



Clasificación Supervisada

Árboles de Decisión

Jesús Ariel Carrasco Ochoa

Instituto Nacional de Astrofísica, Óptica y Electrónica



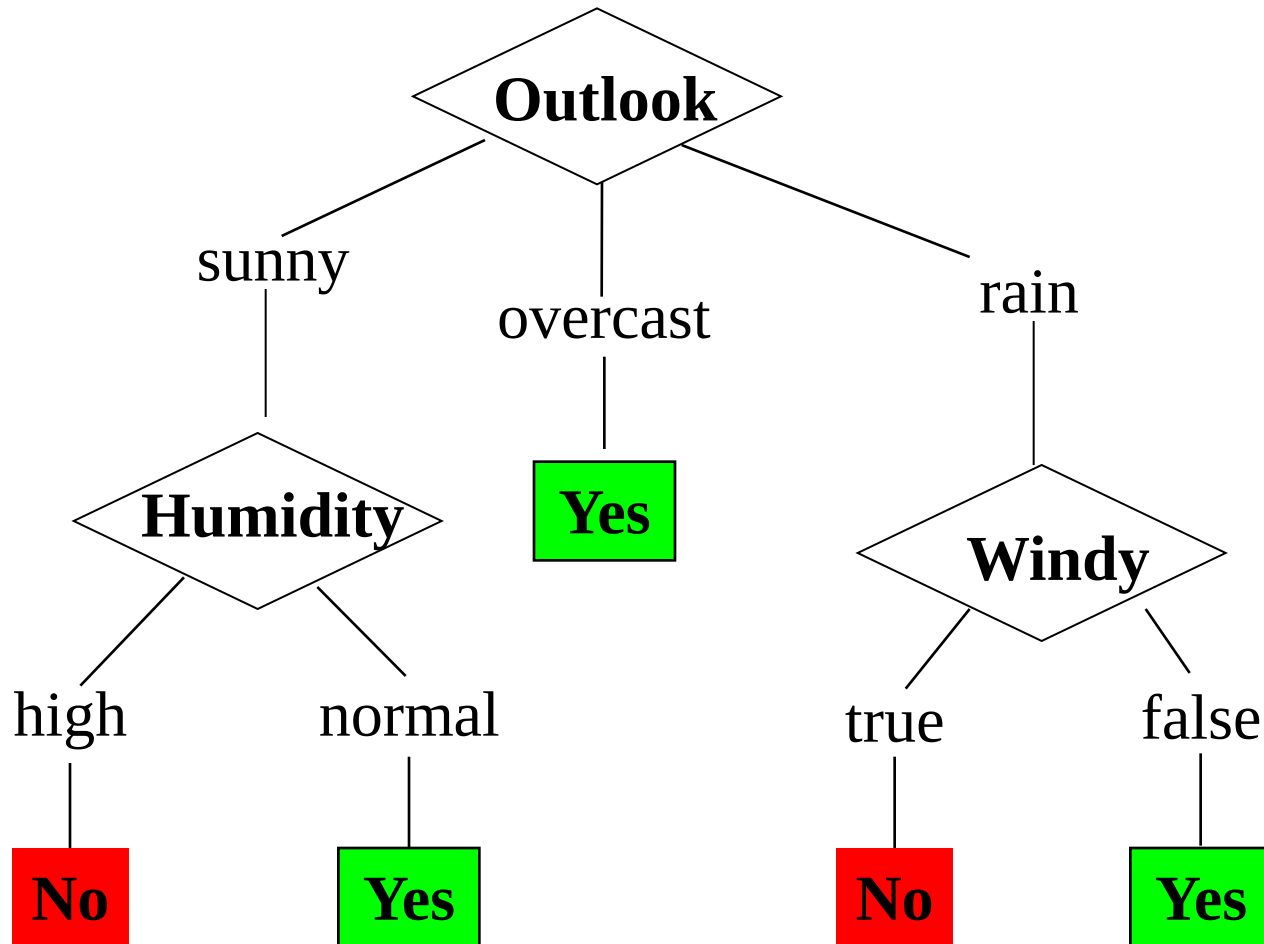
Árboles de Decisión

- Es un árbol en el cual
 - Cada nodo interno es una prueba sobre un atributo
 - Cada arista representa un posible resultado de la prueba
 - Cada hoja representa una clase
- Un nuevo objeto es clasificado siguiendo un camino de la raíz a una hoja.

