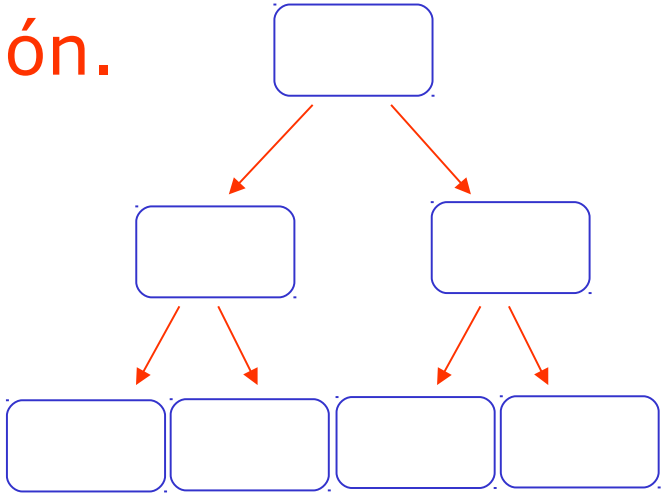
Inteligencia Artificial
IIC 2612

Un Agente Inductivo
Árboles de Decisión

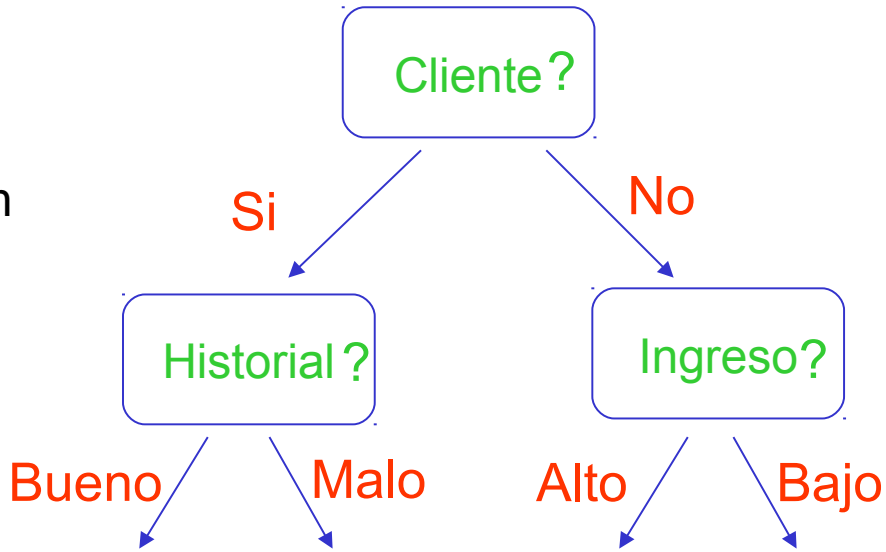
Árboles de decisión

- Es una técnica de clasificación.
- Como su nombre lo dice, consiste en un árbol.
- Por tanto contiene nodos y uniones dirigidas entre los nodos (links).
- Lo anterior permite una fácil visualización del clasificador.



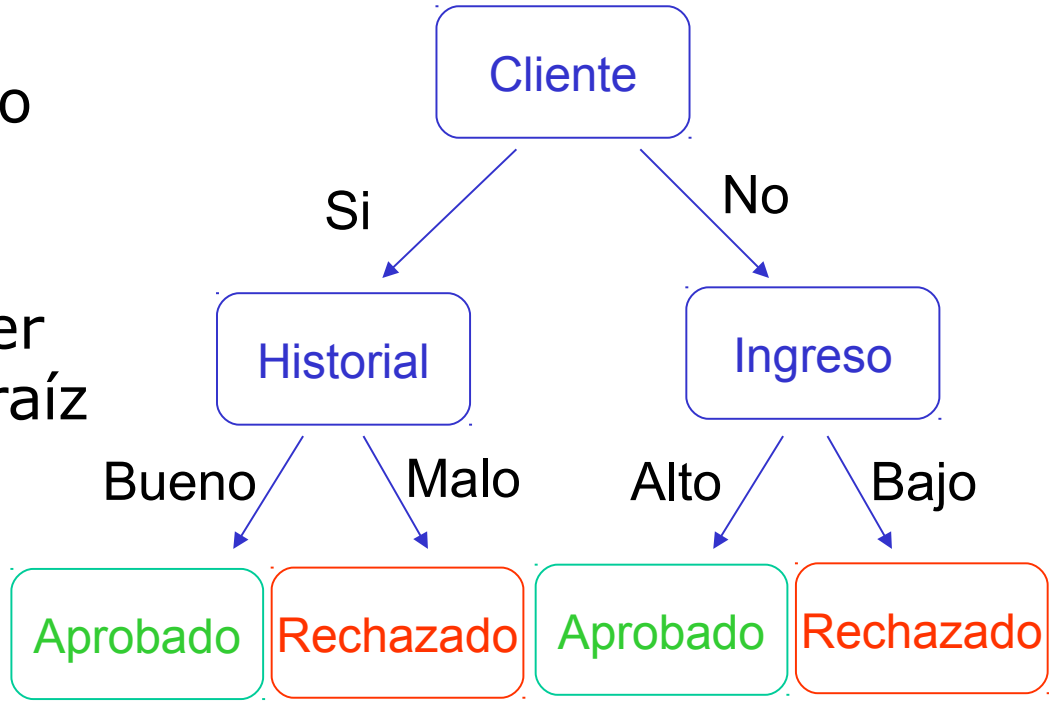
Árboles de decisión

- Cada nodo interno representa un atributo.
- En cada nodo interno se realiza un test basado en los valores del atributo.
- Los links representan el resultado del test.
- Un detalle importante es que el árbol sólo puede implementar decisiones discretas (categóricas), ¿Por qué?
- Por ende, en el caso de atributos continuos será necesario discretizarlos o utilizar un criterio discreto de decisión.



Árboles de decisión

- Los nodos hoja representan el resultado de la clasificación.
- Así, para clasificar un registro se debe recorrer el árbol desde el nodo raíz a la hoja resultante.
- El camino recorrido dependerá de los valores del registro.
- Ej. ¿Cuál es la clasificación para el crédito de un cliente con buen historial y alto ingreso?



