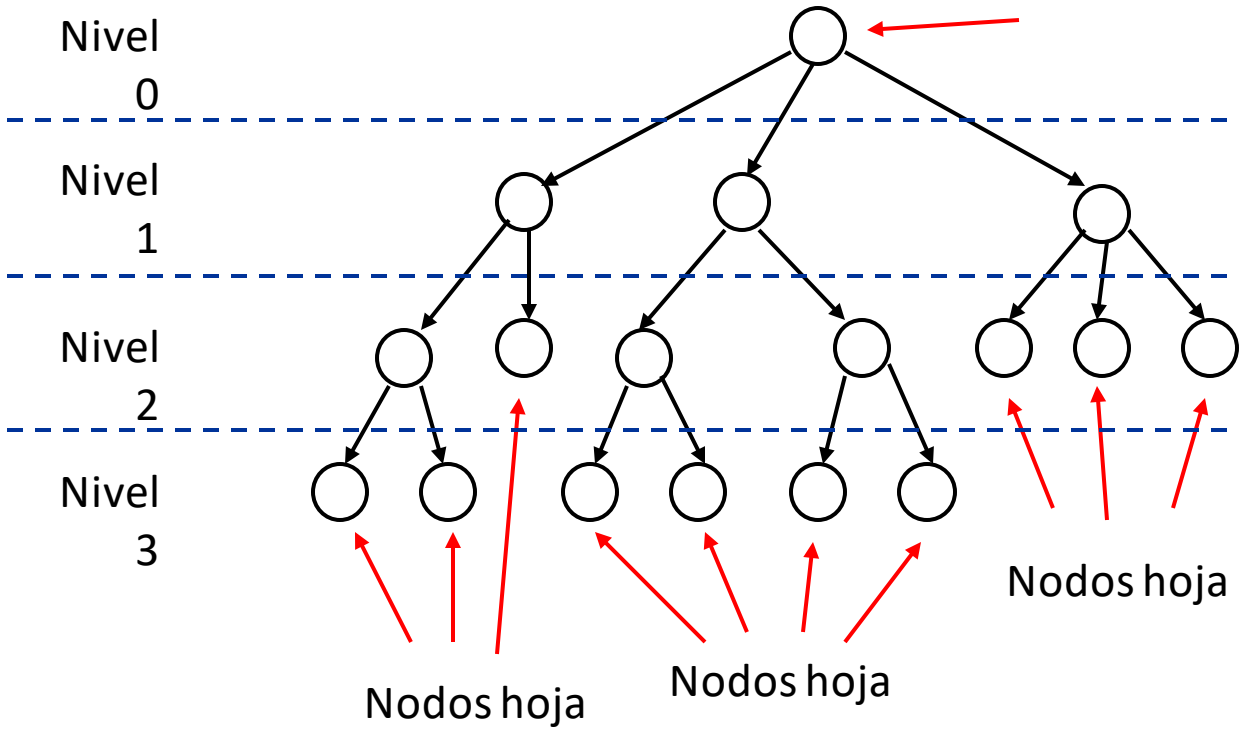
Árboles de Clasificación

Mario Hernández

¿Qué es un árbol? …

… Nos referimos a la estructura de datos …

- Estructura de datos compuesta por nodos y enlaces dirigidos, que unen nodos sin formar ciclos.
- Cada nodo está unido a su “padre” por un enlace.
- Nodo raíz → El único nodo que no tiene nodo padre y corresponde con el nivel 0
- Nodos hoja → Nodos que no tienen nodos hijos.
- Profundidad del árbol → Nivel del nodo hoja más lejano de la raíz
- Balanceado/No Balanceado
- Tipos de árboles (si están balanceados):
  - Binario: 2 hijos/nodo
  - Ternario: 3 hijos/nodo
  - N-ario: n hijos/nodo



Árbol de Decisión

- El problema de decisión se basa en muestras SUPERVISADAS en forma de pares atributo-clase o atributo-valor
- Es tipológico, es decir, es un árbol donde cada nodo contiene una comparación sobre algún atributo del problema.
- El número de enlaces que salen de un determinado nodo (y por tanto el número de hijos) es igual al número de valores que puede tomar el atributo sobre el que se ha hecho la comparación.
- Cada nodo hoja representa una clase o una distribución de clases.
- Cada caso nuevo se clasifica siguiendo los resultados de las comparaciones desde la raíz hasta un nodo hoja.
- Un Árbol de Decisión se usa en dos fases:
  - construcción o aprendizaje
  - explotación o uso

Tipos de árboles de decisión

- **Árboles de Clasificación**
- **Árboles de Regresión**

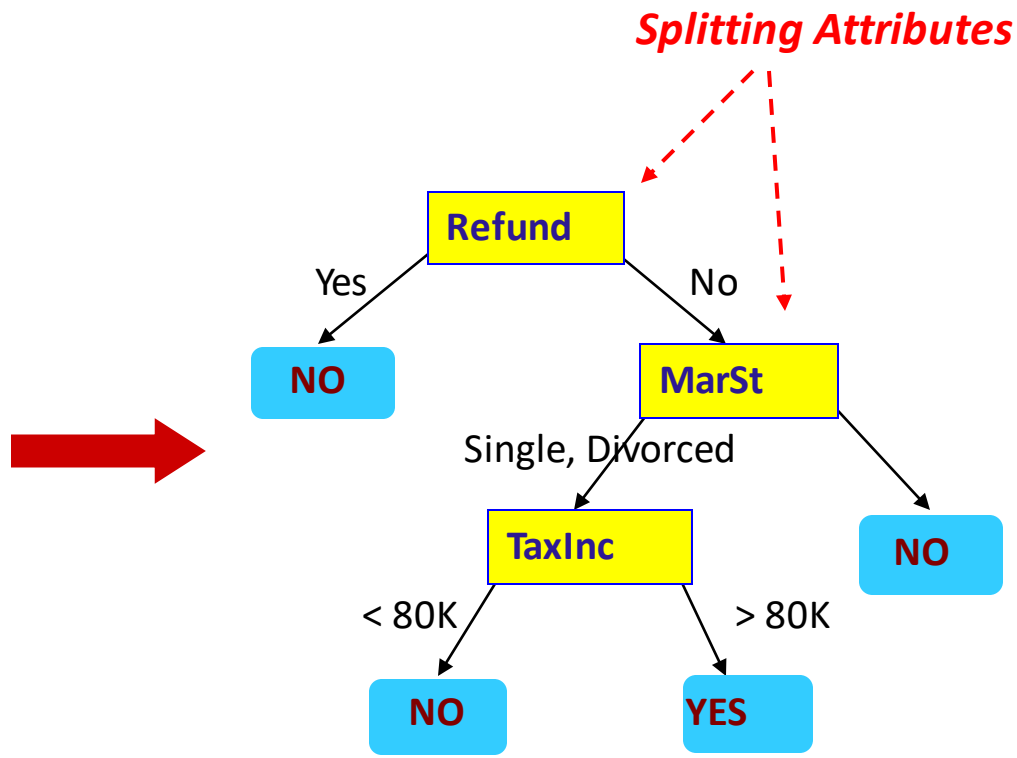
Árboles de Clasificación

- Utilizados con **datos categóricos**
- Asignación de clases $\Rightarrow$ Cada nodo hoja corresponde a una categoría
  - El resultado que se predice es la clase o categoría a la cual pertenecen los datos
  - Pe: para clasificar préstamos como “seguro” o “de riesgo”

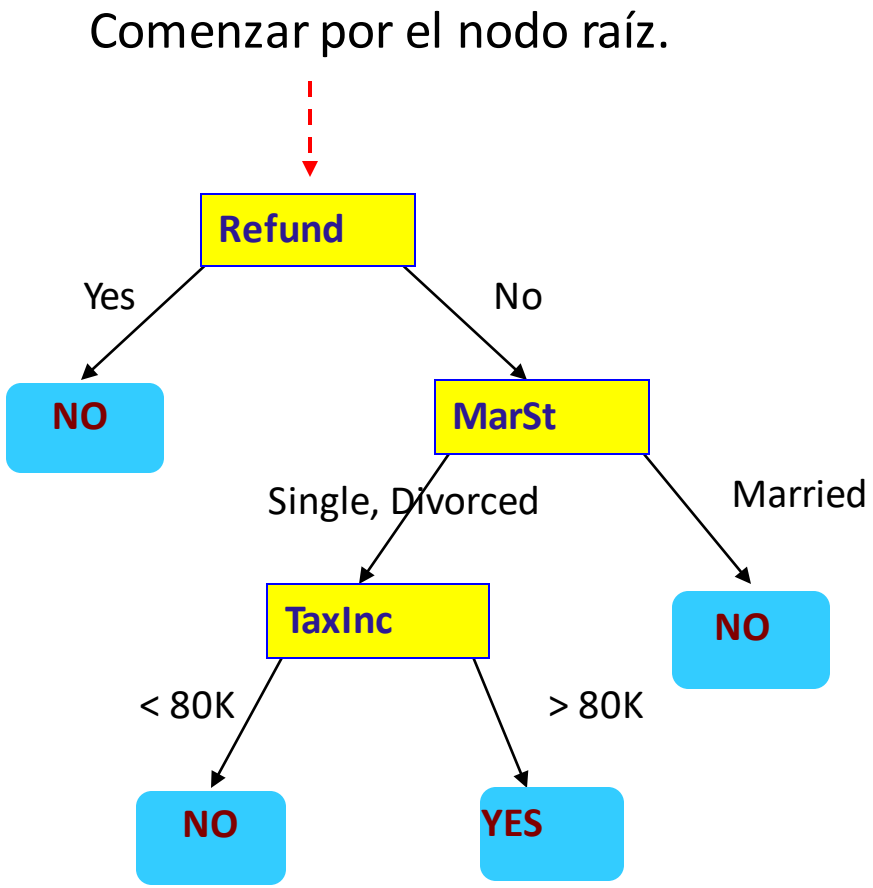
Árboles de Clasificación

Árboles de Clasificación

	categorical	categorical	continuous	class
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**



Clasificación de un nuevo caso

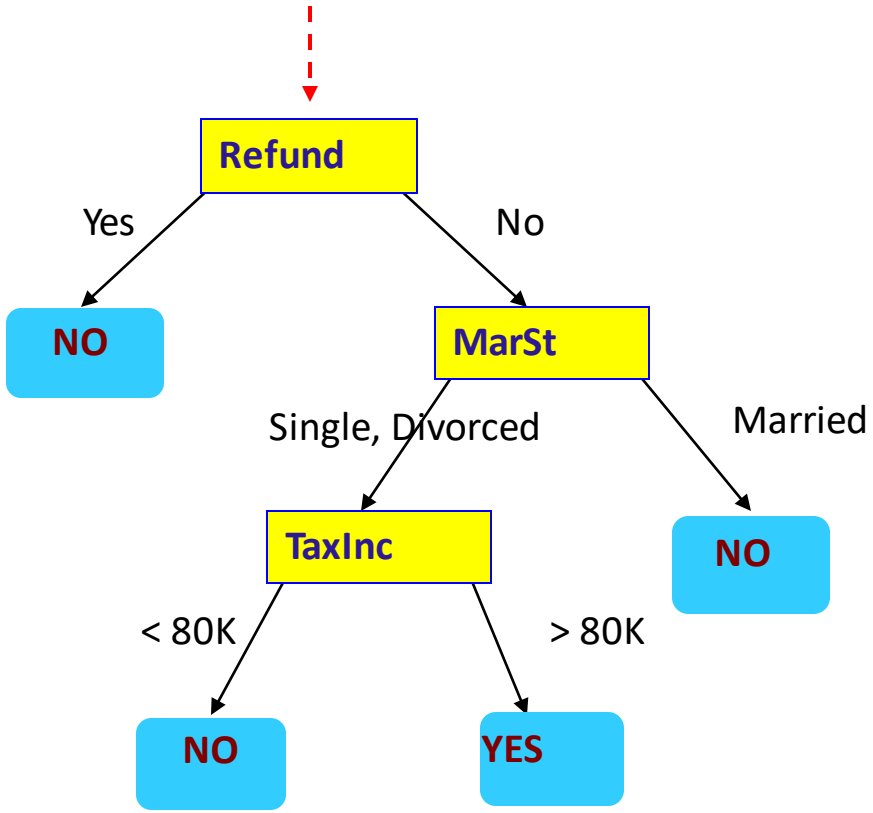


Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Clasificación de un nuevo caso

Comenzar por el nodo raíz.



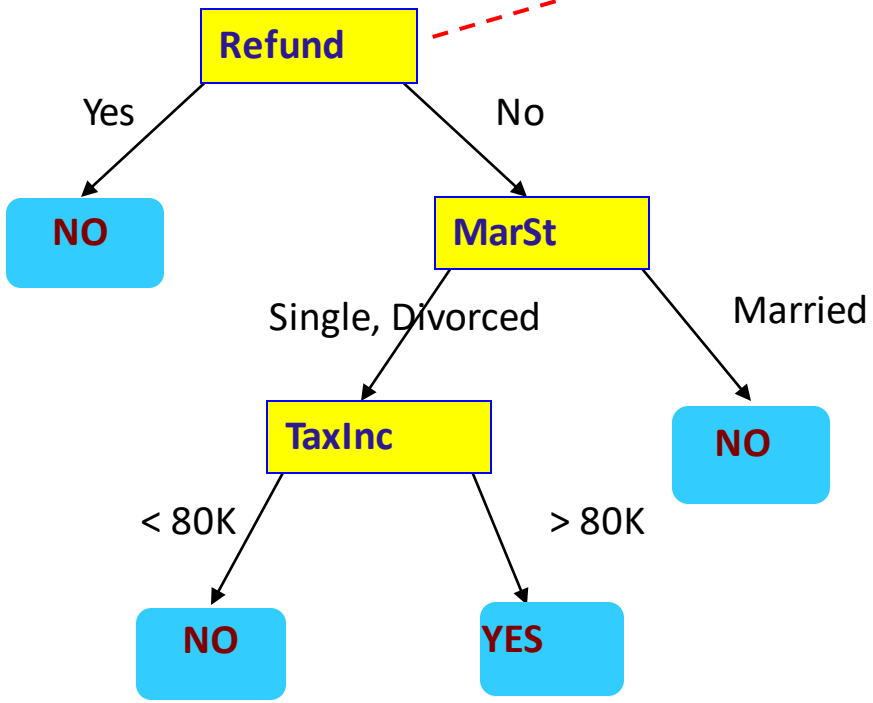
Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Clasificación de un nuevo caso

Test Data

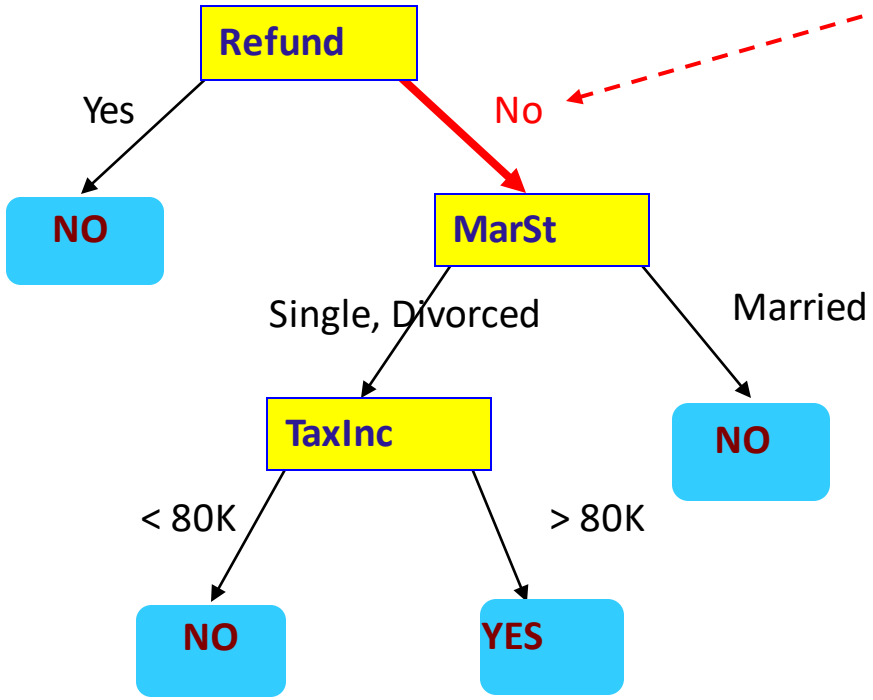
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Clasificación de un nuevo caso

Test Data

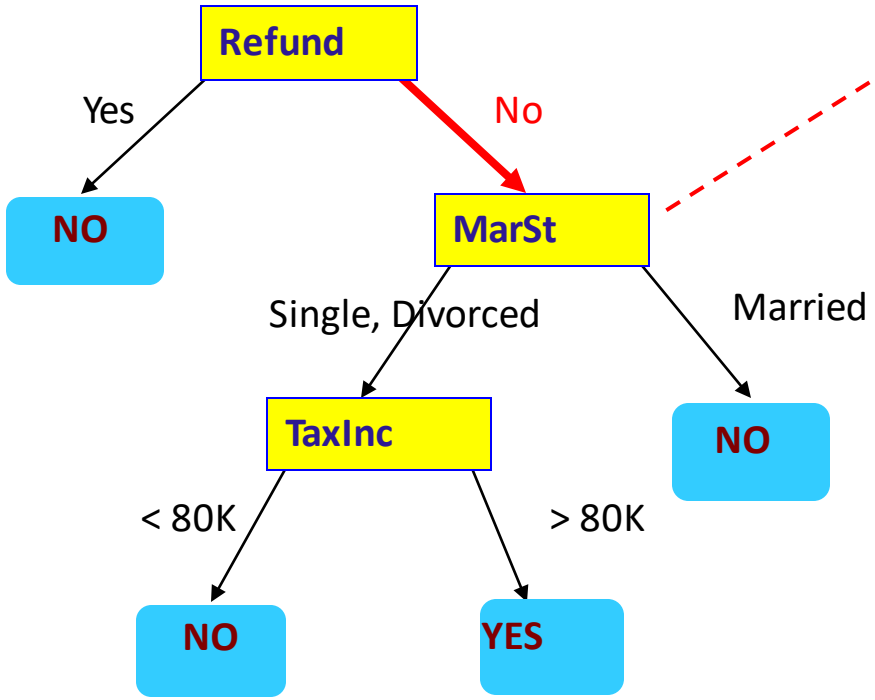
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Clasificación de un nuevo caso

Test Data

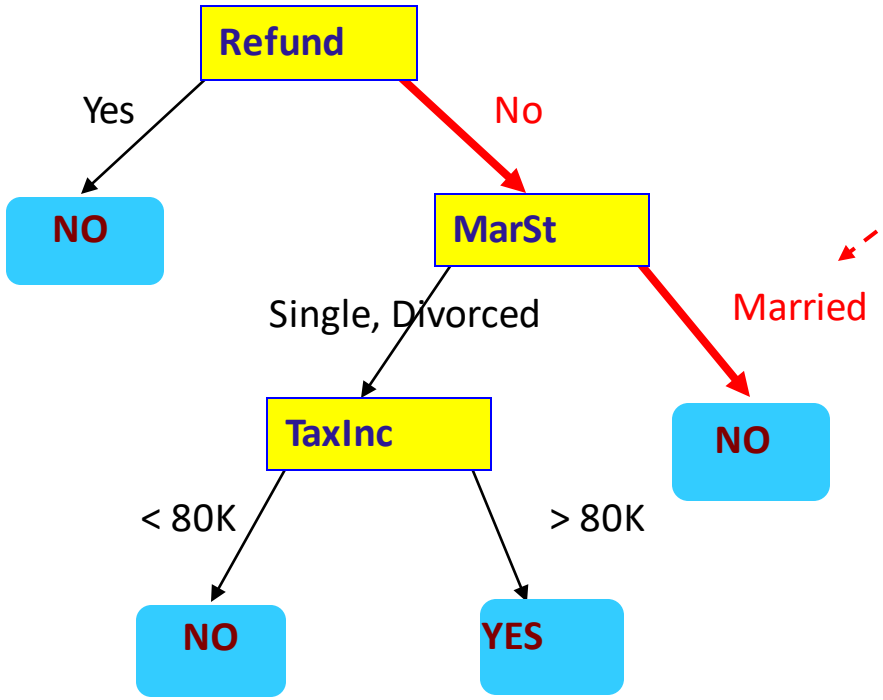
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Clasificación de un nuevo caso

Test Data

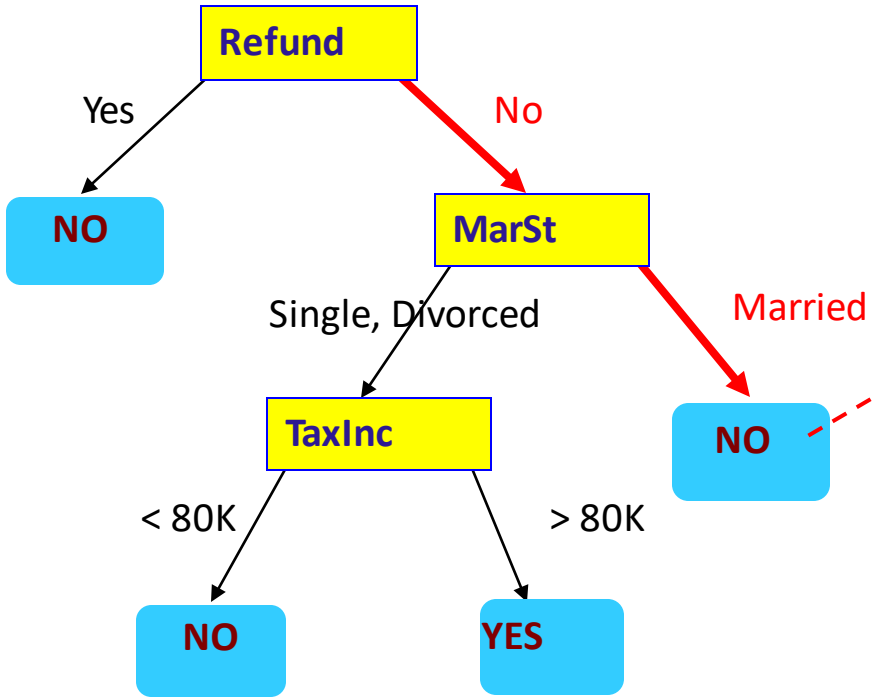
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Clasificación de un nuevo caso

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

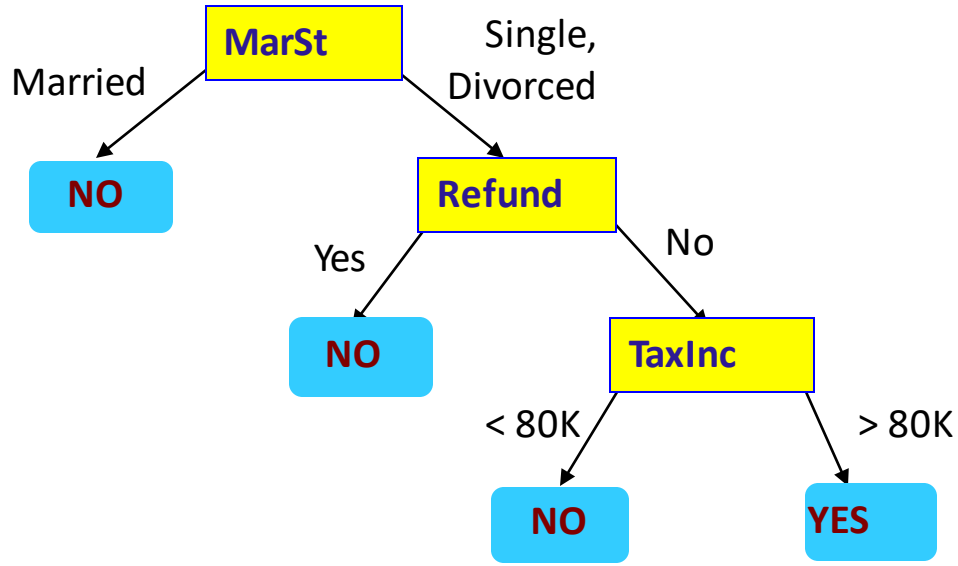


Asignar Cheat a “No”

Otro Árbol de Clasificación para el mismo ejemplo

	categorical	categorical	continuous	class
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

