Universidad Carlos III de Madrid

Aprendizaje Automático para el Análisis de Datos

GRADO EN ESTADÍSTICA Y EMPRESA

Ricardo Aler Mur

MODELOS:
ÁRBOLES DE DECISIÓN Y REGLAS

Datos de entrada

Cielo	Temperatura	Humedad	Viento	Tenis
Sol	85	85	No	**No**
Sol	80	90	Si	**No**
Nublado	83	86	No	**Si**
Lluvia	70	96	No	**No**
Lluvia	68	80	No	**Si**
Nublado	64	65	Si	**Si**
Sol	72	95	No	**No**
Sol	69	70	No	**Si**
Lluvia	75	80	No	**Si**
Sol	75	70	Si	**Si**
Nublado	72	90	Si	**Si**
Nublado	81	75	No	**Si**
Lluvia	71	91	Si	**No**

Esquema general en clasificación

Datos

Cielo	Temperatura	Humedad	Viento	Tenis
Sol	85	85	No	No
Sol	80	90	Si	No
Nubes	83	86	No	Si
Lluvia	70	96	No	So
Lluvia	68	80	No	Si
Nublado	64	65	Si	Si
Sol	72	95	No	No
Sol	69	70	No	Si
Lluvia	75	80	No	Si
Sol	75	70	Si	Si
Nublado	72	90	Si	Si
Nublado	81	75	No	Si
Lluvia	71	91	Si	No

Dato a clasificar

Cielo	Temperatura	Humedad	Viento	Tenis
Sol	60	65	No	?????

Clasificador

Algoritmo MD

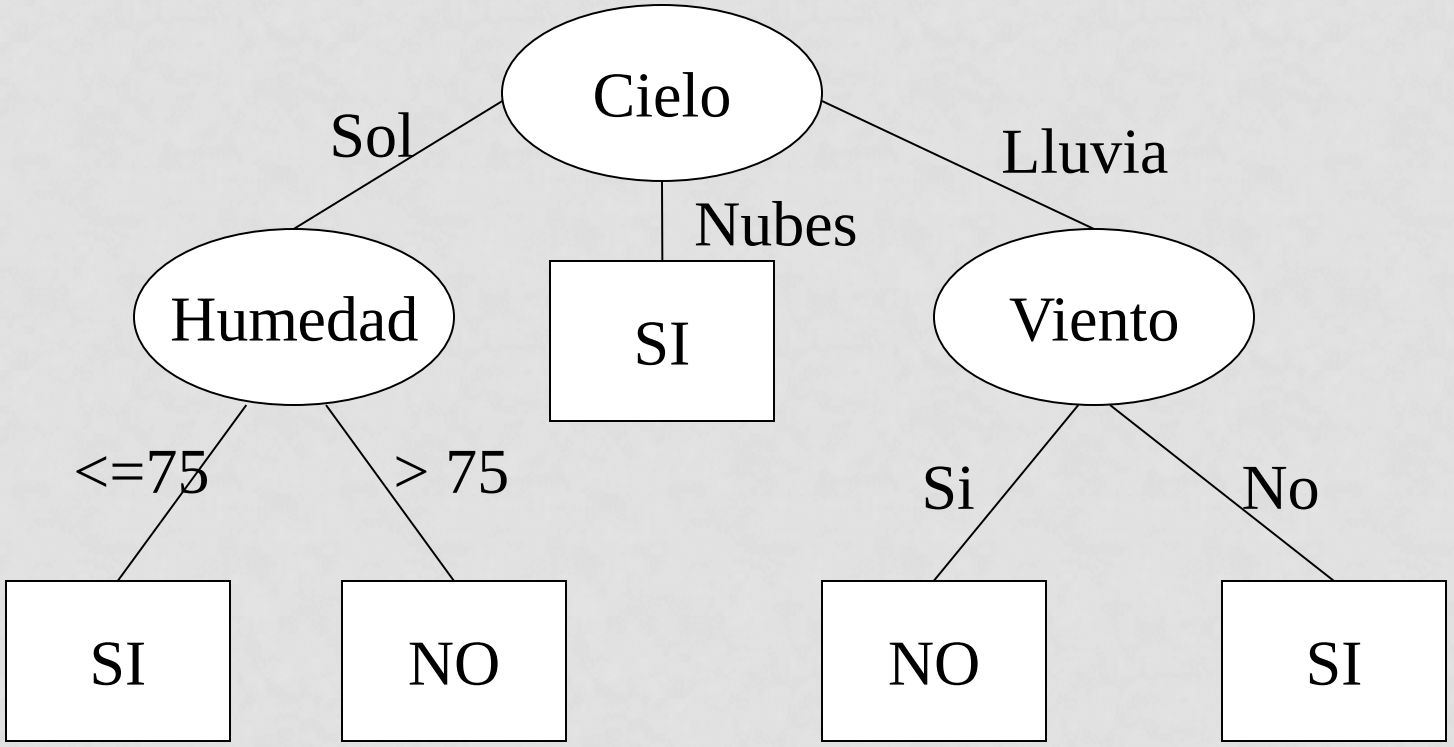
IF Cielo = Sol **Y** Humedad <= 75

THEN Tenis = Si ...