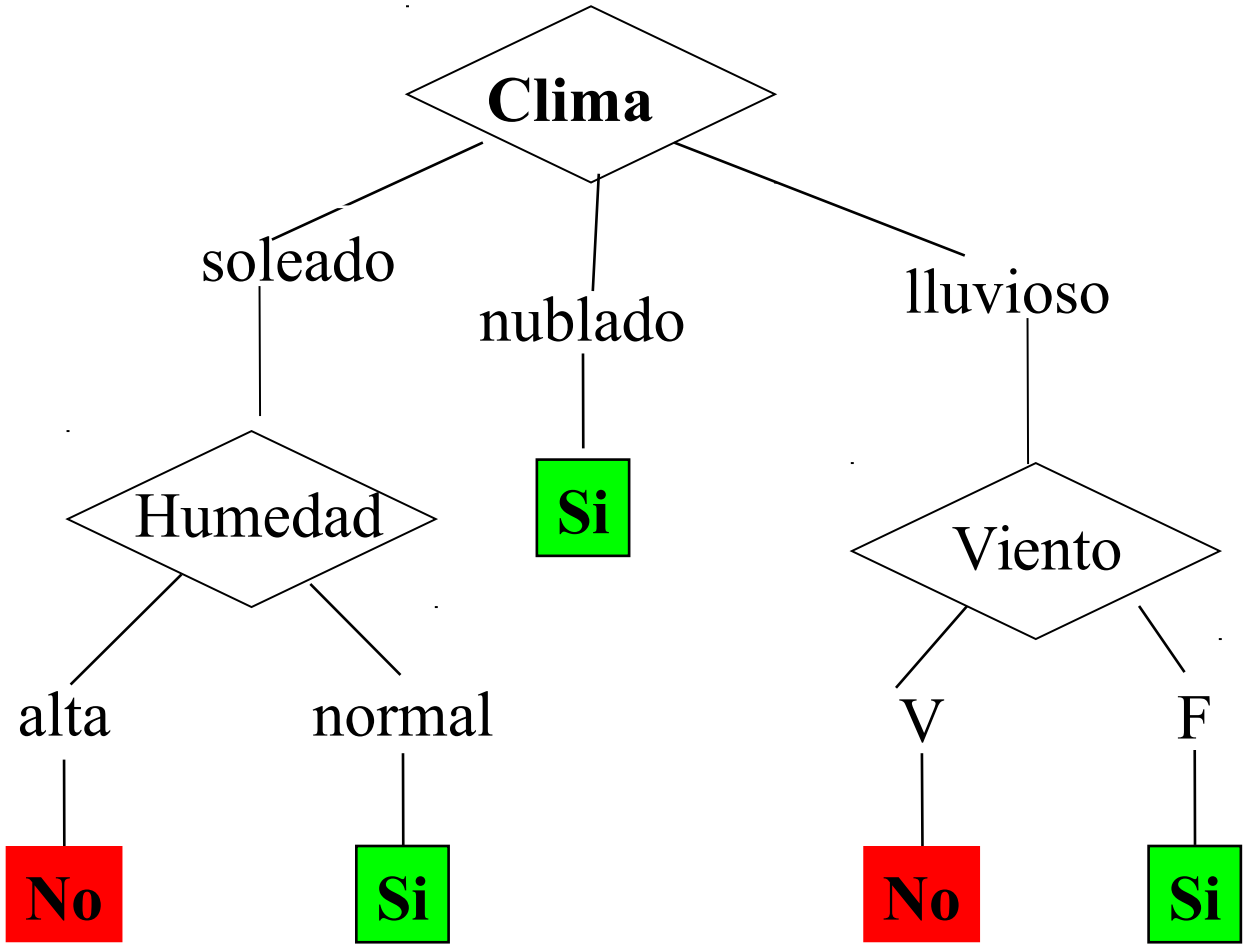
Árboles de decisión

- Los árboles de decisión son una técnica de aprendizaje supervisado.
- Por tanto, su entrenamiento necesita de un set de registros rotulados.
- Como vimos anteriormente, este set es conocido como set de entrenamiento.
- Este set de entrenamiento permite ajustar el modelo, i.e., encontrar una estructura apropiada para el árbol, i.e., explorar el espacio de hipótesis buscando un buen clasificador.

Ej. Jugar tenis?

Clima	Temperatura	Humedad	Viento	Jugar ?
soleado	alta	alta	F	No
soleado	alta	alta	V	No
nublado	alta	alta	F	Si
lluvioso	Agradable	alta	F	Si
lluvioso	frio	normal	F	Si
lluvioso	frio	normal	V	No
nublado	frio	normal	V	Si
soleado	Agradable	alta	F	No
soleado	frio	normal	F	Si
lluvioso	Agradable	normal	F	Si
soleado	Agradable	normal	V	Si
nublado	Agradable	alta	V	Si
nublado	alta	normal	F	Si
lluvioso	Agradable	alta	V	No

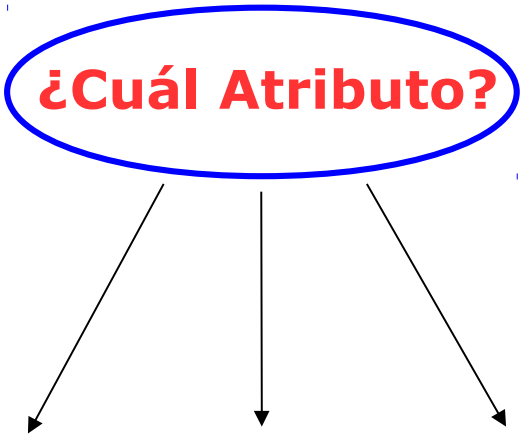
Ej. Jugar tenis?



¿Cómo determinar el árbol?

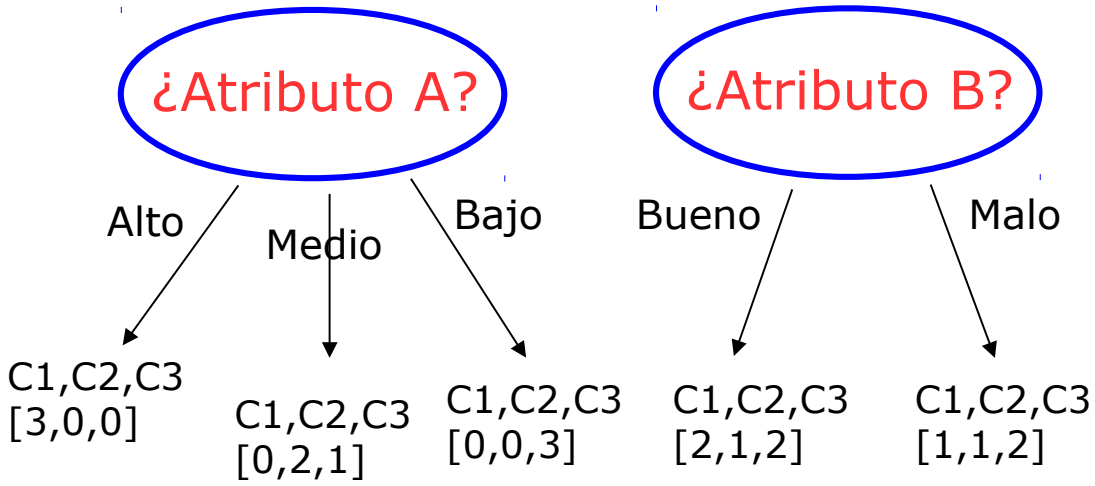
- La construcción de un árbol de decisión es incremental partiendo por el nodo raíz.
- Por tanto, el desafío inicial es decidir qué atributo utilizar para el nodo raíz, alguna idea?:

Clima	Temperatura	Humedad	Viento	Jugar?
soleado	alta	alta	F	No
soleado	alta	alta	V	No
nublado	alta	alta	F	Si
lluvioso	Agradable	alta	F	Si
lluvioso	frio	normal	F	Si
lluvioso	frio	normal	V	No
nublado	frio	normal	V	Si
soleado	Agradable	alta	F	No
soleado	frio	normal	F	Si
lluvioso	Agradable	normal	F	Si
soleado	Agradable	normal	V	Si
nublado	Agradable	alta	V	Si
nublado	alta	normal	F	Si
lluvioso	Agradable	alta	V	No



¿Cuál atributo?

