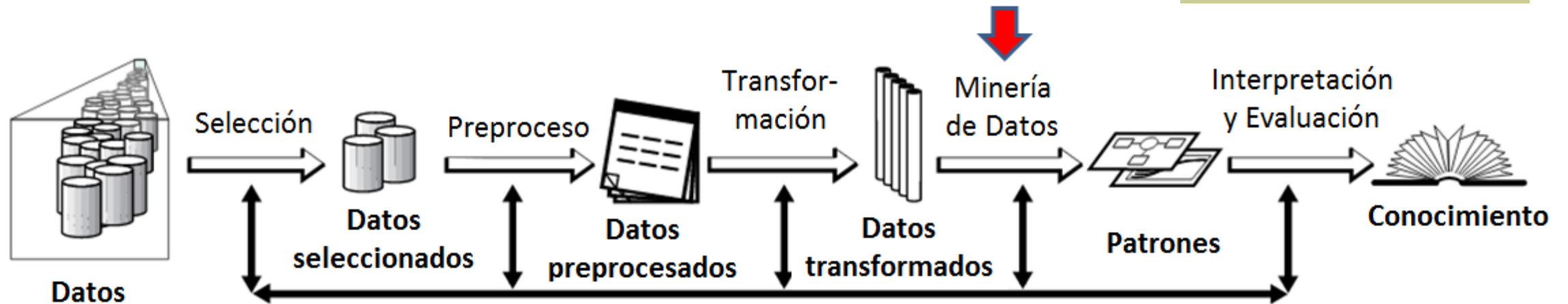


Minería de Datos y el proceso de KDD



Fayyad (1996)

□ Técnicas de Minería de Datos

ARBOL



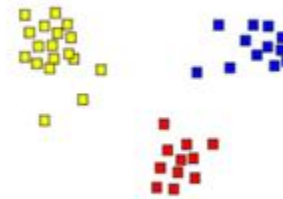
REGLAS

IF (TIPO = CC) AND (SODIO > 470)
ENTONCES (COSTO=BAJO)

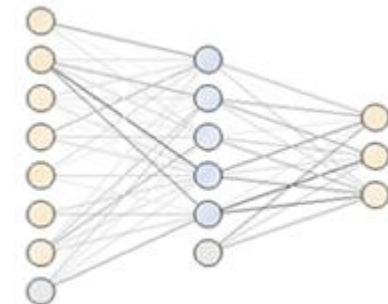
IF (TIPO = CR) AND (PRODUCTO = CN)
ENTONCES (COSTO=ALTO)

IF (TIPO = DC)
ENTONCES (COSTO=MEDIO)

AGRUPAMIENTO



RED NEURONAL



Tipo de conocimiento a extraer

■ Descriptivo

- Muestran nuevas relaciones entre las variables.
- Por ejemplo se buscará describir:
 - Tipos de clientes para diseñar campañas de marketing
 - Transacciones en una tarjeta de crédito para detectar casos anómalos.

■ Predictivo

- En base al modelo construido es posible predecir hechos futuros.
- Por ejemplo se busca predecir:
 - Cuál medicamento suministrar a un paciente dado.
 - Si un mail recibido es spam o no.

Modelos predictivos

- Utiliza **entrenamiento supervisado**

- Se debe contar con un conjunto de **datos etiquetados previamente**. Es decir que, para cada ejemplo, debe conocerse la respuesta esperada.

- Una vez entrenado, es capaz de predecir el valor del atributo indicado previamente como la respuesta esperada (**label**).

- Ejemplos

- Modelo para predecir a qué clase de flor de iris corresponde una flor dada en base a la información suministrada en *Iris.csv*
 - Modelo para predecir la probabilidad de que una persona padezca COVID en base a los datos de su historia clínica y sus síntomas.

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 en Minería de Datos.
- 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

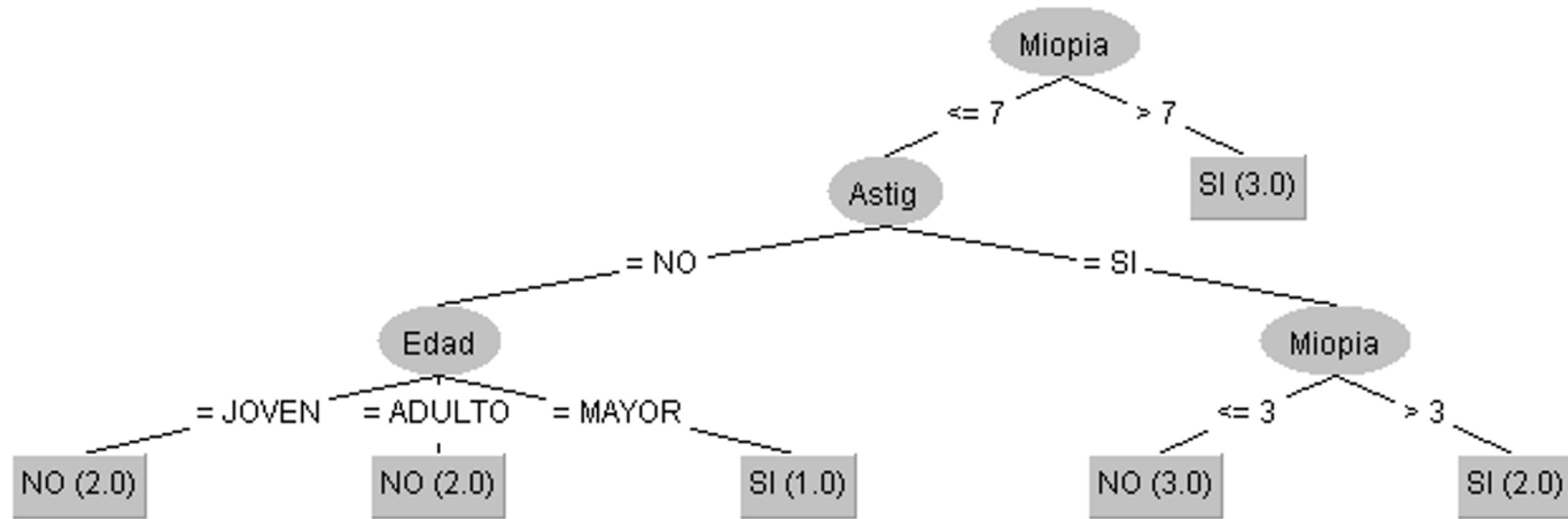
- Su estructura jerárquica les permite mostrar cómo está organizada la información disponible.

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)

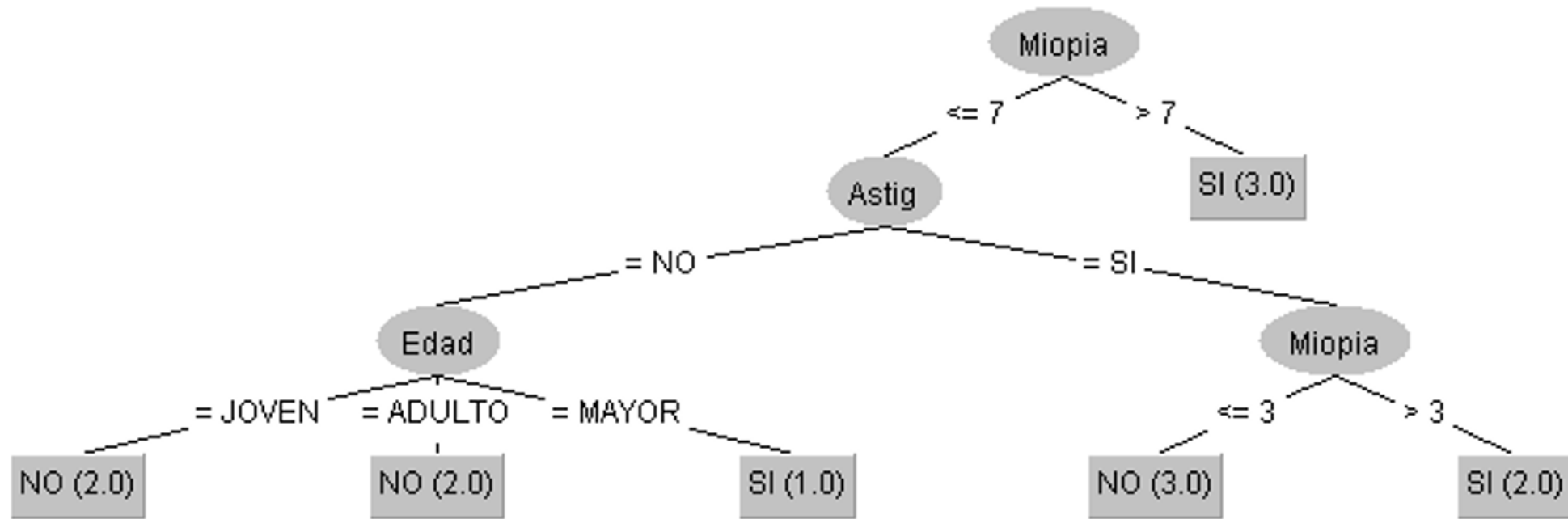
A partir de esta información puede obtenerse un **modelo** en forma de **árbol** que resuma el criterio seguido para recomendar si debe operarse o no.

Arbol de decisión. Ejemplo



Note que las opciones son excluyentes

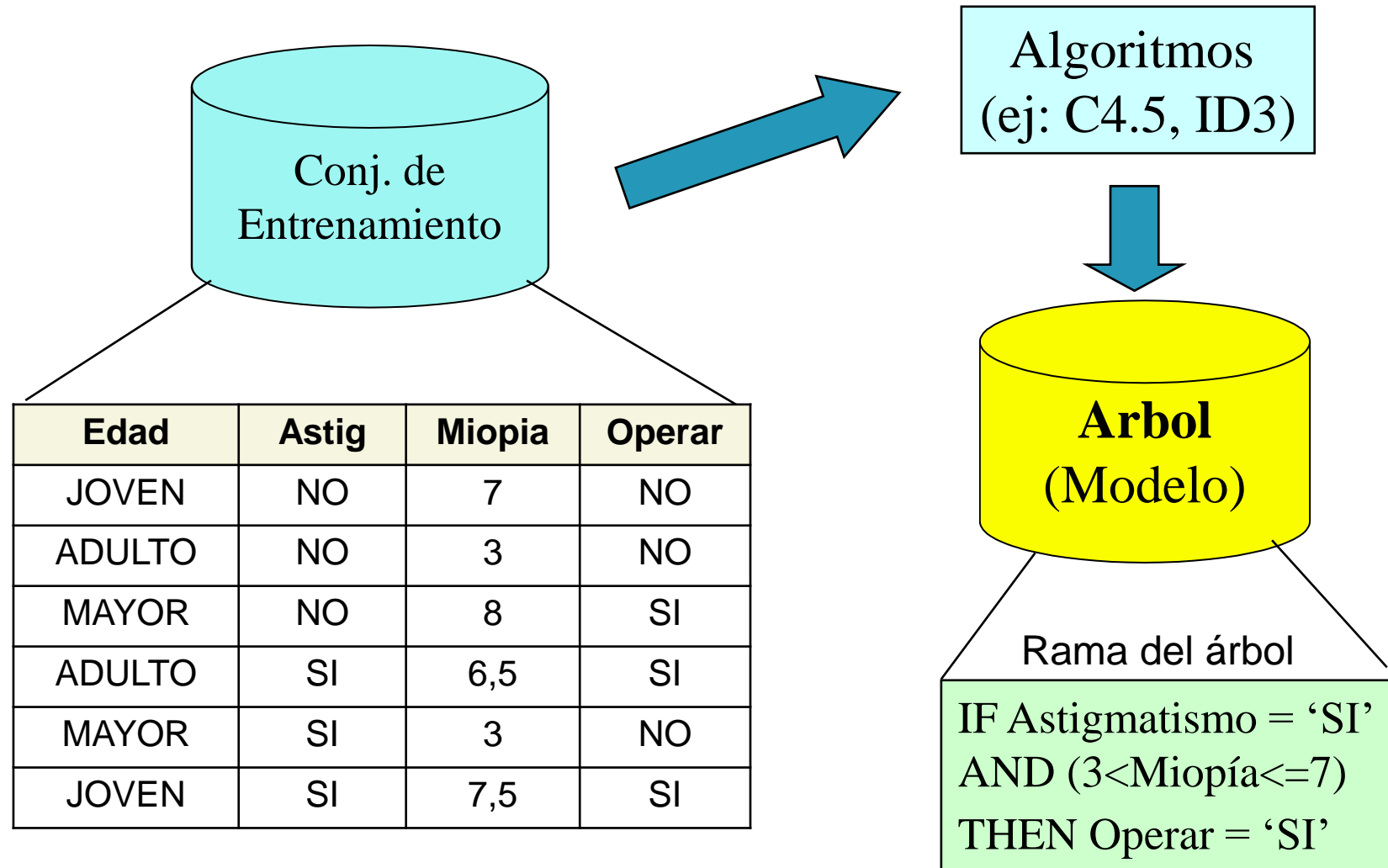
Arbol de decisión. Ejemplo



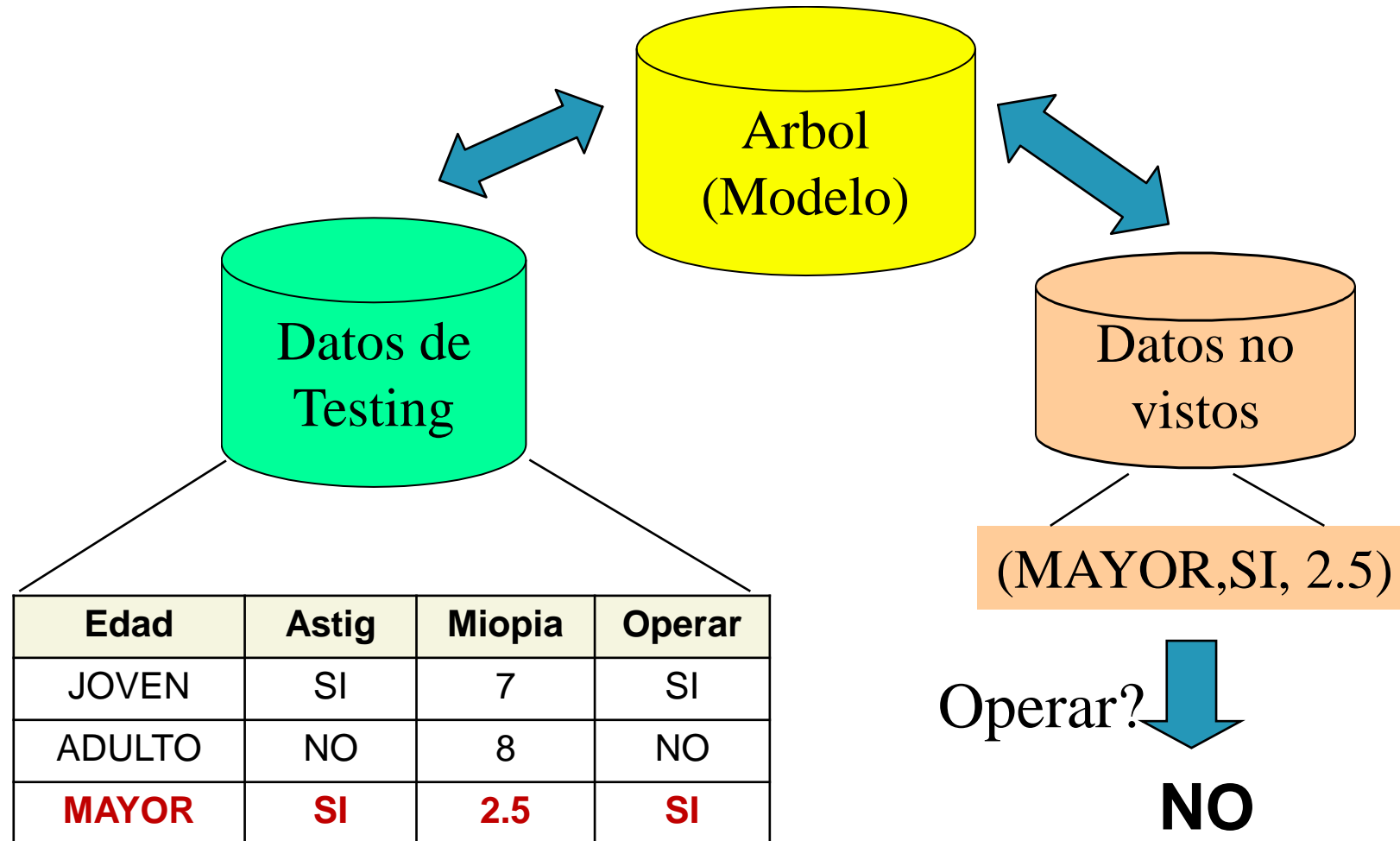
Un atributo cualitativo no puede aparecer más de una vez en la misma rama

y un atributo cuantitativo?

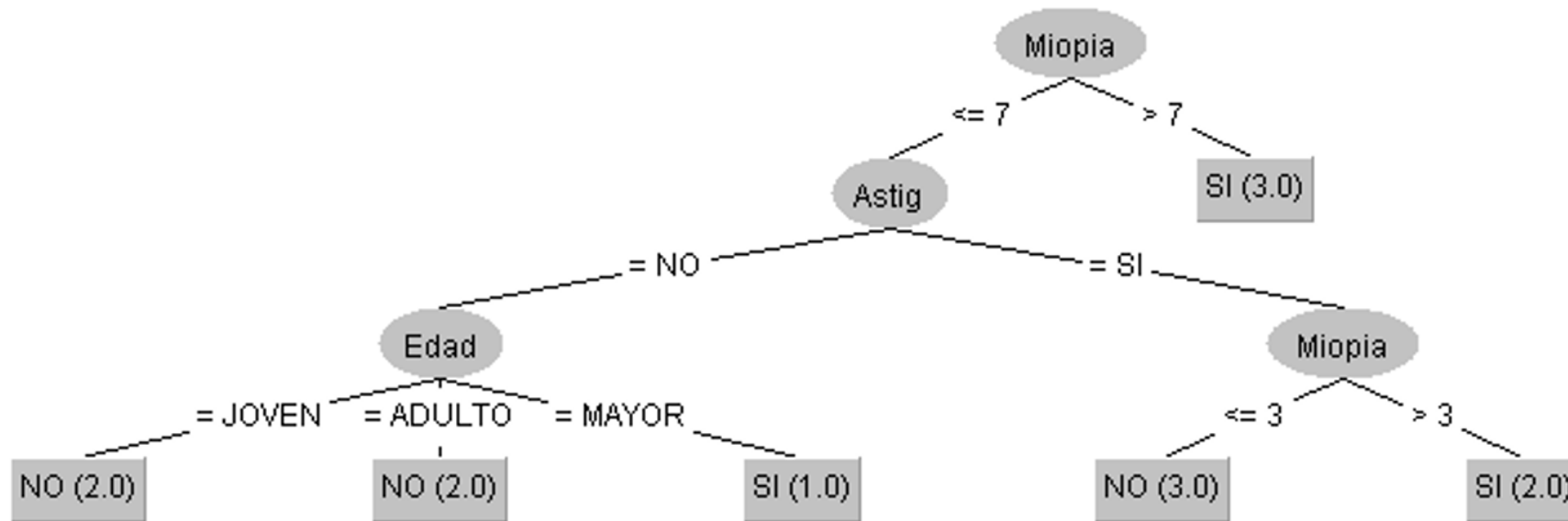
Obtención del modelo



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 recursive divide-and-conquer**.
- 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 heurística o una medida estadística (p.ej., **ganancia de información**).

Obtención del árbol de decisión

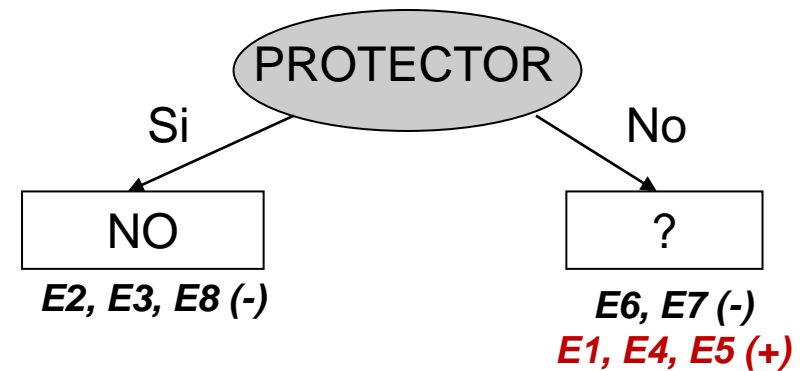
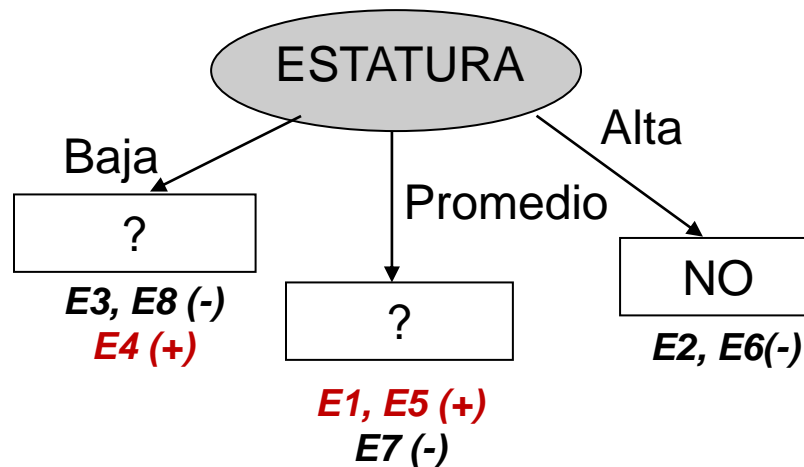
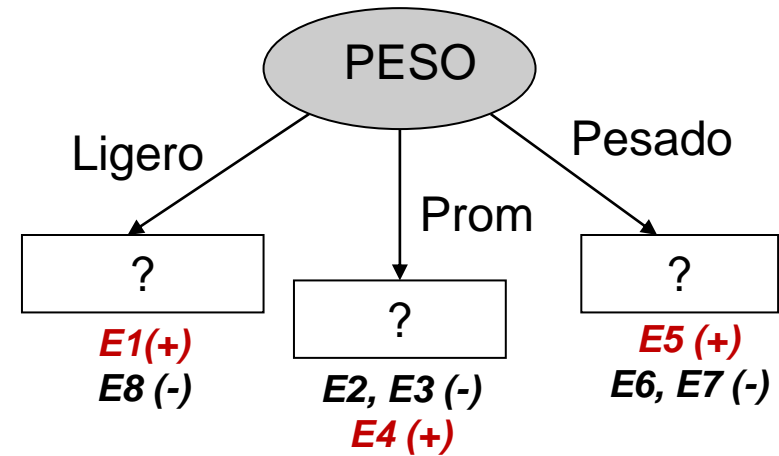
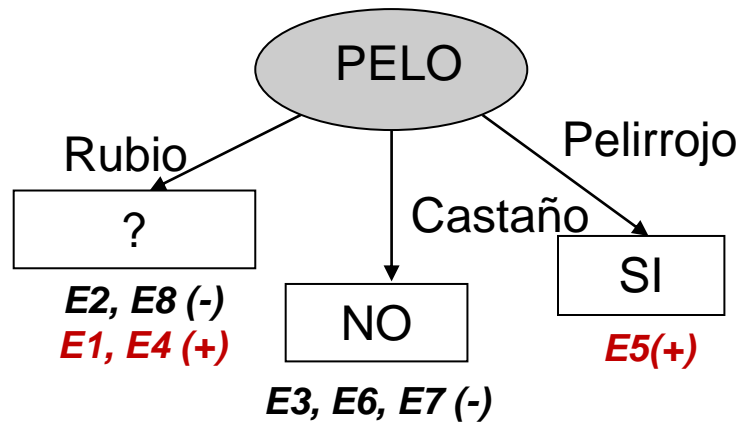
- Condiciones para detener el particionamiento
 - Todas las muestras, para un nodo dado, corresponden a la misma clase.
 - No hay atributos restantes para particionar. Se usa **voto mayoritario** para clasificar la hoja.
 - No quedan más muestras (registros del conjunto de entrenamiento).

Ejemplo 1: AlSol.csv

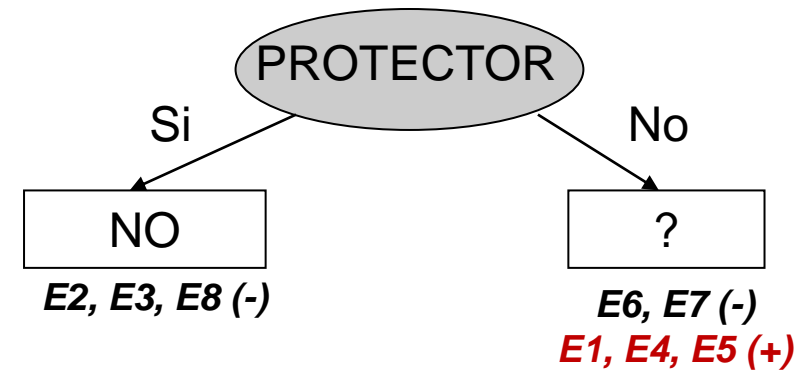
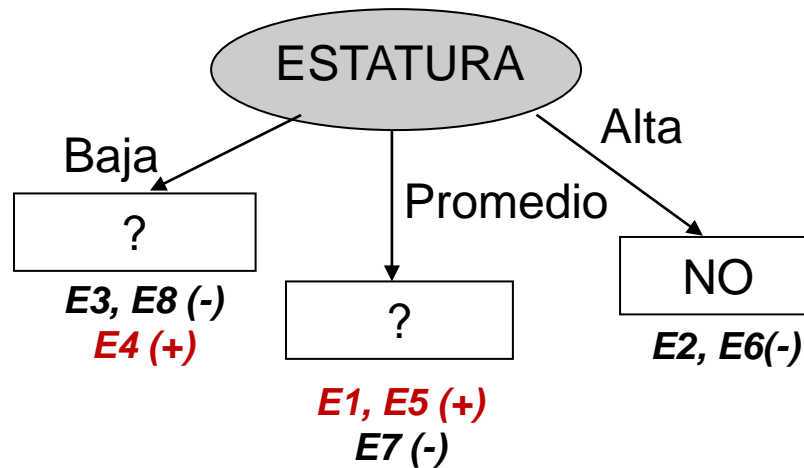
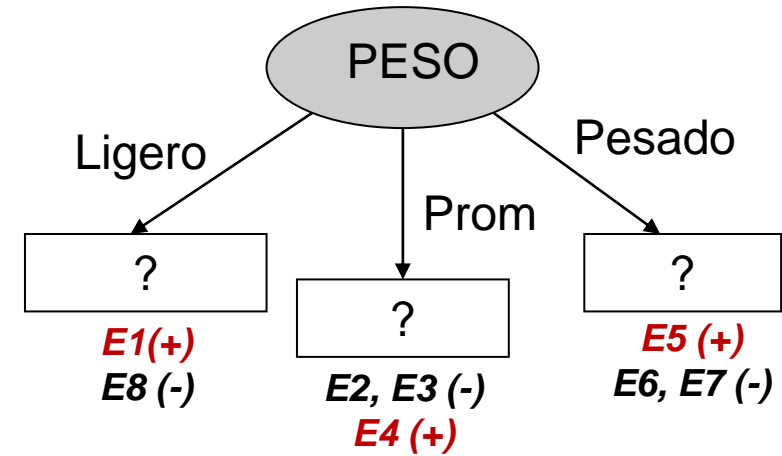
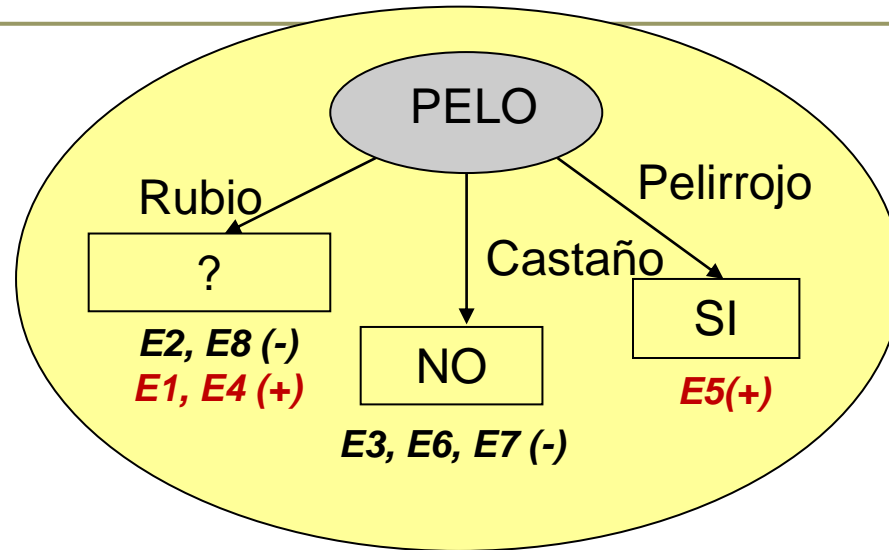
id	Nombre	Pelo	Estatura	Peso	Protector	Quemado
E1	Sara	Rubio	Promedio	Ligero	No	SI
E2	Diana	Rubio	Alta	Promedio	Si	NO
E3	Alexis	Castaño	Baja	Promedio	Si	NO
E4	Ana	Rubio	Baja	Promedio	No	SI
E5	Emilia	Pelirrojo	Promedio	Pesado	No	SI
E6	Pedro	Castaño	Alta	Pesado	No	NO
E7	Juan	Castaño	Promedio	Pesado	No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

- ¿Cuál atributo elegiría como raíz del árbol?

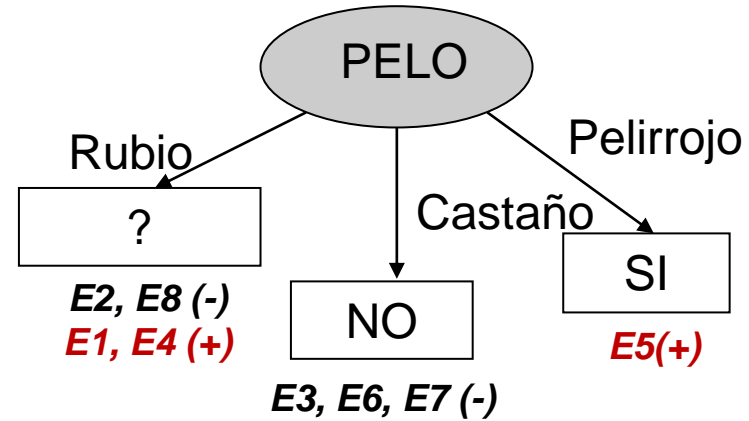
¿Qué pasaría si eligiera?