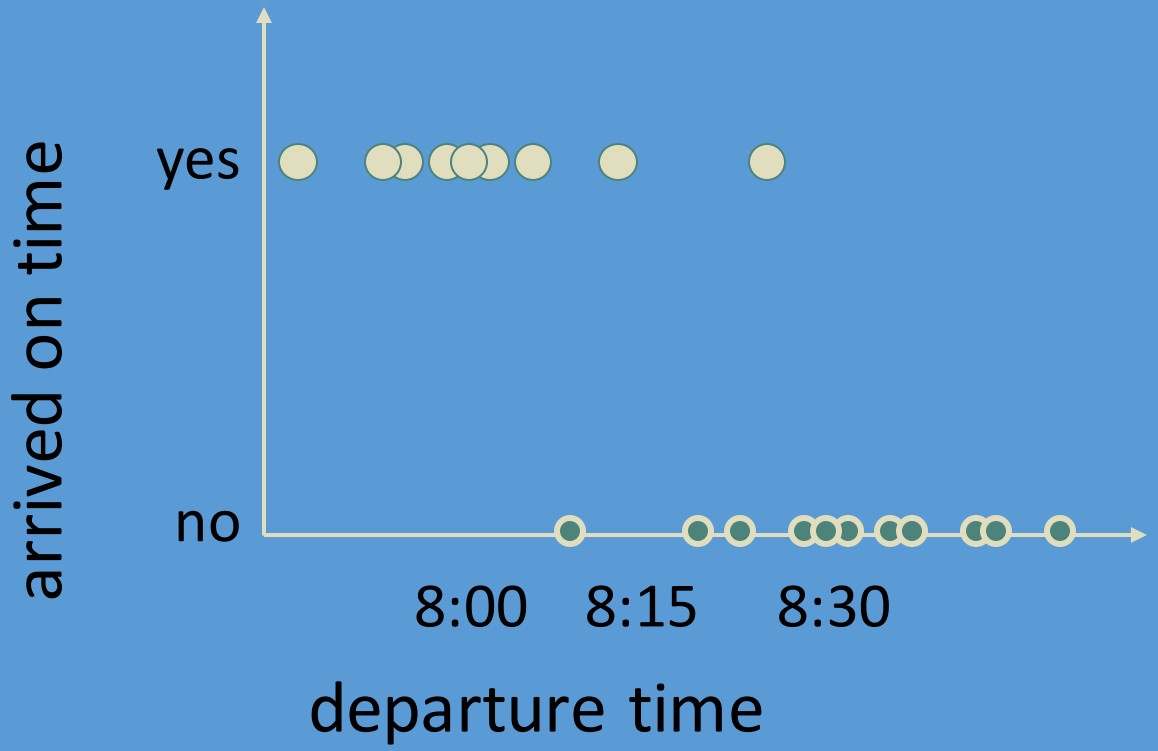
Conjunto de atributos valor

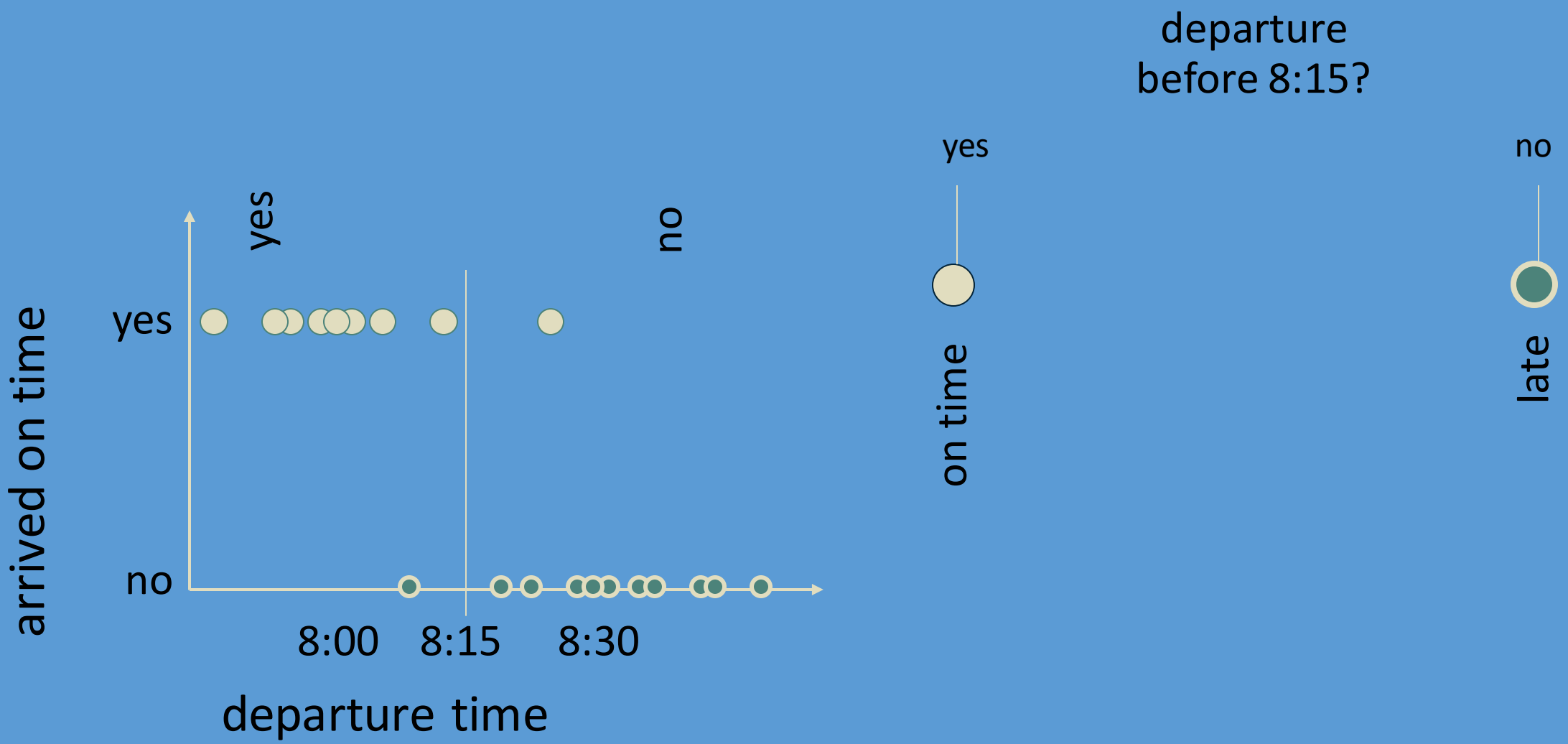


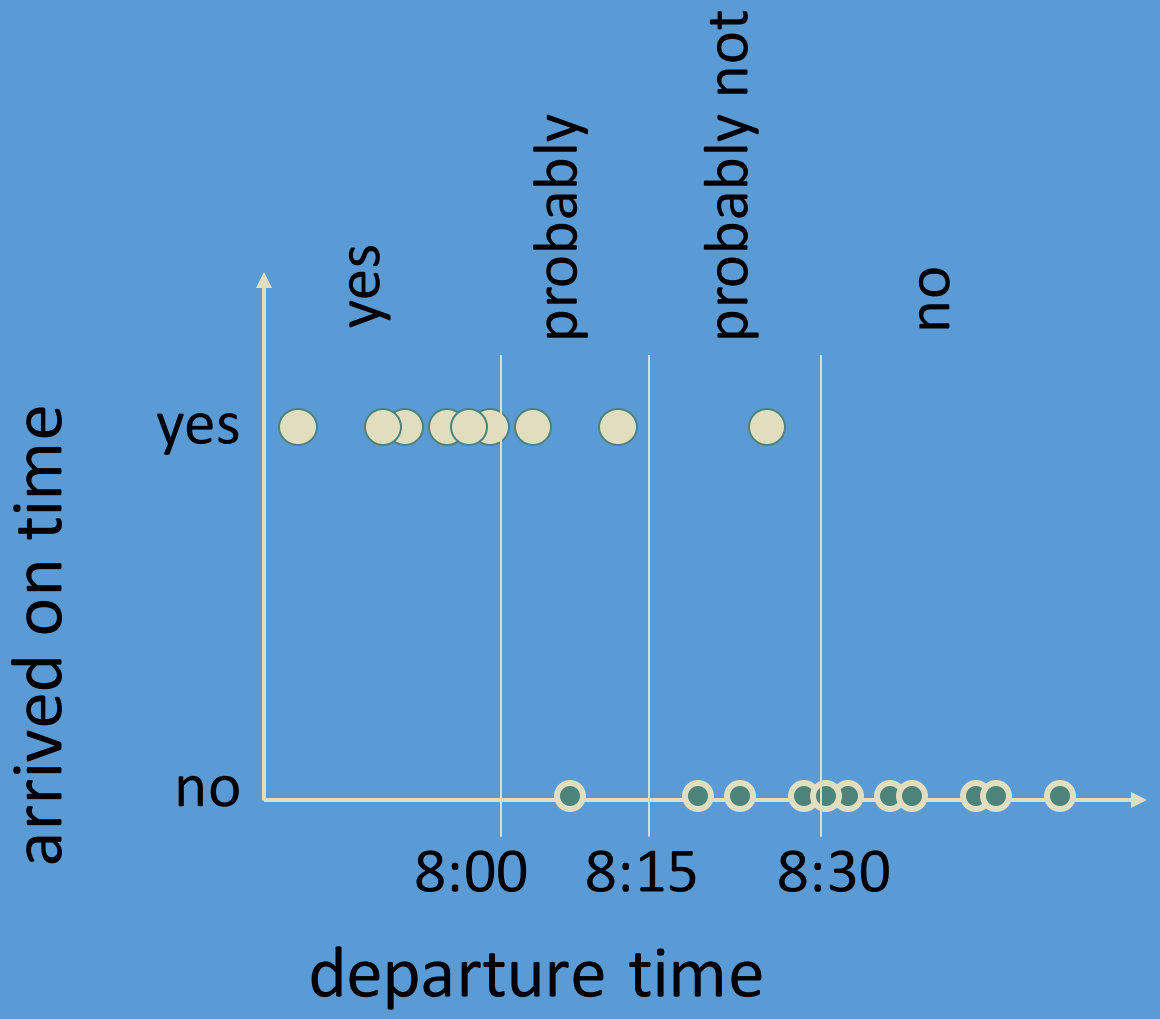
Puede obtenerse más de un árbol para el mismo conjunto de datos cambiando las estrategias!

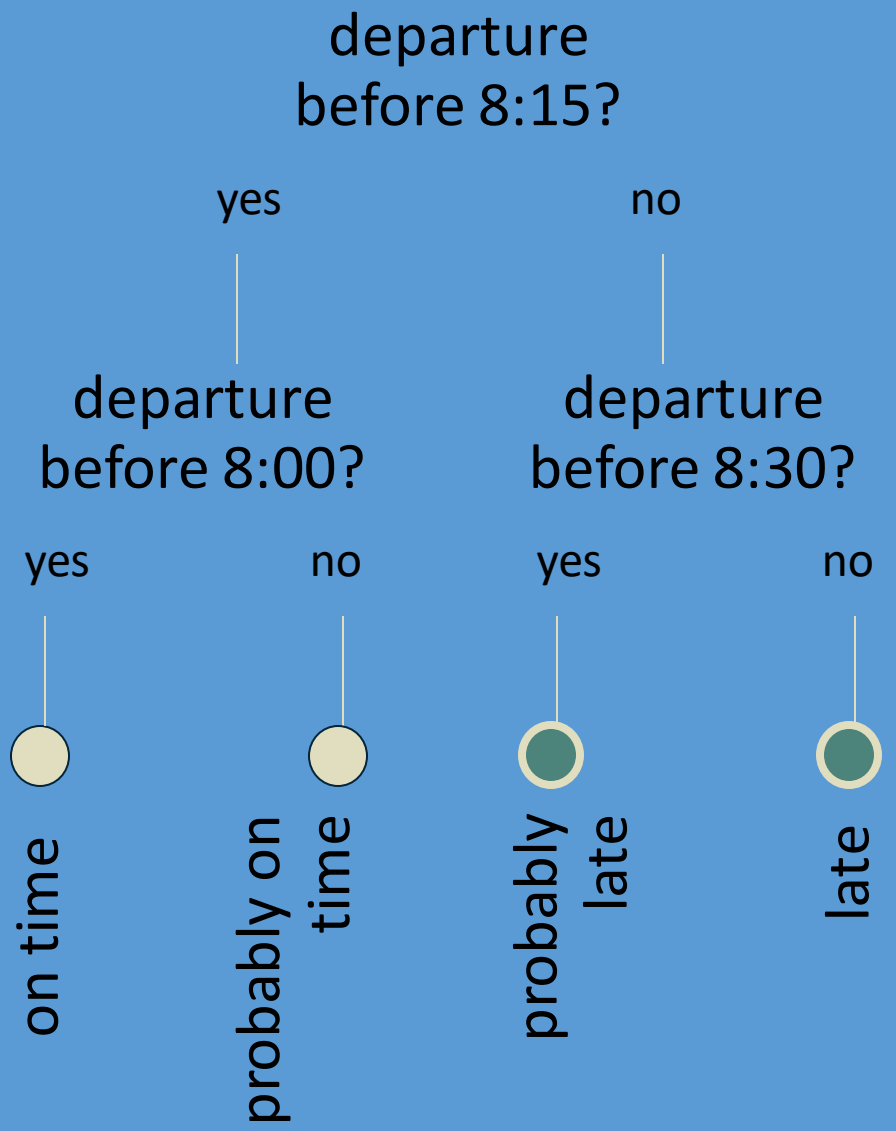
Otro ejemplo de árbol de clasificación

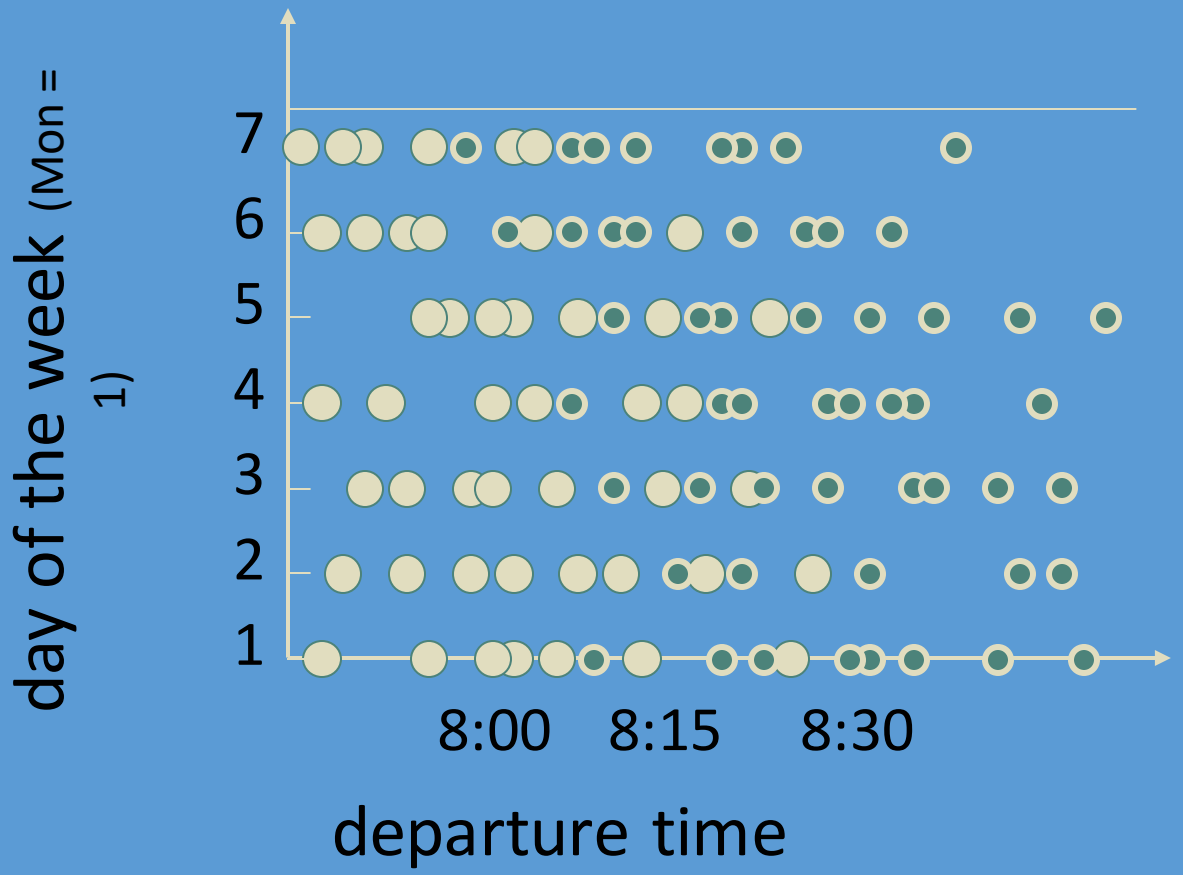
Ejemplo: ¿Se puede modelar con un árbol si una persona llega a tiempo al trabajo en relación a la hora en que salió de su casa?

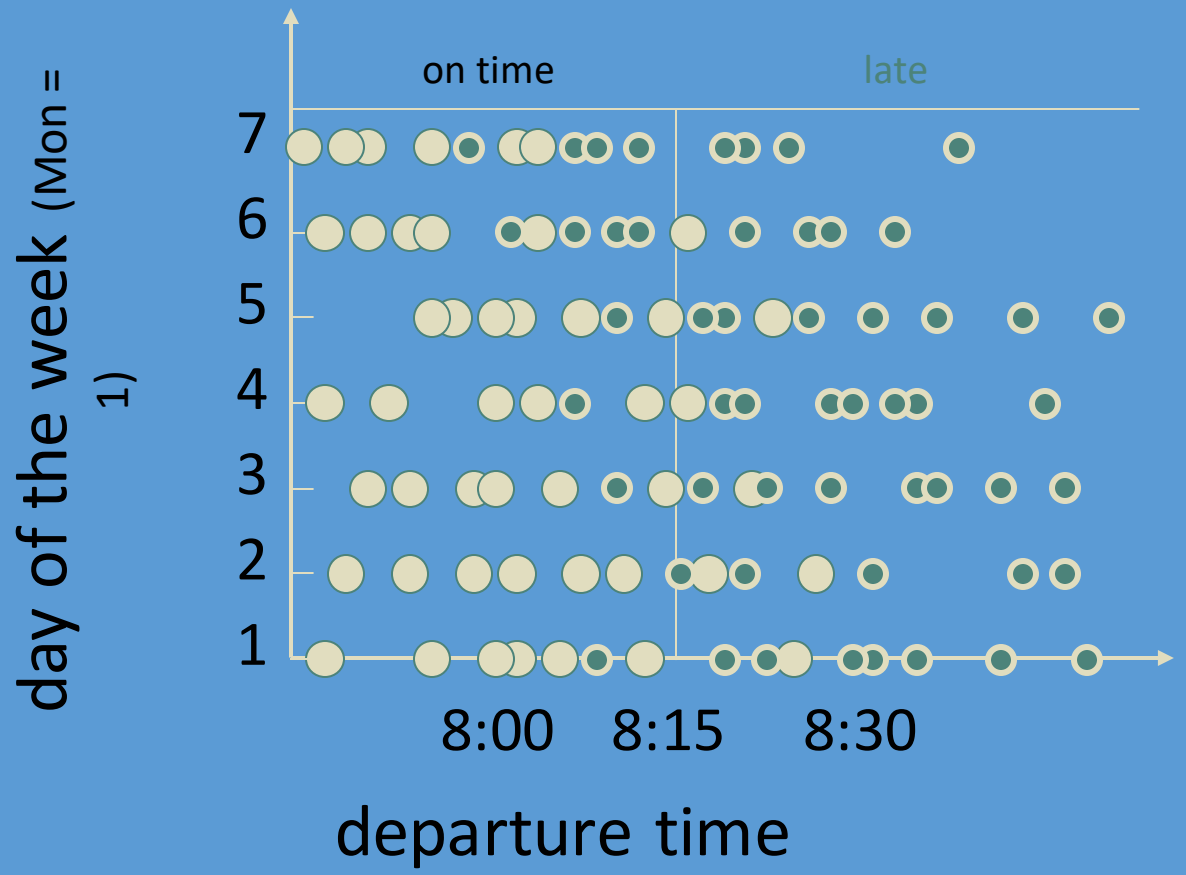

arrived on time
yes
no
8:00
8:15
8:30
departure time

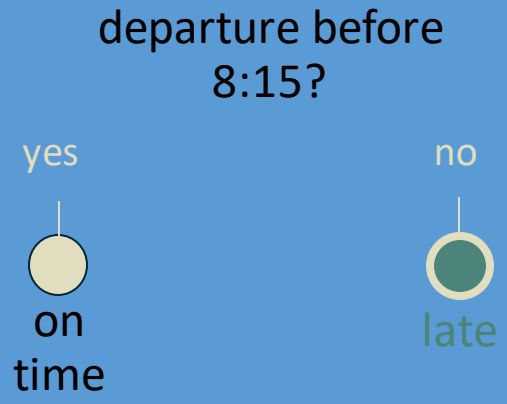

departure
before 8:15?
yes
no
on time
late
yes
no
arrived on time
yes
no
8:00
8:15
8:30
departure time

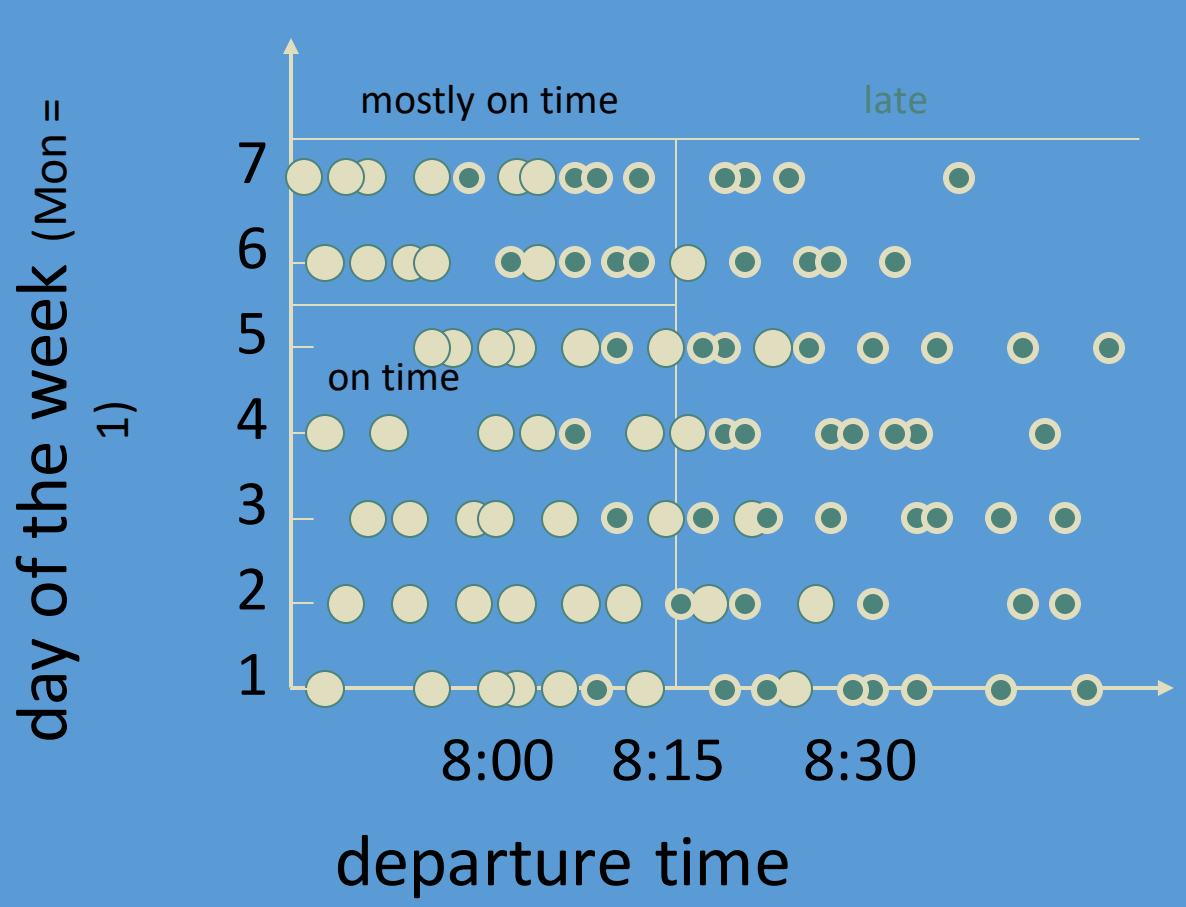

yes
probably
probably not
no
arrived on time
yes
no
8:00
8:15
8:30
departure time

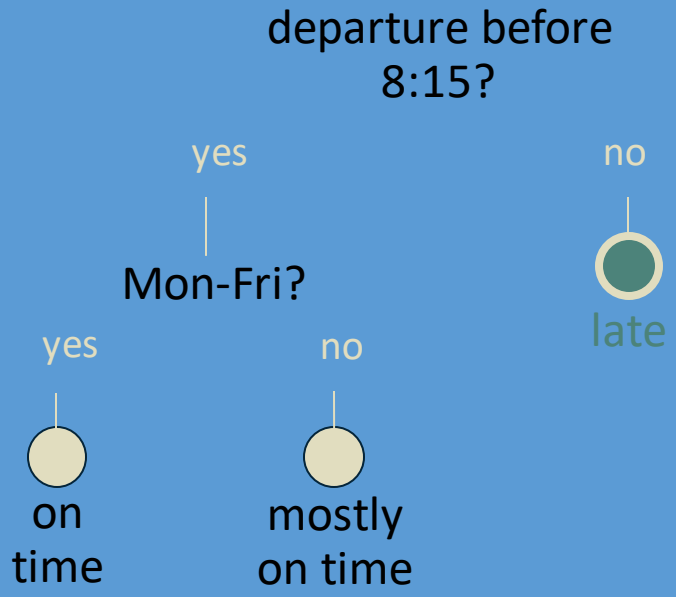

departure before 8:15?
yes
no
departure before 8:00?
departure before 8:30?
yes
no
yes
no
on time
probably on time
probably late
late

