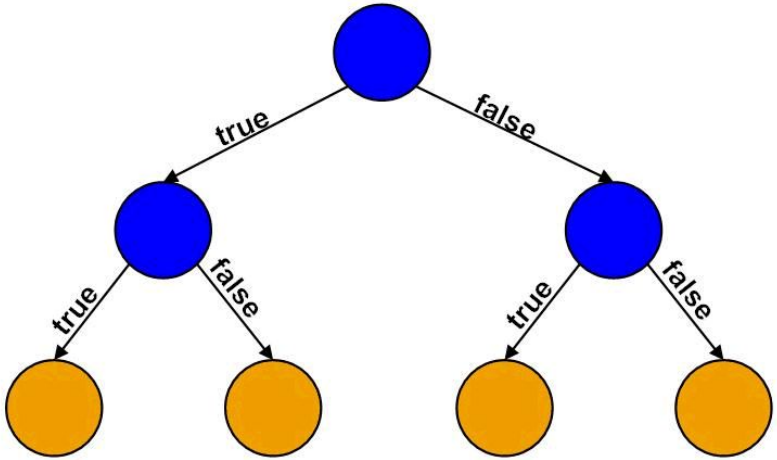
Ayudantía 7: Árboles de Decisión

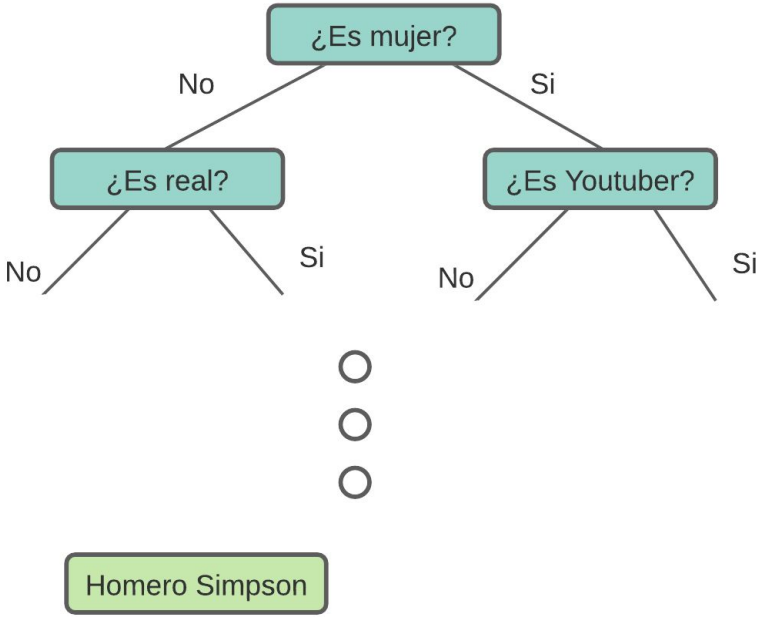
Sebastián Pérez Masri - sperezmasri@uc.cl
José Manuel Domínguez - jndominguez@uc.cl

¿Qué es un Árbol de Decisión?

- Clasificador con estructura de árbol
- Cada nodo representa un atributo
- Los nodos hojas representan el resultado de la clasificación



Akinator (árbol de decisión)



