Aprendizaje Automático

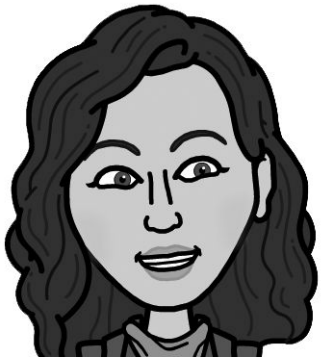
Árboles de Decisión

Árboles de Clasificación

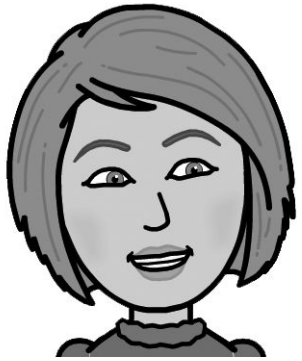
- En este capítulo estudiaremos algoritmos de aprendizaje de máquina que tratan de construir modelos predictivos usando sólo las características más informativas.
- En este contexto, una característica informativa es aquella que divide las instancias en conjuntos homogéneos respecto a la característica objetivo (target).

Brian

John



Aphra

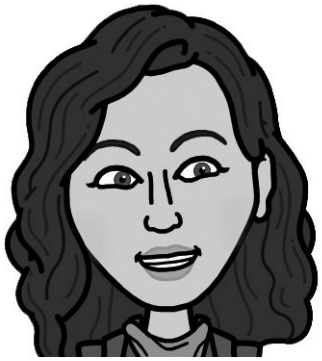


Aoife

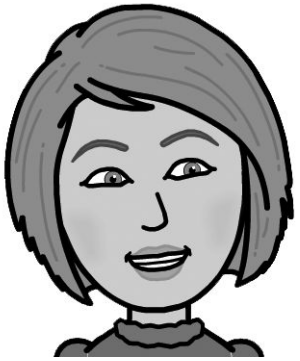
Usa lentes	Pelo largo	hombre	Nombre
Yes	No	Yes	Brian
no	No	yes	John
No	Yes	No	Aphra
No	No	No	Aoife

