# | Árboles de decisión

Microsoft

# Árboles de decisión

- Dada una base de datos se fabrican diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema.

# Ejemplo: Conjunto de datos

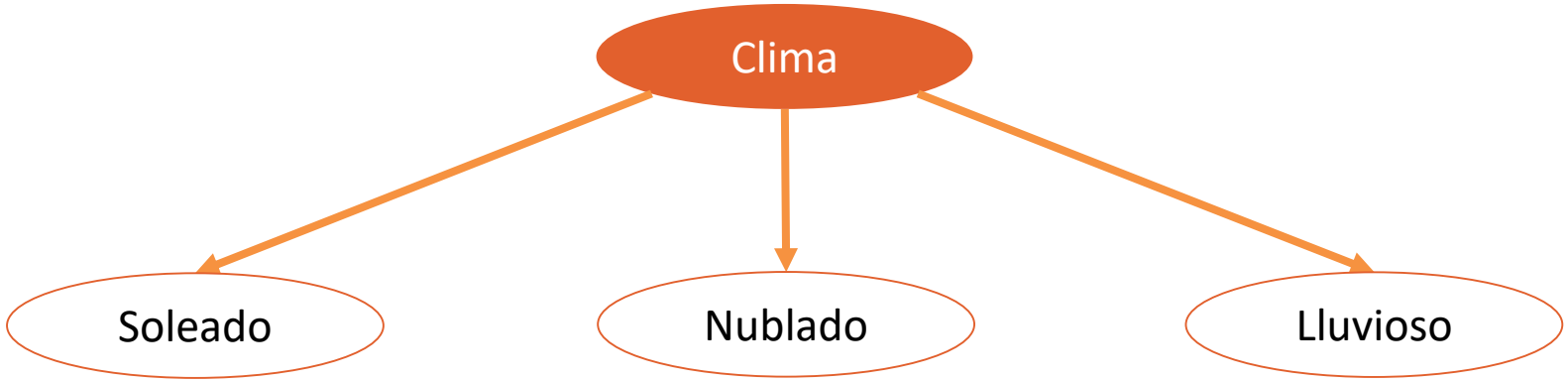
Ejemplos:
9 Sí / 5 No

| Día | Clima | Humedad | Viento | Jugó |
|---|---|---|---|---|
| D1 | Soleado | Alta | Poco | No |
| D2 | Soleado | Alta | Fuerte | No |
| D3 | Nublado | Alta | Poco | Sí |
| D4 | Lluvioso | Alta | Poco | Sí |
| D5 | Lluvioso | Normal | Poco | Sí |
| D6 | Lluvioso | Normal | Fuerte | No |
| D7 | Nublado | Normal | Fuerte | Sí |
| D8 | Soleado | Alta | Poco | No |
| D9 | Soleado | Normal | Poco | Sí |
| D10 | Lluvioso | Normal | Poco | Sí |
| D11 | Soleado | Normal | Fuerte | Sí |
| D12 | Nublado | Alta | Fuerte | Sí |
| D13 | Nublado | Normal | Poco | Sí |
| D14 | Lluvioso | Alta | Fuerte | No |

# Ejemplo: Definición del problema

- Si el día 15 está lloviendo, la humedad es alta y hay poco viento, ¿Juan jugará o no?
  - Es difícil adivinar simplemente viendo los datos
  - Podemos seguir una estrategia divide y vencerás
    - Dividir en subconjuntos
    - Si no hay incertidumbre en el subconjunto paramos, si no seguimos dividiendo
    - Verificamos en que subconjunto cae la pregunta que estamos realizando

# Ejemplo: Creación del árbol




| Día | Clima | Humedad | Viento |
|---|---|---|---|
| D1 | Soleado | Alta | Poco |
| D2 | Soleado | Alta | Fuerte |
| D8 | Soleado | Alta | Poco |
| D9 | Soleado | Normal | Poco |
| D11 | Soleado | Normal | Fuerte |