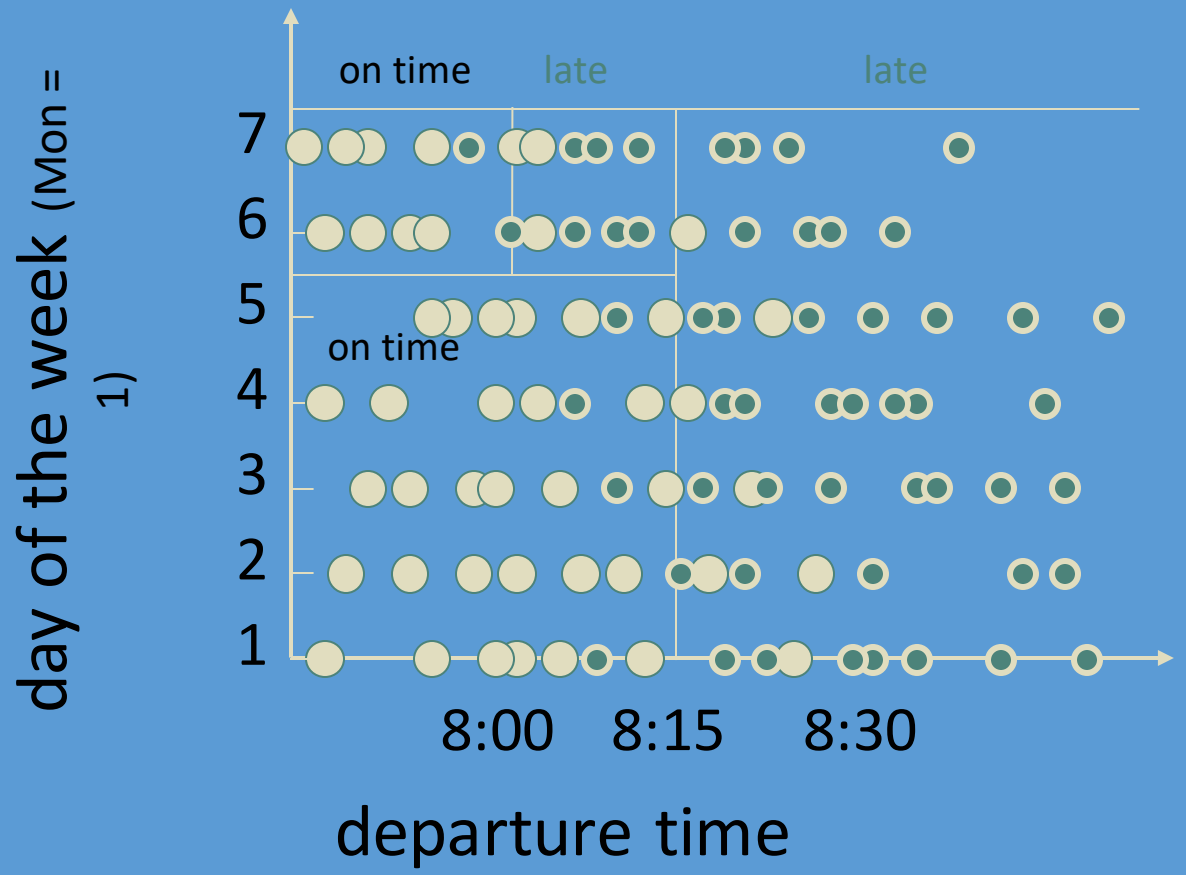

day of the week (Mon = 1)
7
6
5
4
3
2
1
8:00
8:15
8:30
departure time

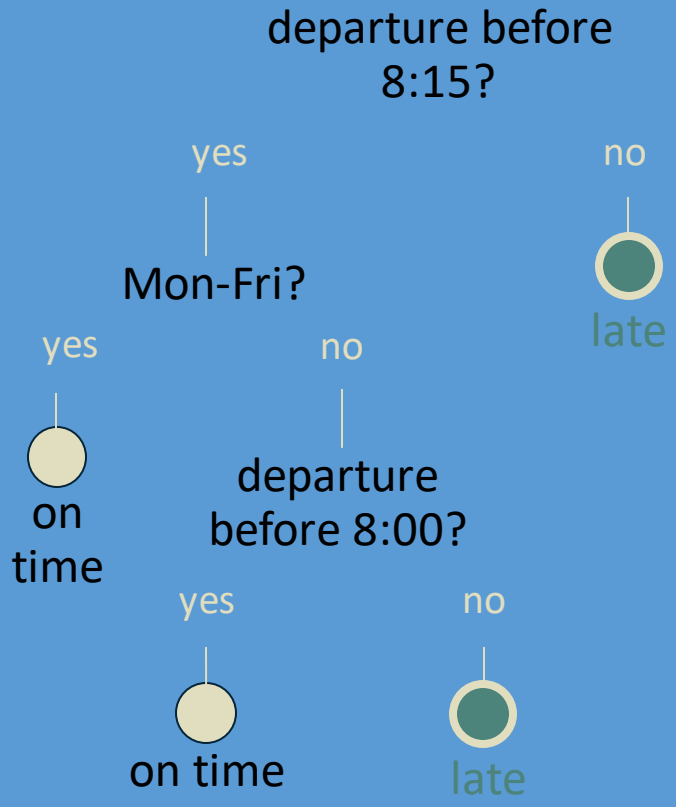

on time
late
7
6
5
4
3
2
1
day of the week (Mon = 1)
8:00
8:15
8:30
departure time


departure before 8:15?
yes
no
on time
late


day of the week (Mon = 1)
mostly on time
late
on time
7
6
5
4
3
2
1
8:00
8:15
8:30
departure time


departure before 8:15?
yes
no
late
Mon-Fri?
yes
no
on time
mostly on time

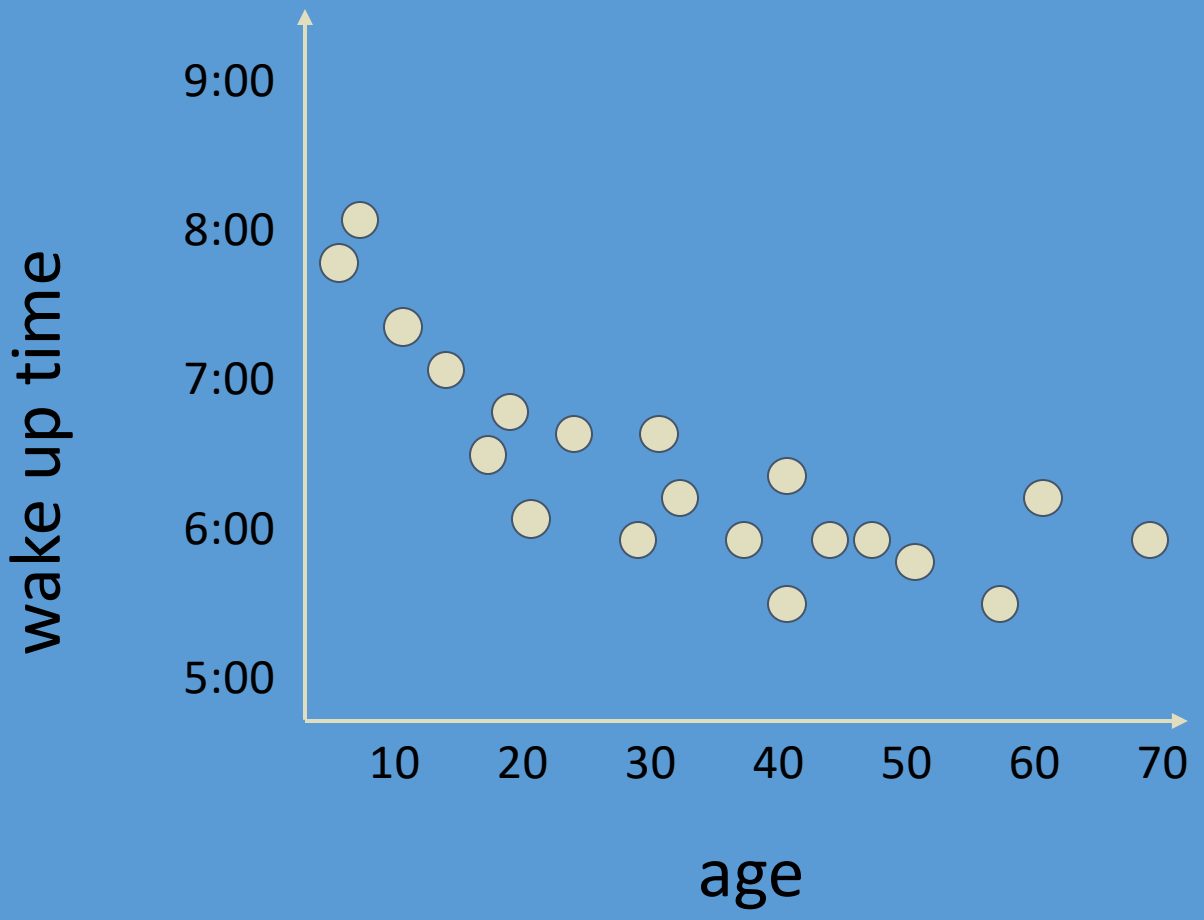

day of the week (Mon = 1)
7
6
5
4
3
2
1
on time
late
late
on time
8:00
8:15
8:30
departure time


departure before 8:15?
yes
no
late
Mon-Fri?
yes
no
on time
departure before 8:00?
yes
no
on time
late

Árboles de Regresión

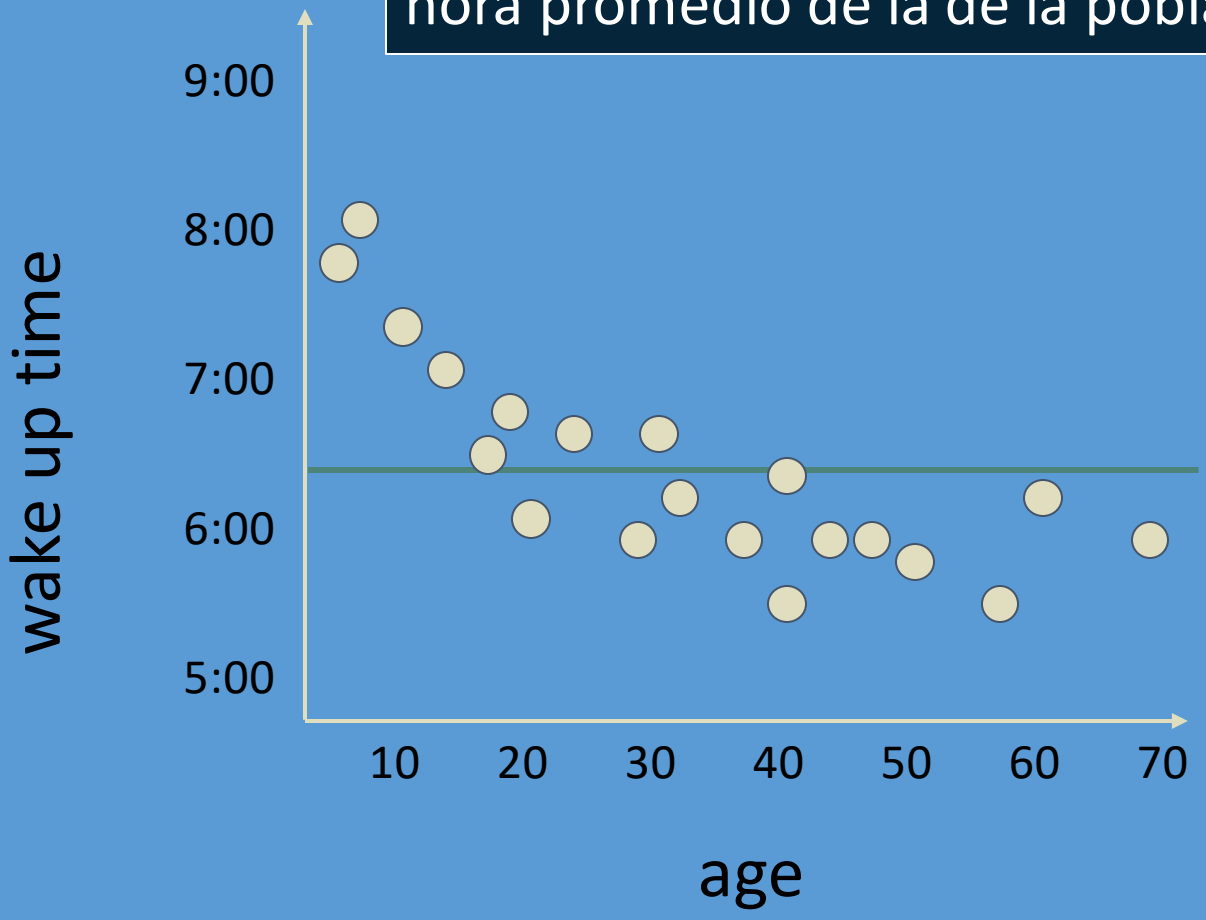
- Utilizados con **datos contínuos** (números reales)
- El resultado que se predice es una cantidad numérica (en general, un número real)
- Cada nodo hoja es un valor numérico que corresponde al promedio de las muestras que llegan al nodo hoja.

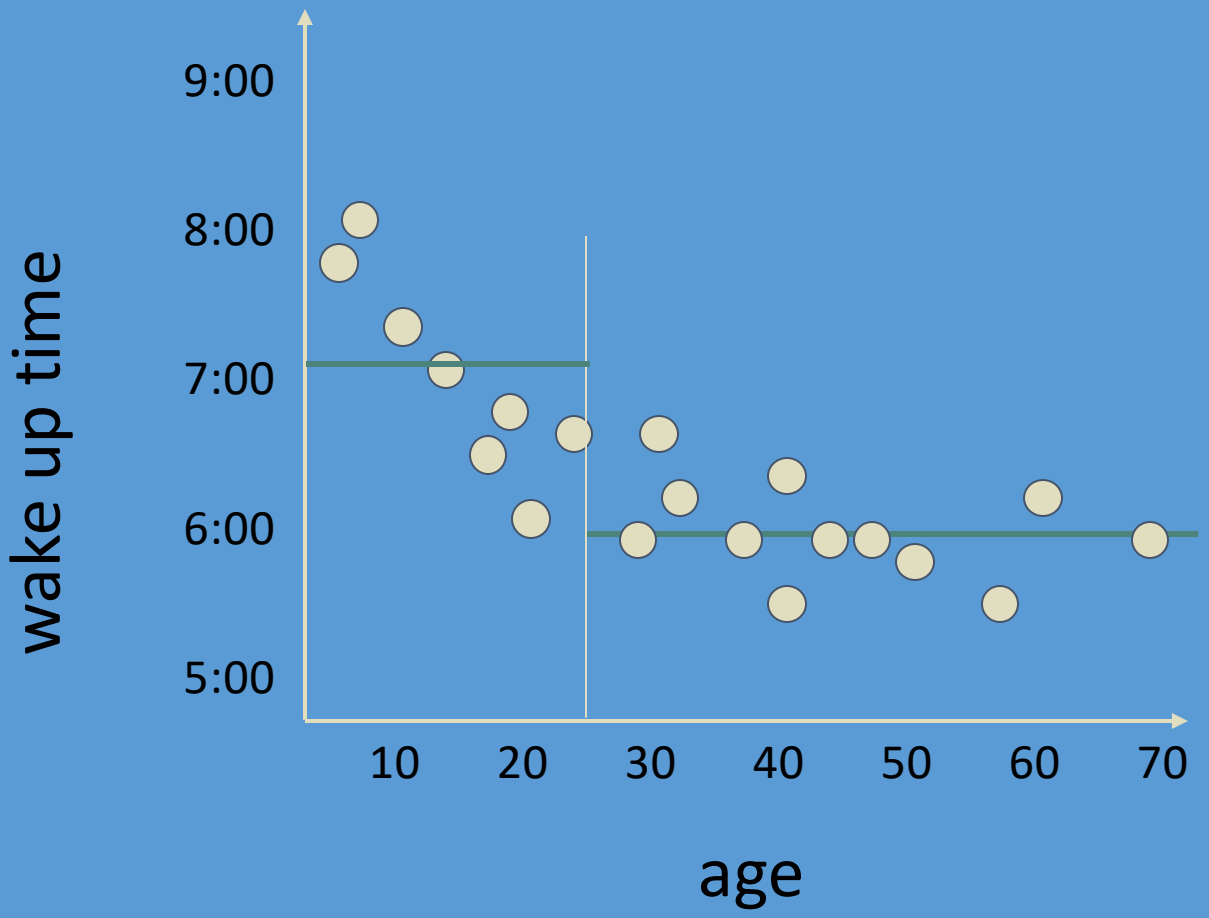
Ejemplo: ¿cómo predice la edad de una persona la hora a la que se despierta?


wake up time
9:00
8:00
7:00
6:00
5:00
10
20
30
40
50
60
70
age

Raiz del árbol - estimación sin saber la edad:
hora promedio de la de la población considerada

6:25





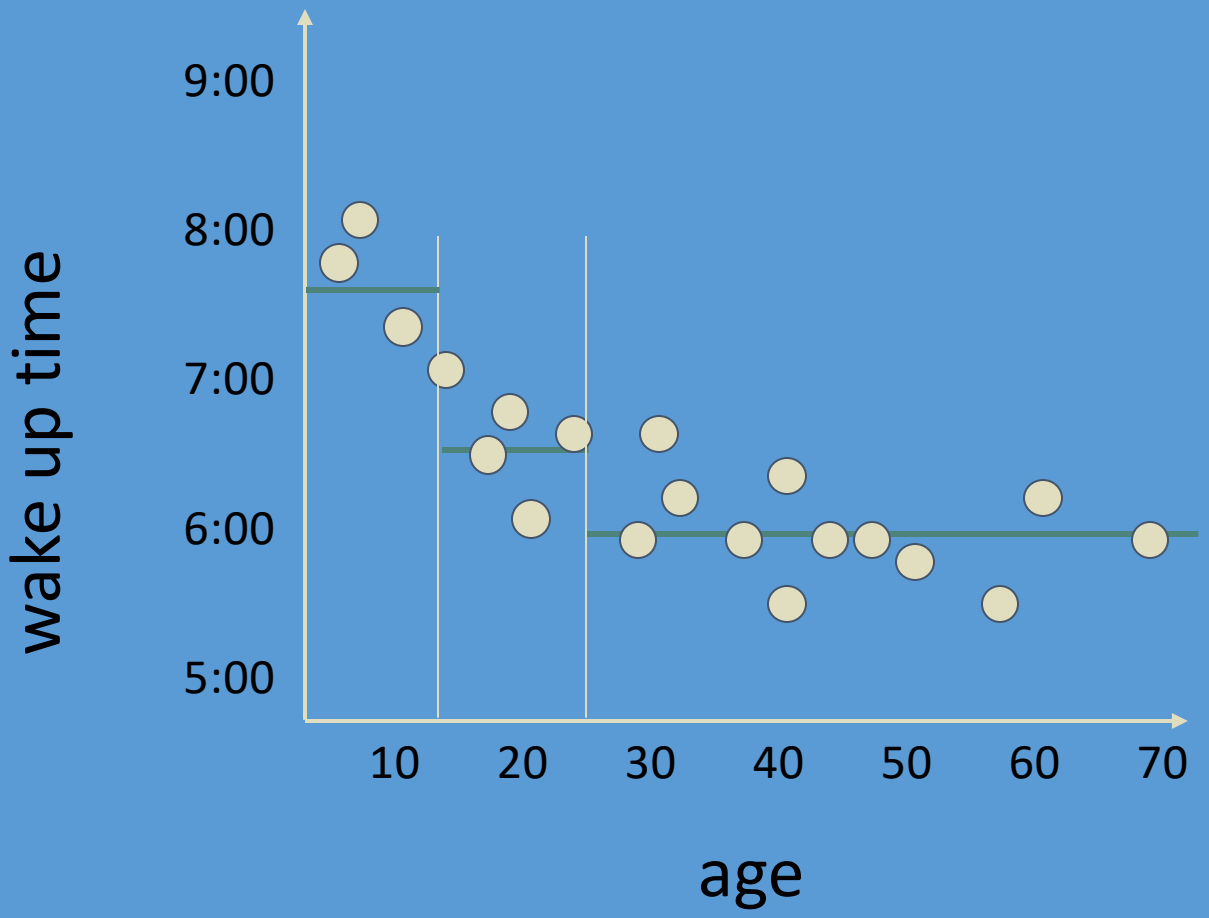
younger than 25?