Brian

John

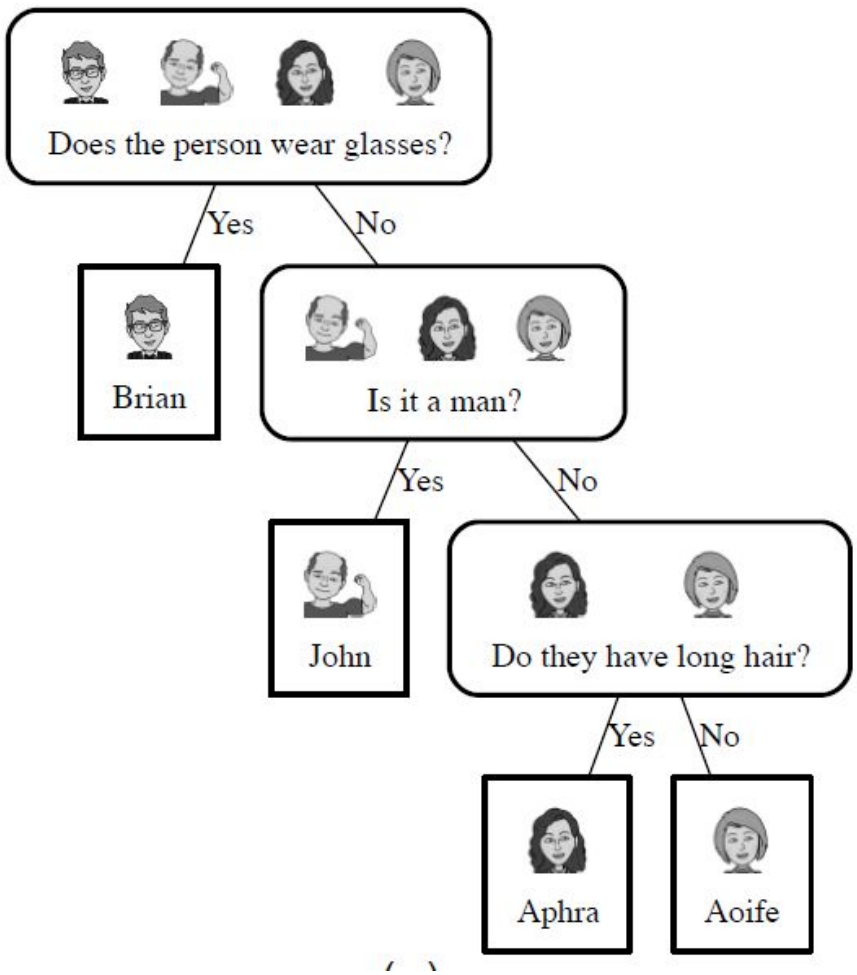


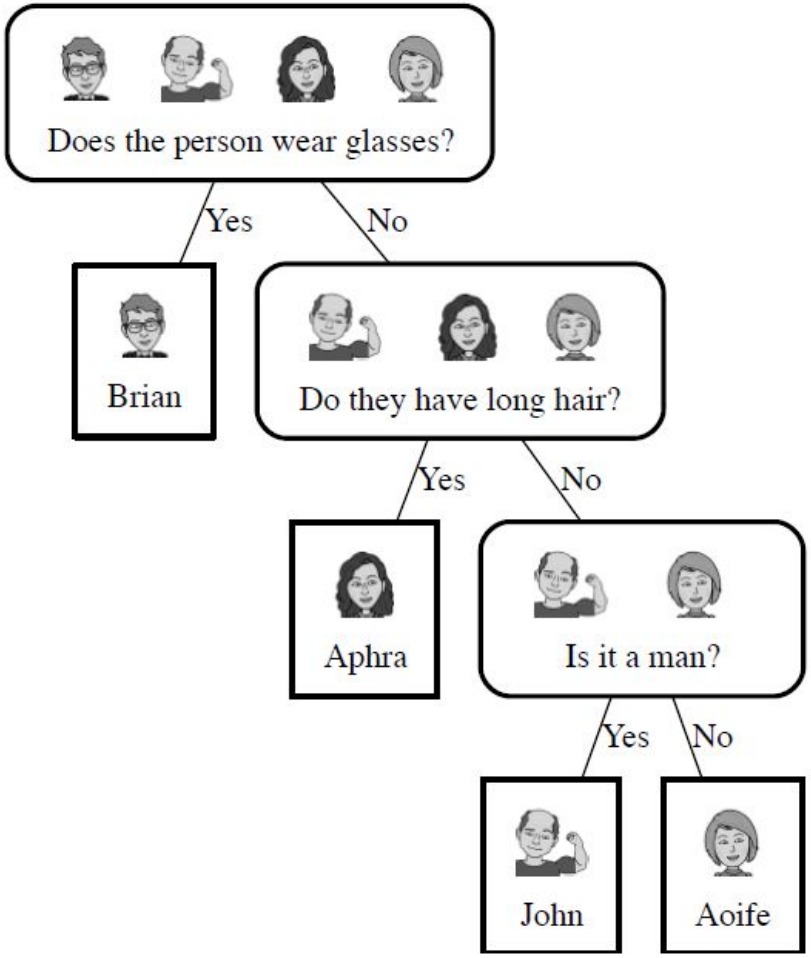
Aphra



Aoife

- Qué pregunta hariamos primero?
  1. Es hombre?
  2. La persona usa lentes?





Dos diferentes secuencias de preguntas que se forman cuando se inicia con la pregunta ¿**La persona usa lentes**?

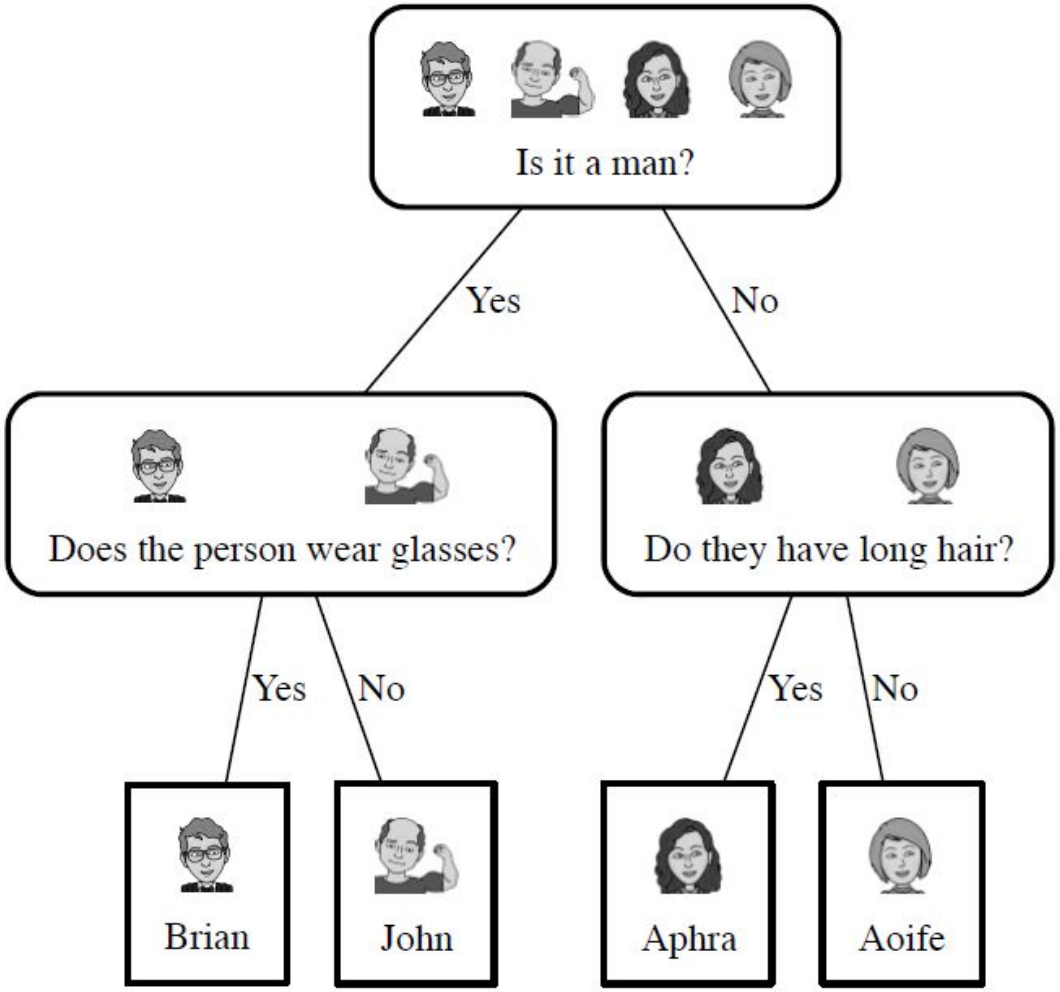
Caminos formados

En ambos diagramas tenemos:

- Un camino tiene una pregunta
- Un camino tiene dos preguntas
- Dos caminos tienen tres preguntas

Entonces, en promedio tenemos:

$(1+2+3+3)/4 = 2.25$



Secuencia de preguntas cuando se comienza con la pregunta:
¿**es hombre**?