Es la seleccionada por tener la mayor cantidad de elementos en subconjuntos homogéneos

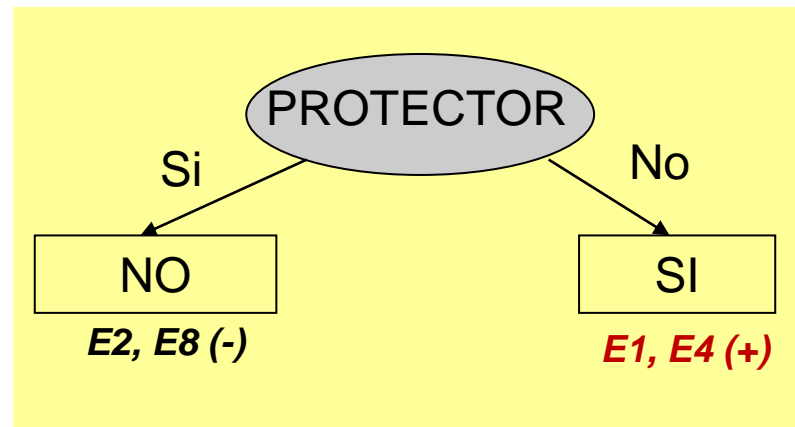
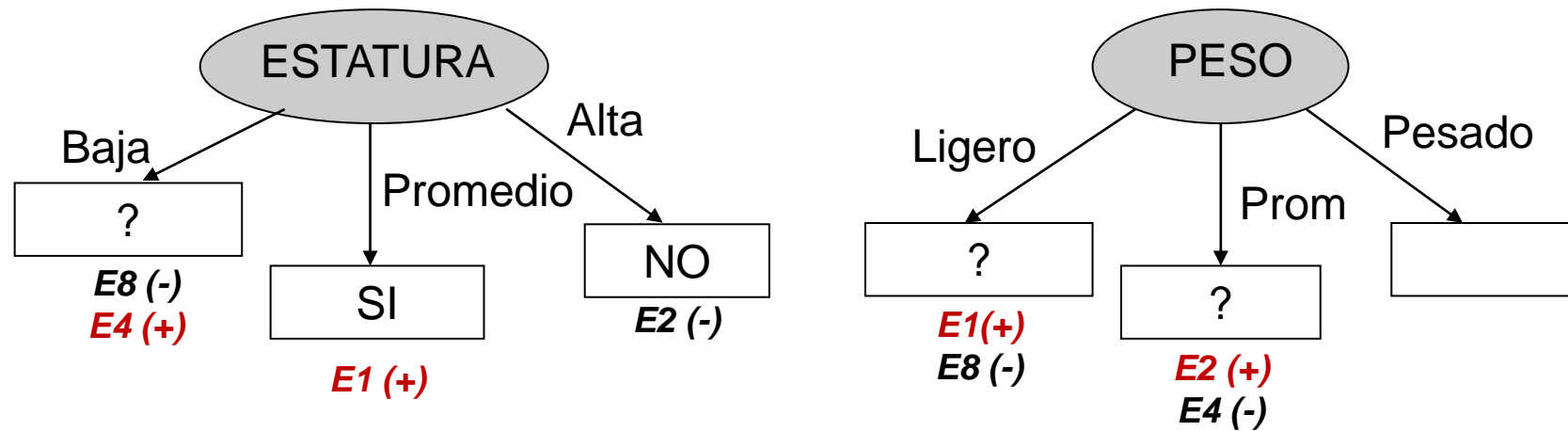


¿Cómo sigue?



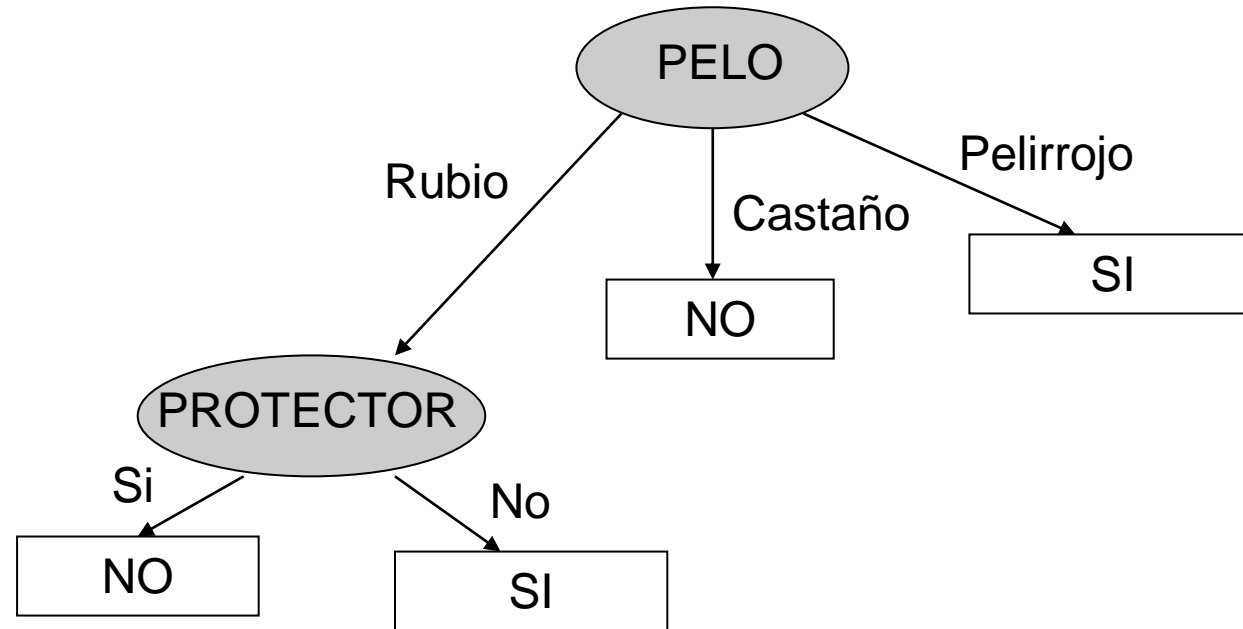
Analizar la repuesta del resto de los atributos para los ejemplos que aún no pertenecen a un subconjunto homogéneo

¿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 su clasificación objetivo.
- Existen distintos criterios
 - Ganancia de Información (Algoritmo Id3) 
 - Tasa de Ganancia (Algoritmo C4.5)
 - Índice 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:

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

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

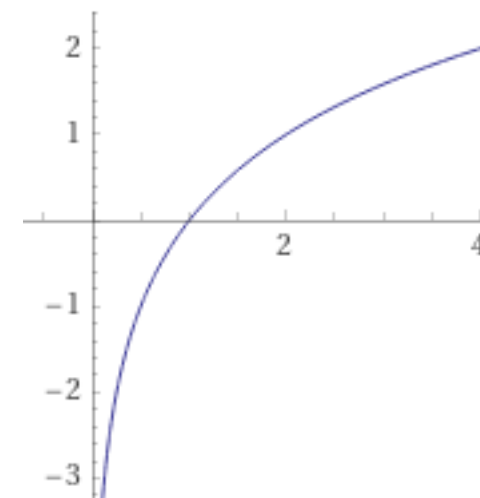
Entropía

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

x	$\log_2(x)$
1/8	-3,00
2/8	-2,00
3/8	-1,42
4/8	-1,00
5/8	-0,68
6/8	-0,42
7/8	-0,19
8/8	0,00

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

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



Entropía

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

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

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



Este valor mide la heterogeneidad del conjunto de ejemplos.
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:

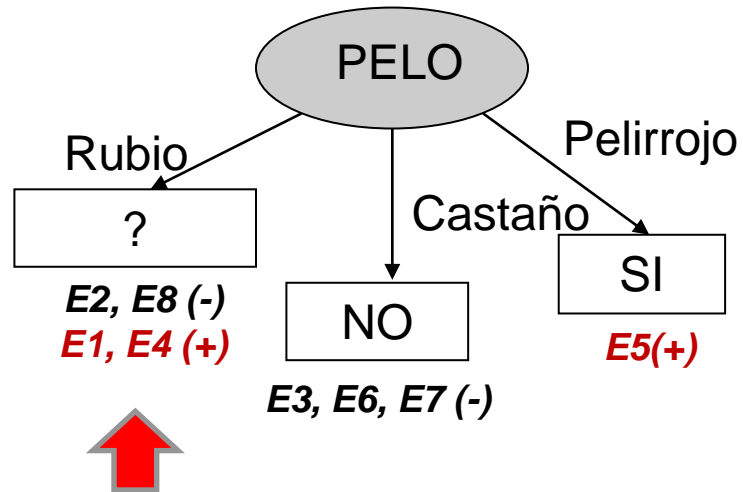
$$Entropia(E, A) = \sum_{v \in V_a} \frac{|E_v|}{E} Entropia(E_v)$$

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 (-)

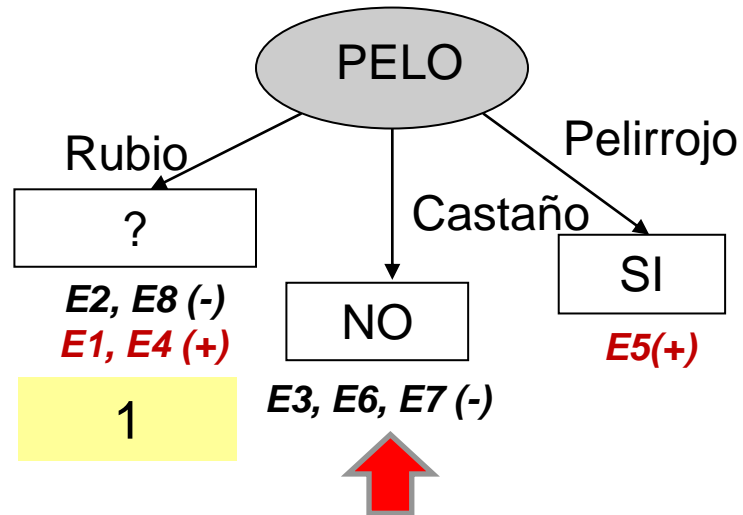


$$Entropia(E_{rubio}) = -\left(\frac{2}{4}\right) * \log_2\left(\frac{2}{4}\right) - \left(\frac{2}{4}\right) * \log_2\left(\frac{2}{4}\right) = 1$$

Entropía de PELO

E

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

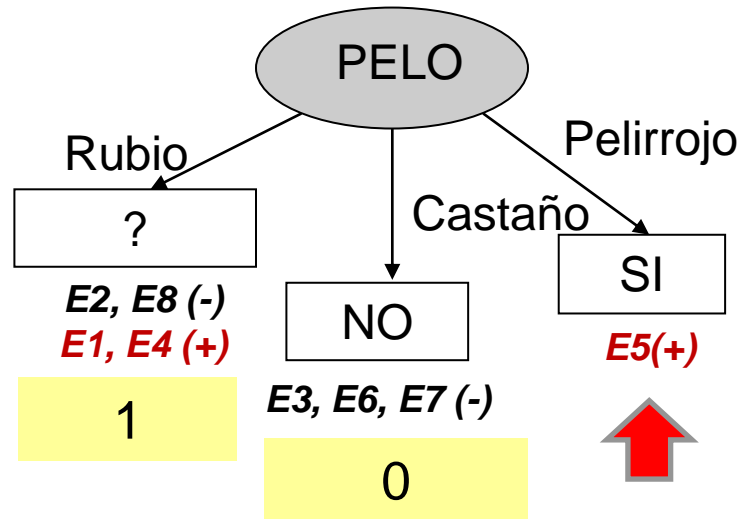


$$Entropia(E_{castaño}) = - \left(\frac{3}{3} \right) * \log_2 \left(\frac{3}{3} \right) = 0$$

Entropía de PELO

E

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

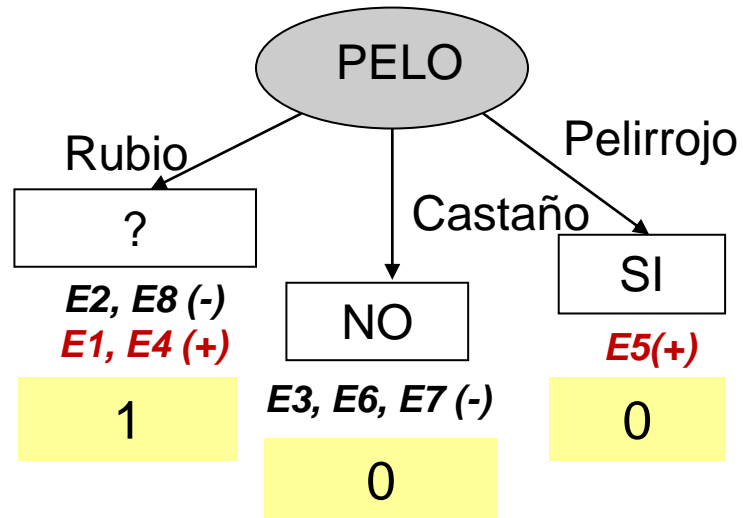


$$Entropia(E_{pelirrojo}) = -\left(\frac{1}{1}\right) * \log_2\left(\frac{1}{1}\right) = 0$$

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

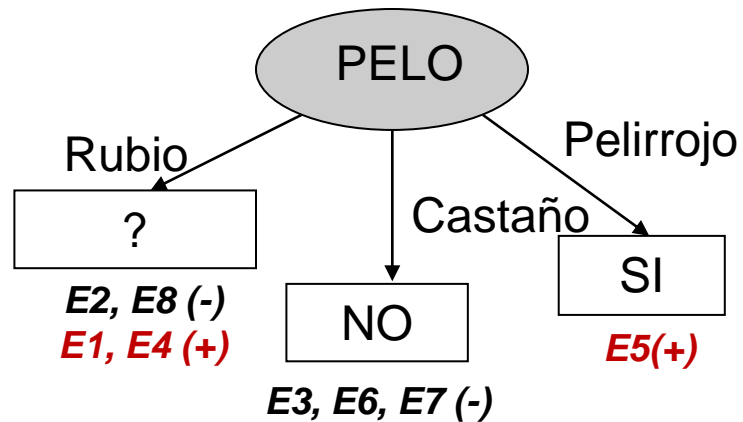
- Dado un atributo A y un conjunto de ejemplos E , la Ganancia de Información de dicho atributo se calcula así:

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

Ganancia de Información

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

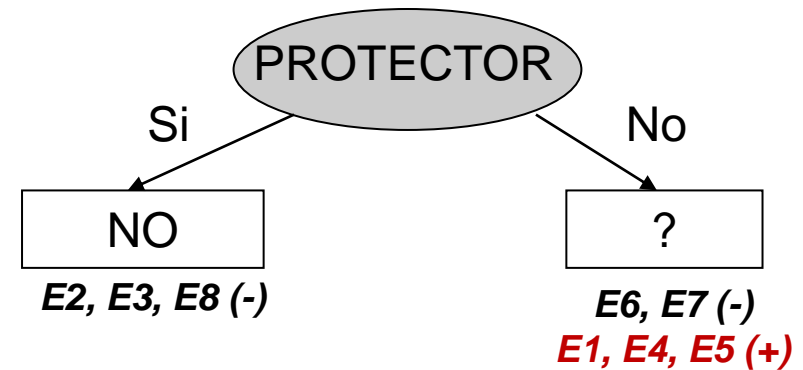
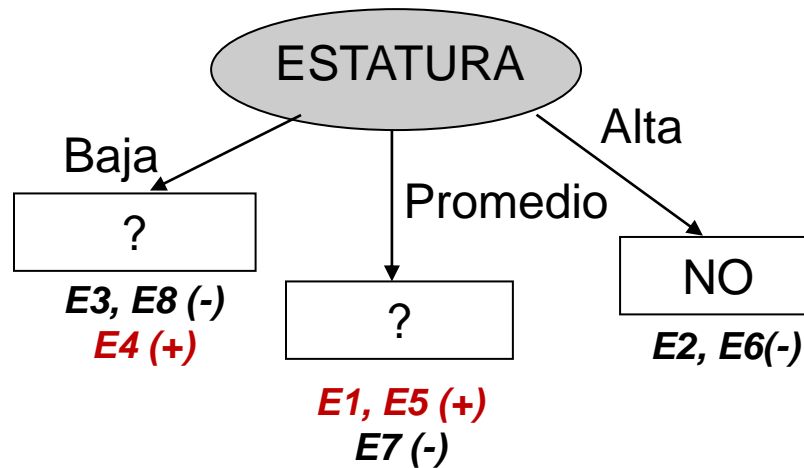
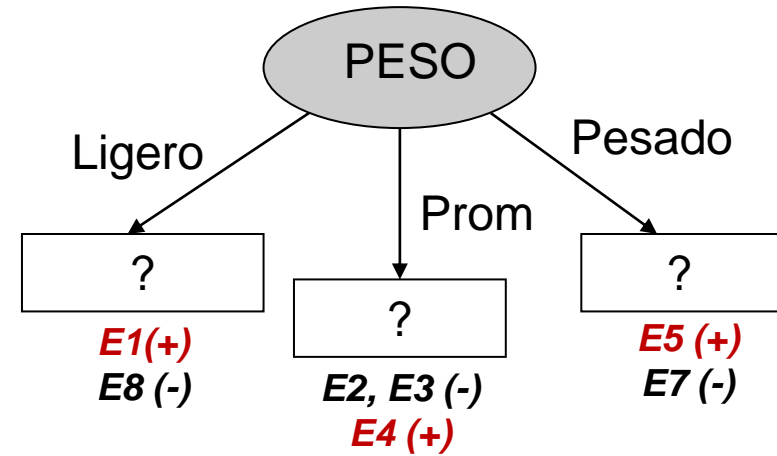
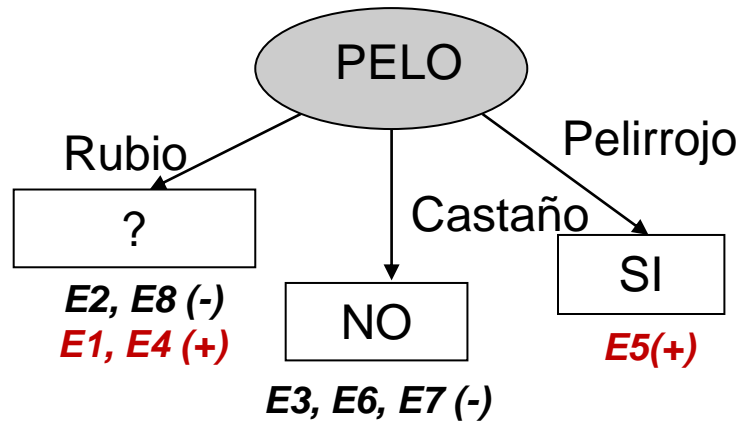
← $Entropía(E)=0.9544$



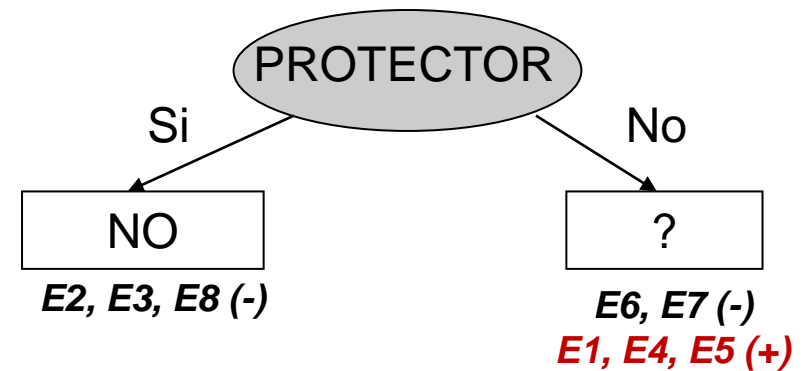
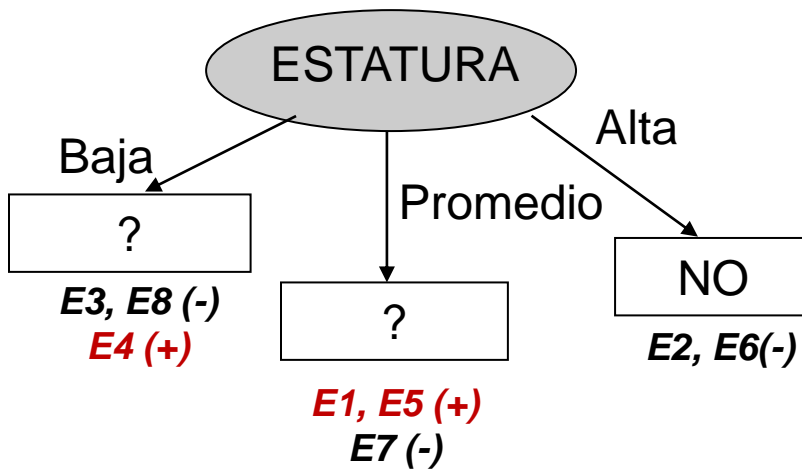
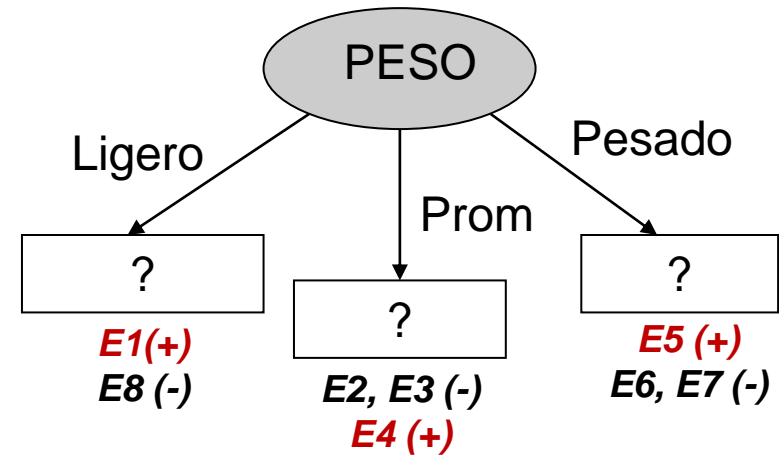
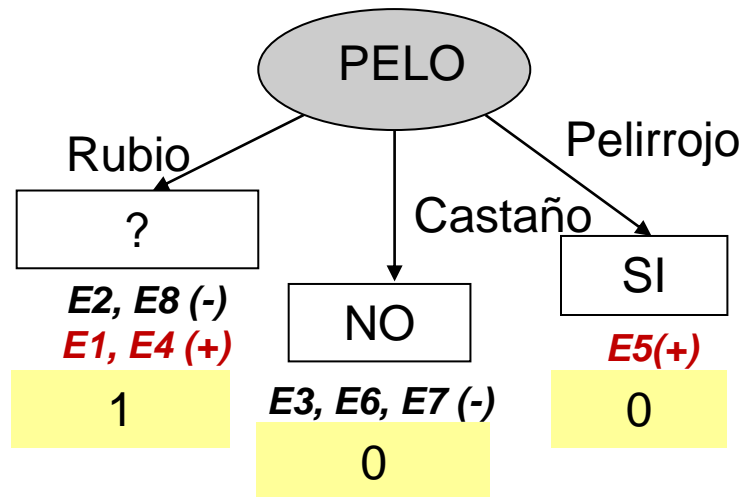
← $Entropía(E, PELO)=0.5$

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

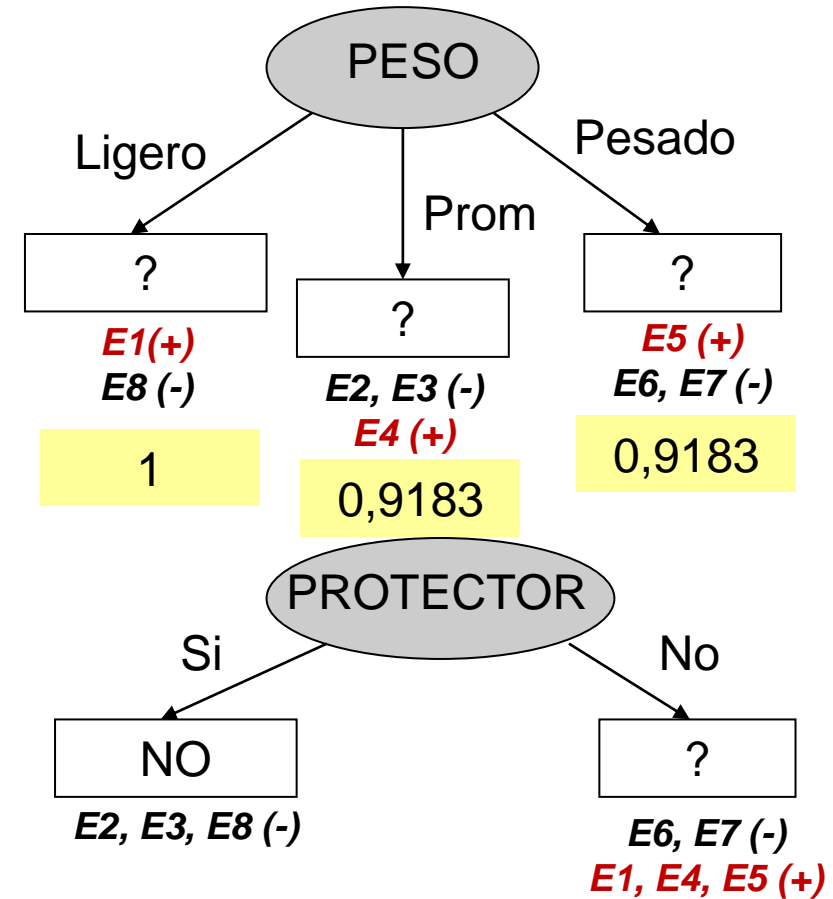
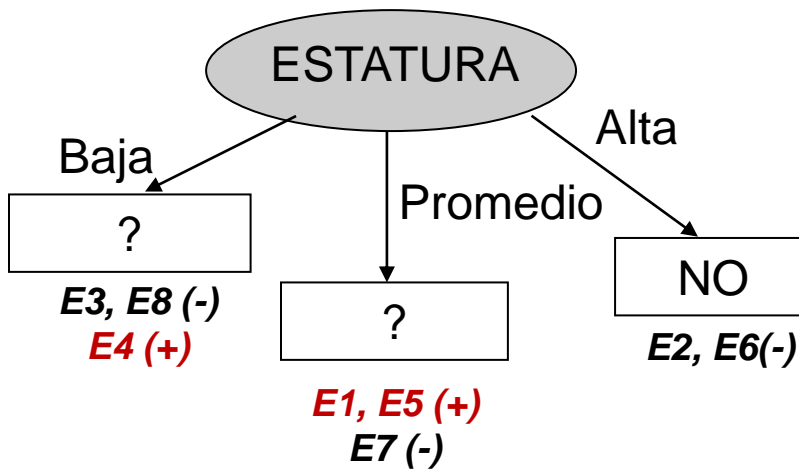
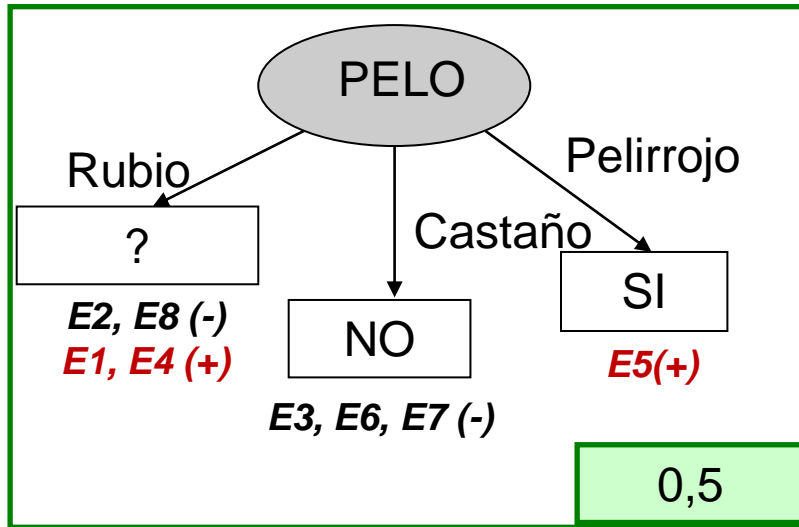
¿Qué pasaría si eligiera?



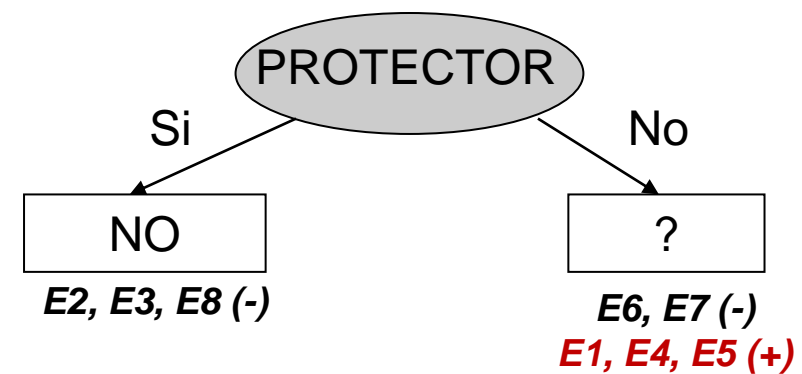
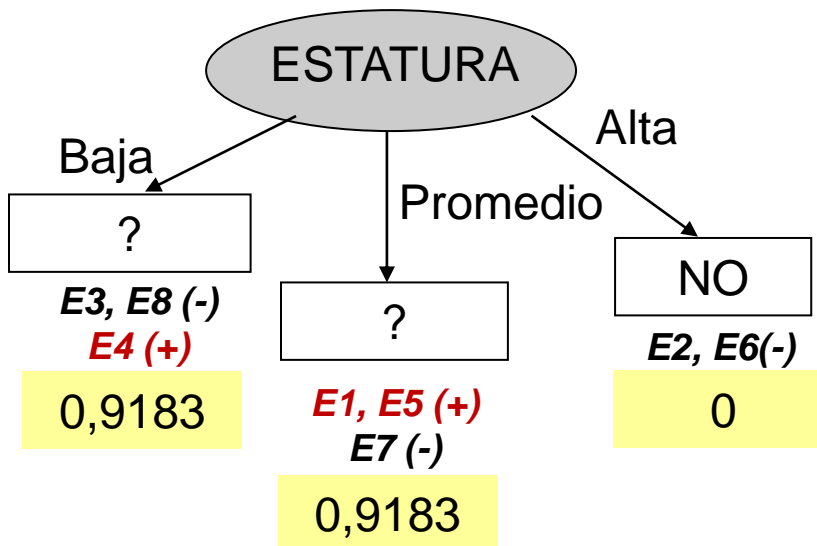
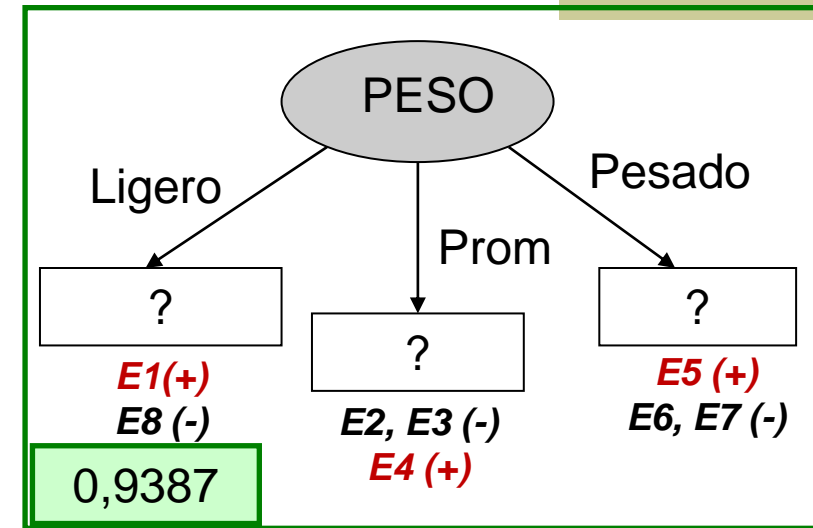
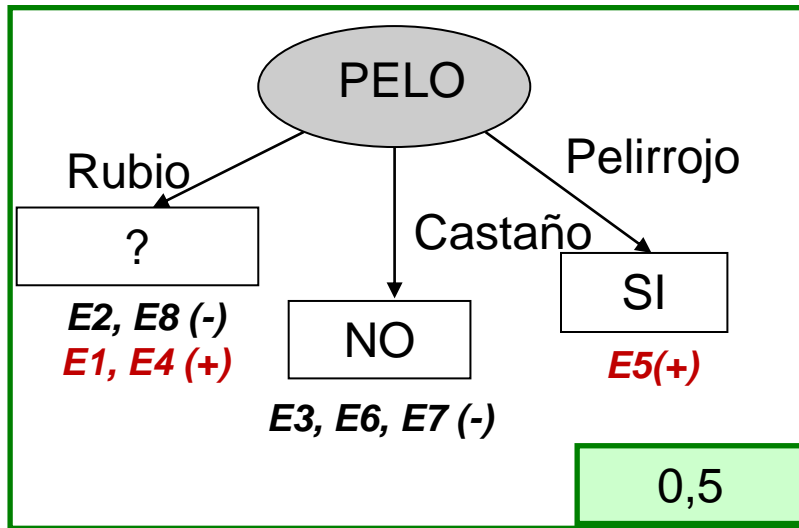
¿Qué pasaría si eligiera?



