M2.2.2 Modelos Supervisados y No Supervisados

Programa Big Data y Business Intelligence

Enrique Onieva
enrique.onieva@deusto.es
https://twitter.com/EnriqueOnieva
https://www.linkedin.com/in/enriqueonieva/

Deusto
Facultad de Ingeniería
Ingeniaritza Fakultatea

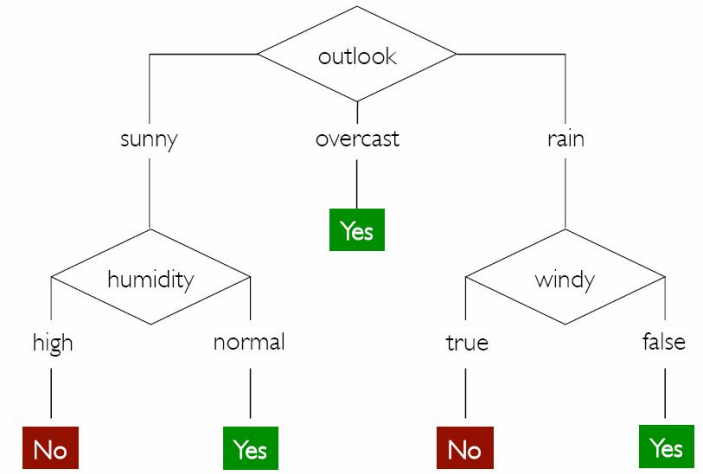
Árboles de Decisión

- Motivación y esquema de funcionamiento
- Construcción del árbol
  - ¿Cómo partir los datos?
  - ¿Cuándo parar de partir los datos?
- Evitando el overfitting
- Aplicación a Regresión

Motivación

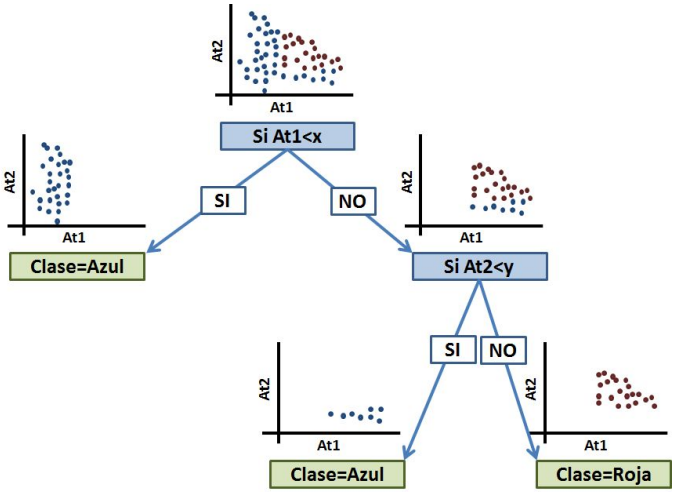
- Se basan en el principio “divide y vencerás”
  - Toman los datos de entrada y los van partiendo en pedazos para los que sea “más fácil“ ajustar un modelo
  - Los datos se van partiendo en función de una condición aplicada sobre uno (o más) atributos

Outlook	Temp	Humidity	Windy	Play
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



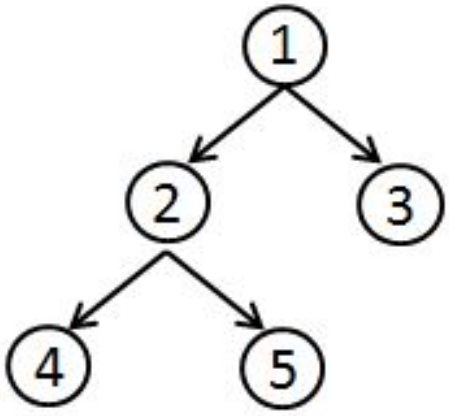
Motivación

- Se basan en dos procesos iterativos
  - ¿Cómo elegir la manera de partir los datos?
    - ¿Qué atributo escojo para dividir?
    - ¿Cómo construyo la condición para dividir en partes?
  - ¿Cuándo dejar de partir los datos?
    - ¿Hasta que obtenga un resultado perfecto? (overfitting)
  - Modificaciones de éstos dan lugar a diferentes métodos de creación de árboles