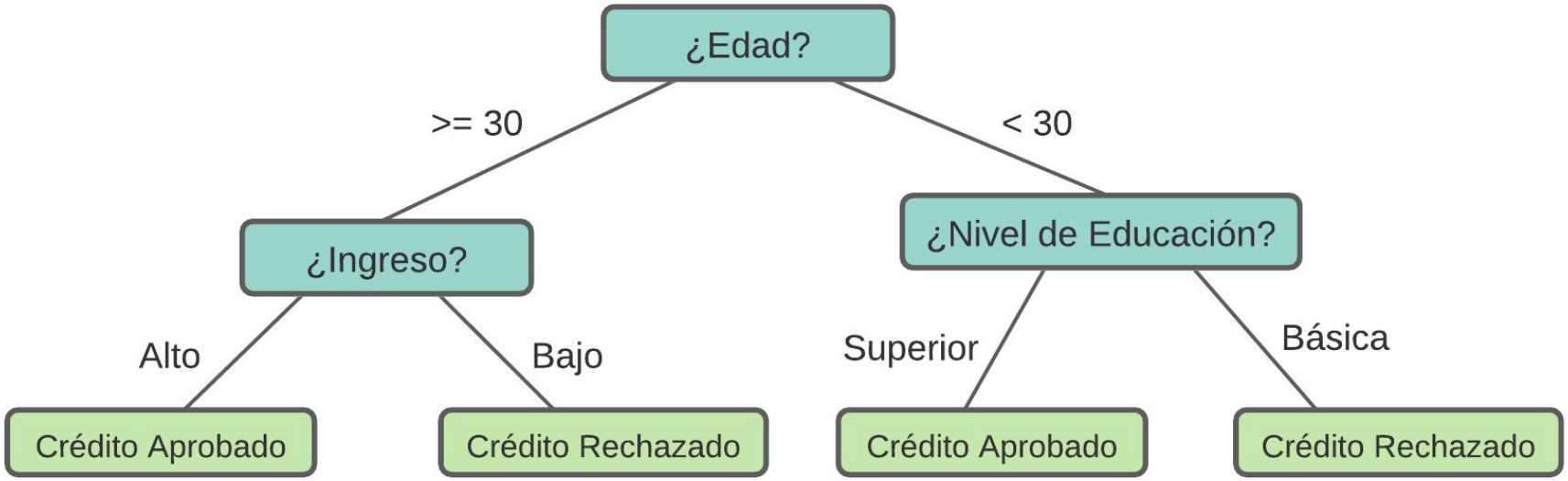
¿Cómo usarlo?

- **Discretización de los atributos. ¿Cómo trabajamos con datos como la edad?**
- Se usa como aprendizaje supervisado por lo tanto necesitamos datos previamente etiquetados
- Elección de qué atributos van primero según alguna métrica (por ej: ganancia de información)

Discretización de los atributos

Edad	Educación	Ingreso/mes	Crédito
20	Básica	100.000	Rechazado
25	Superior	500.000	Aprobado
50	Superior	200.000.000	Aprobado
30	Superior	100.000	Rechazado
80	Básica	500.000.000	Aprobado

Discretización de los atributos



¿Cómo usarlo?

- Discretización de los atributos. ¿Cómo trabajamos con datos como la edad?

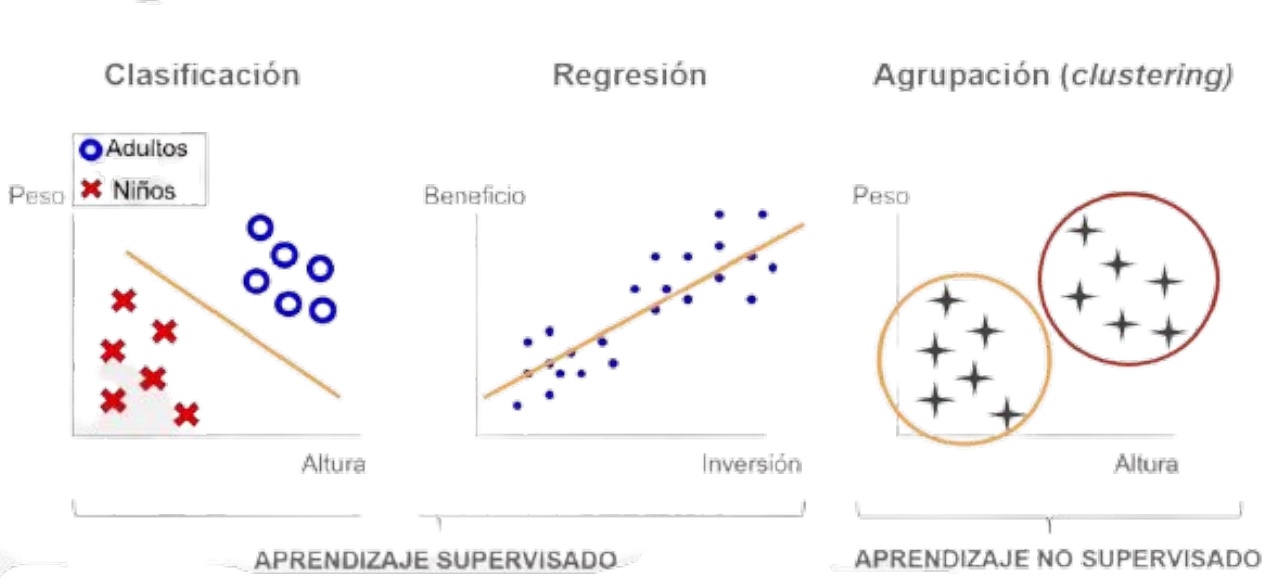
- **Se usa como aprendizaje supervisado por lo tanto necesitamos datos previamente etiquetados**