Caminos formados

- Todos los caminos poseen dos preguntas
- Entonces, el número promedio de preguntas será:

  $(2+2+2+2)/4 = 2.0$

¿Cuál es la idea?

Definir qué características son las más informativas para responder a las preguntas, considerando los distintos efectos de las respuestas en función de:

1. cómo es dividido el dominio luego de cada respuesta
2. y la semejanza de cada respuesta

Árboles de Clasificación

- Entrada: Objetos caracterizables mediante propiedades.
- Salida:
  - En árboles de decisión: una decisión (sí o no).
  - En árboles de clasificación: una clase.
- Conjunto de reglas.

Árboles de Clasificación

- Se clasifican las instancias desde la raíz hacia las hojas, las cuales proveen la clasificación.
- Cada nodo especifica el test de algún atributo.
- Ejemplo: Si

(Outlook = Sunny, Temperature = Hot, Humidity = High, Wind = Strong)

Juego al tenis?

Problemas Apropiados

- Las instancias pueden ser representadas por pares (atributo, valor) .
- La función objetivo tiene valores discretos (o pueden ser discretizados).
- Pueden ser requeridas descripciones en forma de disjunción.
- Posiblemente existen errores en los datos de entrenamiento (robustos al ruido).
- Posiblemente falta información en algunos de los datos de entrenamiento.

Ejemplo de generación de árbol:

Problema juego de tenis

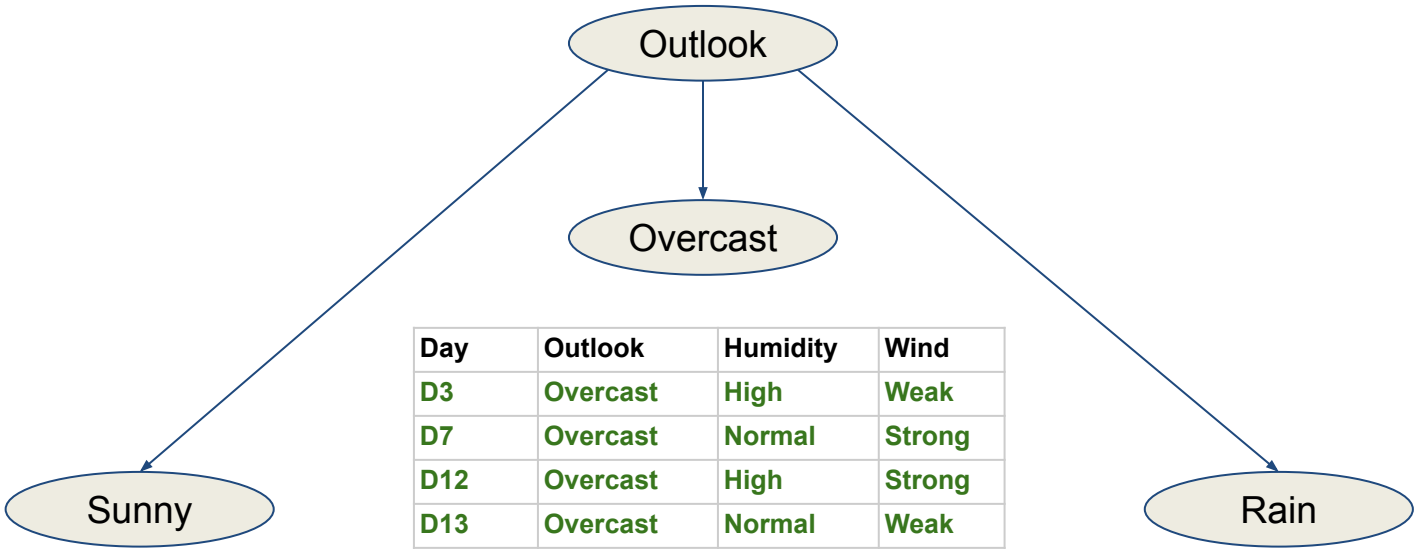
Ejemplos de entrenamiento 9/YES, 5/NO

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

Nuevo dato

D15	Rain	High	Weak	?

9/YES, 5/NO



Day	Outlook	Humidity	Wind
D3	Overcast	High	Weak
D7	Overcast	Normal	Strong
D12	Overcast	High	Strong
D13	Overcast	Normal	Weak

