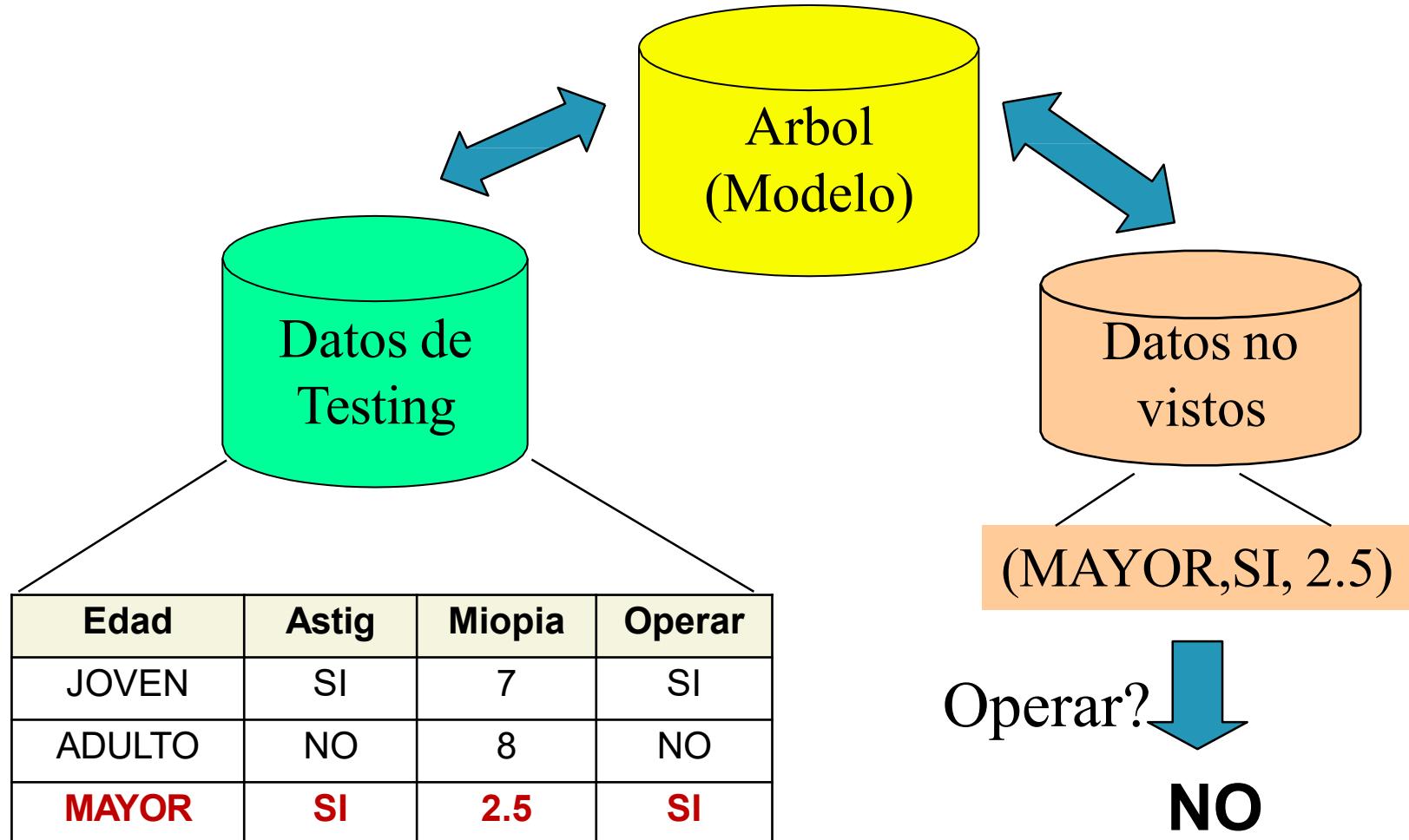
Un atributo cualitativo no
puede aparecer más de una
vez en la misma rama

y un atributo
cuantitativo?

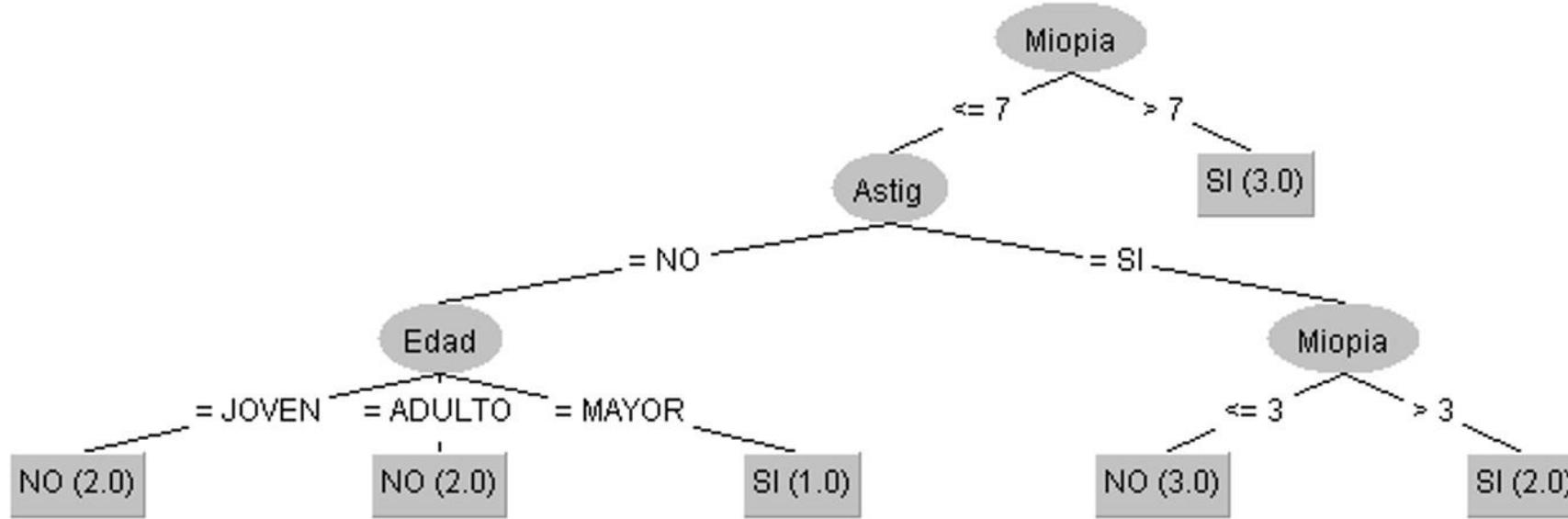
OBTENCIÓN DEL MODELO



USO DEL MODELO



ARBOLES COMO REGLAS



Si (Miopia>7) entonces (SeOpera=SI)

Si (Miopia<=7) y (Astig=SI) y (Miopía >3) entonces (SeOpera=SI)

Si (Miopia<=7) y (Astig=NO) y (Edad=MAYOR) entonces (SeOpera=SI)

EN OTRO CASO NO

OBTENCIÓN DEL ÁRBOL DE DECISIÓN - ALGORITMO BÁSICO

- El árbol **se construye de la forma top-down y recursiva**. Bajo la filosofía **divide-and-conquer** (**divide y vencerás**).
- Al comienzo, todos los ejemplos de entrenamiento están en el **nodo raíz**.
- Los ejemplos se **particionan recursivamente** basado en los atributos seleccionados.
- Los **atributos se seleccionan en base a una medida estadística** (p.ej., **ganancia de información**).

OBTENCIÓN DEL ÁRBOL DE DECISIÓN

■ **Condiciones para detener el particionamiento**

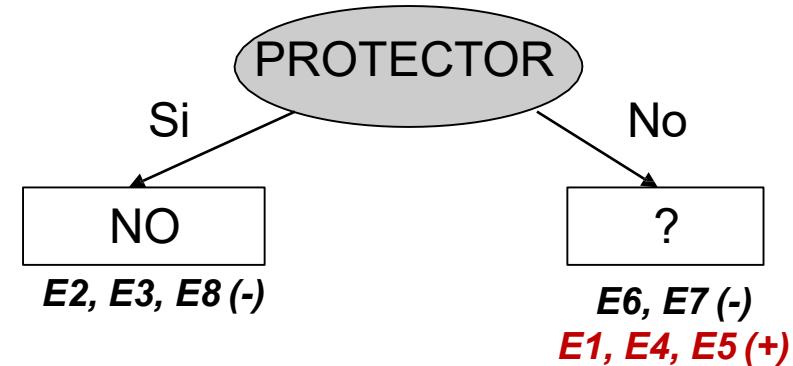
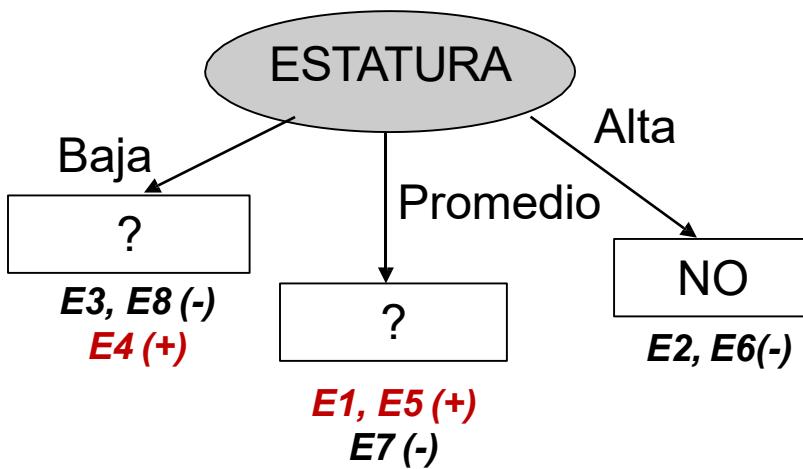
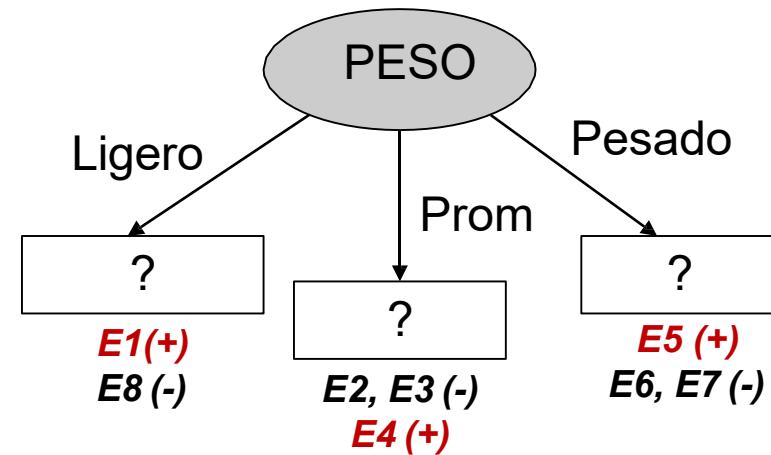
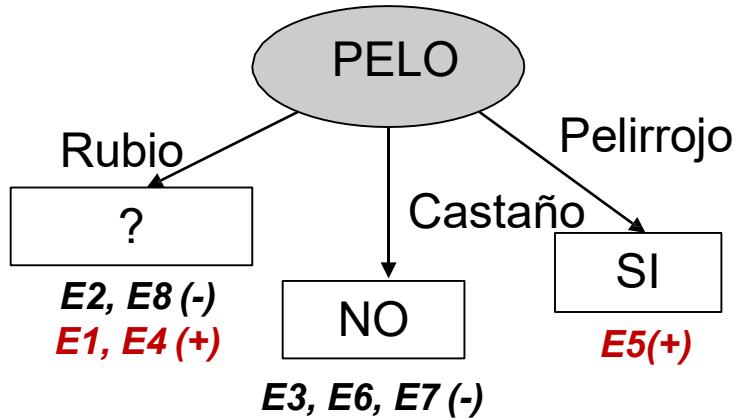
- **Todas las muestras, para un nodo dado, corresponden a la misma clase.**
- **No hay atributos restantes para particionar.**
- **No quedan más muestras** (registros del conjunto de entrenamiento).

EJEMPLO 1: ALSOL.CSV

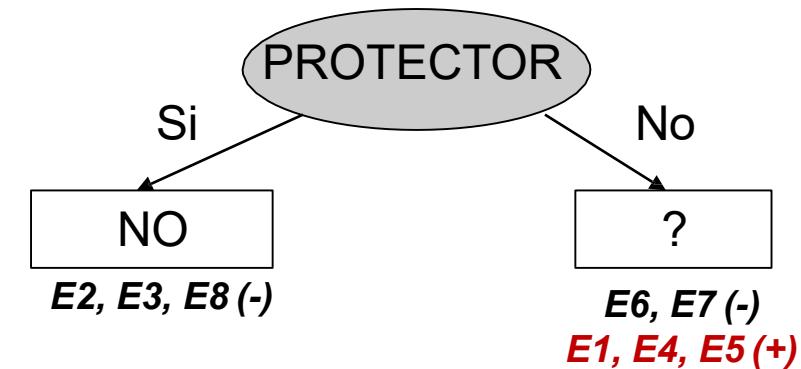
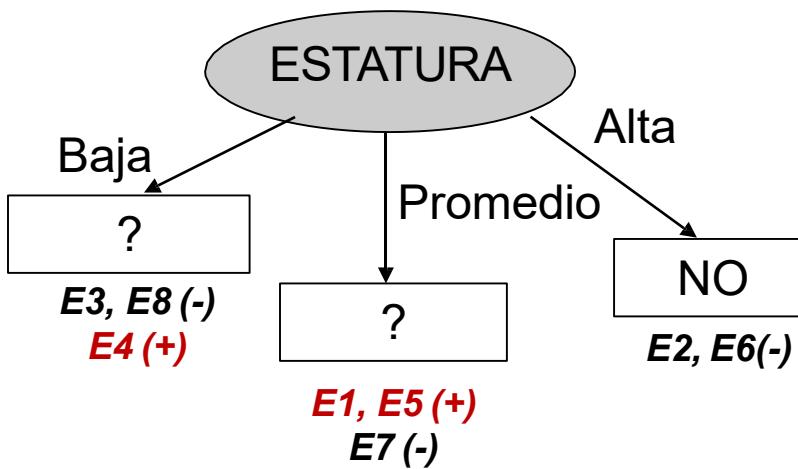
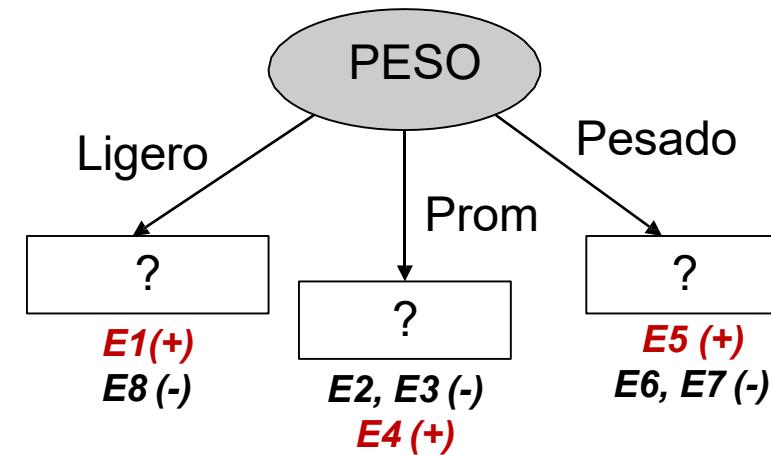
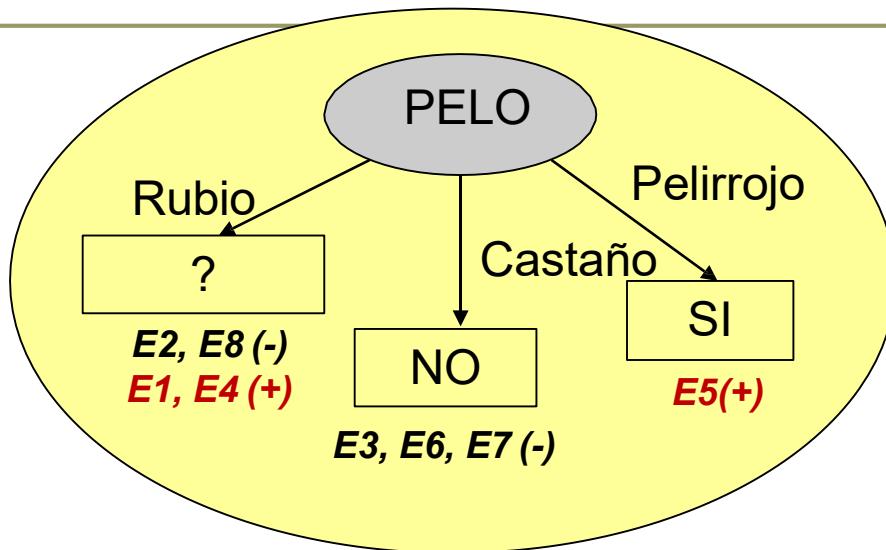
id	Nombre	Pelo	Estatura	Peso	Protector	Quemado
E1	Sara	Rubio	Promedio	Ligero	No	SI
E2	Diana	Rubio	Alta	Promedio	Si	NO
E3	Alexis	Castaño	Baja	Promedio	Si	NO
E4	Ana	Rubio	Baja	Promedio	No	SI
E5	Emilia	Pelirrojo	Promedio	Pesado	No	SI
E6	Pedro	Castaño	Alta	Pesado	No	NO
E7	Juan	Castaño	Promedio	Pesado	No	NO
E8	Catalina	Rubio	Baja	Ligero	Si	NO

■ ¿Cuál atributo elegiría como raíz del árbol?