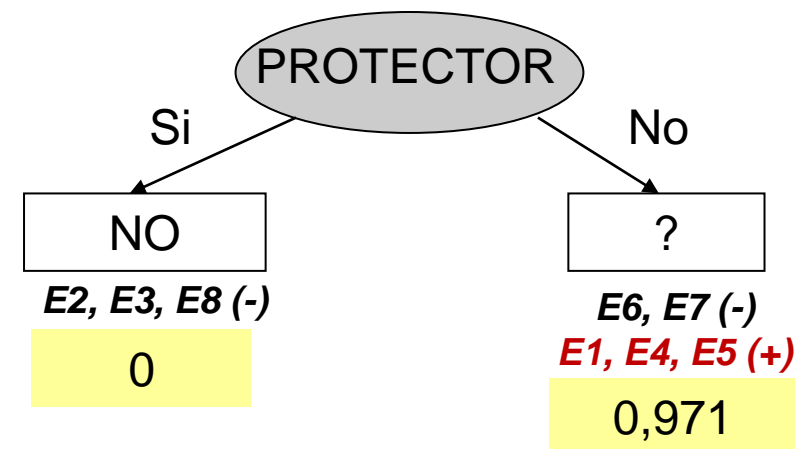
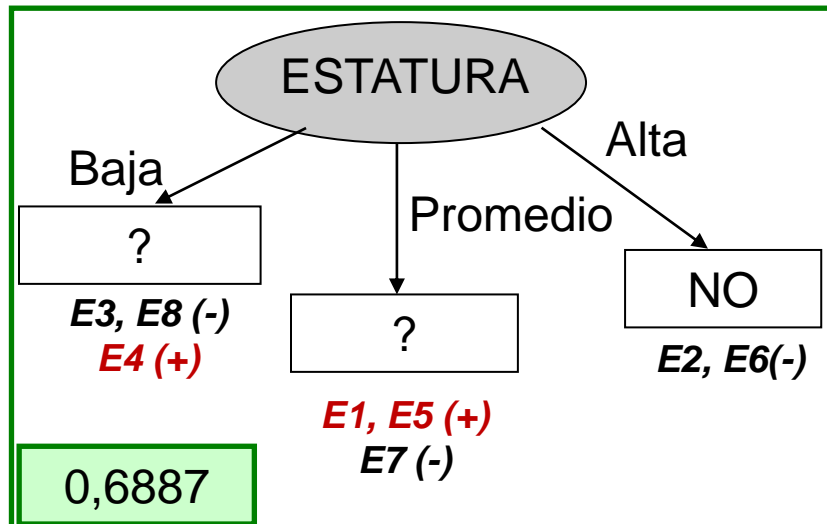
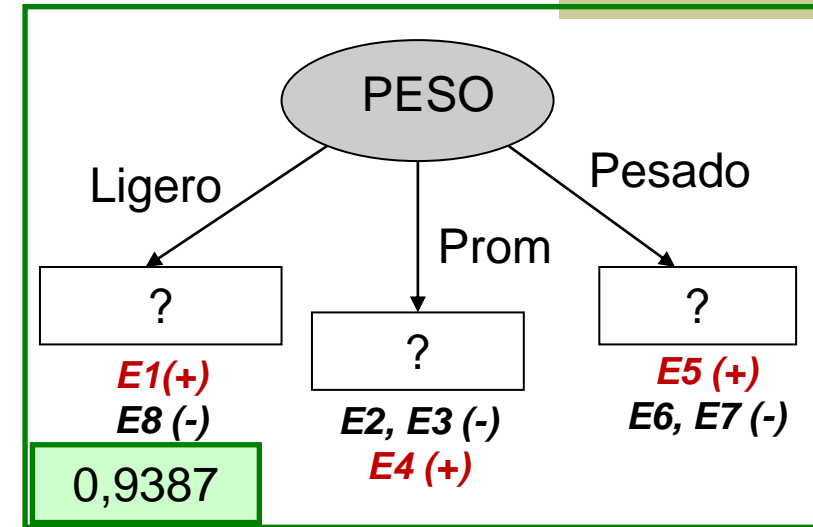
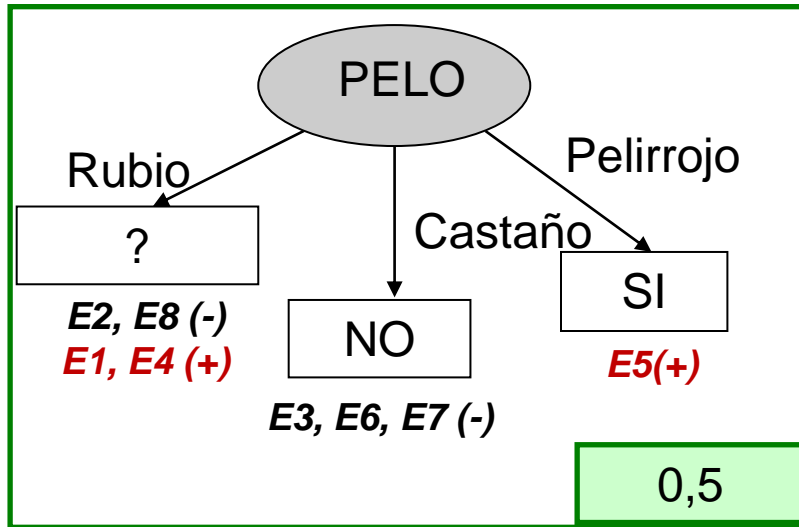
¿Qué pasaría si eligiera?



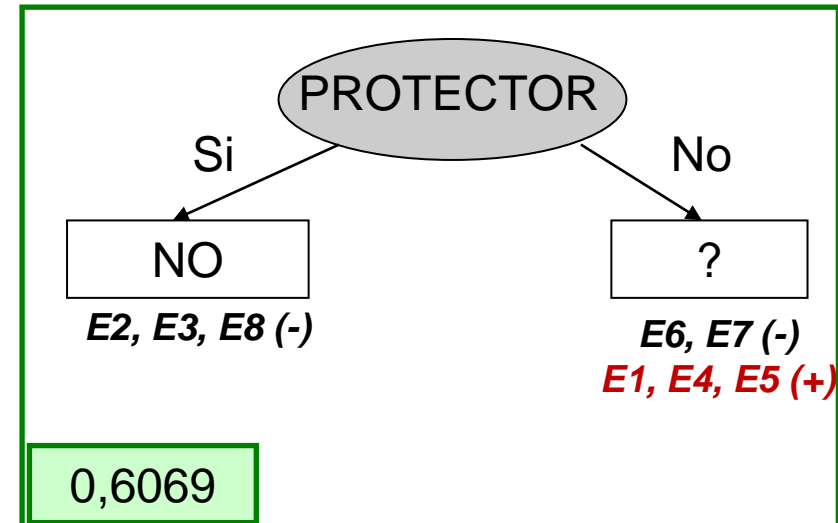
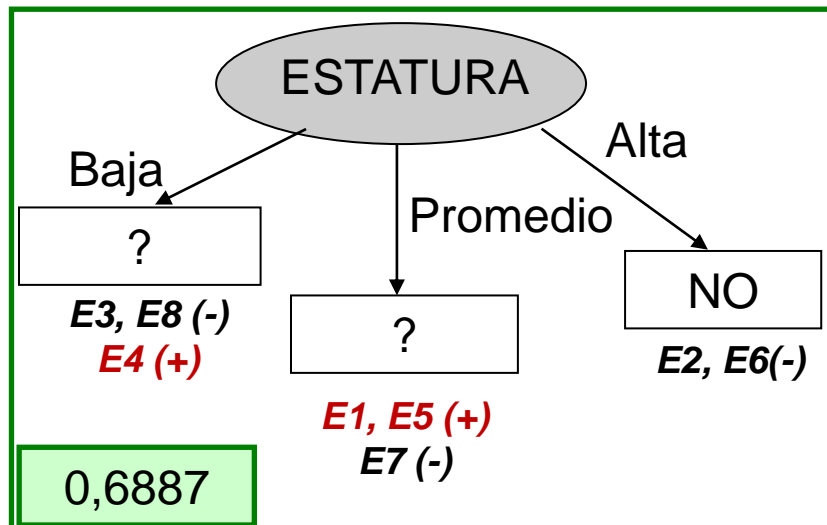
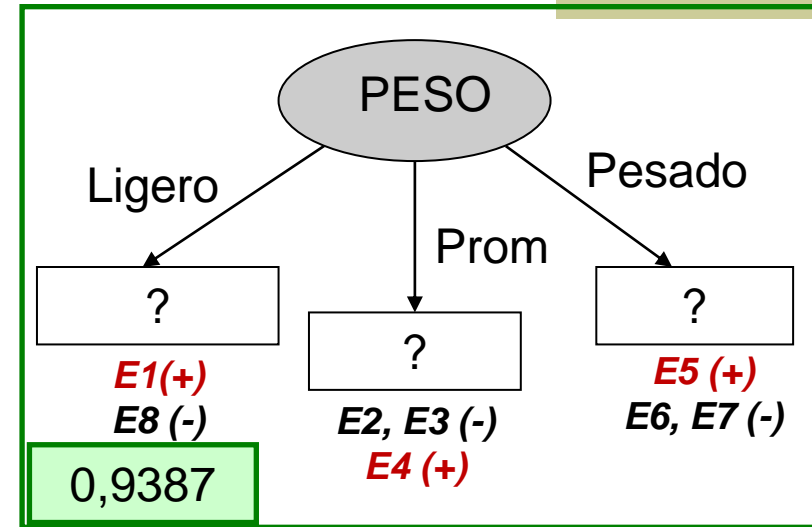
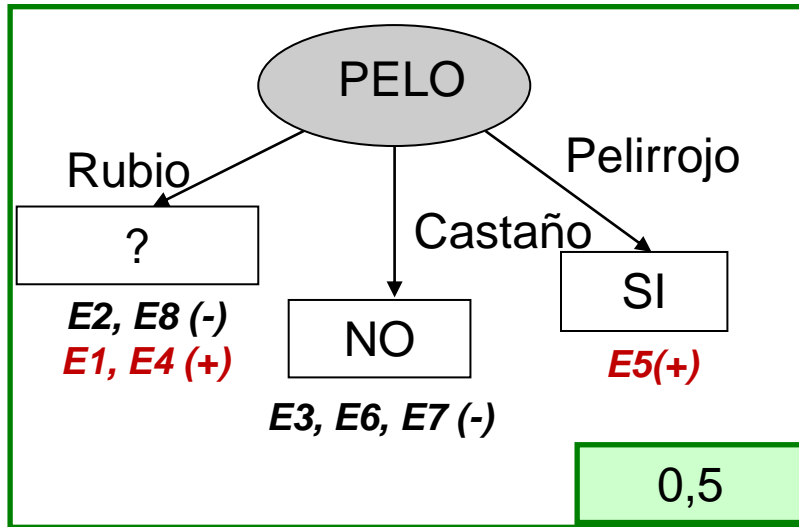
¿Qué pasaría si eligiera?



¿Qué pasaría si eligiera?



¿Qué pasaría si eligiera?



Selección por Ganancia de Información

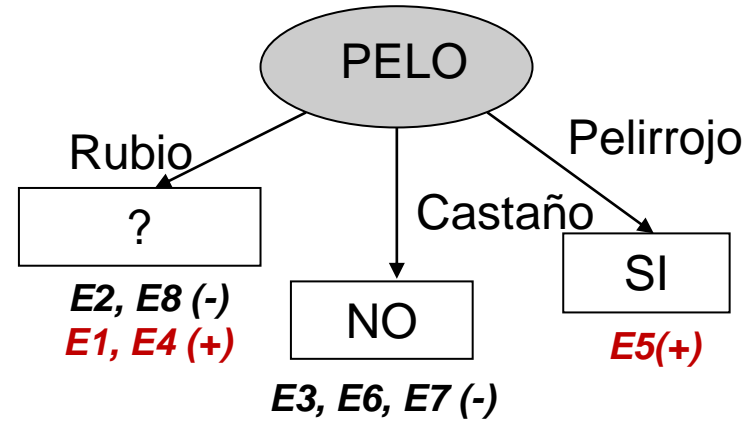
- **$Entropía(E) = 0.9544$**
- **$Ganancia(E,A) = Entropía(E) - Entropía(E,A)$**

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



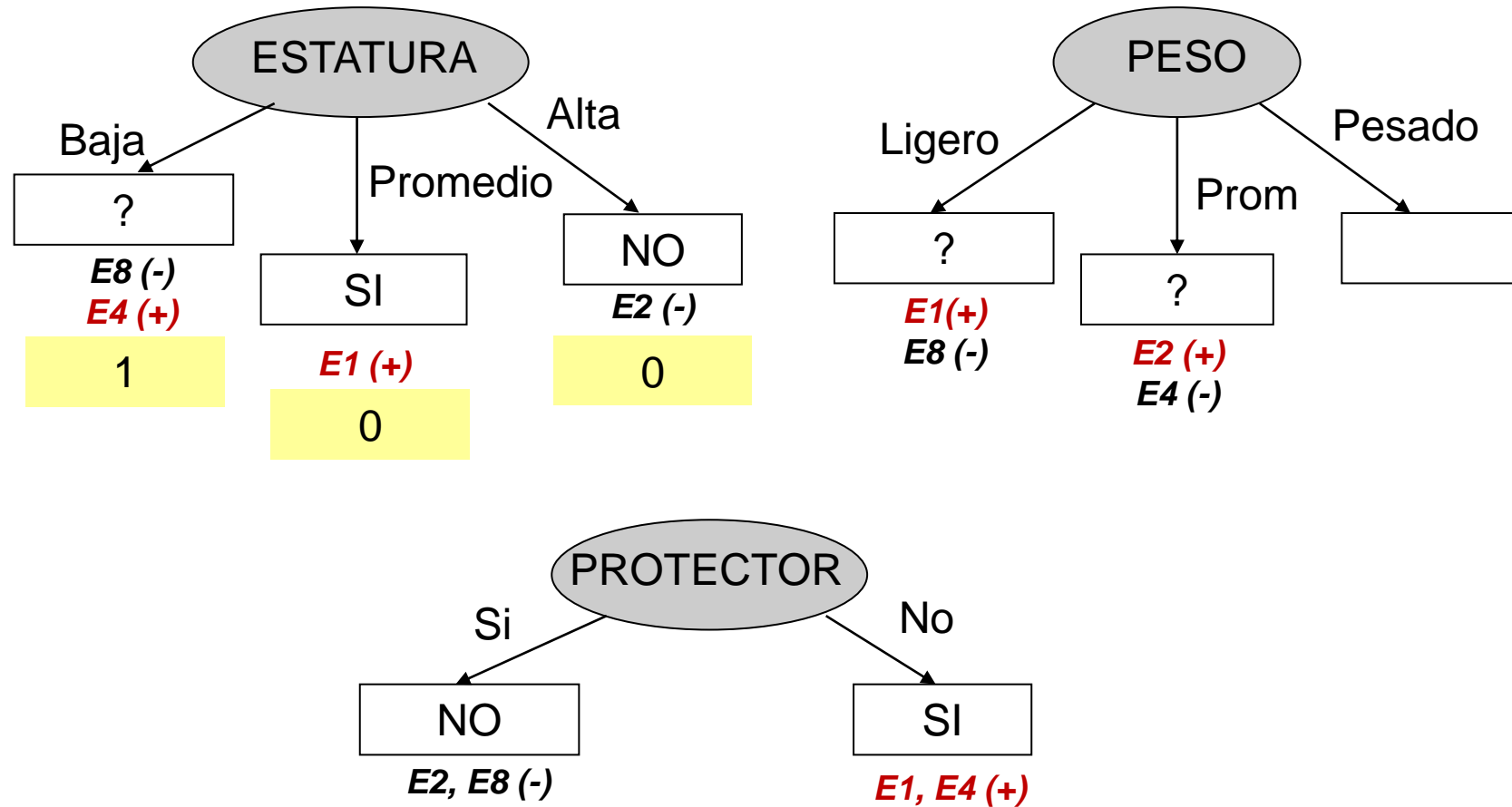
Es el
seleccionado
por ser el de
mayor
Ganancia

¿Cómo sigue?

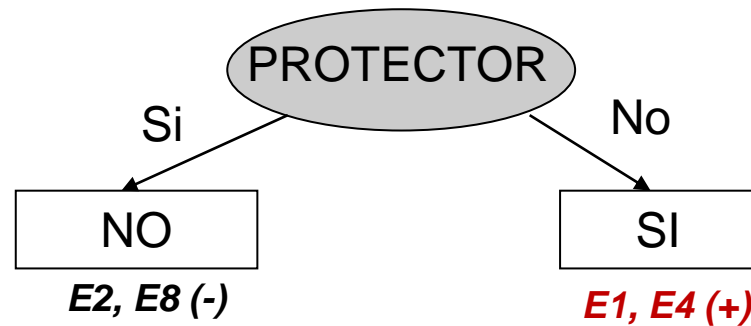
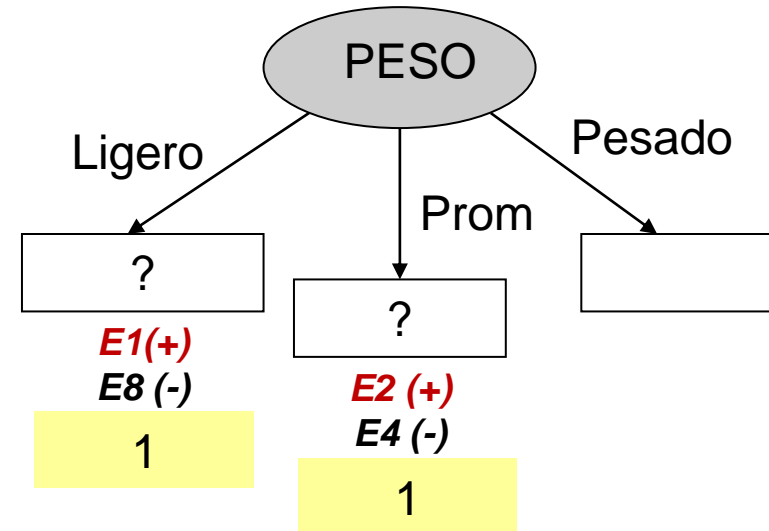
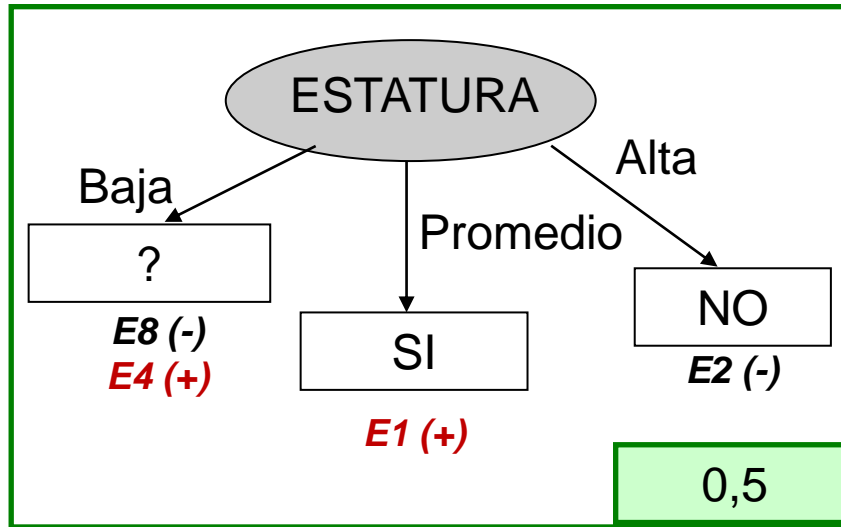


Analizar la repuesta del resto de los atributos para los ejemplos que aún no pertenecen a un subconjunto homogéneo

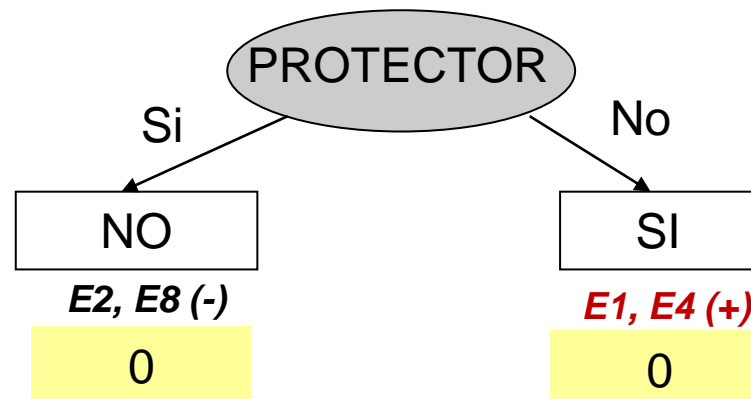
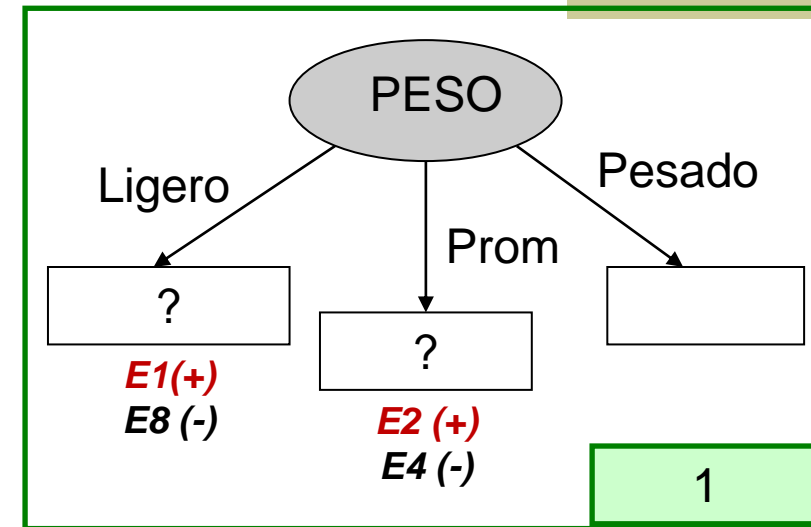
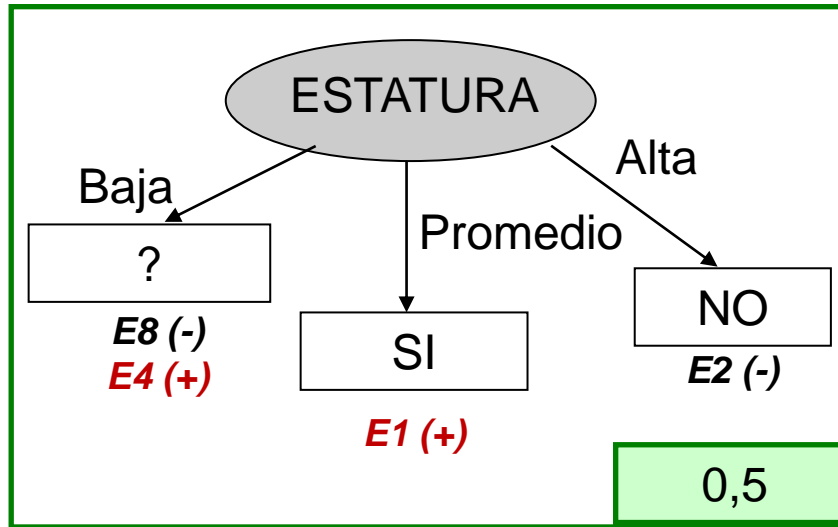
¿Qué pasaría si eligiera?



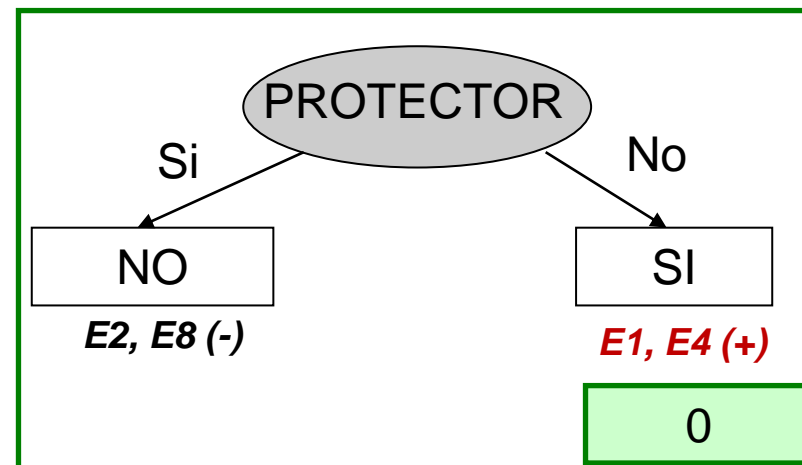
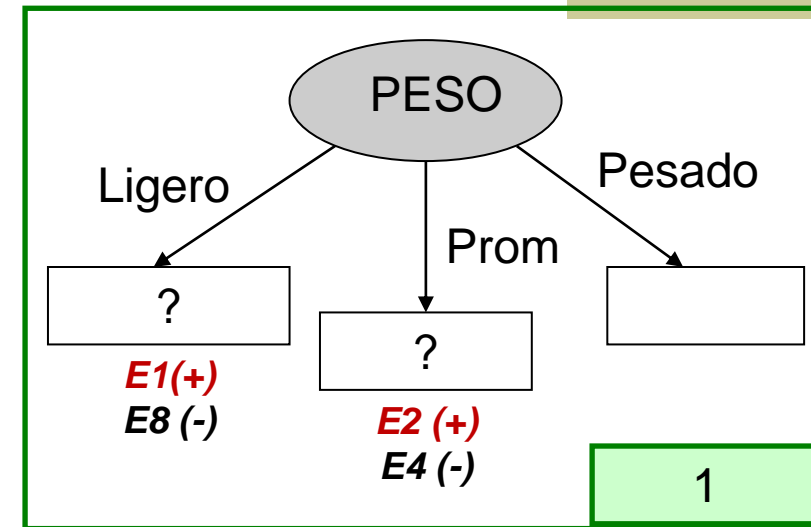
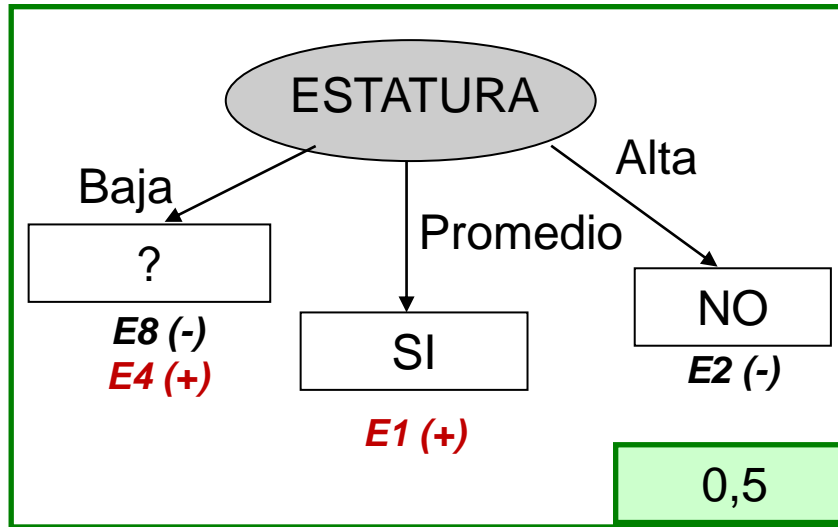
¿Qué pasaría si eligiera?



¿Qué pasaría si eligiera?



¿Qué pasaría si eligiera?