Árboles de Decisión

Object	Outlook	Temperature	Humidity	Windy	Play?
O_1	sunny	hot	high	false	No
O_2	sunny	hot	high	true	No
O_3	overcast	hot	high	false	Yes
O_4	rain	mild	high	false	Yes
O_5	rain	cool	normal	false	Yes
O_6	rain	cool	normal	true	No
O_7	overcast	cool	normal	true	Yes
O_8	sunny	mild	high	false	No
O_9	sunny	cool	normal	false	Yes
O_{10}	rain	mild	normal	false	Yes
O_{11}	sunny	mild	normal	true	Yes
O_{12}	overcast	mild	high	True	Yes
O_{13}	overcast	hot	normal	False	Yes
O_{14}	rain	mild	high	True	No

Árboles de Decisión



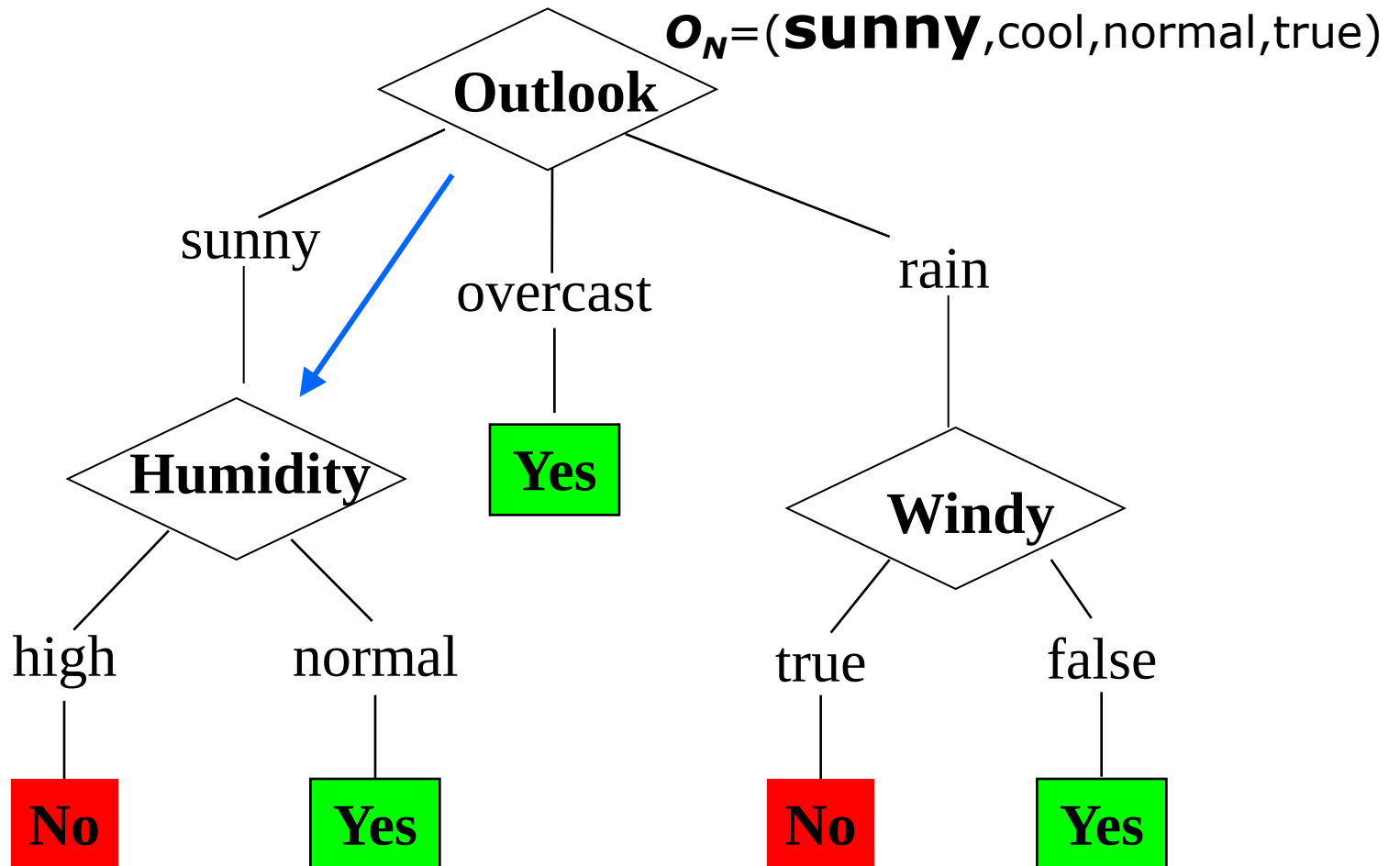
Árboles de Decisión

Para Clasificar un nuevo objeto

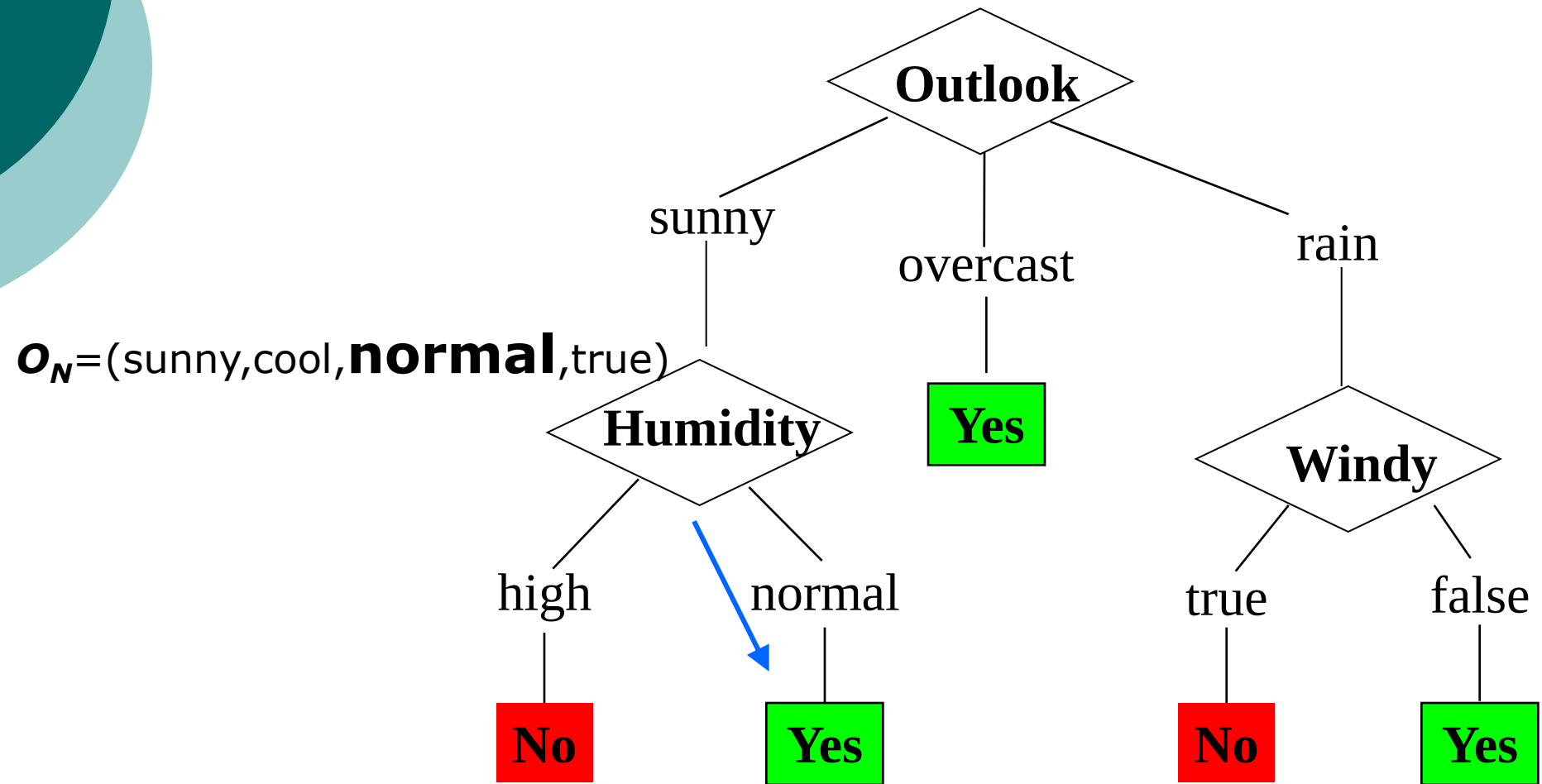
$\mathbf{O}_N = (\text{sunny}, \text{cool}, \text{normal}, \text{true})$

- Se inicia desde la raíz.
- Se verifica el valor de la variable asociada con el nodo que se esté comparando
- Se sigue la rama que corresponda al valor del objeto a clasificar
- Cuando se llega a una hoja se asigna la clase correspondiente