4/YES, 0/NO
Subconjunto Puro

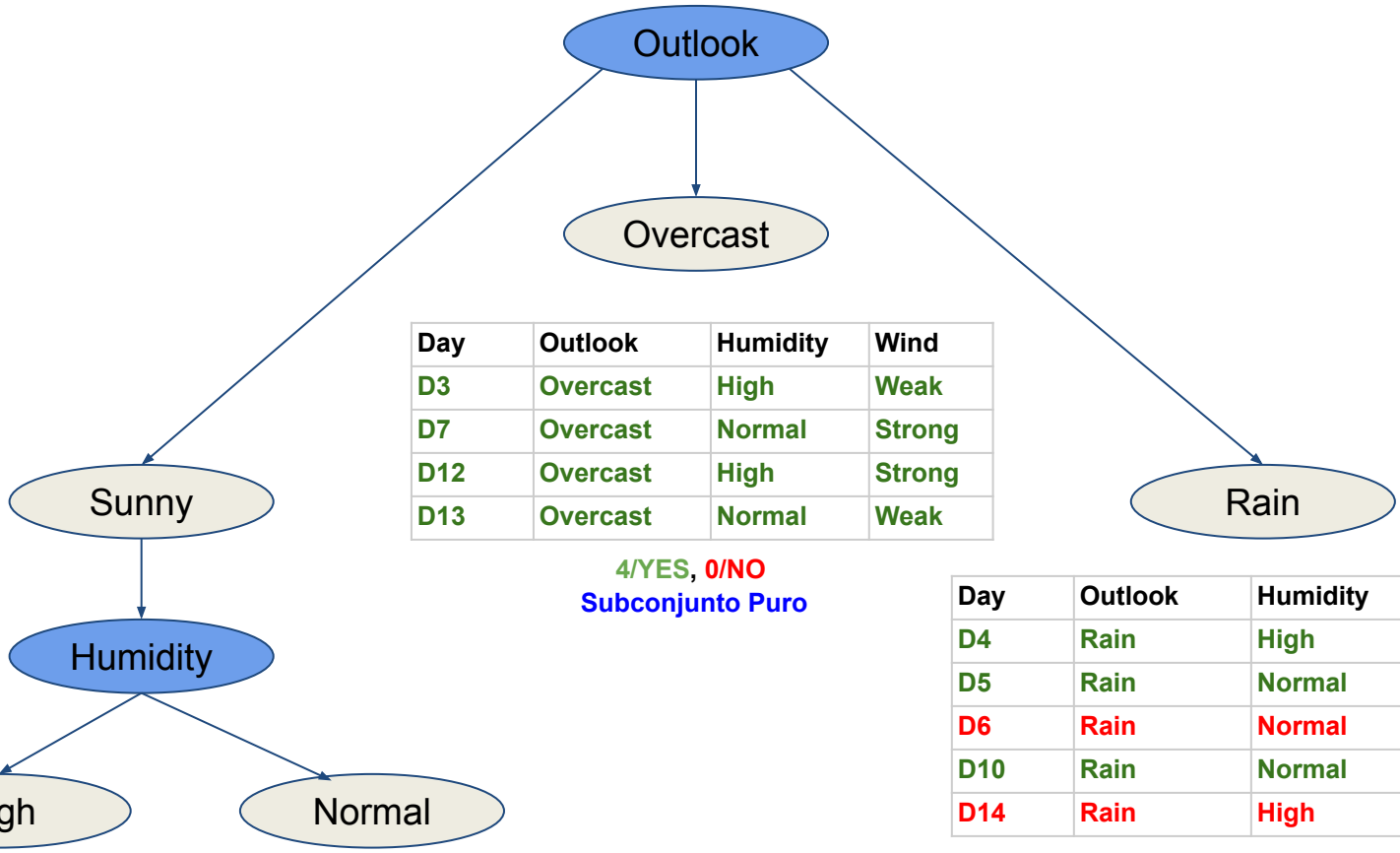
Day	Outlook	Humidity	Wind
D1	Sunny	High	Weak
D2	Sunny	High	Strong
D8	Sunny	High	Weak
D9	Sunny	Normal	Weak
D11	Sunny	Normal	Strong

2/YES, 3/NO
Para dividir

Day	Outlook	Humidity	Wind
D4	Rain	High	Weak
D5	Rain	Normal	Weak
D6	Rain	Normal	Strong
D10	Rain	Normal	Weak
D14	Rain	High	Strong

3/YES, 2/NO
Para dividir

9/YES, 5/NO



Day	Outlook	Humidity	Wind
D3	Overcast	High	Weak
D7	Overcast	Normal	Strong
D12	Overcast	High	Strong
D13	Overcast	Normal	Weak