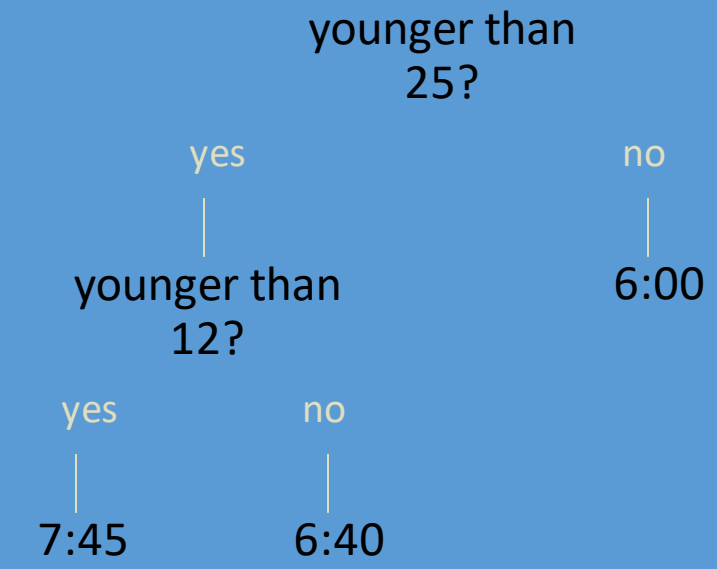
yes

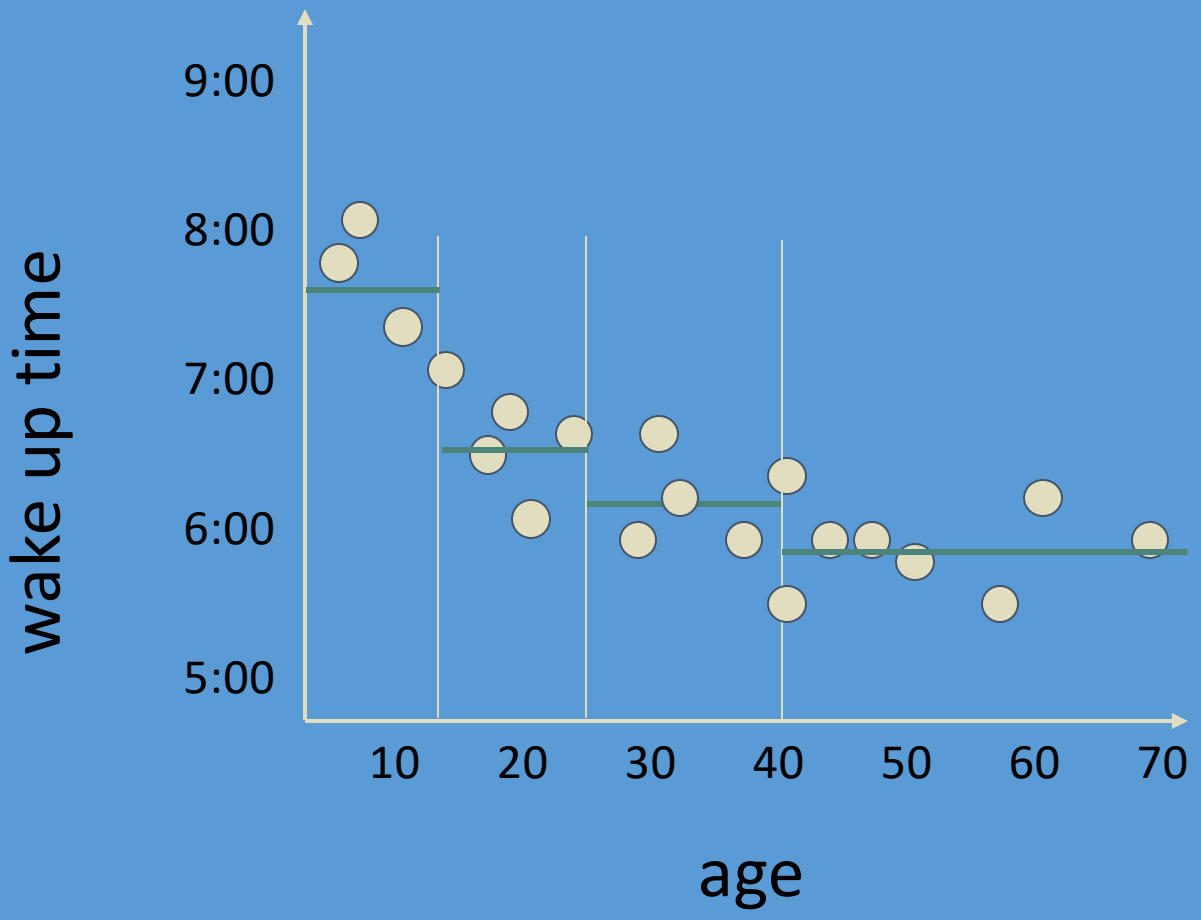
7:05

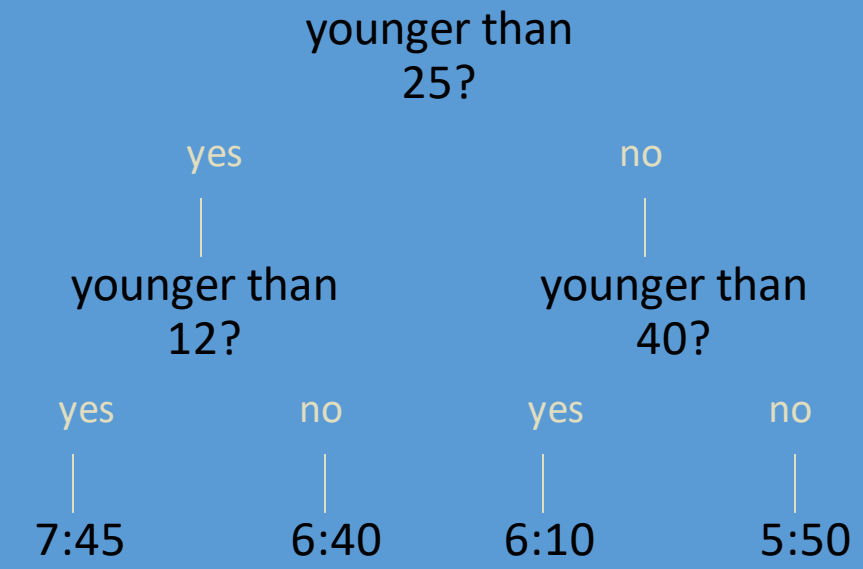
no

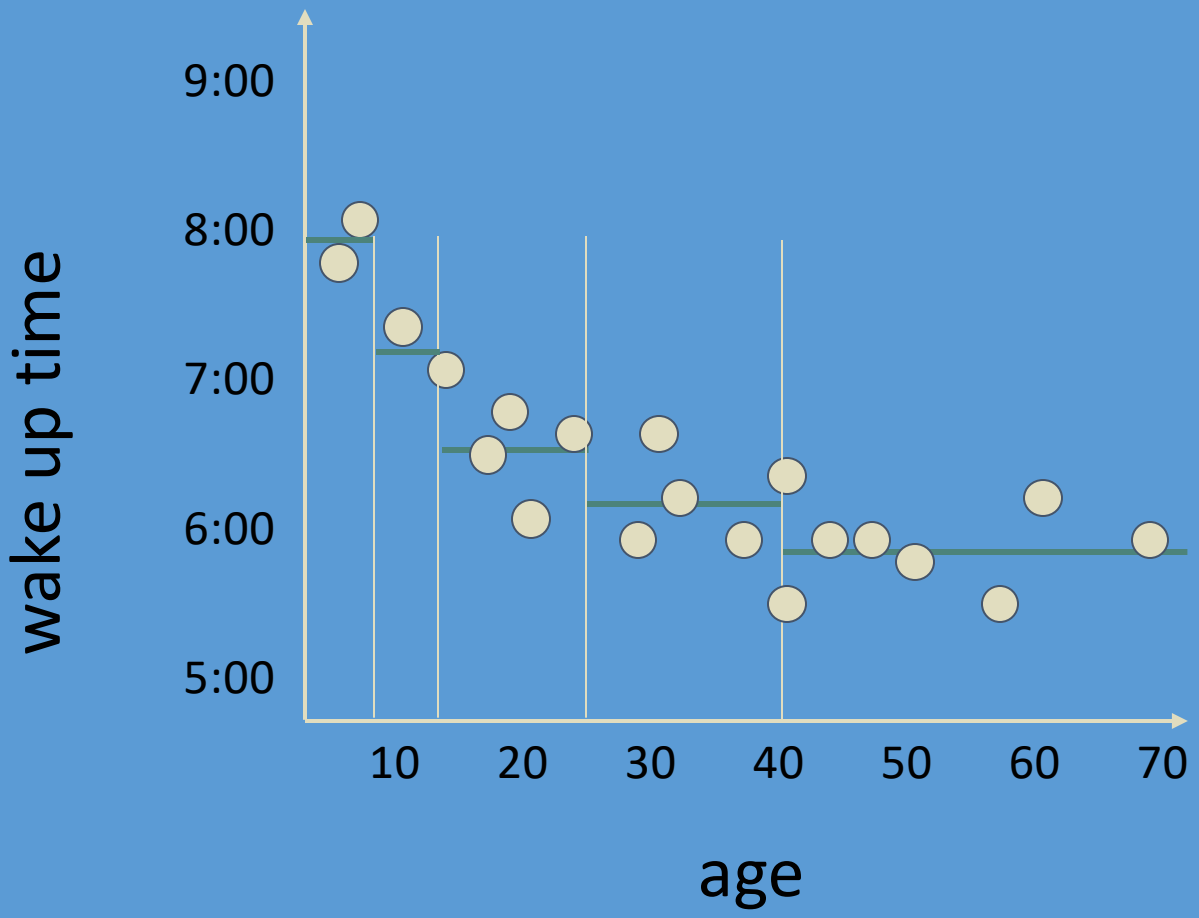
6:00

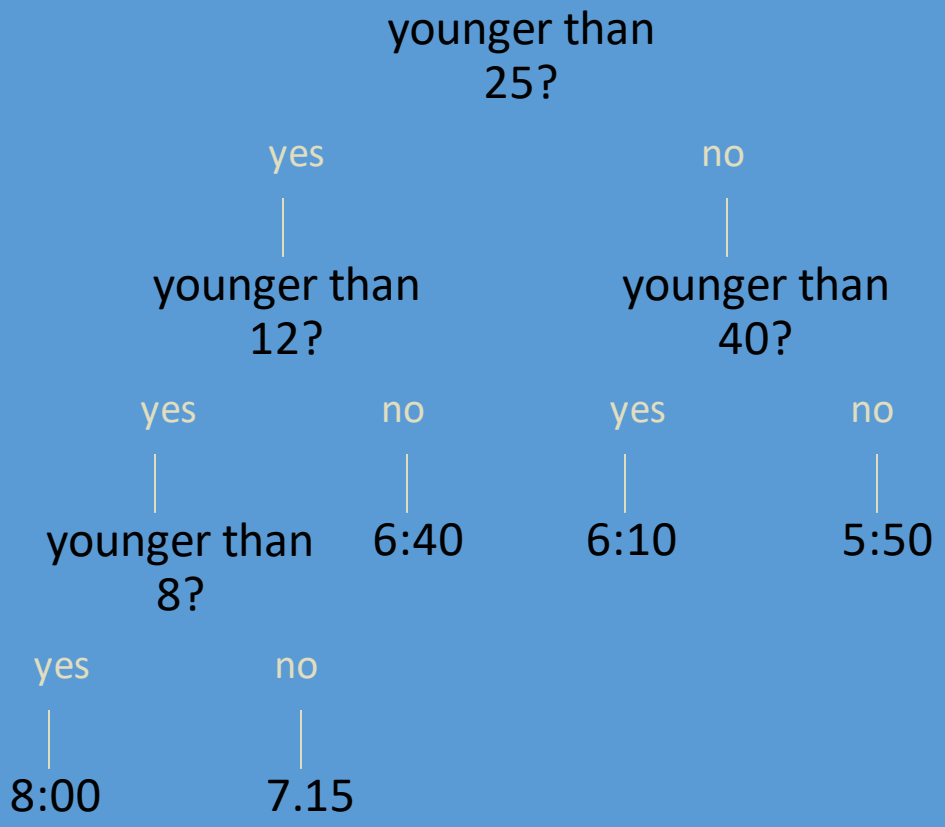

wake up time
9:00
8:00
7:00
6:00
5:00
10
20
30
40
50
60
70
age

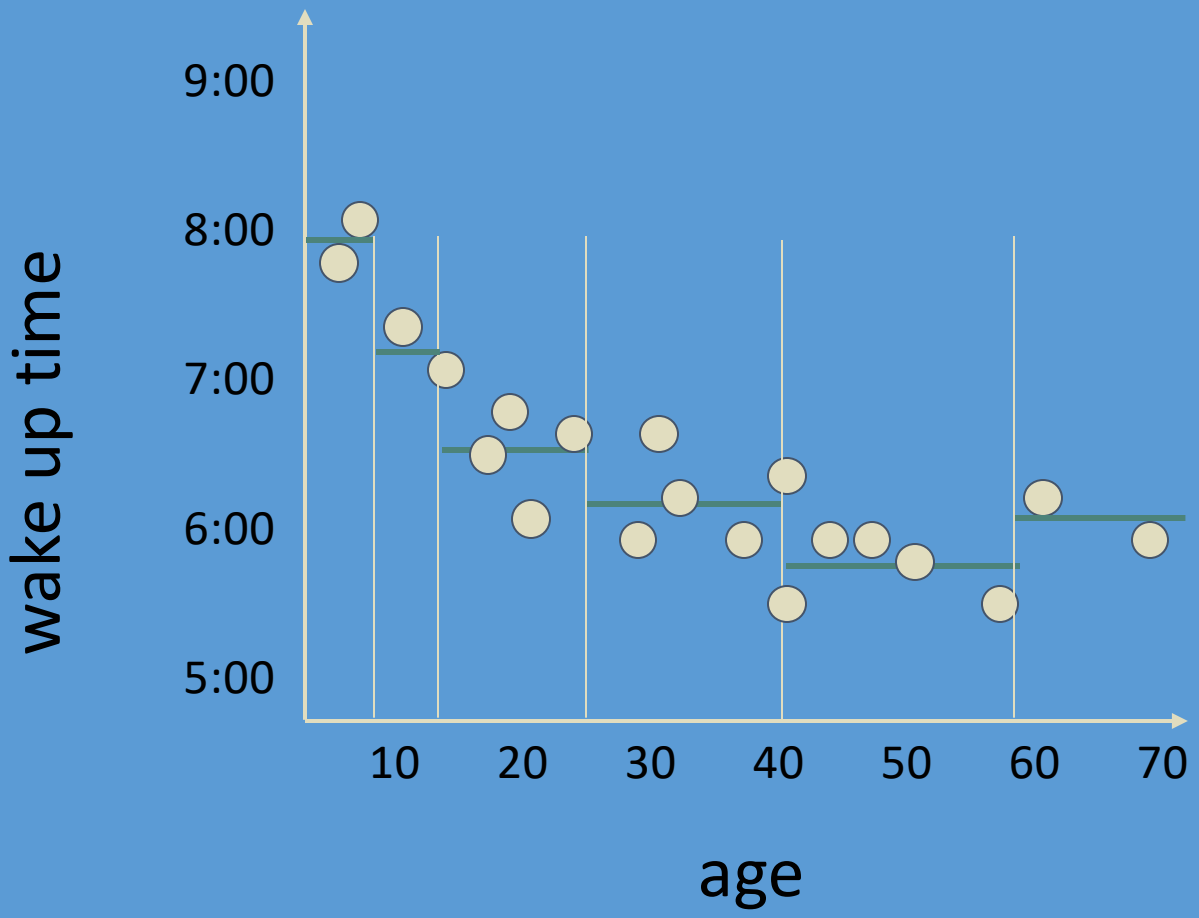

younger than 25?
yes
no
younger than 12?
6:00
yes
no
7:45
6:40

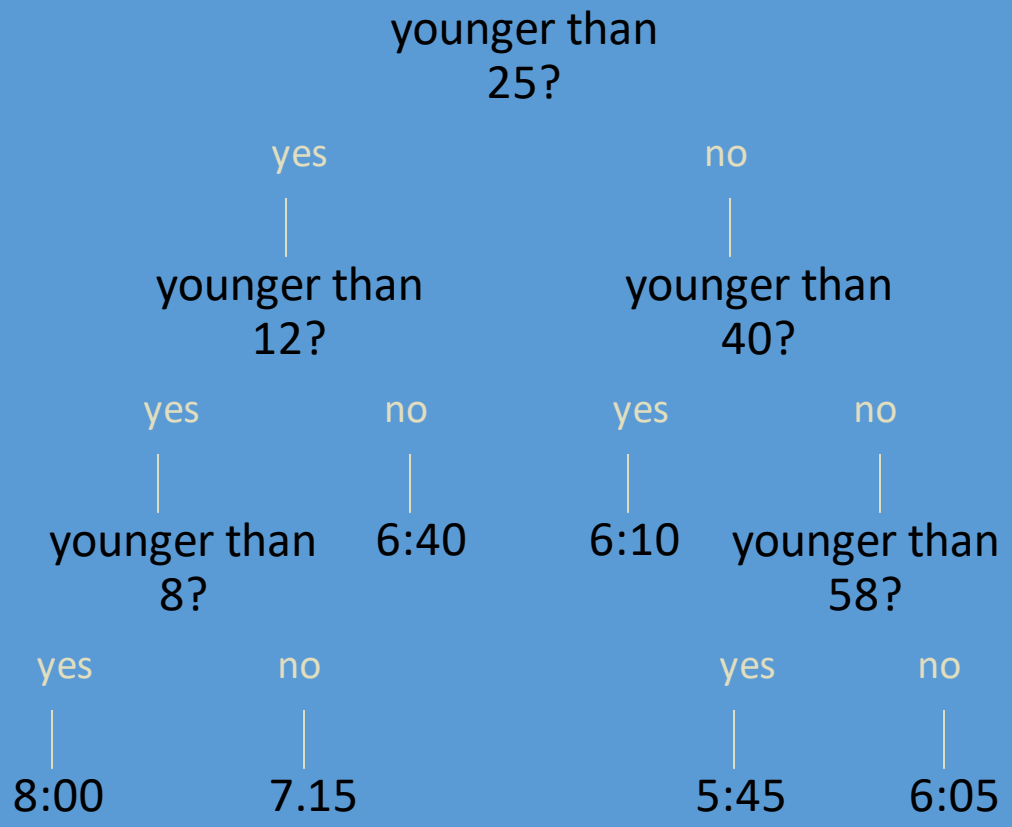

wake up time
9:00
8:00
7:00
6:00
5:00
10
20
30
40
50
60
70
age


younger than 25?
yes
no
younger than 12?
younger than 40?
yes
no
yes
no
7:45
6:40
6:10
5:50


wake up time
9:00
8:00
7:00
6:00
5:00
10
20
30
40
50
60
70
age

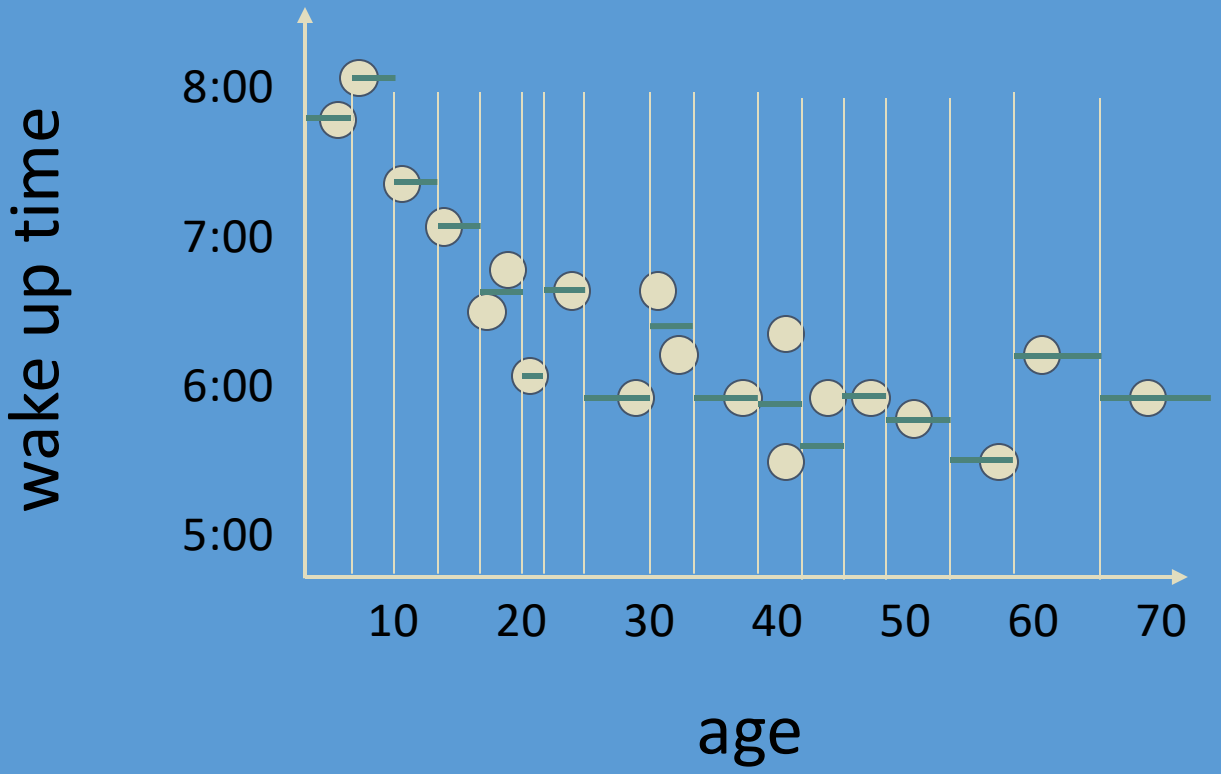

younger than 25?
yes
no
younger than 12?
younger than 40?
yes
no
yes
no
younger than 8?
6:40
6:10
5:50
yes
no
8:00
7.15


wake up time
9:00
8:00
7:00
6:00
5:00
10
20
30
40
50
60
70
age

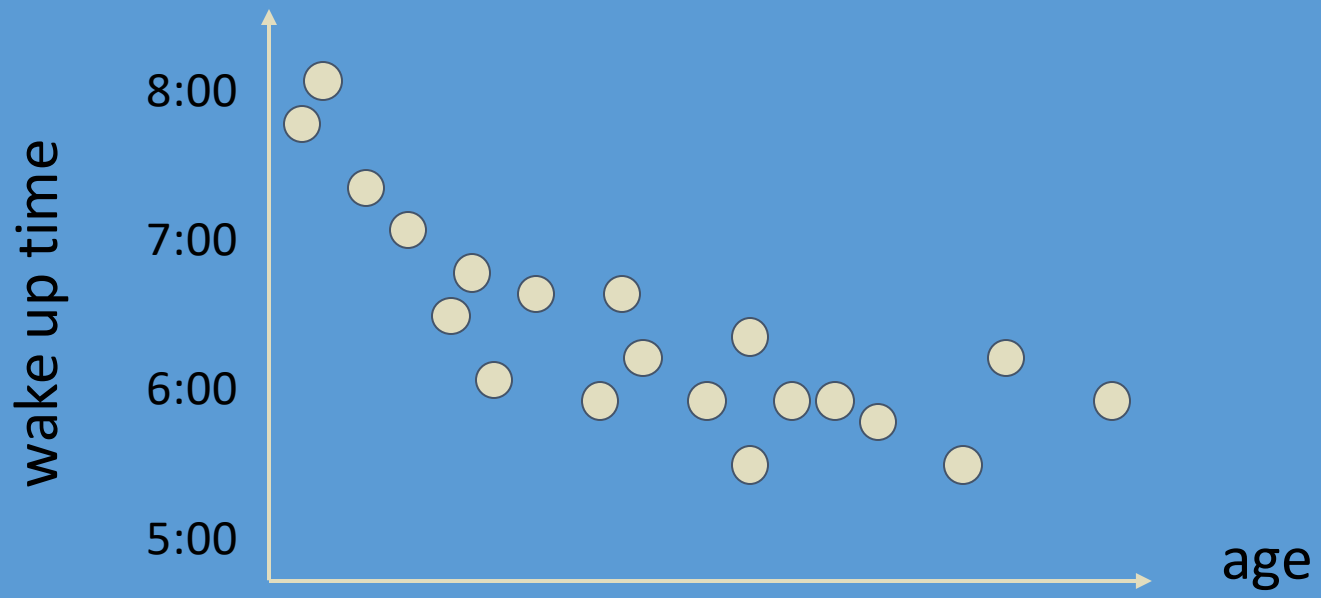

younger than 25?
yes
no
younger than 12?
younger than 40?
yes
no
yes
no
younger than 8?
6:40
6:10
younger than 58?
yes
no
yes
no
8:00
7.15
5:45
6:05