- Elección de qué atributos van primero según alguna métrica (por ej: ganancia de información)

Datos etiquetados

Edad	Educación	Ingreso/mes	Crédito
20	Básica	100.000	Rechazado
25	Superior	500.000	Aprobado
50	Superior	200.000.000	Aprobado
30	Superior	100.000	Rechazado
80	Básica	500.000.000	Aprobado

Datos etiquetados



¿Cómo usarlo?

- Discretización de los atributos. ¿Cómo trabajamos con datos como la edad?
- Se usa como aprendizaje supervisado por lo tanto necesitamos datos previamente etiquetados
- **Elección de qué atributos van primero según alguna métrica (por ej: ganancia de información)**

Elección de atributos (construcción del árbol)

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Entropía y ganancia de información

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

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

Entropía y ganancia de información

Y : Jugar tenis (Yes)

N: No jugar tenis (No)

Entropía y ganancia de información

Entropía del sistema o de PlayTennis:

$$Entropy(S) = - (P(Y) * Log_2(P(Y)) + P(N) * Log_2(P(N))$$

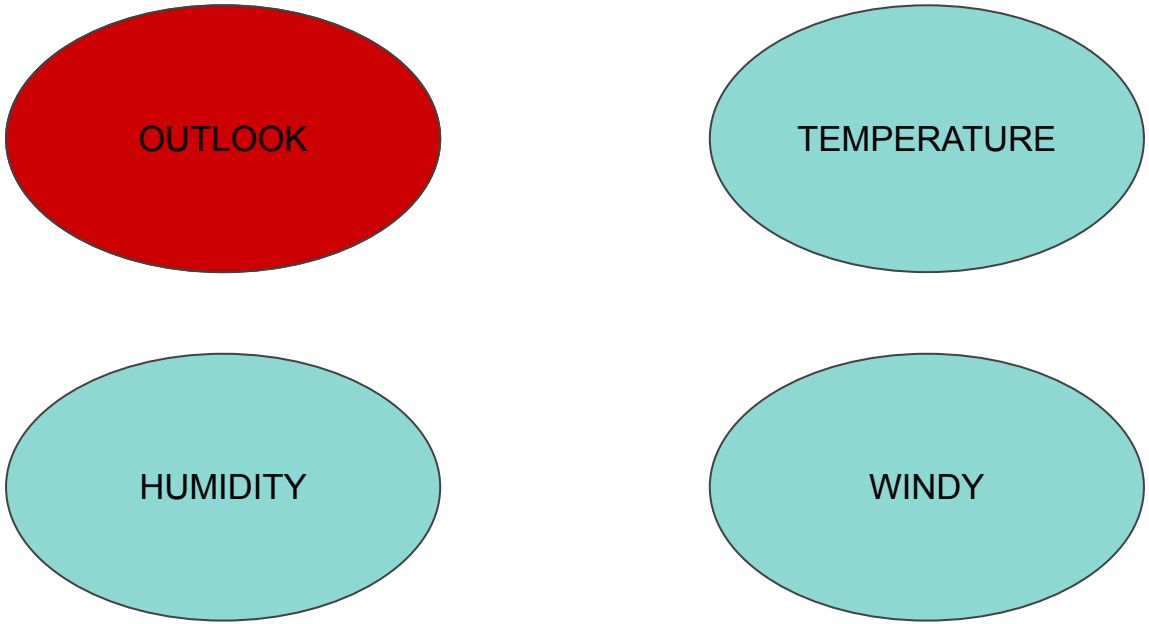
Viendo la tabla...