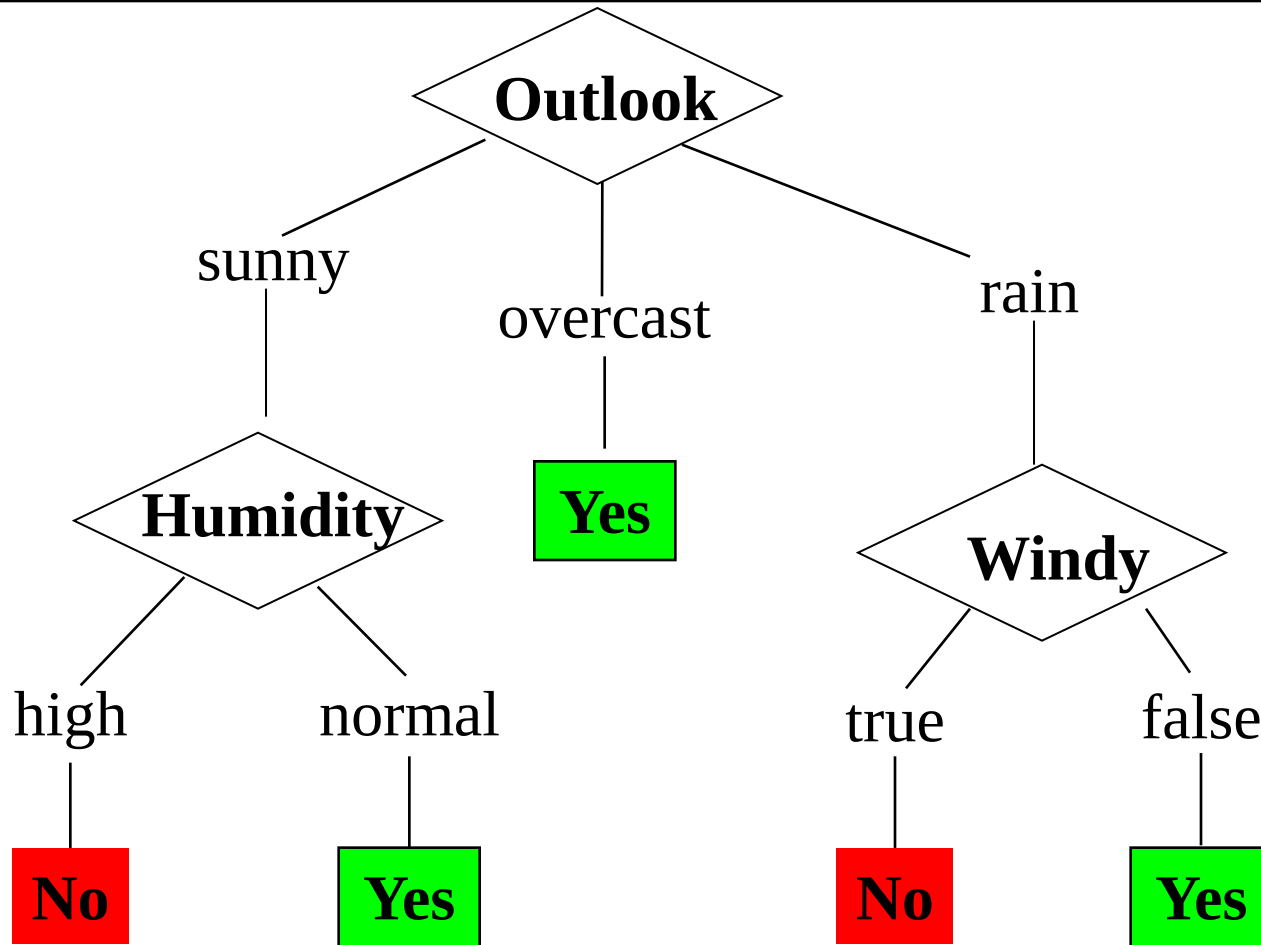
Árboles de Decisión



Árboles de Decisión



Árboles de Decisión



Clase de $\mathbf{o}_N = \mathbf{Yes}$



Árboles de Decisión

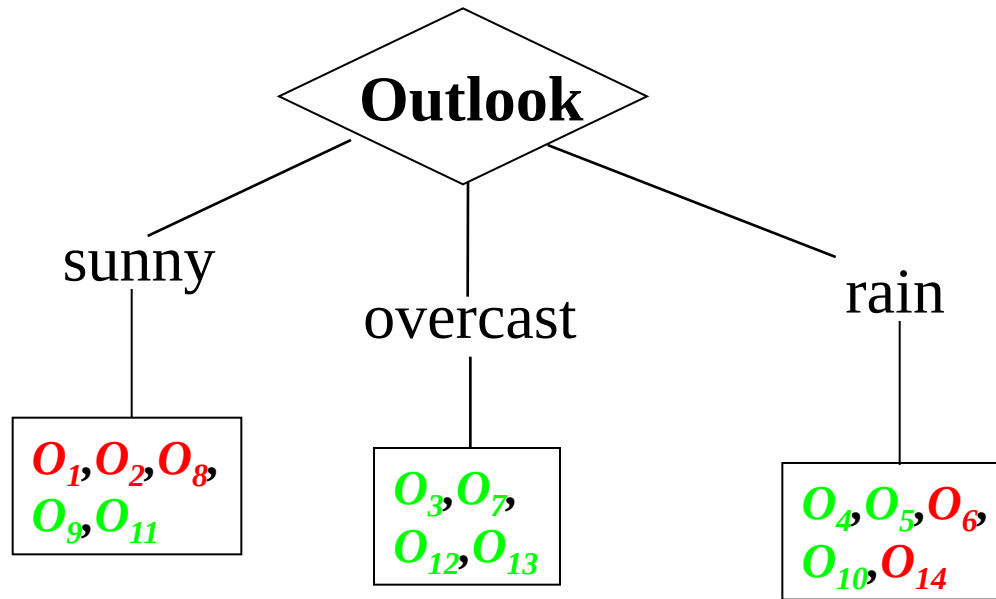
Cómo construir el árbol?