Cuidado con el exceso de particiones

Overfitting / insufciencia de datos

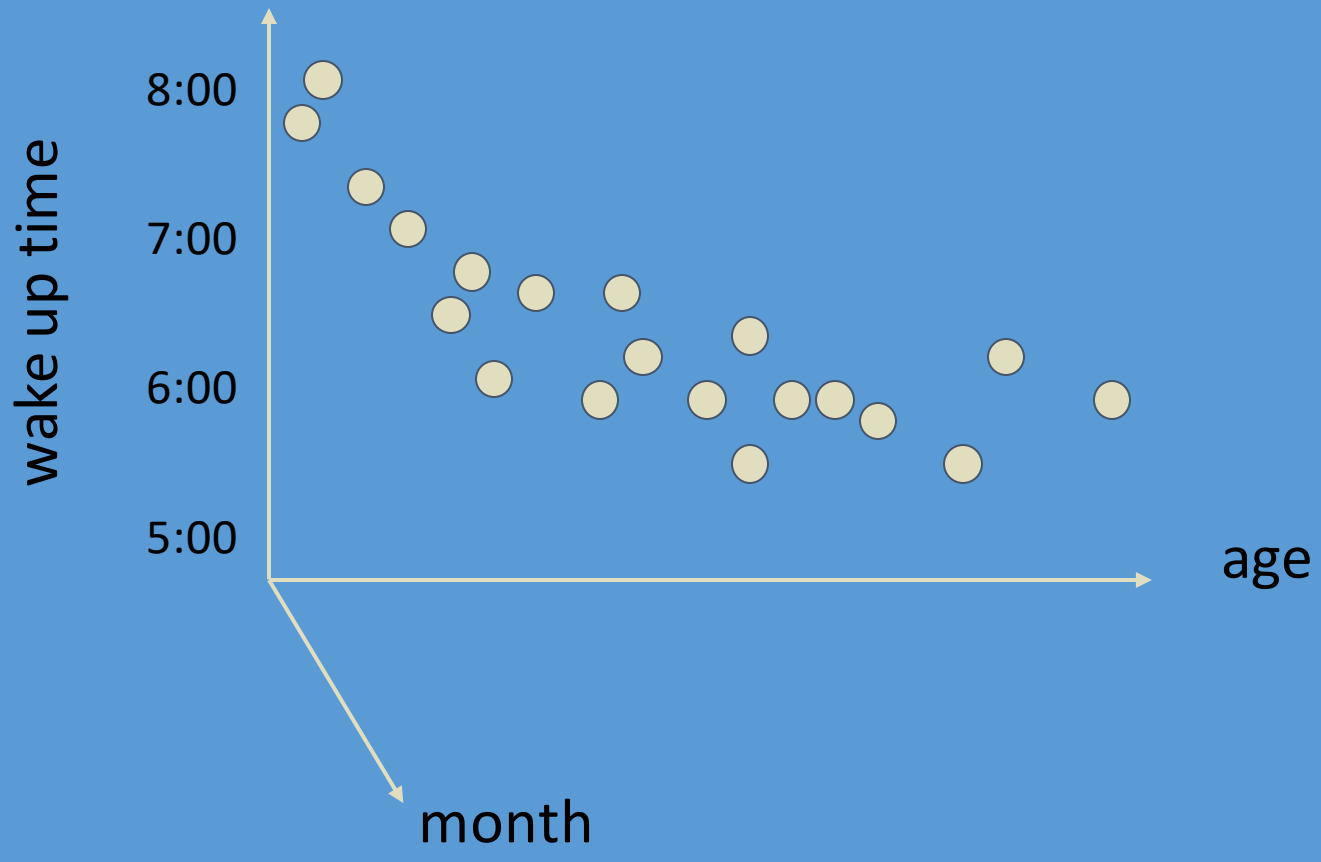


Observar más variables



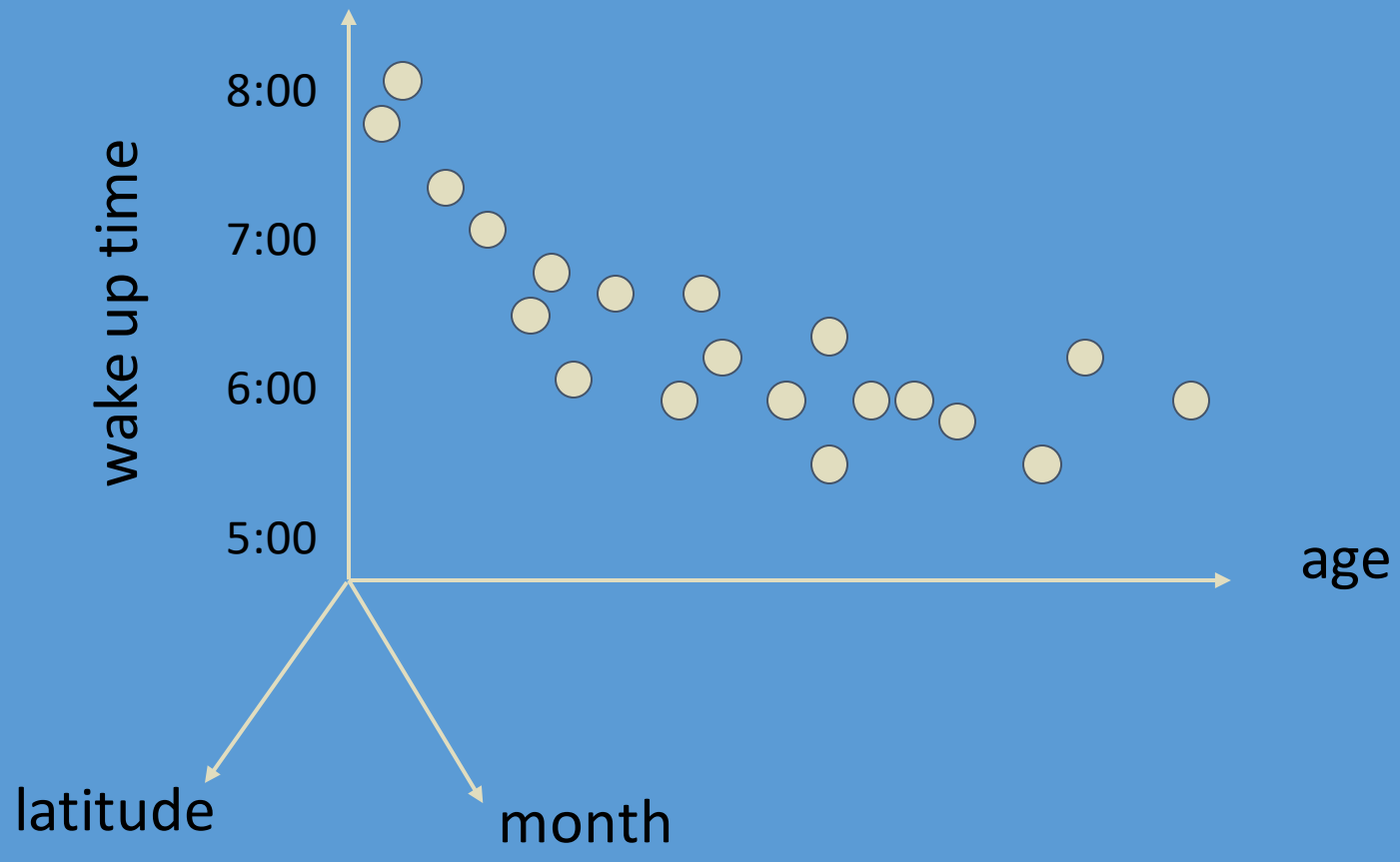
Watch out for

Muchas variables posibles



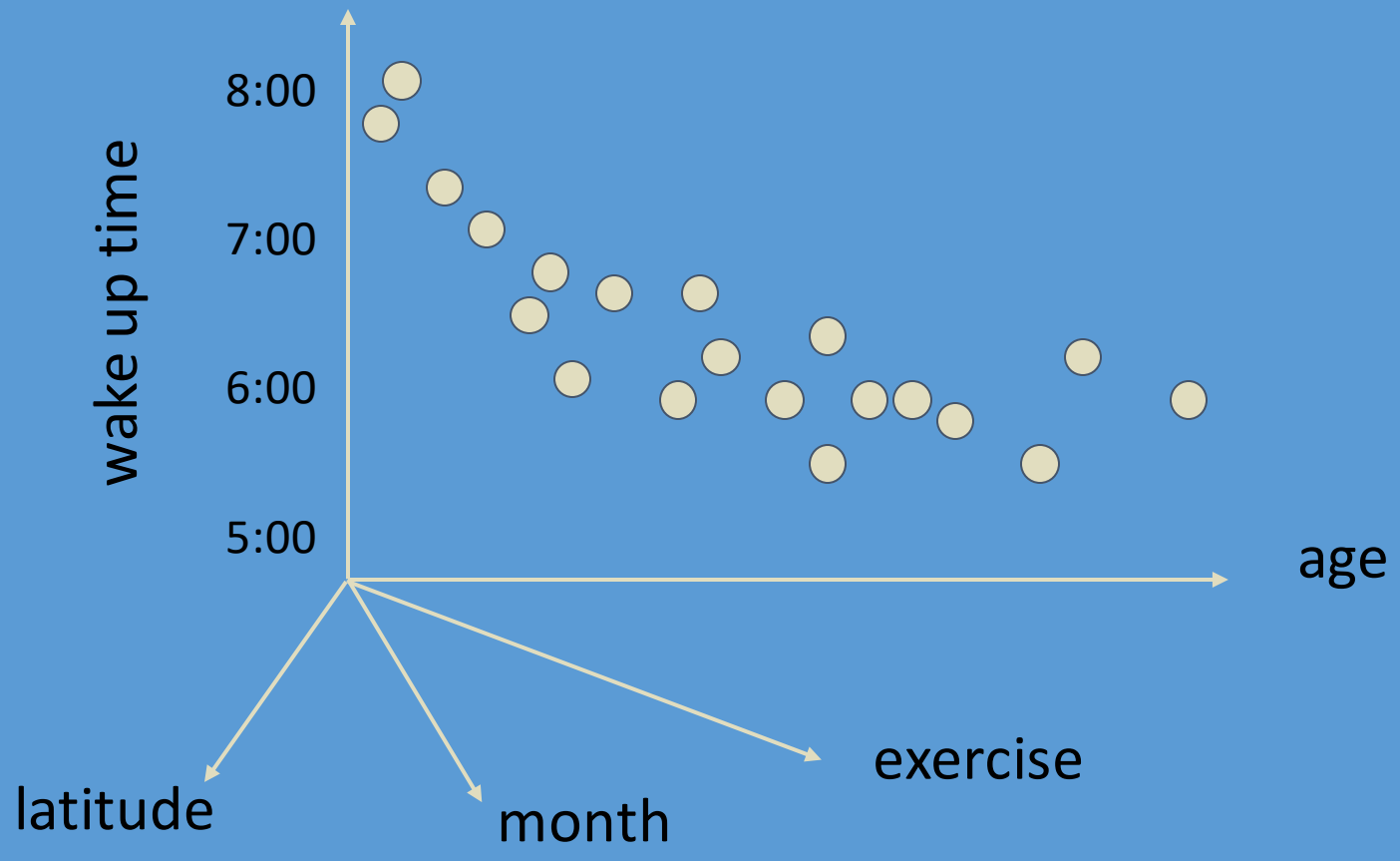
Observar más variables

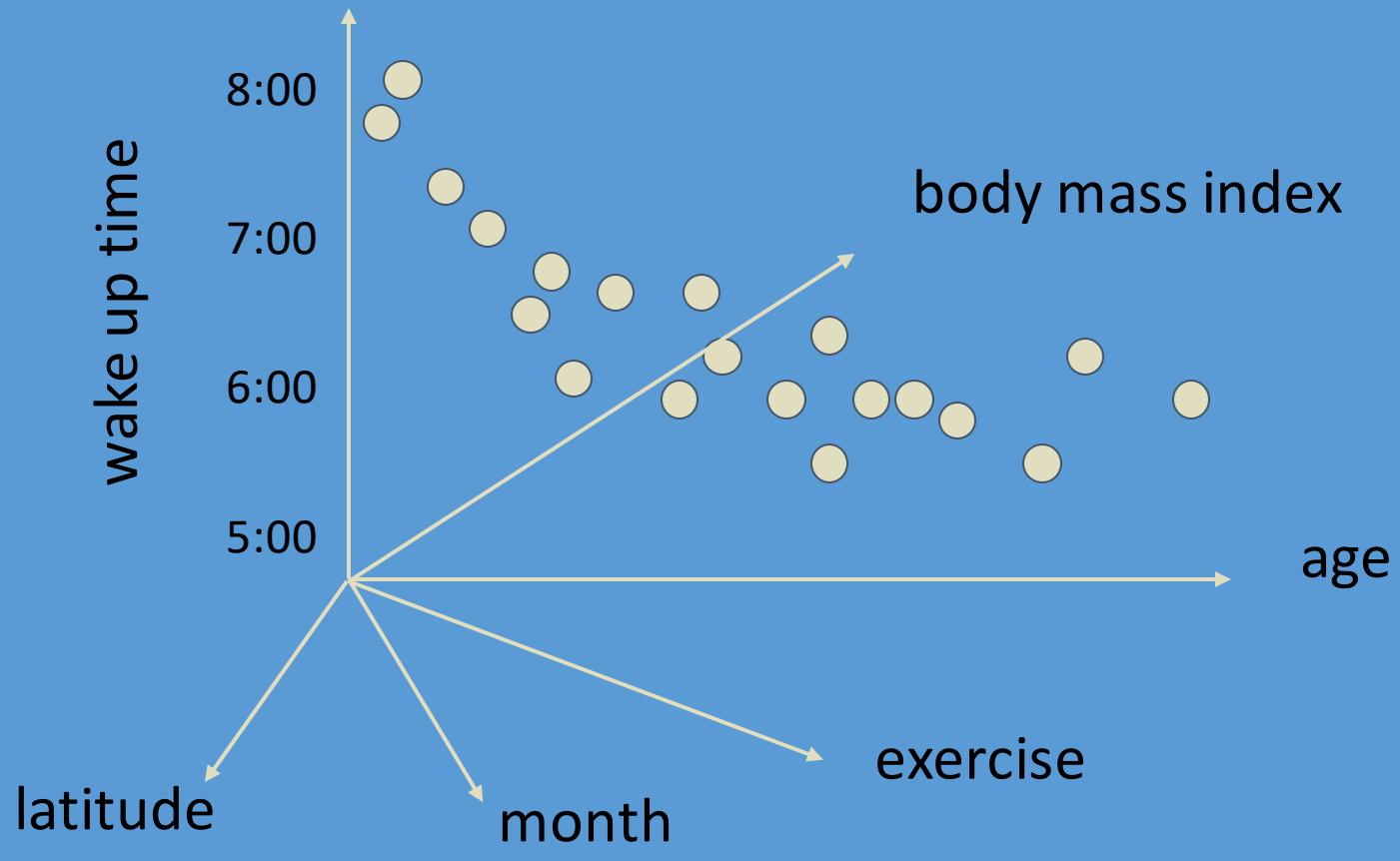
Muchas variables posibles



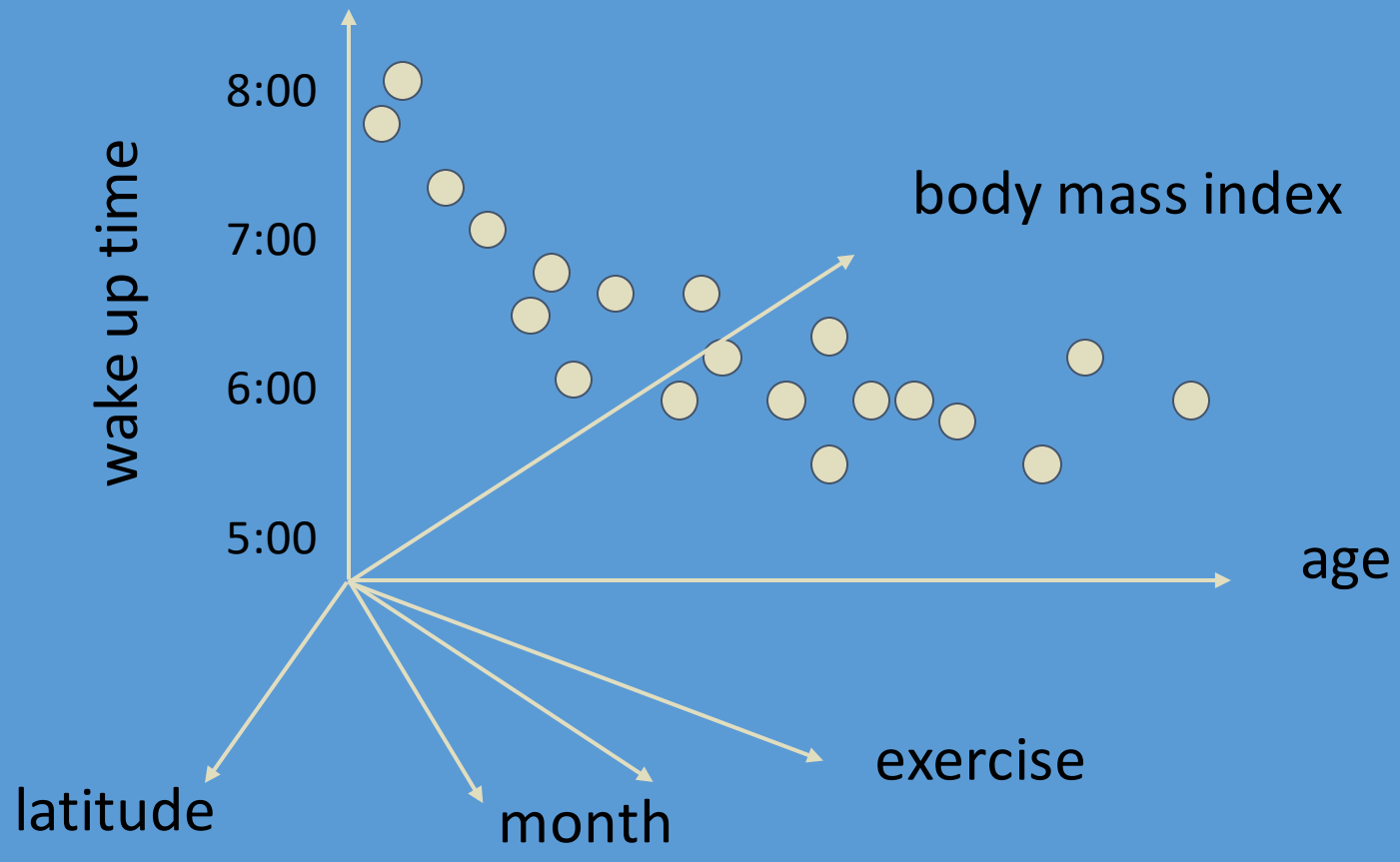
Observar más variables

Muchas variables posibles

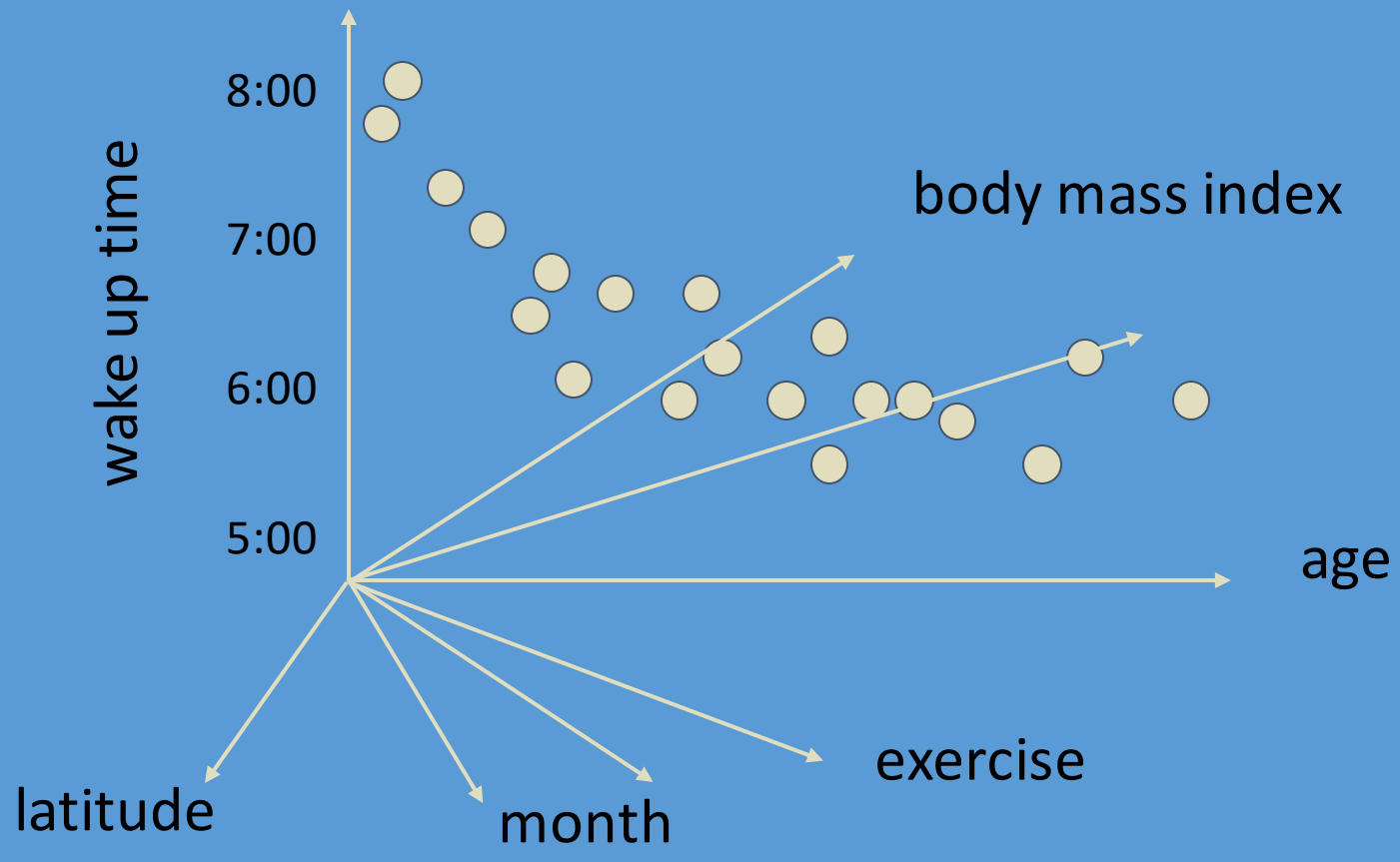



wake up time
8:00
7:00
6:00
5:00
body mass index
age
exercise
month
latitude

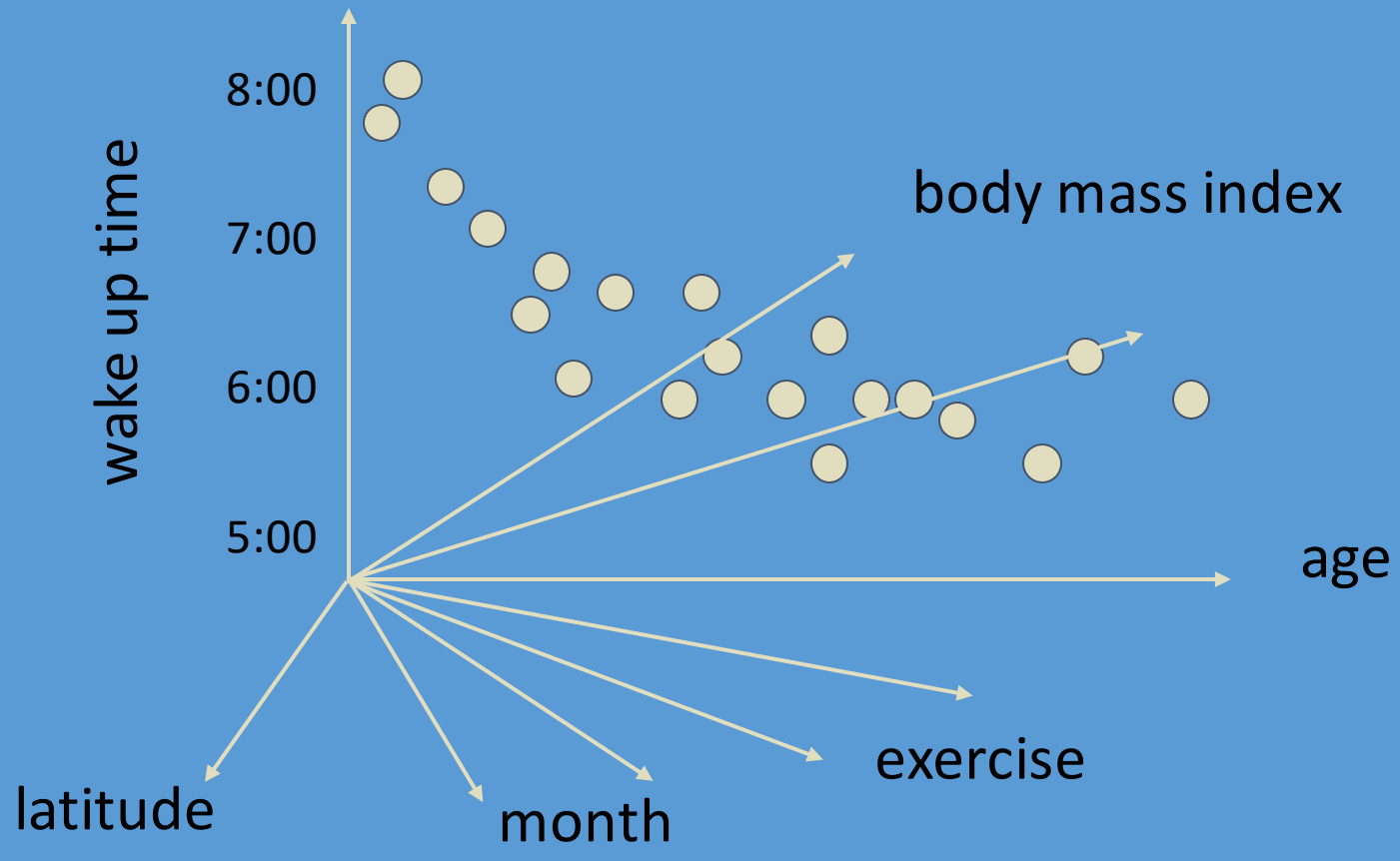
Observar más variables



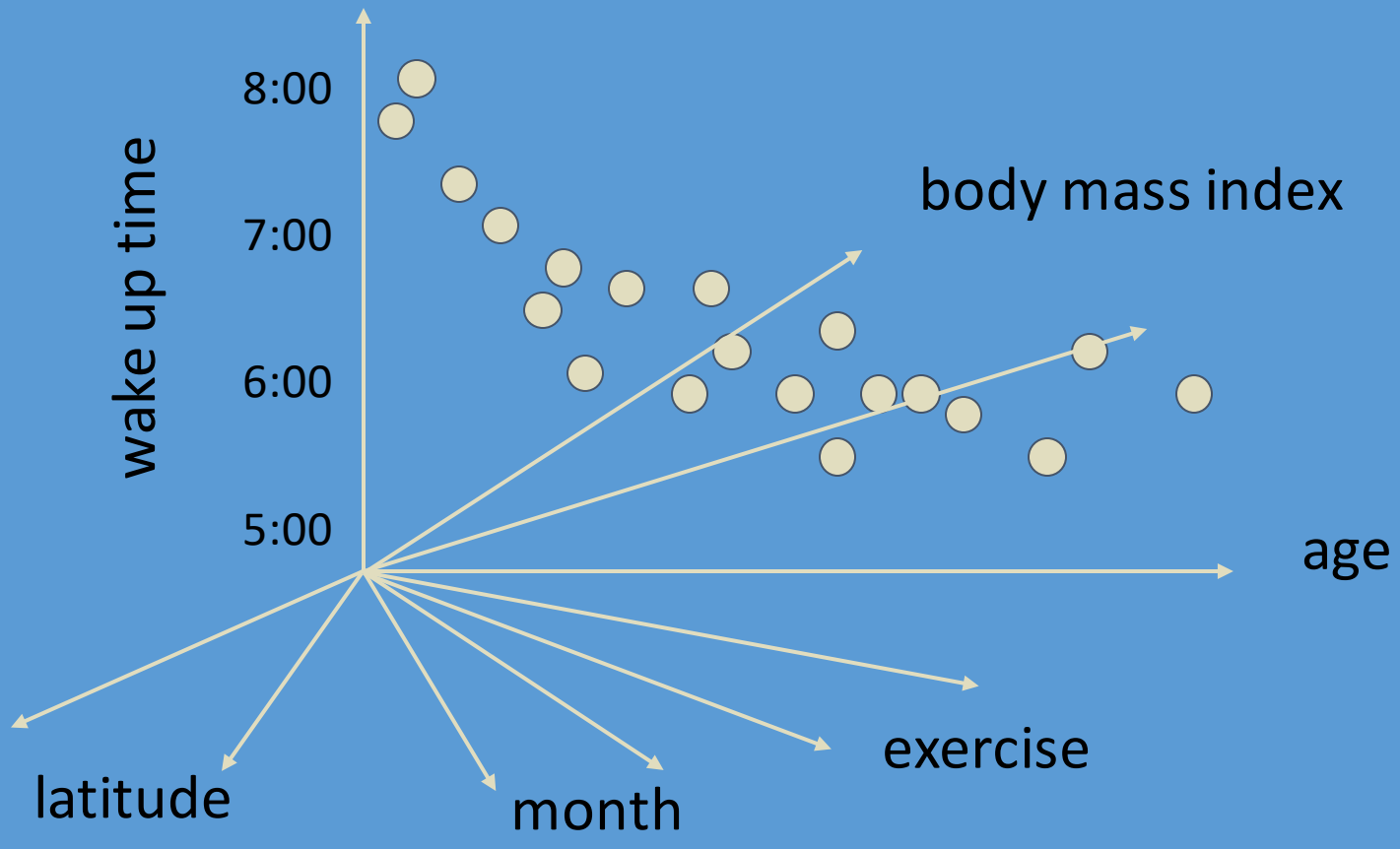
Observar más variables



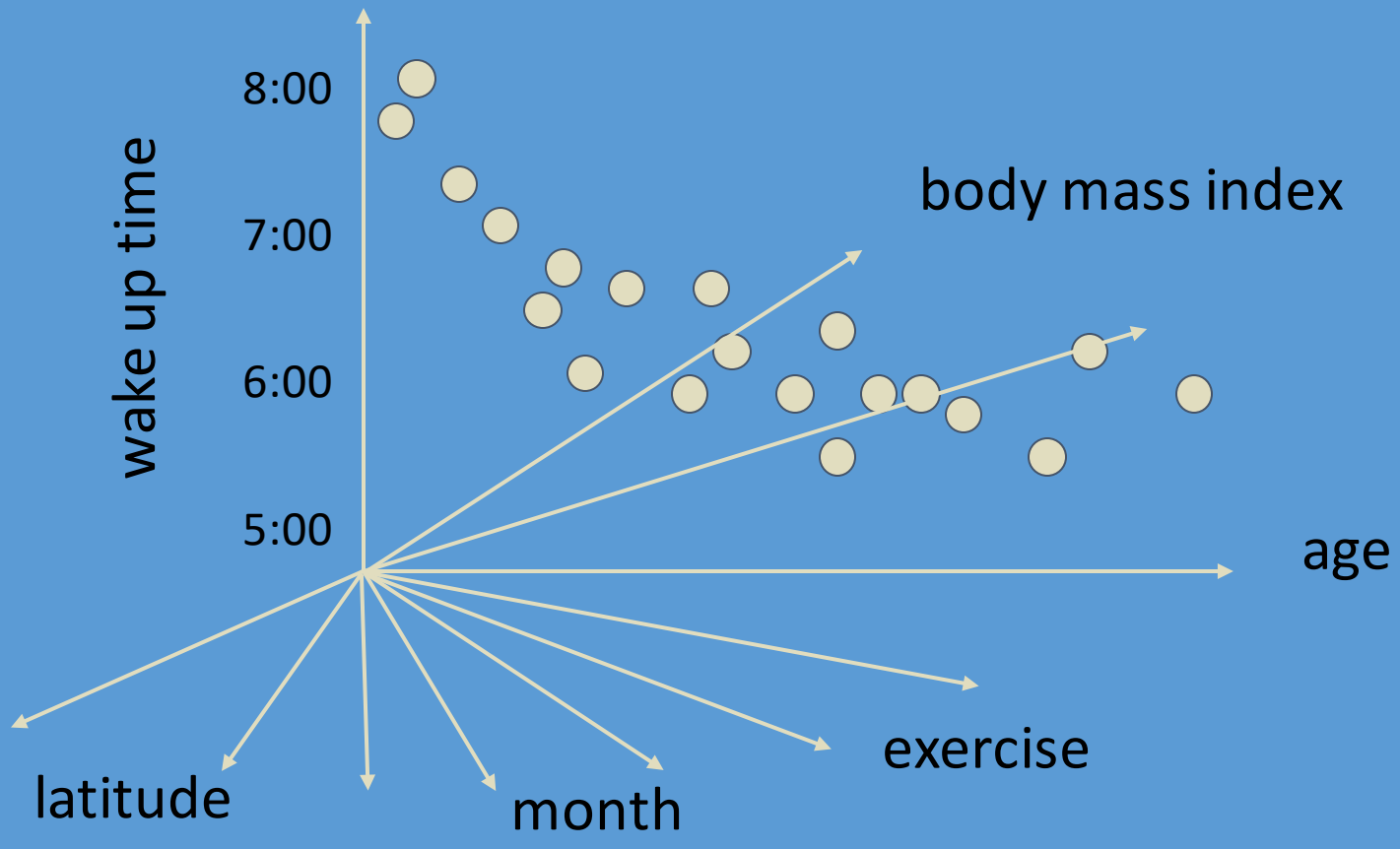
Observar más variables



Observar más variables



Observar más variables



Árboles de clasificación y regresión. Procedimiento

- Dos pasos:
  1. Aprendizaje, es decir, Creación del árbol
  2. Uso para decisión
- Se hace crecer un árbol partiendo del nodo raiz.
- En cada nodo se hace partición de datos entre nodos hijos.
- La partición se realiza según un cierto criterio
- Los nodos hoja son los terminales
- Para regresión, el valor predicho en el nodo es el promedio, media o mediana de lasobservaciones de la variable respuesta en el conjunto de aprendizaje que caen en el segmento
- Para clasificación, la clase predicha es la más común en el nodo, el decir el voto mayoritario
- Para árboles de clasificación también puede estimarse la probabilidad de pertenencia en cada una de las clases

- En este tema nos dedicaremos a Árboles de Clasificación.

¿Cuándo usar árboles de clasificación?

- Los ejemplos están descritos por pares de atributo-valor.
  - Existen extensiones para atributos continuos.
- La función/concepto a aprender es discreta
- Se obtienen hipótesis disyuntivas.

Ventajas

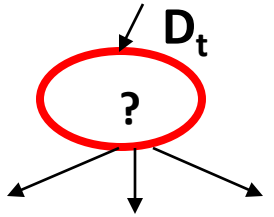
- Construcción sin excesivo coste computacional
- Clasificación rápida de nuevas muestras
- Interpretabilidad sencilla
- Exactitud comparable a la de otras técnicas para conjuntos de datos simples

Algoritmos

- Hay muchos algoritmos específicos para árboles de decisión. Entre ellos son resaltables:
  - **ID3** (Iterative Dichotomiser 3)
  - **C4.5** (successor of ID3)
  - **CART** (Classification And Regression Tree)
  - **CHAID** (Chi-square automatic interaction detection). Realiza divisions multinivel .
  - **MARS** algoritmo para manejar mejor datos numéricos.
  - **Árboles de Inferencia Condicional** (Conditional Inference Trees) Aproximación basada en la estadística que utiliza tests no paramétricos para las decisions de division (splits) en el árbol, corregidas por multiples tests para evitar overfitting.

Algoritmo ID3

- Inicializamos cargando el conjunto de muestras de aprendizaje $\mathbf{D_0}$ y creamos el nodo raíz $\mathbf{n_0}$
- Mientras no se den las condiciones de acabar:
  - Sea $\mathbf{D_t}$ el conjunto de muestras de entrenamiento que alcanzan el nodo t
  - Si $\mathbf{D_t}$ solo contiene muestras que pertenecen a una misma clase $\mathbf{y_t}$, entonces $\mathbf{n_t}$ es un nodo hoja etiquetado como $\mathbf{y_t}$
  - Si $\mathbf{D_t}$ contiene muestras que pertenecen a más de una clase, utilizar una comparación sobre un atributo para dividir el conjunto de datos en subconjuntos más pequeños. Aplicar recursivamente el procedimiento a cada subconjunto resultante.