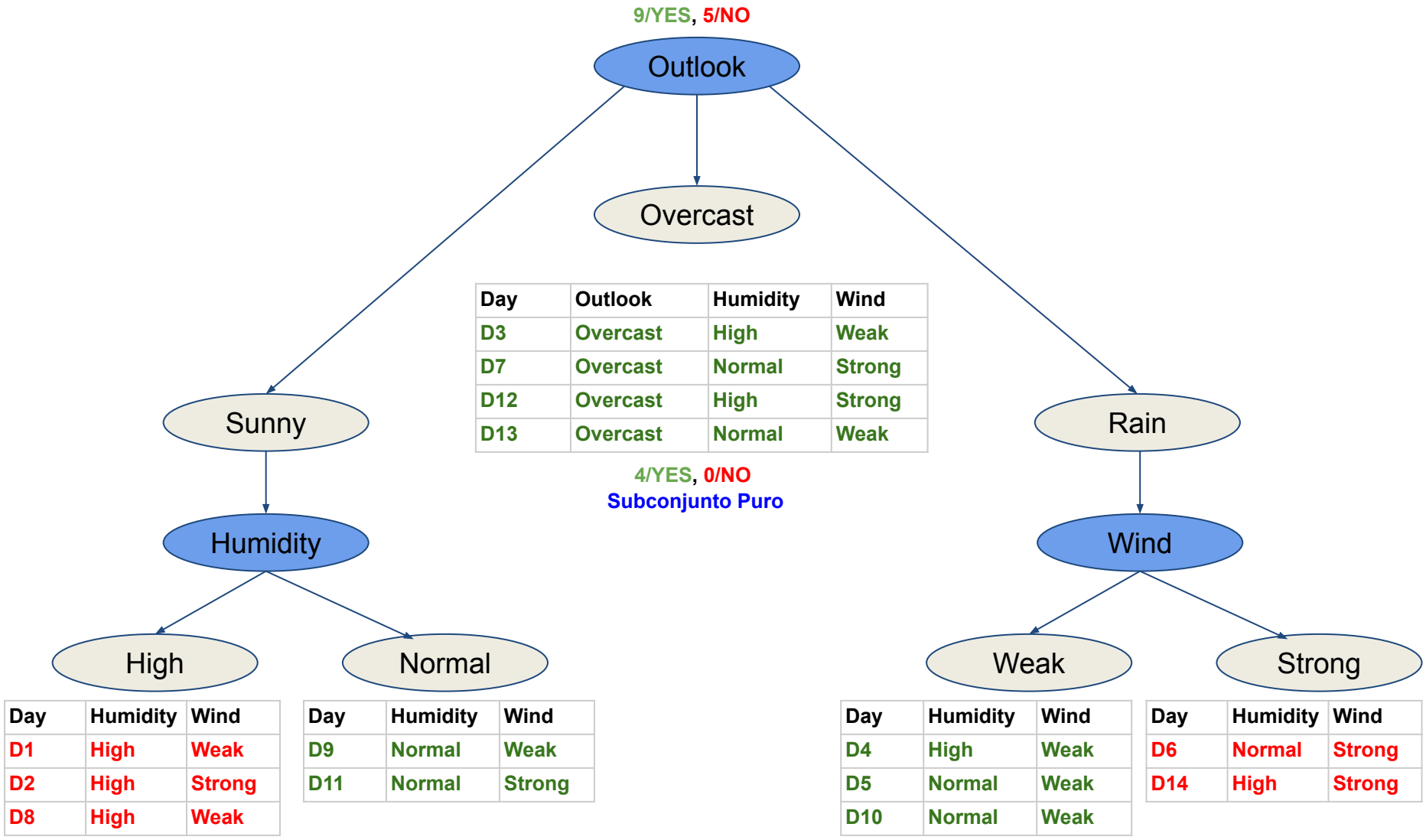
4/YES, 0/NO
Subconjunto Puro

Day	Outlook	Humidity	Wind
D4	Rain	High	Weak
D5	Rain	Normal	Weak
D6	Rain	Normal	Strong
D10	Rain	Normal	Weak
D14	Rain	High	Strong

3/YES, 2/NO
Para dividir

Day	Humidity	Wind
D1	High	Weak
D2	High	Strong
D8	High	Weak

Day	Humidity	Wind
D9	Normal	Weak
D11	Normal	Strong



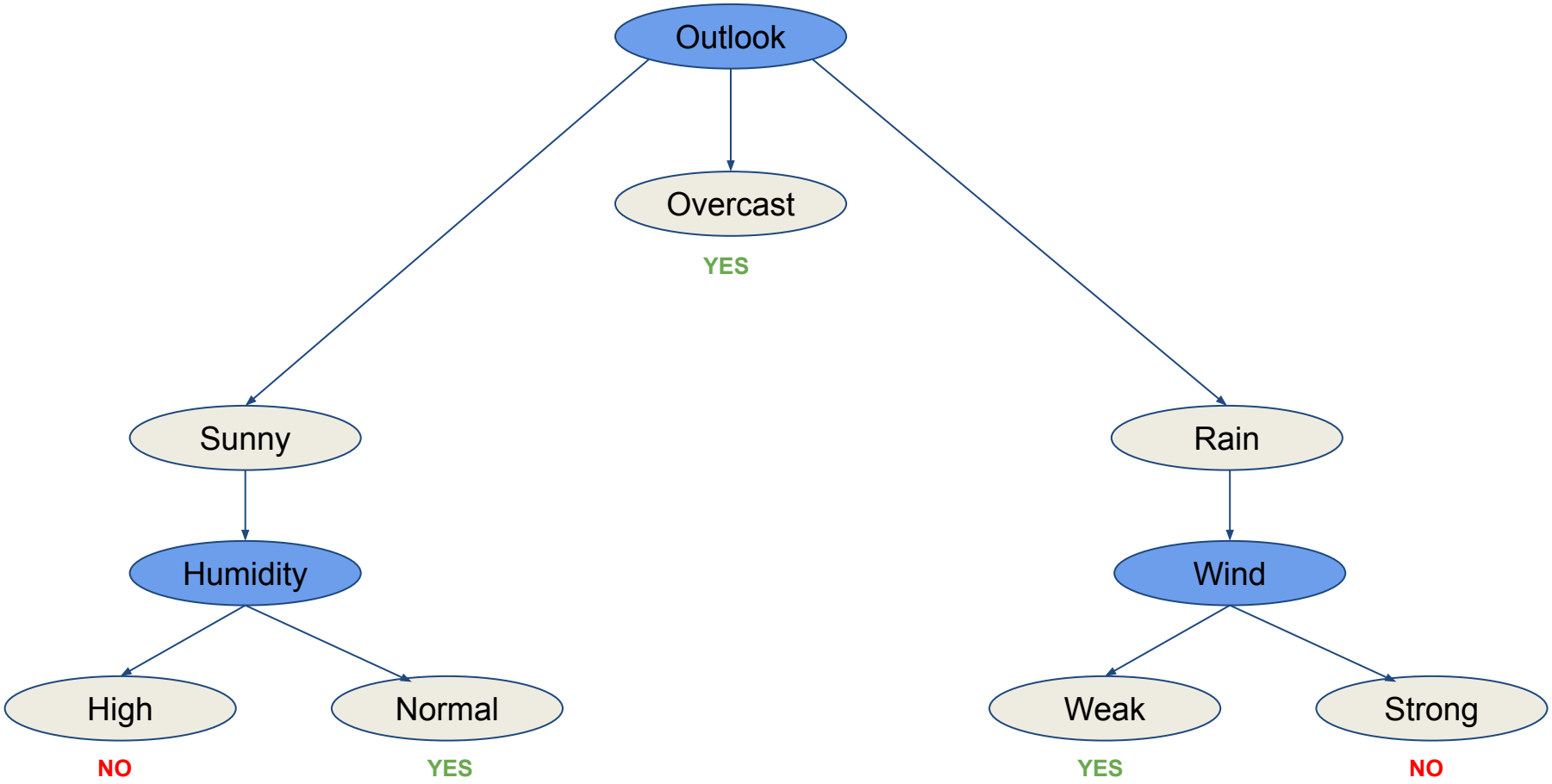
Day	Outlook	Humidity	Wind
D3	Overcast	High	Weak
D7	Overcast	Normal	Strong
D12	Overcast	High	Strong
D13	Overcast	Normal	Weak