- Se sigue una estrategia “Top-Down”
 - Al inicio todos los objetos de entrenamiento están en la raíz
 - Se escoge un atributo y se separan los objetos
 - Se repite hasta que cada hoja tiene objetos de una sola clase
 - Se asigna a las hojas la clase de los objetos que contiene

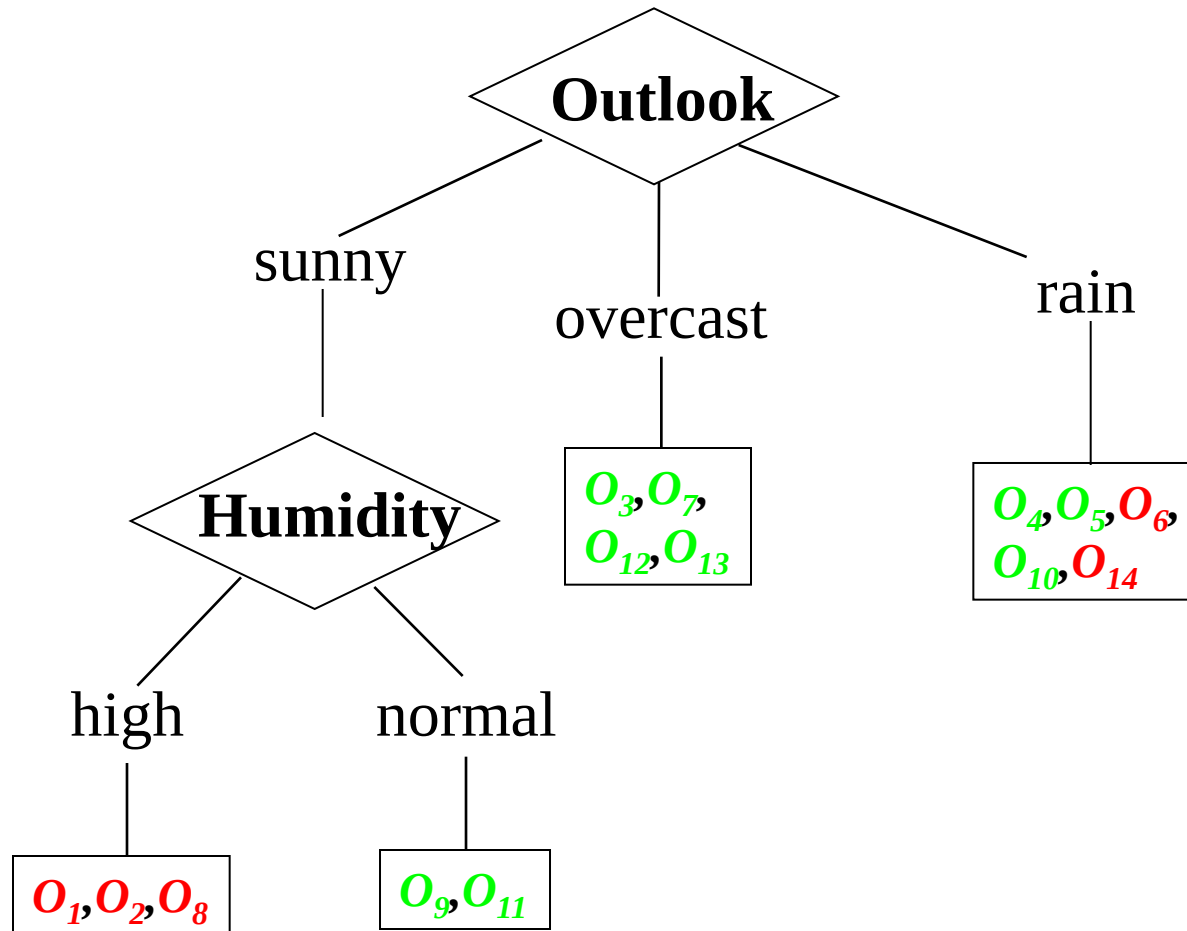
Árboles de Decisión

$O_1, O_2, O_3, O_4, O_5, O_6, O_7, O_8,$