Selección por Ganancia de Información

- ***Entropía(E) = 1***

E

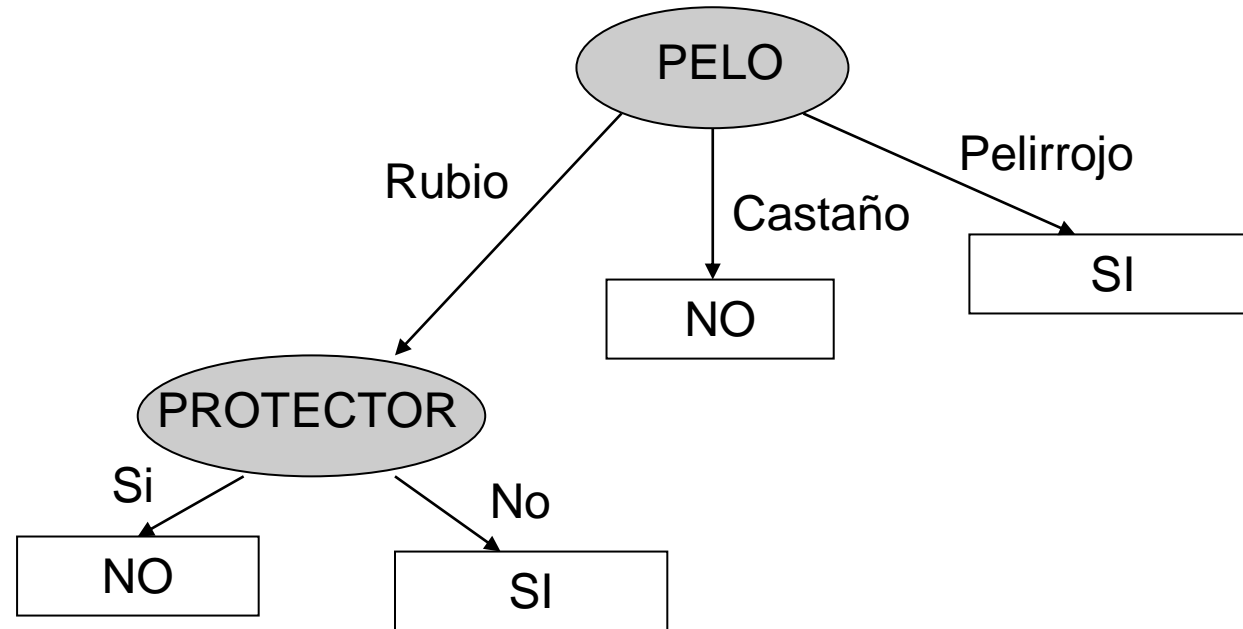
E2, E8 (-)

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

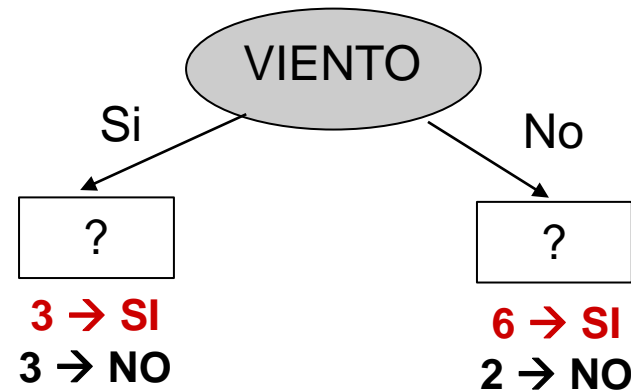
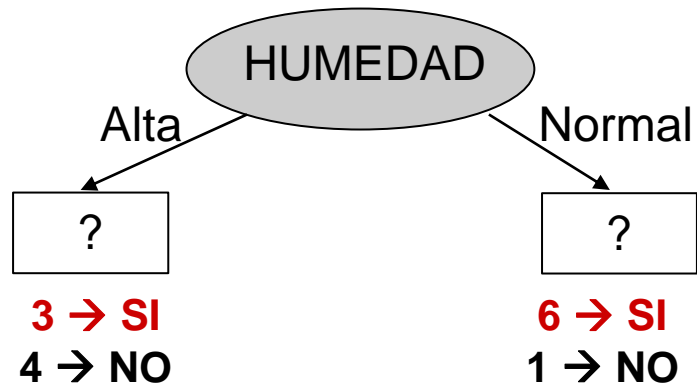
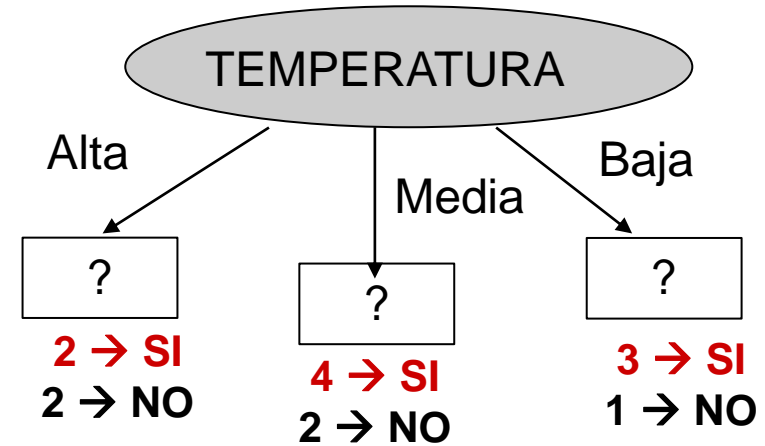
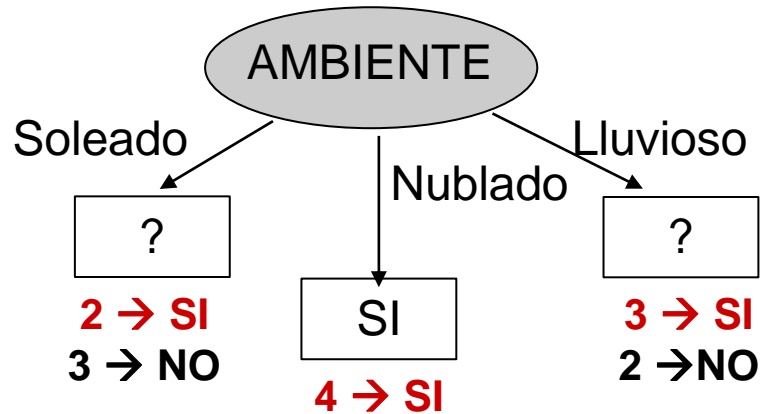
Arbol de clasificación



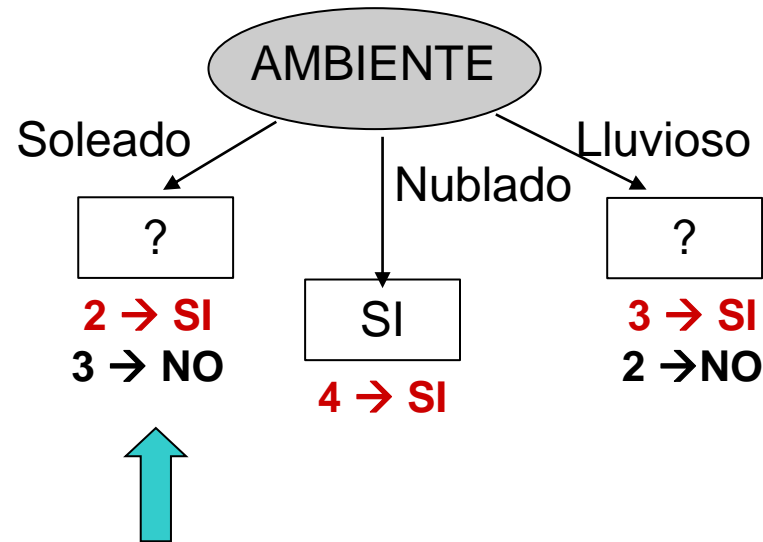
Ejemplo 2: Construir el árbol a partir de los siguientes datos

N°	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

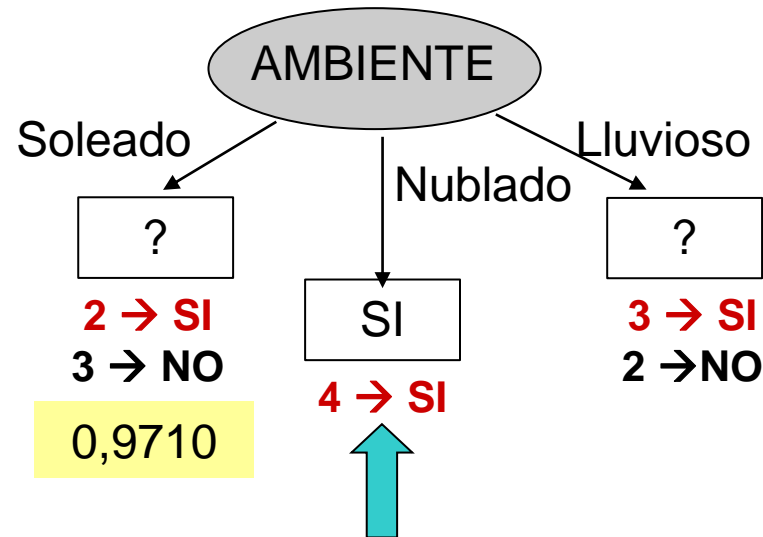


Analizando cada rama de AMBIENTE



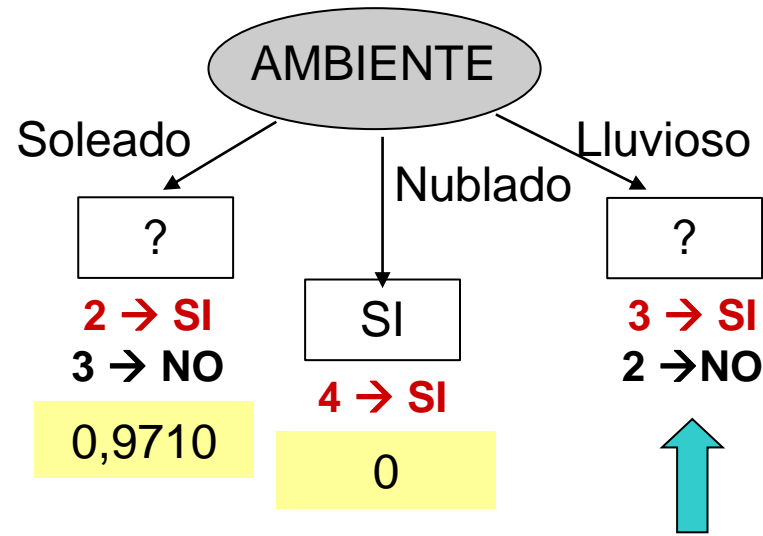
$$Entropia_{Soleado} = -\frac{2}{5} \log_2 \left(\frac{2}{5} \right) - \frac{3}{5} \log_2 \left(\frac{3}{5} \right) = 0,9710$$

Analizando cada rama de AMBIENTE



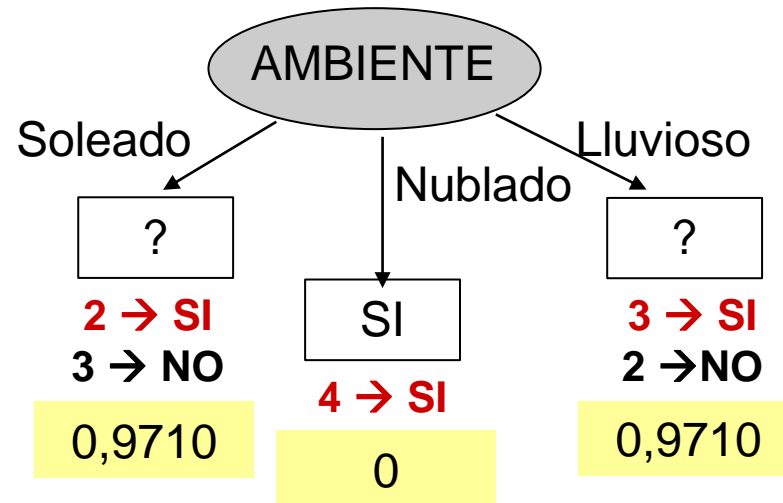
$$Entropia_{Nublado} = -\frac{4}{4} \log_2 \left(\frac{4}{4} \right) = 0$$

Analizando cada rama de AMBIENTE



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

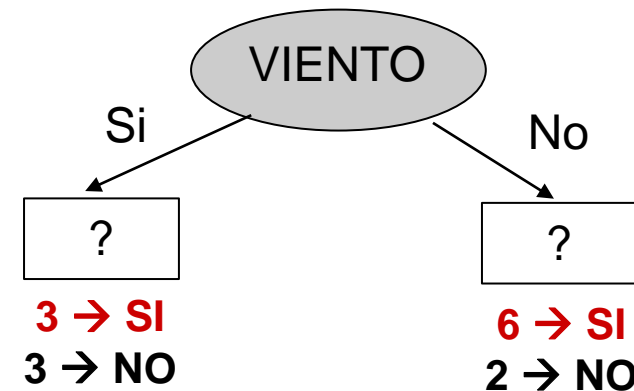
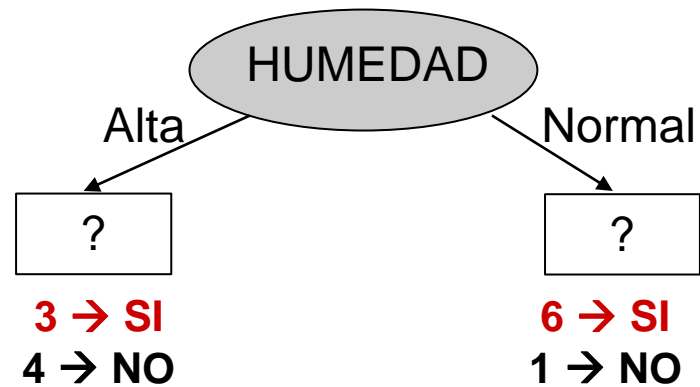
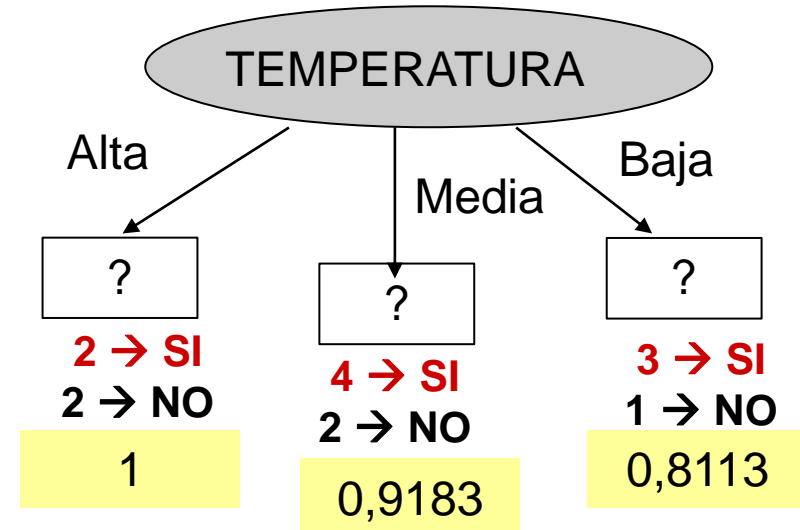
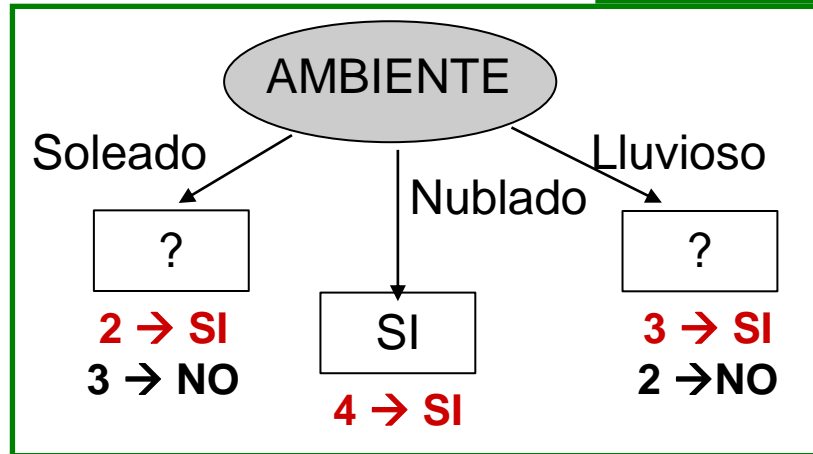
Analizando cada rama de AMBIENTE



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

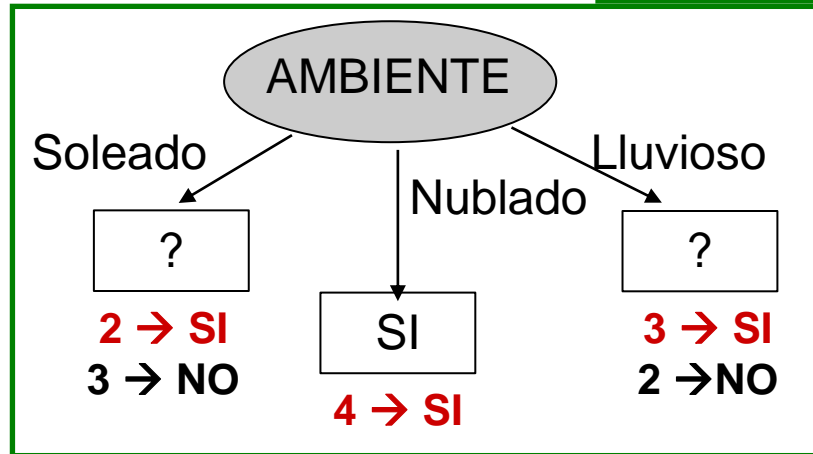
Analizando el atributo para la raíz

0,6935

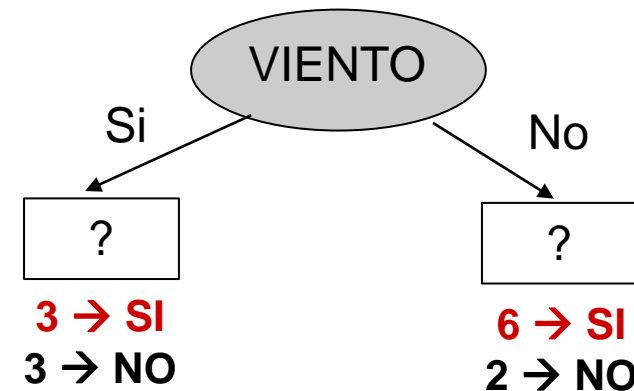
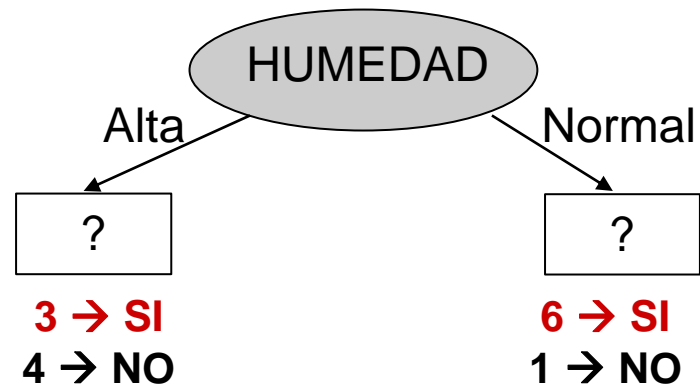
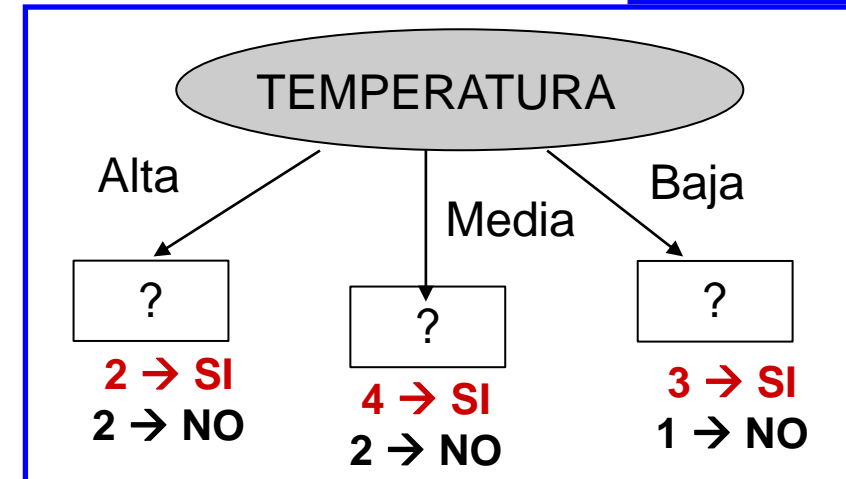


Analizando el atributo para la raíz

0,6935

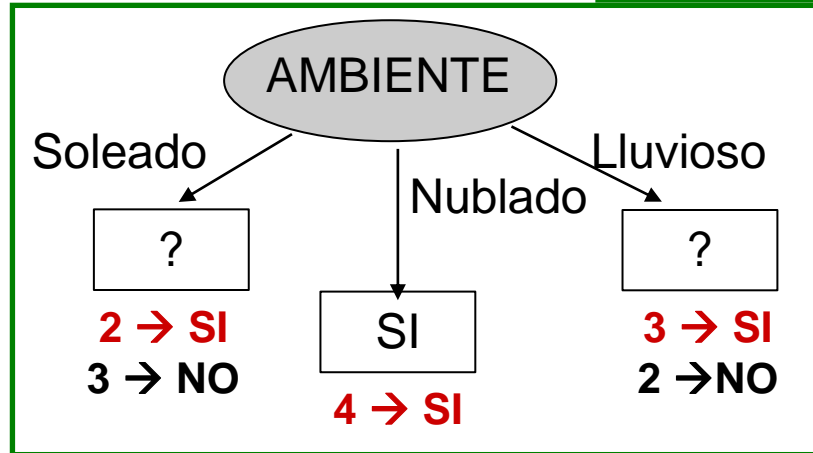


0,9164

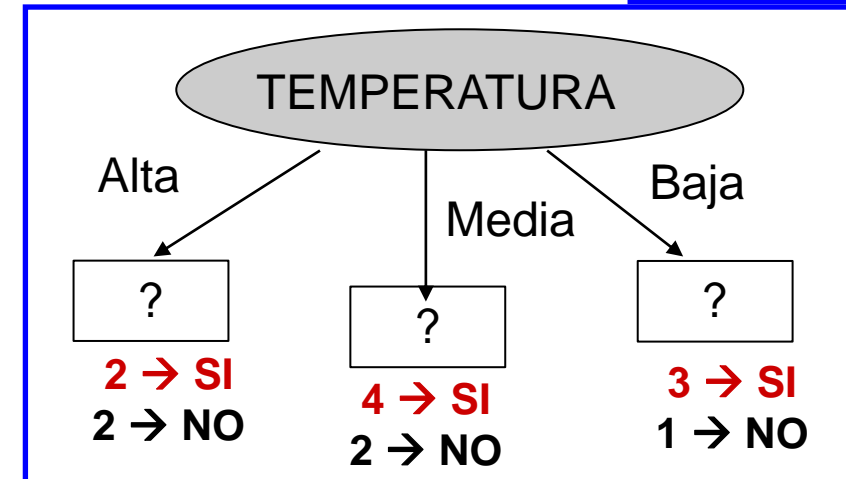


Analizando el atributo para la raíz

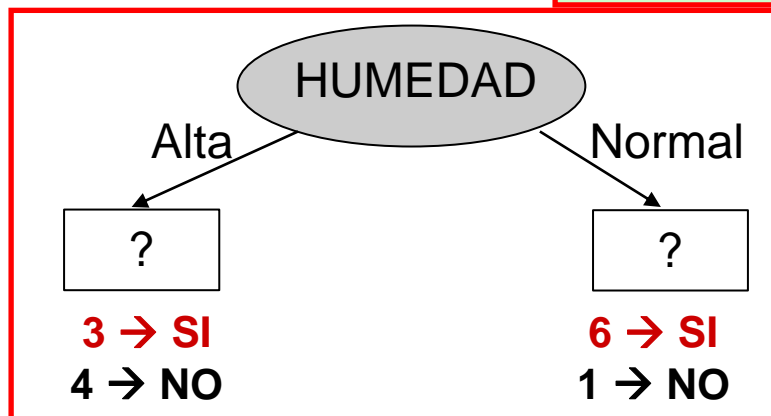
0,6935



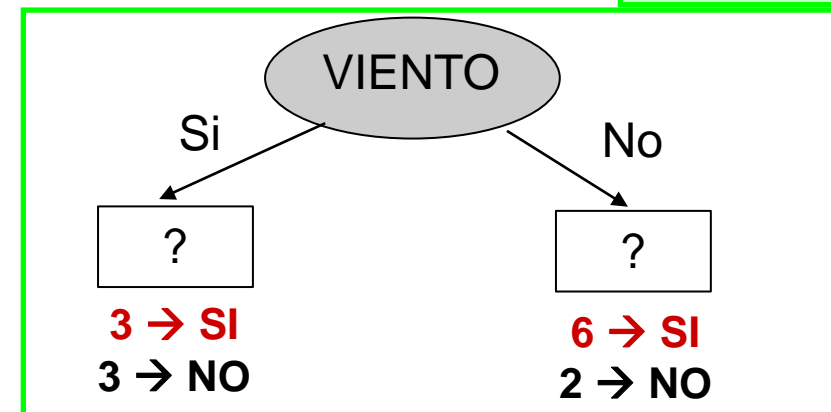
0,9164



0,7885

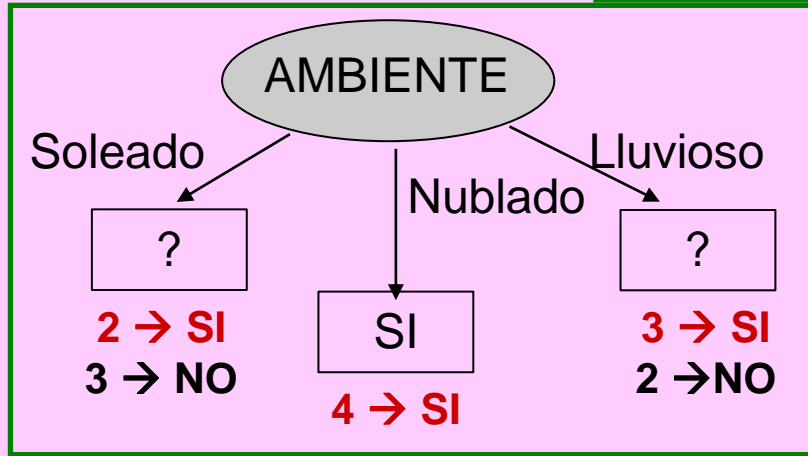


0,8922

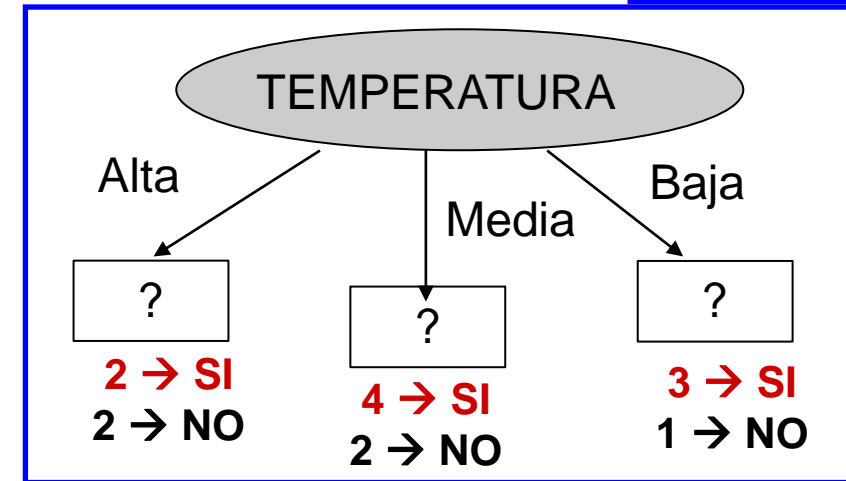


Es el seleccionado por tener el menor valor de **Entropia**,
es decir, la mayor cantidad de elementos en
subconjuntos homogéneos

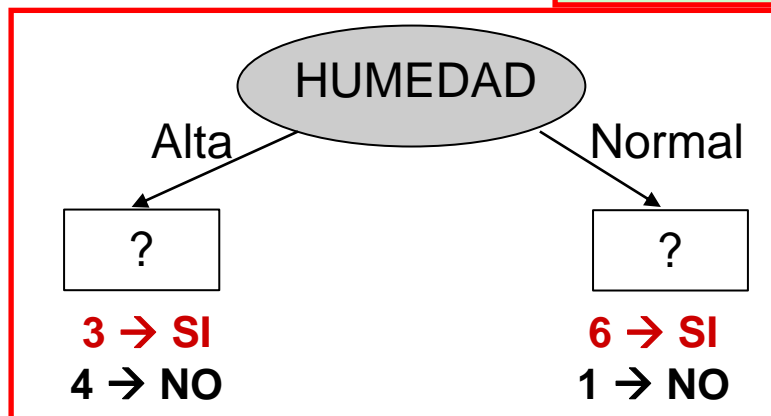
0,6935



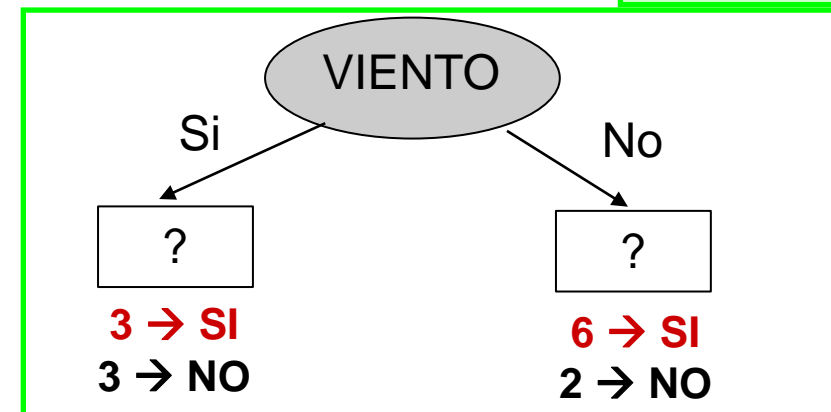
0,9164



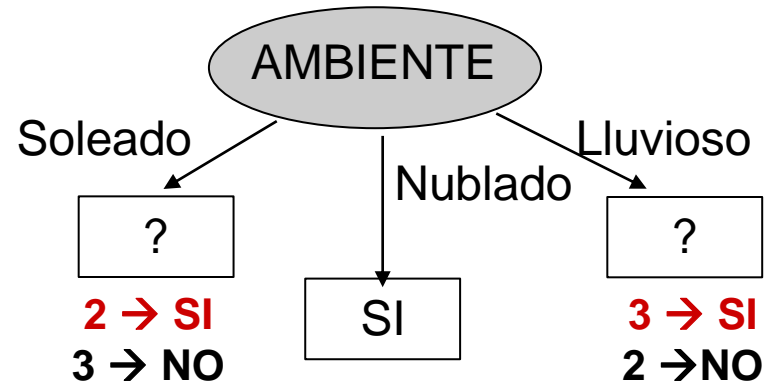
0,7885



0,8922

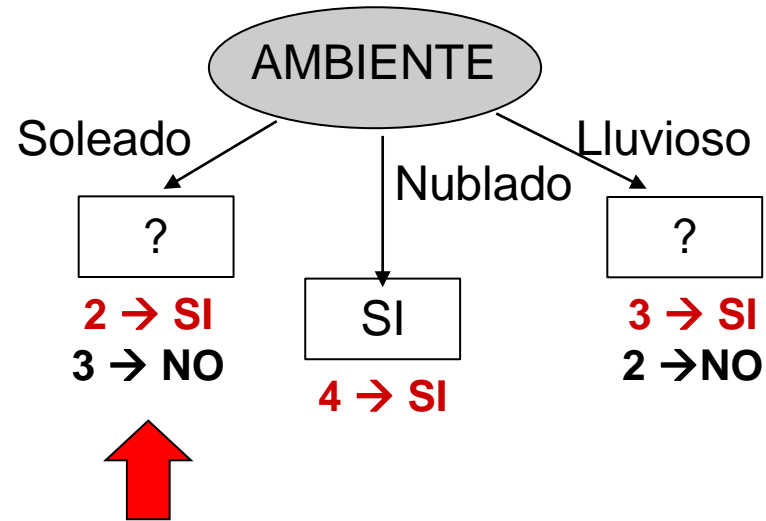


Ya tenemos el nodo raíz



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

Buscando los nodos del 1er. nivel del árbol



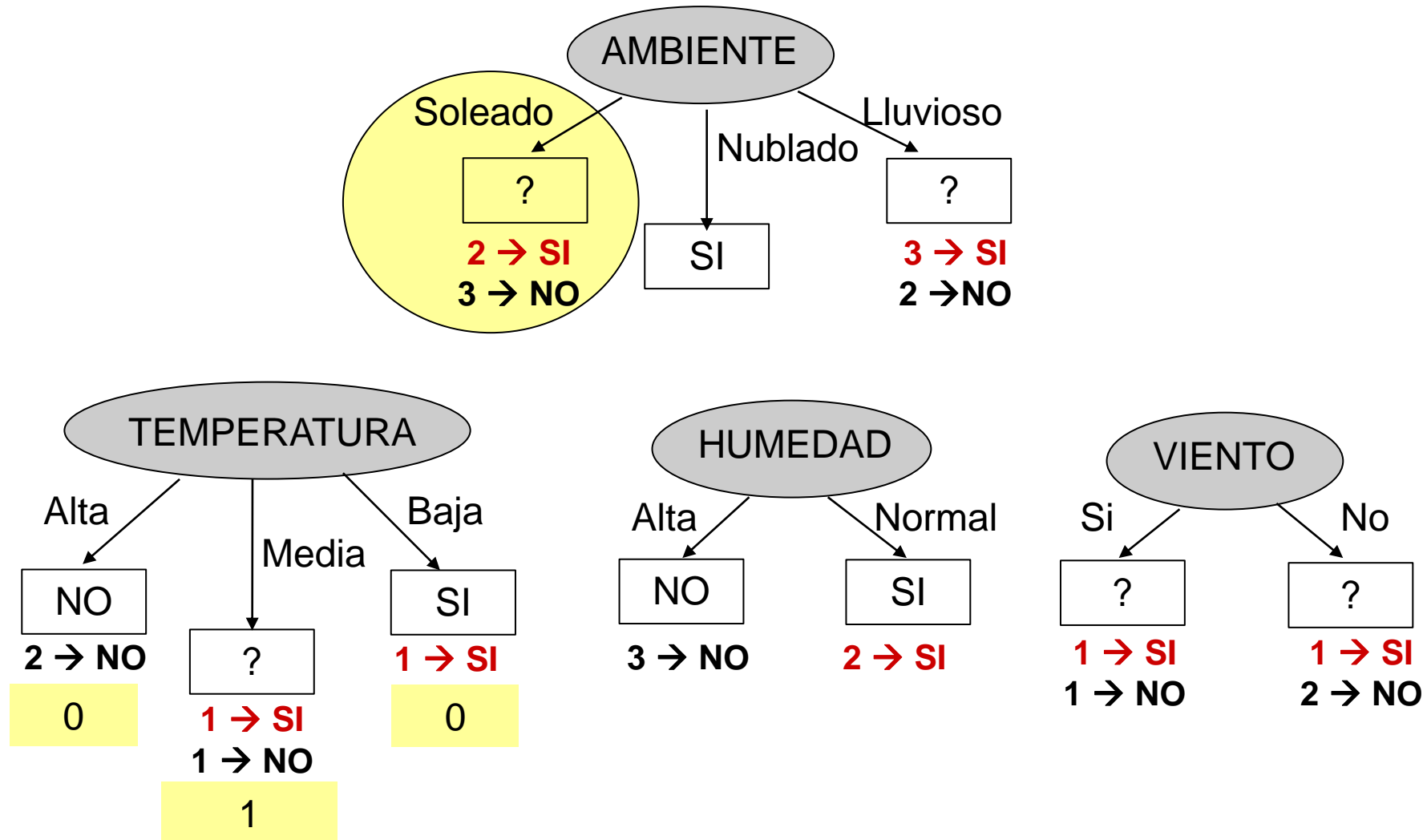
Para estas 5 muestras,
calcular el desorden de los
3 atributos restantes