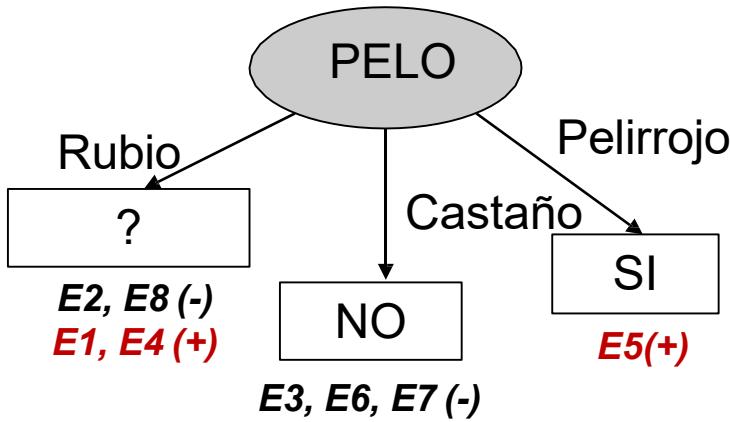
¿QUÉ PASARÍA SI ELIGIERA?



ES LA SELECCIONADA POR TENER LA MAYOR CANTIDAD DE ELEMENTOS EN SUBCONJUNTOS HOMOGÉNEOS

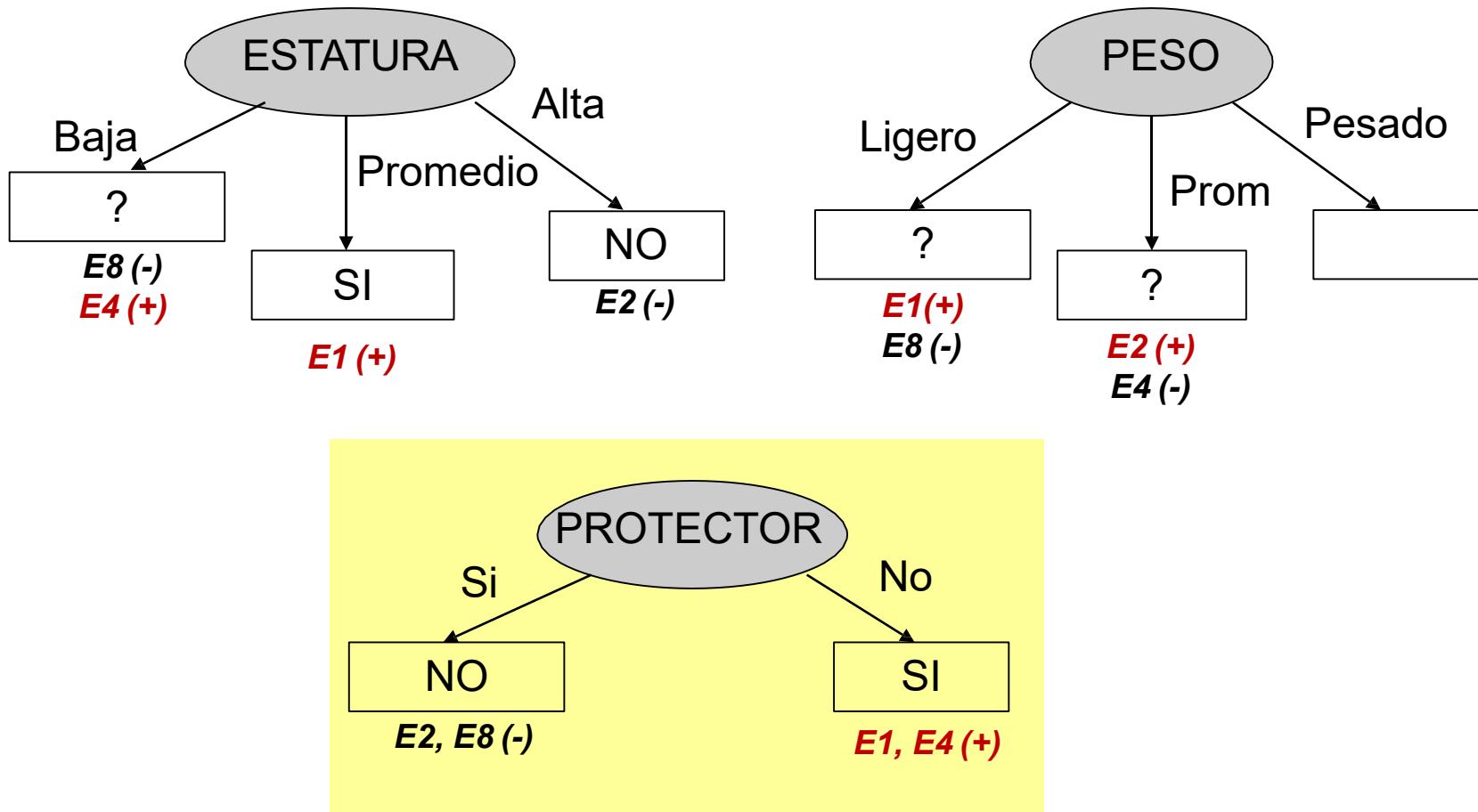


¿CÓMO SIGUE?



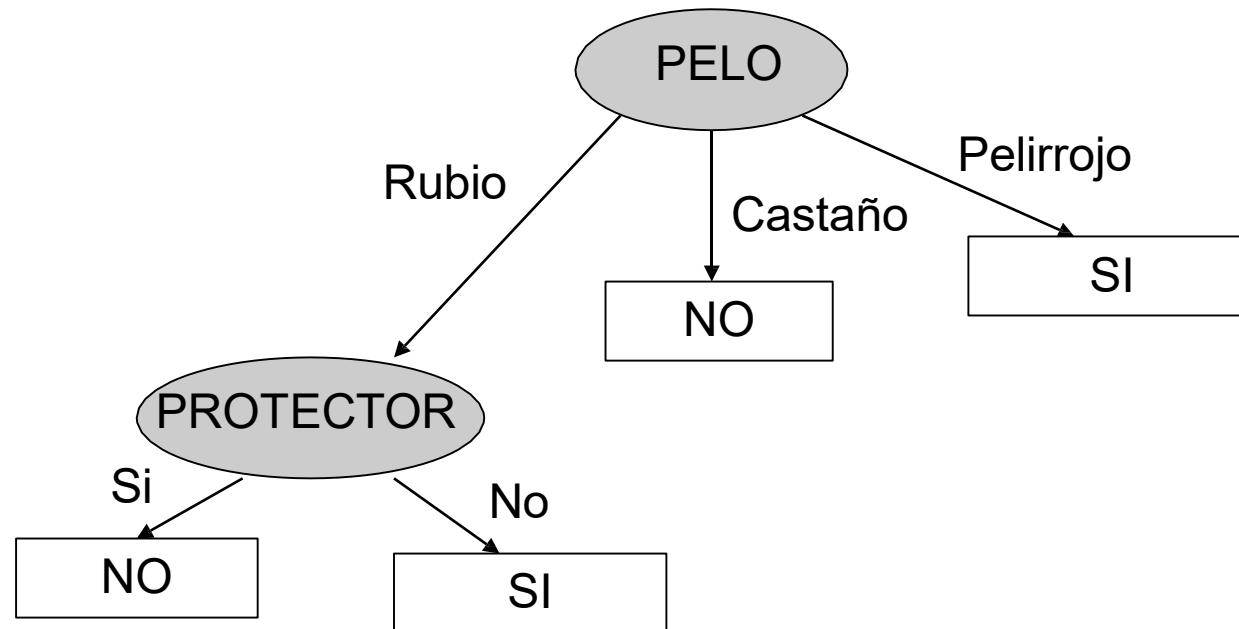
Analizar la respuesta del resto de los atributos para los ejemplos que aún no pertenecen a un subconjunto homogéneo

¿QUÉ PASARÍA SI ELIGIERA?



Es la seleccionada ¿Por qué?

ARBOL DE CLASIFICACIÓN



MÉTODOS DE SELECCIÓN DE ATRIBUTOS

- Utilizados por los algoritmos de construcción para seleccionar en cada paso, según se va generando el árbol, aquel atributo que mejor distribuye los ejemplos de acuerdo a la clasificación objetivo.
- Existen distintos criterios
 - Ganancia de Información (Algoritmo Id3)
 - Tasa de Ganancia (Algoritmo C4.5)
 - Indice Gini (Algoritmos CART)

GANANCIA DE INFORMACIÓN

- Detecta el atributo con menor incertidumbre para identificar la clase.
- La **Ganancia de Información** es la diferencia entre la **incertidumbre inicial** y la del atributo seleccionado.
- Luego, se elegirá aquél que brinde la mayor **Ganancia de Información**.

-
- Comencemos por revisar el concepto de **entropía**

ENTROPÍA

- La entropía caracteriza la heterogeneidad de un conjunto de valores.
- Si en el conjunto E hay n valores distintos, la entropía de E se define como:

$$\text{Entropia}(E) = \sum_{i=1}^n -p_i \log_2(p_i)$$

siendo p_i la proporción de ejemplos de E que coinciden con el i -ésimo valor.

ENTROPÍA

E1, E4, E5 (+)
E2, E3, E6, E7, E8 (-)

- En *AISol.csv*, el conjunto inicial de ejemplos E está formado por 3 casos positivos y 5 negativos. Luego

$$\begin{aligned} \text{Entropia}(E) &= -\left(\frac{3}{8}\right) * \log_2\left(\frac{3}{8}\right) - \left(\frac{5}{8}\right) * \log_2\left(\frac{5}{8}\right) \\ &= 0.9544 \end{aligned}$$



Valdría 0 si todos pertenecieran a la misma clase y 1 si hubiera la misma cantidad de ejemplos positivos que negativos (caso de 2 clases)

ENTROPÍA DE UN ATRIBUTO

- Utilizaremos el concepto de **entropía** para **medir la incertidumbre** de cada **atributo**.
- La **incertidumbre** del atributo es la **cantidad de información que necesita para identificar la clase**.
- Permite **medir cuán heterogénea es la distribución de los ejemplos** luego de usar el **atributo**.
- Valdrá **0 si en cada rama los ejemplos pertenecen a una única clase**.

ENTROPÍA DE UN ATRIBUTO

- La entropía de un atributo A con V_A valores distintos y un conjunto de valores E se calcula de la siguiente forma:

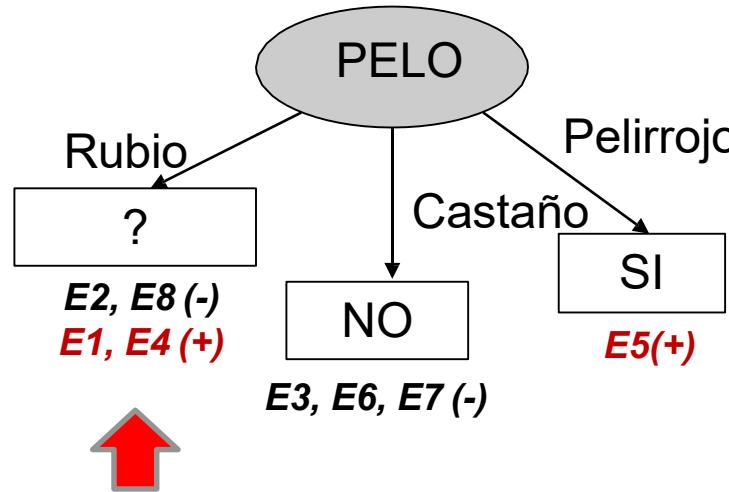
$$Entropia(E, A) = \sum_{v \in V_A} \frac{|E_v|}{E} Entropia(E_v)$$

— siendo E_v el subconjunto de ejemplos para los que el atributo A toma el valor v .

ENTROPÍA DE PELO

E

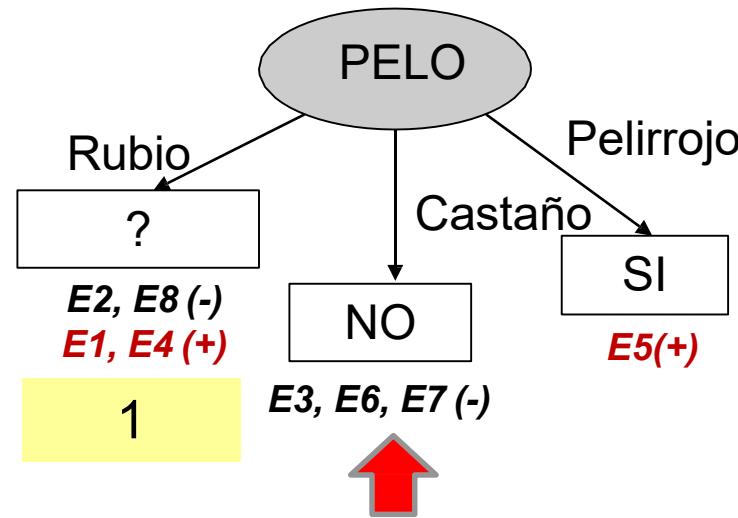
E1, E4, E5 (+)
E2, E3, E6, E7, E8 (-)



$$Entropia(E_{rubio}) = - \left(\frac{2}{4} \right) * \log_2 \left(\frac{2}{4} \right) - \left(\frac{2}{4} \right) * \log_2 \left(\frac{2}{4} \right) = 1$$

ENTROPÍA DE PELO

E
E1, E4, E5 (+)
E2, E3, E6, E7, E8 (-)

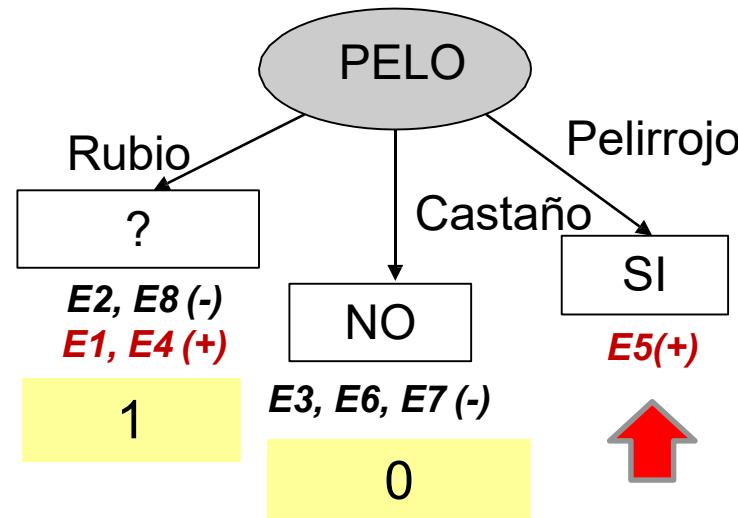


$$Entropia(E_{casta\tilde{n}o}) = - \left(\frac{3}{3} \right) * \log_2 \left(\frac{3}{3} \right) = 0$$

ENTROPÍA DE PELO

E

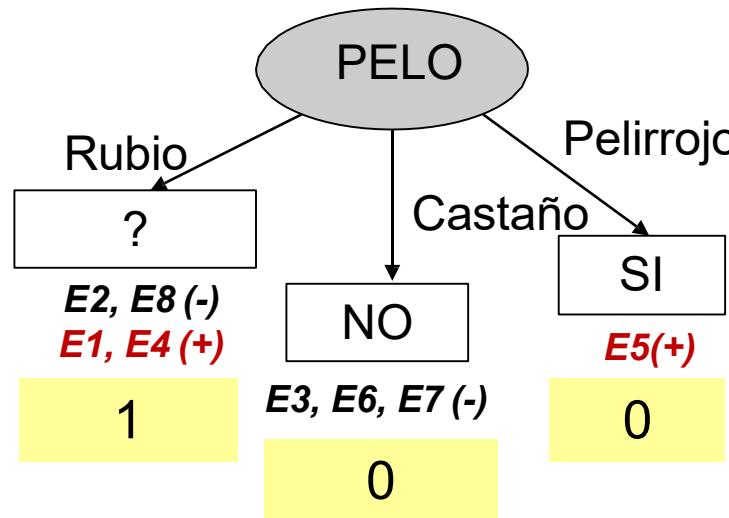
E1, E4, E5 (+)
E2, E3, E6, E7, E8 (-)



$$Entropia(E_{pelirrojo}) = - \left(\frac{1}{1} \right) * \log_2 \left(\frac{1}{1} \right) = 0$$

ENTROPÍA DE PELO

E
E1, E4, E5 (+)
E2, E3, E6, E7, E8 (-)



$$Entropia(E, Pelo) = \left(\frac{4}{8}\right) * 1 + \left(\frac{3}{8}\right) * 0 + \left(\frac{1}{8}\right) * 0 = 0.5$$