Atributo A	Atributo B	Clase
Alto	Bueno	C1
Alto	Malo	C1
Bajo	Malo	C3
Medio	Malo	C2
Alto	Bueno	C1
Bajo	Malo	C3
Bajo	Bueno	C3
Medio	Bueno	C3
Medio	Bueno	C2

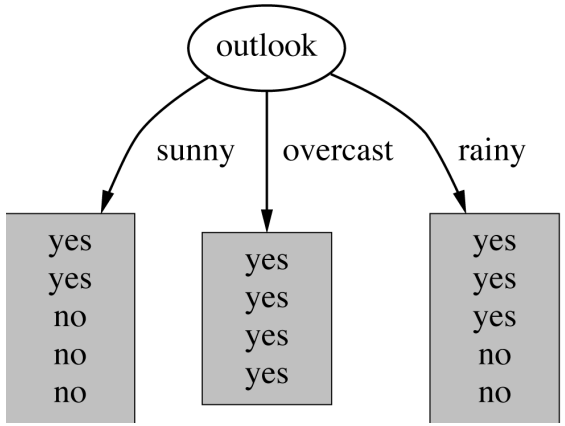


El algoritmo de construcción selecciona el atributo que mejor separa los registros de acuerdo al valor de las clases, i.e., el atributo más discriminativo.

¿Cómo determinar el árbol?

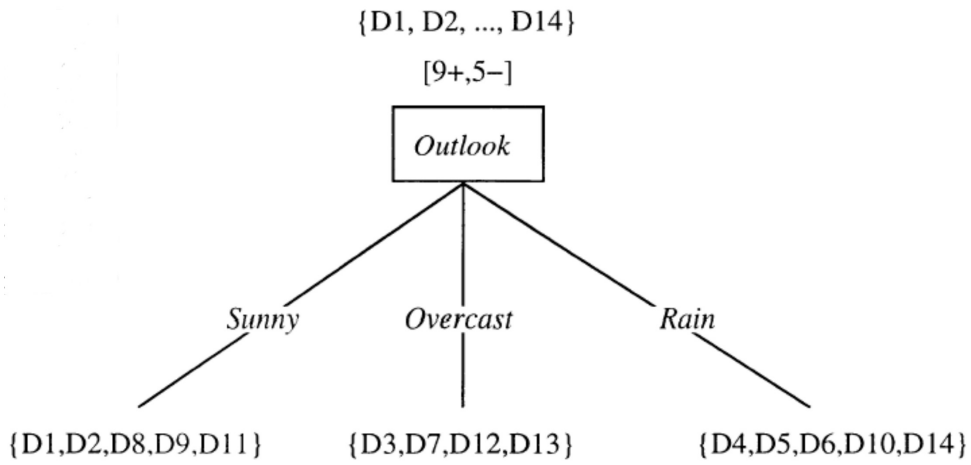
- Más adelante veremos un criterio matemático para determinar el atributo más discriminativo.
- Por ahora, asumamos que para el set de datos de la figura, el atributo más discriminativo es: "outlook".

Day	Outlook	Temperature	Humidity	Wind	PlayTenn
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



¿Cómo determinar el árbol?

Day	Outlook	Temperature	Humidity	Wind	PlayTenn
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



- Al analizar las ramas resultantes es posible apreciar que cada una de ellas recibe un subconjuntos de los datos originales.

- ¿Cómo podriamos seguir construyecto el árbol?

Algoritmo resultante:

- Pertenecen todos los registros a la misma clase?
  - Si → Retornar marcando el nodo hoja con la clase respectiva.
- Tienen todos los registros el mismo valor para todos los atributos que determinan su clase.
  - Si → Retornar marcando nodo hoja con la clase más común.
- De lo contrario:
  - Seleccionar el atributo que "*mejor*" separa los registros de las distintas clases.
  - Usar ese atributo como nodo raíz.
  - Dividir el set de entrenamiento de acuerdo a este atributo y para cada rama resultante continuar la construcción del árbol en forma recursiva.

¿Atributo más discriminativo?

- Cada atributo separa los datos en subgrupos.
- Idealmente cada subgrupo debe ser lo más homogéneo posible respecto a la clase.
- Por tanto, se necesita una métrica de homogeneidad de cada subgrupo.
- Ejemplo:
  - 2 clases: +/-
  - 100 registros (50+ y 50-)
  - A y B son dos atributos binarios.
    - Registros con A=0: 48+, 2-
Registros con A=1: 2+, 48-
    - Registros con B=0: 26+, 24-
Registros con B=1: 24+, 26-
  - Separar usando A es mejor que separar usando B.
    - A produce una mejor separación de los ejemplos **+** y **-**
    - B produce una pobre separación de los ejemplos **+** y **-**

Entropía

- Entropía es una buena manera de medir homogeneidad.
- La entropía mide el número de bits promedio que se necesita para codificar en forma óptima un conjunto de datos.
- Breve paréntesis de teoría de la información.

( )

Entropía

- Entropía:

$$H(S) = -\sum_{c_i} p_i \log_2 p_i$$

  - H(S) es la entropía del set S
  - c_i son las posibles clases
  - pi = fracción de registros de S que posee la clase C_i

- Ejemplo de entropía:
  - 3 clases (A,B,C)
  - A ocurre en la mitad de los ejemplos
  - B y C ocurren en un ¼ de los ejemplos
  - Codificación óptima: A = 0, B = 10, C = 11
  - Entropía = número de bits promedio/registro = 1.5 bits

Entropía ejemplos

- Ejemplo 1:
  - 10 registros con clase A
  - 20 registros con clase B
  - 30 registros con clase C
  - 40 registros con clase D
  - Entropía= -[(.1 log .1) + (.2 log .2) + (.3 log .3) + (.4 log .4)]
  - Entropía = 1.85

- Ejemplo 2:
  - 2 clases: +/-
  - 100 registros (50+ y 50-)
  - A y B son dos atributos binarios
    - Registros con A=0: 48+, 2- Entropía=?
    - Registros con A=1: 2+, 48- Entropía=?