Predicción

Clase = Si

Árboles de decisión



Algoritmos de construcción de árboles de decisión

- El más básico es el ID3: construye árboles de decisión de manera recursiva, de la raíz hacia las hojas, seleccionando en cada momento el mejor nodo para poner en el árbol
- El C4.5 (o J48), trata con valores continuos y utiliza criterios estadísticos para impedir que el árbol se sobreadapte (que "crezca demasiado", que se aprenda los datos en lugar de generalizar)

Algoritmo ID3 simplificado

1. Detener la construcción del árbol si:
   1. Todos los ejemplos pertenecen a la misma clase
   2. Si no quedan ejemplos o atributos
2. Si no, elegir el mejor atributo para poner en ese nodo (el que minimice la entropía media)
3. Crear de manera recursiva tantos subárboles como posibles valores tenga el atributo seleccionado

Algoritmo ID3 detallado

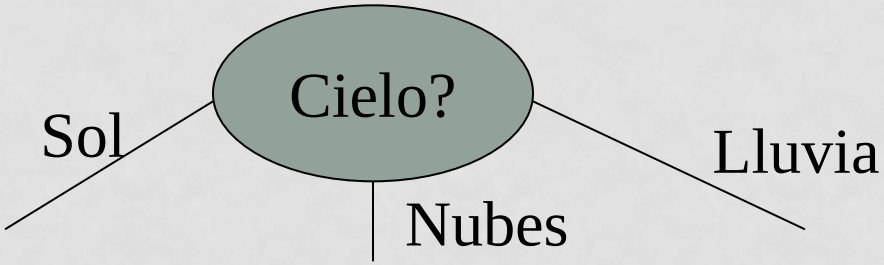
- **ID3(Ejemplos, Atributo-objetivo, Atributos)**

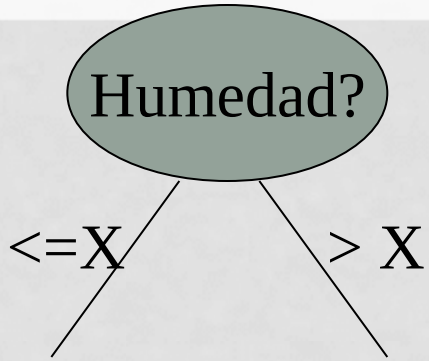
```
1. Si todos los Ejemplos son positivos, devolver un nodo etiquetado con +
2. Si todos los Ejemplos son negativos, devolver un nodo etiquetado con -
3. Si Atributos está vacío, devolver un nodo etiquetado con el valor más
   frecuente de Atributo-objetivo en Ejemplos.
4. En otro caso:
   4.1. Sea A el atributo de Atributos que MEJOR clasifica Ejemplos
   4.2. Crear Árbol, con un nodo etiquetado con A.
   4.3. Para cada posible valor v de A, hacer:
     * Añadir un arco a Árbol, etiquetado con v.
     * Sea Ejemplos(v) el subconjunto de Ejemplos con valor
       del atributo A igual a v.
     * Si Ejemplos(v) es vacío:
       - Entonces colocar debajo del arco anterior un nodo etiquetado con
         el valor más frecuente de Atributo-objetivo en Ejemplos.
       - Si no,  colocar debajo del arco anterior el subárbol
         ID3(Ejemplos(v), Atributo-objetivo, Atributos-{A}).
   4.4  Devolver Árbol
```

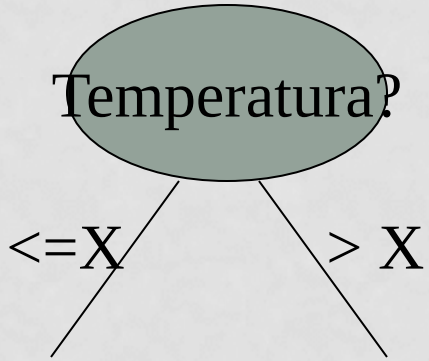
Algoritmo C4.5 simplificado

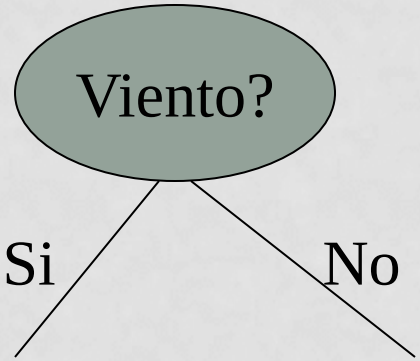
1. Detener la construcción del árbol si:
    1. Todos los ejemplos pertenecen a la misma clase
    2. Si no quedan ejemplos o atributos
    3. Si no se espera que se produzcan mejoras continuando la subdivisión
2. Si no, elegir el mejor atributo para poner en ese nodo (el que minimice la entropía media)
3. Crear de manera recursiva tantos subárboles como posibles valores tenga el atributo seleccionado

¿Qué nodo es el mejor para poner en la raíz del árbol?



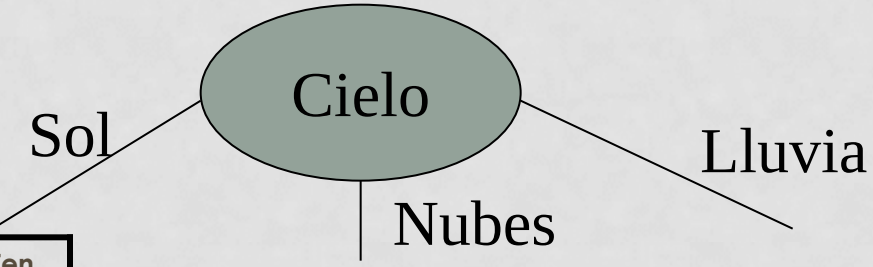






Supongamos que usamos Cielo

Cielo nos genera tres particiones de los datos, tantas como valores posibles tiene