GANANCIA DE INFORMACIÓN

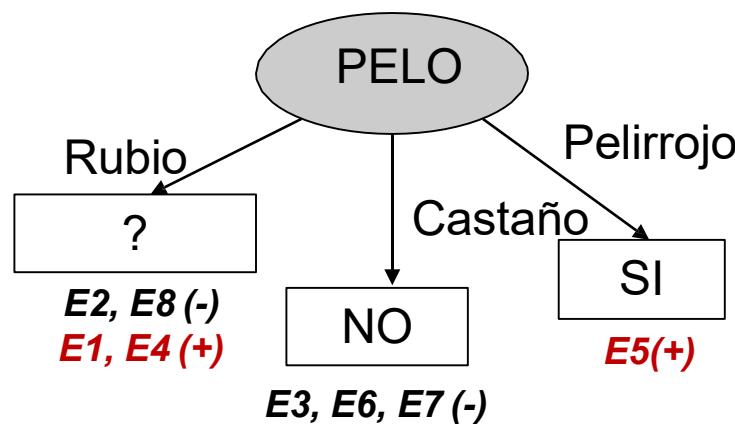
- Dado un atributo A y un conjunto de ejemplos E, la Ganancia de Información de dicho atributo se calcula así:

$$Ganancia(E,A) = Entropia(E) - Entropia(E,A)$$

GANANCIA DE INFORMACIÓN

**E1, E4, E5 (+)
E2, E3, E6, E7, E8 (-)**

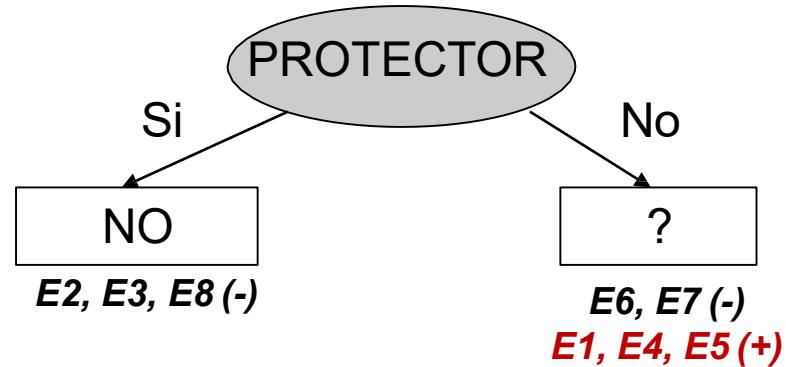
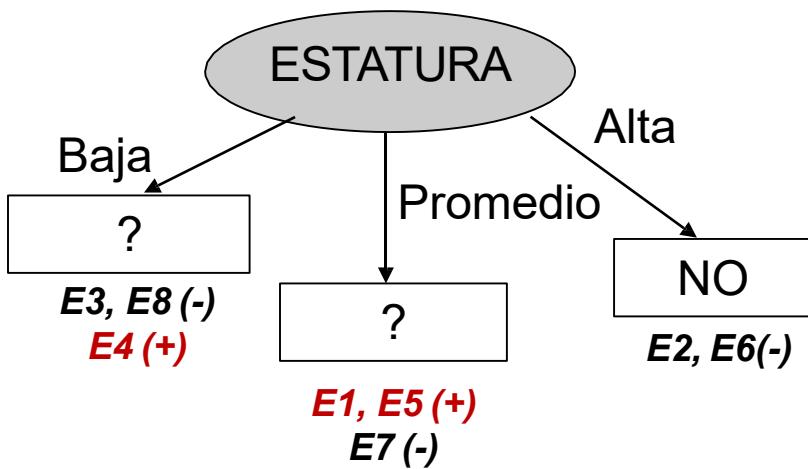
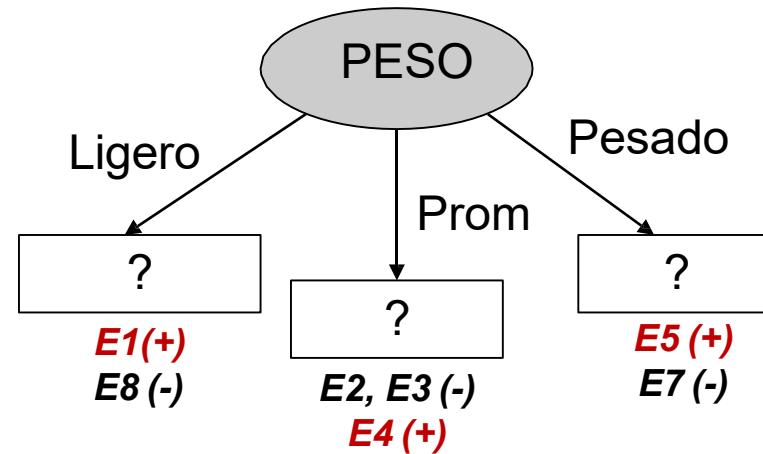
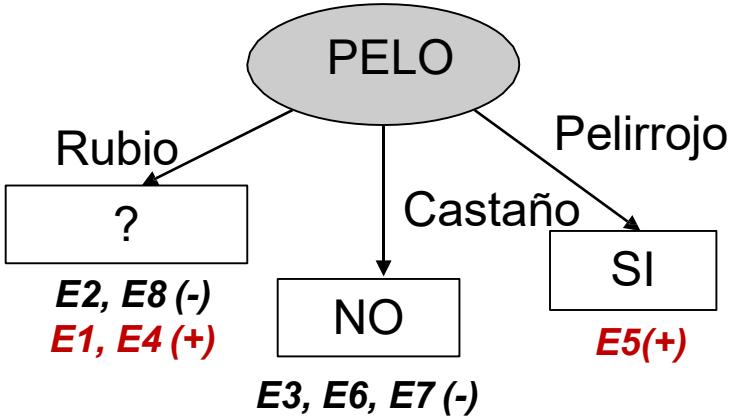
← Entropía(E)=0.9544



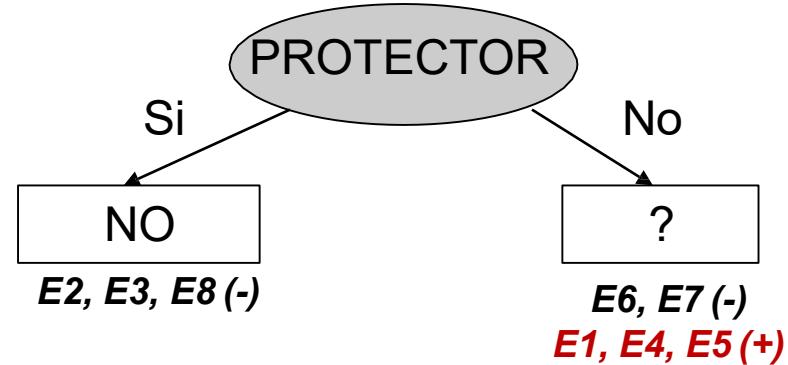
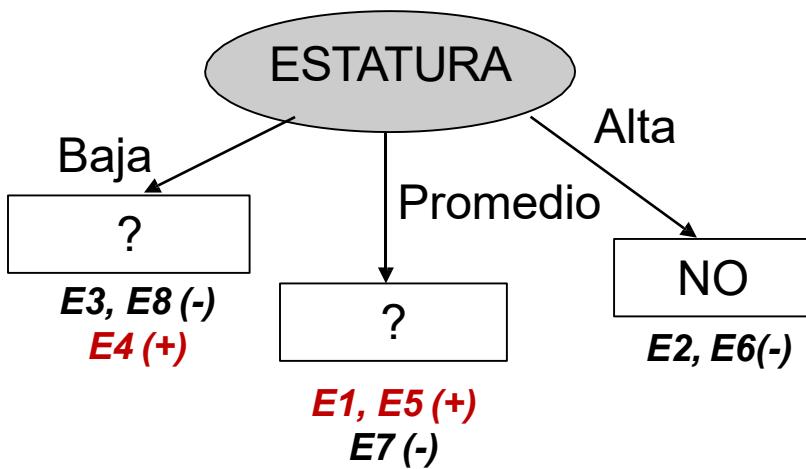
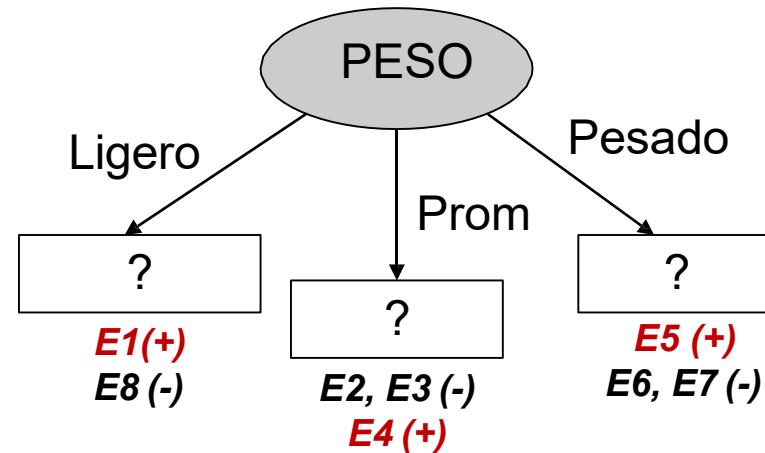
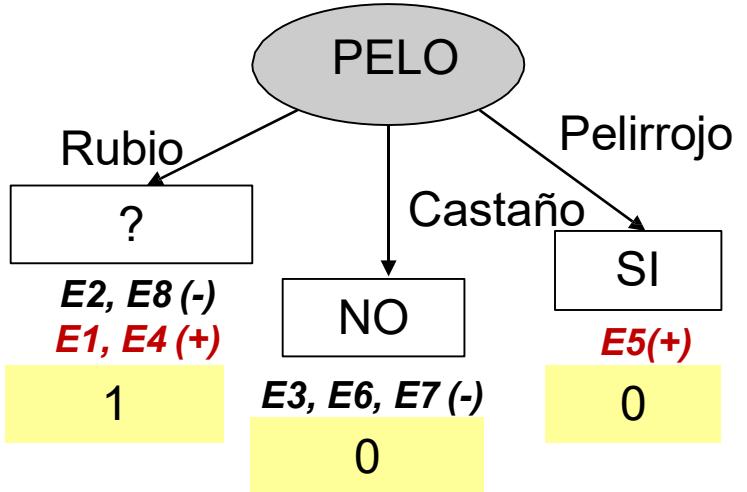
← Entropía($E, Pelo$)=0.5

$$Ganancia(E, Pelo) = 0.9544 - 0.5 = 0.4544$$

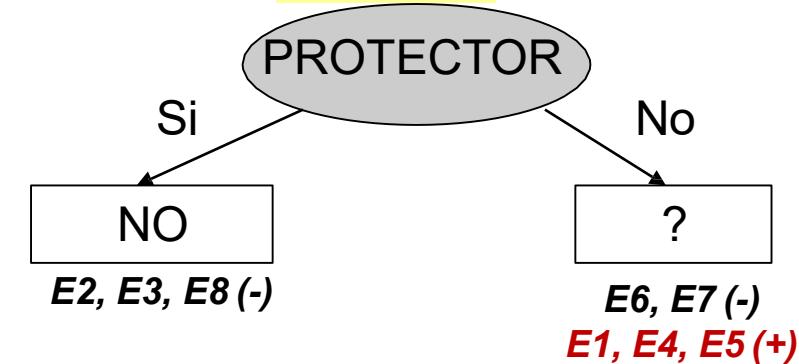
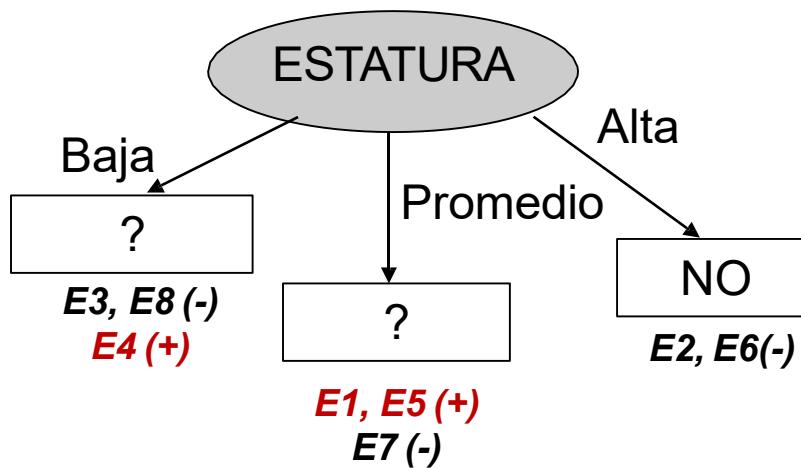
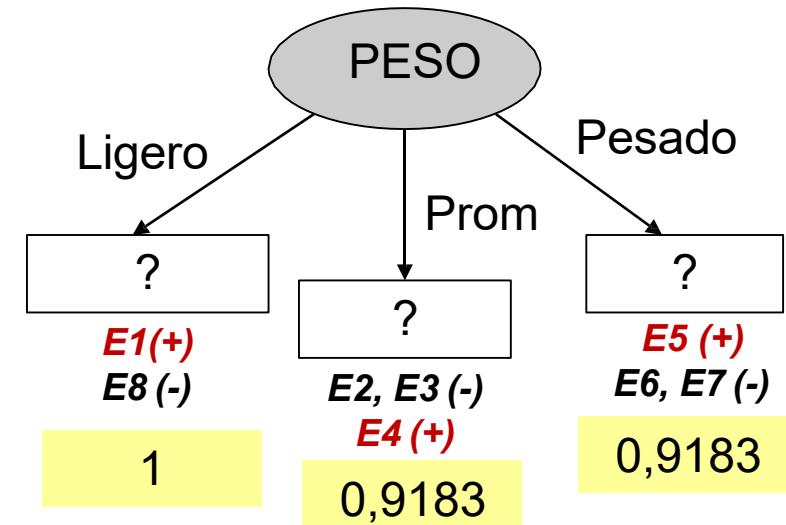
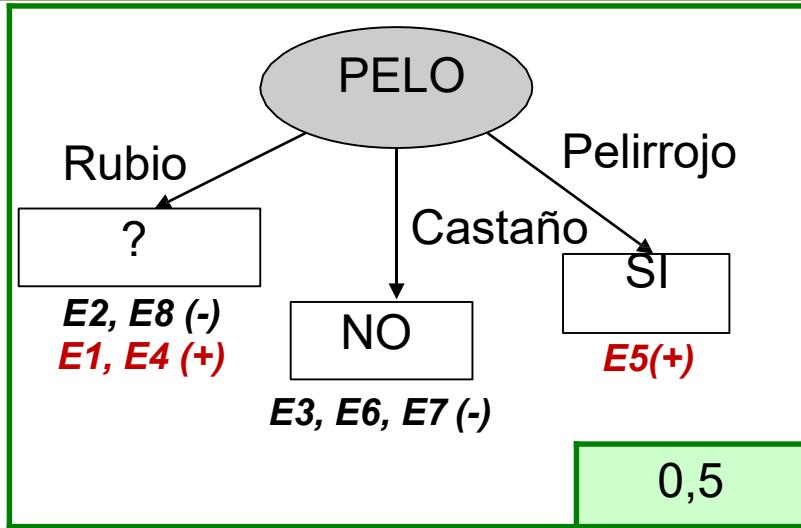
¿QUÉ PASARÍA SI ELIGIERA?