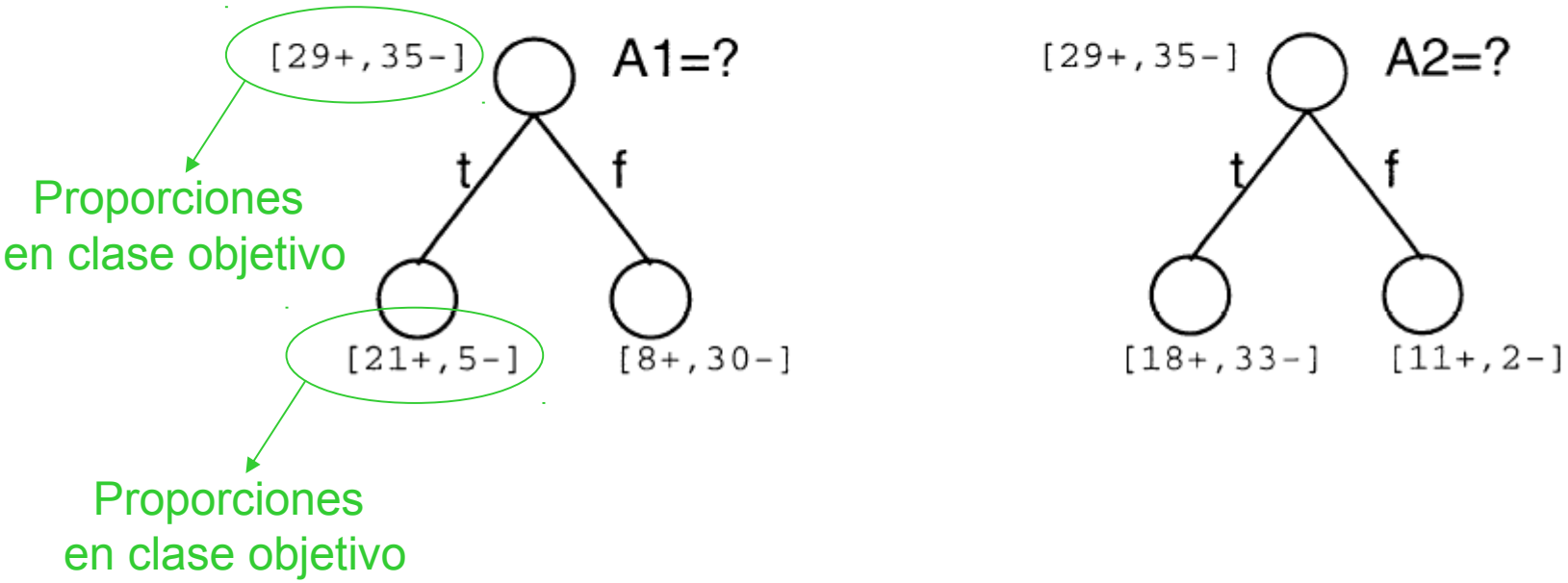
    - Registros con B=0: 26+, 24- Entropía=?
    - Registros con B=1: 24+, 26- Entropía=?

  - Atributo A es mejor pues entropía promedio de los subgrupos resultantes es menor.

Ganancia de información

- La ganancia de información es la reducción esperada en entropía al separar según cierto atributo, digamos A:

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

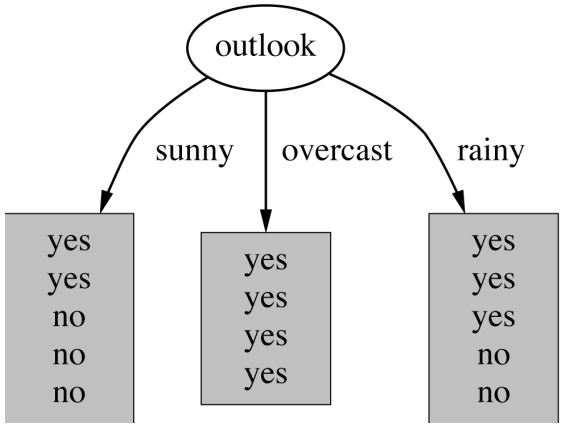


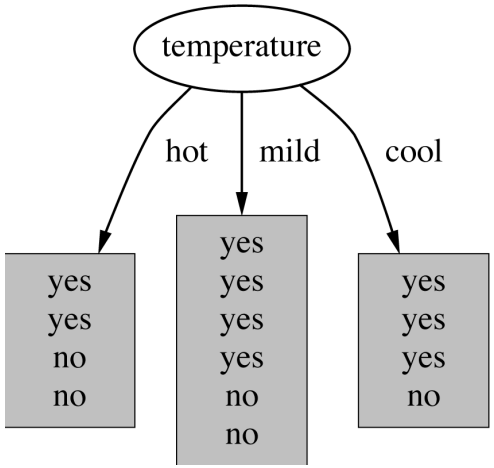
Ganancia de información

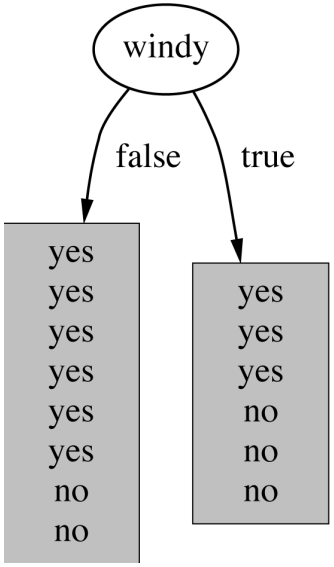
Day	Outlook	Temperature	Humidity	Wind	PlayTenn
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	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

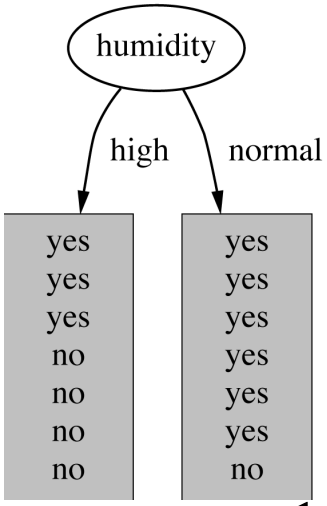
¿Qué atributo seleccionar?

Day	Outlook	Temperature	Humidity	Wind	PlayTenn
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



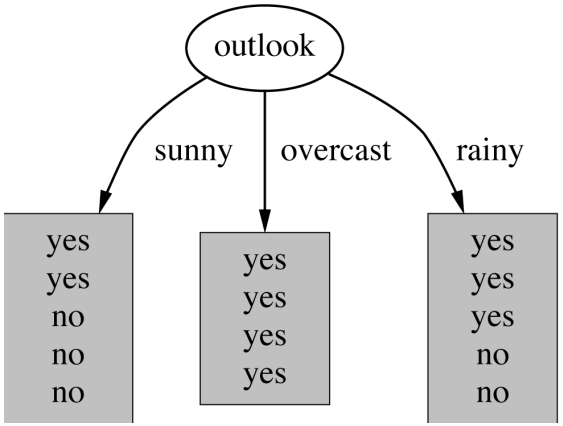






Ganancia de información

¿Cuál atributo?



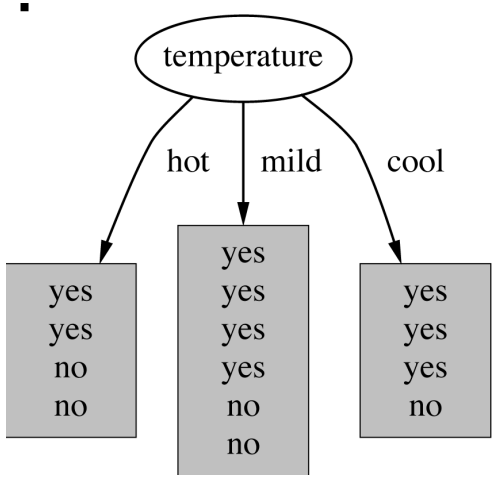
S:[9+,5-]
E=0.940

Outlook

Sunny **Overc.** **Rainy**

[2+,3-] E=0.971 | [4+,0-] E=0 | [3+,2-] E=0.971

Gain(S, Outlook)=0.940-(5/14)0.971-0-(5/14)0.971 =0.266



S:[9+,5-]
E=0.940

Temperat.