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Aprendizaje del árbol

- Cuestiones
  - Determinar cómo dividir las muestras
    - Cómo especificar la comparación sobre el atributo?
    - Cómo determinar la mejor división de las muestras?
  - Determinar cuando detener la división de las muestras

Aprendizaje del árbol

- Cuestiones
  - Determinar cómo dividir las muestras
    - Cómo especificar la comparación sobre el atributo?
    - Cómo determinar la mejor división de las muestras?
  - Determinar cuando detener la división de las muestras

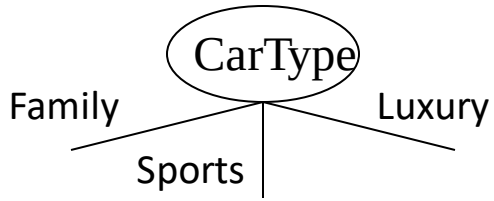
¿Cómo especificar la comparación?

- Depende del tipo de atributo
  - Nominal
  - Ordinal
  - Continuo

- Depende del número de formas de dividir
  - División binaria
  - División múltiple

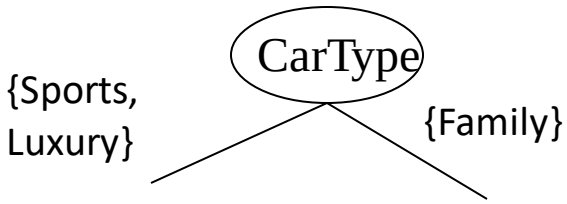
División de atributos nominales

- División múltiple: Generar tantas particiones como valores diferentes tiene el atributo.

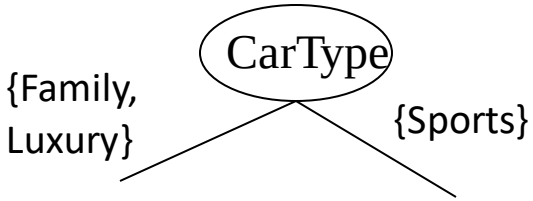


- División binaria: Dividir los valores en dos subconjuntos

Encontrar la partición óptima.

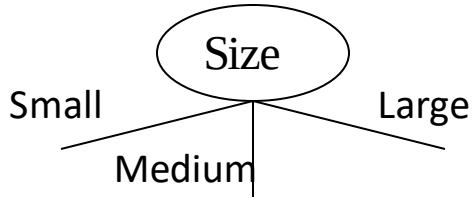


o



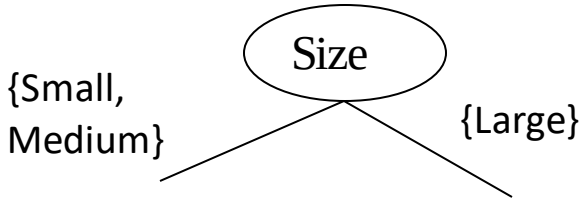
División de atributos ordinales

- División múltiple: Generar tantas particiones como valores diferentes tiene el atributo.

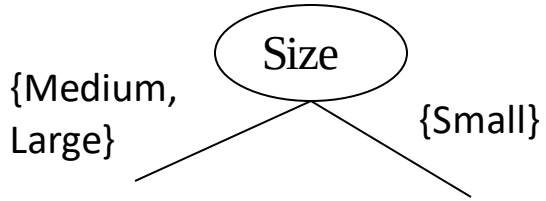


- División binaria: Dividir los valores en dos subconjuntos.

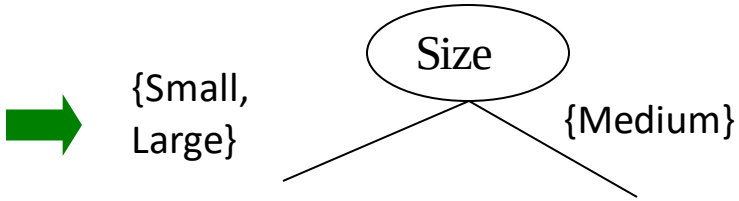
Encontrar la partición óptima.



o



Qué ocurriría con esta partición?

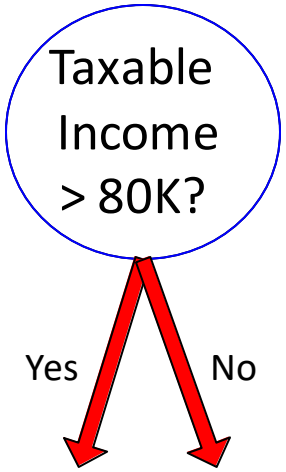


División de atributos continuos

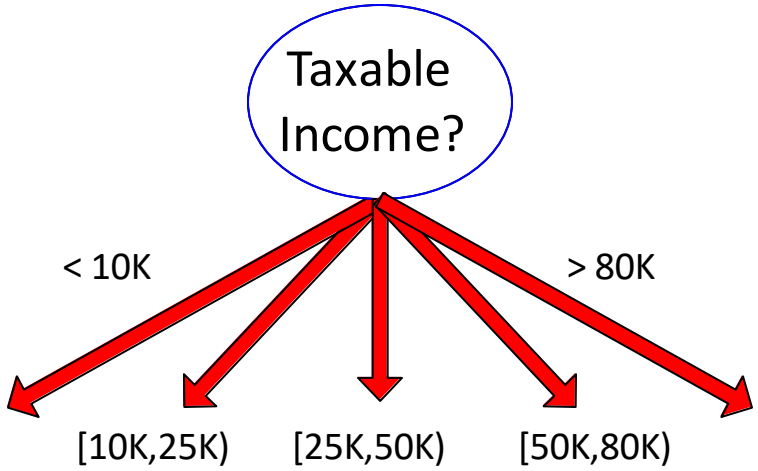
- Discretización para convertirlo en un atributo ordinal
  - Estática – se discretizan una sola vez
  - Dinámica – los intervalos se pueden obtener por igualdad de frecuencia (percentiles), o clustering.

- División binaria: $(A < v)$ o $(A \geq v)$
  - Se consideran todos las posibles divisiones y se busca el mejor punto de corte
  - Puede ser más costoso computacionalmente

División de atributos continuos



(i) División binaria



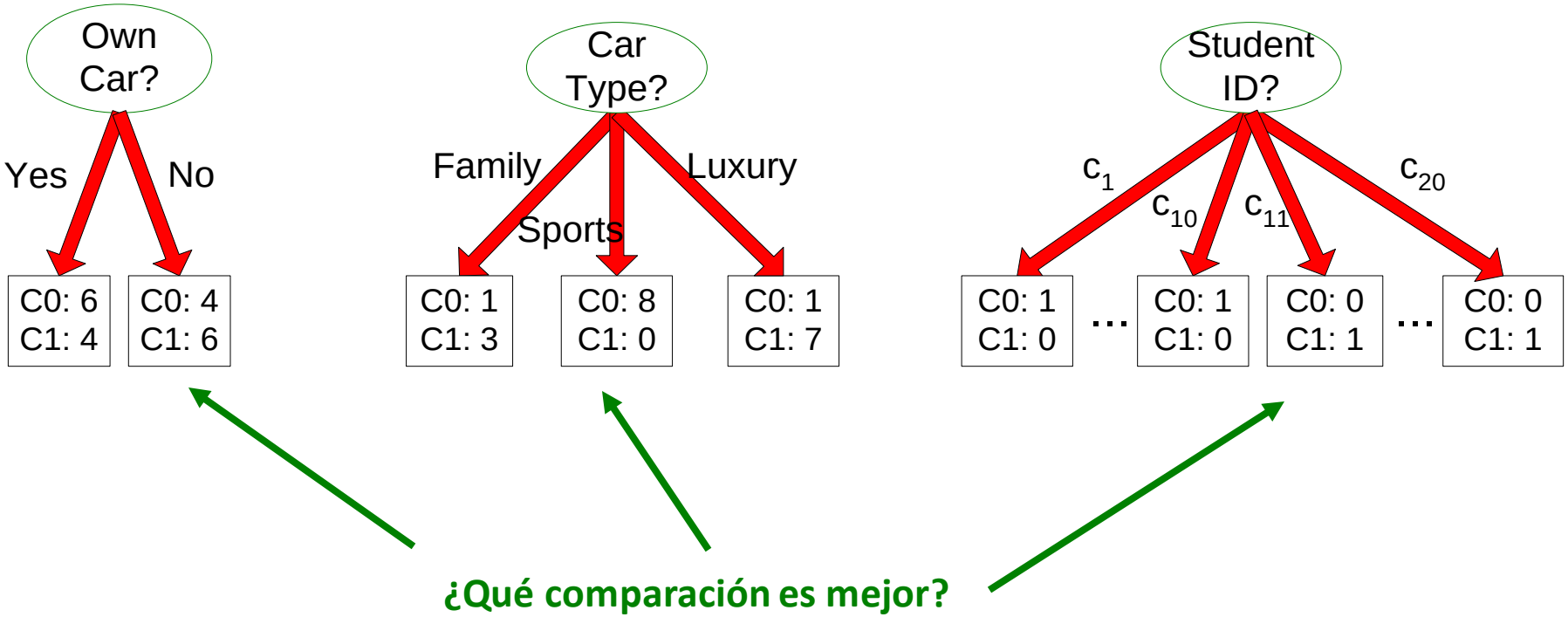
(ii) División múltiple

Aprendizaje del árbol

- Cuestiones
  - Determinar cómo dividir las muestras
    - ¿Cómo especificar la comparación sobre el atributo?
    - ¿Cómo determinar la mejor división de las muestras?
  - Determinar cuando detener la división de las muestras

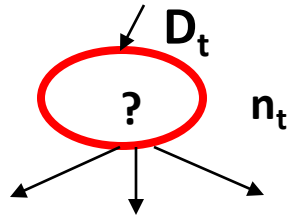
Cómo determinar la mejor partición

Antes de la partición: 10 muestras clase 0,
10 muestras clase 1



Estructura general del algoritmo ID3 (recordatorio)

- Inicializamos cargando el conjunto de muestras de aprendizaje $\mathbf{D_0}$ y creamos el nodo raíz $\mathbf{n_0}$
- Mientras no se den las condiciones de acabar:
  - Sea $\mathbf{D_t}$ el conjunto de muestras de entrenamiento que alcanzan el nodo t
  - Si $\mathbf{D_t}$ solo contiene muestras que pertenecen a una misma clase $\mathbf{y_t}$, entonces $\mathbf{n_t}$ es un nodo hoja etiquetado como $\mathbf{y_t}$
  - Si $\mathbf{D_t}$ contiene muestras que pertenecen a más de una clase, utilizar una comparación sobre un atributo para dividir el conjunto de datos en subconjuntos más pequeños. Aplicar recursivamente el procedimiento a cada subconjunto resultante.



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Cómo determinar el mejor atributo para realizar la partición

- Estrategia greedy:
  - Se prefieren nodos con una distribución de clases homogéneas

C0: 5 C1: 5	C0: 9 C1: 1
No homogéneo, **Alto grado de impureza**	**Homogéneo,** **Bajo grado de impureza**

Necesidad de una medida de la "pureza" de un nodo

Cómo determinar la mejor partición

- Propiedades que debe cumplir la medida de pureza del nodo:
  - Cuando un nodo es puro (solo muestras pertenecentes a una clase), la medida debe dar cero.
  - Cuando la impureza es máxima (todas las clases son equiprobables), la medida debe ser máxima.

Medida para decidir cuál es el mejor atributo para realizar la partición

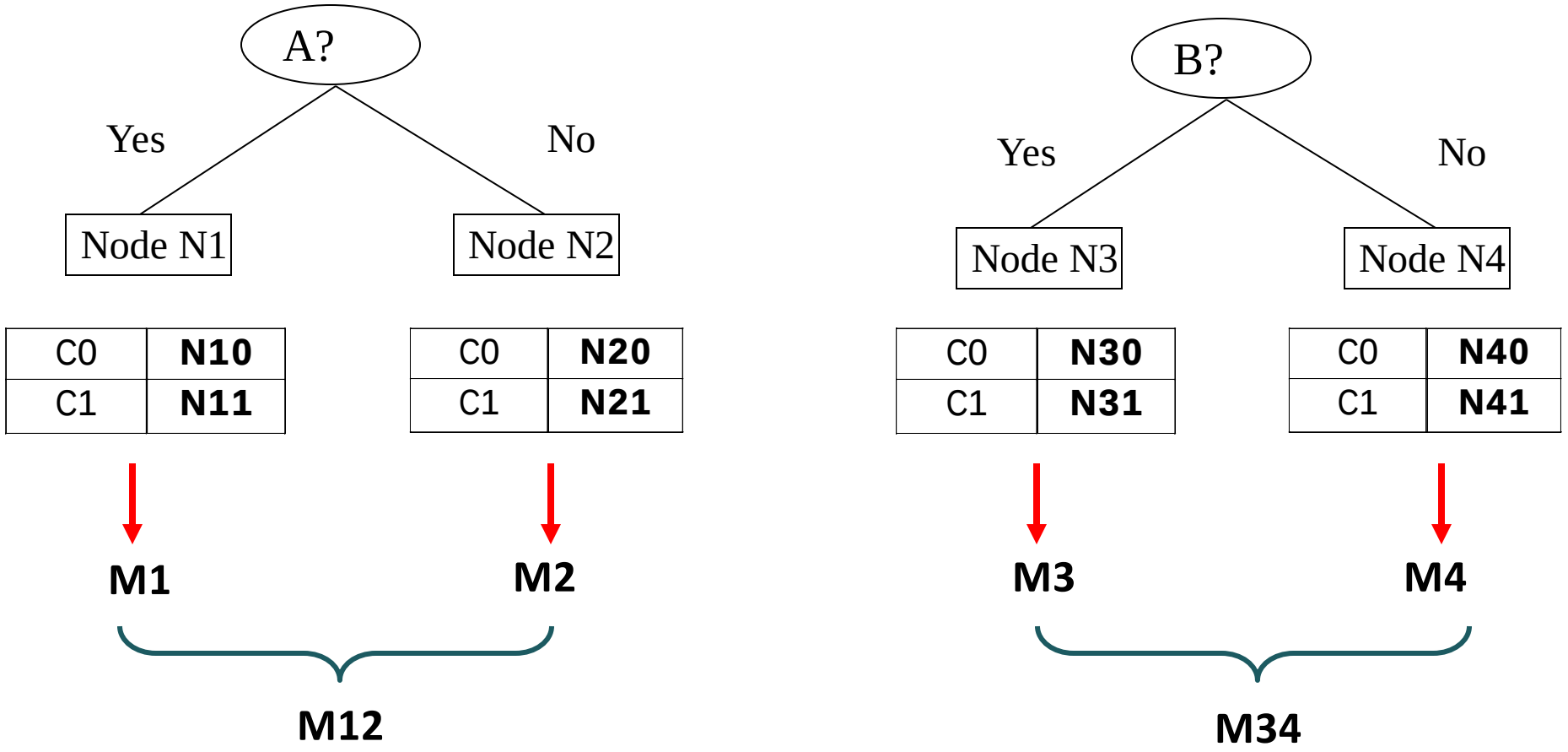
- Aquel que genere un particionado más homogéneo, es decir más puro
- La medida base es la ENTROPÍA
- Determinamos el atributo por el que hacer la mejor partición a aquel que produce más GANACIA DE INFORMACIÓN

Cómo determinar la mejor partición

Antes de la división:

C0	**N00**
C1	**N01**

→ **M0**



Gain = M0 – M12 **vs** **Gain = M0 – M34**

Medida de impureza de nodos

- Basada en Teoría de la Información: Entropía

División basada en Teoría de la Información: La Entropía

La entropía en un nodo *t*:

$$Entropy(t) = -\sum_j p(j \mid t) \log p(j \mid t)$$

- $p(j \mid t)$ es la frecuencia relativa de la clase j en el nodo t.
- Medida de la homogeneidad de un nodo.
  - Máximo = (log n_c) cuando todas las muestras están igualmente distribuidas entre todas las clases $\Rightarrow$ información mínima
  - Mínimo = (0.0) cuando todas las muestras pertenecen a una sola clase $\Rightarrow$ más información

Ejemplo de cálculo de la entropía

$$Entropy(t) = -\sum_j p(j \mid t) \log p(j \mid t)$$

C1	**0**
C2	**6**

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Entropy = – 0 log 0 – 1 log 1 = – 0 – 0 = 0

C1	**1**
C2	**5**

P(C1) = 1/6 P(C2) = 5/6

Entropy = – (1/6) $\log_2$ (1/6)– (5/6) $\log_2$ (1/6) = 0.65

C1	**2**
C2	**4**

P(C1) = 2/6 P(C2) = 4/6

Entropy = – (2/6) $\log_2$ (2/6)– (4/6) $\log_2$ (4/6) = 0.92

Atributos con una alta cardinalidad

- Los atributos con un gran número de valores tienden a tener un valor más alto de la ganancia (p.e. identificador de muestra)
- Los subconjuntos tienen más probabilidad de ser “puros” si tienen un número muy alto de valores.
  - $\Rightarrow$ La ganancia de información tiene un sesgo en favor de los atributos con mayor número de valores.
  - $\Rightarrow$ Puede dar lugar a sobreajuste (seleccionar atributos que no generalizan bien)

Ejemplo – Jugar al Tenis

Se pretende construir un árbol de clasificación a partir de los ejemplos de la tabla adjunta para determinar si el día será adecuado para jugar al tenis en función de los valores de cuatro atributos: **Outlook**, **Temperature**, **Humidity** y **Wind**

outlook	temperature	humidity	wind	play
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rain	mild	high	weak	yes
rain	cool	normal	weak	yes
rain	cool	normal	strong	no
overcast	mild	high	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rain	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	hot	normal	weak	yes
overcast	cool	normal	strong	yes
rain	mild	high	strong	no

Cálculo de la entropía para la tabla del ejemplo

$$Entropy(t) = -\sum_j p(j \mid t) \log p(j \mid t)$$

C1 (yes)	**9**
C2 (no)	**5**

$$Entropy([9+,\ 5-]) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = .940$$

$$Donde:\ log_b(x) = \frac{log_c(x)}{log_c(b)}$$

outlook	temperature	humidity	wind	play
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rain	mild	high	weak	yes
rain	cool	normal	weak	yes
rain	cool	normal	strong	no
overcast	mild	high	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rain	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	hot	normal	weak	yes
overcast	cool	normal	strong	yes
rain	mild	high	strong	no

División basada en Teoría de la Información

Ganancia de información:

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

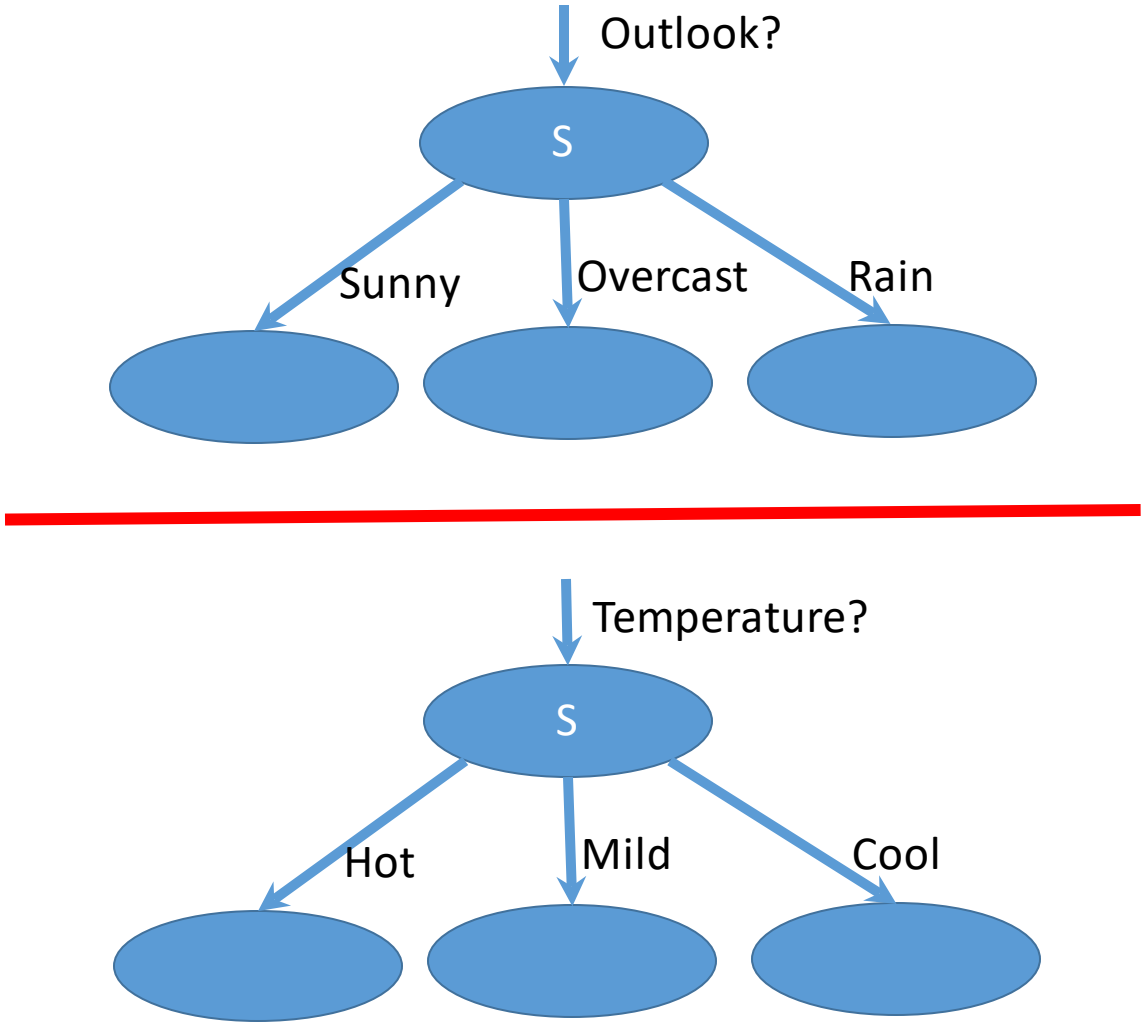
donde, n_i = número de muestras en el hijo i,

n = número de muestras en el nodo p.

- Mide la reducción de la entropía debida a la división. Elegir el atributo con mayor valor de ganancia
- Se utiliza en ID3 y C4.5
- Inconveniente: Tiende a preferir divisiones que dan lugar a un número grande de particiones (nodos hijos), siendo cada una pequeña pero “pura”.

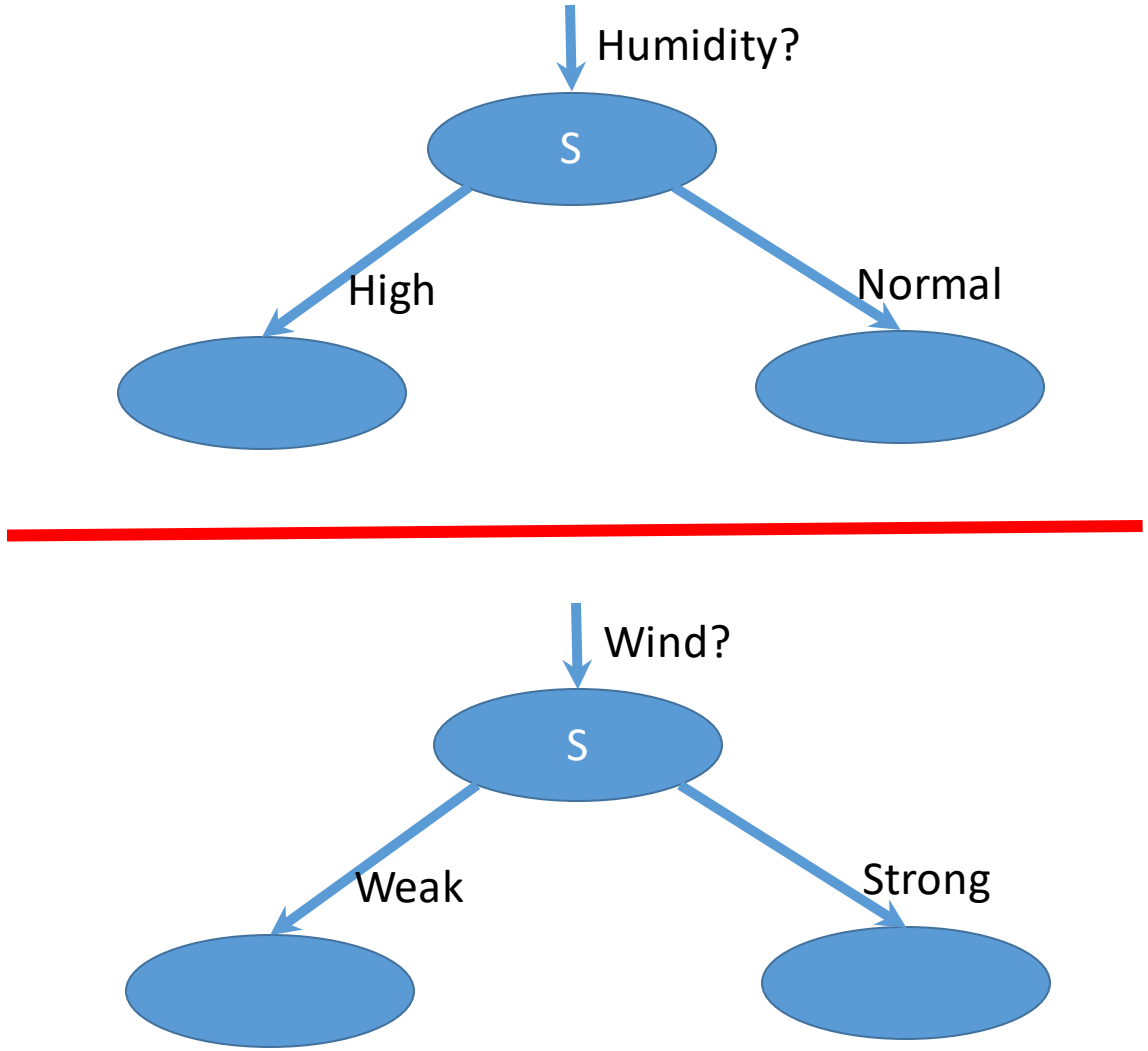
Divisiones posibles del nodo inicial: Tabla S (I)

outlook	temperature	humidity	wind	play
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rain	mild	high	weak	yes
rain	cool	normal	weak	yes
rain	cool	normal	strong	no
overcast	mild	high	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rain	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	hot	normal	weak	yes
overcast	cool	normal	strong	yes
rain	mild	high	strong	no



Divisiones posibles de la tabla S (II)

outlook	temperature	humidity	wind	play
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rain	mild	high	weak	yes
rain	cool	normal	weak	yes
rain	cool	normal	strong	no
overcast	mild	high	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rain	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	hot	normal	weak	yes
overcast	cool	normal	strong	yes
rain	mild	high	strong	no

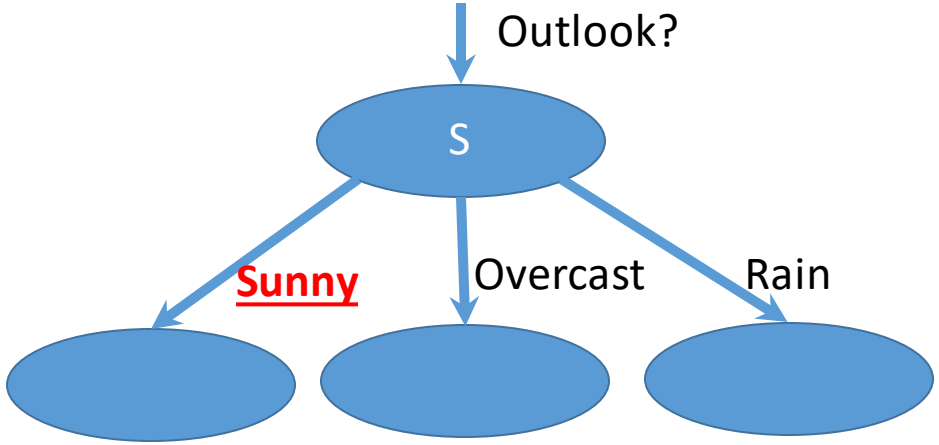


¿Cómo dividir el conjunto de datos inicial S?