$P(Y) = 9 / 14 = 0.643$

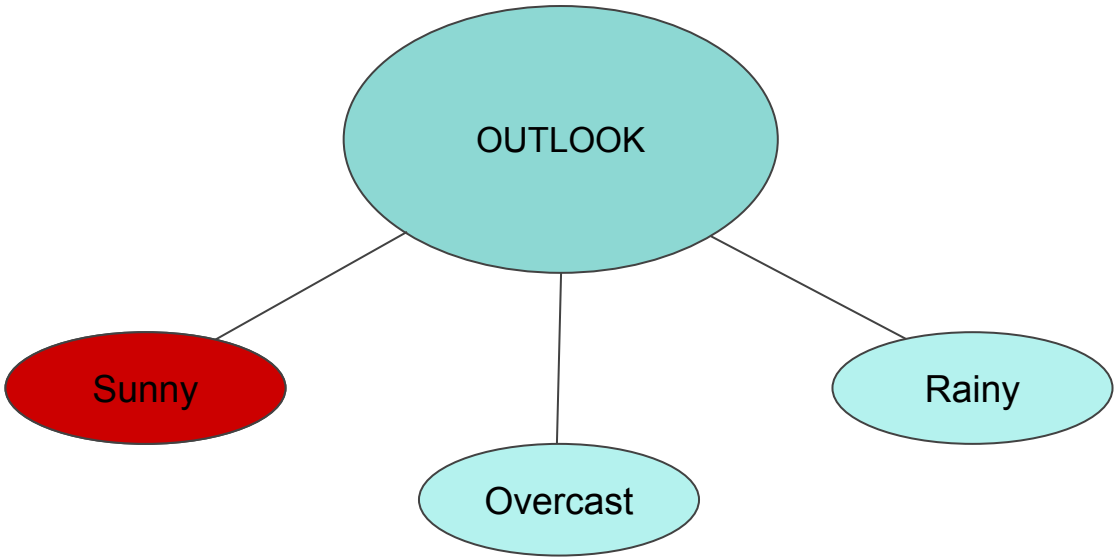
$P(N) = 5 / 14 = 0.357$

$$Entropy(S) = 0.94$$

Entropía y ganancia de información



Entropía y ganancia de información



Entropía y ganancia de información

Entropía cuando está soleado (sunny):

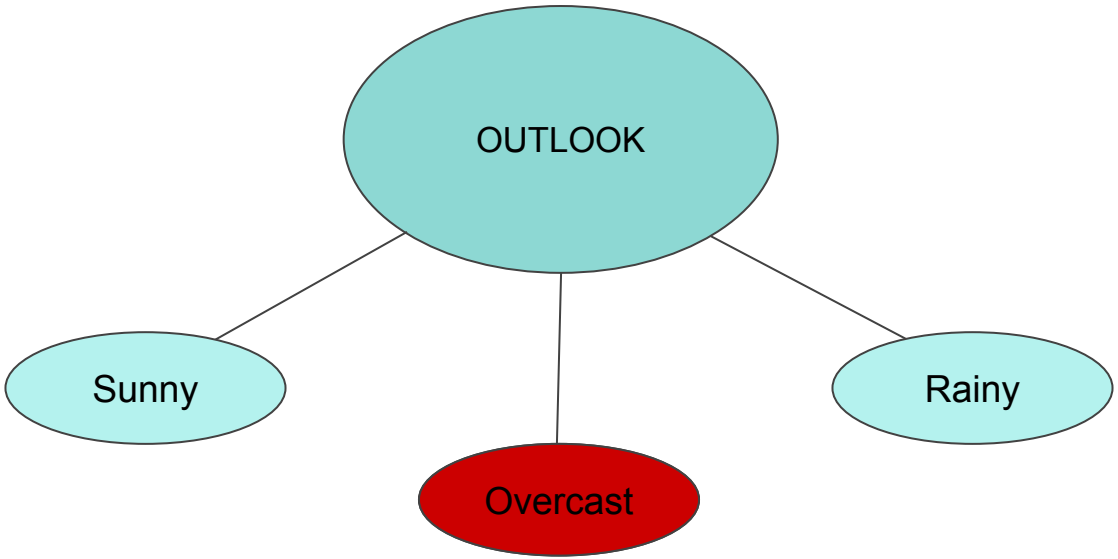
Viendo la tabla...