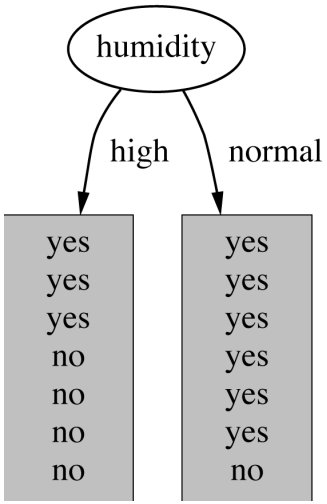
Hot **Mild** **Cool**

[2+,2-] E=1 | [4+,2-] E=0.918 | [3+,1-] E=0.811

Gain(S, Temp.)=0.940-(4/14)-(6/14)0.918 - (4/14)0.811 =0.029

Ganancia de información

¿Cuál atributo?



S:[9+,5-]
E=0.940

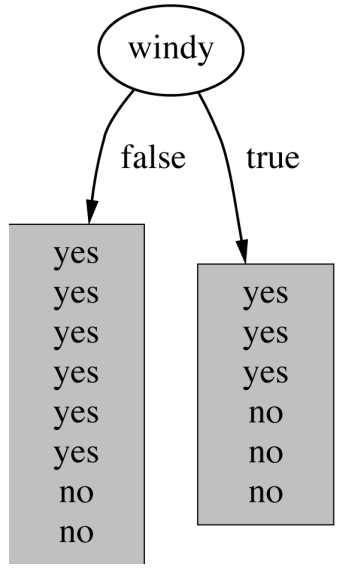
Humidity

High **Normal**

[3+,4-]
E=0.985

[6+,1-]
E=0.592

Gain(S, Humidity)=0.940-(7/14)0.985-(7/14)0.592
=0.151



S:[9+,5-]
E=0.940

Windy

False **True**

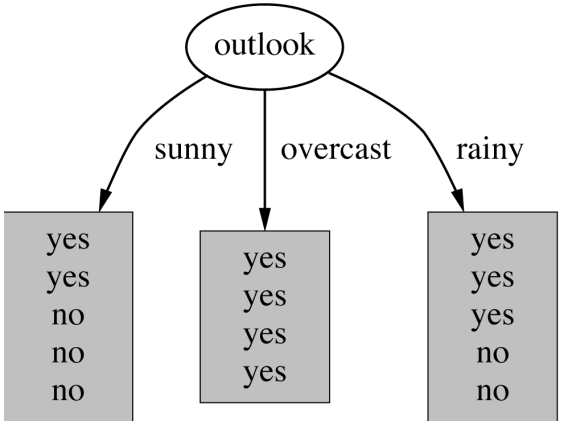
[6+,2-]
E=0.811

[3+,3-]
E=1

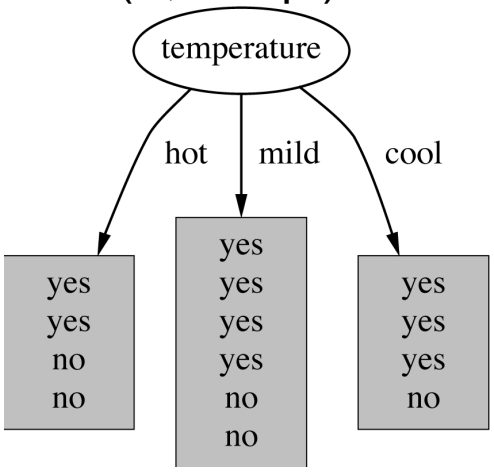
Gain(S, Windy)=0.940-(8/14)0.985-(6/14)
=0.048

¿Qué atributo seleccionar?