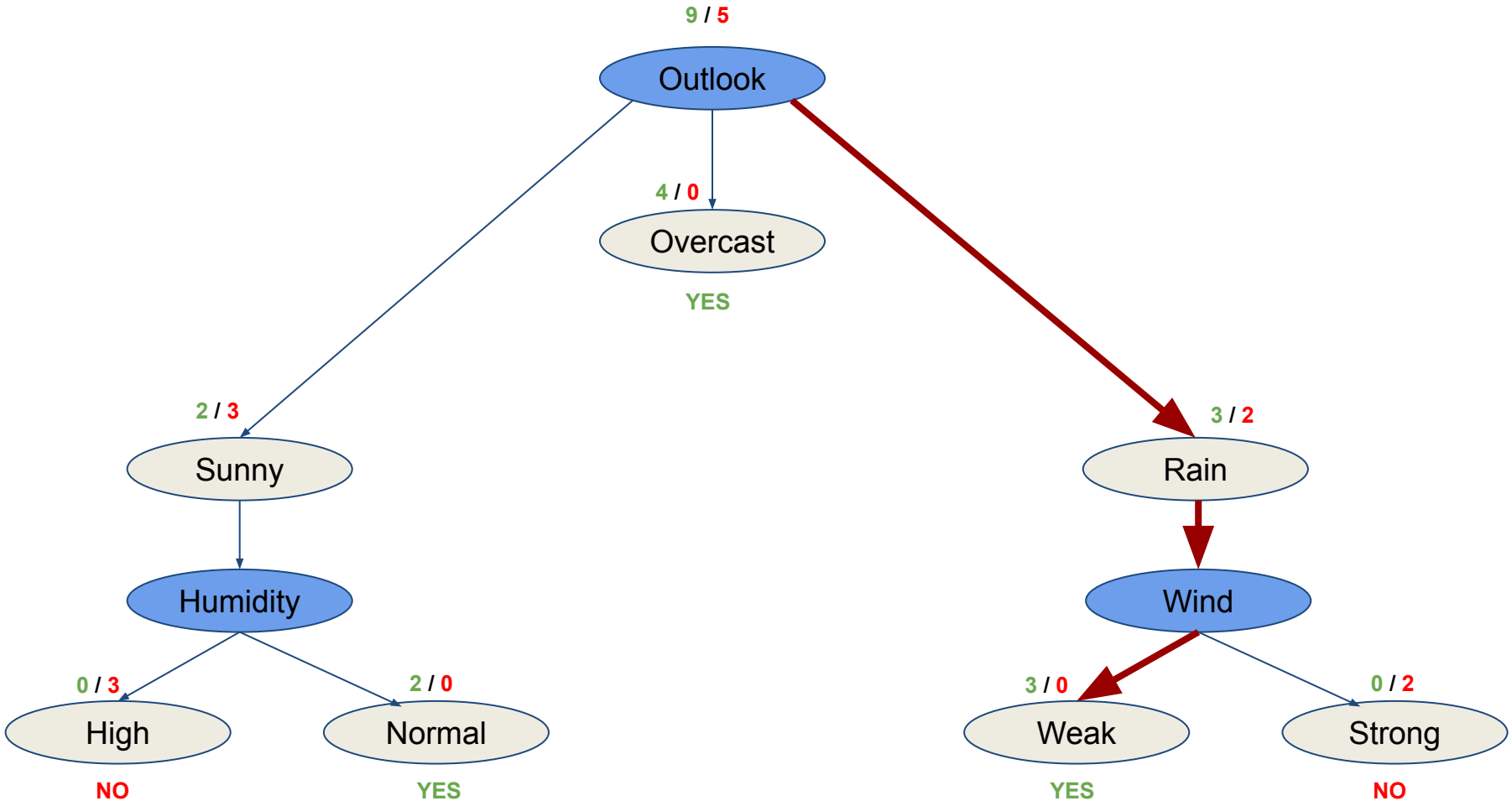
Day	Humidity	Wind
D1	High	Weak
D2	High	Strong
D8	High	Weak

Day	Humidity	Wind
D9	Normal	Weak
D11	Normal	Strong

Day	Humidity	Wind
D4	High	Weak
D5	Normal	Weak
D10	Normal	Weak

Day	Humidity	Wind
D6	Normal	Strong
D14	High	Strong


Outlook
Overcast
YES
Sunny
Rain
Humidity
Wind
High
Normal
Weak
Strong
NO
YES
YES
NO



Nuevo dato

Day	Outlook	Humidity	Wind	
D15	**Rain**	**High**	**Weak**	**?**

-> YES

Algoritmo básico para obtener un árbol de decisión

- Búsqueda exhaustiva, en profundidad (de arriba hacia abajo), a través del espacio de posibles árboles de decisión (ID3 y C4.5).
- Raíz: el atributo que mejor clasifica los datos

  Cuál atributo es el mejor clasificador?