C	T	H	V	Ten
Sol	85	85	No	No
Sol	80	90	Si	No
Sol	72	95	No	No
Sol	69	70	No	Si
Sol	75	70	Si	Si

C	T	H	V	Ten
Nubes	83	86	No	Si
Nubes	64	65	Si	Si
Nubes	72	90	Si	Si
Nubes	81	75	No	Si

C	T	H	V	Ten
Lluvia	70	96	No	No
Lluvia	68	80	No	Si
Lluvia	75	80	No	Si
Lluvia	65	70	Si	No
Lluvia	71	91	Si	No

"3 No, 2 Si"

Tendencia al "no"

"0 No, 4 Si"

Partición perfecta

"3 No, 2 Si"

Tendencia al "no"

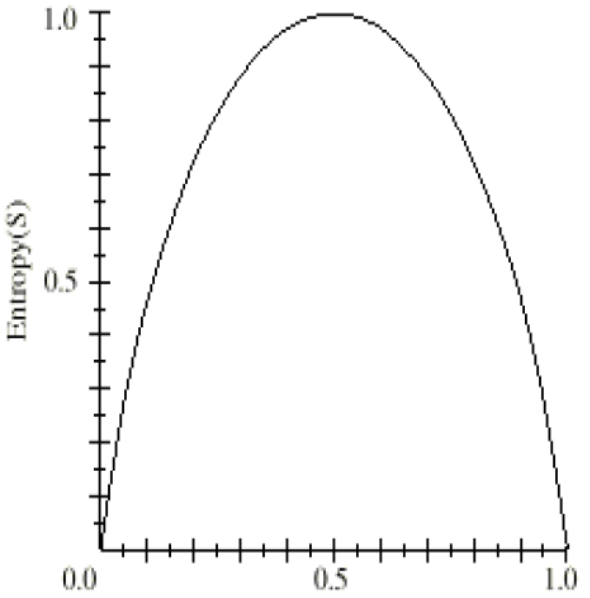
¿Cómo medimos lo bueno que es Cielo como atributo para clasificar?

- Usaremos una medida que obtenga el mejor valor cuando el atributo me obtenga particiones lo mas homogéneas posible, en media
  - Homogénea: "0 No, todo Si"; o bien "todo No, 0 Si"
  - Indecisión: "50% No, 50% Si"
- Una medida que me dice lo lejana que está una partición de la perfección es la entropía
- A mayor entropía, peor es la partición

$$H(P) = -\sum_{Ci} p_{Ci} \log_2(p_{Ci})$$

$$H(P) = -(p_{si} \log_2(p_{si}) + p_{no} \log_2(p_{no}))$$

$$p_{no} = (1 - p_{si})$$



Entropía media de Cielo

- Cielo genera tres particiones cuya entropía es:
  1. "3 No, 2 Si": $H=-((3/5)*\log_2(3/5)+(2/5)*\log_2(2/5) = 0.97$
  2. "0 No, 4 Si": $H=-((0/4)*\log_2(0/4)+1*\log_2(1))=0$
  3. "3 No, 2 Si": $H=-((3/5)*\log_2(3/5)+(2/5)*\log_2(2/5))= 0.97$

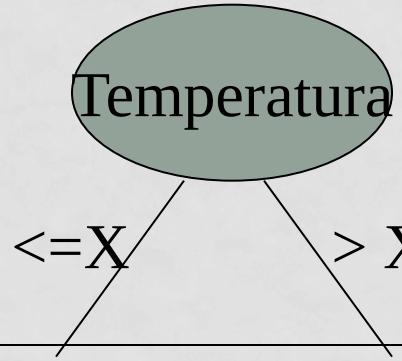
La entropía media ponderada de Cielo será:

- $HP=(5/14)*0.97+(4/14)*0+(5/14)*0.97 =$ **0.69**
- Nota: hay 14 datos en total

¿Y si el atributo es contínuo?

Hay que partir por el valor X, donde sea mas conveniente, minimizando la entropía

Nota: solo hemos probado algunas de las posibles particiones, entre las que se encuentra la mejor



X<=70

64 – Si, 65 – No, 68 – Si, 69 – Si, 70 – Si, 71 – No, 72 – No Si, 75 – Si Si, 80 – No, 81 – Si, 83 – Si, 85 - No

1 No, 4 Si

4 No, 5 Si

HP = 0.89

64 – Si, 65 – No, 68 – Si, 69 – Si, 70 – Si, 71 – No, 72 – No Si, 75 – Si Si, 80 – No, 81 – Si, 83 – Si, 85 - No

3 No, 5 Si

2 No, 4 Si

HP = 0.93

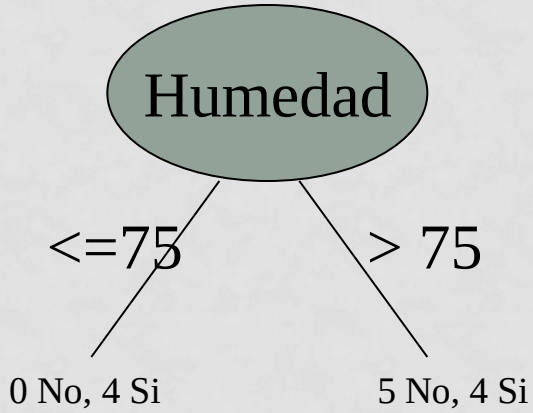
64 – Si, 65 – No, 68 – Si, 69 – Si, 70 – Si, 71 – No, 72 – No Si, 75 – Si Si, 80 – No, 81 – Si, 83 – Si, 85 - No