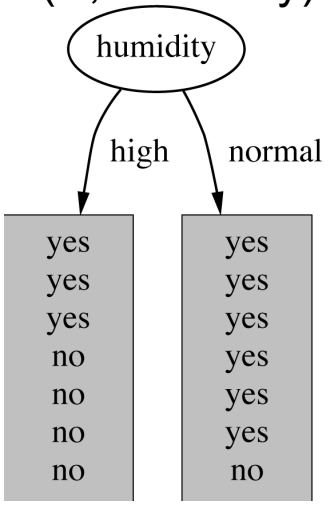
Gain(S, Outlook)=0.266



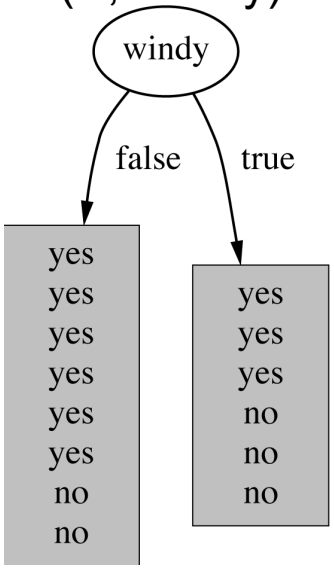
Gain(S, Temp.)=0.029



Gain(S, Humidity)=0.151

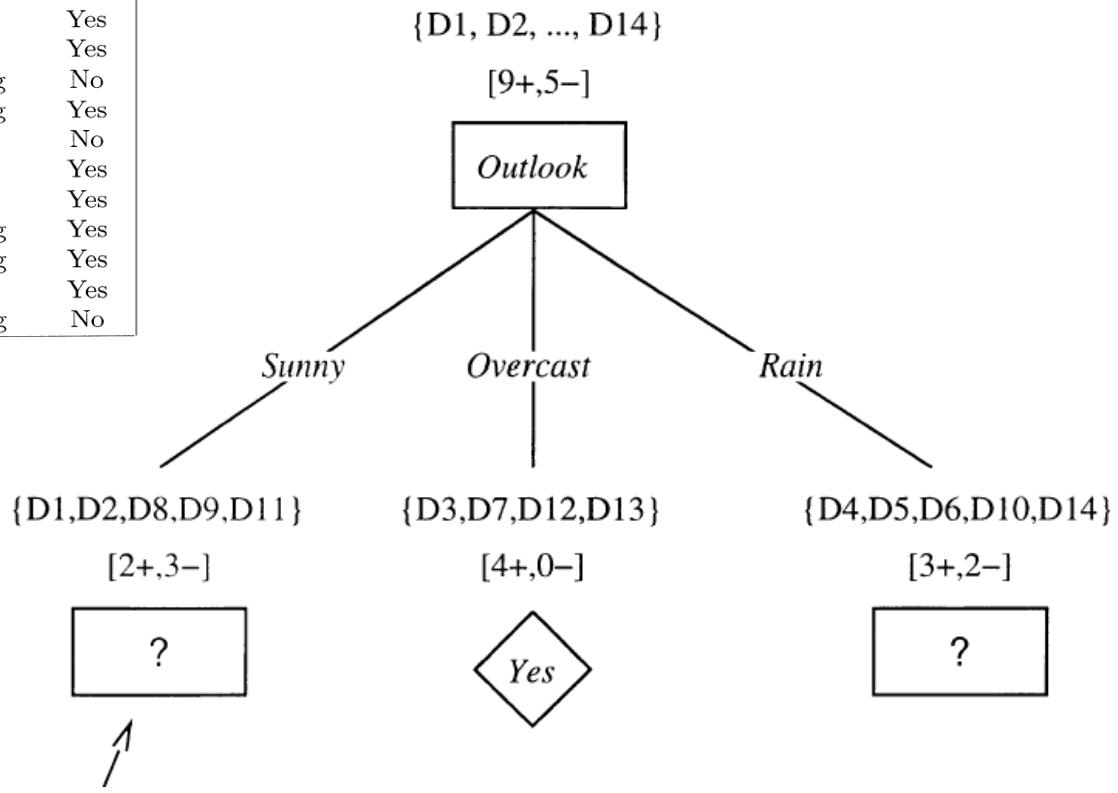


Gain(S, Windy)=0.048



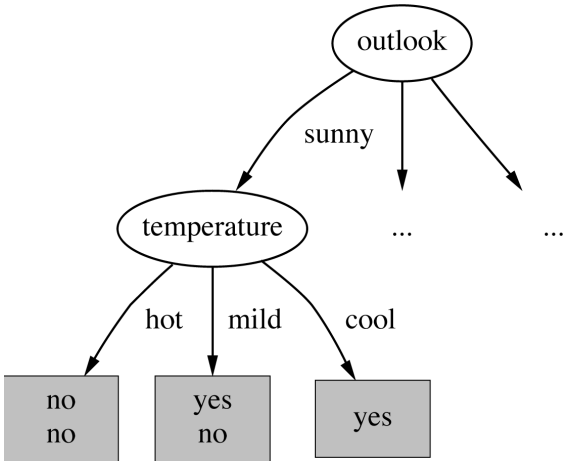
¿Qué atributo seleccionar?

Day	Outlook	Temperature	Humidity	Wind	PlayTenn
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

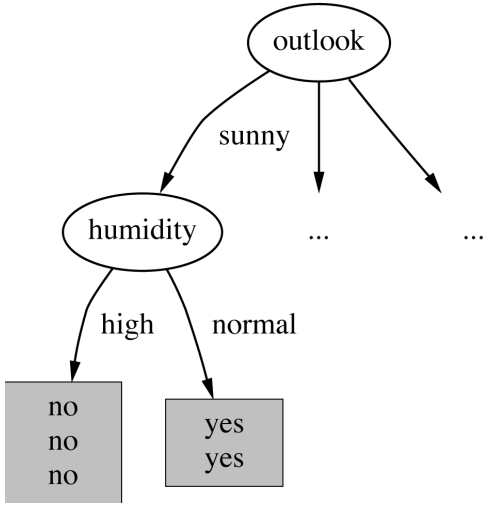


¿Cuál atributo?

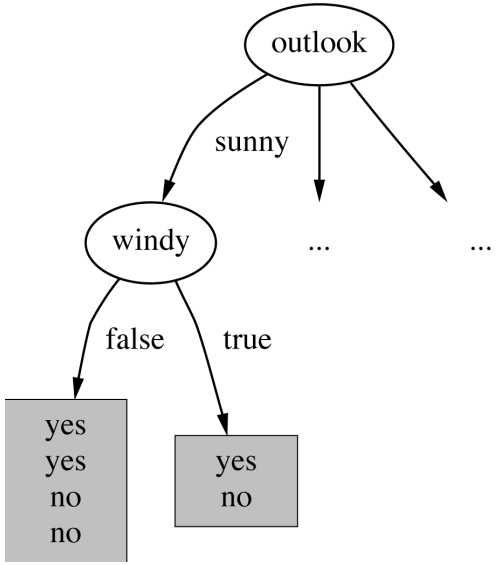
¿Qué atributo seleccionar?



Gain(S,Temp.)=?

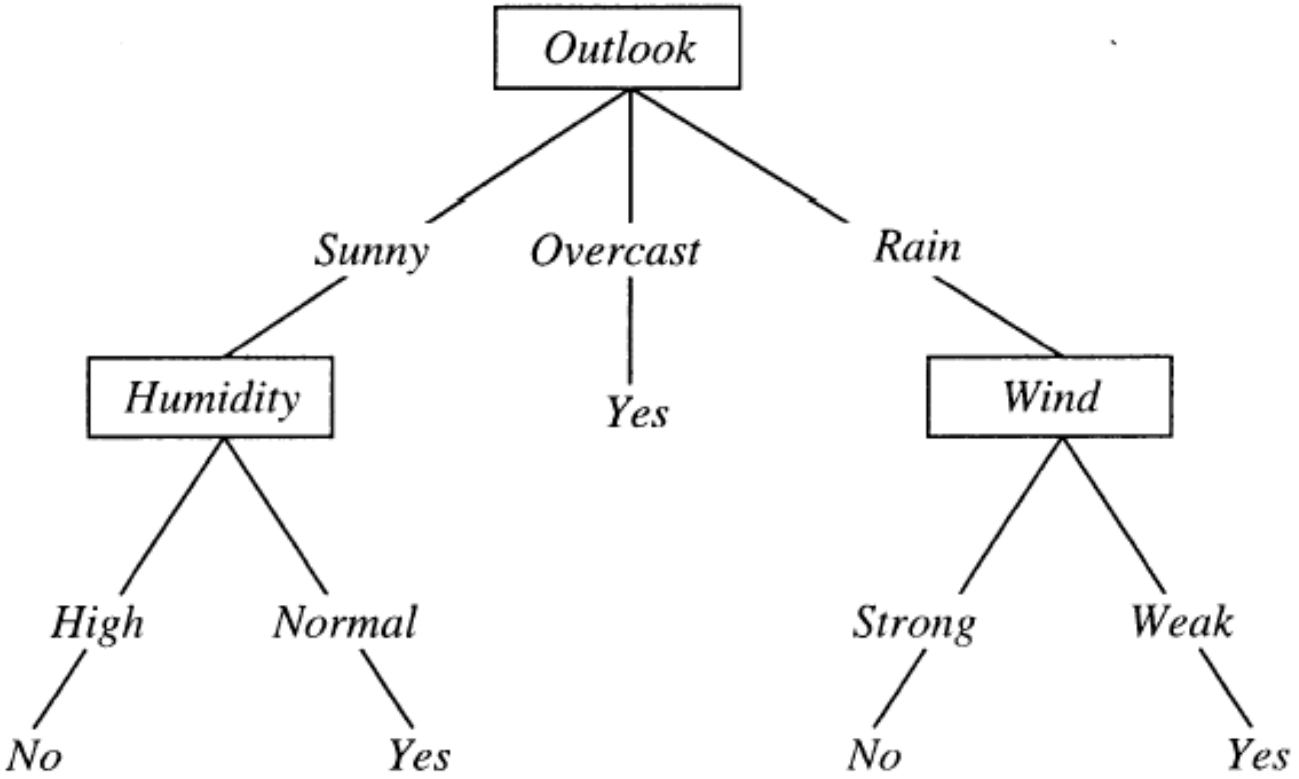


Gain(humid.,Temp.)=?