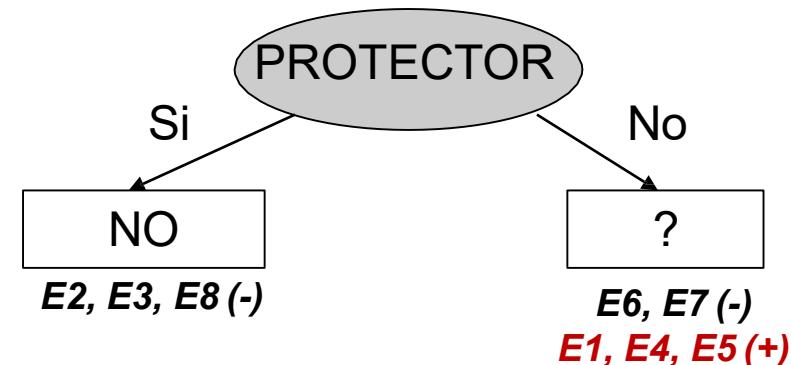
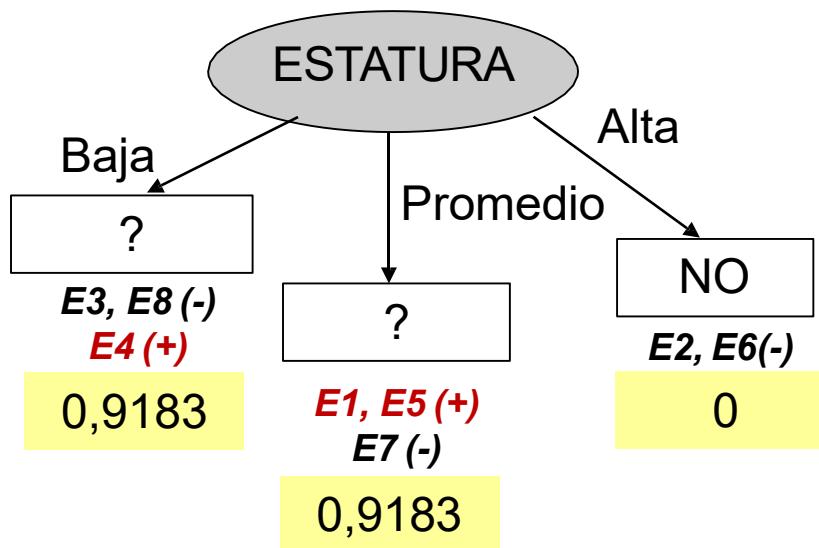
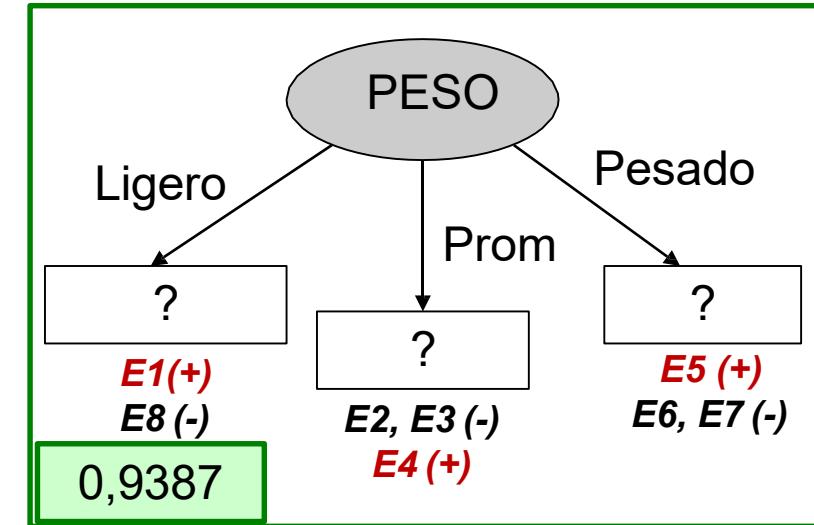
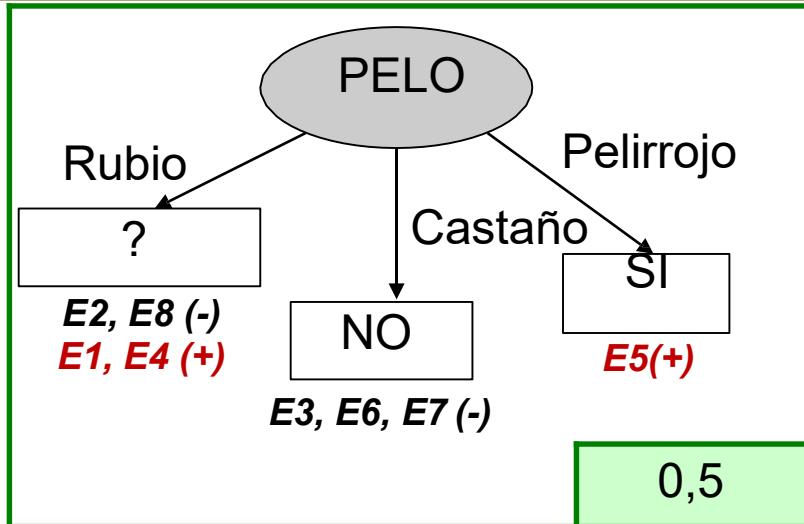
¿QUÉ PASARÍA SI ELIGIERA?



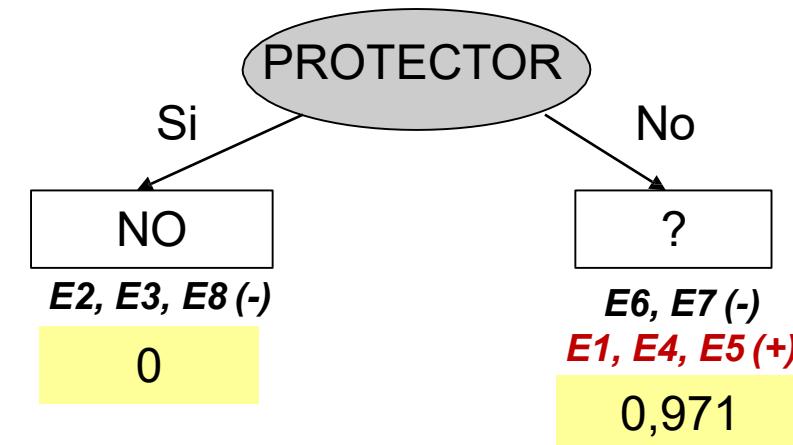
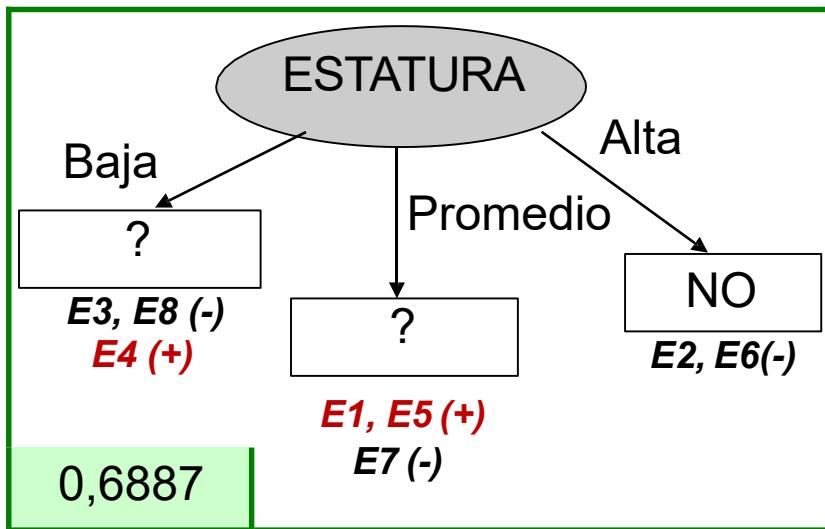
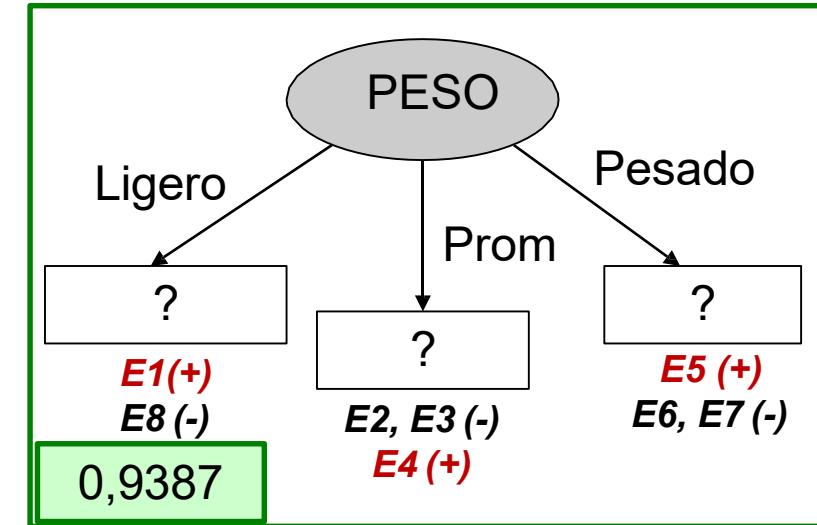
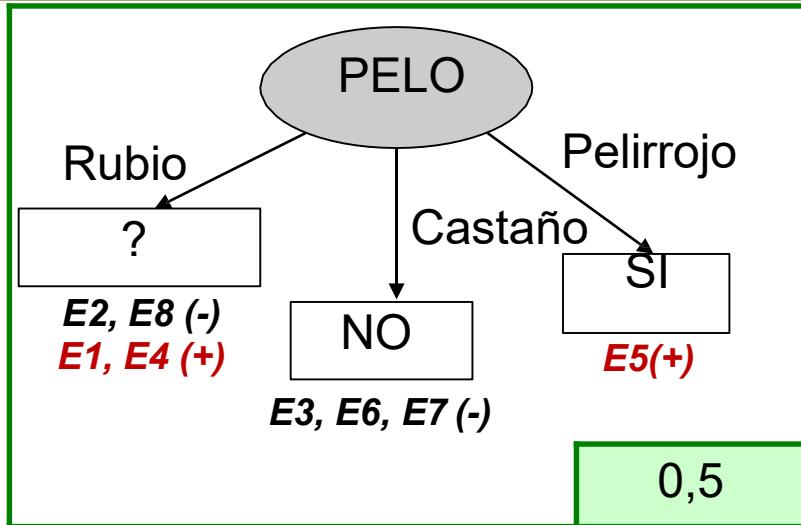
¿QUÉ PASARÍA SI ELIGIERA?



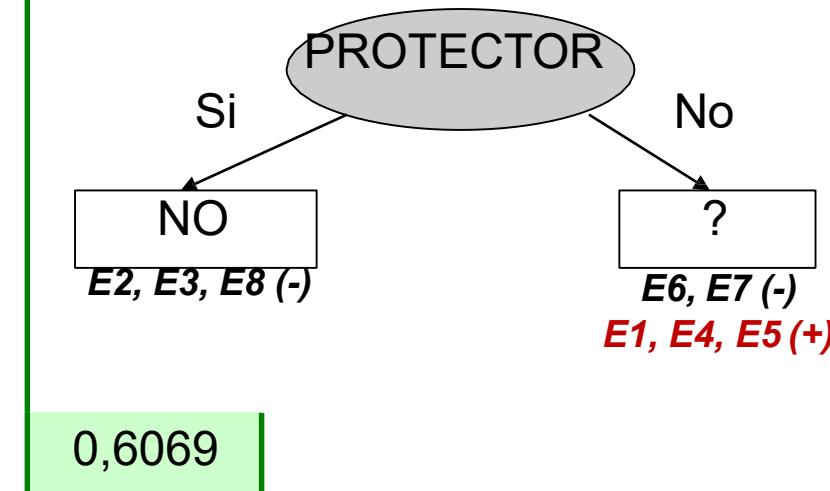
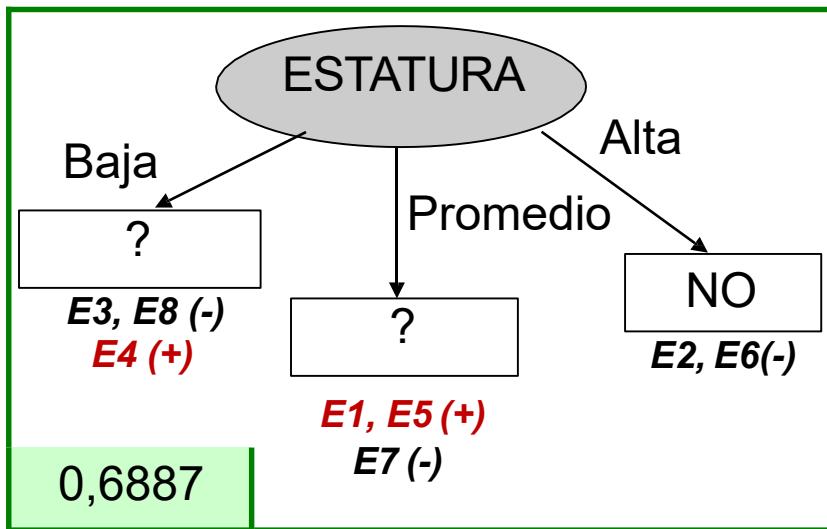
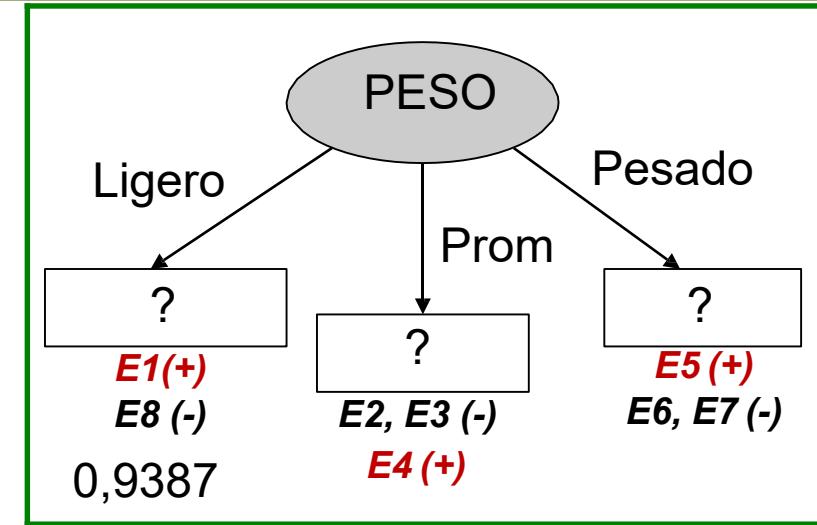
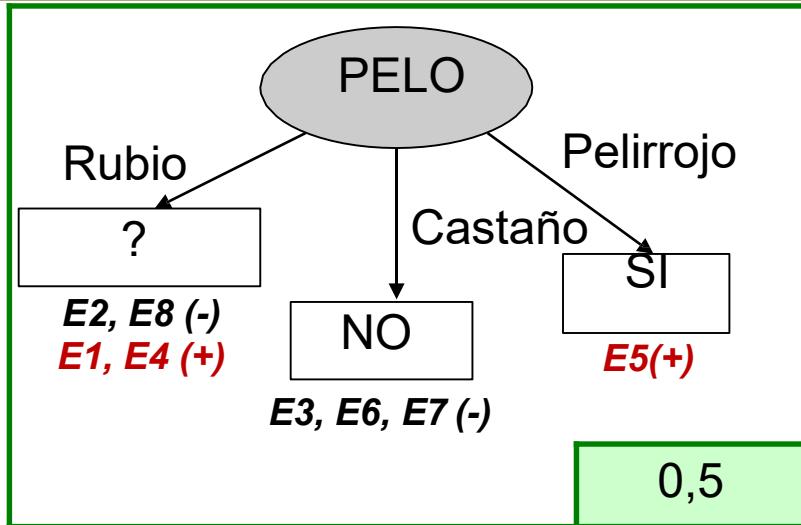
¿QUÉ PASARÍA SI ELIGIERA?



¿QUÉ PASARÍA SI ELIGIERA?



¿QUÉ PASARÍA SI ELIGIERA?



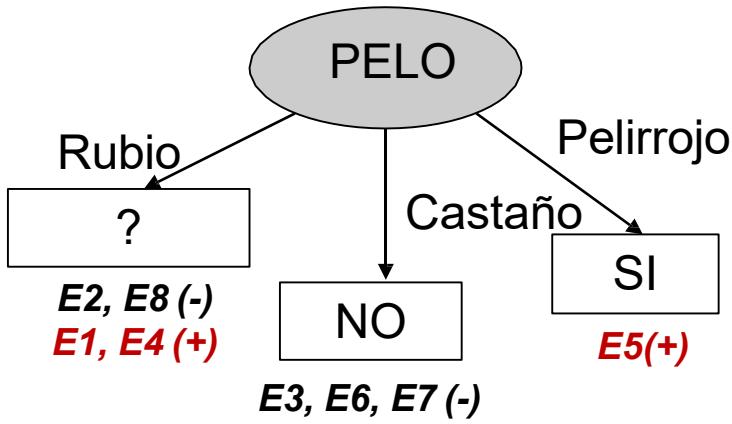
SELECCIÓN POR GANANCIA DE INFORMACIÓN

- $Entropía(E) = 0.9544$
- $Ganancia(E,A) = Entropía(E) - Entropía(E,A)$

Atributo	Entropía(E,A)	Ganancia(E,A)
Pelo	0,5	0,4544
Protector	0,6069	0,3475
Estatura	0,6887	0,2657
Peso	0,9387	0,0157

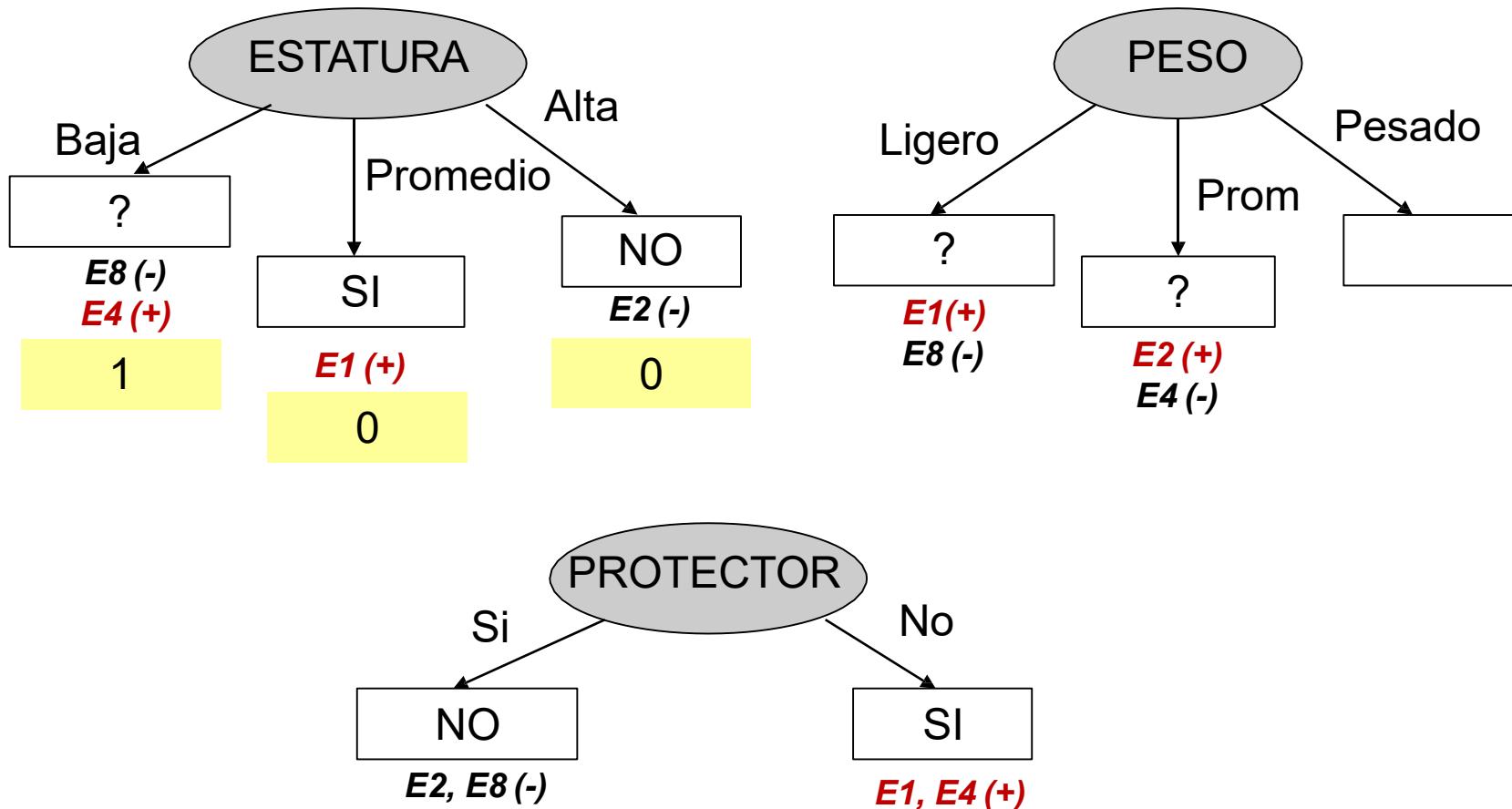
Es el seleccionado por ser el de mayor Ganancia

¿CÓMO SIGUE?