Terminología

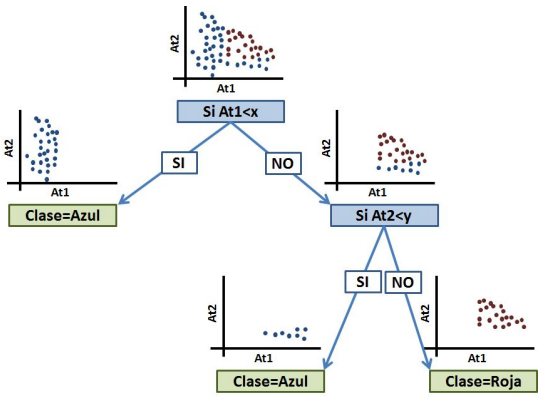
- Relaciones
  - Hijo/Padre:
  - Ascendente/Descendiente:
- Nodos:
  - Raíz: Nodo inicial (sin padres). (1)
  - Hoja: Nodo final (sin hijos). (3) (4) (5)
  - Intermedio: Ni inicial ni final. (2) (3)



- Profundidad de un nodo:
  - Número de pasos (ascendentes) necesarios para llegar al nodo raíz. Profundidad de (4) = 2

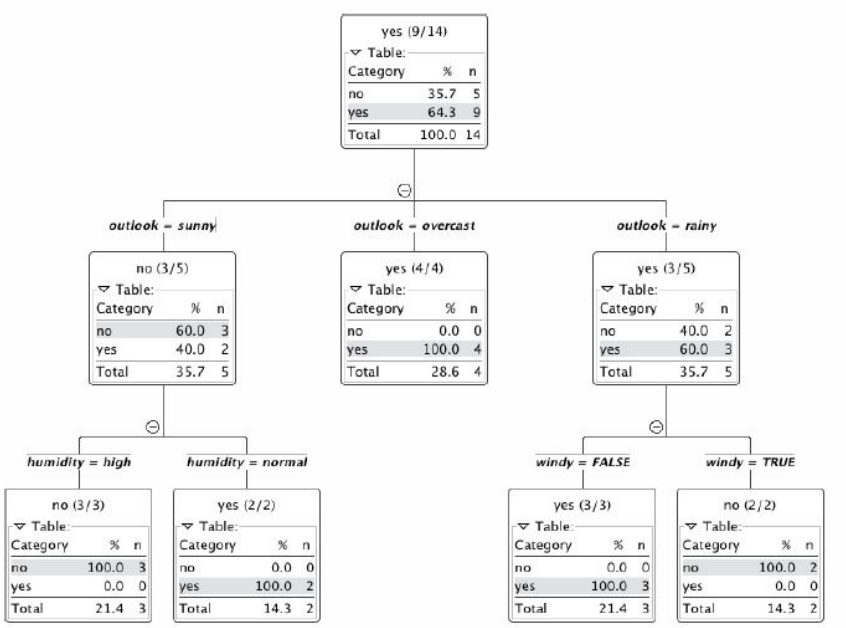
Introducción

- Por lo general, cada nodo interno es condición sobre un atributo (Si At1<x)
- Cada rama representa representa uno de los resultados posibles de esa prueba (SI y NO)
- Un caso nuevo se resuelve siguiendo las condiciones hasta alcanzar un nodo hoja.
- Los nodos hoja representan:
  - Clasificación: la clase a predecir, o la distribución de clases
  - Regresión: la manera de calcular el valor de salida



Introducción

- El Modelo se puede representar de una manera gráfica, o por medio de texto (condiciones)



outlook = overcast: yes {no=0, yes=4}
outlook = rainy
| windy = FALSE: yes {no=0, yes=3}
| windy = TRUE: no {no=2, yes=0}
outlook = sunny
| humidity = high: no {no=3, yes=0}
| humidity = normal: yes {no=0, yes=2}

¿Cómo se construye el árbol?