- Escoger cada atributo (fila) de la columna de la tabla y generar nuevas tablas para cada uno de los valores del atributo en cuestión, eliminando la columna de ese atributo en cada subtabla

Tabla para Outlook= Sunny

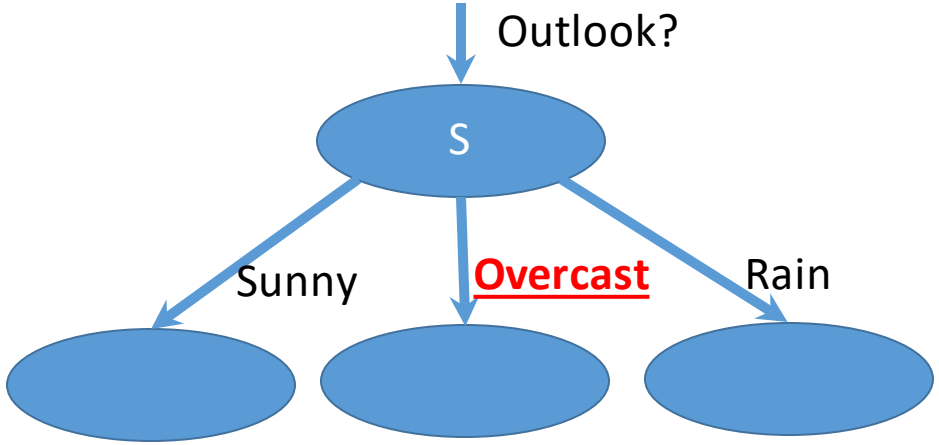
outlook	temperature	humidity	wind	play
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rain	mild	high	weak	yes
rain	cool	normal	weak	yes
rain	cool	normal	strong	no
overcast	mild	high	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rain	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	hot	normal	weak	yes
overcast	cool	normal	strong	yes
rain	mild	high	strong	no



temperature	humidity	wind	play
hot	high	weak	no
hot	high	strong	no
mild	high	weak	no
cool	normal	weak	yes
mild	normal	strong	yes

Tabla para Outlook= Overcast

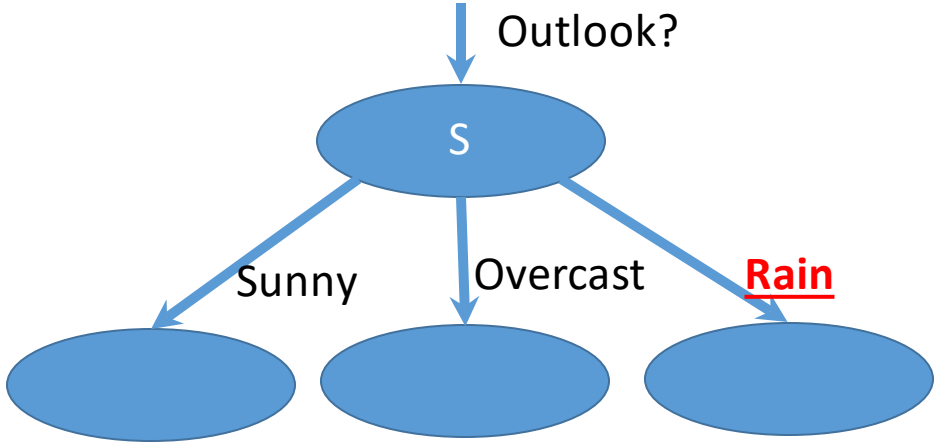
outlook	temperature	humidity	wind	play
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rain	mild	high	weak	yes
rain	cool	normal	weak	yes
rain	cool	normal	strong	no
overcast	mild	high	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rain	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	hot	normal	weak	yes
overcast	cool	normal	strong	yes
rain	mild	high	strong	no



temperature	humidity	wind	play
hot	high	weak	yes
mild	high	strong	yes
hot	normal	weak	yes
cool	normal	strong	yes

Tabla para Outlook= Rain

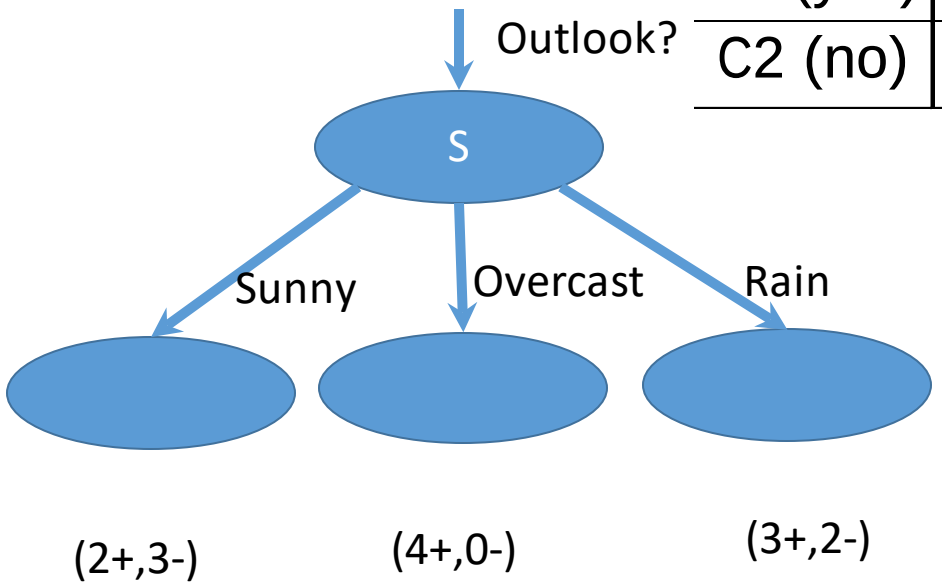
outlook	temperature	humidity	wind	play
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rain	mild	high	weak	yes
rain	cool	normal	weak	yes
rain	cool	normal	strong	no
overcast	mild	high	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rain	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	hot	normal	weak	yes
overcast	cool	normal	strong	yes
rain	mild	high	strong	no



temperature	humidity	wind	play
mild	high	weak	yes
cool	normal	weak	yes
cool	normal	strong	no
mild	normal	weak	yes
mild	high	strong	no

Cálculo de la Ganancia de Información para Outlook

C1 (yes)	**9**
C2 (no)	**5**



$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

GAIN = Entropía antes de dividir–Entropía después

Entropía antes:

$$Entropy([9+,\ 5-]) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = .940$$

$$GAIN(S,Outlook) = Entropy\,([9+,5-]) - \Sigma\,(n_i/n)(-\,p_{yes}\,\log_2 p_{yes} - p_{no}\,\log_2 p_{no})$$

$$GAIN(S,\ Outlook) = 0.940$$

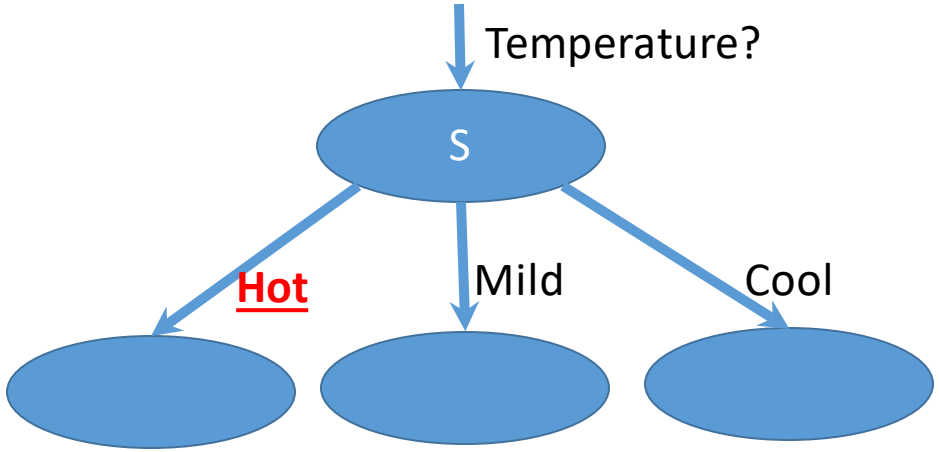
$$- \{\ (\ 5/14)(-2/5\ \log_2 2/5 - 3/5\ \log_2 3/5)$$

$$-\ (\ 4/14)(-4/4\ \log_2 4/4 - 0/4\ \log_2 0/4)$$

$$-\ (\ 5/14)(-3/5\ \log_2 3/5 - 2/5\ \log_2 2/5)\} = 0.246$$

Tabla para Temperature = Hot

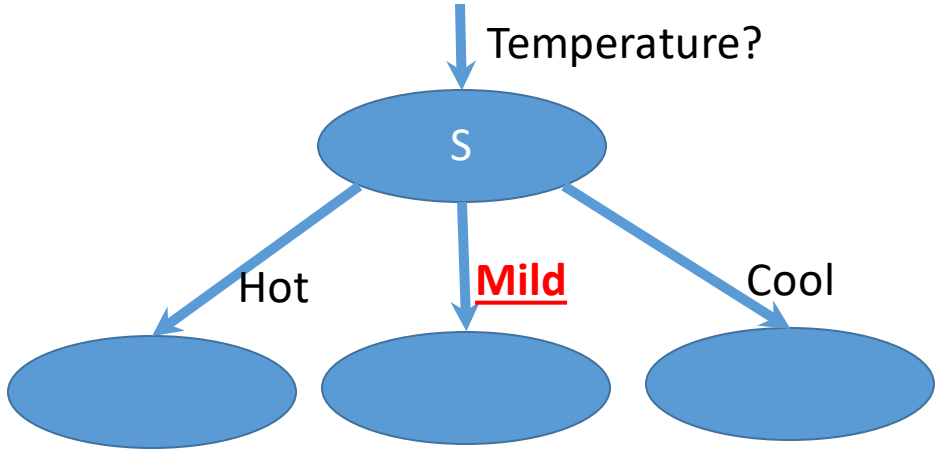
outlook	temperature	humidity	wind	play
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rain	mild	high	weak	yes
rain	cool	normal	weak	yes
rain	cool	normal	strong	no
overcast	mild	high	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rain	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	hot	normal	weak	yes
overcast	cool	normal	strong	yes
rain	mild	high	strong	no



outlook	humidity	wind	play
sunny	high	weak	no
sunny	high	strong	no
overcast	high	weak	yes
overcast	normal	weak	yes

Tabla para Temperature = Mild

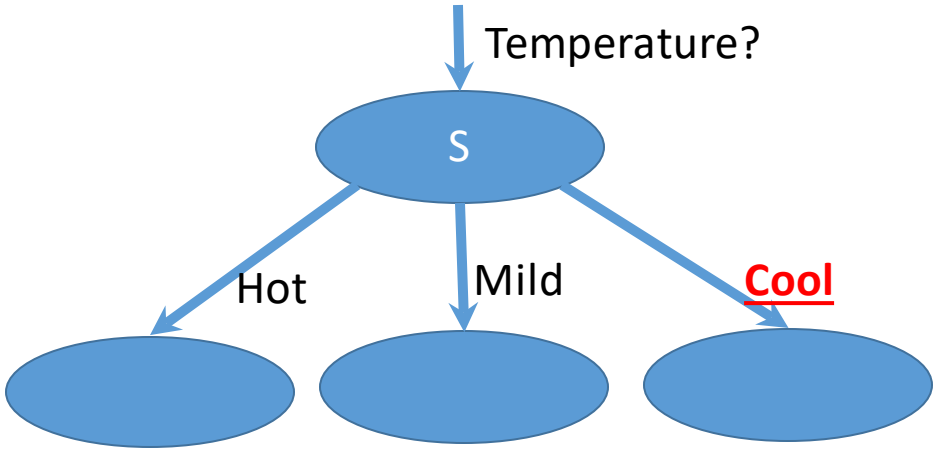
outlook	temperature	humidity	wind	play
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rain	mild	high	weak	yes
rain	cool	normal	weak	yes
rain	cool	normal	strong	no
overcast	mild	high	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rain	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	hot	normal	weak	yes
overcast	cool	normal	strong	yes
rain	mild	high	strong	no



outlook	humidity	wind	play
rain	high	weak	yes
overcast	high	strong	yes
sunny	high	weak	no
rain	normal	weak	yes
sunny	normal	strong	yes
rain	high	strong	no

Tabla para Temperature = Cool

outlook	temperature	humidity	wind	play
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rain	mild	high	weak	yes
rain	cool	normal	weak	yes
rain	cool	normal	strong	no
overcast	mild	high	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rain	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	hot	normal	weak	yes
overcast	cool	normal	strong	yes
rain	mild	high	strong	no

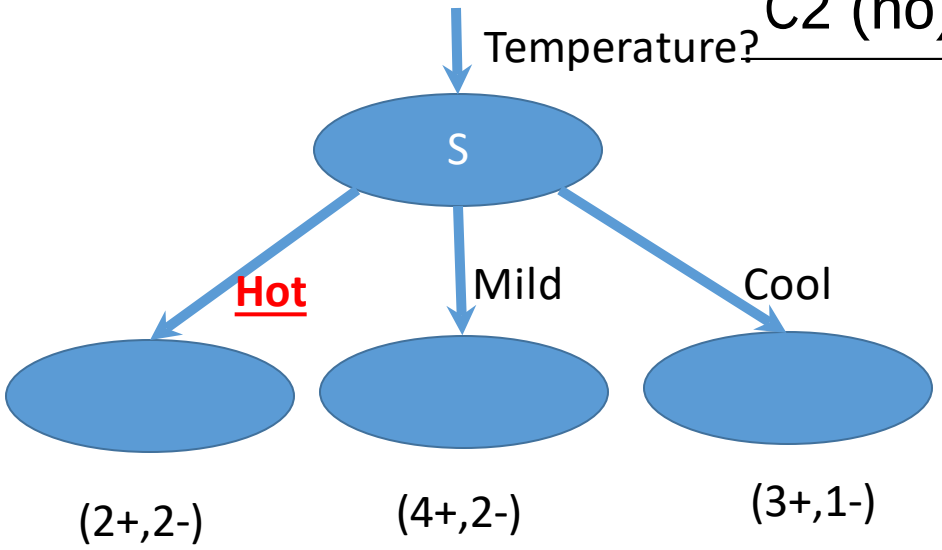


outlook	humidity	wind	play
rain	normal	weak	yes
rain	normal	strong	no
sunny	normal	weak	yes
overcast	normal	strong	yes

Cálculo de la Ganancia de Información para Temperature

C1 (yes)	**9**
C2 (no)	**5**

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$



GAIN = Entropía antes de dividir–Entropía después

Entropía antes:

$$Entropy([9+,\ 5-]) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = .940$$

$GAIN(S,Temperature) = Entropy\ ([9+,5-]) - \Sigma\,(n_i/n)(\ - p_{yes}\ log_2\ p_{yes}\ - p_{no}\ log_2\ p_{no}\ )$

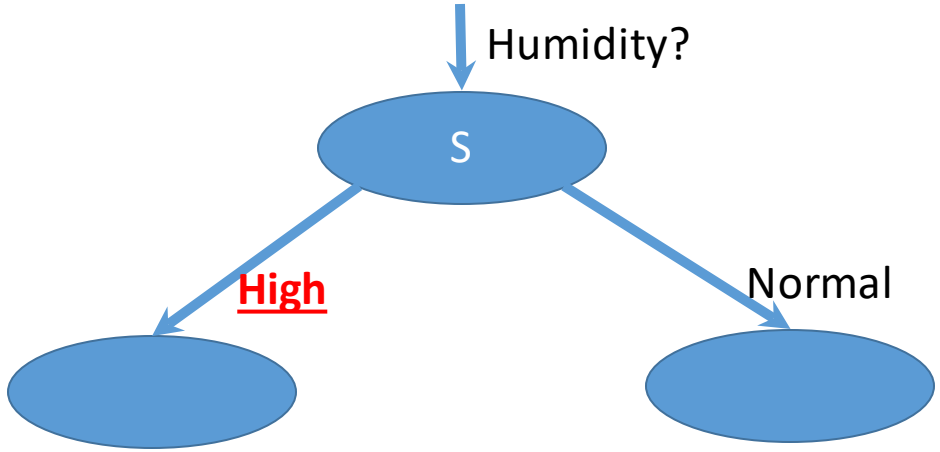
$GAIN(S,\ Temperature) = 0.940 - \{\ (4/14)(-2/4\ log_2\ 2/4\ - 2/4\ log_2\ 2/4)$

$- (\ 6/14)(-4/6\ log_2\ 4/6\ - 2/6\ log_2\ 2/6)\ \ - (\ 4/14)(-3/4\ log_2\ 3/4\ - 1/4\ log_2\ 1/4)\}$

$Gain(D,Temperature) = \mathbf{0.029}$

Tabla para Humidity = High

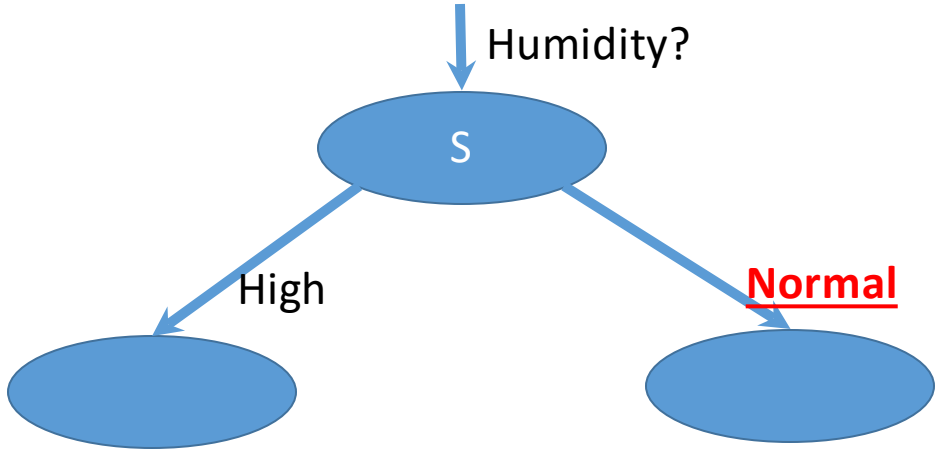
outlook	temperature	humidity	wind	play
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rain	mild	high	weak	yes
rain	cool	normal	weak	yes
rain	cool	normal	strong	no
overcast	mild	high	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rain	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	hot	normal	weak	yes
overcast	cool	normal	strong	yes
rain	mild	high	strong	no



outlook	temperature	wind	play
sunny	hot	weak	no
sunny	hot	strong	no
overcast	hot	weak	yes
rain	mild	weak	yes
overcast	mild	strong	yes
sunny	mild	weak	no
rain	mild	strong	no

Tabla para Humidity = Normal

outlook	temperature	humidity	wind	play
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rain	mild	high	weak	yes
rain	cool	normal	weak	yes
rain	cool	normal	strong	no
overcast	mild	high	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rain	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	hot	normal	weak	yes
overcast	cool	normal	strong	yes
rain	mild	high	strong	no

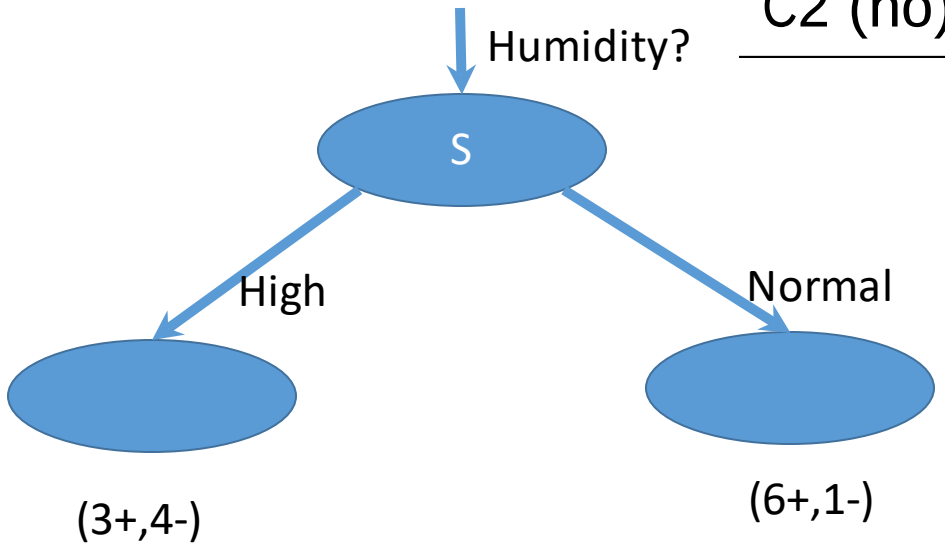


outlook	temperature	wind	play
rain	cool	weak	yes
rain	cool	strong	no
sunny	cool	weak	yes
rain	mild	weak	yes
sunny	mild	strong	yes
overcast	hot	weak	yes
overcast	cool	strong	yes

Cálculo de la Ganancia de Información para Humidity

C1 (yes)	**9**
C2 (no)	**5**

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$



GAIN = Entropía antes de dividir–Entropía después

Entropía antes:

$$Entropy([9+,\ 5-]) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = .940$$

$GAIN(S,Humidity) = Entropy\,([9+,5-]) - \Sigma\,(n_i/n)(\,-p_{yes}\,\log_2 p_{yes} - p_{no}\,\log_2 p_{no}\,)$

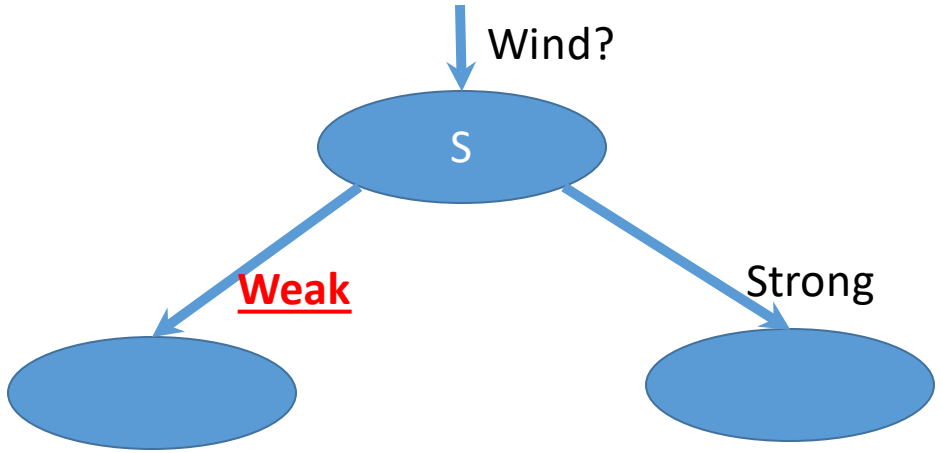
$GAIN(S, Humidity) = 0.940 - \{\,(7/14)(-3/7\,\log_2 3/7 - 4/7\,\log_2 4/7)$

$-\,(7/14)(-6/7\,\log_2 6/7 - 1/7\,\log_2 1/7)\}$

$Gain(D, Humidity) = \mathbf{0.151}$

Tabla para Wind =Weak

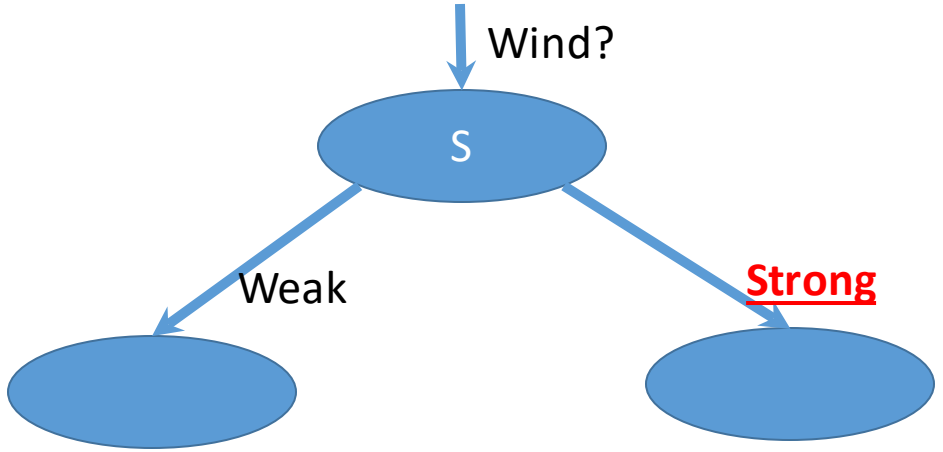
outlook	temperature	humidity	wind	play
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rain	mild	high	weak	yes
rain	cool	normal	weak	yes
rain	cool	normal	strong	no
overcast	mild	high	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rain	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	hot	normal	weak	yes
overcast	cool	normal	strong	yes
rain	mild	high	strong	no



outlook	temperature	humidity	play
sunny	hot	high	no
overcast	hot	high	yes
rain	mild	high	yes
rain	cool	normal	yes
sunny	mild	high	no
sunny	cool	normal	yes
rain	mild	normal	yes
overcast	hot	normal	yes