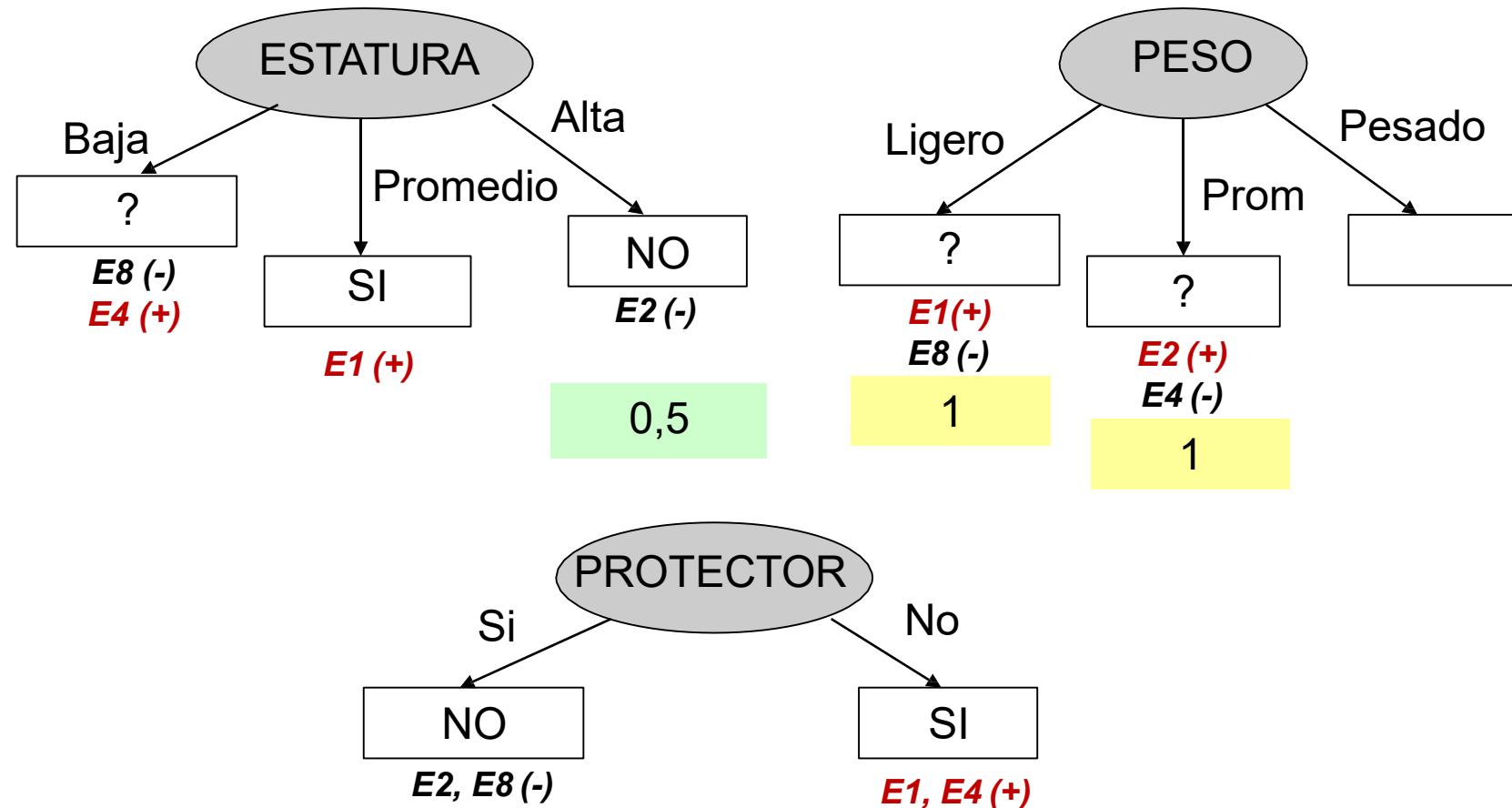


Analizar la respuesta del resto de los atributos para los ejemplos de un subconjunto heterogéneo

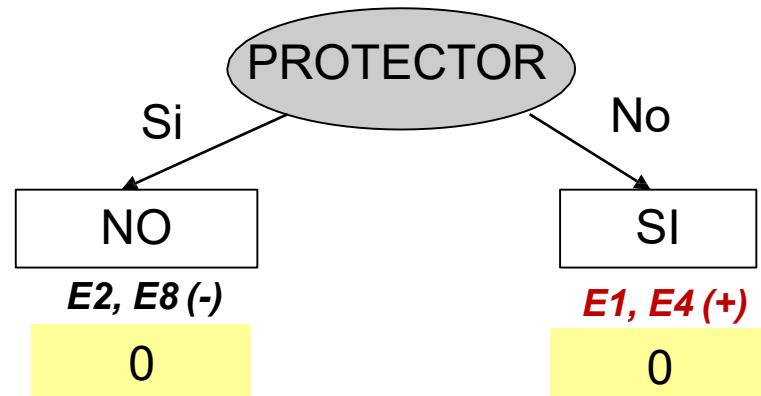
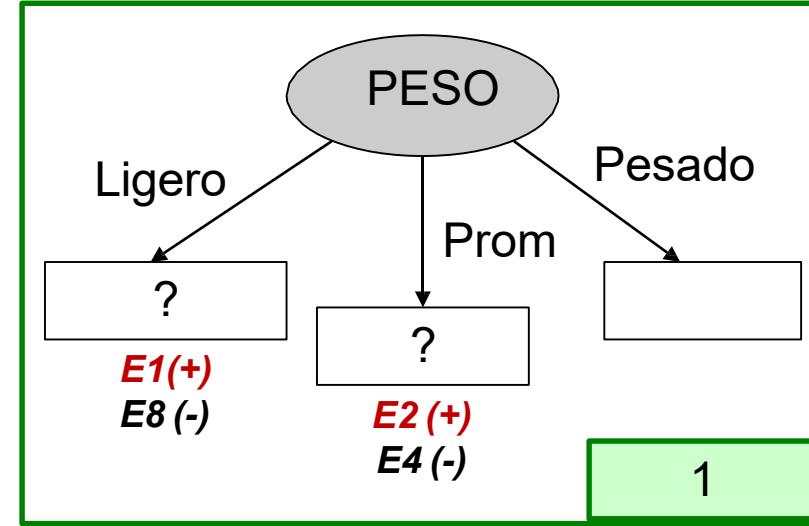
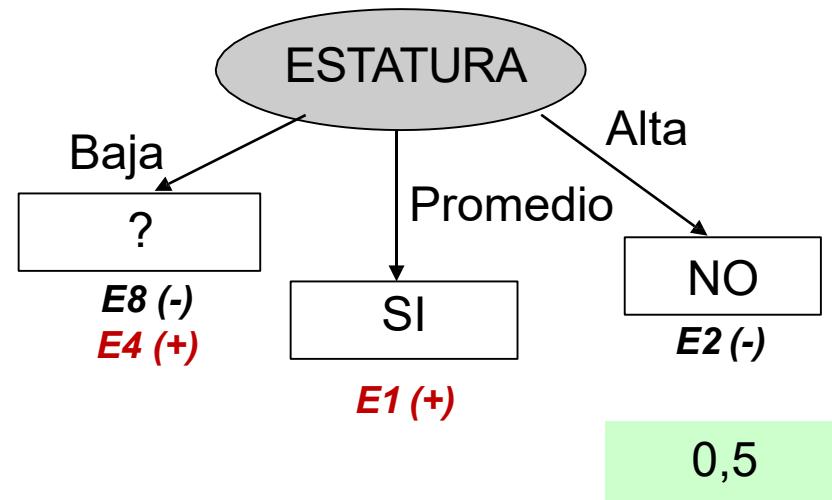
¿QUÉ PASARÍA SI ELIGIERA?



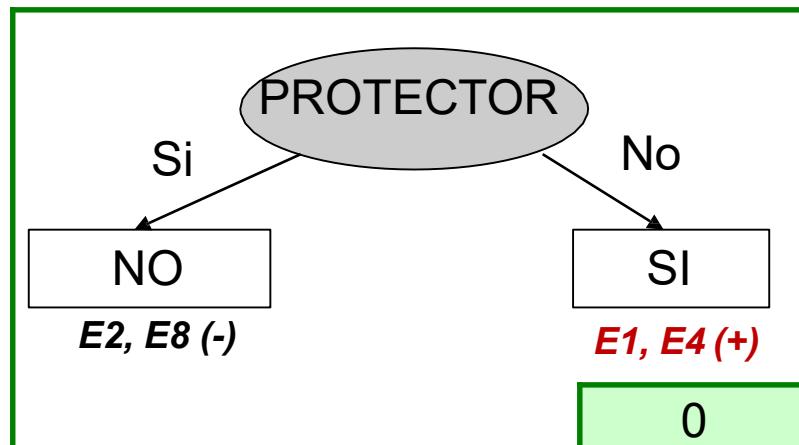
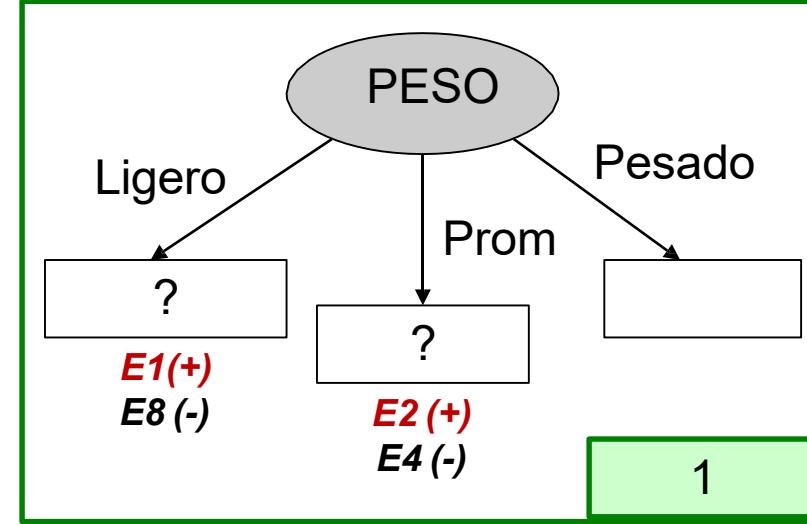
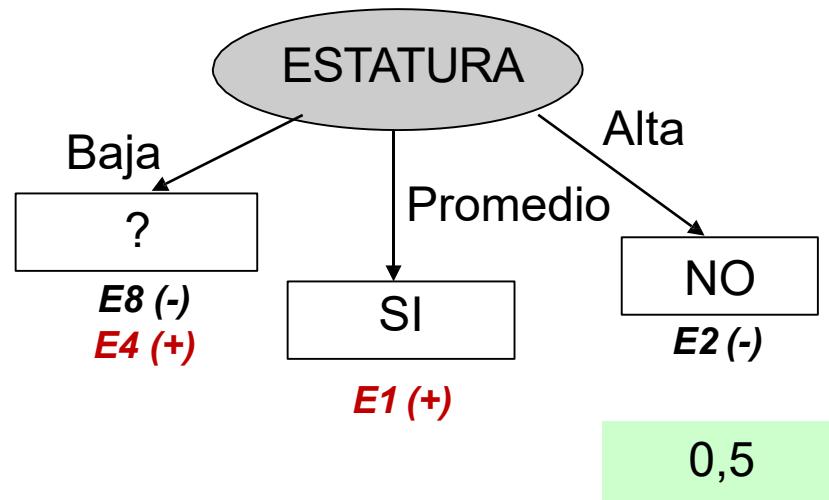
¿QUÉ PASARÍA SI ELIGIERA?



¿QUÉ PASARÍA SI ELIGIERA?



¿QUÉ PASARÍA SI ELIGIERA?