  ⇒ respuesta basada en la **ganancia de información.**

Ganancia de información

- Mide la reducción esperada de entropía sabiendo el valor del atributo A

$$Gain(S,A) \equiv Entropía(S) - \sum_{v \in Valores(A)} (|S_v|/|S|)Entropía(S_v)$$

Valores(A): Conjunto de posibles valores del atributo A

S_v: Subconjunto de S en el cual el atributo A tiene el valor *v*

Algoritmos: ID3 (Interactive Dichotomizer Version 3)

- Entropía

$$Entropía(S) \equiv - p_{\oplus} \log_2 p_{\oplus} - p_{\Theta} \log_2 p_{\Theta}$$

$p_{\oplus}$ = proporción de ejemplos positivos.

p_{Θ} = proporción de ejemplos negativos.

S: conjunto de datos actual.

Por ejemplo, en el conjunto de datos Play Tennis

$p_{\oplus} = 9/14$, $p_{\Theta} = 5/14$ y $E(S) = 0.940$

En general: $Entropía(S) = - \sum_{i=1,c} p_i \log_2 p_i$

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

Algoritmos: ID3 (Interactive Dichotomizer Version 3)

- Por ejemplo:

Si S_1 es el subconjunto de S en el cual

Humidity = **High**

Entonces:

$p_{\oplus} = 3/7$

$p_{\ominus} = 4/7$

Entropía(S_1) = $-3/7 \log_2 3/7 - 4/7 \log_2 4/7$

= 0.985

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

Algoritmos: ID3 (Interactive Dichotomizer Version 3)

- Por ejemplo:

Si S_2 es el subconjunto de S en el cual

Humidity = **Normal**

Entonces:

$p_{\oplus} = 6/7$

$p_{\ominus} = 1/7$

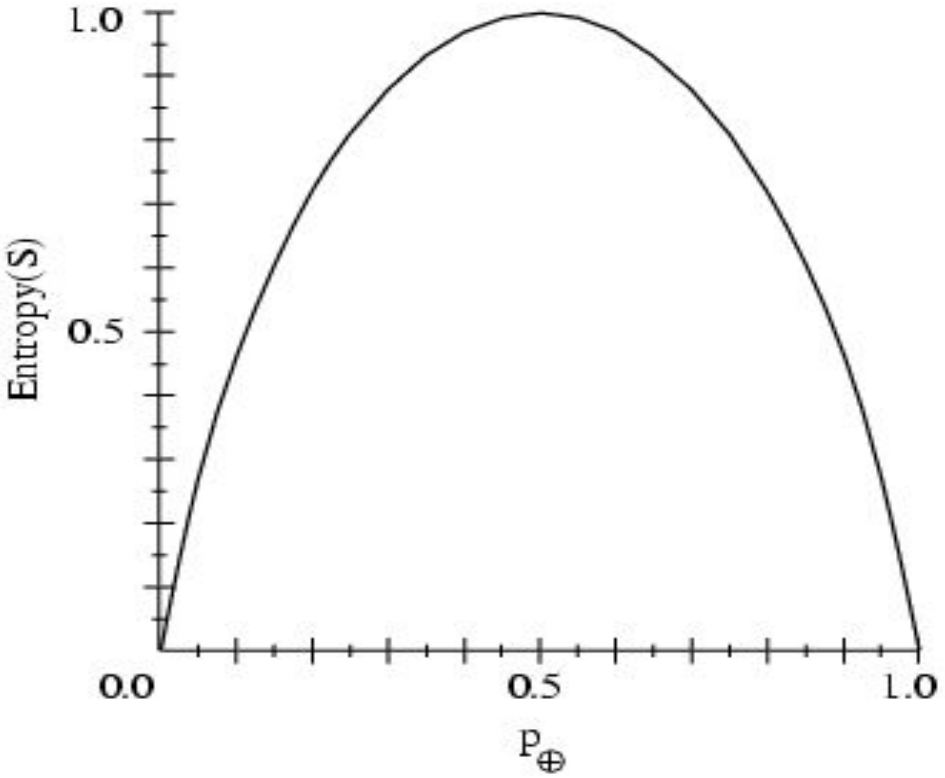
Entropía(S_2) = $-6/7 \log_2 6/7 - 1/7 \log_2 1/7$

= 0.592

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

Entropía y proporción de positivos

Cuando la Entropia es próxima a 1, se forman conjuntos homogéneos



Ganancia de información

- Mide la reducción esperada de entropía sabiendo el valor del atributo A

$$Gain(S,A) \equiv Entropía(S) - \sum_{v \in Valores(A)} (|S_v|/|S|) Entropía(S_v)$$

Valores(A): Conjunto de posibles valores del atributo A

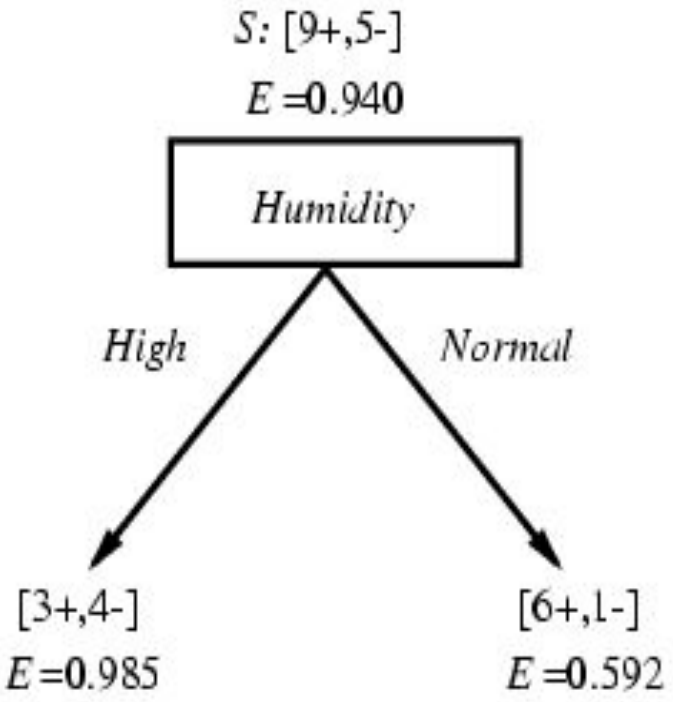
S_v: Subconjunto de S en el cual el atributo A tiene el valor *v*