- Construcción de arriba hacia abajo
  - Inicialmente, los ejemplos están en el nodo raíz
  - Entonces, se toma la decisión de cómo dividirlos
  - Este proceso se realiza de una manera recursiva
  - ¿Cuándo paramos?
    - Todos los ejemplos son de la misma clase
    - No ganamos nada dividiendo
- Poda del árbol hacia arriba
  - Eliminamos ramas, de manera que se obtengan mejores resultados ante ejemplos nuevos

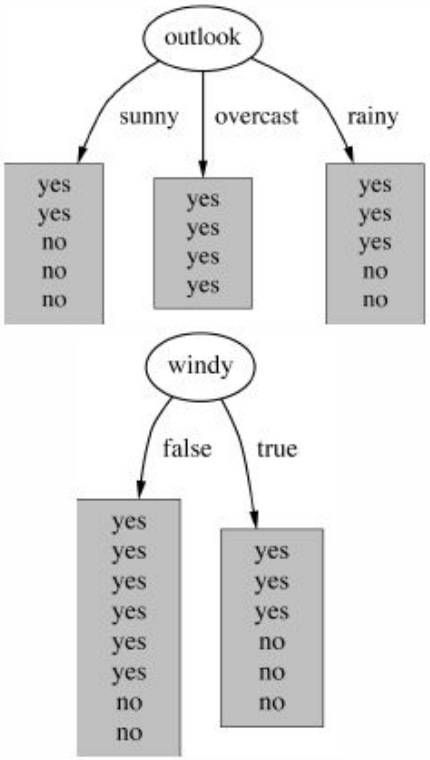
¿Cómo determinar el atributo?

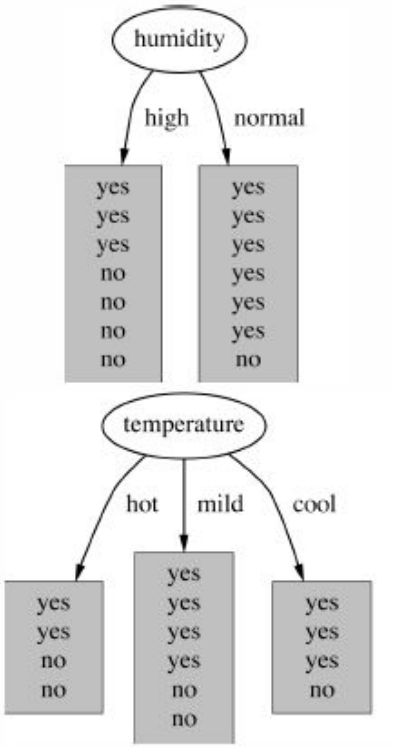
- En cada nodo, los atributos se evalúan según cómo de bien separan las clases de los datos
  - Se utilizan medidas de pureza o impureza
  - Ganancia de Información: función que crece conforme crece la pureza de los sub-conjuntos que produce el atributo
- Estrategia: Elegir el atributo con mejor ganancia de información
- Ejemplos:
  - Usar Ganancia de Información → Método ID3
  - Usar el Ratio de Ganancia de Información → Método C4.5 *
  - Usar el Índice Gini → Método CART

¿Cómo determinar el atributo?

- ¿Cuál escogerías?

Outlook	Temp	Humidity	Windy	Play
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





Ganancia de Información

- Es la diferencia entre la información antes y después de cortar según un atributo

$$gain(A) = info(D) - info_A(D)$$

- La información antes de cortar, es la entropía

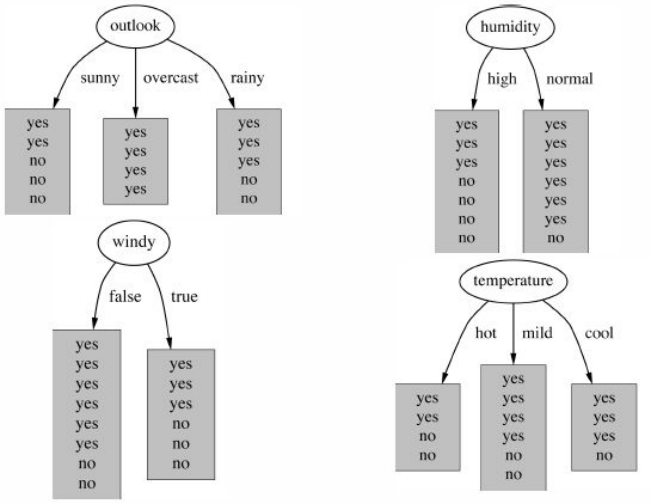
$$info(D) = -p_1 \log p_1 - \ldots - p_n \log p_n$$