SELECCIÓN POR GANANCIA DE INFORMACIÓN

■ $Entropia(E) = 1$

E

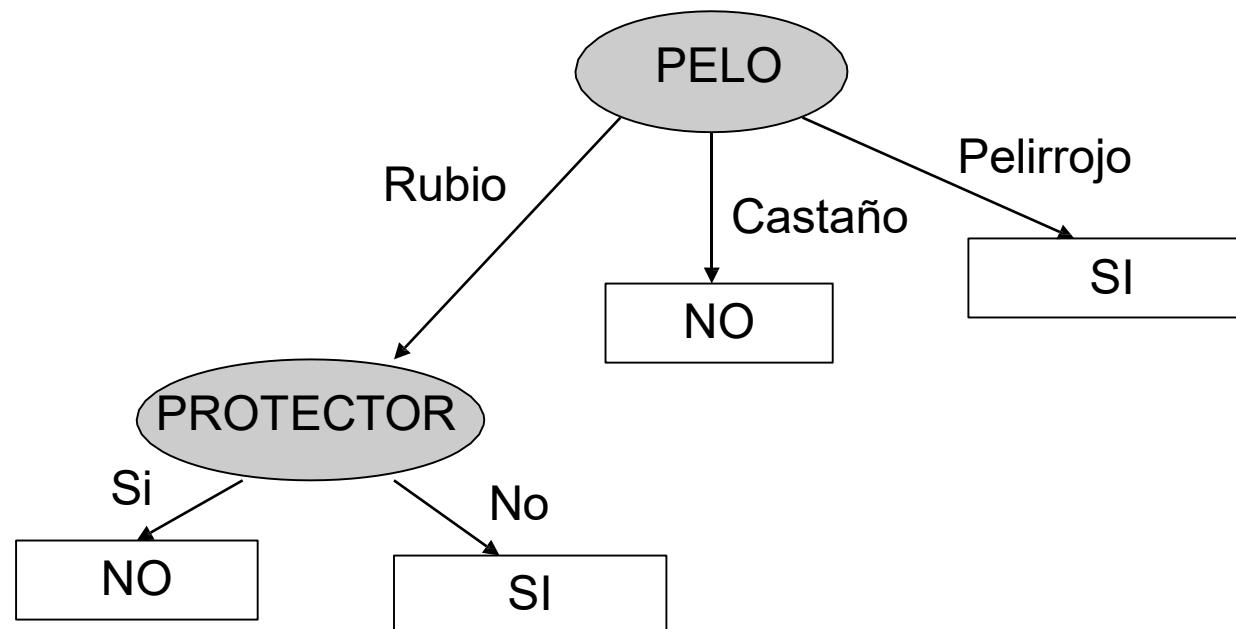
$E2, E8 (-)$
 $E1, E4 (+)$

Atributo	Entropía(E, A)	Ganancia(E, A)
Protector	0	1
Estatura	0,5	0,5
Peso	1	0



Es el
seleccionado
por ser el de
mayor
Ganancia

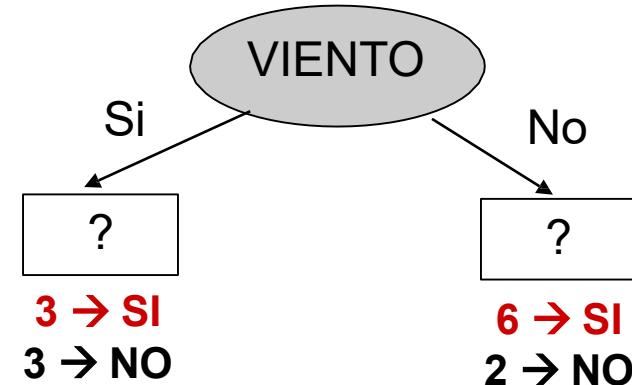
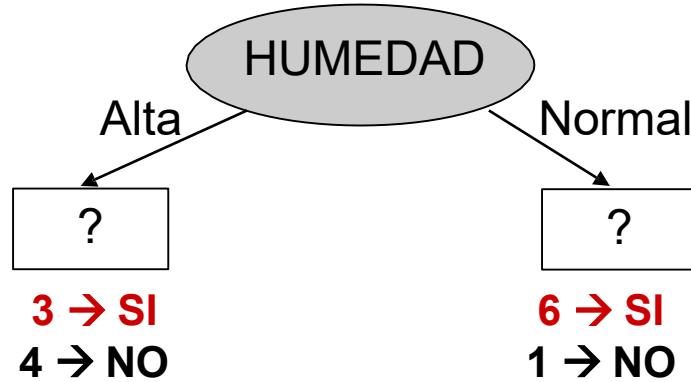
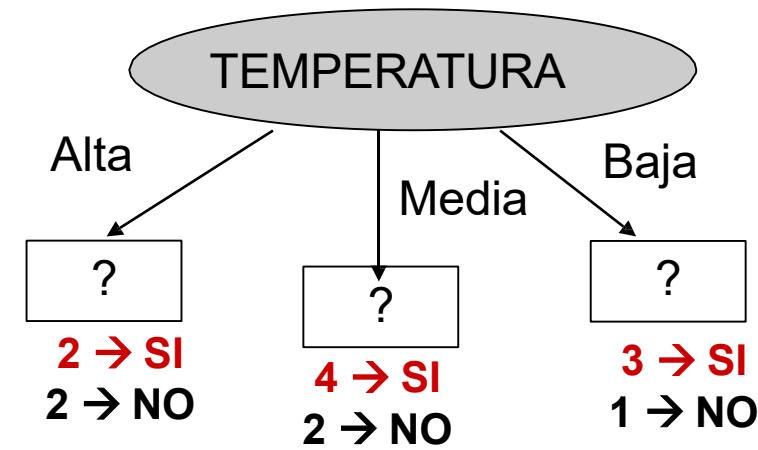
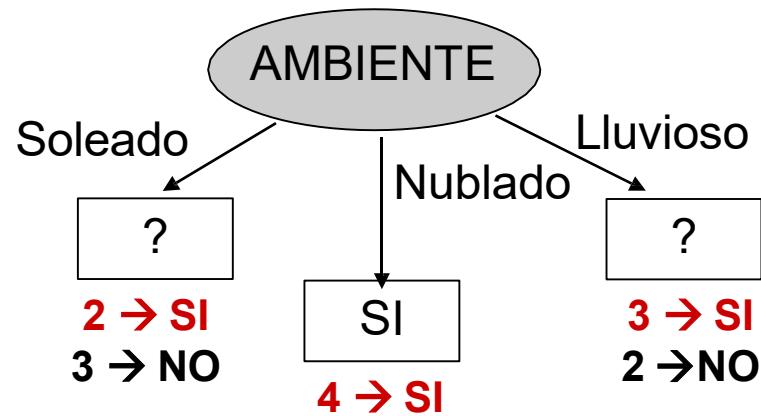
ARBOL DE CLASIFICACIÓN



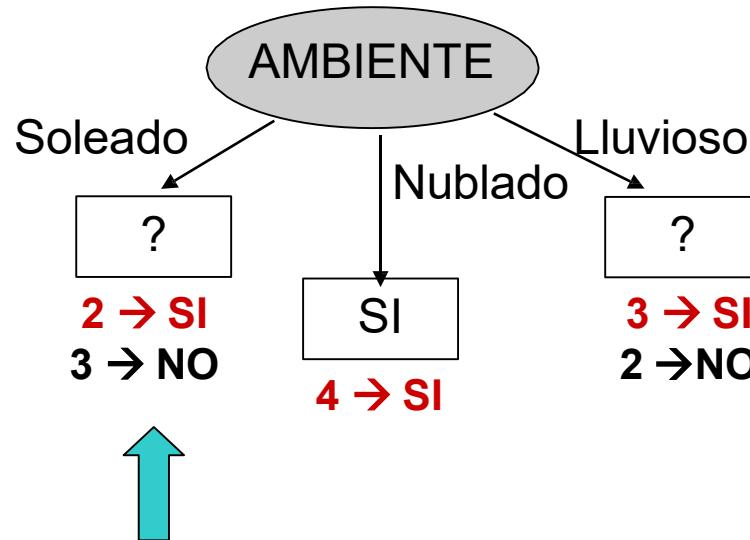
Ejemplo 2: Construir el árbol a partir de los siguientes datos

Nº	Ambiente	Temperatura	Humedad	Viento	Juega?
1	soleado	alta	alta	no	No
2	soleado	alta	alta	si	No
3	nublado	alta	alta	no	Si
4	lluvioso	media	alta	no	Si
5	lluvioso	baja	normal	no	Si
6	lluvioso	baja	normal	si	No
7	nublado	baja	normal	si	Si
8	Soleado	media	alta	no	No
9	Soleado	baja	normal	no	Si
10	lluvioso	media	normal	no	Si
11	Soleado	media	normal	si	Si
12	Nublado	media	alta	si	Si
13	Nublado	alta	normal	no	Si
14	lluvioso	media	alta	si	No

ANALIZANDO EL ATRIBUTO PARA LA RAÍZ

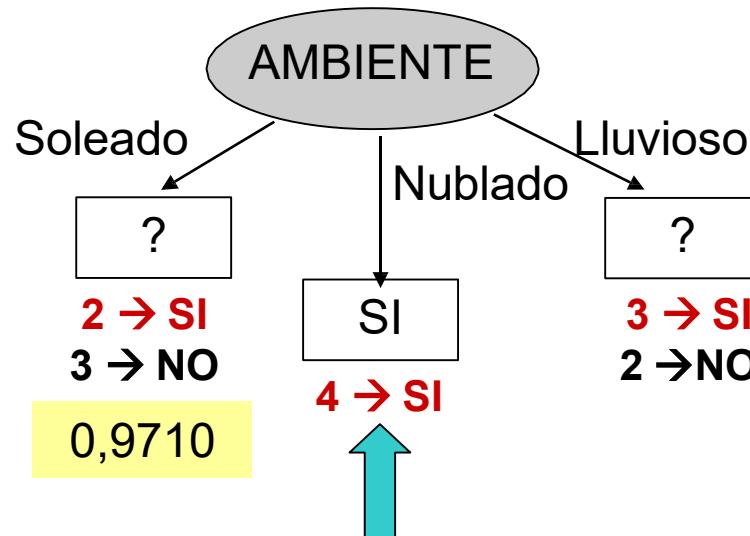


ANALIZANDO CADA RAMA DE AMBIENTE



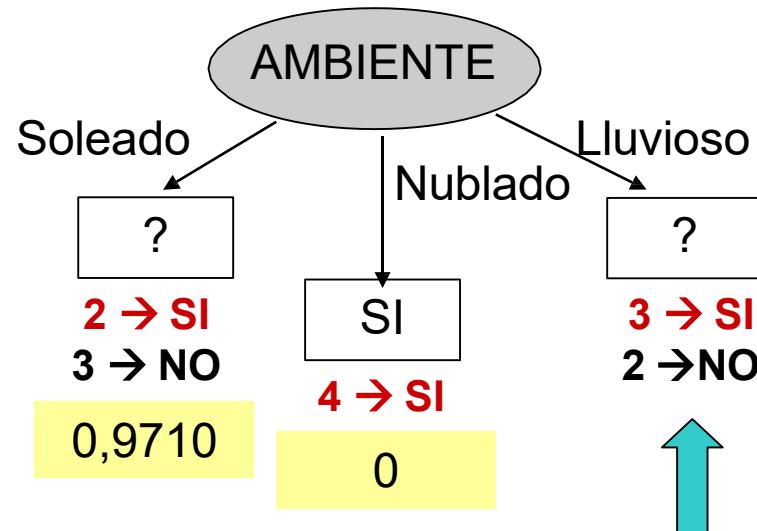
$$Entropia_{Soleado} = -\frac{2}{5} \log_2 \left(\frac{2}{5} \right) - \frac{3}{5} \log_2 \left(\frac{3}{5} \right) = 0,9710$$

ANALIZANDO CADA RAMA DE AMBIENTE



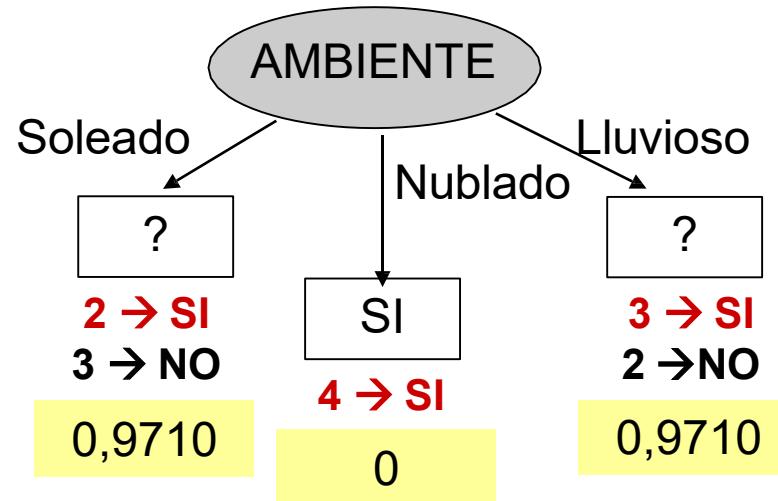
$$Entropia_{Nublado} = -\frac{4}{4} \log_2 \left(\frac{4}{4} \right) = 0$$

ANALIZANDO CADA RAMA DE AMBIENTE



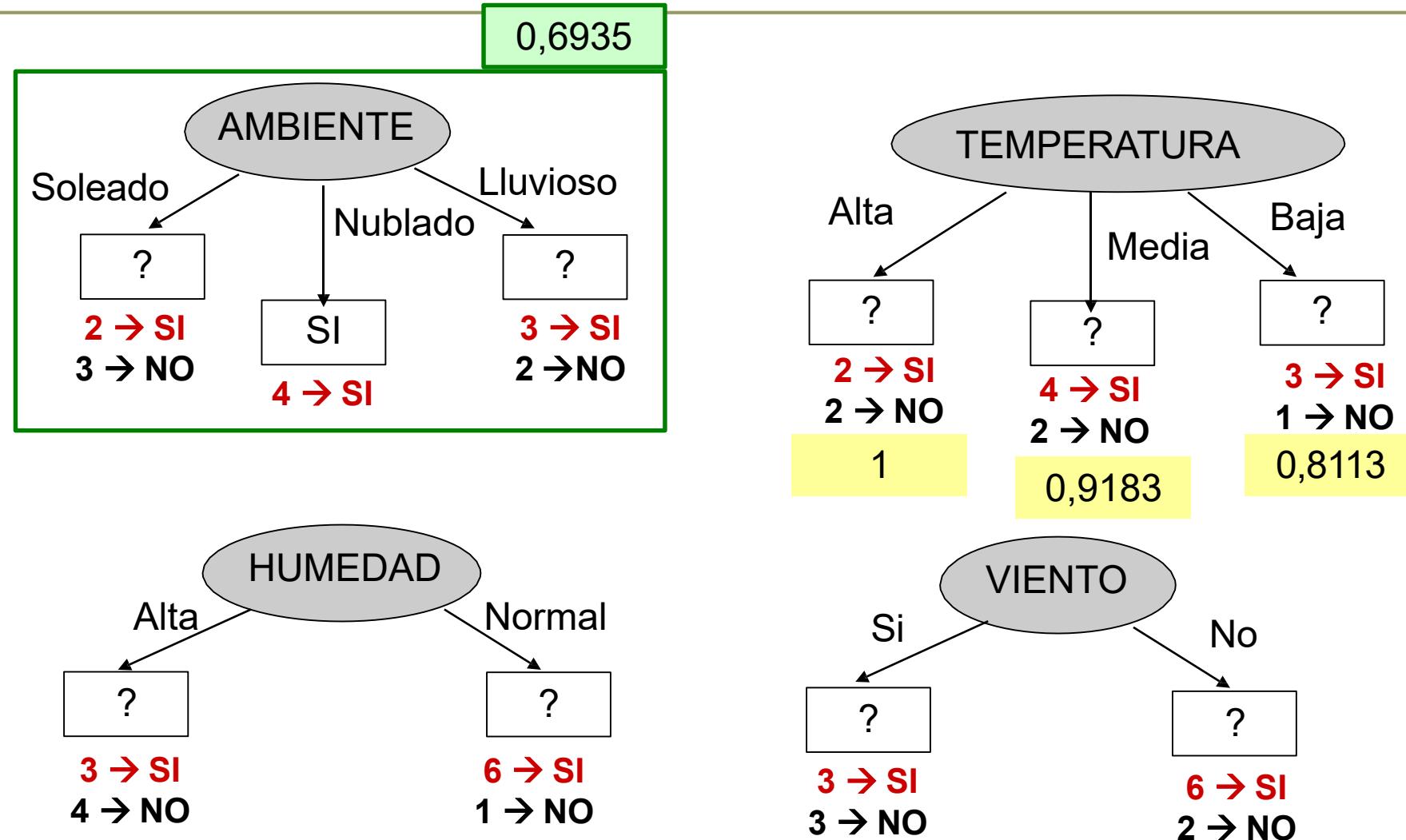
$$Entropia_{Lluvioso} = -\frac{3}{5} \log_2\left(\frac{3}{5}\right) - \frac{2}{5} \log_2\left(\frac{2}{5}\right) = 0,9710$$

ANALIZANDO CADA RAMA DE AMBIENTE

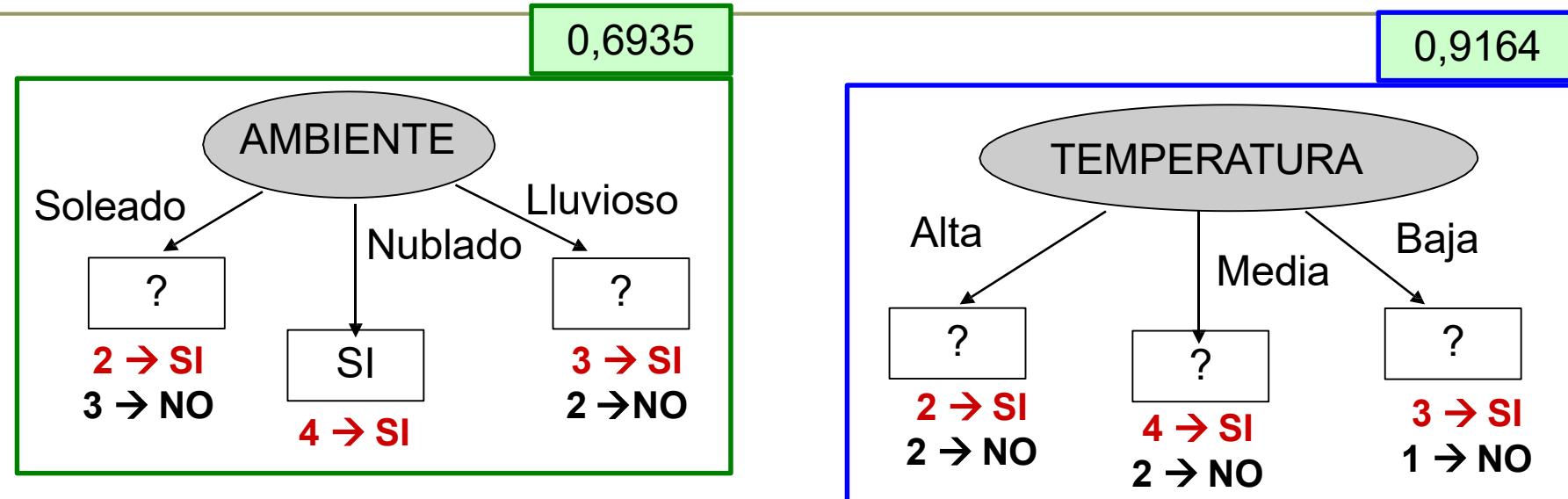


$$Entropia_{AMBIENTE} = \frac{5}{14} * 0,971 + \frac{4}{14} * 0 + \frac{5}{14} * 0,971 = 0,6935$$