3 No, 7 Si

2 No, 2 Si

HP = 0.91

Caso de humedad



65-Si, 70-No Si Si, 75–Si, 80-Si Si, | 85-No, 86-Si, 90-No Si, 91-No, 95-No, 96-Si,

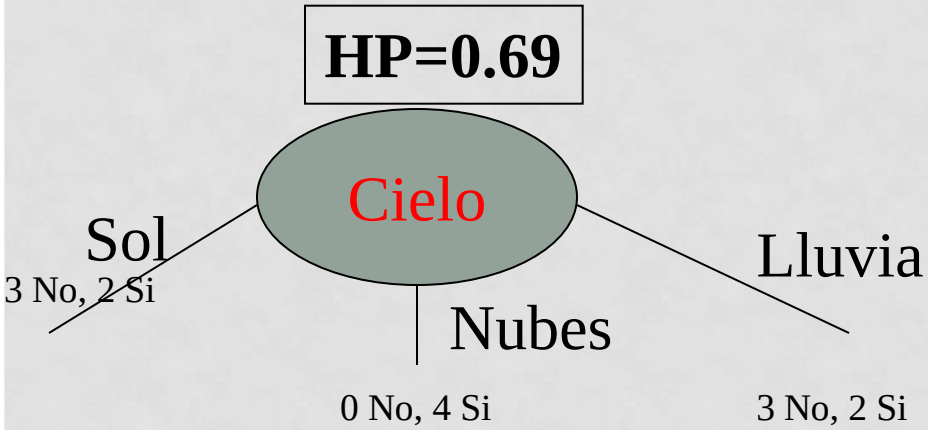
1 No, 6 Si

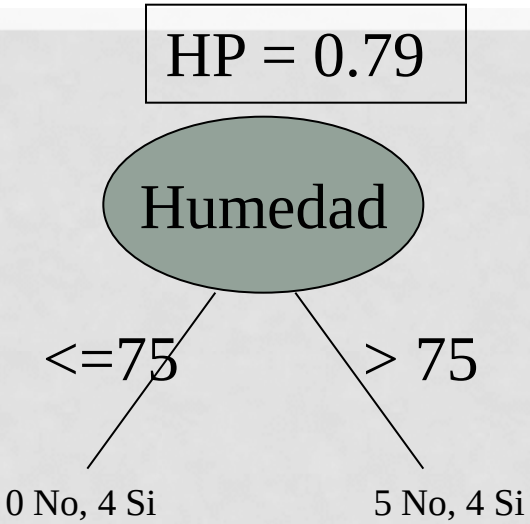
4 No, 3 Si

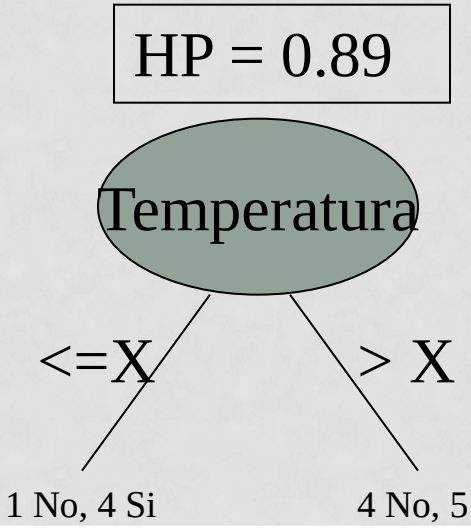
HP = 0.79

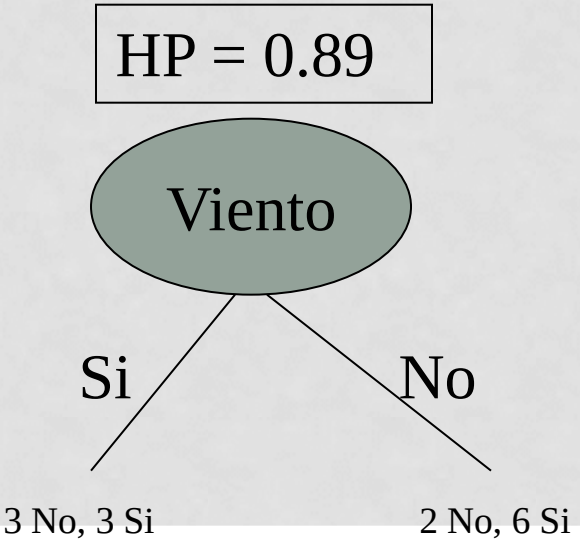
Nota: hay otras posibilidades de particiones, pero esta es la mejor

¿Qué nodo es el mejor para poner en la raíz?