P(Y/Sunny) = 2 / 5 = 0.4

P(N/Sunny) = 3 / 5 = 0.6

$$\text{Entropy}(S_{\text{sunny}}) = -\left(0.4 * \text{Log}_2(0.4) + 0.6 * \text{Log}_2(0.6)\right) = 0.97$$

Entropía y ganancia de información



Entropía y ganancia de información

Entropía cuando está nublado (overcast):

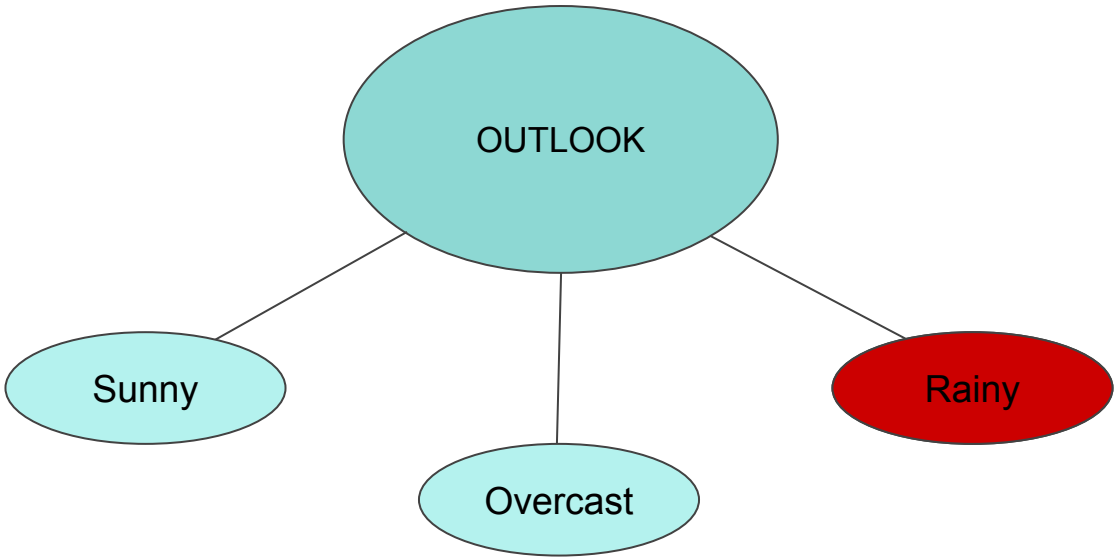
Viendo la tabla...

P(Y/Overcast) = 4 / 4 = 1

P(N/Overcast) = 0 / 4 = 0

$$\text{Entropy}(S_{\text{overcast}}) = -\ (1 * \text{Log}_2(1) + 0 * \text{Log}_2(0)) = 0$$

Entropía y ganancia de información



Entropía y ganancia de información

Entropía cuando está lloviendo (rainy):

Viendo la tabla...

P(Y/Rainy) = 3 / 5 = 0.6

P(N/Rainy) = 2 / 5 = 0.4

$$\text{Entropy}(S_{\text{rainy}}) = -\ (0.6 * \text{Log}_2(0.6) + 0.4 * \text{Log}_2(0.4)) = 0.97$$

Entropía y ganancia de información

Ganancia de Información:

n_s: cantidad de días que fueron soleados

n_o: cantidad de días que fueron nublados

n_r: cantidad de días con lluvia

n_t: cantidad de días totales

$$Gain_{outlook} = E(S) - ((n_s / n_t) * E(S_{sunny}) + (n_o / n_t) * E(S_{overcast}) + (n_r / n_t) * E(S_{rainy})$$