- La información tras el corte se calcula como la suma ponderada de las entropías en cada corte

$$info_A(D) = \frac{|D_1|}{|D|} info(D_1) + \ldots + \frac{|D_n|}{|D|} info(D_n)$$

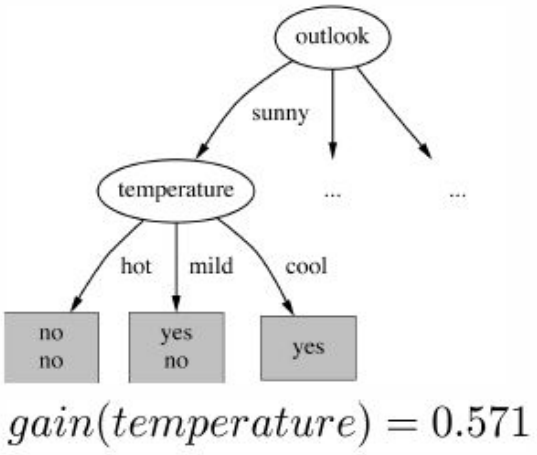
Ganancia de Información

- Haciendo cuentas:
  - Ganancia de Información("outlook") = 0.247
  - Ganancia de Información("humidity") = 0.029
  - Ganancia de Información("windy") = 0.152
  - Ganancia de Información("temperature") = 0.048
  - En un primer paso, cortaremos según "outlook"

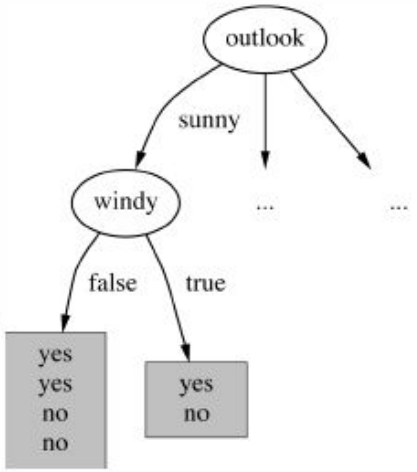


Ganancia de Información

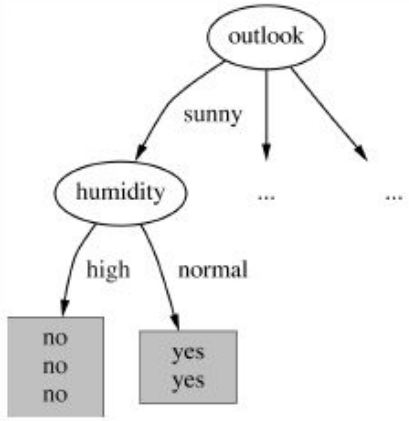
- Un paso más allá
  - ¿Cómo cortaremos los datos que cumplen outlook=sunny?



$gain(temperature) = 0.571$



$gain(windy) = 0.020$

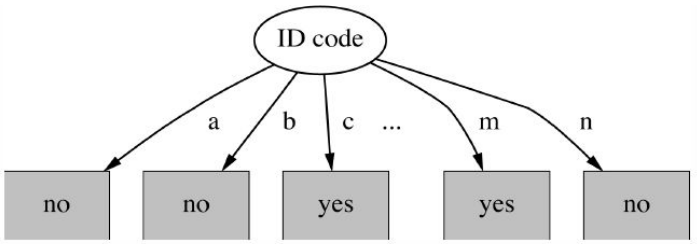


$gain(humidity) = 0.971$

Ganancia de Información

- Presenta problemas para atributos (categóricos) con muchos valores
  - En un caso extremo, con un ID, utilizar dicho campo para cortar nos da ganancia de información MÁXIMA
  - La ganancia de información presenta “preferencia” por atributos con muchos valores posibles
  - Ésto puede resultar en Overfitting

ID Code	Outlook	Temp	Humidity	Windy	Play
A	Sunny	Hot	High	False	No
B	Sunny	Hot	High	True	No
C	Overcast	Hot	High	False	Yes
D	Rainy	Mild	High	False	Yes
E	Rainy	Cool	Normal	False	Yes