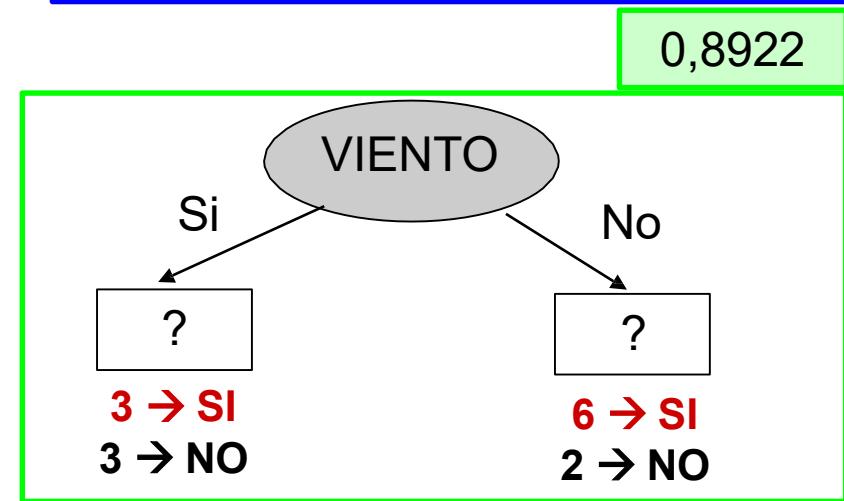
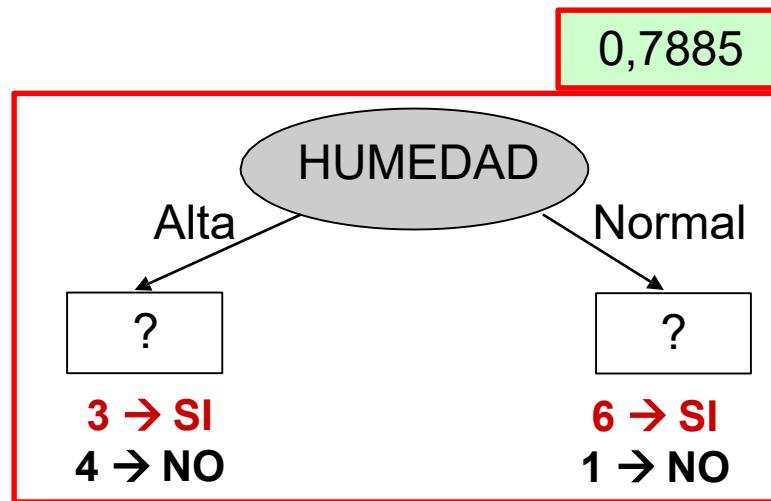
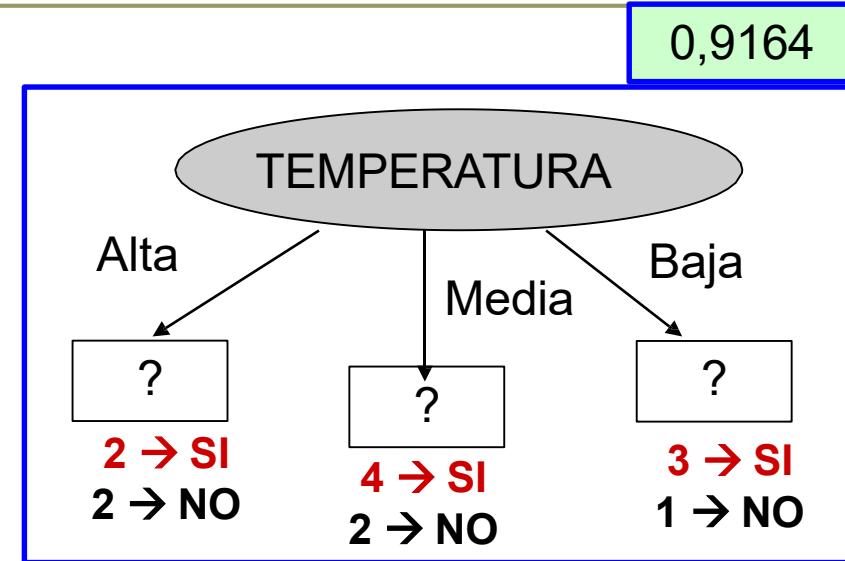
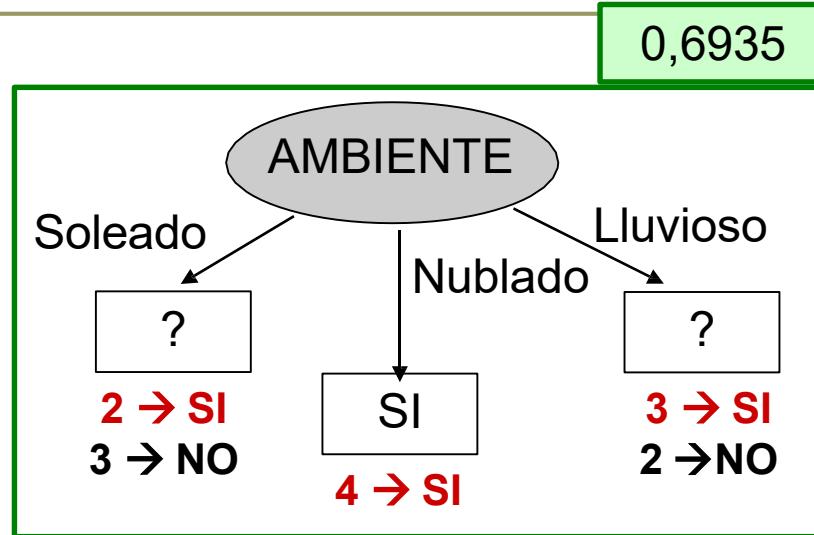
ANALIZANDO EL ATRIBUTO PARA LA RAÍZ



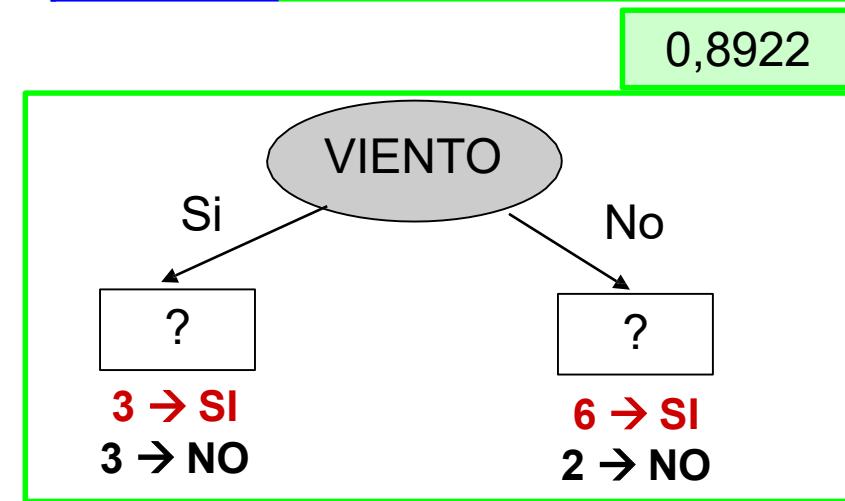
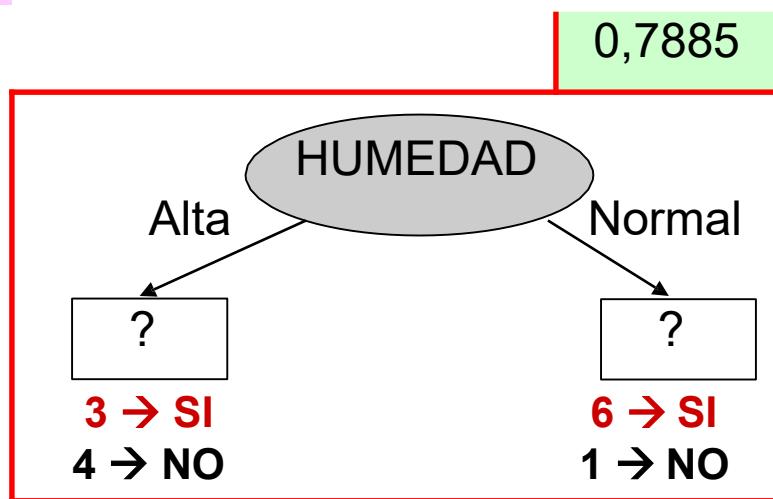
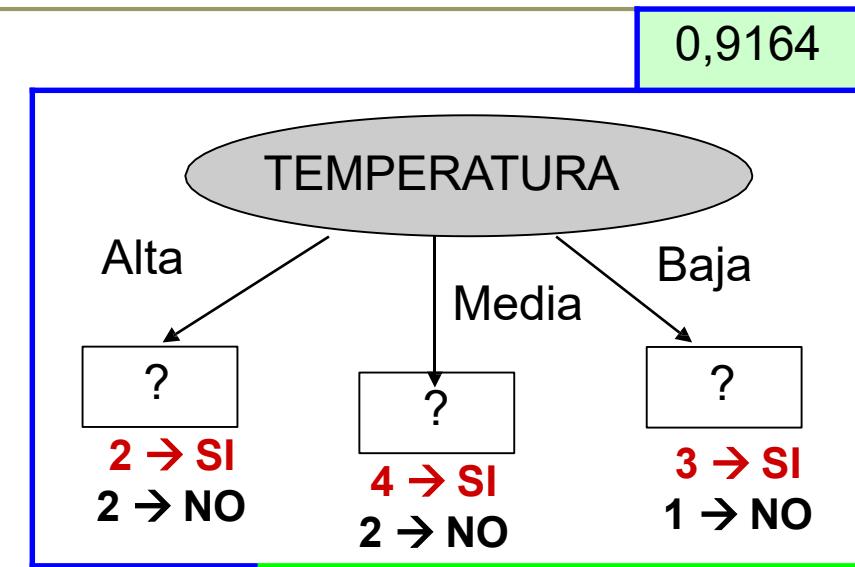
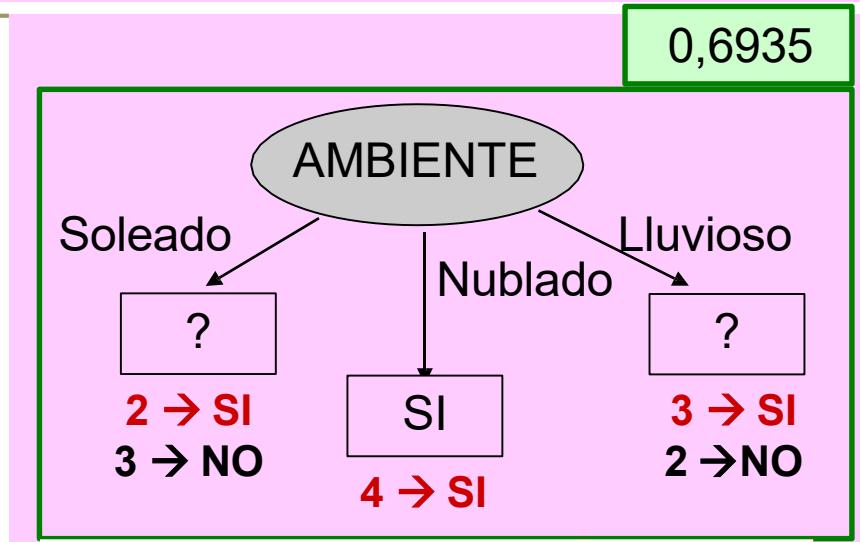
ANALIZANDO EL ATRIBUTO PARA LA RAÍZ



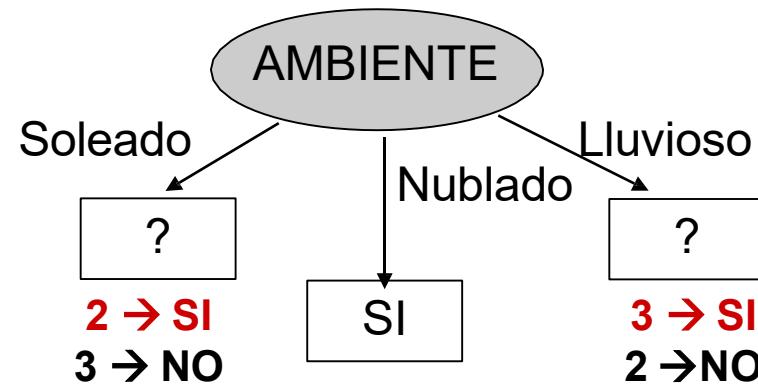
ANALIZANDO EL ATRIBUTO PARA LA RAÍZ



ES EL SELECCIONADO POR TENER EL MENOR VALOR DE **ENTROPIA**, ES DECIR, LA MAYOR CANTIDAD DE ELEMENTOS EN SUBCONJUNTOS HOMOGÉNEOS

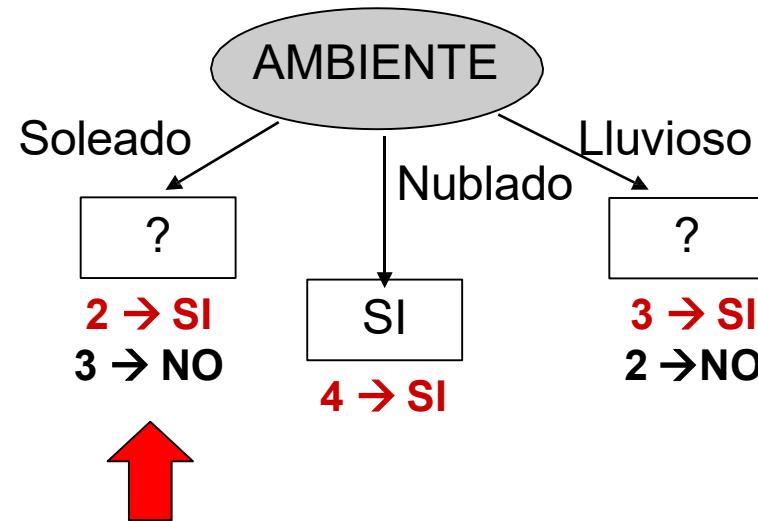


YA TENEMOS EL NODO RAÍZ



- Si está nublado, SI juega.
- Ahora falta analizar las dos ramas que no son puras.

BUSCANDO LOS NODOS DEL 1ER. NIVEL DEL ÁRBOL



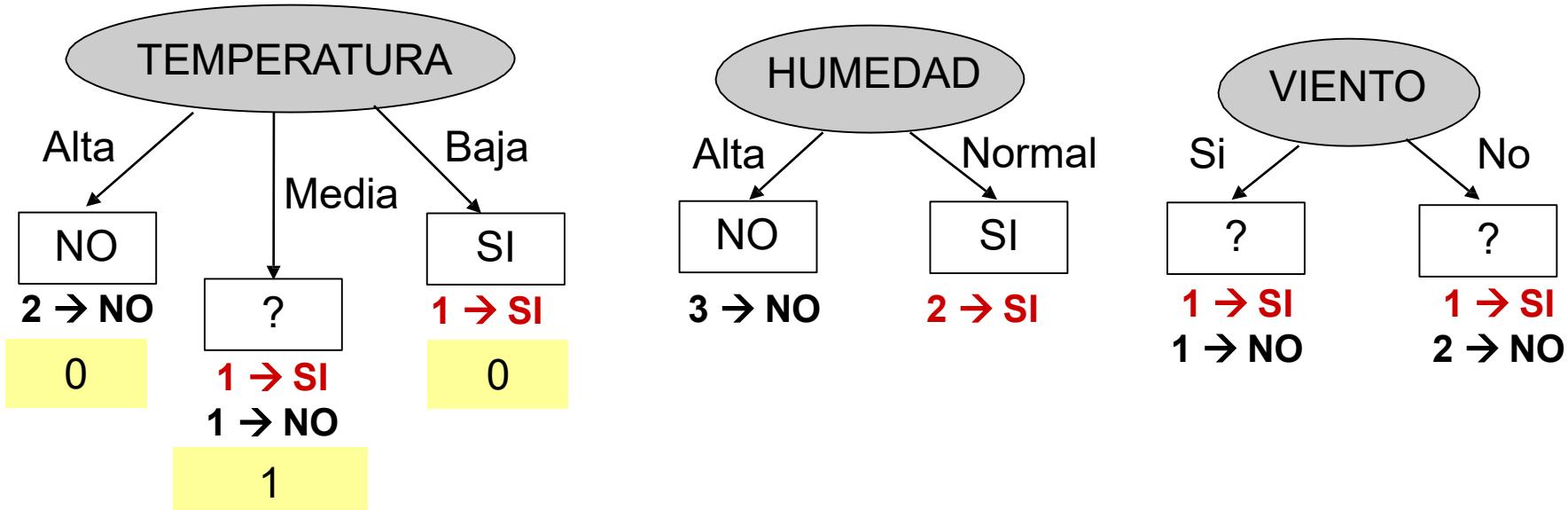
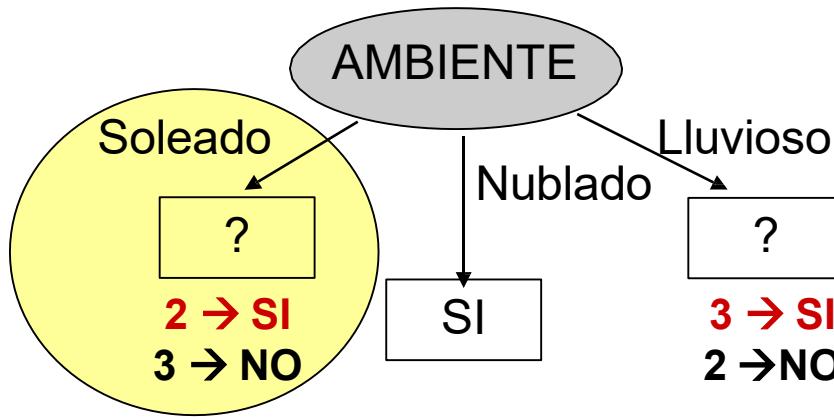
Para estas 5 muestras,
calcular la incertidumbre
de los 3 atributos
restantes

Muestras a considerar para la rama *SOLEADO* del atributo AMBIENTE

Nº	Ambiente	Temperatura	Humedad	Viento	Juega?
1	soleado	alta	alta	no	No
2	soleado	alta	alta	si	No
8	Soleado	media	alta	no	No
9	Soleado	baja	normal	no	Si
11	Soleado	media	normal	si	Si