 $O_9, O_{10}, O_{11}, O_{12}, O_{13}, O_{14}$

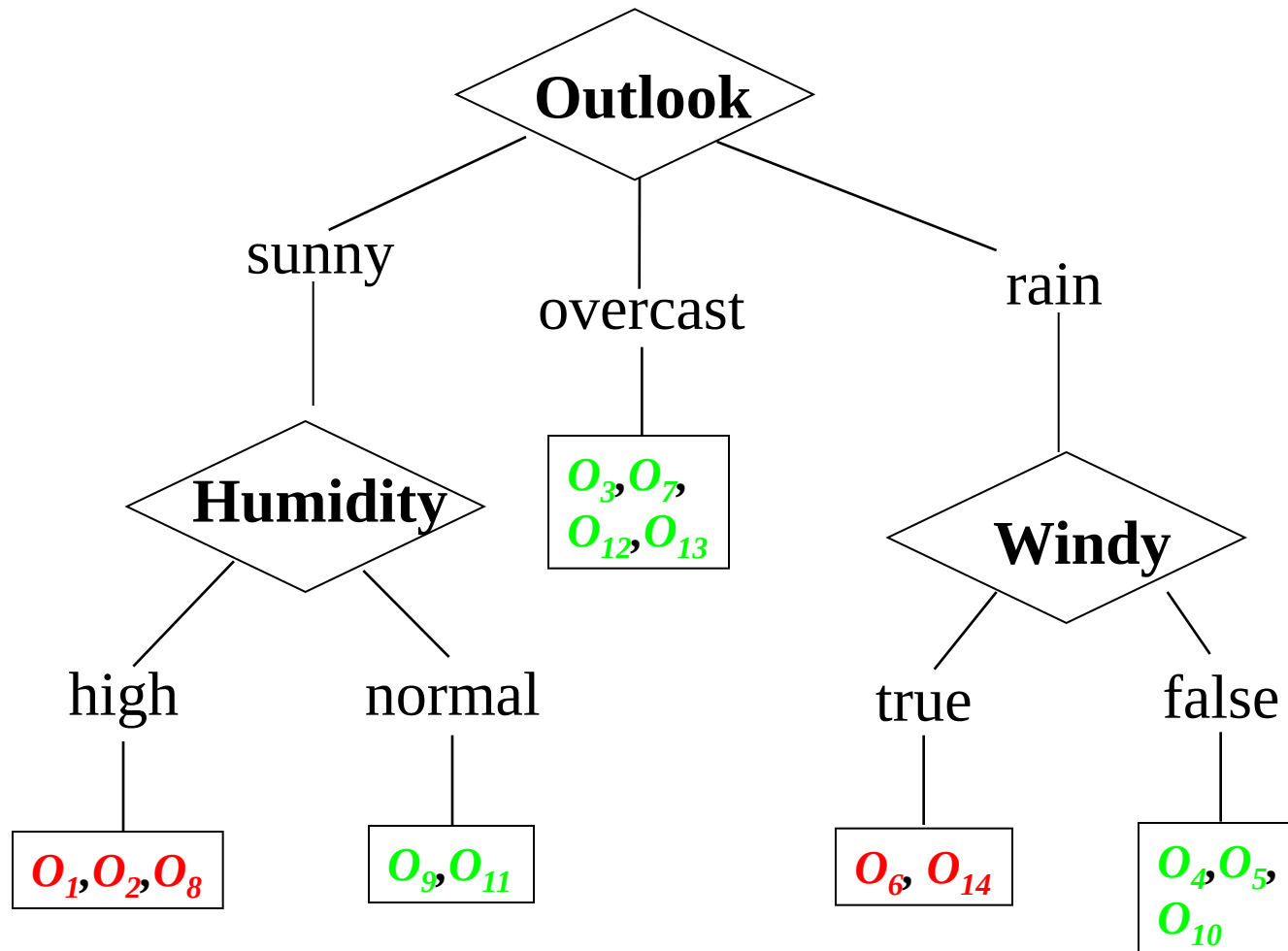
Árboles de Decisión



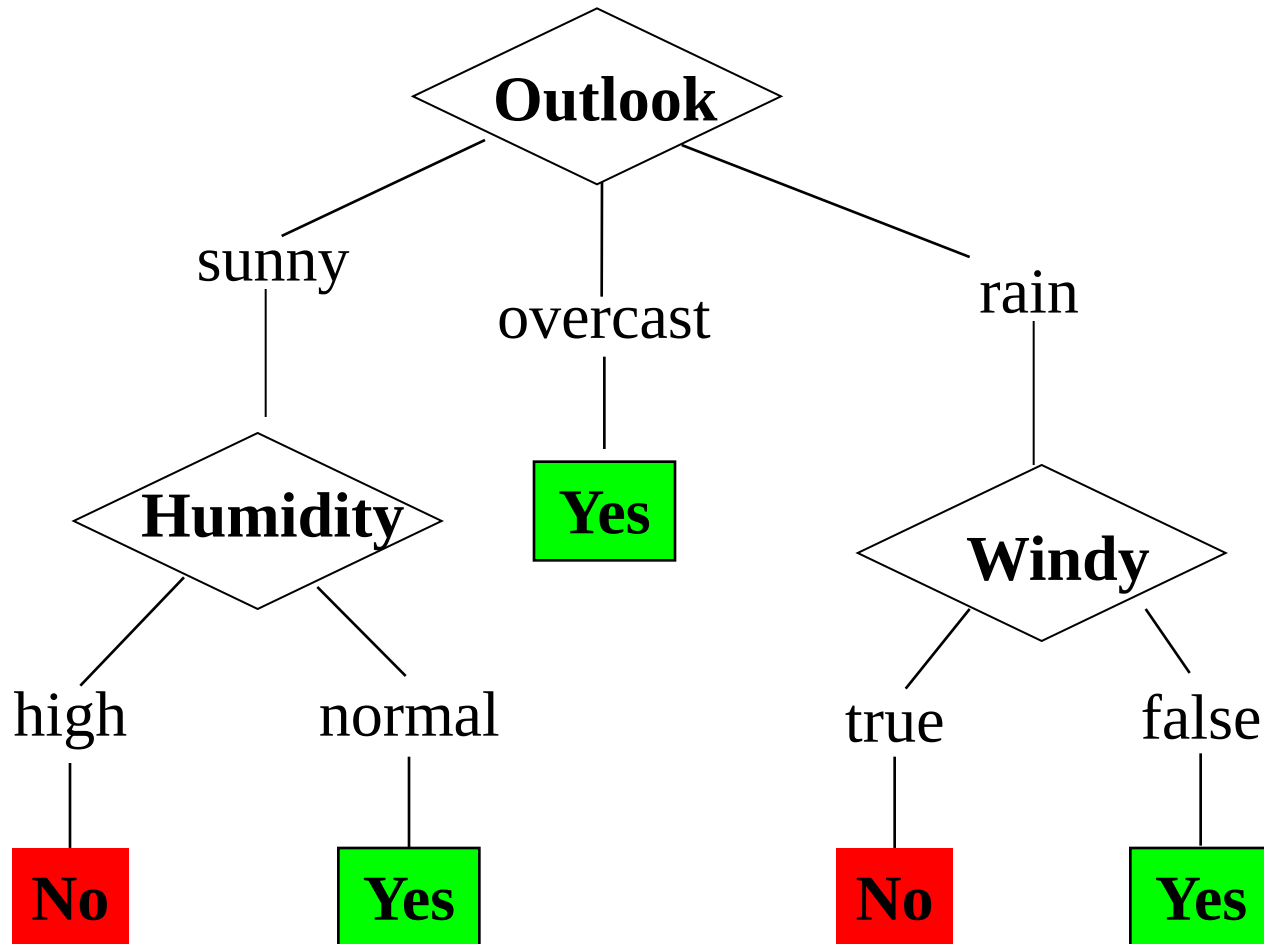
Árboles de Decisión



Árboles de Decisión



Árboles de Decisión





Árboles de Decisión

Como seleccionar el atributo para separar?

- Se usa una función de calidad
 - Ganancia de información
 - Razón de Ganancia de información
 - Índice “gini”

Árboles de Decisión

Ganancia de información

$$\begin{aligned}\text{ganancia}(T, x_i) &= (\text{info antes de separar}) - (\text{info después de separar}) \\ &= \text{info}([\#C_1, \#C_2]) - \text{info}([\#C_1^{v_1}, \#C_2^{v_1}], \dots, [\#C_1^{v_s}, \#C_2^{v_s}])\end{aligned}$$

$$\text{info}([x, y]) = \text{entropía}\left(\frac{x}{x+y}, \frac{y}{x+y}\right)$$