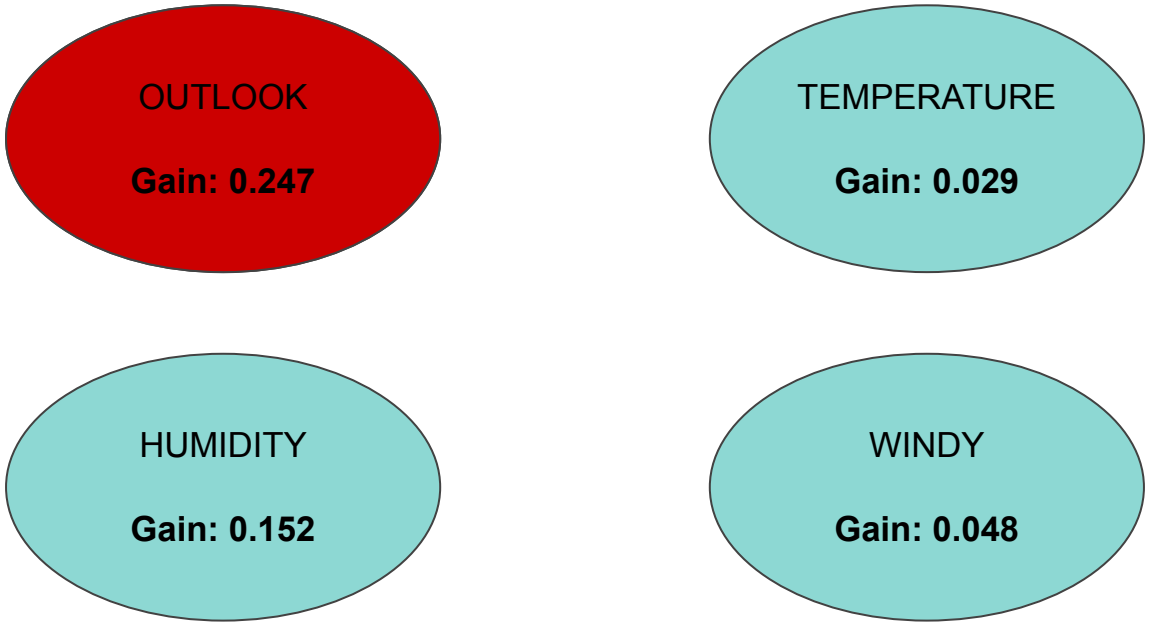
Entropía y ganancia de información

$$Gain_{outlook} = E(S) - ((n_s / n_t) * E(S_{sunny}) + (n_o / n_t) * E(S_{overcast}) + (n_r / n_t) * E(S_{rainy})$$