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

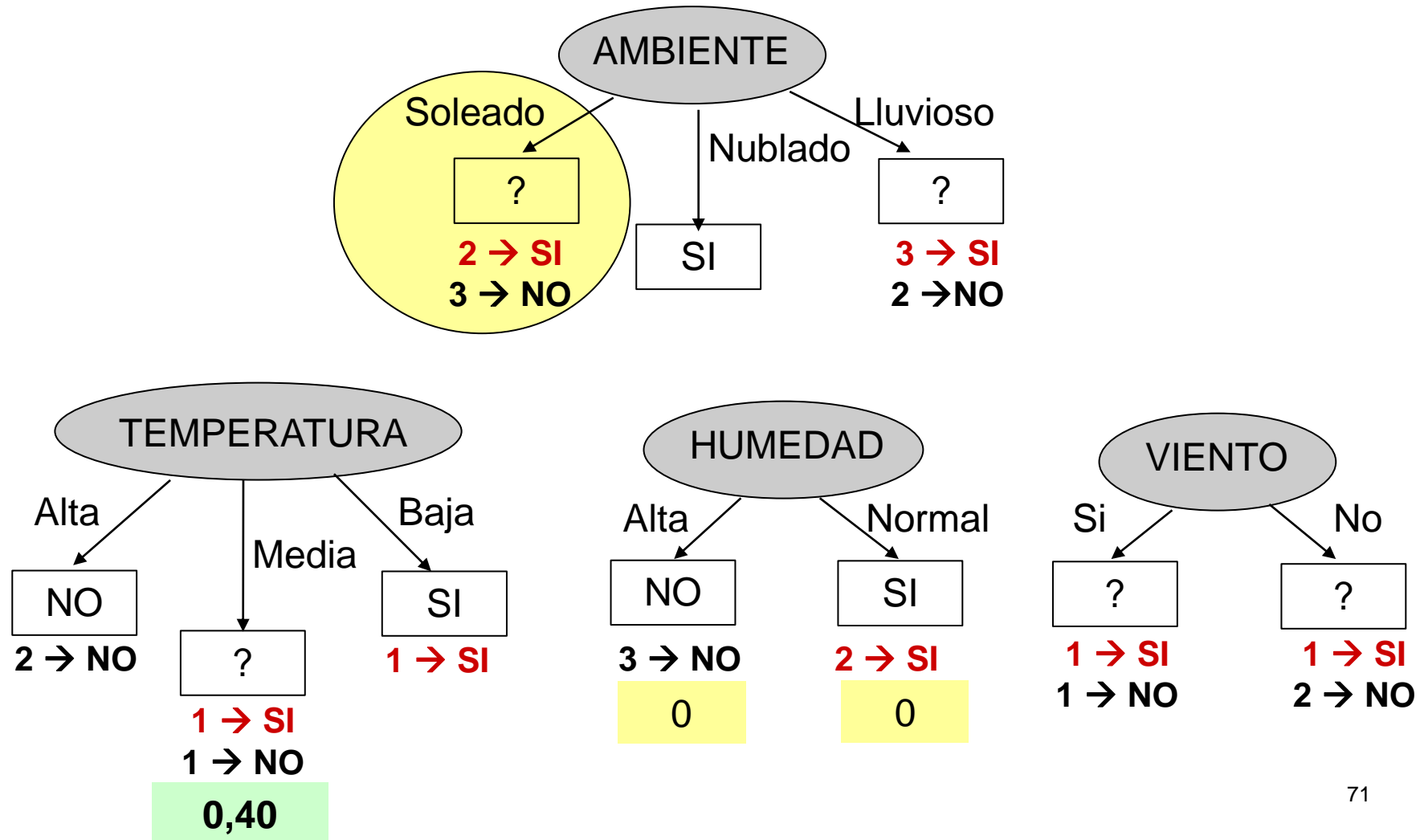
Buscando el atributo que mejor clasifica la rama

Soleado de Ambiente



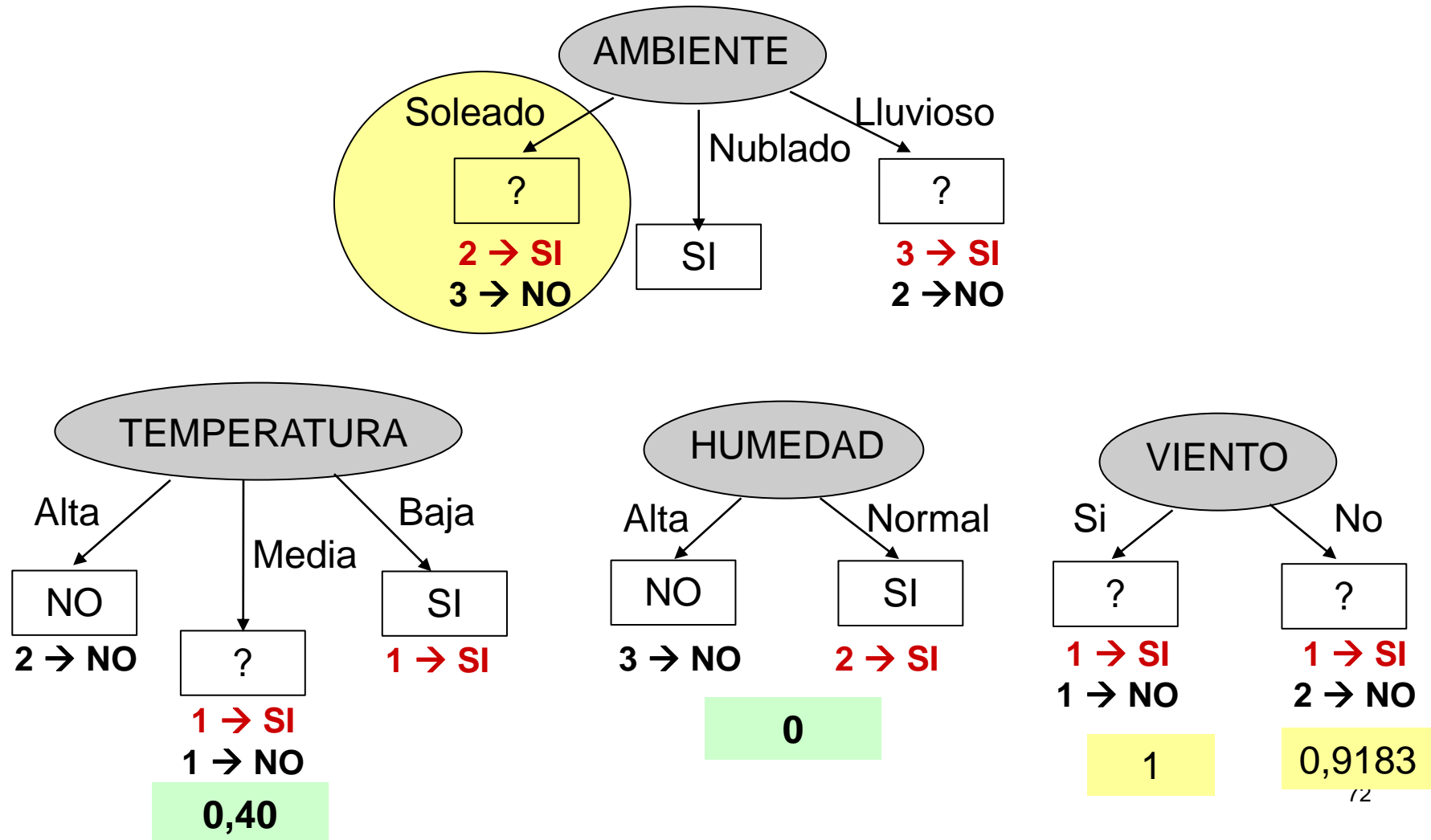
Buscando el atributo que mejor clasifica la rama

Soleado de Ambiente



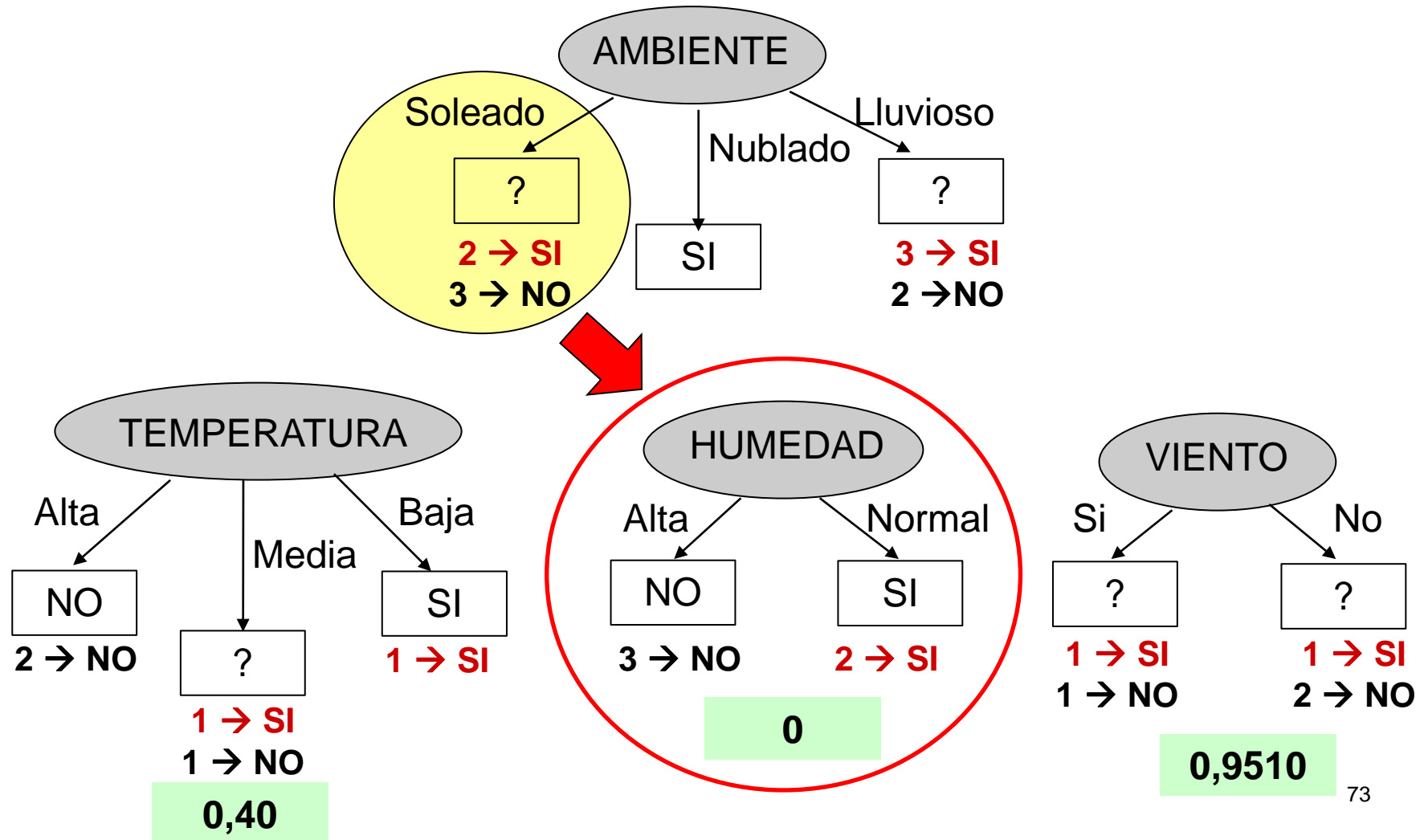
Buscando el atributo que mejor clasifica la rama

Soleado de Ambiente

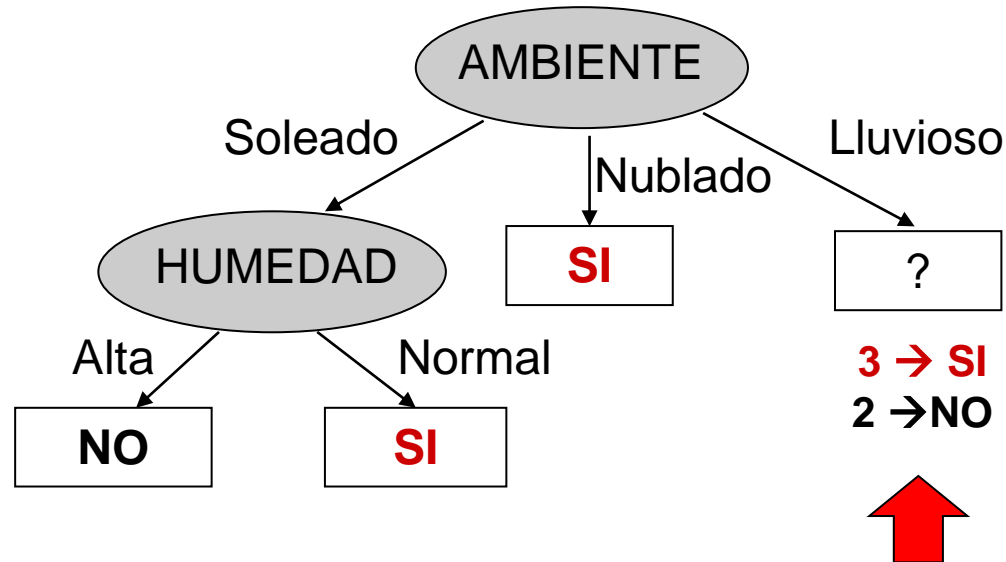


Buscando el atributo que mejor clasifica la rama

Soleado de Ambiente

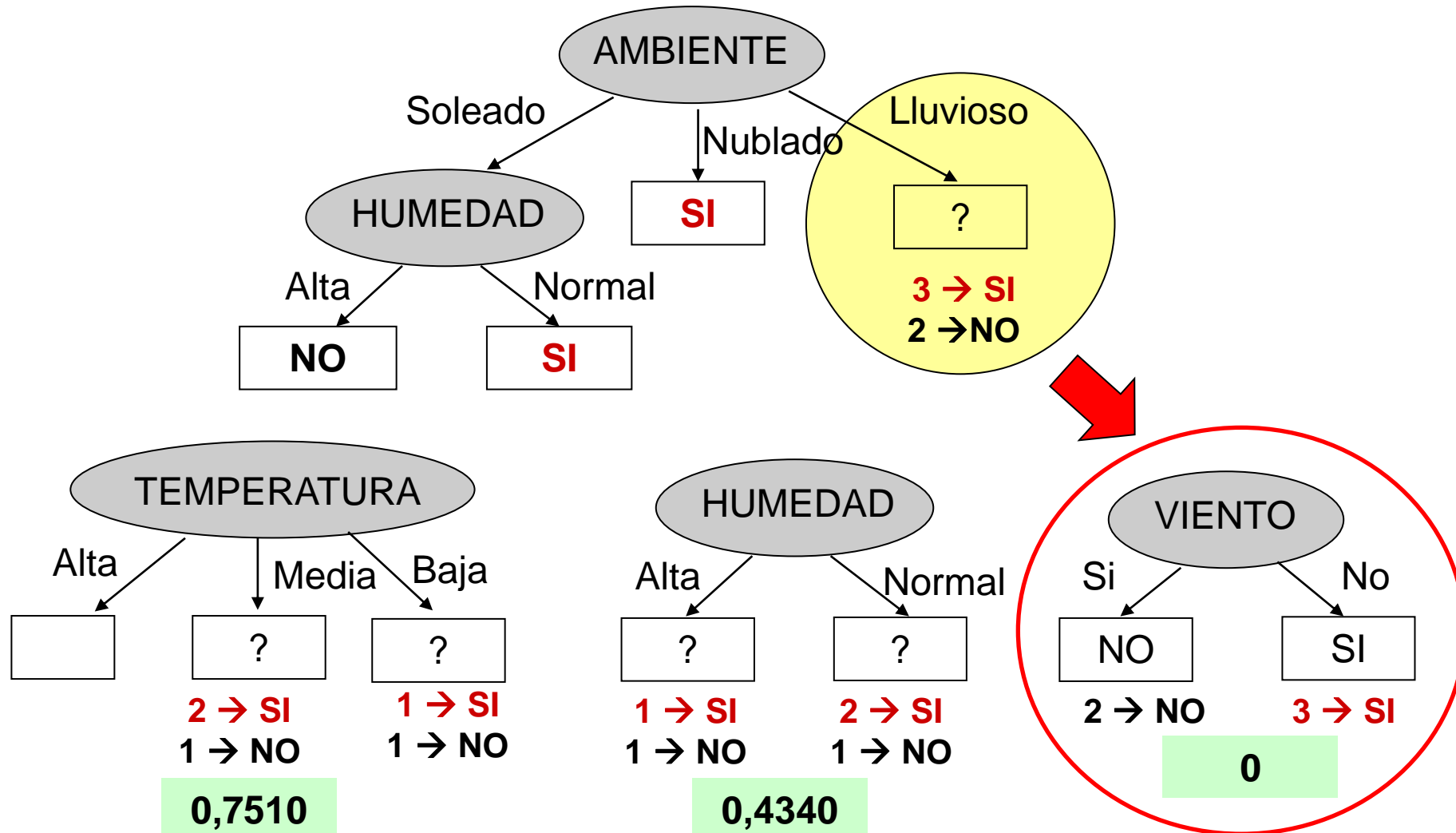


Estado actual del árbol

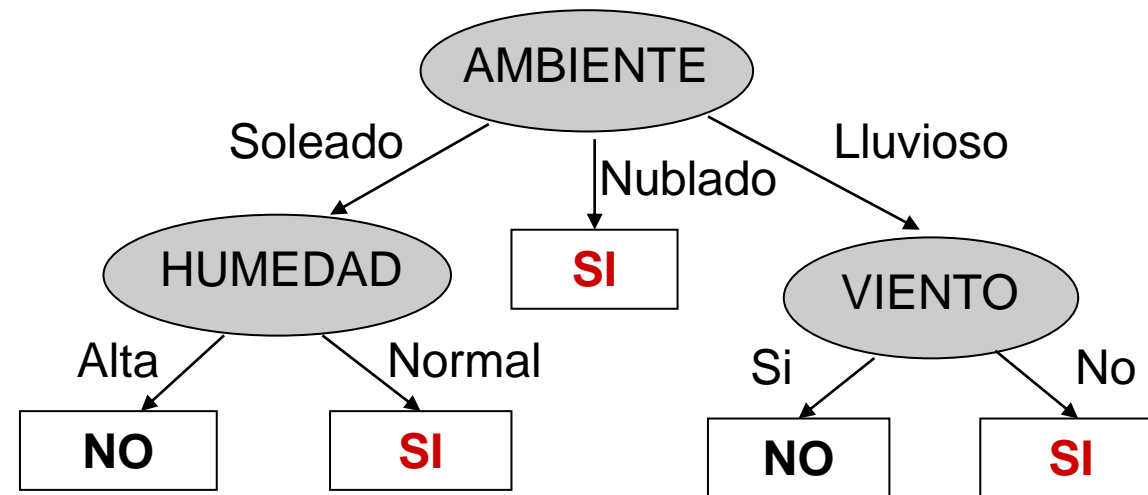


Para estas 5 muestras,
calcular el desorden de los
3 atributos restantes
(sacando AMBIENTE)


Buscando el atributo que mejor clasifica la rama *Lluvioso de Ambiente*



Arbol de clasificación



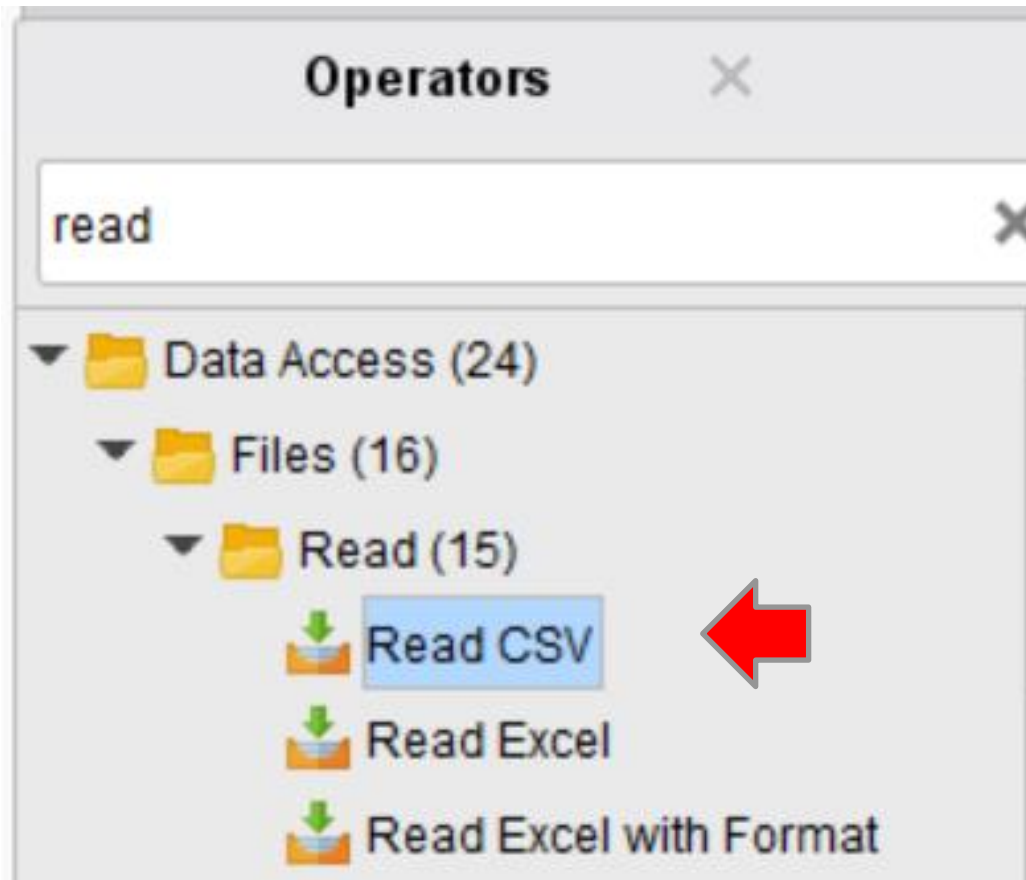
Arboles de clasificación

- Algoritmos de construcción
 - ID3 – sólo para atributos cualitativos 
 - C4.5 – opera con atributos cuantitativos y cualitativos
- Utilizaremos las métricas
 - Ganancia de Información
 - Tasa de Ganancia

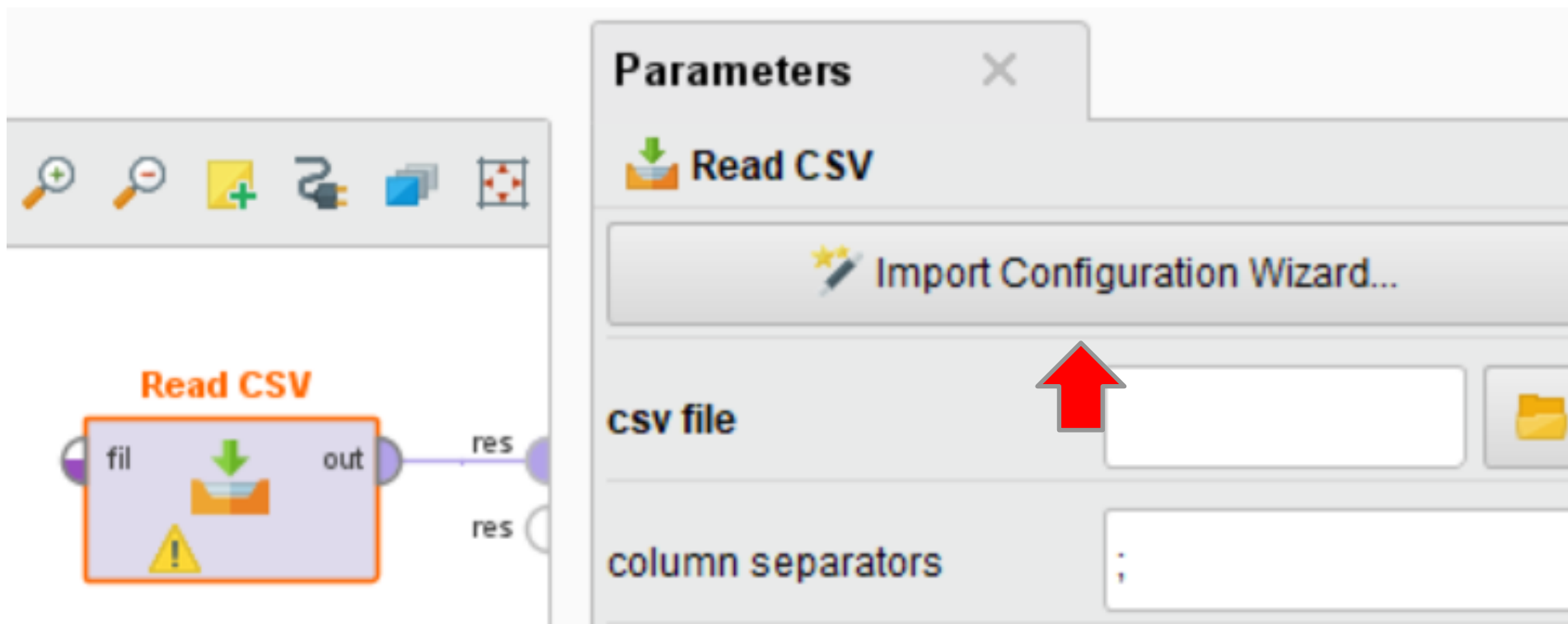
Ejemplo: Utilice Rapid Miner para construir el árbol que modelice cuando jugar al golf usando el algoritmo ID3

N°	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

Ejemplo



Ejemplo



Select the data location.

Datos

Bookmarks	File Name	Size	Type	Last Modified
★ --- Last Directory	Ataques a Redes		File Folder	Mar 5, 2019
	AlSol.csv	1 KB	Archivo de valores s...	Sep 12, 2018
	Drug5.csv	8 KB	Archivo de valores s...	Sep 27, 2018
	Drug5_numerico.csv	5 KB	Archivo de valores s...	Sep 27, 2018
	Ejemplo.csv	1 KB	Archivo de valores s...	Mar 5, 2019
	Globos.csv	1 KB	Archivo de valores s...	Mar 5, 2019
	Globos_nros.csv		valores s...	Mar 5, 2019
	Golf.csv		valores s...	Aug 24, 2018
	Golf_Numerico.csv	1 KB	Archivo de valores s...	Nov 14, 2018
	Golf_V2.csv	1 KB	Archivo de valores s...	Aug 24, 2018
	iris.csv	4 KB	Archivo de valores s...	Aug 9, 2017
	lentes.csv	1 KB	Archivo de valores s...	Feb 25, 2019
	Premios.csv	36 KB	Archivo de valores s...	Mar 1, 2019
	Sonar.csv	85 KB	Archivo de valores s...	Feb 27, 2019

Golf.csv

Golf.csv

CSV (.tsv, .csv)

← Previous

→ Next

✕ Cancel

Select the data location.

Datos

Bookmarks

★ --- Last Directory

File Name

Size

Type

Last Modified

Ataques a Redes

File Folder

Mar 5, 2019

AISol.csv

1 KB

Archivo de valores s... Sep 12, 2018

Drug5.csv

8 KB

Archivo de valores s... Sep 27, 2018

Drug5_numerico.csv

5 KB

Archivo de valores s... Sep 27, 2018

Ejemplo.csv

1 KB

Archivo de valores s... Mar 5, 2019

Globos.csv

1 KB

Archivo de valores s... Mar 5, 2019

Globos_nros.csv

1 KB

Archivo de valores s... Mar 5, 2019

Golf.csv

1 KB

Archivo de valores s... Aug 24, 2018

Golf_Numerico.csv

1 KB

Archivo de valores s... Nov 14, 2018

Golf_V2.csv

1 KB

Archivo de valores s... Aug 24, 2018

iris.csv

4 KB

Archivo de valores s... Aug 9, 2017

lentes.csv

1 KB

Archivo de valores s... Feb 25, 2019

Premios.csv

36 KB

Archivo de valores s... Mar 1, 2019

Sonar.csv

85 KB

Archivo de valores s... Feb 27, 2019

Golf.csv

Golf.csv

CSV (.tsv, .csv)

← Previous

→ Next

✕ Cancel

Specify your data format

☒ Header Row

1

Start Row

1

Column Separator

Comma ","

File Encoding

windows-1252

Escape Character

\

Decimal Character

.

☒ Use Quotes

"

☐ Trim Lines☒ Skip Comments

#

1	Ambiente	Temperatura	Humedad	Viento	Juega
2	soleado	alta	alta	no	no
3	soleado	alta	alta	si	no
4	nublado	alta	alta	no	si
5	lluvioso	media	alta	no	si
6	lluvioso	baja	normal	no	si
7	lluvioso	baja	normal	si	no
8	nublado	baja	normal	si	si
9	soleado	media	alta	no	no
10	soleado	baja	normal	no	si
11	lluvioso	media	normal	no	si



no problems.

Previous

Next



Cancel


Format your columns.


Date format

Enter value...

☐ Replace errors with missing values ⓘ

	Ambiente <i>polynomial</i>	Temperatura <i>polynomial</i>	Humedad <i>polynomial</i>	Viento <i>polynomial</i>	Juega <i>polynomial</i>
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

 no problems.

 Previous

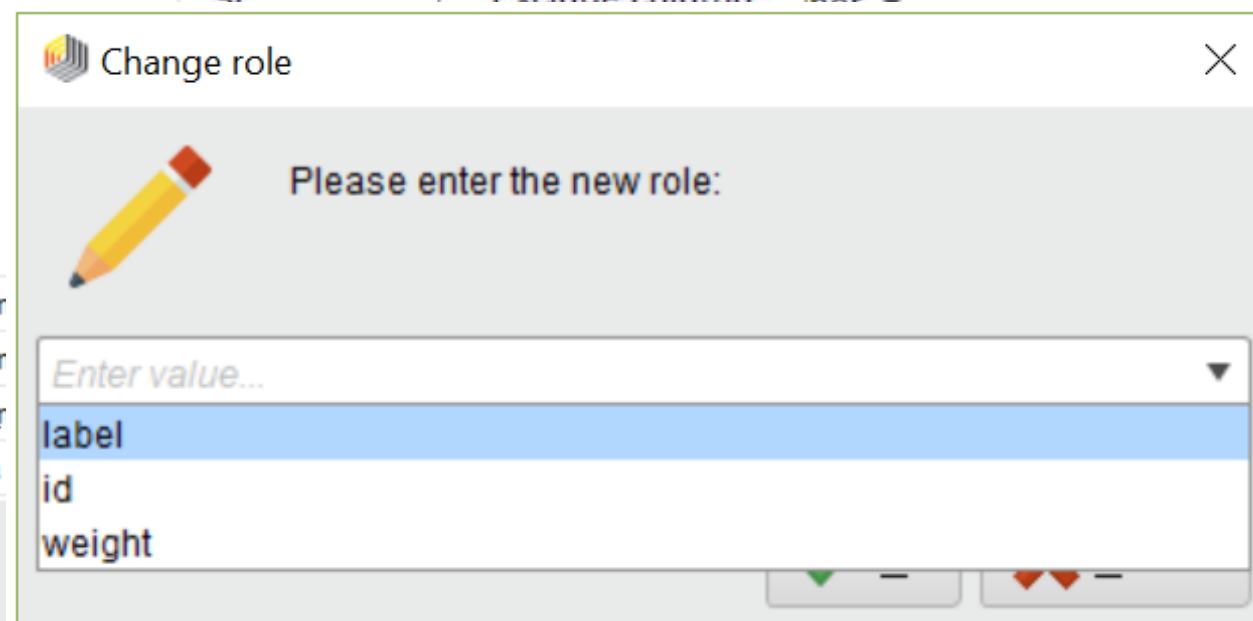
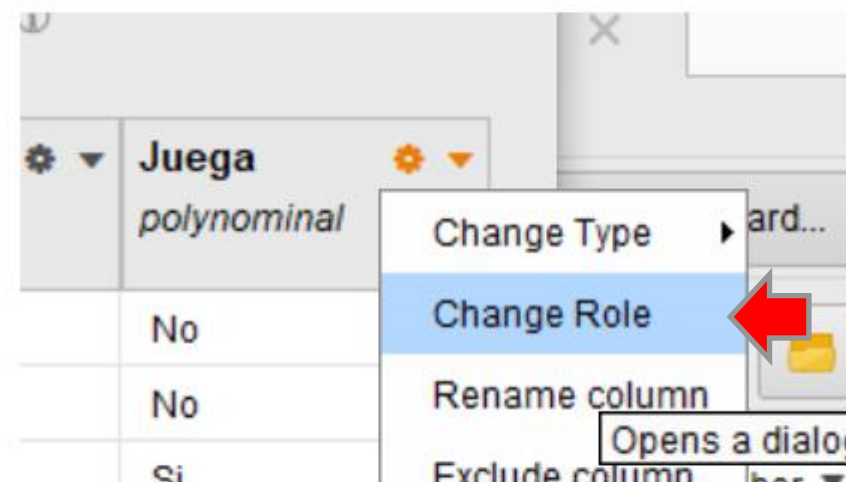
 Finish

 Cancel

Date format

	Ambiente <i>polynomial</i>	Temperatura <i>polynomial</i>
1	soleado	alta
2	soleado	alta
3	nublado	alta
4	lluvioso	media
5	lluvioso	baja
6	lluvioso	baja
7	nublado	baja
8	soleado	media
9	soleado	baja
10	lluvioso	norr
11	soleado	norr
12	nublado	alta

Marcar
como label



Format your columns.