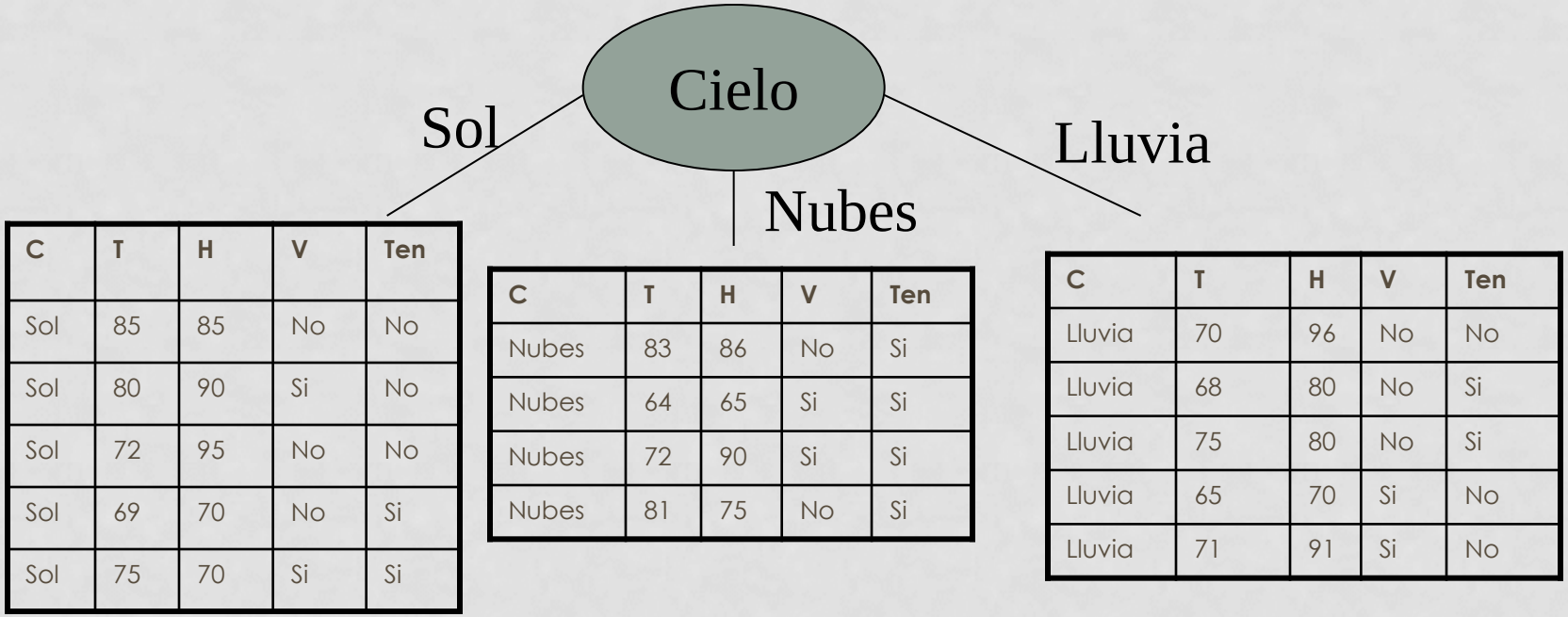






Construcción recursiva del árbol

Ahora que ya tenemos el nodo raíz, el proceso continua recursivamente: hay que construir tres subárboles con los datos que se muestran en cada rama



C	T	H	V	Ten
Sol	85	85	No	No
Sol	80	90	Si	No
Sol	72	95	No	No
Sol	69	70	No	Si
Sol	75	70	Si	Si

C	T	H	V	Ten
Nubes	83	86	No	Si
Nubes	64	65	Si	Si
Nubes	72	90	Si	Si
Nubes	81	75	No	Si

C	T	H	V	Ten
Lluvia	70	96	No	No
Lluvia	68	80	No	Si
Lluvia	75	80	No	Si
Lluvia	65	70	Si	No
Lluvia	71	91	Si	No

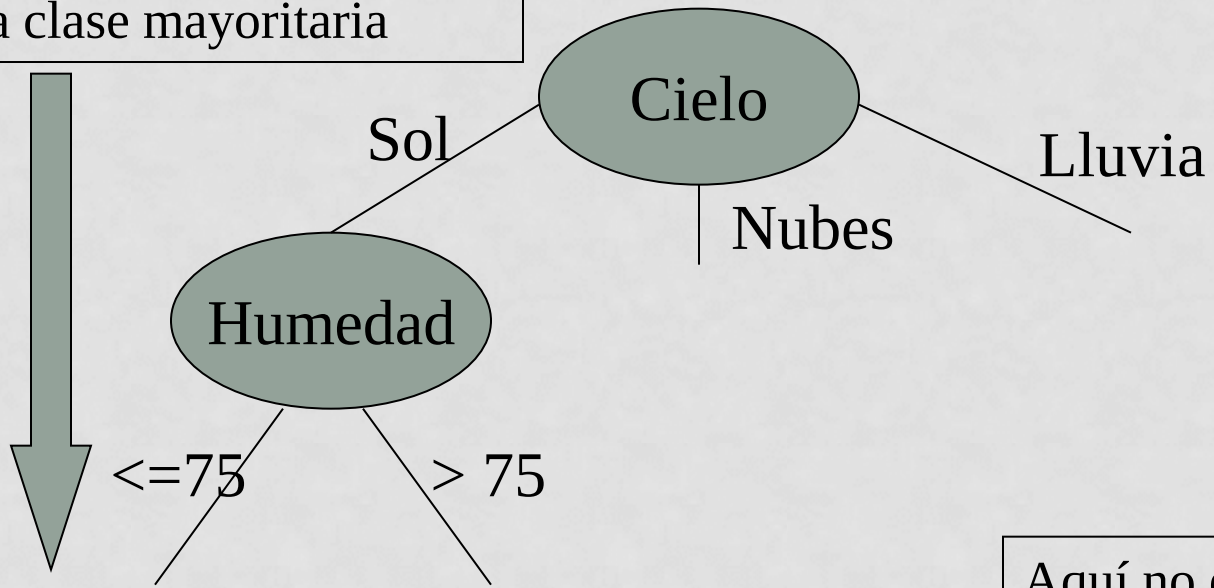
“3 No, 2 Si”

“0 No, 4 Si”

“3 No, 2 Si”