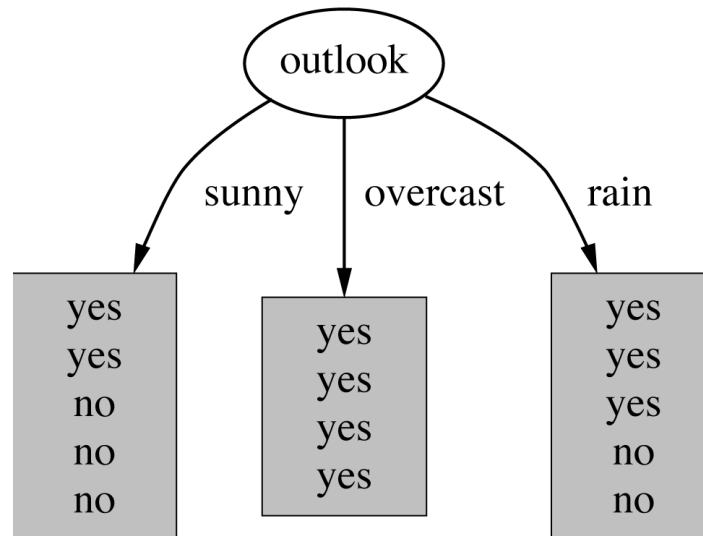
$$\text{entropía}(x, y) = -x \log(x) - y \log(y) \quad (\text{asumimos } 0 \log(0) = 0)$$

$$\text{info}([x_1, y_1], \dots, [x_s, y_s]) = \sum_{j=1}^s \left(\frac{x_j + y_j}{\sum_{k=1}^s (x_k + y_k)} \text{info}([x_j, y_j]) \right)$$

Árboles de Decisión

Ganancia de información

Para el atributo Outlook



Árboles de Decisión

Ganancia de información

- “Outlook” = “Sunny”:

$$\begin{aligned}\text{info}([2,3]) &= \text{entropía}(2/5, 3/5) \\ &= -2/5 \log(2/5) - 3/5 \log(3/5) = 0.971\end{aligned}$$

- “Outlook” = “Overcast”:

$$\begin{aligned}\text{info}([4,0]) &= \text{entropía}(1, 0) \\ &= -1 \log(1) - 0 \log(0) = 0\end{aligned}$$

- “Outlook” = “Rain”:

$$\begin{aligned}\text{info}([3,2]) &= \text{entropía}(3/5, 2/5) \\ &= -3/5 \log(3/5) - 2/5 \log(2/5) = 0.971\end{aligned}$$

Árboles de Decisión

Ganancia de información

$$\text{info}([2,3],[4,0],[3,2]) = (5/14) \times 0.971 + (4/14) \times 0 + (5/14) \times 0.971 \\ = 0.693$$

$$\text{info}([9,5]) = \text{entropía}(9/14, 5/14) \\ = -9/14 \log(9/14) - 5/14 \log(5/14) = 0.940$$

$$\text{ganancia}(T, \text{Outlook}) = \text{info}([9,5]) - \text{info}([2,3],[4,0],[3,2]) \\ = 0.940 - 0.693 = 0.247$$

Árboles de Decisión

Ganancia de información

$$\text{ganancia}(T, \text{Outlook}) = 0.247 \quad \leftarrow \text{Seleccionada}$$

$$\text{ganancia}(T, \text{Temperature}) = 0.029$$

$$\text{ganancia}(T, \text{Humidity}) = 0.152$$

$$\text{ganancia}(T, \text{Windy}) = 0.048$$

Árboles de Decisión

Razón de Ganancia de información

$$\text{razon_ganancia}(T, x_i) = \frac{\text{ganancia}(T, x_i)}{\text{info_atributo}(T, x_i)}$$

$$\text{info_atributo}(T, x_i) \equiv - \sum_{v_j \in V_{x_i}} \left(\frac{|T_{v_j}|}{|T|} \log \left(\frac{|T_{v_j}|}{|T|} \right) \right)$$

donde V_{x_i} = conjunto de valores que toma x_i en T

T_{v_j} = conjunto de objetos de T que toman valor v_j en x_i

Árboles de Decisión

Índice “*gini*”

$$gini(T) = 1 - \sum_{j=1}^r p_j^2$$

donde p_j = frecuencia de la clase j en T

$$gini_{Partición}(T_1, \dots, T_s) = \sum_{k=1}^s \left(\frac{|T_k|}{|T|} gini(T_k) \right)$$



Árboles de Decisión

- ID3 – datos categóricos
- C4.5 – permite datos mezclados, ausencia de información
- CART – Datos con ruido
- C5.0 – Versión comercial

RandomForest

- Para $i = 1 \dots T$ (número de árboles)
 - Seleccionar aleatoriamente con reemplazo $n^* < n$ objetos del conjunto de entrenamiento
 - Crear un árbol de decisión C_i con los objetos seleccionados, pero para expandir cada nodo seleccionar aleatoriamente m atributos y con éstos seleccionar la mejor división.
- Combinar los resultados de $C_1 \dots C_T$
 - Usualmente por votación



Clasificación Supervisada

Árboles de Decisión

Jesús Ariel Carrasco Ochoa

Instituto Nacional de Astrofísica, Óptica y Electrónica