Date format ☐ Replace errors with missing values ⓘ

	Ambiente <i>polynomial</i>	Temperatura <i>polynomial</i>	Humedad <i>polynomial</i>	Viento <i>polynomial</i>	Juega <i>polynomial label</i>
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

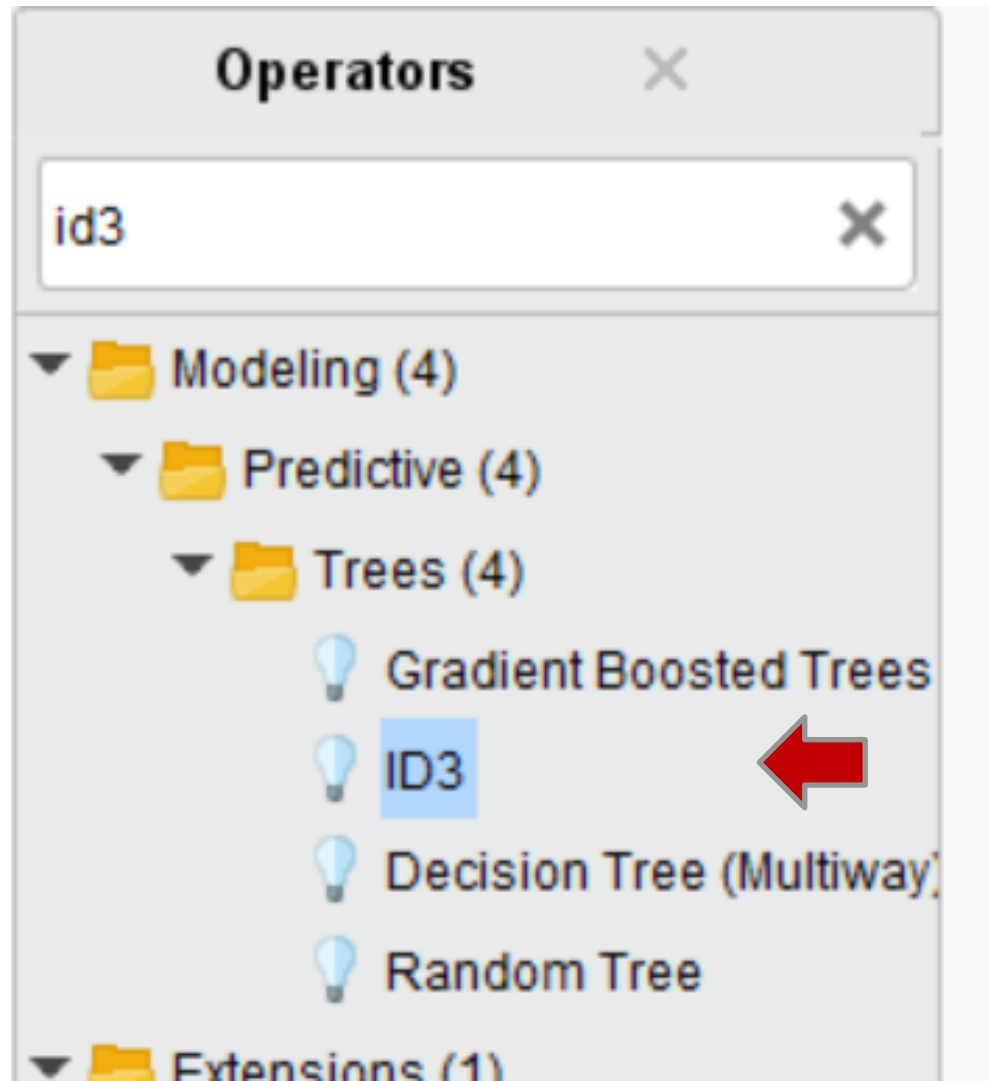
✓ no problems.

← Previous

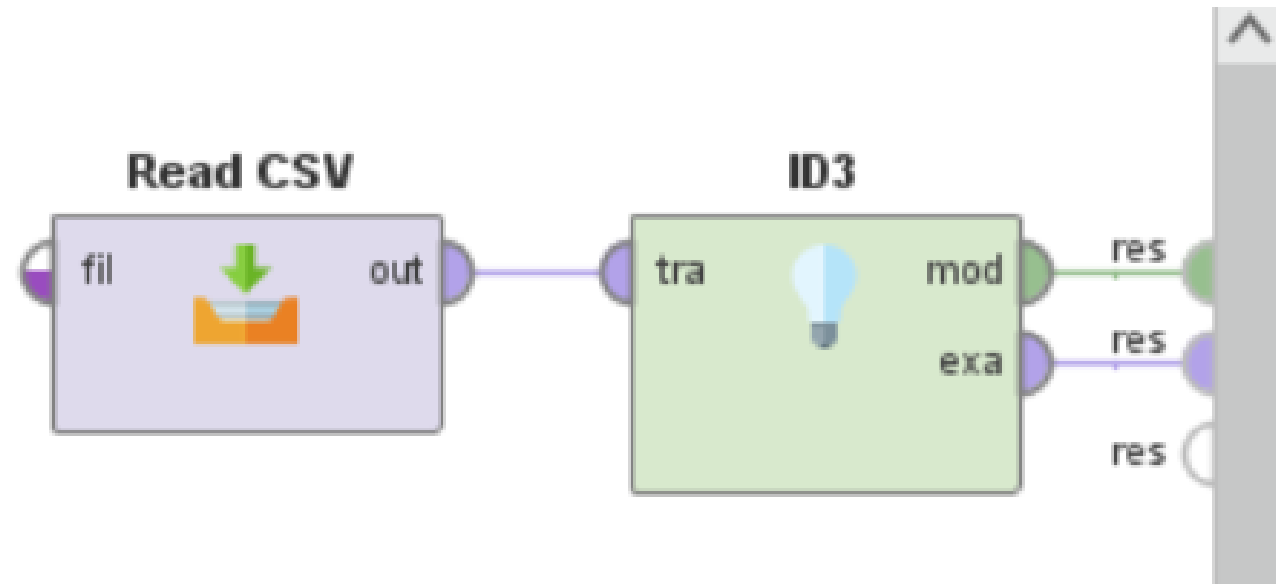
Finish

✕ Cancel

Arbol con ID3



Arbol con ID3



Ejecutar y visualizar el
árbol

Arbol obtenido con ID3





Arbol obtenido con ID3

Result History

Tree (ID3)

ExampleSet (Read E


Graph


Description

Tree

```
Ambiente = lluvioso
|   Viento = NO: Si {No=0, Si=3}
|   Viento = SI: No {No=2, Si=0}
Ambiente = nublado: Si {No=0, Si=4}
Ambiente = soleado
|   Humedad = Normal: Si {No=0, Si=2}
|   Humedad = alta: No {No=3, Si=0}
```

Características del algoritmo ID3

- Sólo opera con atributos cualitativos.
- Utiliza la **Ganancia de Información** para seleccionar los atributos a insertar en el árbol.
- Crea el árbol de decisión de tal manera que clasifica correctamente los datos de entrenamiento.
- Esto puede limitar la **capacidad de generalización** del árbol ya que los datos de entrenamiento podrían contener ruido o no ser lo suficientemente representativos.

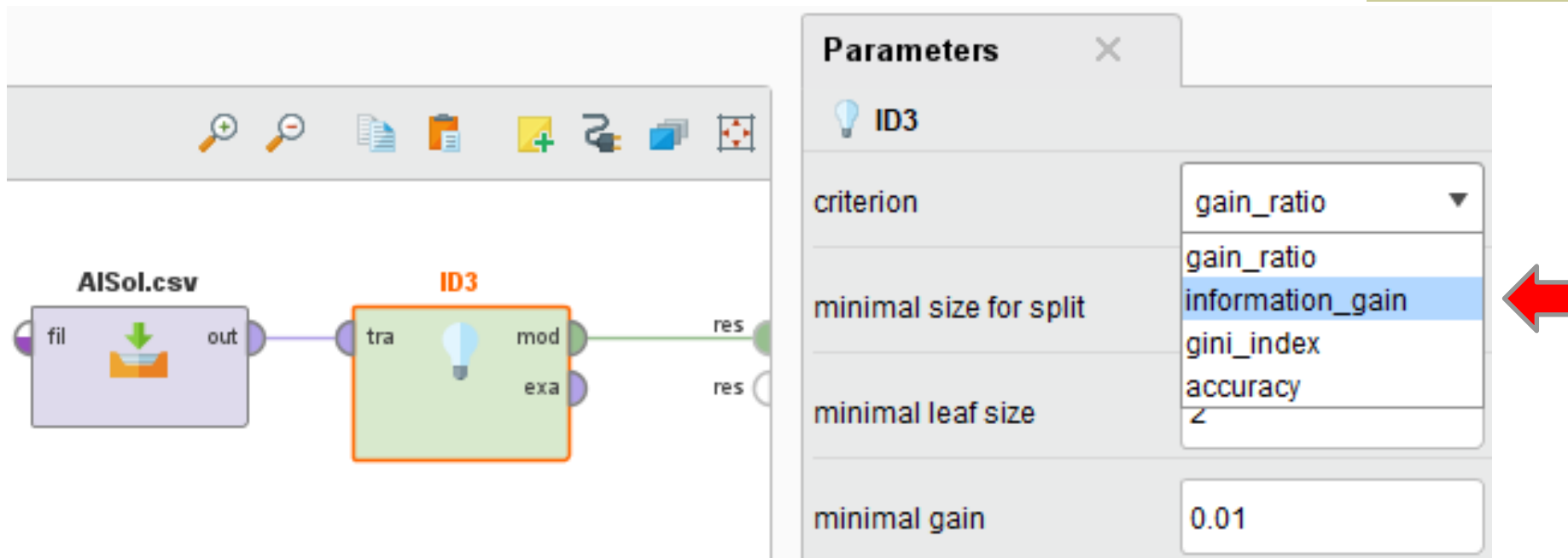
Algoritmo ID3

- Seleccionar el atributo A_i con mayor **ganancia de información**.
- Agregar el nodo al árbol.
- Para cada rama
 - Si sólo hay ejemplos de una clase C_k , etiquetarlo como C_k
 - Si no, llamar a ID3 con los ejemplos de la rama y eliminando al atributo A_i

Ejercicio: Utilice Rapid Miner para construir el árbol que modelice la información de **AlSol.csv**

Nombre	Pelo	Estatura	Peso	Protector	Resultado
Sara	Rubio	Promedio	Ligero	No	Quemado
Diana	Rubio	Alta	Promedio	Si	Ninguno
Alexis	Castaño	Baja	Promedio	Si	Ninguno
Ana	Rubio	Baja	Promedio	No	Quemado
Emilia	Pelirrojo	Promedio	Pesado	No	Quemado
Pedro	Castaño	Alta	Pesado	No	Ninguno
Juan	Castaño	Promedio	Pesado	No	Ninguno
Catalina	Rubio	Baja	Ligero	Si	Ninguno

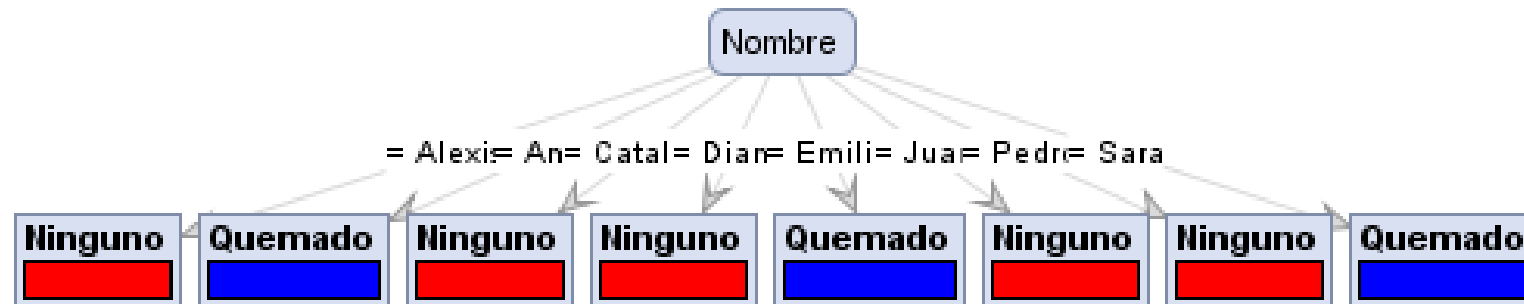
Ejercicio



- Note que se obtienen árboles diferentes según el criterio utilizado

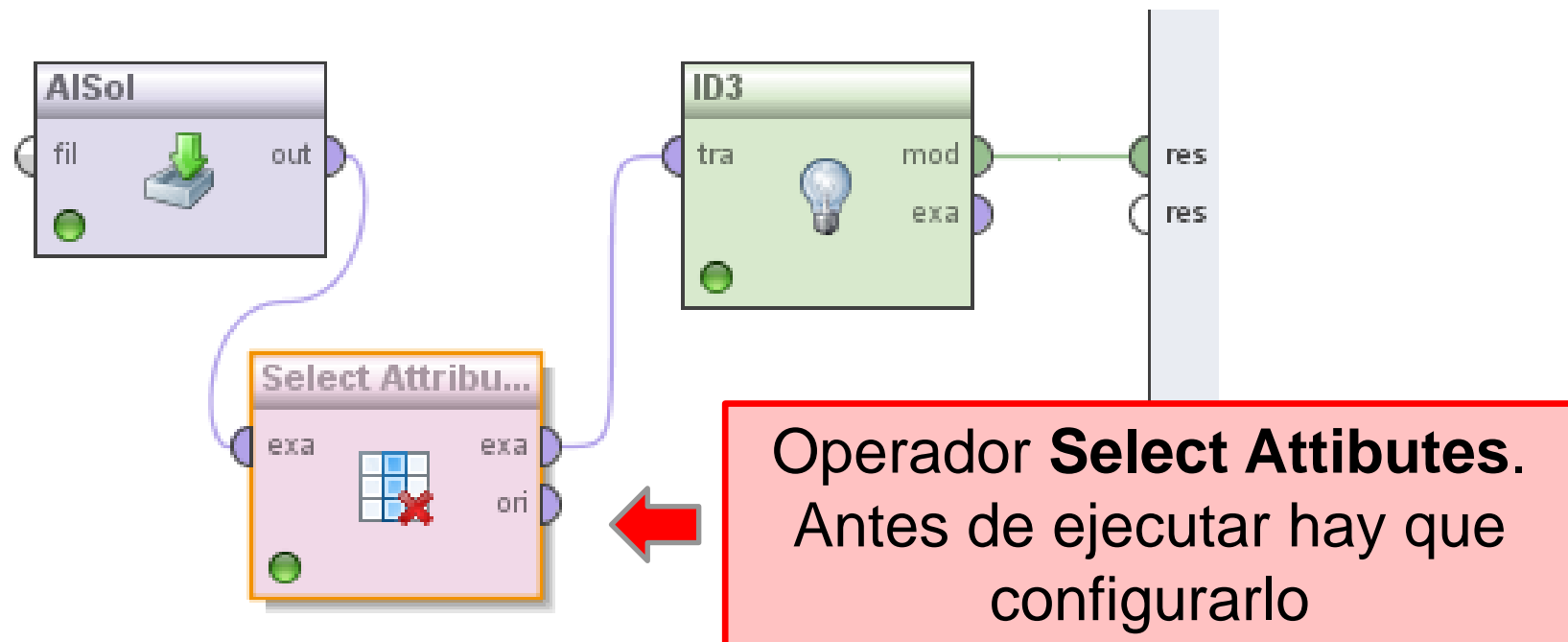
Ejercicio

- Utilizando como métrica (parámetro ***criterion***) la Ganancia de Información (opción **information_gain**)



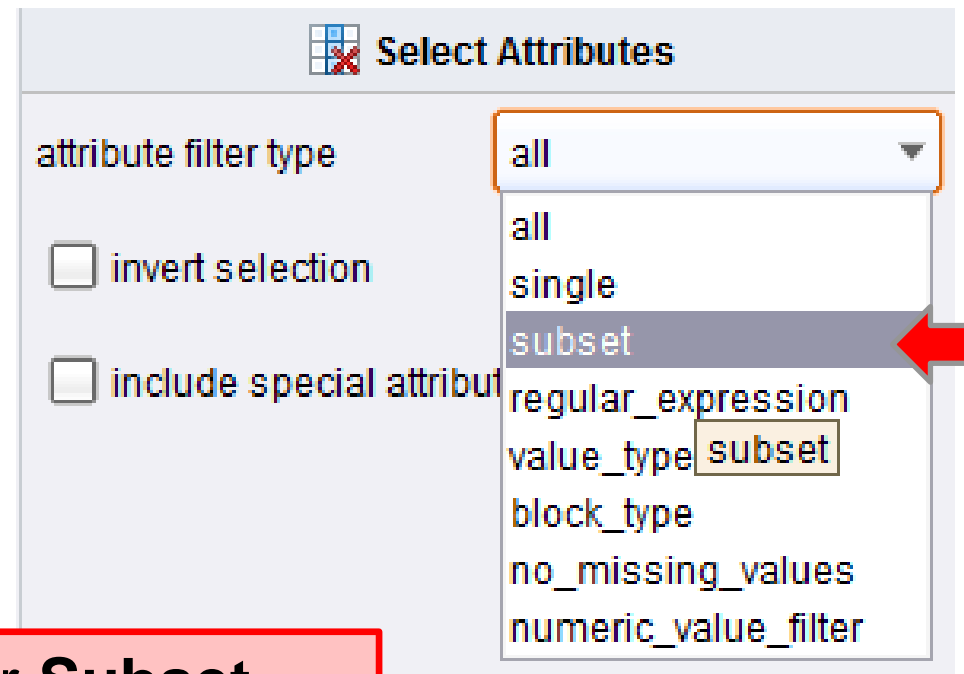
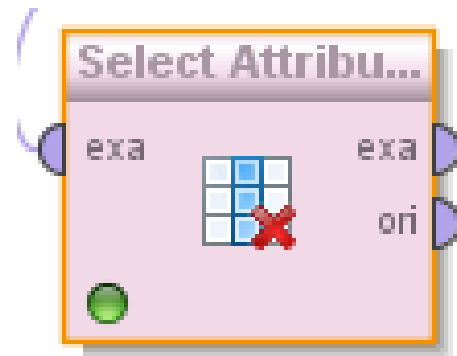
Ejercicio 1

- Utilizando como métrica la Ganancia de Información (opción **information_gain**) pero **sin considerar el atributo Nombre**



Ejercicio

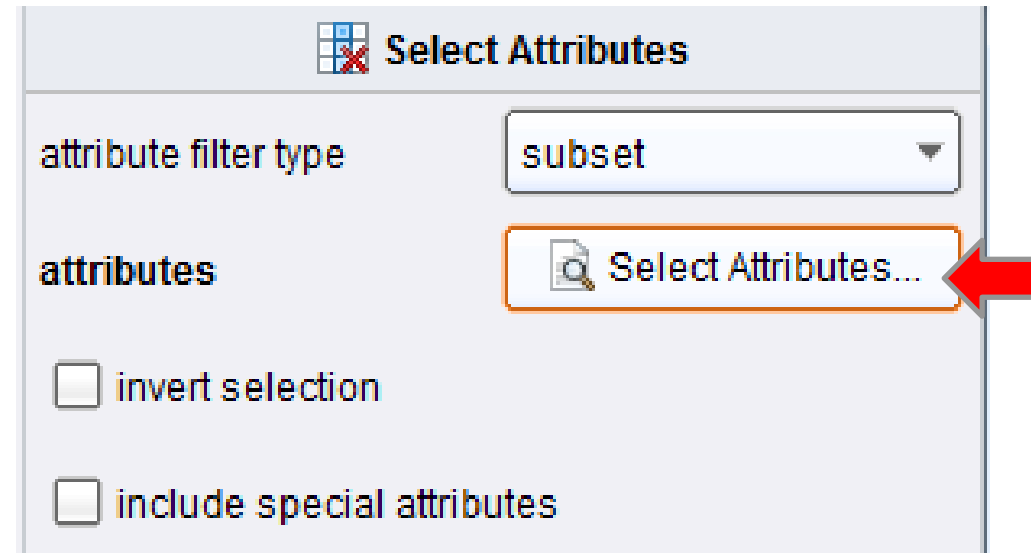
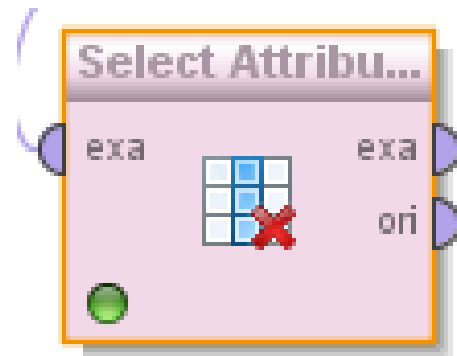
- Operador **Select Attributes** para descartar el atributo **Nombre**



Elegir **Subset**

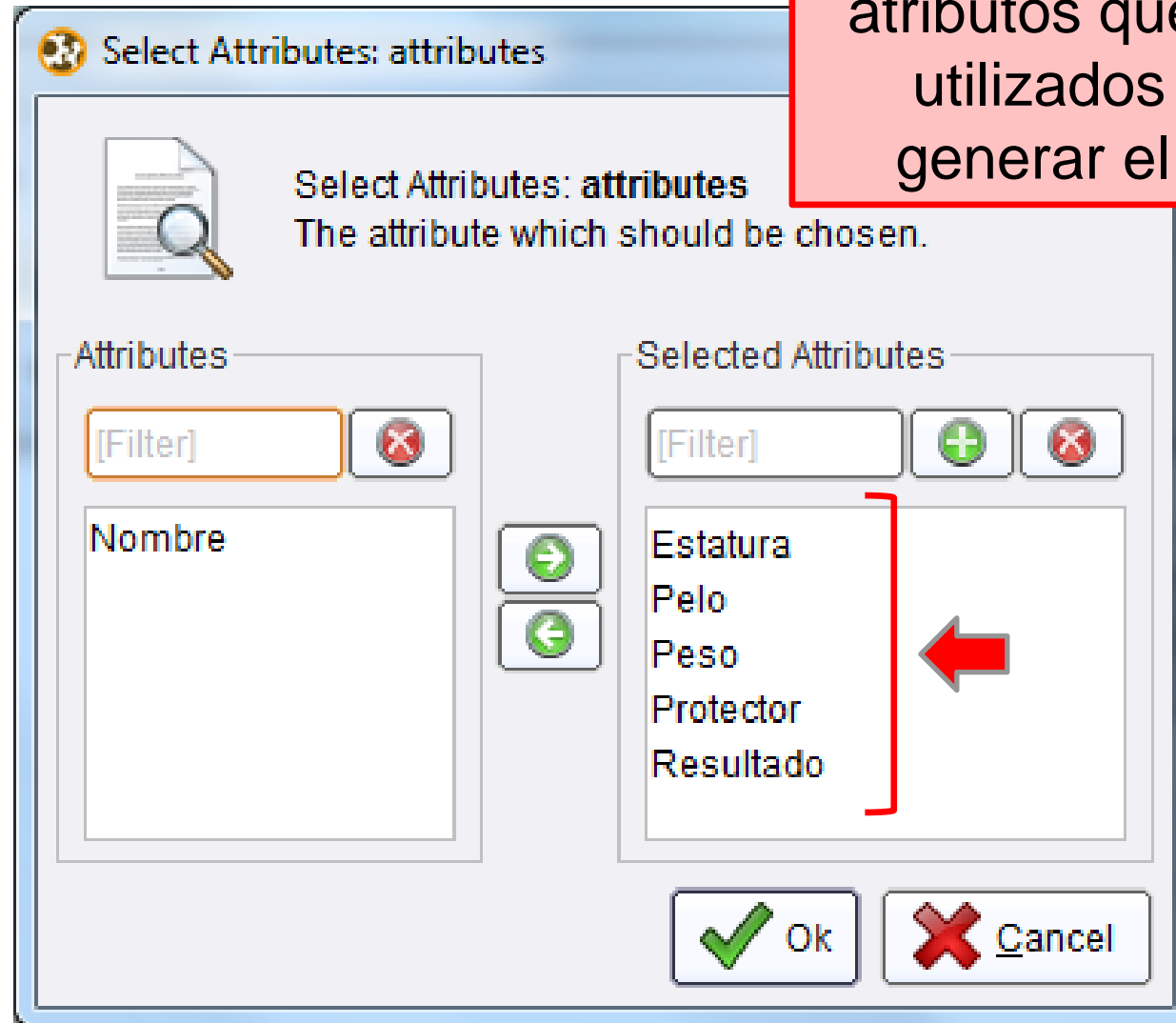
Ejercicio

- Operador **Select Attributes** para descartar el atributo **Nombre**



Ingresa aquí para indicar los atributos que deben considerarse

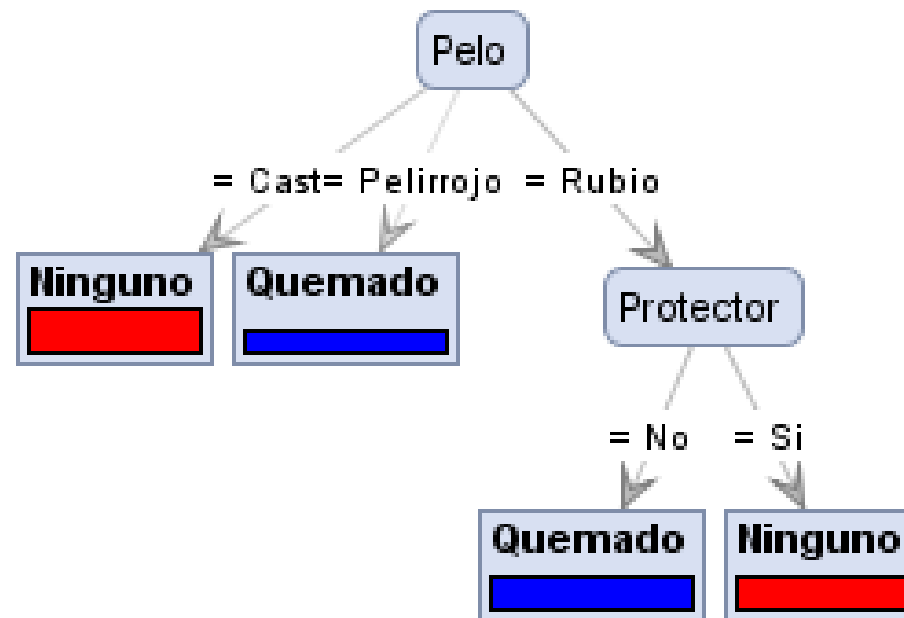
Ejercicio



Estos son los atributos que serán utilizados para generar el árbol

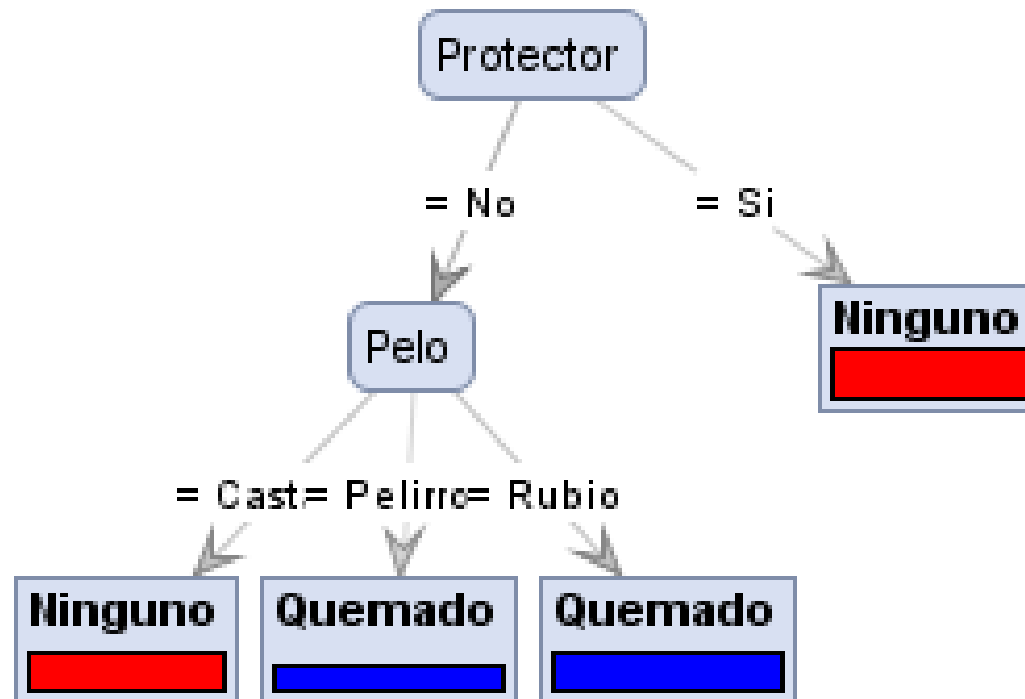
Ejercicio

- Utilizando como métrica la Ganancia de Información (opción **information_gain**) pero sin considerar el atributo **Nombre**



Ejercicio

- Utilizando como métrica (parámetro ***criterion***) la Tasa de Ganancia (opción **gain_ratio**)



Tasa de Ganancia (Gain Ratio)

- La Ganancia de Información (**Information Gain**) favorece a los atributos con muchos valores.