Tabla para Wind = Strong

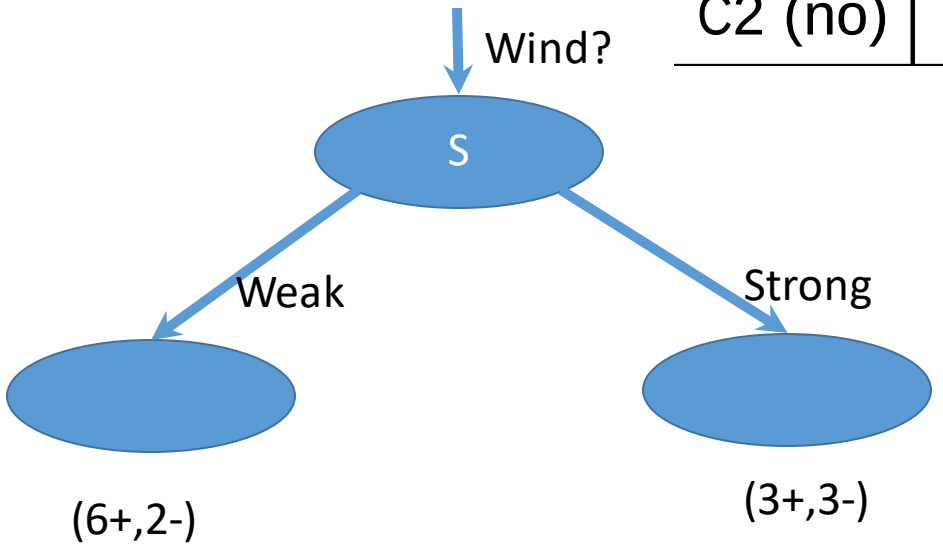
outlook	temperature	humidity	wind	play
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rain	mild	high	weak	yes
rain	cool	normal	weak	yes
rain	cool	normal	strong	no
overcast	mild	high	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rain	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	hot	normal	weak	yes
overcast	cool	normal	strong	yes
rain	mild	high	strong	no



outlook	temperature	humidity	play
sunny	hot	high	no
rain	cool	normal	no
overcast	mild	high	yes
sunny	mild	normal	yes
overcast	cool	normal	yes
rain	mild	high	no

Cálculo de la Ganancia de Información para Wind

C1 (yes)	**9**
C2 (no)	**5**



$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

GAIN = Entropía antes de dividir–Entropía después

Entropía antes:

$$Entropy([9+,\ 5-]) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = .940$$

$GAIN(S,Wind) = Entropy\ ([9+,5-]) - \Sigma\,(n_i/n)(\ -\ p_{yes}\ log_2\ p_{yes}\ -\ p_{no}\ log_2\ p_{no}\ )$

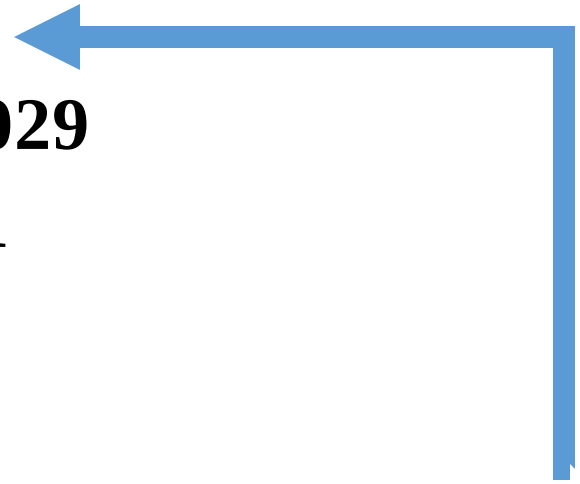
$GAIN(S,\ Wind) = 0.940 - \{\ (\ 8/14)(-6/8\ log_2\ 6/8\ -\ 2/8\ log_2\ 2/8)$

$-\ (\ 6/14)(-3/6\ log_2\ 3/6\ -\ 3/6\ log_2\ 3/6)\}$

$Gain(D,\ Wind) =$ **0.048**

Resumen de las GAIN de los 4 atributos

1. *Gain*(*S*,*Outlook*) = **0.246**
2. *Gain*(*D*,*Temperature*) = **0.029**
3. *Gain*(*D*, *Humidity*) = **0.151**
4. *Gain*(*D*, *Wind*) = **0.048**

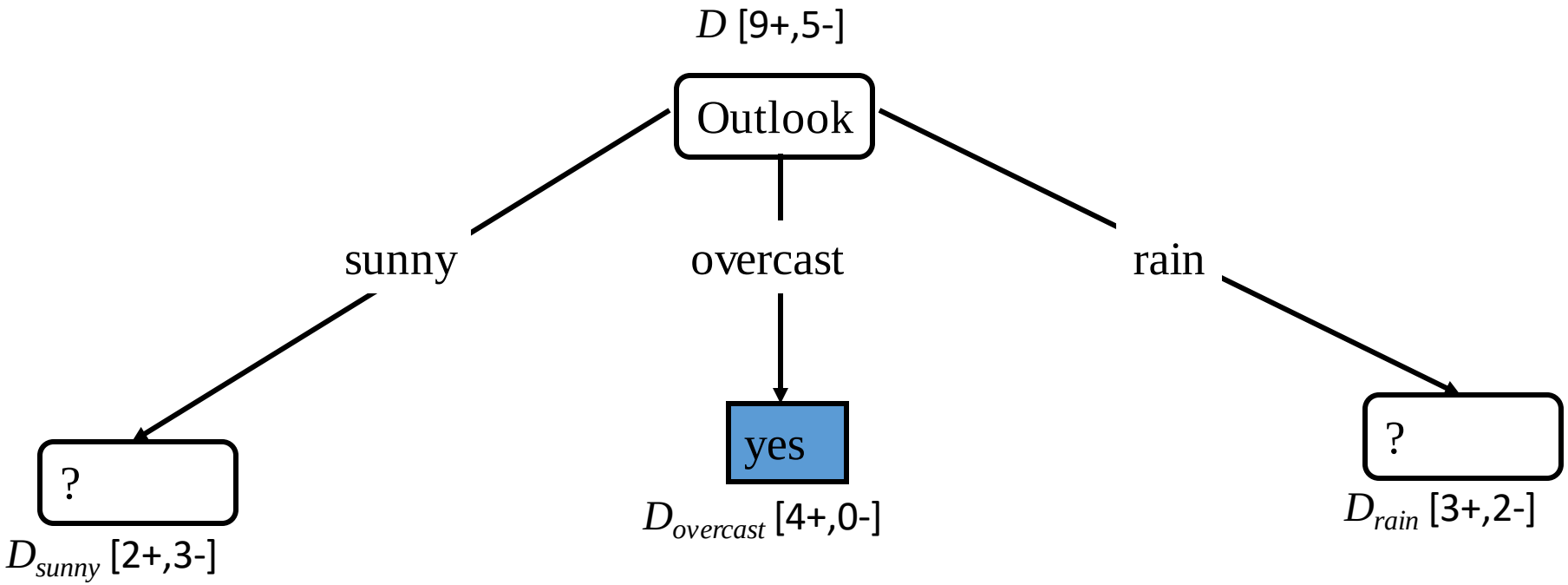


¿Cuál es la Máxima Ganancia de Información de entre los cuatro atributos?

⇩

Por tanto, la primera decisión se tomará por el atributo Outlook

Primer nodo de decisión del árbol

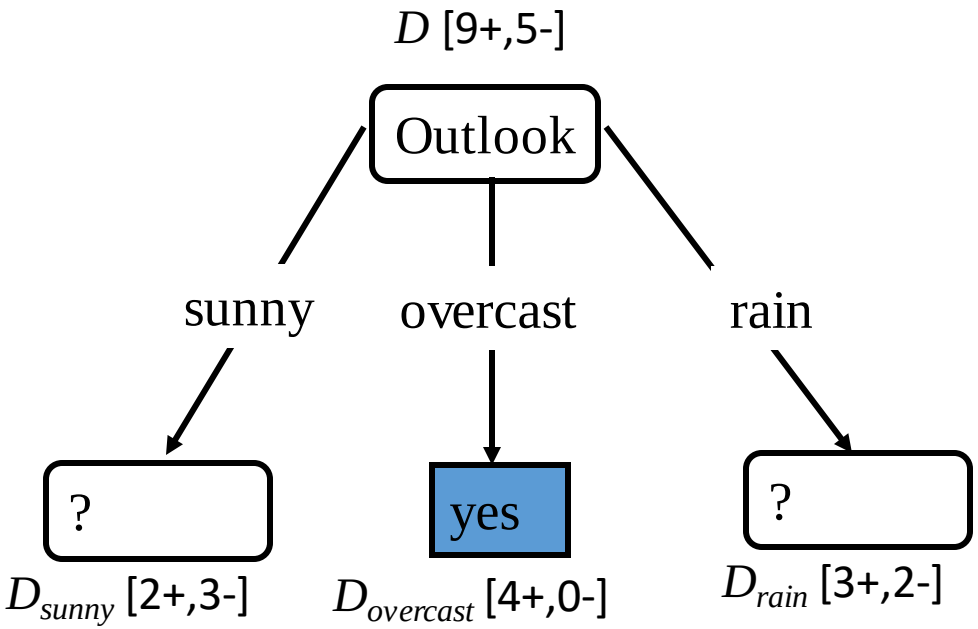


temperat.	humidity	wind	play
hot	high	weak	no
hot	high	strong	no
mild	high	weak	no
cool	normal	weak	yes
mild	normal	strong	yes

temperat.	humidity	wind	play
hot	high	weak	yes
mild	high	strong	yes
hot	normal	weak	yes
cool	normal	strong	yes

temperat.	humidity	wind	play
mild	high	weak	yes
cool	normal	weak	yes
cool	normal	strong	no
mild	normal	weak	yes
mild	high	strong	no

Nodo de decisión al que se dirige Outlook=Sunny



temperat.	humidity	wind	play
hot	high	weak	no
hot	high	strong	no
mild	high	weak	no
cool	normal	weak	yes
mild	normal	strong	yes

$D_{sunny} = [2+, 3-]$

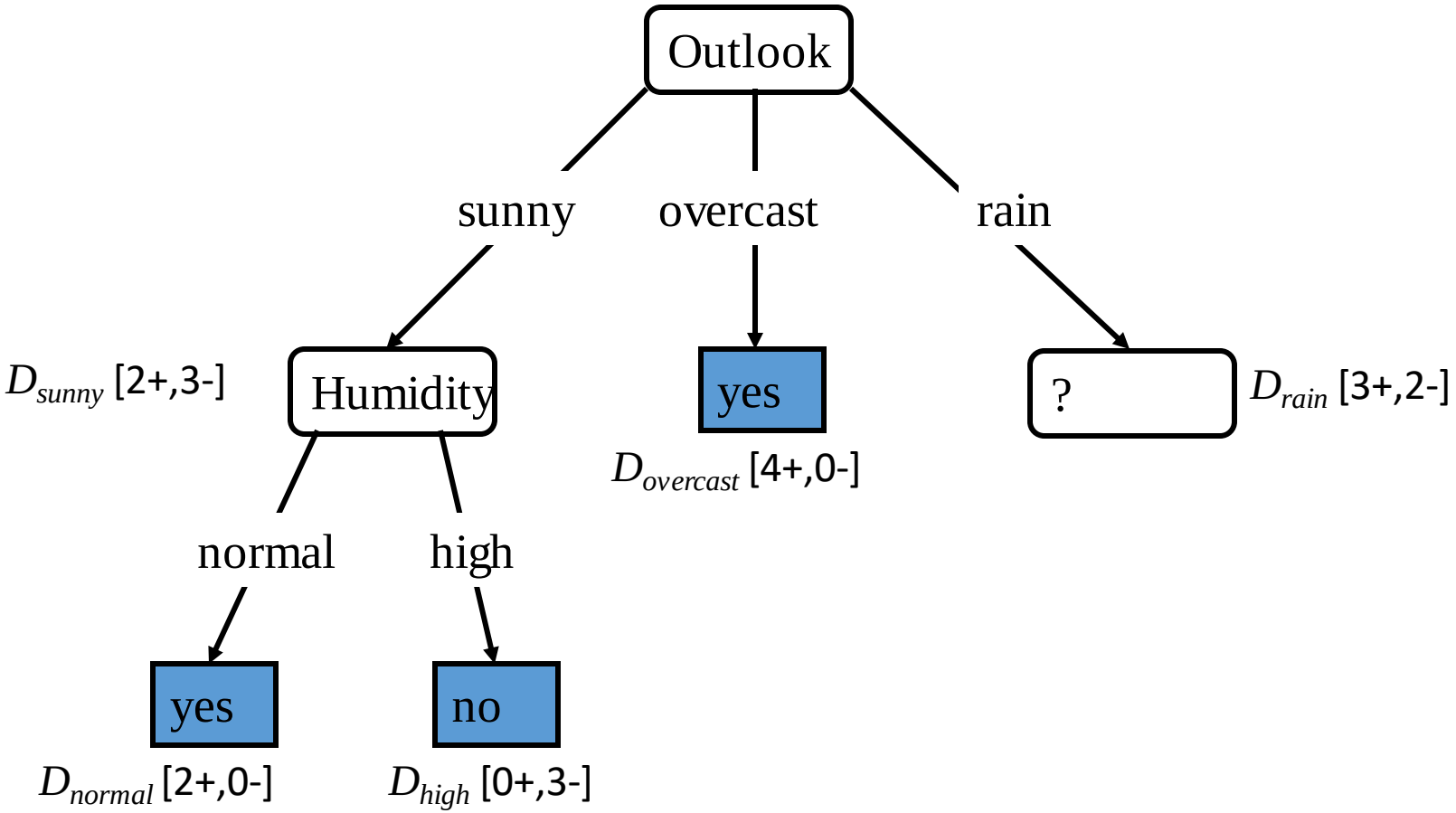
$E(D_{sunny}) = 0.970$

$Gain(D_{sunny}, Wind) = 0.019$

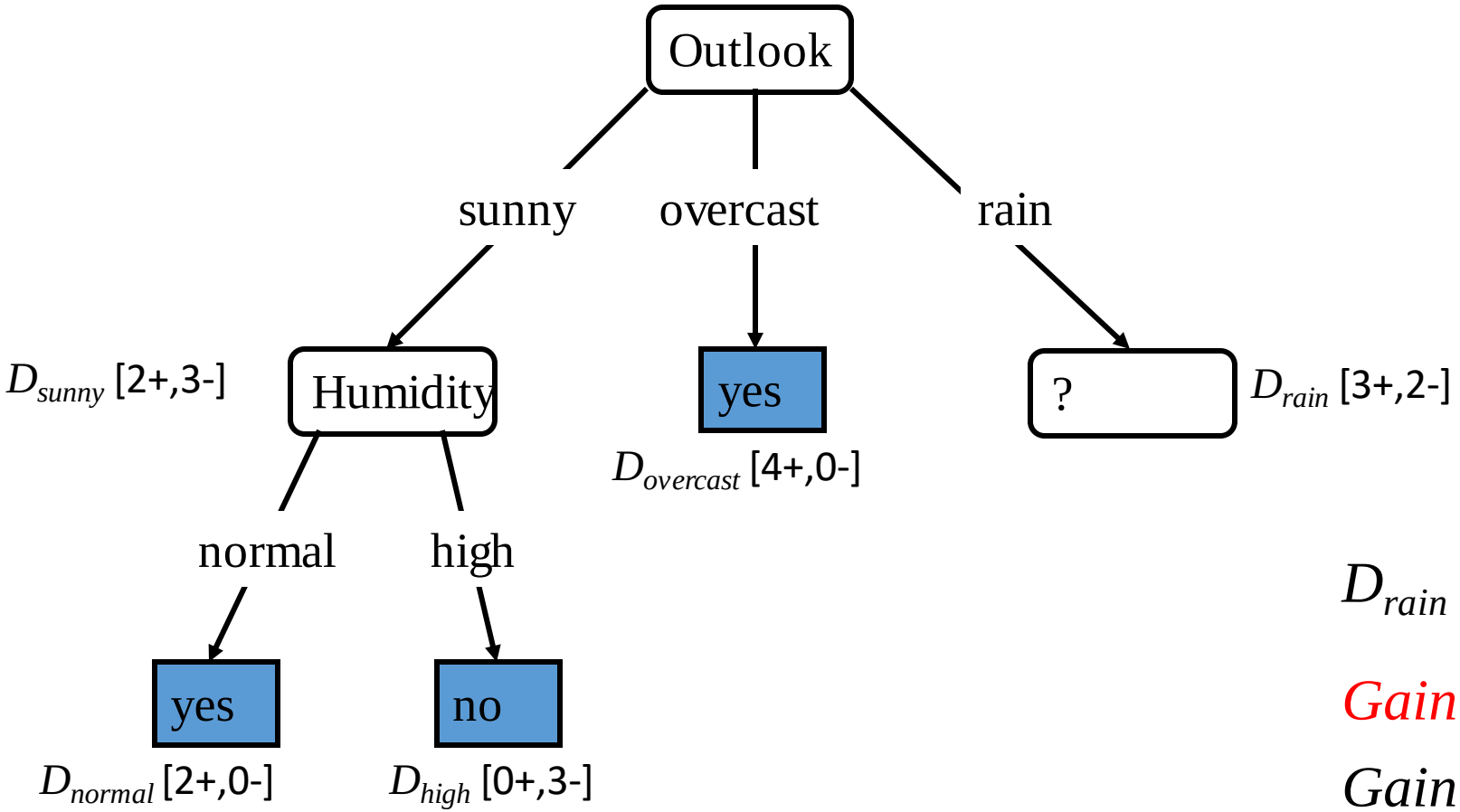
$Gain(D_{sunny}, Humidity) = 0.970 \leftarrow$

$Gain(D_{sunny}, Temperature) = 0.570$

Segundo Nodo resultante



Nodo de decisión al que se dirige Outlook=Rain



temperature	humidity	wind	play
mild	high	weak	yes
cool	normal	weak	yes
cool	normal	strong	no
mild	normal	weak	yes
mild	high	strong	no

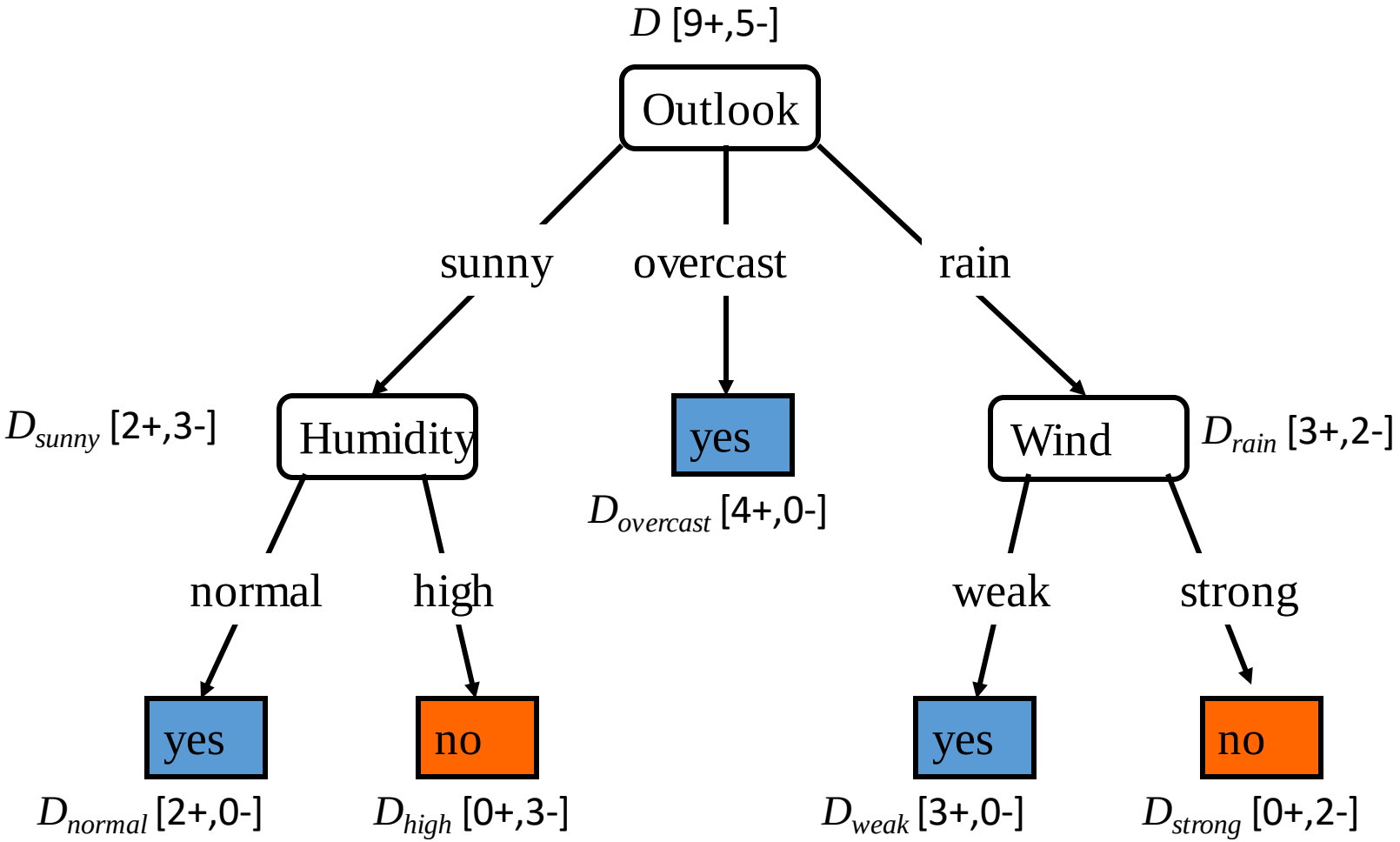
D_{rain} = [3+, 2-], $E(D_{rain})$ = 0.970

$Gain(D_{sunny}, Wind)$ = 0.970 ←

$Gain(D_{sunny}, Humidity)$ = 0.0192

$Gain(D_{sunny}, Temperature)$ = 0.0192

Árbol Final


D [9+,5-]
Outlook
sunny
overcast
rain
D_sunny [2+,3-]
Humidity
yes
D_overcast [4+,0-]
Wind
D_rain [3+,2-]
normal
high
weak
strong
yes
no
yes
no
D_normal [2+,0-]
D_high [0+,3-]
D_weak [3+,0-]
D_strong [0+,2-]

División basada en Teoría de la Información (cont)

Gain Ratio:

$$GainRATIO_{split} = \frac{GAIN_{Split}}{SplitINFO}$$

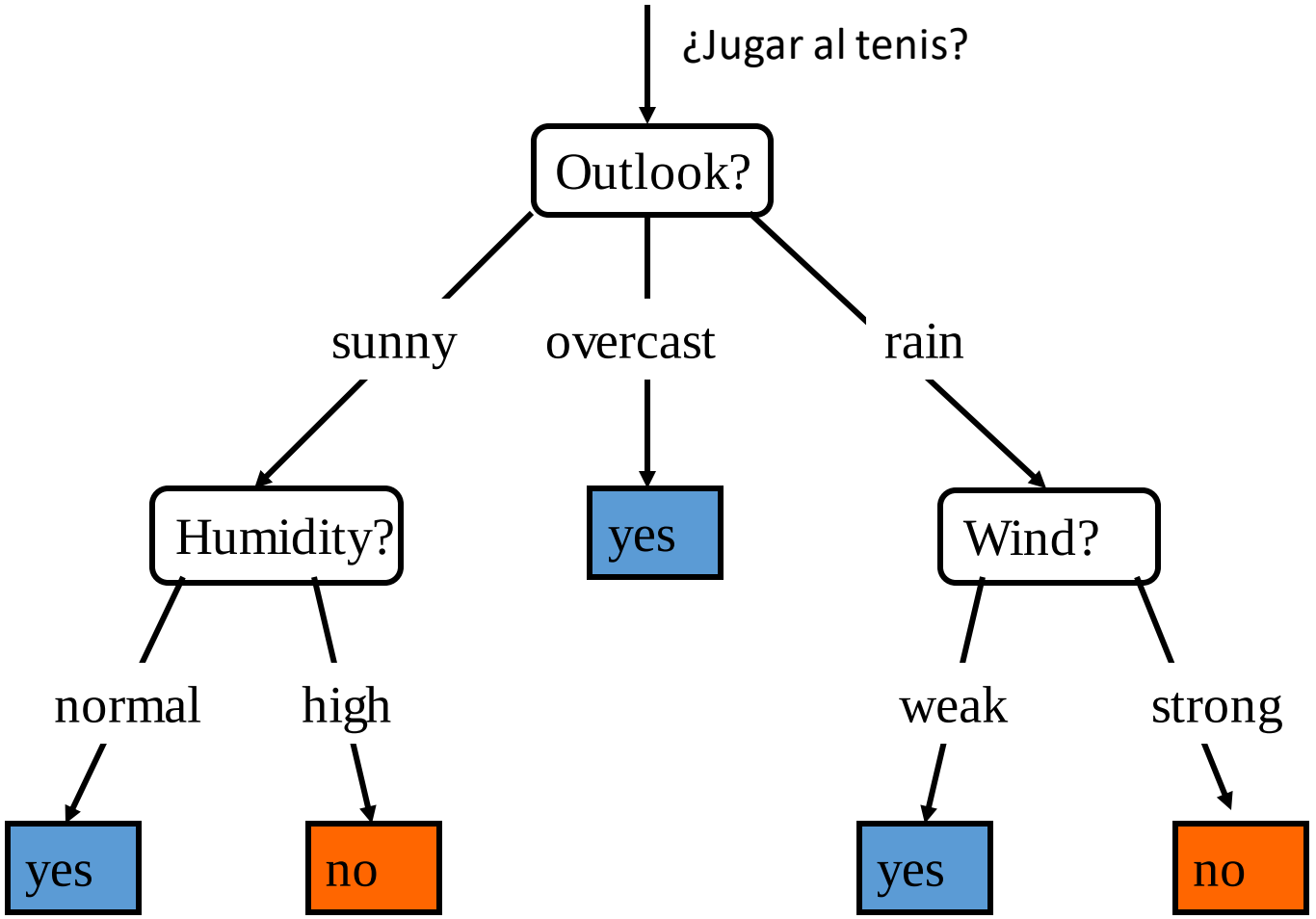
donde

$$SplitINFO = -\sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n}$$

- Pondera la ganancia de información por la entropía del particionado (SplitINFO) → Se penalizan particiones con alta entropía (es decir, muchas particiones pequeñas)!!

- Diseñado para resolver el inconveniente de la ganancia de información

Interpretabilidad de los árboles



- Automáticamente transformables en reglas:

1) IF Outlook=sunny AND

   Humidity = normal THEN Play = yes

2) IF Outlook=sunny AND

   Humidity = high THEN Play = no

3) IF Outlook=overcast THEN Play = yes

4) IF Outlook=rain AND

   Wind = weak THEN Play =yes

5) IF Outlook=rain AND

   Wind = strong THEN Play =no

FIN