Ratio de “Information Gain”

- Modificación de la ganancia de información que reduce la “preferencia” por atributos con muchos valores posibles
  - Tiene en cuenta el número y tamaño de cada una de las posibles ramas resultantes a la hora de tomar la decisión de escoger el atributo para partir

Ratio de “Information Gain”

- Información intrínseca

$$IntrinsicInfo(S, A) = -\sum \frac{|S_i|}{|S|} \log \frac{|S_i|}{|S|}$$

- El ratio de ganancia de información normaliza la ganancia de información

$$GainRatio(S, A) = \frac{Gain(S, A)}{IntrinsicInfo(S, A)}$$

¿Y con atributos numéricos?

- Para el atributo Temperatura
  - Ordenar el según los valores del atributo numérico (incluyendo la clase)
  - Comprobar todos los posibles puntos de corte, y elegir aquél con mejor medida
    - {Temperatura<71.5 || Temperatura≥71.5} …
  - Poner los puntos de corte en puntos intermedios entre valores

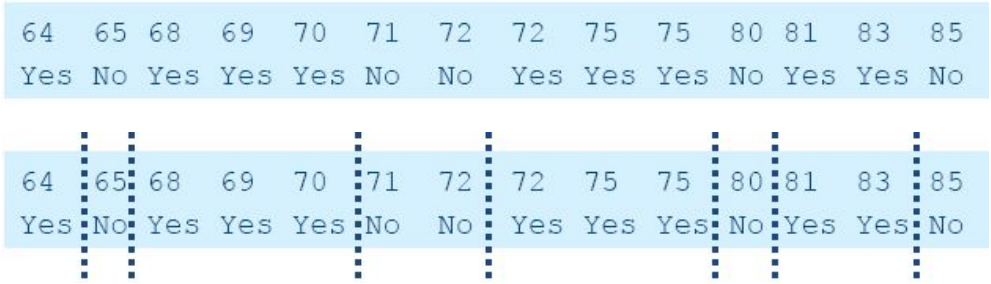
Outlook	Temp	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	78	False	Yes
Rainy	70	96	False	Yes
Rainy	68	80	False	Yes
Rainy	65	70	True	No
Overcast	64	65	True	Yes
Sunny	72	95	False	No
Sunny	69	70	False	Yes
Rainy	75	80	False	Yes
Sunny	75	70	True	Yes
Overcast	72	90	True	Yes
Overcast	81	75	False	Yes
Rainy	71	80	True	No

64	65	68	69	70	71	72	72	75	75	80	81	83	85
Yes	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	No

¿Y con atributos numéricos?

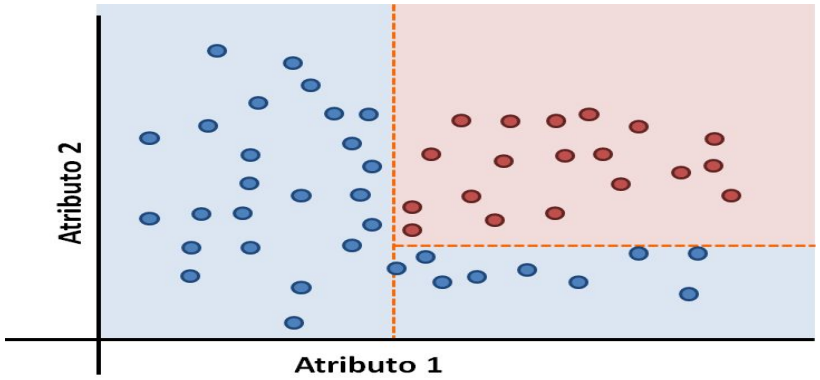
- Para el atributo Temperatura
  - Ordenar el según los valores del atributo numérico (incluyendo la clase)
  - ¿Tiene sentido probar en 68.5?
  - Poner los puntos de corte en puntos intermedio entre valores, siempre que las clases sean diferentes