Ej: atributo NOMBRE en AISol.csv

- La Tasa de Ganancia compensa el hecho de que un atributo pueda tener muchos valores dividiendo la Ganancia de Información por la medida denominada **información de la división**.

$$\text{Gain_Ratio}(E, \text{Atrib}) = \frac{\text{Ganancia}(E, \text{Atrib})}{\text{Info_Division}(E, \text{Atrib})}$$

*Note que la **tasa de ganancia** se calcula para un atributo no para el conjunto de ejemplos completo*

E

Ganancia de Información de PELO

- Dado el conjunto E, la ganancia de información del atributo PELO será

Sara
Diana
Ana
Catalina
Alexis
Pedro
Juan
Emilia

$$Ganancia(E, Pelo) = \underbrace{Entropia(E)}_{\text{Cantidad de información requerida para dar una respuesta en base los ejemplos del conjunto E}} - \underbrace{Entropia(E, PELO)}_{\text{Cantidad de información requerida para dar una respuesta en base los ejemplos del conjunto E divididos por el atributo PELO}}$$

Cantidad de **información** requerida para dar una respuesta en base los ejemplos del conjunto E
o
Cantidad de **incertidumbre** que se tiene al momento de dar una respuesta

Cantidad de **información** requerida para dar una respuesta en base los ejemplos del conjunto E divididos por el atributo PELO

Entropía del conjunto E

- Inicialmente, la cantidad de información requerida para dar una respuesta es

$$Entropia(E) = -\frac{5}{8} \log_2 \left(\frac{5}{8} \right) - \frac{3}{8} \log_2 \left(\frac{3}{8} \right) = 0.9544$$

Sara

Diana

Ana

Catalina

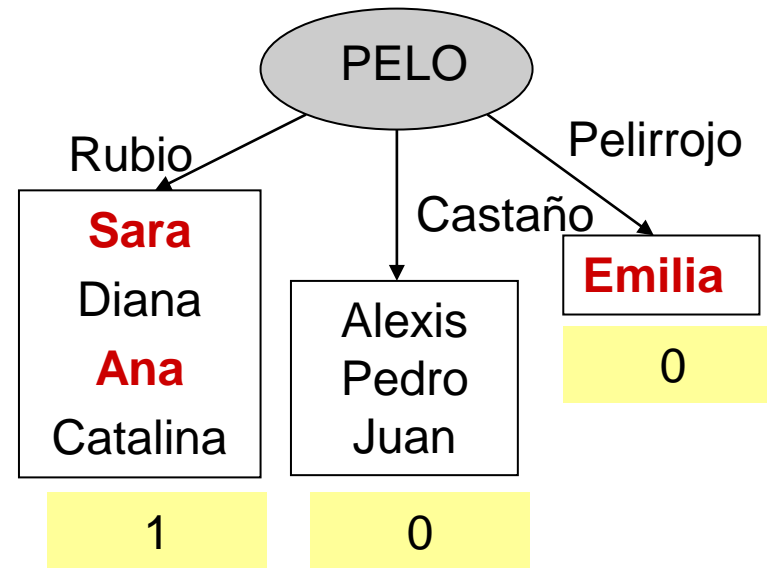
Alexis

Pedro

Juan

Emilia

Entropía del atributo PELO



Cantidad de información requerida por PELO para responder

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

Ganancia de Información de PELO

- Luego, la ganancia de información si se elige PELO será

$$\begin{aligned} \text{Ganancia}(E, \text{Pelo}) &= \text{Entropia}(E) - \text{Entropia}(E, \text{PELO}) \\ &= 0.9544 - 0.5 = 0.4544 \end{aligned}$$

Sara

Diana

Ana

Catalina

Alexis

Pedro

Juan

Emilia

Información de la división

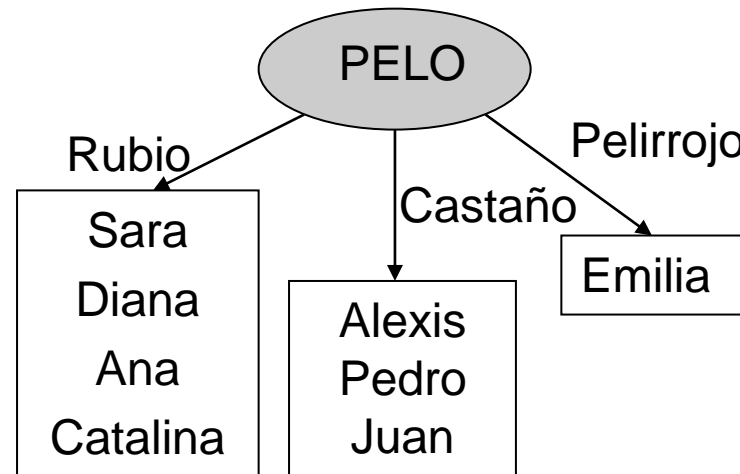
- La información de la división se calcula así

$$Info_Division(E, A) = - \sum_{v_i}^{v_n} \frac{|E_i|}{|E|} \log_2 \frac{|E_i|}{|E|}$$

- Siendo E_i, E_{i+1}, \dots, E_n las diferentes particiones que resultan de dividir el conjunto E de ejemplos teniendo en cuenta los valores $v_i \dots v_n$ que toma el atributo.

Info_Division(E, Pelo)

- Dentro de las ramas del atributo no se distinguen las clases



$$Info_Division(E, Pelo) = -\frac{4}{8} \log_2 \left(\frac{4}{8} \right) - \frac{3}{8} \log_2 \left(\frac{3}{8} \right) - \frac{1}{8} \log_2 \left(\frac{1}{8} \right)$$

$$Info_Division(E, Pelo) = 1.4056$$

Tasa de Ganancia de PELO

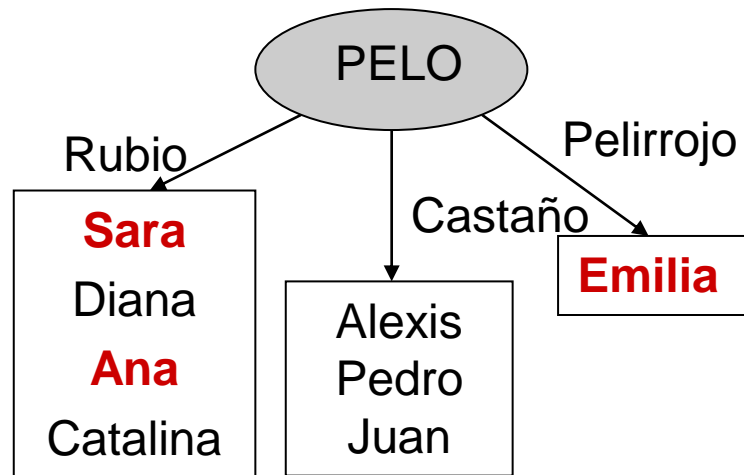
$$Gain_Ratio(E, P_{elo}) = \frac{Ganancia(E, P_{elo})}{Info_Division(E, P_{elo})}$$

$$Gain_Ratio(E, P_{elo}) = \frac{0.4544}{1.4056} = 0.3233$$

Ganancia de información

E

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



$$Entropia(E) = 0.9544$$

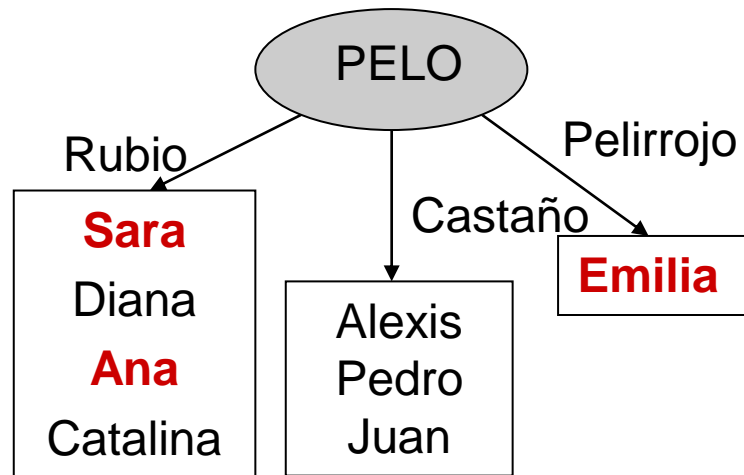
$$Entropia(E, Pelo) = \frac{4}{8} * 1 + \frac{3}{8} * 0 + \frac{1}{8} * 0 = 0.5$$

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

Ganancia de información

E

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



$$Entropia(E) = 0.9544$$

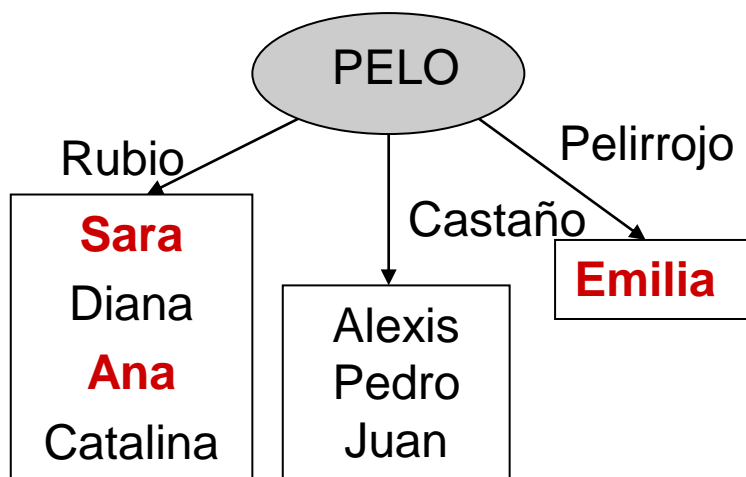
$$Entropia(E, Pelo) = 0.5$$

$$Ganancia(E, PELO) = 0.4544$$

Ganancia de información

E

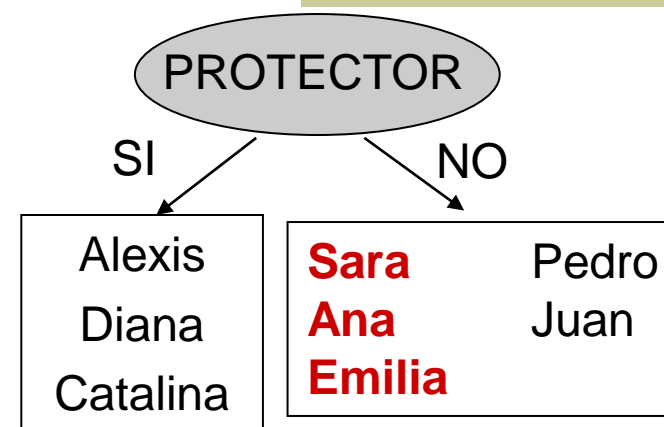
E1, E4, E5 (+)
E2, E3, E6, E7, E8 (-)



$$Entropia(E) = 0.9544$$

$$Entropia(E, Pelo) = 0.5$$

$$Ganancia(E, PELO) = 0.4544$$



$$Entropia(E) = 0.9544$$

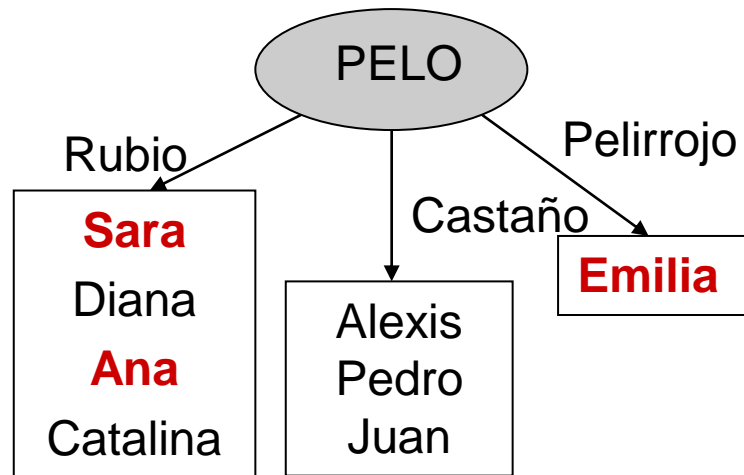
$$Entropia(E, Protector) = \frac{3}{8} * 0 + \frac{5}{8} * 0.971 = 0.6069$$

$$Ganancia(E, Protector) = 0.9544 - 0.6069 = 0.3475$$

Ganancia de información

E

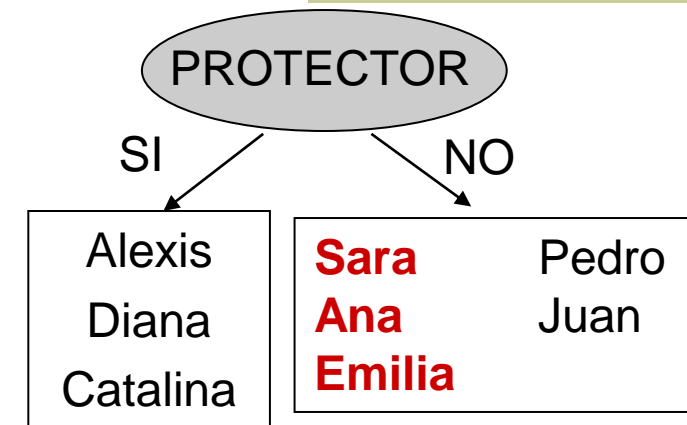
E1, E4, E5 (+)
E2, E3, E6, E7, E8 (-)



$$Entropia(E) = 0.9544$$

$$Entropia(E, Pelo) = 0.5$$

$$Ganancia(E, PELO) = 0.4544$$

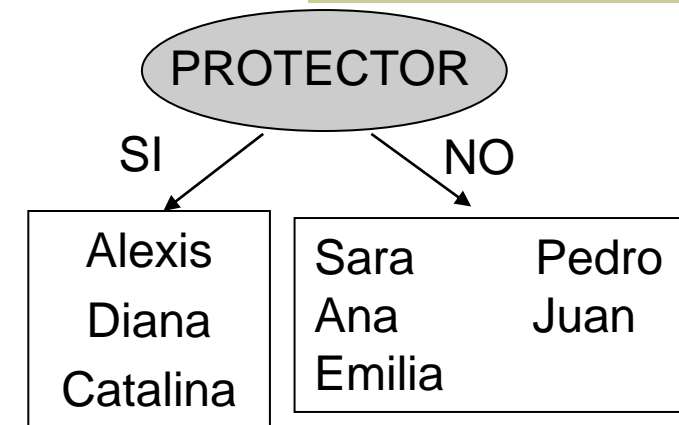
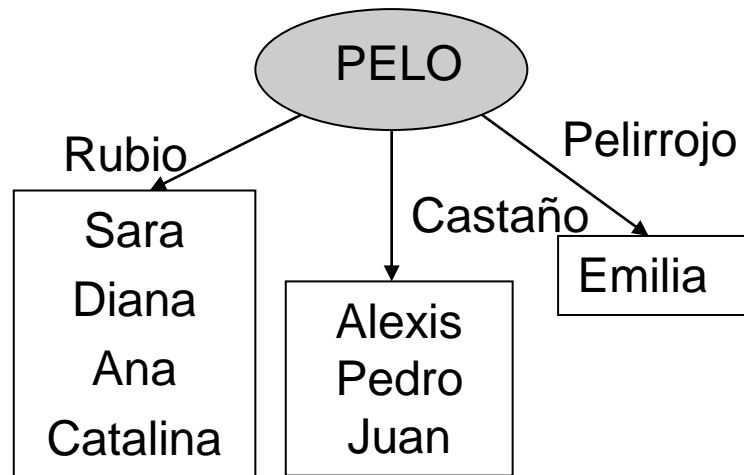


$$Entropia(E) = 0.9544$$

$$Entropia(E, Protector) = 0.6069$$

$$Ganancia(E, Protector) = 0.3475$$

Información por la división



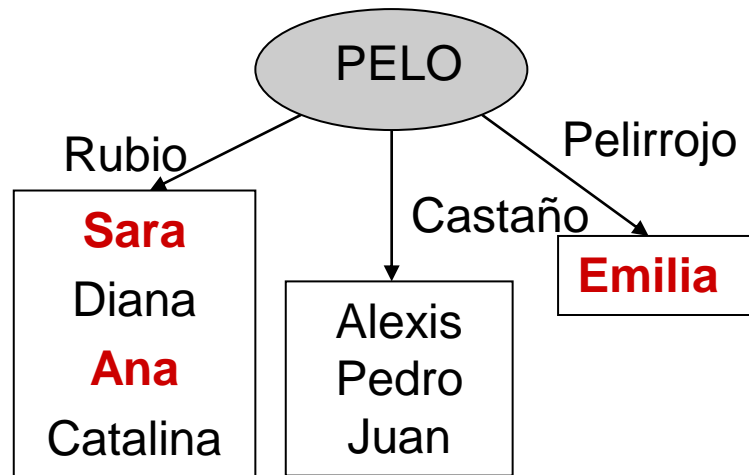
$$\text{InfoDiv}(E, \text{PELO}) = -\frac{4}{8}\log_2\left(\frac{4}{8}\right) - \frac{3}{8}\log_2\left(\frac{3}{8}\right) - \frac{1}{8}\log_2\left(\frac{1}{8}\right) = 1.4056$$

$$\text{InfoDiv}(E, \text{Protector}) = -\frac{3}{8}\log_2\left(\frac{3}{8}\right) - \frac{5}{8}\log_2\left(\frac{5}{8}\right) = 0.9544$$

Tasa de Ganancia

E

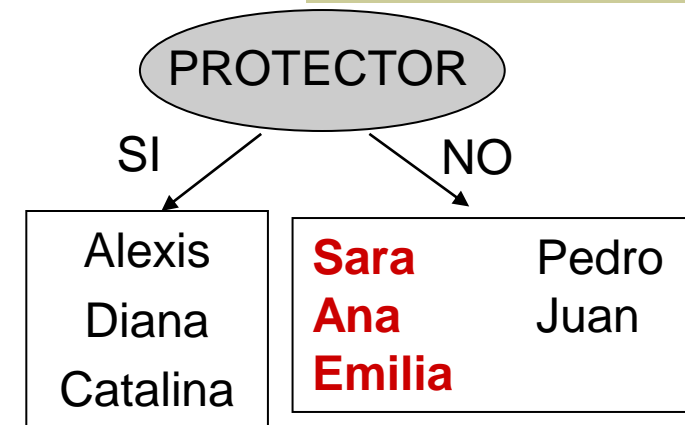
E1, E4, E5 (+)
E2, E3, E6, E7, E8 (-)



$$Ganancia(E, Pelo) = 0.4544$$

$$InfoDiv(E, Pelo) = 1.4056$$

$$TasaGanancia(E, Pelo) = \frac{0.4544}{1.4056} = \mathbf{0.3232}$$

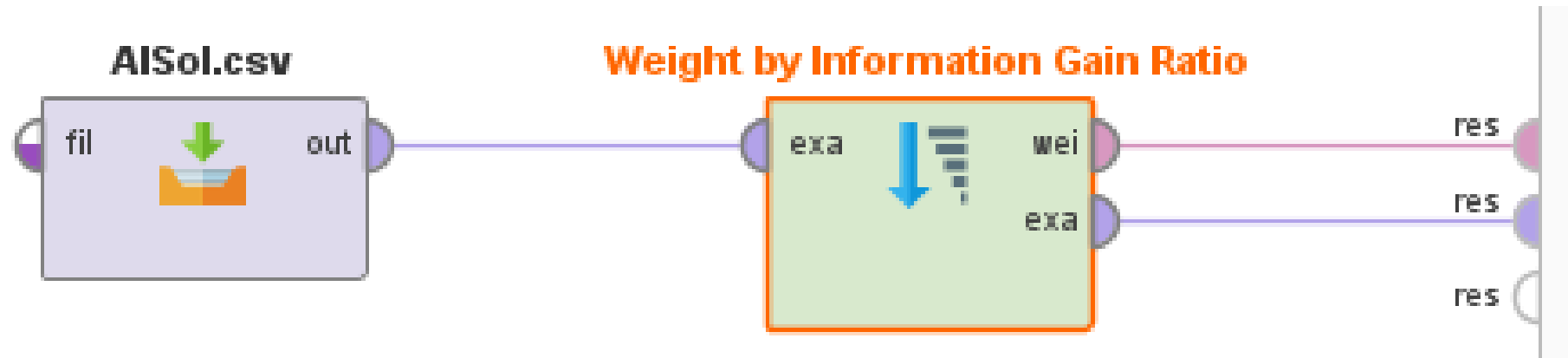


$$Ganancia(E, Protector) = 0.3475$$

$$InfoDiv(E, Protector) = 0.9544$$

$$TasaGanancia(E, Protector) = \frac{0.3475}{0.9544} = \mathbf{0.3641}$$

Tasa de Ganancia con Rapid Miner



Tasa de Ganancia con Rapid Miner

Result History

AttributeWeights (Weight by Information Gain Ratio)

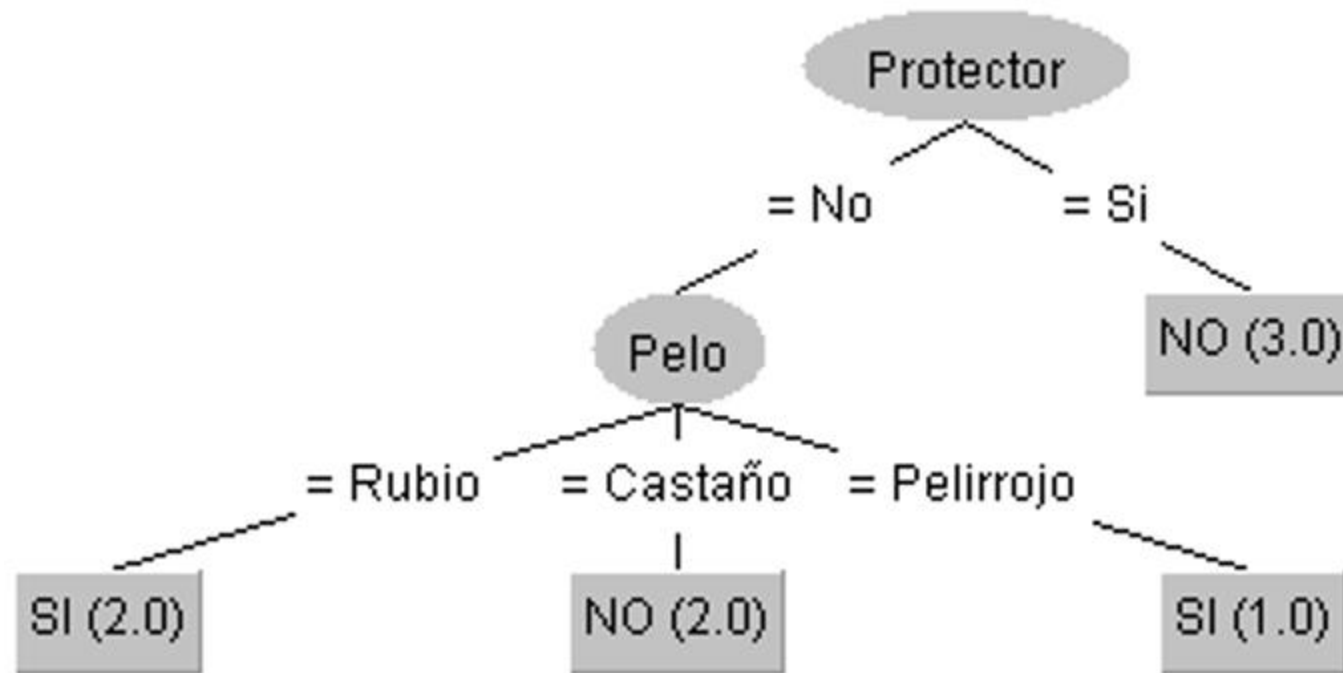
Data

Weight Visualizations

attribute	weight ↓
Protector	0.364
Pelo	0.323
Nombre	0.318
Estatura	0.170
Peso	0.010

Se elige el de mayor valor

Arbol obtenido usando como criterio Tasa de Ganancia

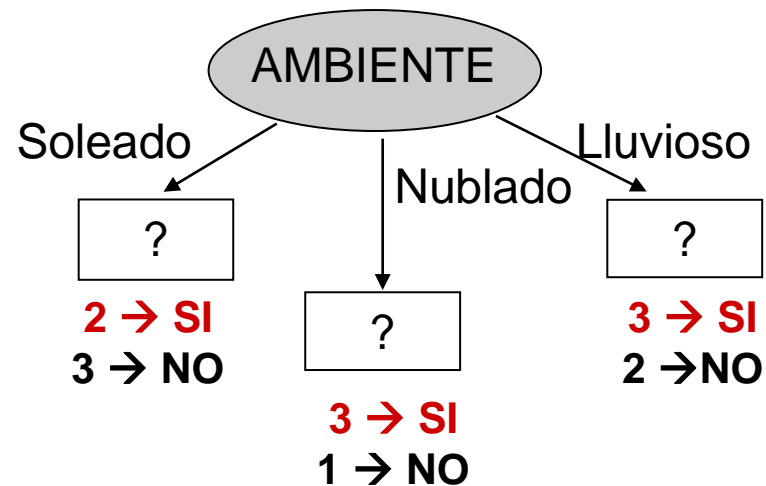


Sobreajuste

- Es el efecto de entrenar excesivamente un modelo (en este caso el árbol) con ciertos datos para los que se conoce el resultado deseado.
- Cuando el modelo se ajusta excesivamente (sobreajusta) a los datos de entrenamiento su desempeño a la hora de clasificarlos es muy superior los obtenidos al aplicarlo sobre ejemplos nuevos. Esto evidencia el **sobreajuste**.
- Construir modelos complejos (en este caso árboles con demasiados nodos) a veces se debe a **sobreajustar** dicho modelo a los datos de entrenamiento.

Ejemplo

- En el archivo **Golf_V2.csv** se ha cambiado el ejemplo 7 a Juega=NO.



- Rehaga el árbol utilizando Id3 y compárelo con el anterior.

Arboles usando sólo Ganancia de Información

Golf.csv

```
Ambiente = lluvioso
| Viento = no: si {no=0, si=3}
| Viento = si: no {no=2, si=0}
Ambiente = nublado: si {no=0, si=4}
Ambiente = soleado
| Humedad = alta: no {no=3, si=0}
| Humedad = normal: si {no=0, si=2}
```

Note que los datos de
entrenamiento sólo difieren en
un ejemplo

Golf_V2.csv

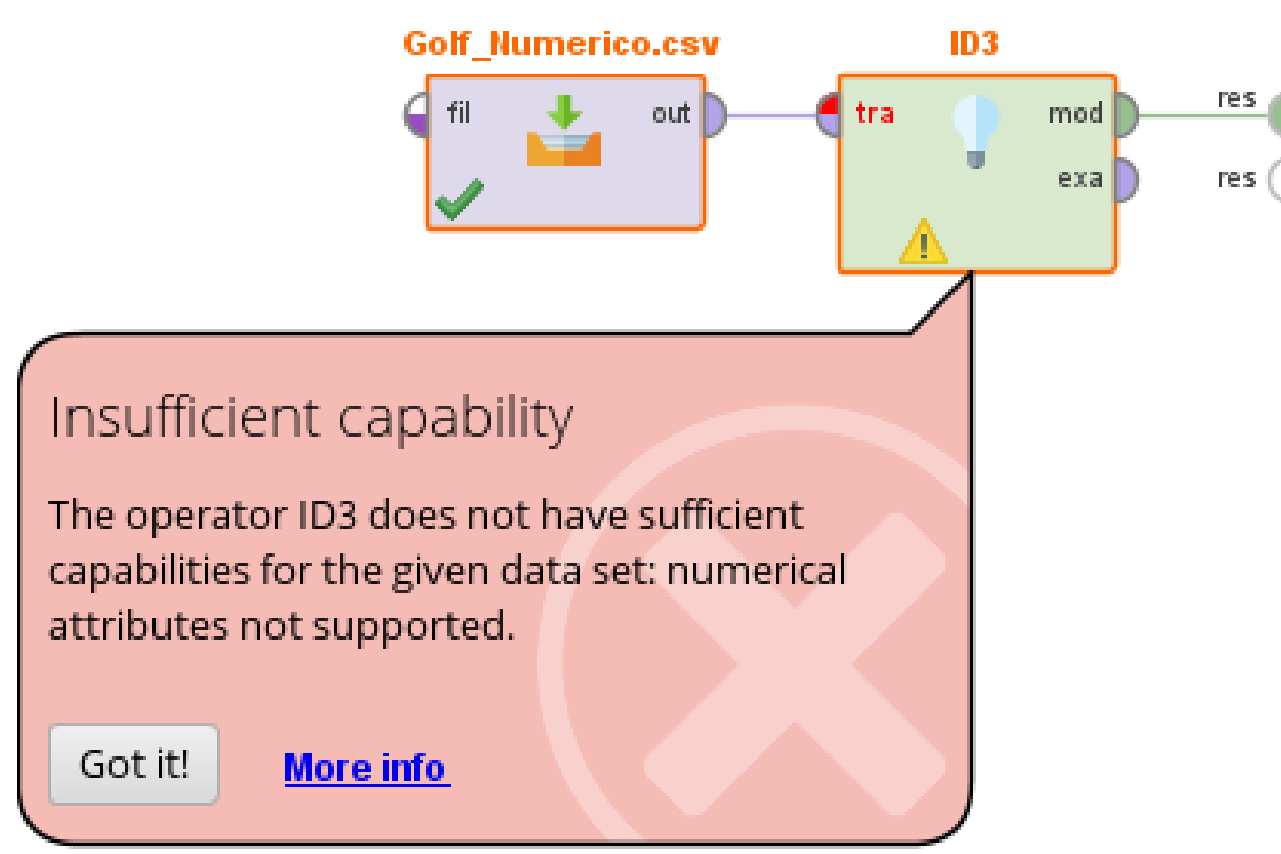
```
Viento = no
| Ambiente = lluvioso: si {no=0, si=3}
| Ambiente = nublado: si {no=0, si=2}
| Ambiente = soleado
| | Temperatura = alta: no {no=1, si=0}
| | Temperatura = baja: si {no=0, si=1}
| | Temperatura = media: no {no=1, si=0}
Viento = si
| Temperatura = alta: no {no=1, si=0}
| Temperatura = baja: no {no=2, si=0}
| Temperatura = media
| | Ambiente = lluvioso: no {no=1, si=0}
| | Ambiente = nublado: si {no=0, si=1}
| | Ambiente = soleado: si {no=0, si=1}
```


Cómo construir el árbol con atributos numéricos

#	Ambiente	Temperatura	Humedad	Viento	Juega
1	soleado	85	85	NO	No
2	soleado	80	90	SI	No
3	nublado	83	86	NO	Si
4	lluvioso	70	96	NO	Si
5	lluvioso	68	80	NO	Si
6	lluvioso	65	70	SI	No
7	nublado	64	65	SI	Si
8	soleado	72	95	NO	No
9	soleado	69	70	NO	Si
10	lluvioso	75	80	NO	Si
11	soleado	75	70	SI	Si
12	nublado	72	90	SI	Si
13	nublado	81	75	NO	Si
14	lluvioso	71	91	SI	No

Operando con atributos numéricos

- El algoritmo ID3 sólo opera con atributos cualitativos



Arbol de clasificación. Algoritmo C4.5

The screenshot shows the WEKA graphical user interface. On the left, a workflow is visible: a file named 'Golf_Numerico.csv' is imported into a 'W-J48' classifier. The classifier has two output ports labeled 'res'. On the right, the 'Parameters' dialog for 'W-J48' is open. It shows a lightbulb icon, a checkbox for 'U', and two numeric parameters: 'C' with a value of 0.25 and 'M' with a value of 2.0. A red arrow points from the 'M' parameter field to a yellow box containing the text 'A mayor valor, menor será la poda'. Another red arrow points from a pink box at the bottom, which says 'Operador de WEKA Importar la extension', to the 'W-J48' classifier box. At the bottom right, there is a logo for 'Weka Extension 7.3.0' with the text 'All modeling methods and evaluation methods from'.