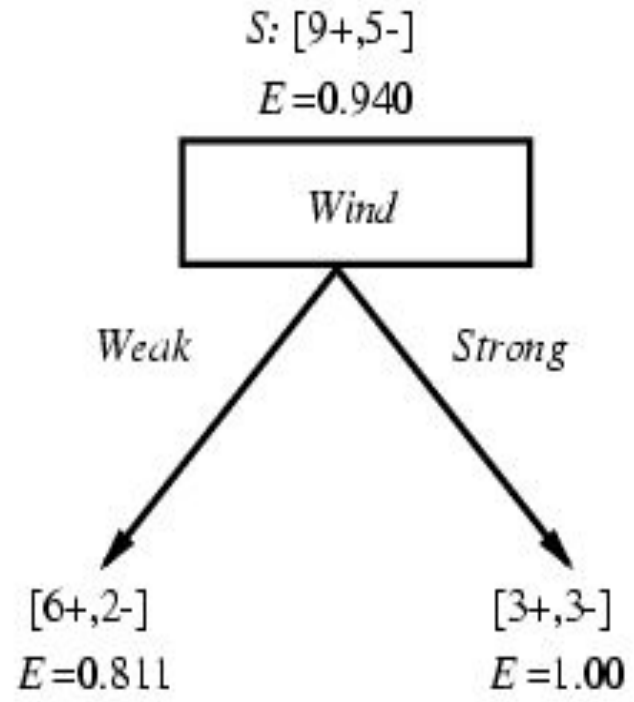
Ej: Gain(S, Humedad) = 0.940 - (7/14)0.985 - (7/14)0.592 = **0.151**

proporción de humedad **High**

prop. de humedad **normal**

Play Tennis





Gain (S, Humidity)

= .940 - (7/14).985 - (7/14).592
= .151

Gain (S, Wind)

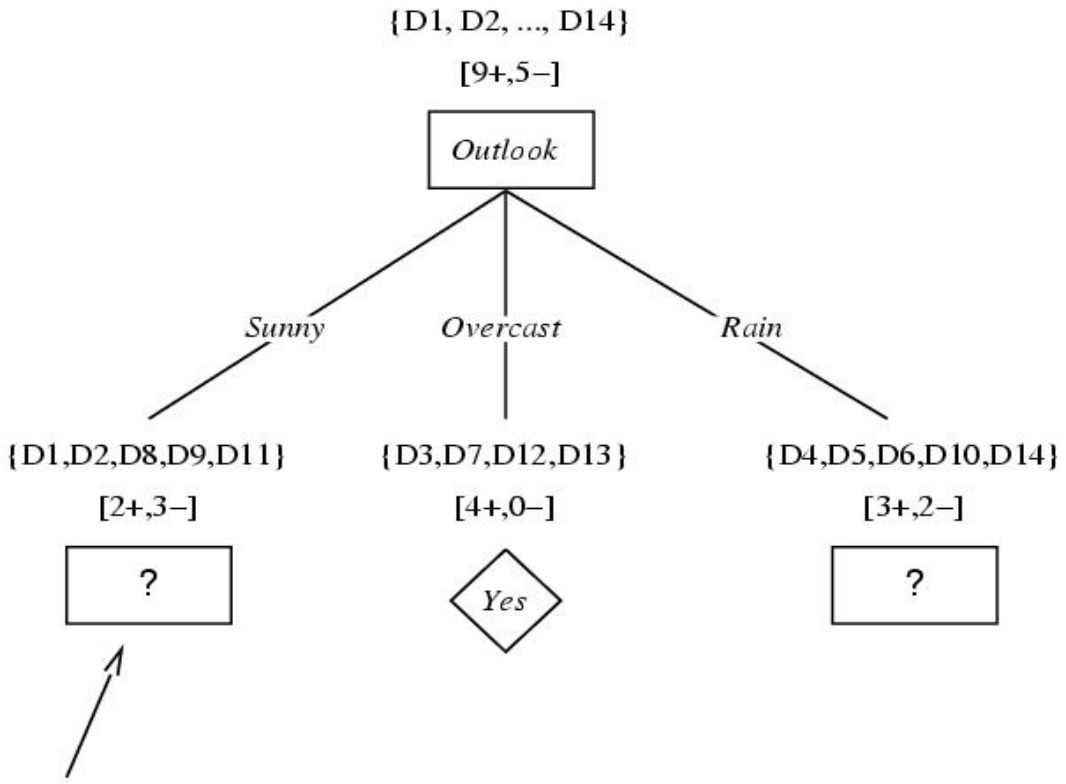
= .940 - (8/14).811 - (6/14)1.0
= .048

Play Tennis

- Gain(*S*,Outlook) = 0.246
- Gain(S,Humidity) = 0.151
- Gain(S,Wind) = 0.048
- Gain(S,Temperature) = 0.029

⇒ Outlook es el atributo del nodo raíz.

Play Tennis



Which attribute should be tested here?

S_{sunny} = {D1,D2,D8,D9,D11}

Gain (S_{sunny}, *Humidity*) = .970 − (3/5) 0.0 − (2/5) 0.0 = .970

Gain (S_{sunny}, *Temperature*) = .970 − (2/5) 0.0 − (2/5) 1.0 − (1/5) 0.0 = .570

Gain (S_{sunny}, *Wind*) = .970 − (2/5) 1.0 − (3/5) .918 = .019