$$Gain_{outlook} = 0.94 - ((5 / 14) * 0.97 + (4 / 14) * 0 + (5 / 14) * 0.97$$

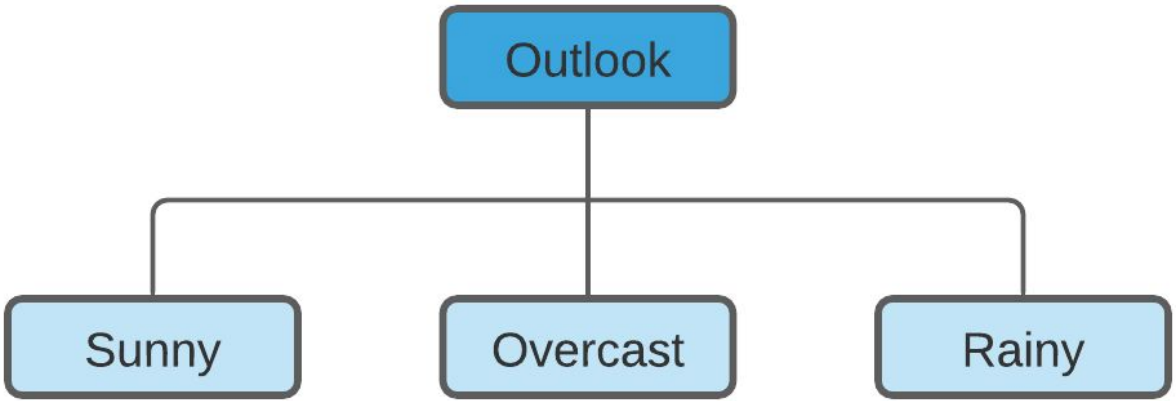
$$Gain_{outlook} = 0.247$$

Entropía y ganancia de información



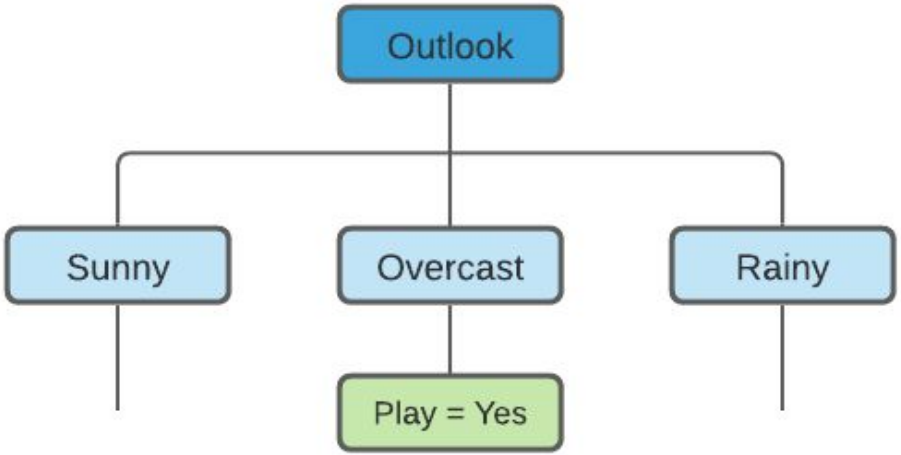
Entropía y ganancia de información

Por lo tanto nuestro árbol queda (por el momento) así:



Entropía y ganancia de información

Viendo que Entropía(S_{sunny}) != 0, Entropía($S_{overcast}$) = 0, Entropía(S_{rainy}) != 0



Entropía y ganancia de información

¿Cómo seguir? Analizar cada caso por separado

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Sunny	Mild	Normal	True	Yes

Entropía y ganancia de información

¿Cómo seguir? Analizar cada caso por separado

Outlook	Temperature	Humidity	Windy	PlayTennis
Overcast	Hot	High	False	Yes
Overcast	Cool	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes

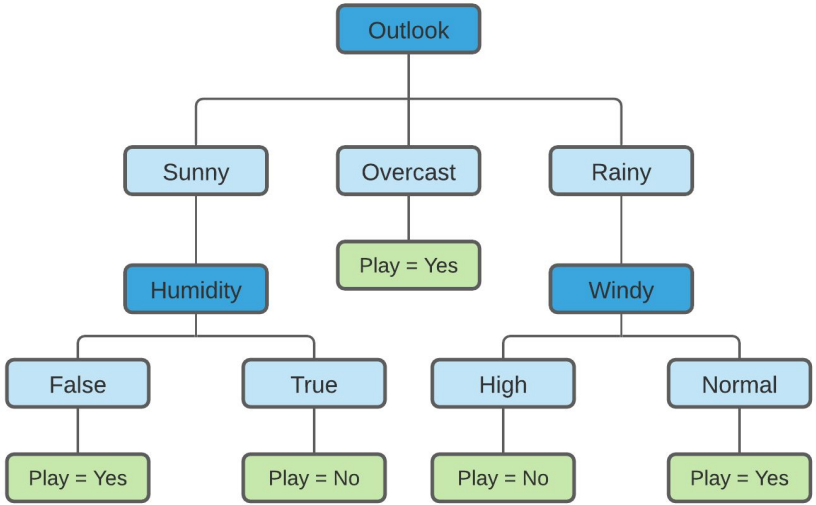
Entropía y ganancia de información

¿Cómo seguir? Analizar cada caso por separado

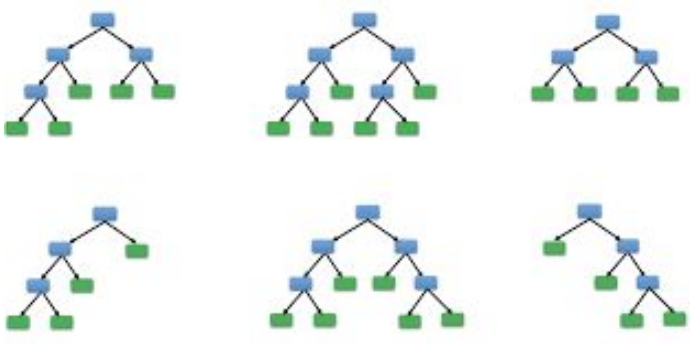
Outlook	Temperature	Humidity	Windy	PlayTennis
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Rainy	Mild	Normal	False	Yes
Rainy	Mild	High	True	No