Parameters

W-J48

☐ U

C 0.25

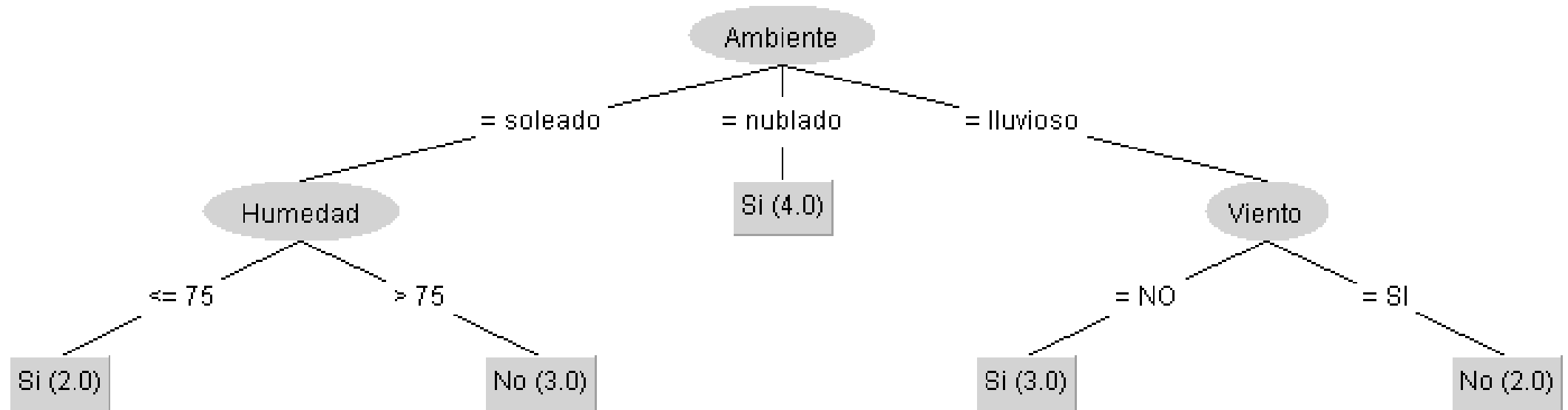
M 2.0

A mayor valor, menor será la poda

Operador de WEKA
Importar la extension

Weka Extension 7.3.0
All modeling methods and
evaluation methods from

Golf_Numérico.csv – Algoritmo C4.5 (W-J48)



Golf_V2.csv

W-J48 (C=0.25)

Precisión 10/14

```
Viento = NO: Si (8.0/2.0)
Viento = SI: No (6.0/2.0)
```

W-J48 (C=0.5)

Precisión 12/14

```
Viento = NO
|  Humedad = alta
|  |  Ambiente = soleado: No
|  |  Ambiente = nublado: Si
|  |  Ambiente = lluvioso: Si
|  Humedad = Normal: Si (4.0)
Viento = SI: No (6.0/2.0)
```

Id3 (precisión 100%)

```
Viento = NO
|  Ambiente = soleado
|  |  Temperatura = alta: No
|  |  Temperatura = media: No
|  |  Temperatura = baja: Si
|  Ambiente = nublado: Si
|  Ambiente = lluvioso: Si
Viento = SI
|  Temperatura = alta: No
|  Temperatura = media
|  |  Ambiente = soleado: Si
|  |  Ambiente = nublado: Si
|  |  Ambiente = lluvioso: No
|  Temperatura = baja: No
```

Arboles. Atributos Numéricos

- Para cada atributo numérico
 - Se ordenan sus valores de menor a mayor
 - Se calculan TODOS los valores de desorden
 - Se elige el menor.



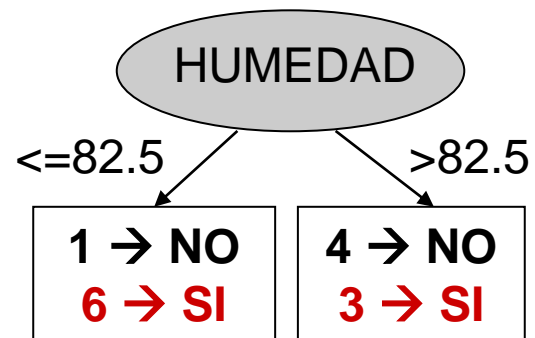
Humedad	65	70	75	80	85	86	90	91	95	96
	Si	No	Si	Si	No	Si	No	No	No	Si
		Si		SI			Si			
		Si								

Analicemos (**Humedad \leq 82.5**) vs (**Humedad $>$ 82.5**)

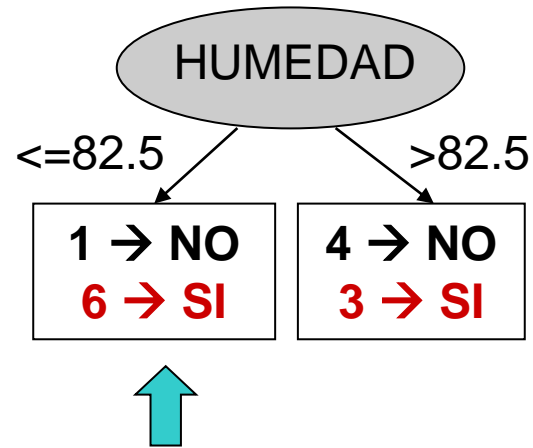
Atributo HUMEDAD



Humedad	65	70	75	80	85	86	90	91	95	96
	Si	No	Si	Si	No	Si	No	No	No	Si
		Si		Si			Si			
		Si								

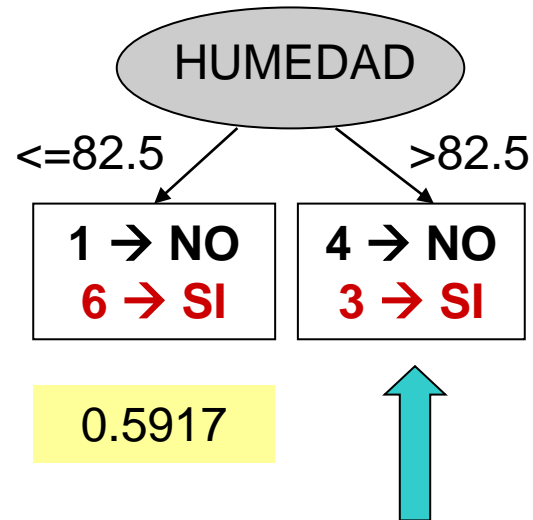


Atributo HUMEDAD



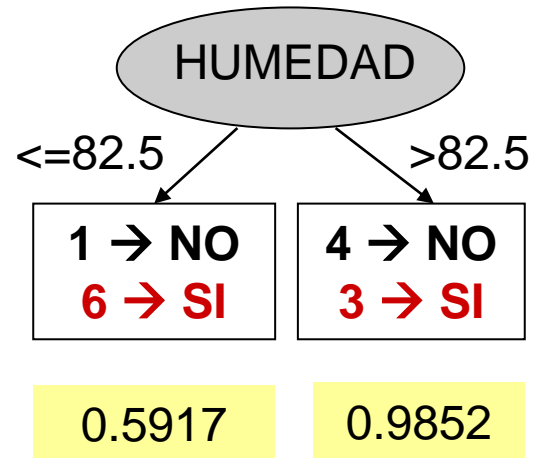
$$Entropia_{\leq 82.5} = -\frac{1}{7} \log_2 \left(\frac{1}{7} \right) - \frac{6}{7} \log_2 \left(\frac{6}{7} \right) = 0.5917$$

Atributo HUMEDAD



$$Entropia_{>82.5} = -\frac{4}{7} \log_2 \left(\frac{4}{7} \right) - \frac{3}{7} \log_2 \left(\frac{3}{7} \right) = 0.9852$$

Atributo HUMEDAD



$$Entropia_{HUMEDAD} = \frac{7}{14} * 0.5917 + \frac{7}{14} * 0.9852 = 0.7885$$

Atributo HUMEDAD

67,5	0,8926
72,5	0,9253
77,5	0,8950
82,5	0,7885
85,5	0,8922
88	0,8380
90,5	0,8610
93	0,9300
95,5	0,8926



Si calculamos TODOS los cortes posibles veremos que 0.7885 es el menor valor y por lo tanto es el nivel de entropía elegido para el atributo HUMEDAD

Atributo TEMPERATURA

64,5	0,8926
66,5	0,9300
68,5	0,9398
69,5	0,9253
70,5	0,8950
71,5	0,9389
73,5	0,9389
77,5	0,9152
80,5	0,9398

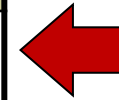


este es el nivel de
entropía elegido para el
atributo
TEMPERATURA

Desorden Promedio de cada atributo

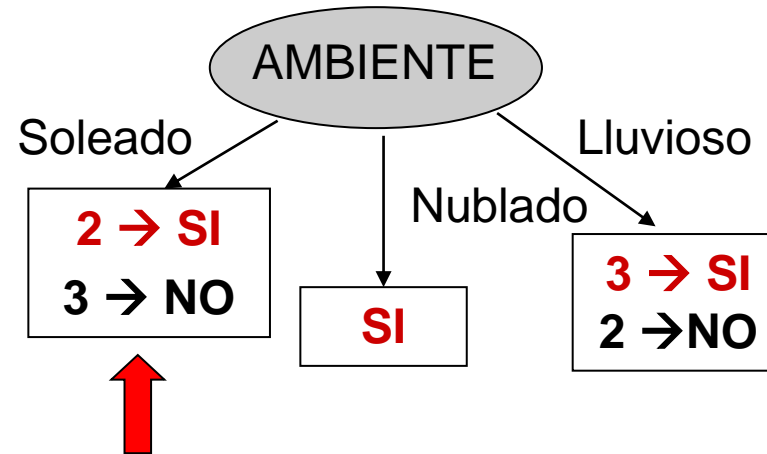
- Repitiendo el mismo proceso para el resto de los atributos puede completarse la siguiente tabla:

Atributo	Entropía
Ambiente	0.6935
Temperatura	0.8926
Humedad	0.7885
Viento	0.8922



Es el
seleccionado
por ser el de
menor valor

Buscando los nodos del 1er. nivel del árbol



Para estas 5 muestras,
calcular el desorden de los
3 atributos restantes

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

#	Ambiente	Temperatura	Humedad	Viento	Juega
1	soleado	85	85	NO	No
2	soleado	80	90	SI	No
8	soleado	72	95	NO	No
9	soleado	69	70	NO	Si
11	soleado	75	70	SI	Si

TEMPERATURA

70,5	73,5	77,5	82,5
0,6490	0,9510	0,5510	0,80

HUMEDAD

77,5	87,5	92,5
0	0,5510	0,8

VIENTO

0.9510

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

#	Ambiente	Temperatura	Humedad	Viento	Juega
1	soleado	85	85	NO	No
2	soleado	80	90	SI	No
8	soleado	72	95	NO	No
9	soleado	69	70	NO	Si
11	soleado	75	70	SI	Si

TEMPERATURA

70,5	73,5	77,5	82,5
0,6490	0,9510	0,5510	0,80

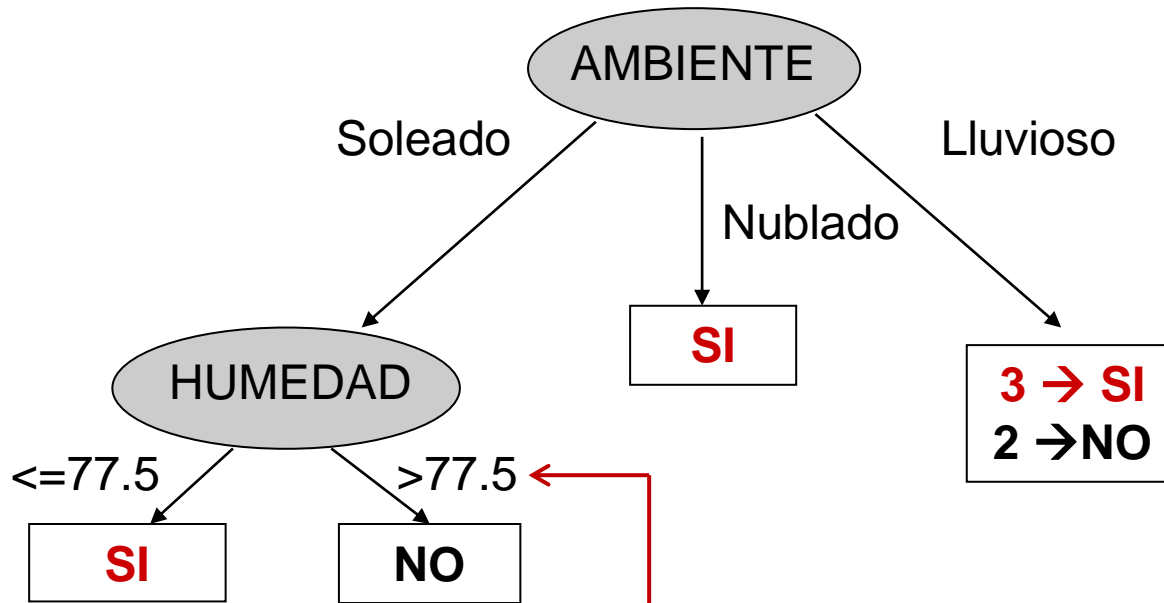
HUMEDAD

77,5	87,5	92,5
0	0,5510	0,8

VIENTO

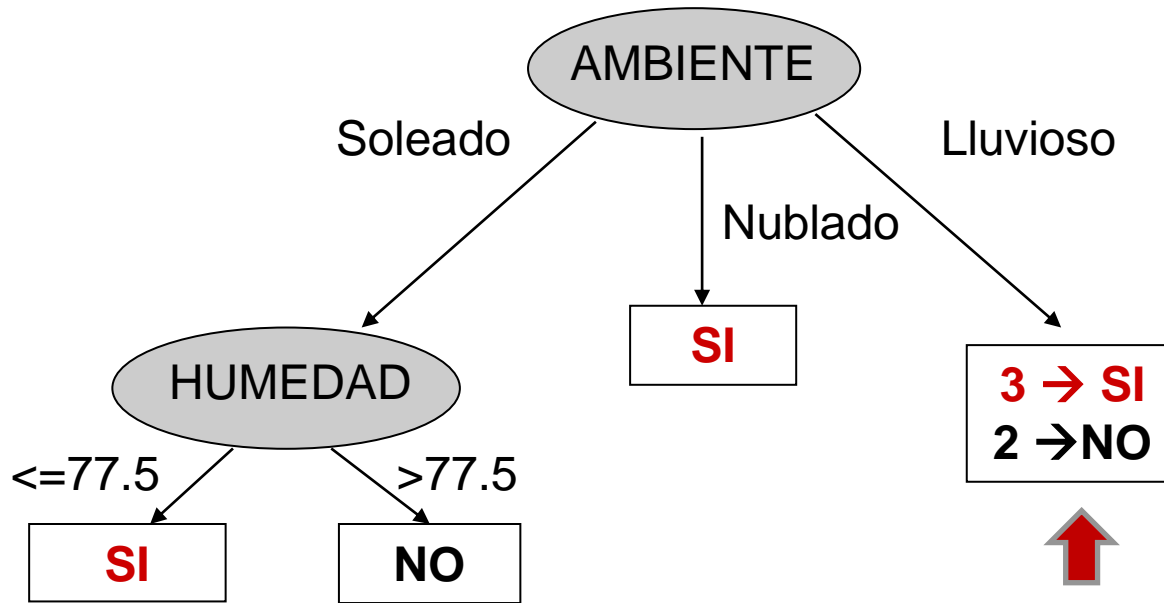
0.9510

Arbol de decisión



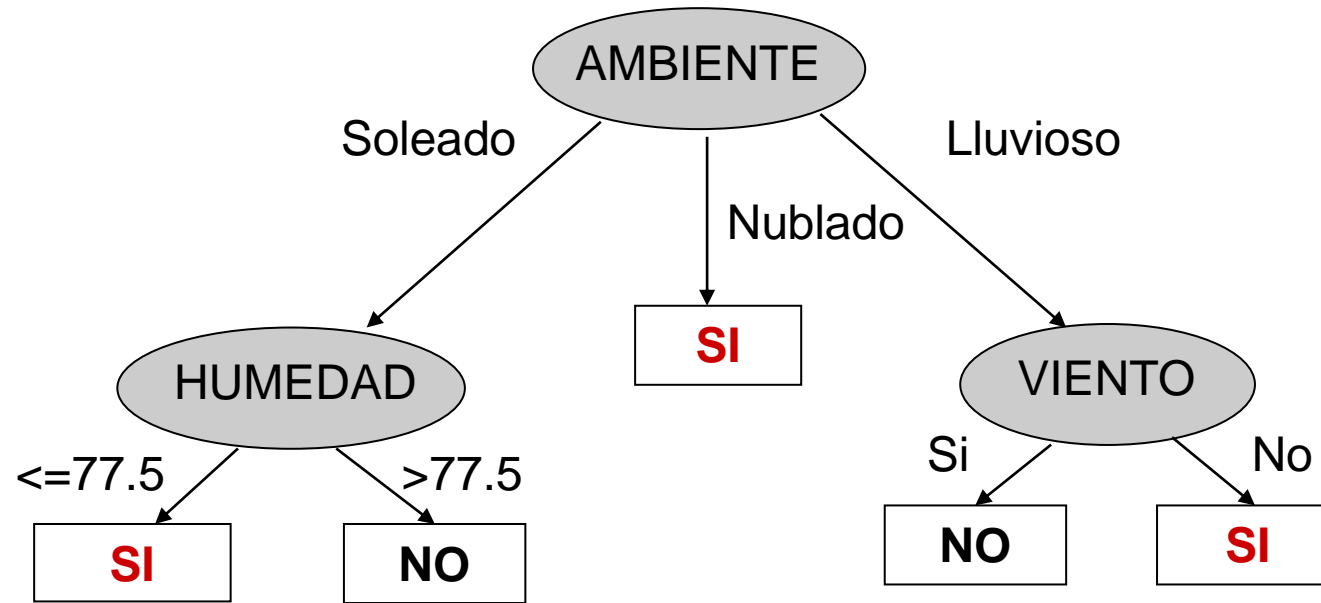
El punto de corte puede variar (ej: sólo se considera el punto medio o se tiene en cuenta la cantidad de repetidos en cada extremo).

Arbol de decisión



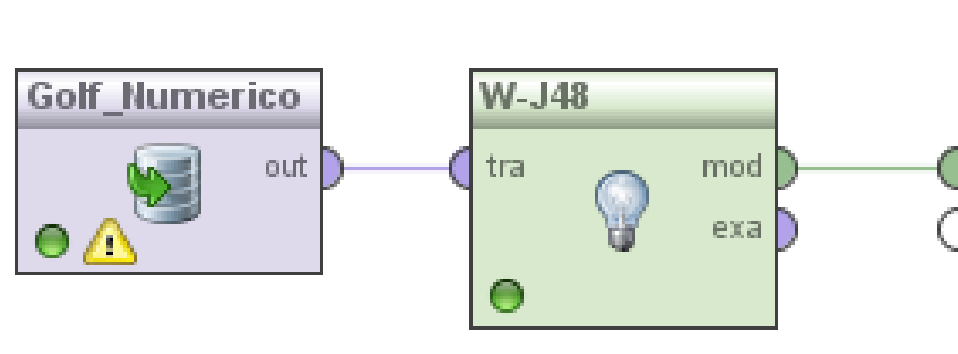
Para estos 5 ejemplos se repite nuevamente el proceso de selección

Arbol de decisión

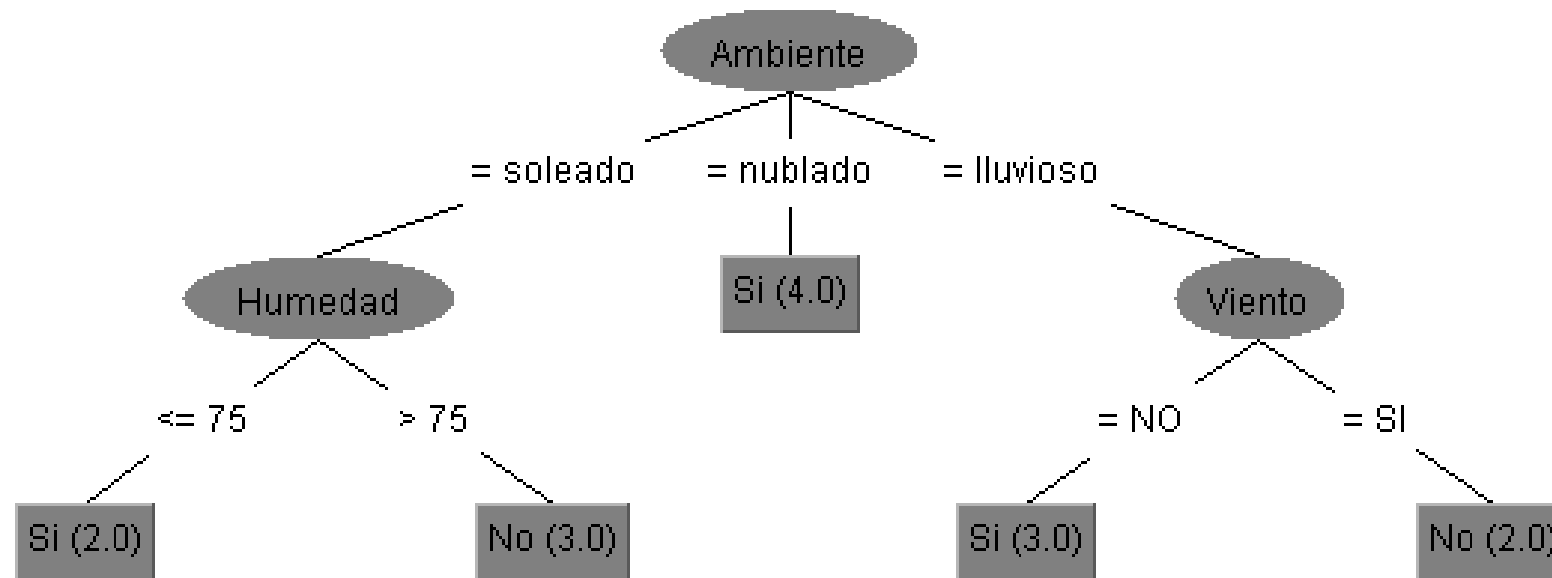


Ejemplo

- Utilice el archivo “**Golf_numerico.csv**” y construya nuevamente el árbol utilizando el operador W-J48 que no requiere atributos discretos



Golf_numérico



Poda

- Estrategias para minimizar el sobreajuste
 - **Prepoda:** limitar el crecimiento del árbol cuando la división de los ejemplos no es estadísticamente significativa.
 - **Pospoda:** podar el árbol luego de haberlo generado. Es lo más usado.
- Cuando se poda un nodo de un árbol, se eliminan las ramificaciones y se convierte a ese nodo en una hoja a la que se puede asignar la clase mayoritaria de los ejemplos vinculados a ese nodo.

Prepoda

```
Viento = no
|  Ambiente = lluvioso: si {no=0, si=3}
|  Ambiente = nublado: si {no=0, si=2}
|  Ambiente = soleado
|  |  Temperatura = alta: no {no=1, si=0}
|  |  Temperatura = baja: si {no=0, si=1}
|  |  Temperatura = media: no {no=1, si=0}
Viento = si
|  Temperatura = alta: no {no=1, si=0}
|  Temperatura = baja: no {no=2, si=0}
|  Temperatura = media
|  |  Ambiente = lluvioso: no {no=1, si=0}
|  |  Ambiente = nublado: si {no=0, si=1}
|  |  Ambiente = soleado: si {no=0, si=1}
```

**Hay varias hojas
con un único
ejemplo**

Prepoda : no generar hojas con 1 ejemplo

```
Viento = no
| Ambiente = lluvioso: si {no=0, si=3}
| Ambiente = nublado: si {no=0, si=2}
| Ambiente = soleado
| | Temperatura = alta: no {no=1, si=0}
| | Temperatura = baja: si {no=0, si=1}
| | Temperatura = media: no {no=1, si=0}
Viento = si
| Temperatura = alta: no {no=1, si=0}
| Temperatura = baja: no {no=2, si=0}
| Temperatura = media
| | Ambiente = lluvioso: no {no=1, si=0}
| | Ambiente = nublado: si {no=0, si=1}
| | Ambiente = soleado: si {no=0, si=1}
```

**Si no se usa TEMPERATURA la
rama AMBIENTE=SOLEADO
responde Juega=no (1 error)**

**Si la rama VIENTO=NO
responde Juega=SI
hay sólo 2 errores**

Prepoda : no generar hojas con 1 ejemplo

```
Viento = no
|  Ambiente = lluvioso: si {no=0, si=3}
|  Ambiente = nublado: si {no=0, si=2}
|  Ambiente = soleado
|  |  Temperatura = alta: no {no=1, si=0}
|  |  Temperatura = baja: si {no=0, si=1}
|  |  Temperatura = media: no {no=1, si=0}
Viento = si
|  Temperatura = alta: no {no=1, si=0}
|  Temperatura = baja: no {no=2, si=0}
|  Temperatura = media
|  |  Ambiente = lluvioso: no {no=1, si=0}
|  |  Ambiente = nublado: si {no=0, si=1}
|  |  Ambiente = soleado: si {no=0, si=1}
```

Si no se usa AMBIENTE la
rama TEMPERATURA = MEDIA
responde Juega=SI (1 error)

Si la rama VIENTO=SI
responde Juega=NO
se producen 2 errores

Golf_V2

W-J48 (C=0.25)

Precisión 10/14

Viento = NO: Si (8.0/2.0)

Viento = SI: No (6.0/2.0)

W-J48 (C=0.5)

Precisión 12/14

Viento = NO

| Humedad = alta

| | Ambiente = soleado: No

| | Ambiente = nublado: Si

| | Ambiente = lluvioso: Si

| Humedad = Normal: Si (4.0)

Viento = SI: No (6.0/2.0)

Id3 (precisión 100%)

Viento = NO

| Ambiente = soleado

| | Temperatura = alta: No

| | Temperatura = media: No

| | Temperatura = baja: Si

| Ambiente = nublado: Si

| Ambiente = lluvioso: Si

Viento = SI

| Temperatura = alta: No

| Temperatura = media

| | Ambiente = soleado: Si

| | Ambiente = nublado: Si

| | Ambiente = lluvioso: No

| Temperatura = baja: No

Prepoda

- Una forma fácil de prepoda es no permitir hojas con un único ejemplo.
- Hay otros criterios. Ej: chi-cuadrado, cota de error

Viento = no: si {no=2, si=6}

Viento = si: no {no=4, si=2}

**No se generan tantas ramas.
Esto facilita la lectura**

Ejemplo: Clasificación de flores de Iris

- Se dispone de información de 3 tipos de flores Iris



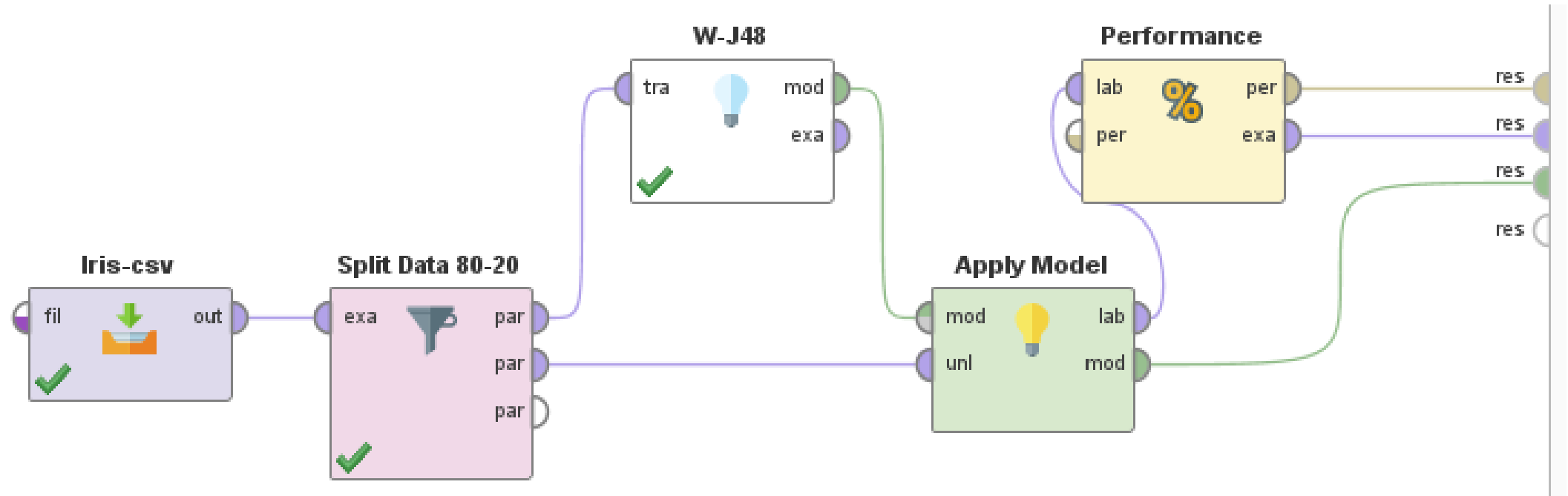
<https://archive.ics.uci.edu/ml/datasets/Iris>

Ejemplo: Clasificación de flores de Iris

Id	sepalength	sepalwidth	petallength	petalwidth	class
1	5,1	3,5	1,4	0,2	Iris-setosa
2	4,9	3,0	1,4	0,2	Iris-setosa
...
95	5,6	2,7	4,2	1,3	Iris-versicolor
96	5,7	3,0	4,2	1,2	Iris-versicolor
97	5,7	2,9	4,2	1,3	Iris-versicolor
...
149	6,2	3,4	5,4	2,3	Iris-virginica
150	5,9	3,0	5,1	1,8	Iris-virginica

<https://archive.ics.uci.edu/ml/datasets/Iris>

Ejemplo 6: Clasificación de flores de Iris



Ejemplo 6 – Iris.csv

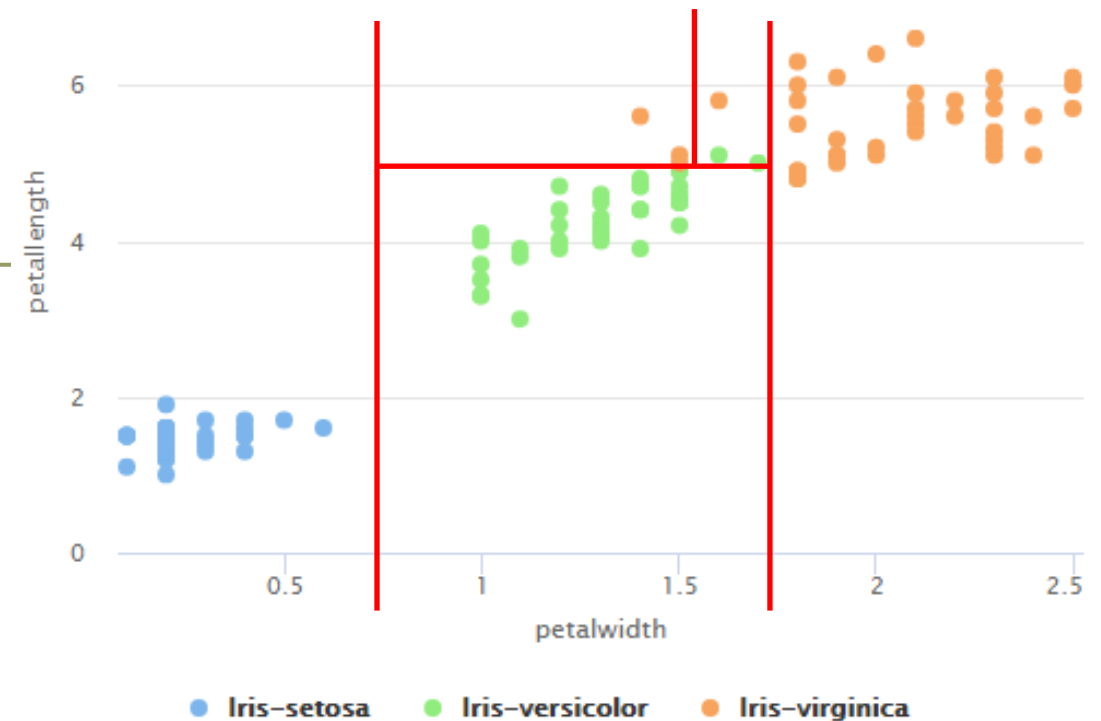
W-J48

J48 pruned tree

```
-----  
petalwidth <= 0.6: Iris-setosa (40.0)  
petalwidth > 0.6  
|   petalwidth <= 1.7  
|   |   petallength <= 4.9: Iris-versicolor (37.0)  
|   |   petallength > 4.9  
|   |   |   petalwidth <= 1.5: Iris-virginica (3.0)  
|   |   |   petalwidth > 1.5: Iris-versicolor (3.0/1.0)  
|   petalwidth > 1.7: Iris-virginica (37.0/1.0)
```

Number of Leaves : 5

Size of the tree : 9



Ejemplo: matriz de confusión

accuracy: 96.67%

	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	10	0	0	100.00%
pred. Iris-versicolor	0	10	1	90.91%
pred. Iris-virginica	0	0	9	100.00%
class recall	100.00%	100.00%	90.00%	

- Verifique que la tasa de acierto (*accuracy*) es algo superior sobre los datos de testeo