Construcción recursiva del árbol

Aquí un criterio estadístico determina que no merece la pena seguir subdividiendo y se asigna la clase mayoritaria

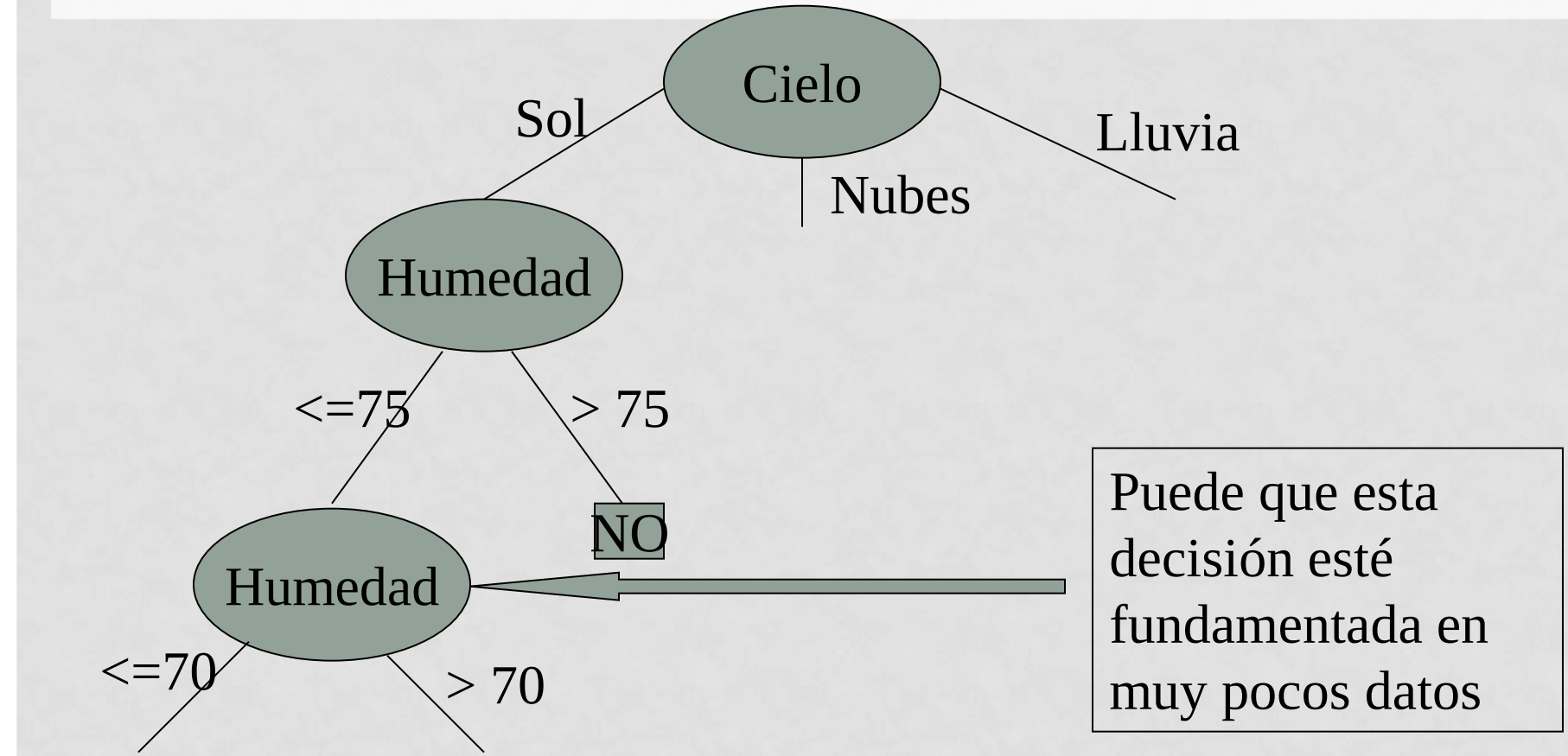


T	H	V	Ten
72	95	No	No
69	70	No	Si
75	70	Si	Si

T	H	V	Ten
85	85	No	No
80	90	Si	No

Aquí no es necesario seguir subdividiendo porque todos los datos son de la misma clase

¿Porqué no seguir subdividiendo?