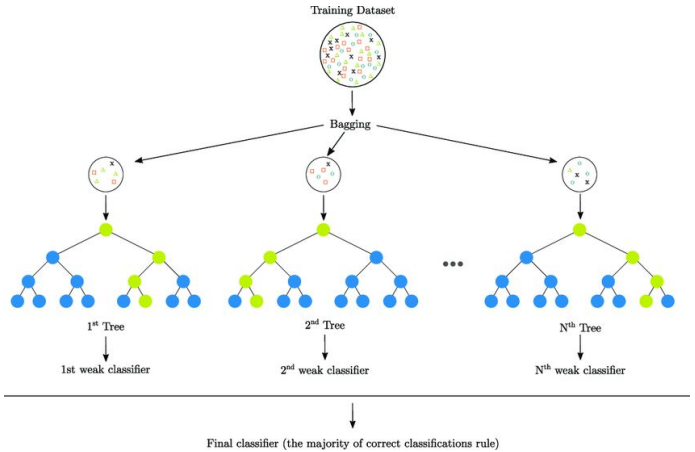
Entropía y ganancia de información

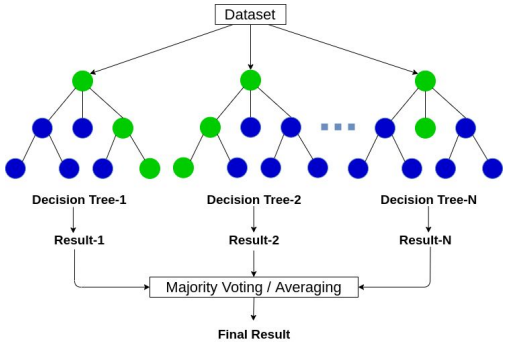
Hacemos esto hasta solo tener hojas, llegando al árbol final



Random Forest







Random Forest

Algoritmo

Sea:

N = número de instancia entrenamiento,

M = número de atributos,

m = número atributos para cada árbol, $m \ll M$:

For (1:L),

- Muestrear N ejemplos del set de entrenamiento con reemplazo.
- Seleccionar aleatoriamente *m* features.
- Entrenar árbol de decisión usando las *N* instancias y *m* atributos.
- Entrenar hasta convergencia sin poda.

End;

- Clasificar nuevas instancias usando votación por mayoría.

Random Forest

¿Cuántos características debería tener cada árbol?

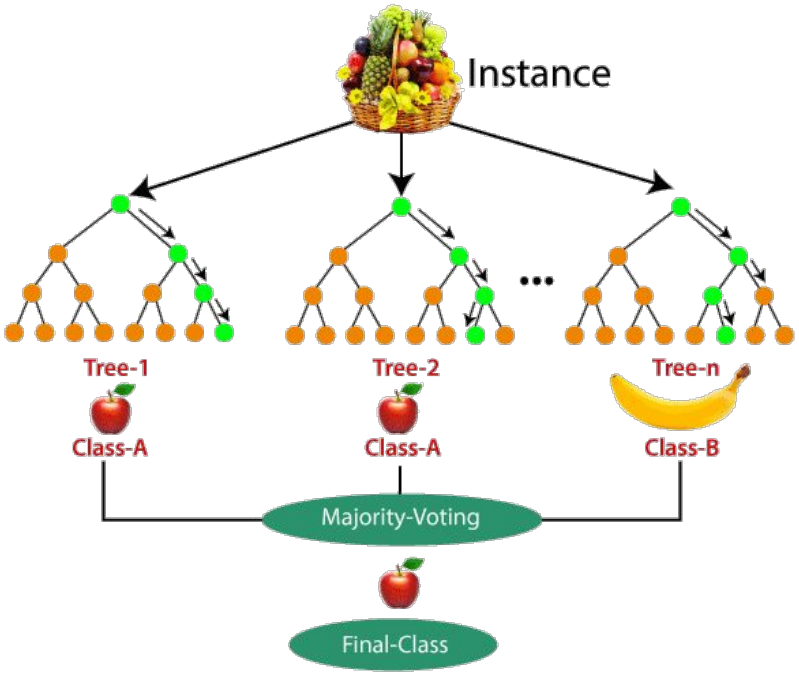
Considerando que tenemos 10 características en total...

¿Qué pasa si cada árbol tiene solo 1 característica?

¿Qué pasa si cada árbol tiene las 10 características?

Trade-off entre diversidad y precisión

Random Forest



Random Forest vs Decision Tree

Ventajas:

- En general suele ser más preciso
- Suele funcionar mejor con datasets grandes

Desventaja:

- Difícil de interpretar
- Más lento de evaluar y entrenar