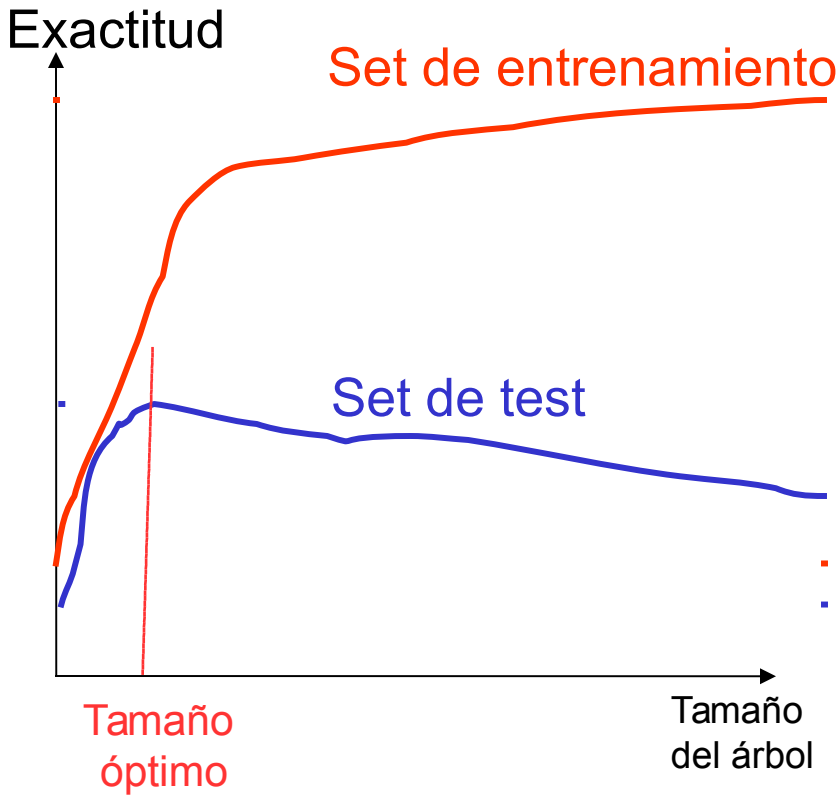
Gain(S,Windy)=?

Ejemplo: árbol final



Overfitting

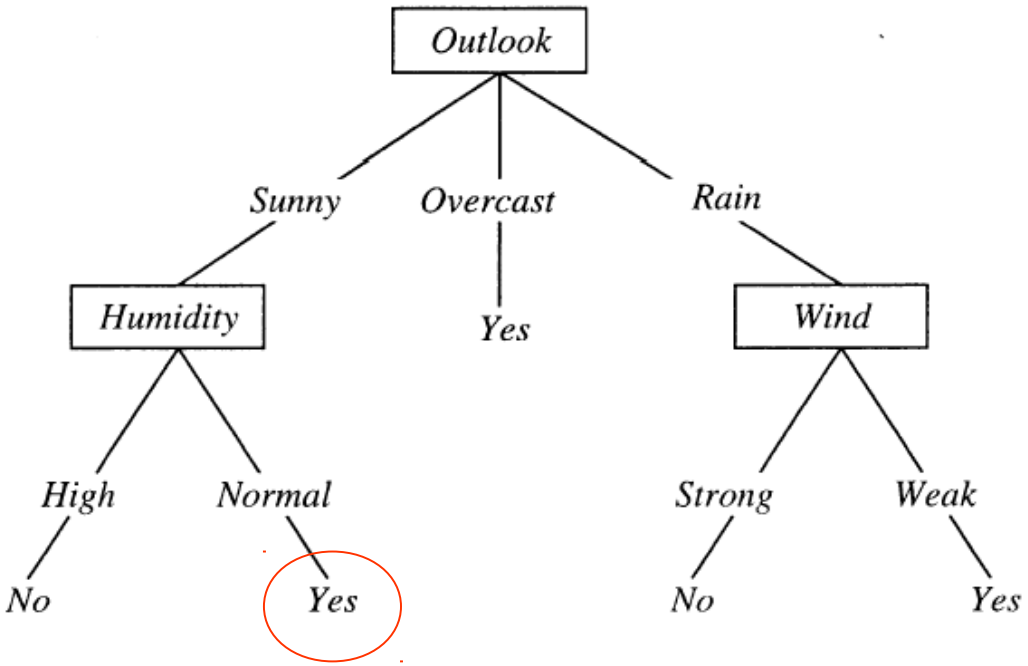
- El gráfico muestra el comportamiento típico en el set de entrenamiento:
  - Exactitud en el set de entrenamiento crece a medida que el árbol de decisión crece.
  - Existe un momento en que la exactitud en el set de test comienza a disminuir.
- ¿Cuál es la causa de este fenómeno?
  - El árbol comienza a sobre modelar los ejemplos del set de entrenamiento, perdiendo su capacidad de generalización.
  - Atributos menos predictivos, agregados al final, introducen ruido.
- Escenario deseado:
  - Detener la construcción del árbol antes que se produzca overfitting.



Overfitting

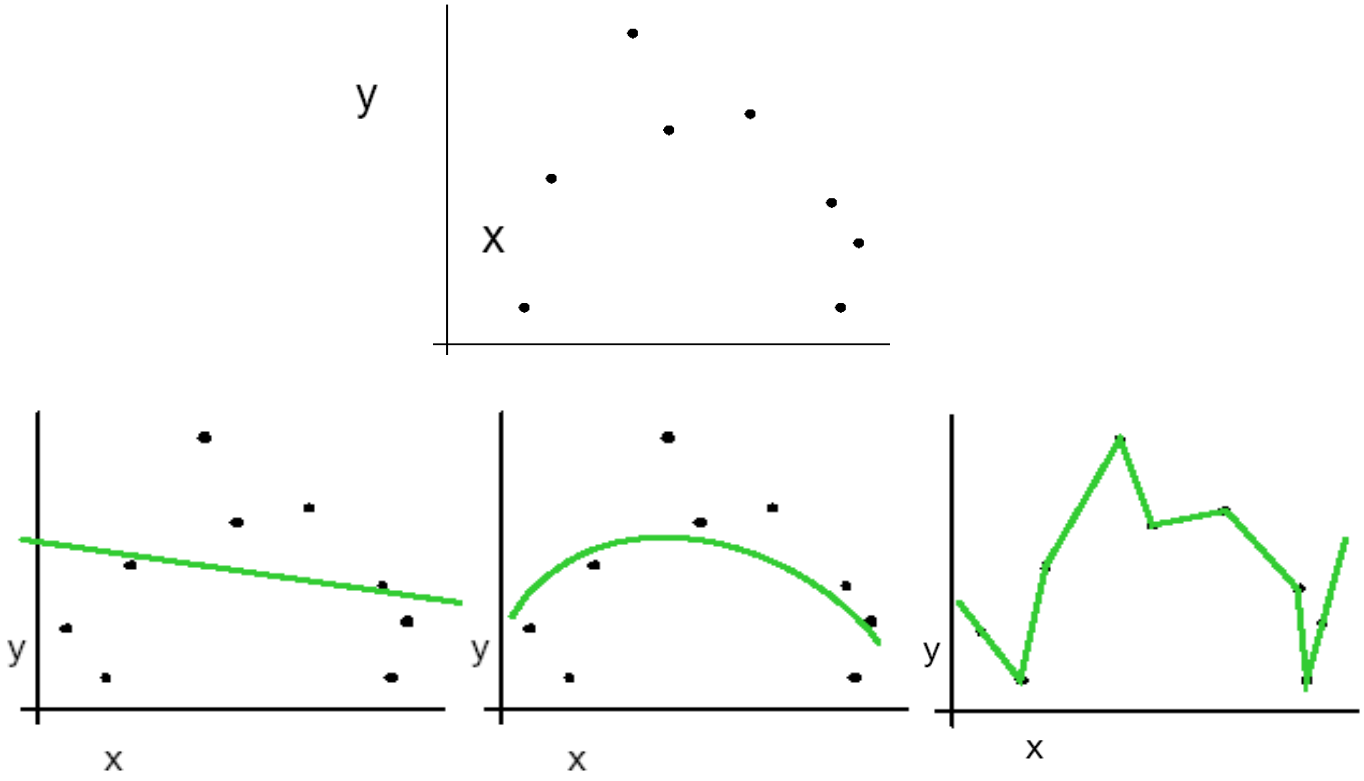
- Por ejemplo, que sucede si agregamos el siguiente registro al ejemplo sobre jugar tenis:

$$Sunny,\ Hot,\ Normal,\ Strong,\ PlayTennis = No$$



Overfitting

- Overfitting es un problema importante para los árboles de decisión y en general para los algoritmo de aprendizaje de máquina



Occam's Razor al rescate

Occam's razor is a logical principle attributed to the mediaeval philosopher William of Occam (or Ockham). The principle states that one should not make more assumptions than the minimum needed. It underlies all scientific modelling and theory building. It admonishes us to choose from a set of otherwise equivalent models of a given phenomenon the simplest one.

¿Cómo evitamos el overfitting?

- Parar la construcción del árbol cuando el número de registros restante no es estadísticamente significativo.

- Construir el árbol sin considerar restricciones de tamaño y luego podar usando rendimiento en set de validación.

- Otras métricas de selección de modelos AIC, BIC,… (IIC-3672).

Ganancia de información

- En casa paso de la construcción del árbol, GI es usada para cuantificar el atributo que presenta mayor discriminatividad, medida como homogeneidad de grupos resultantes respecto a los valores de la clase.
  - GI entrega buenos resultados y es muy usado en la práctica.
  - Sin embargo, existen otras métricas para medir homogeneidad de los grupos resultantes.

- Un inconveniente de GI es que no funciona bien con atributos que toman muchos valores, por qué?

Atributos con muchos valores

- Si un atributo tiene muchos valores, probablemente la métrica de ganancia de información lo seleccionará ¿Por qué?
  - Ej. Día=Julio 7 2005
- Una forma de solucionar el problema es usar la razón de ganancia (GainRatio):

$$GainRatio(S,A) \equiv \frac{Gain(S,A)}{SplitInformation(S,A)}$$

$$SplitInformation(S,A) \equiv -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

SplitInformation: mide entropía respecto al números de instancias $|S_i|$ en cada subgrupo. Mientras más uniforme sea el número de registros que sigue cada link, mayor será el split of information.

Árboles de decisión ¿Cuándo?

- Problemas de clasificación.
- Necesitamos generar reglas de decisión entendibles por personas. Un árbol de decisión es una disyunción de conjunciones.
- Atributos son discretos o discretizables sin gran pérdida de información.

Árboles de decisión en acción

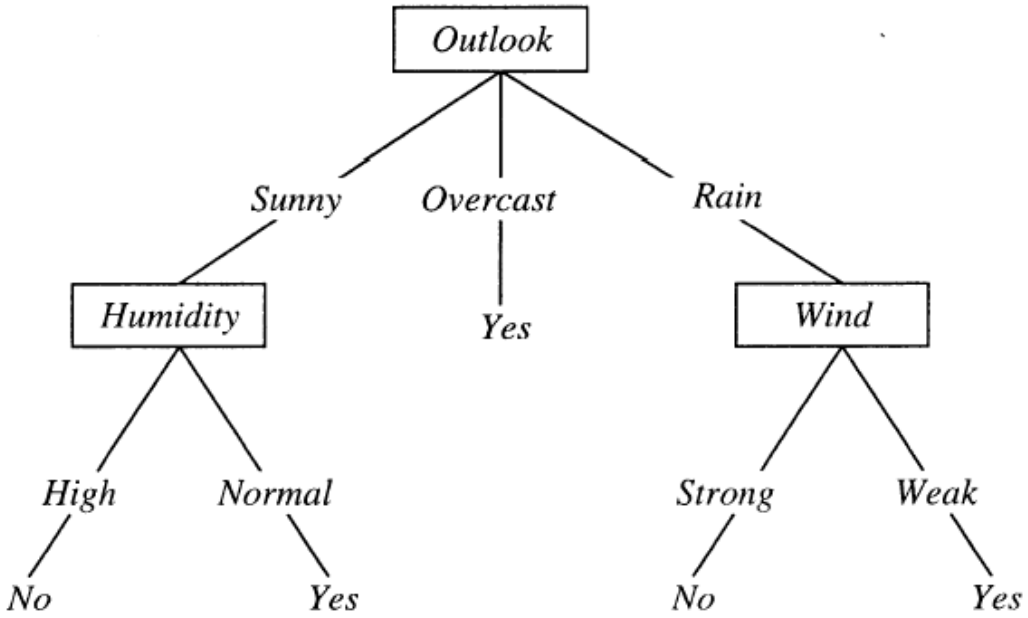
- Evaluación de riesgos al otorgar creditos bancarios utilizando árboles de decisión.

  Tesis de magister: Eduardo Robledo.

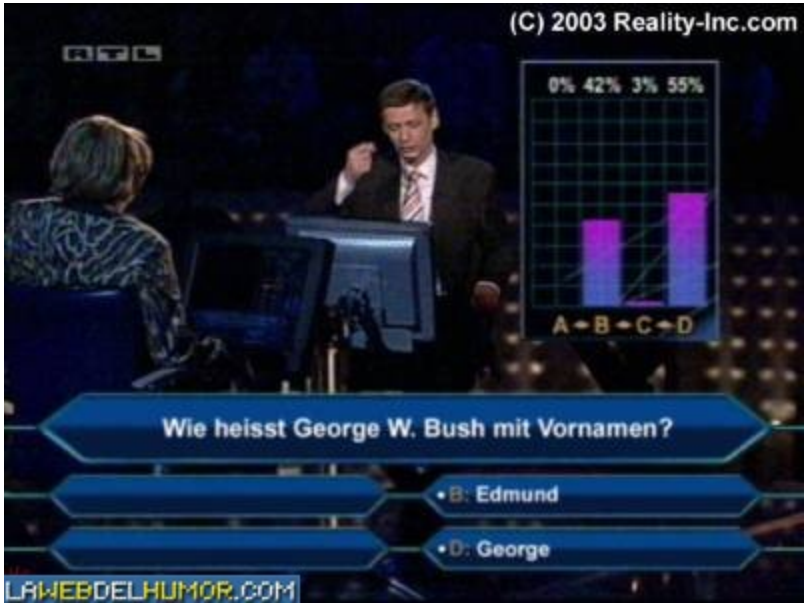
- Modelos de abandono (fuga, "churn") en una compañía de seguros usando árboles de decisión:

  Memoria de título: Pablo Ardiles.

Bonus point: árboles de decisión y selección de variables



Bonus point: Random Forest



"Wisdom of the crow"

Bonus point: Random Forest Microsoft Kinect