2 Sí / 3 No
Subdividir más

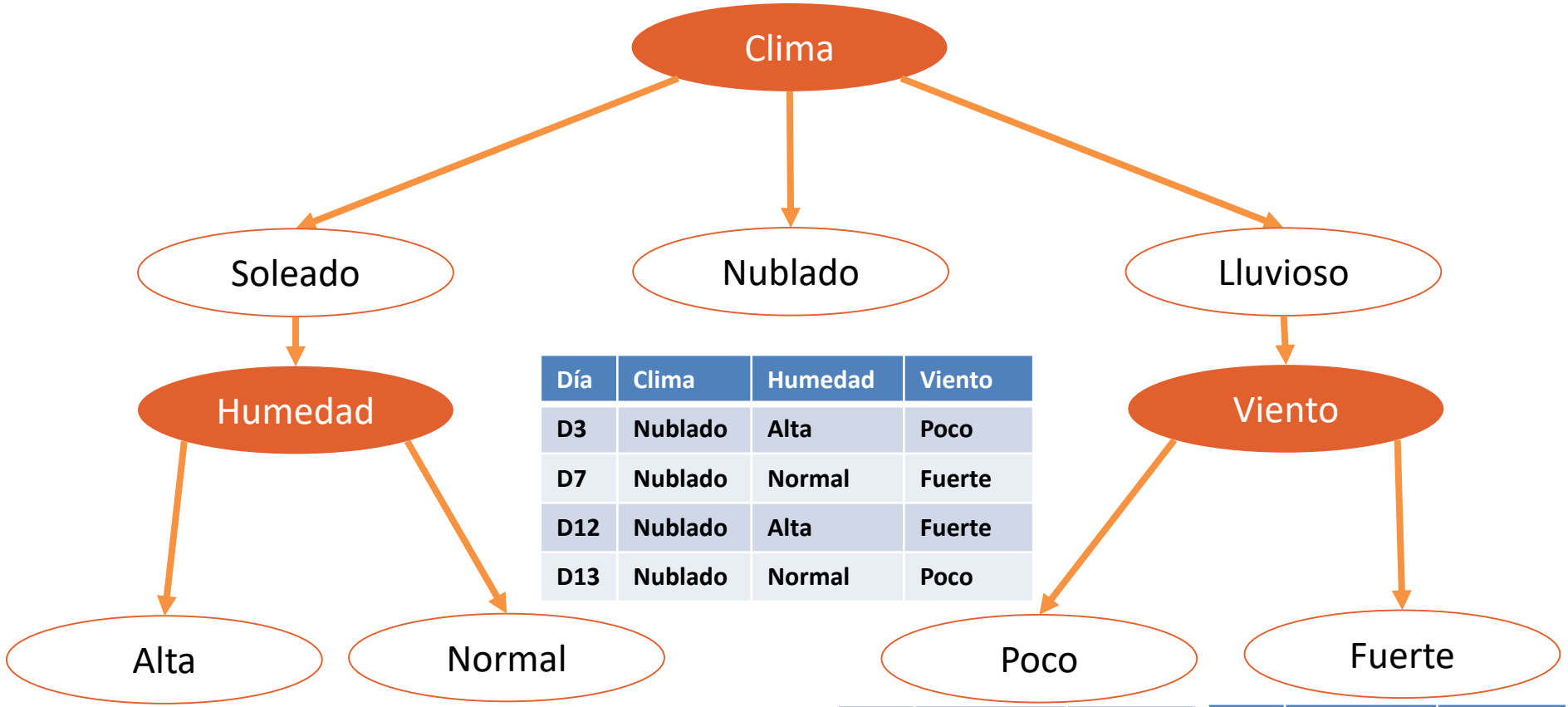
| Día | Clima | Humedad | Viento |
|---|---|---|---|
| D3 | Nublado | Alta | Poco |
| D7 | Nublado | Normal | Fuerte |
| D12 | Nublado | Alta | Fuerte |
| D13 | Nublado | Normal | Poco |

4 Sí / 0 No
Subconjunto puro

| Día | Clima | Humedad | Viento |
|---|---|---|---|
| D4 | Lluvioso | Alta | Poco |
| D5 | Lluvioso | Normal | Poco |
| D6 | Lluvioso | Normal | Fuerte |
| D10 | Lluvioso | Normal | Poco |
| D14 | Lluvioso | Alta | Fuerte |

3 Sí / 2 No
Subdividir más

# Ejemplo: Creación del árbol




| Día | Clima | Humedad | Viento |
|---|---|---|---|
| D3 | Nublado | Alta | Poco |
| D7 | Nublado | Normal | Fuerte |
| D12 | Nublado | Alta | Fuerte |
| D13 | Nublado | Normal | Poco |