Puede que esta decisión esté fundamentada en muy pocos datos

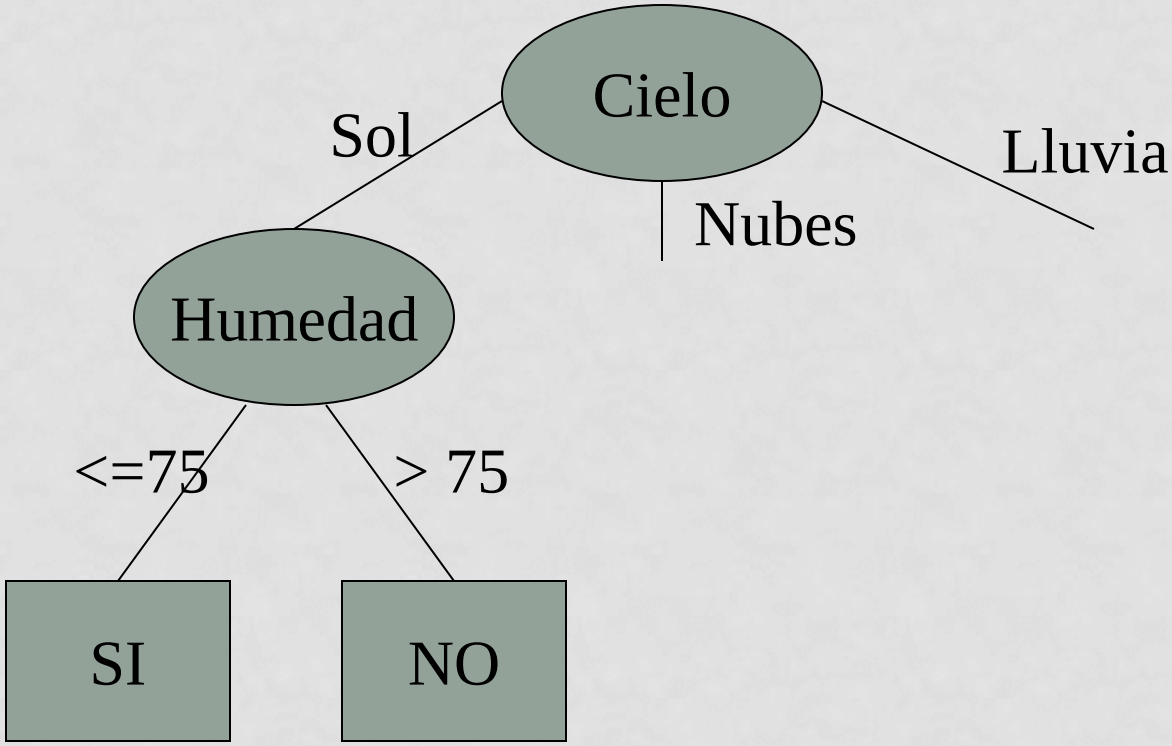
T	H	V	Ten
69	70	No	Si
75	70	Si	Si

T	H	V	Ten
72	95	No	No

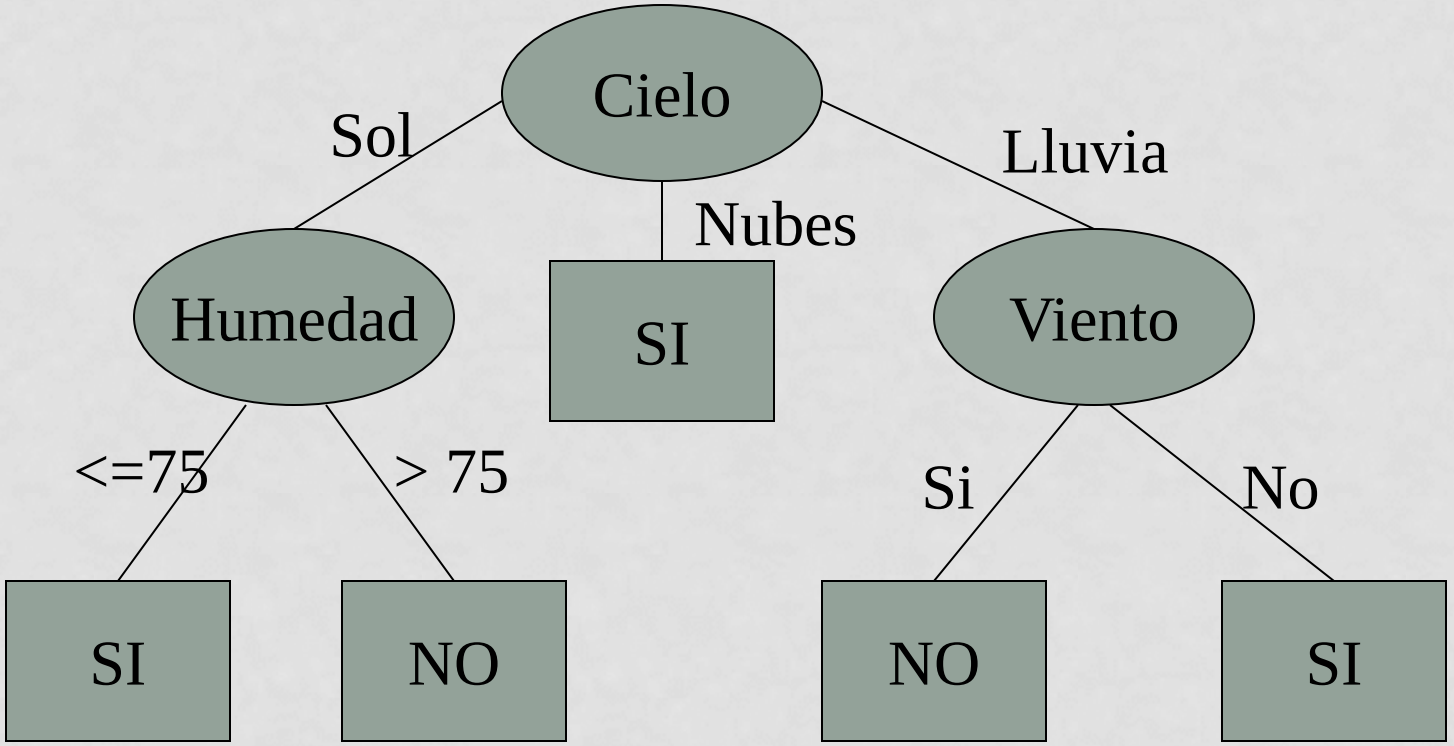
¿Porqué no seguir subdividiendo?