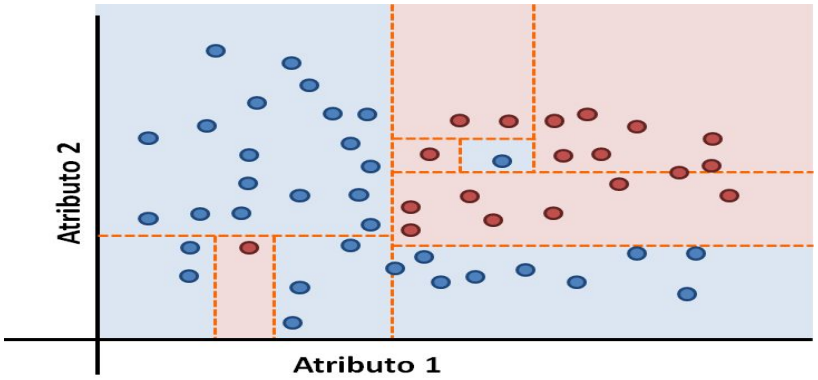
Outlook	Temp	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	78	False	Yes
Rainy	70	96	False	Yes
Rainy	68	80	False	Yes
Rainy	65	70	True	No
Overcast	64	65	True	Yes
Sunny	72	95	False	No
Sunny	69	70	False	Yes
Rainy	75	80	False	Yes
Sunny	75	70	True	Yes
Overcast	72	90	True	Yes
Overcast	81	75	False	Yes
Rainy	71	80	True	No



¿Cuándo dejar de partir?

- Siguiendo el proceso de elegir atributo, partir y repetir recursivamente…
  - Siempre podremos obtener árboles "perfectos" (para el conjunto de entrenamiento)
  - Muy sensibles al ruido y outliers en los datos
  - Sobreajustados (Overfitting)



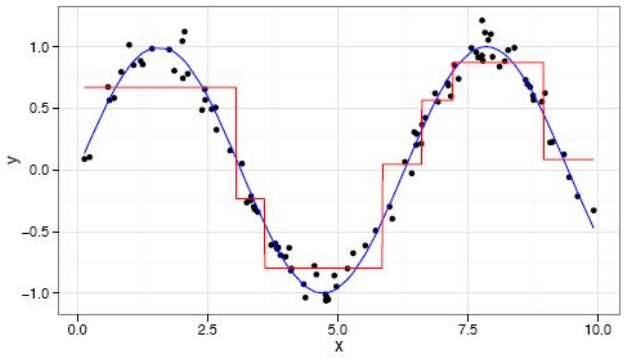


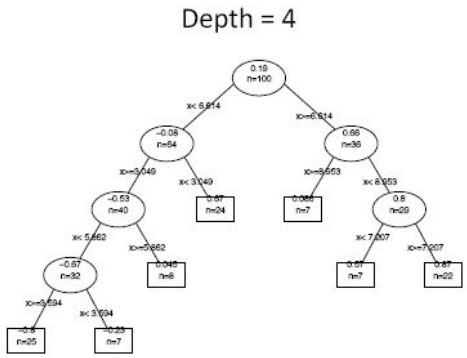
Evitando el Overfitting

- Prepruning ("Poda previa")
  - Dejar de generar ramas
  - No partir un nodo si el resultado en cierta medida cae por debajo del umbral
  - No partir un nodo con menos de cierto número de muestras
- Postpruning ("Poda posterior")
  - Eliminar algunas ramas de un árbol completo
  - Usar datos de validación para podar el árbol

Árboles de Regresión

- En cada hoja, en lugar de la clase, se asigna el valor promedio de los datos en esa hoja
- La medida de la bondad de un corte se basa en medidas de regresión
  - Normalmente el error cuadrático medio





Resúmen

- Ventajas
  - Clasificador válido siempre
  - Muy automatizado, todo tipo de inputs y missing values
    - Missing values → un valor más
  - Descarta variables no importantes
  - Fácil de interpretar
- Desventajas
  - Los cortes se suelen hacer en base a variables que tengan muchos niveles, Fácil overfitting
  - Sensibles a cambios en los datos de entrenamiento
  - Árboles frondosos difíciles de interpretar y pueden repetir variables

Enrique Onieva
enrique.onieva@deusto.es
https://twitter.com/EnriqueOnieva
https://www.linkedin.com/in/enriqueonieva/