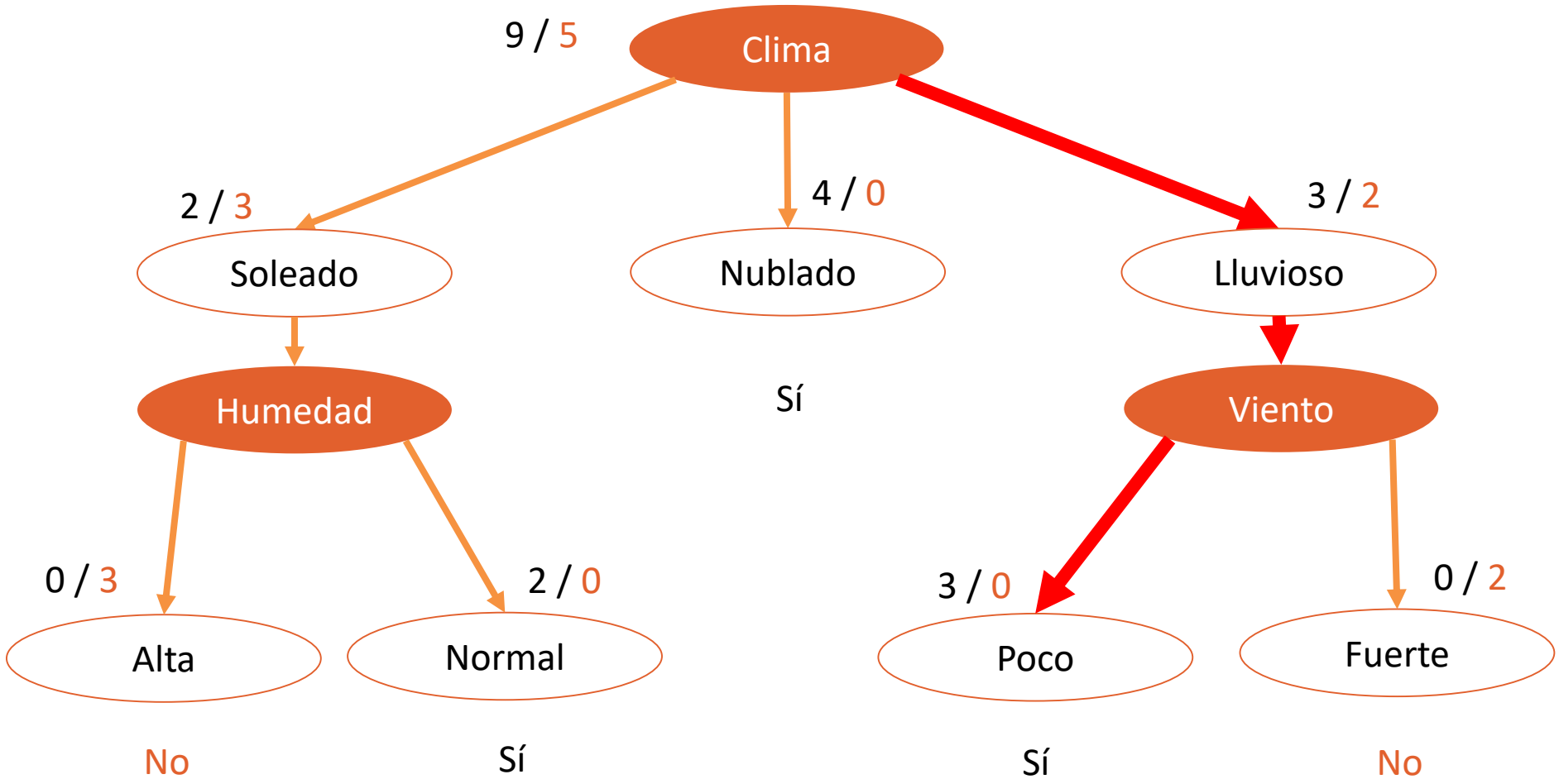
| Día | Humedad | Viento |
|---|---|---|
| D1 | Alta | Poco |
| D2 | Alta | Fuerte |
| D8 | Alta | Poco |

| Día | Humedad | Viento |
|---|---|---|
| D9 | Normal | Poco |
| D11 | Normal | Fuerte |

| Día | Humedad | Viento |
|---|---|---|
| D4 | Alta | Poco |
| D5 | Normal | Poco |
| D10 | Normal | Poco |

| Día | Humedad | Viento |
|---|---|---|
| D6 | Normal | Fuerte |
| D14 | Alta | Fuerte |