- ¿Hay que detener la construcción cuando tenemos “1 no, 2 si”?
- Tal vez. Cuando hay tan pocos datos (y suponiendo que haya ruido) es posible que el “1 no” haya aparecido por azar, e igualmente podriamos tener “2 no, 2 si”
- Pasamos de una situación en la que hay mayoría de “si” a otra en la que están equiparados con los “no”
- Se puede utilizar algún criterio estadístico para saber si es probable que “1 no, 2 si” se deba al azar
- Cuando se manejan pocos datos (3 en este caso), es bastante probable que las regularidades (humedad<=70 en ete caso) sean sólo aparentes y se deban al azar

Construcción recursiva del árbol

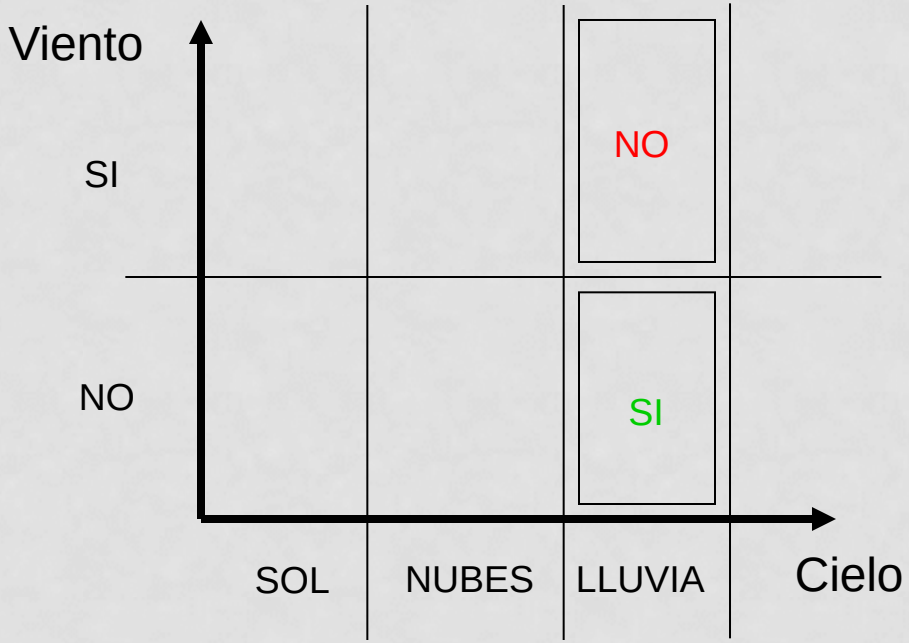


Construcción recursiva del árbol



C4.5 (J48) Tipo de clasificador

- Es no lineal
- Las fronteras de separación entre las clases son rectas paralelas a los ejes



Reglas (desde el árbol de decisión)

IF Cielo = Sol

Humedad <= 75 **THEN** Tenis = Si

ELSE IF Cielo = Sol

Humedad > 75 **THEN** Tenis = No

ELSE IF Cielo = Nubes **THEN** Tenis = Si

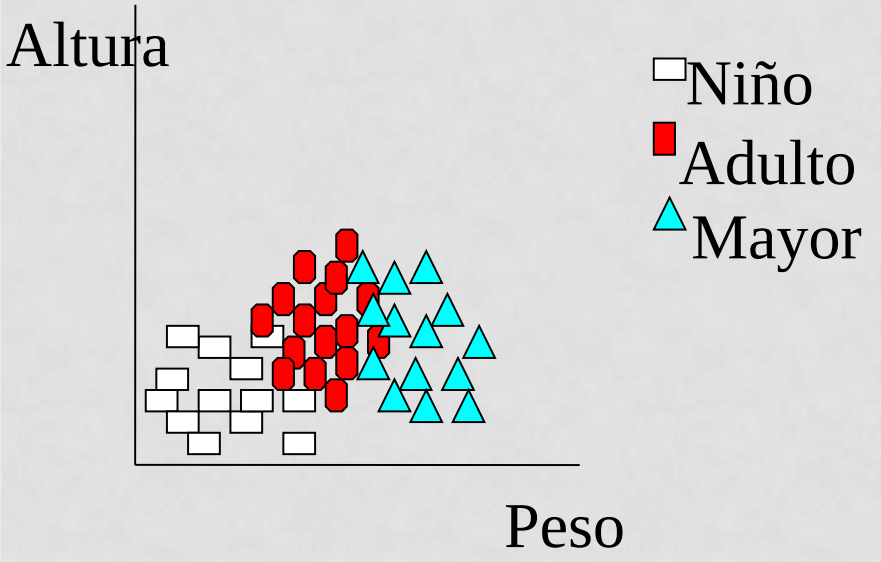
ELSE IF Cielo = Lluvia

Viento = Si **THEN** Tenis = Si

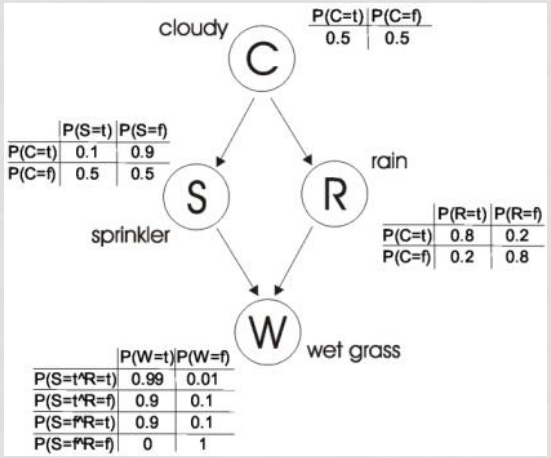
ELSE Tenis = No

OTROS MODELOS

Vecino mas cercano



Redes bayesianas



Basado en prototipos