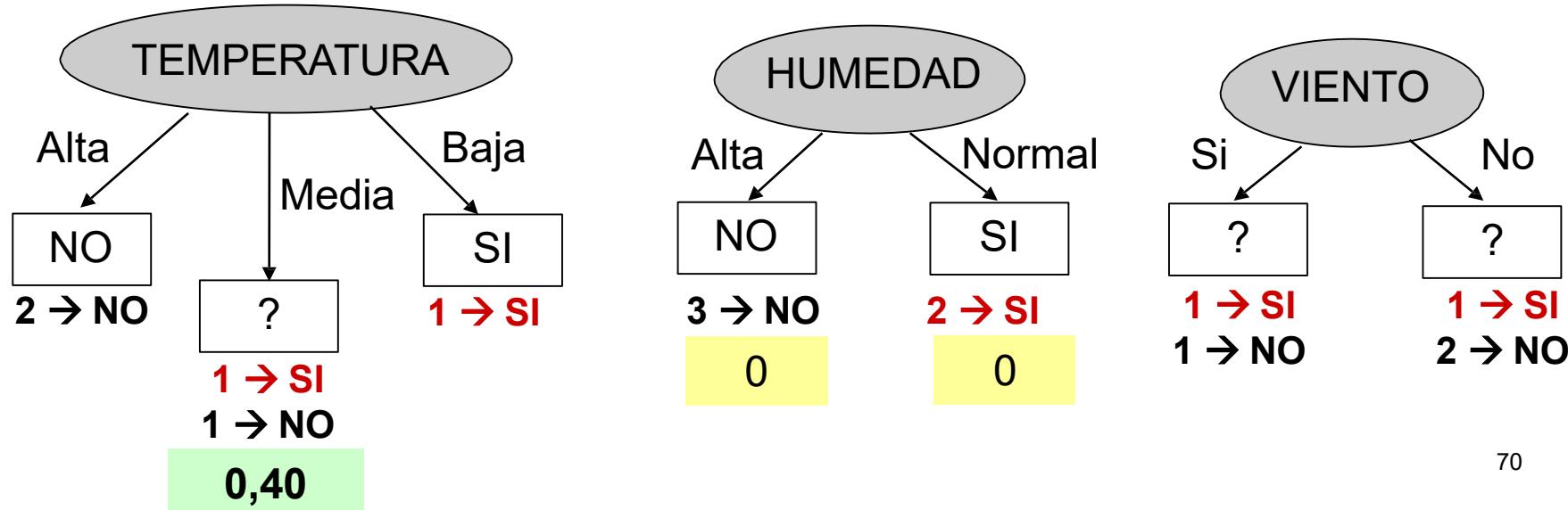
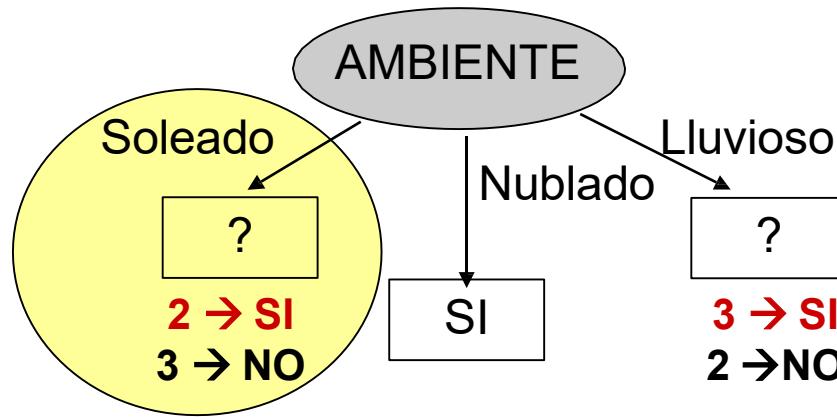
BUSCANDO EL ATRIBUTO QUE MEJOR CLASIFICA LA RAMA

Soleado de Ambiente



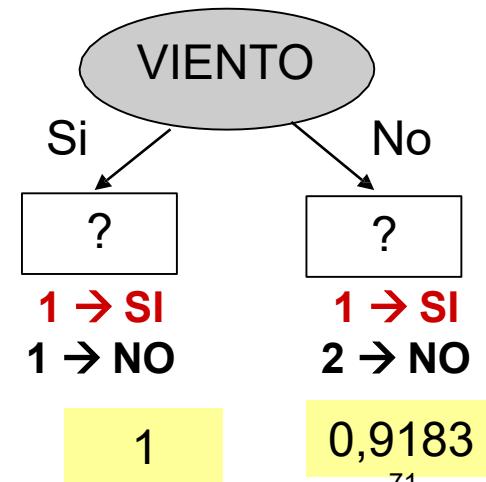
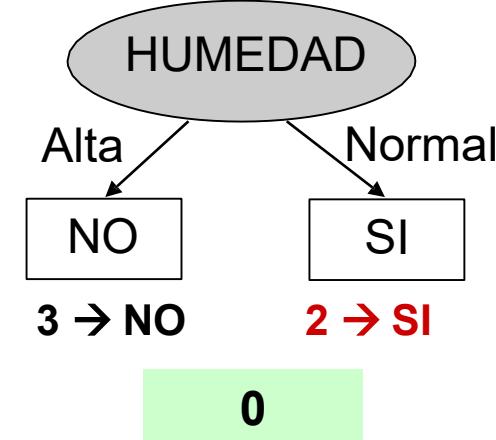
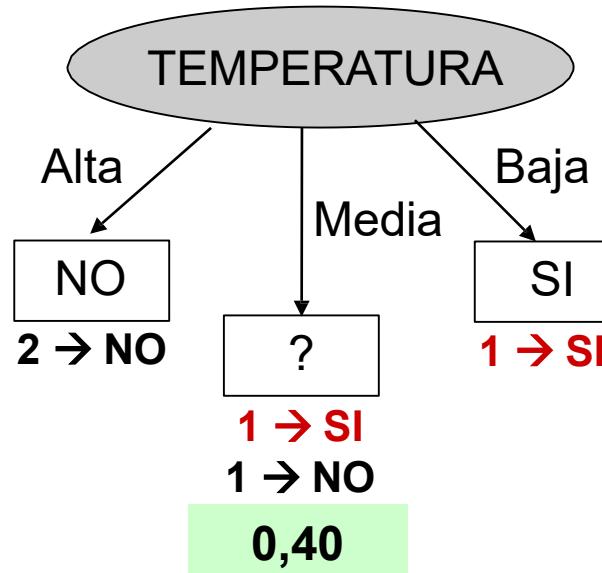
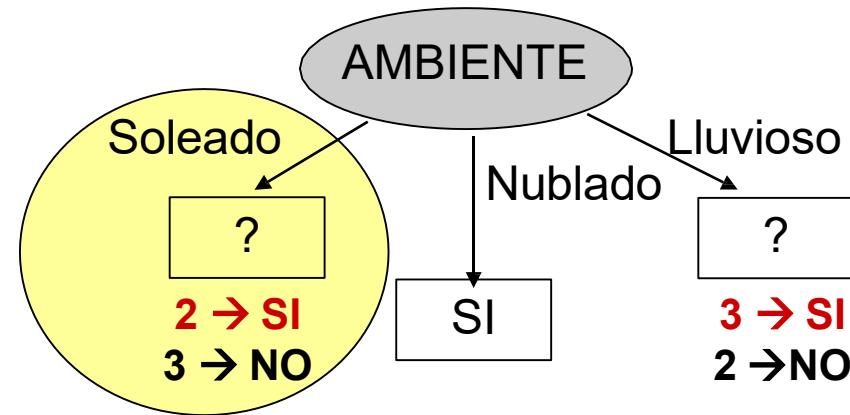
BUSCANDO EL ATRIBUTO QUE MEJOR CLASIFICA LA RAMA

Soleado de Ambiente



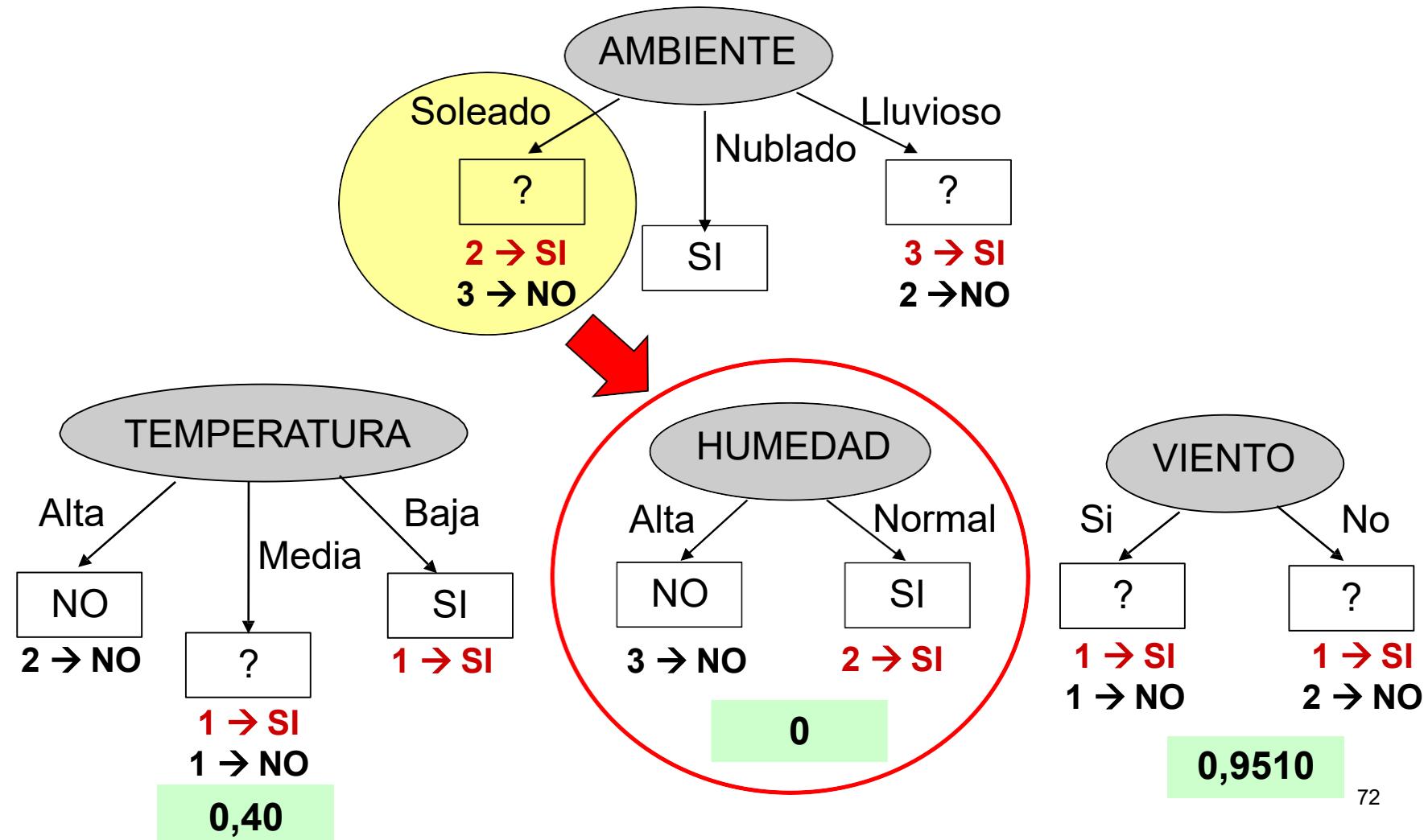
BUSCANDO EL ATRIBUTO QUE MEJOR CLASIFICA LA RAMA

Soleado de Ambiente

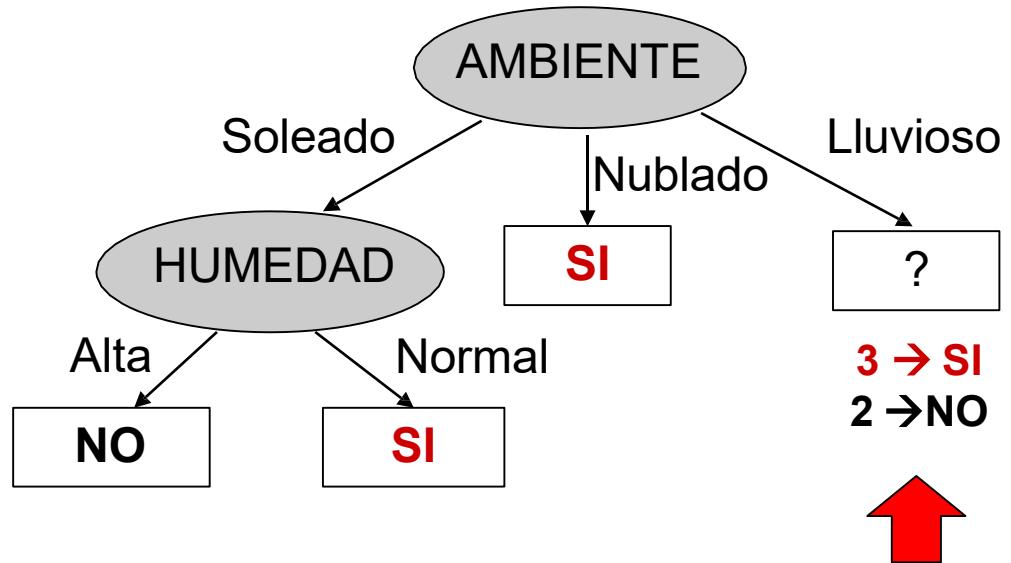


BUSCANDO EL ATRIBUTO QUE MEJOR CLASIFICA LA RAMA

Soleado de Ambiente



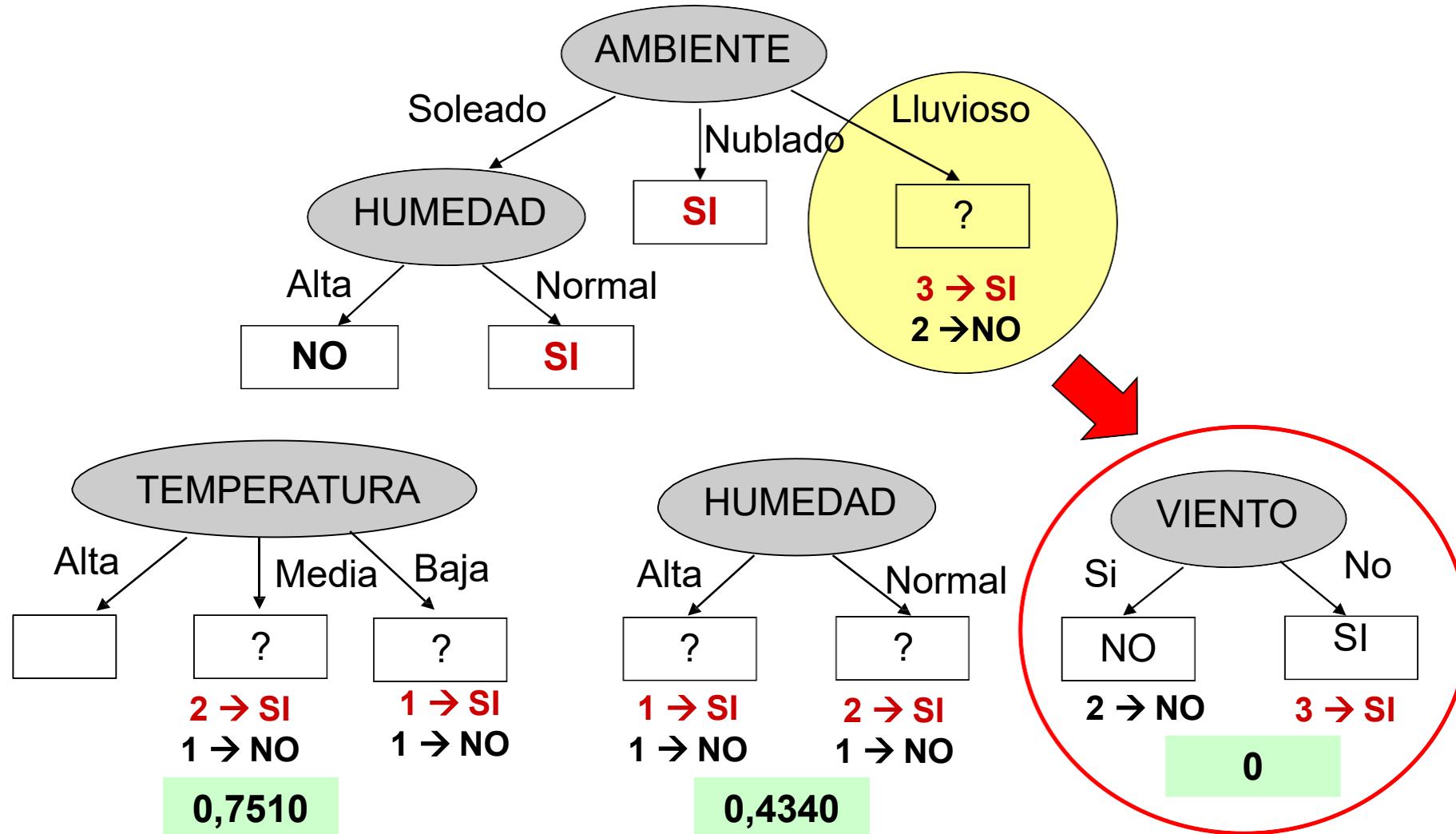
ESTADO ACTUAL DEL ÁRBOL



Para estas 5 muestras,
calcular la entropía de los
3 atributos restantes
(sacando AMBIENTE)

BUSCANDO EL ATRIBUTO QUE MEJOR CLASIFICA LA RAMA

Lluvioso de Ambiente



ARBOL DE CLASIFICACIÓN