# Ejemplo: Árbol final



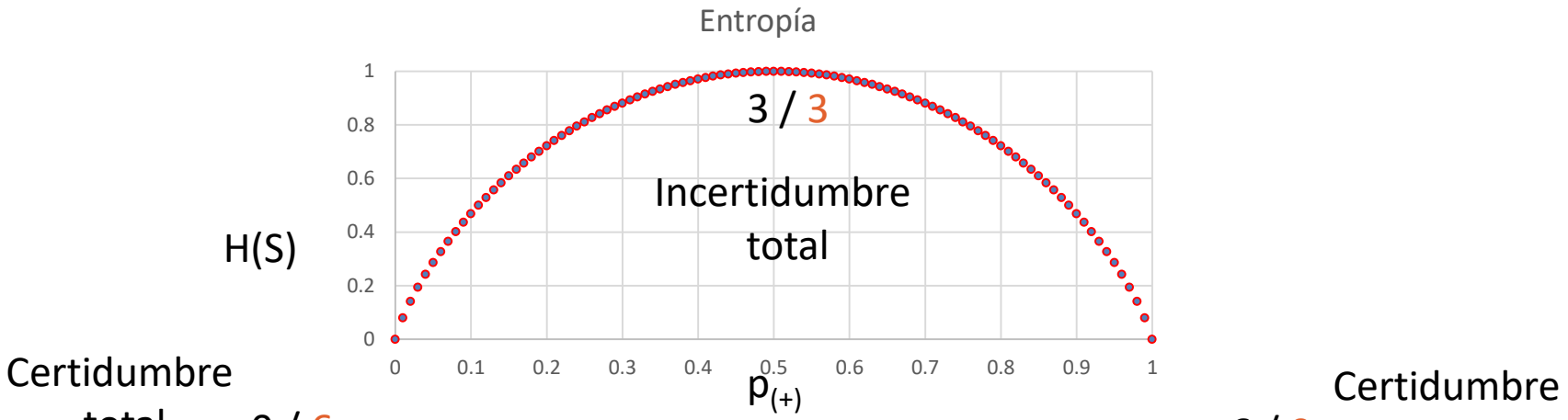

Juan si va a jugar

# Algoritmos

- Algoritmos
  - Ross Quinlan (ID3 Iterative Dichotomiser: 1986), (C4.5: 1993)
  - Breiman et al (CaRT Classification and Regression Trees: 1984)

- ID3
  - Divide(nodo, {ejemplos})
    - A ← el mejor atributo para dividir {ejemplos}
    - Atributo decisor de nodo ← A
    - Para cada valor de A crea un nodo hijo
    - Divide {ejemplos} para cada nodo hijo en subconjuntos
    - Para cada nodo hijo / subconjunto:
      - Si subconjunto es puro: termina
      - Si no: Divide(nodo_hijo,{ subconjunto })

# Entropía

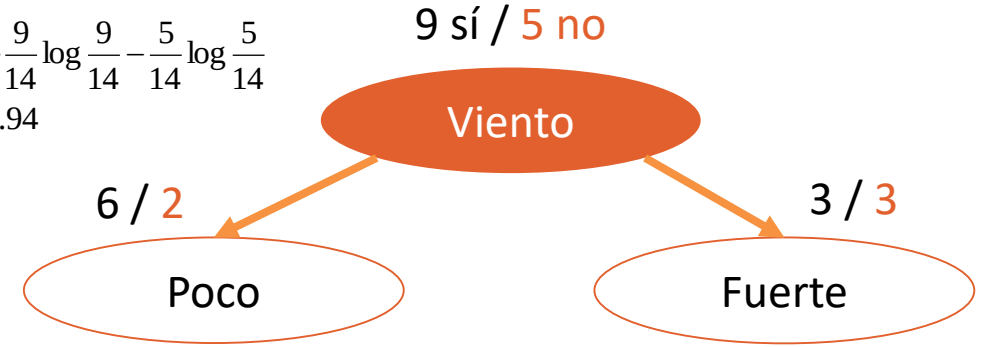

Entropía
H(S)
1
0.8
0.6
0.4
0.2
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
$p_{(+)}$
3 / 3
Incertidumbre total
Certidumbre total
0 / 6
6 / 0
Certidumbre total

# Ganancia

$$\text{Ganancia}(S, A) = H(S) - \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} H(S_v)$$

$$H(S) = -\frac{9}{14}\log\frac{9}{14} - \frac{5}{14}\log\frac{5}{14}$$

$$H(S) = 0.94$$




$$H(S_{poco}) = -\frac{6}{8}\log\frac{6}{8} - \frac{2}{8}\log\frac{2}{8}$$

$$H(S_{poco}) = 0.81$$

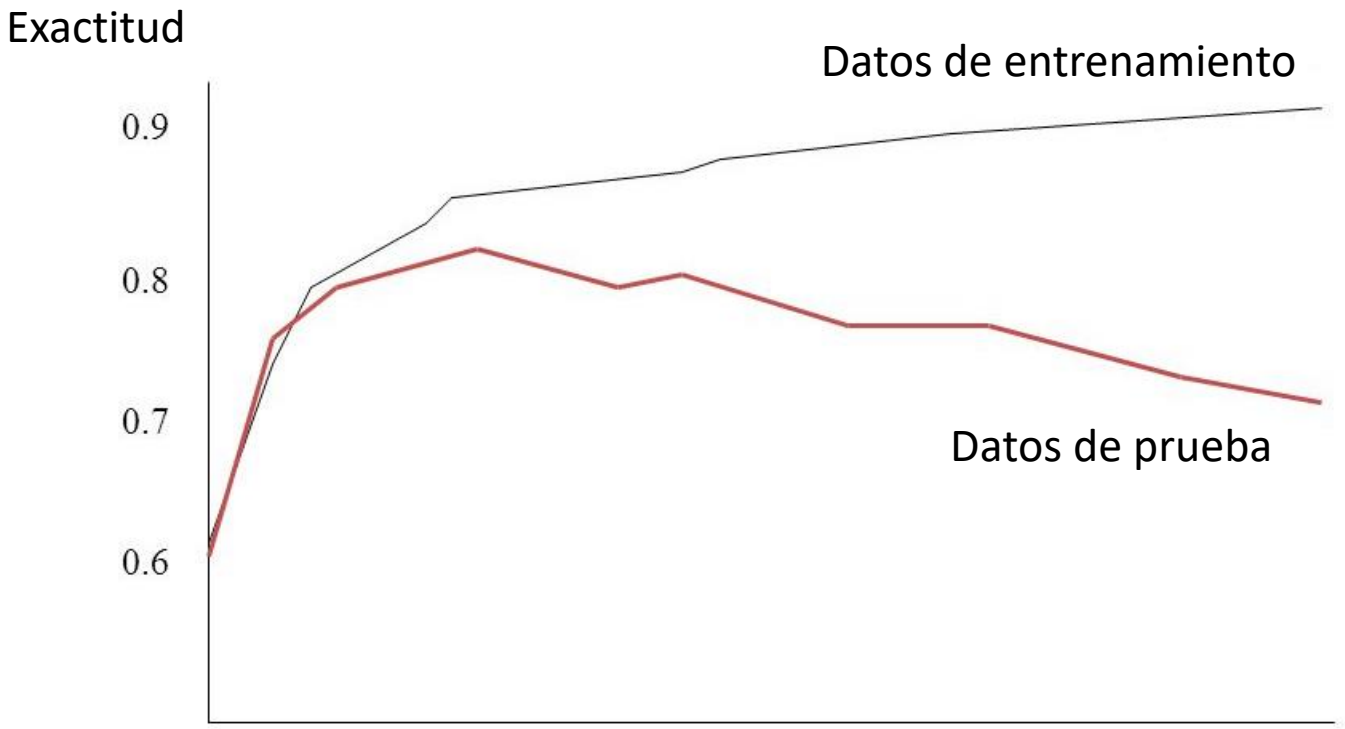
$$H(S_{fuerte}) = -\frac{3}{6}\log\frac{3}{6} - \frac{3}{6}\log\frac{3}{6}$$

$$H(S_{fuerte}) = 1.0$$

$$Ganancia(S, Viento) = H(S) - \frac{8}{14}H(S_{poco}) - \frac{6}{14}H(S_{fuerte})$$

$$Ganancia(S, Viento) = 0.049$$

# Sobre ajuste en árboles de decisión




Siempre pueden clasificar los datos de entrenamiento perfectamente

- Continua la división hasta tener un solo registro (singleton)
- Es bueno clasificando lo que ha visto en el pasado, pero no nuevos datos

# Sobre ajuste en árboles de decisión

- La solución es no dejar crecer mucho los árboles:
  - Parar la división cuando no sea estadísticamente significativa (no hay atributos predictivos cuya correlación con la clase sea significativa)
  - Dejar que crezca y después se poda
    - Basado en los datos de prueba
    - Para cada nodo
      - Probar como se comporta el árbol sin un nodo y todos sus hijos
      - Medimos desempeño con los datos de prueba
    - Quitamos el nodo que resulte en la mejora más grande
    - Repetimos hasta que la poda empeore el desempeño.

Microsoft