HUB IQ

Inteligencia Artificial & Machine Learning

Applicaciones en movilidad

Dr. Iván S. Razo Zapata

EngineeringX
Founded by the Royal Academy of Engineering and Lloyd's Register Foundation

Aprendizaje Supervisado

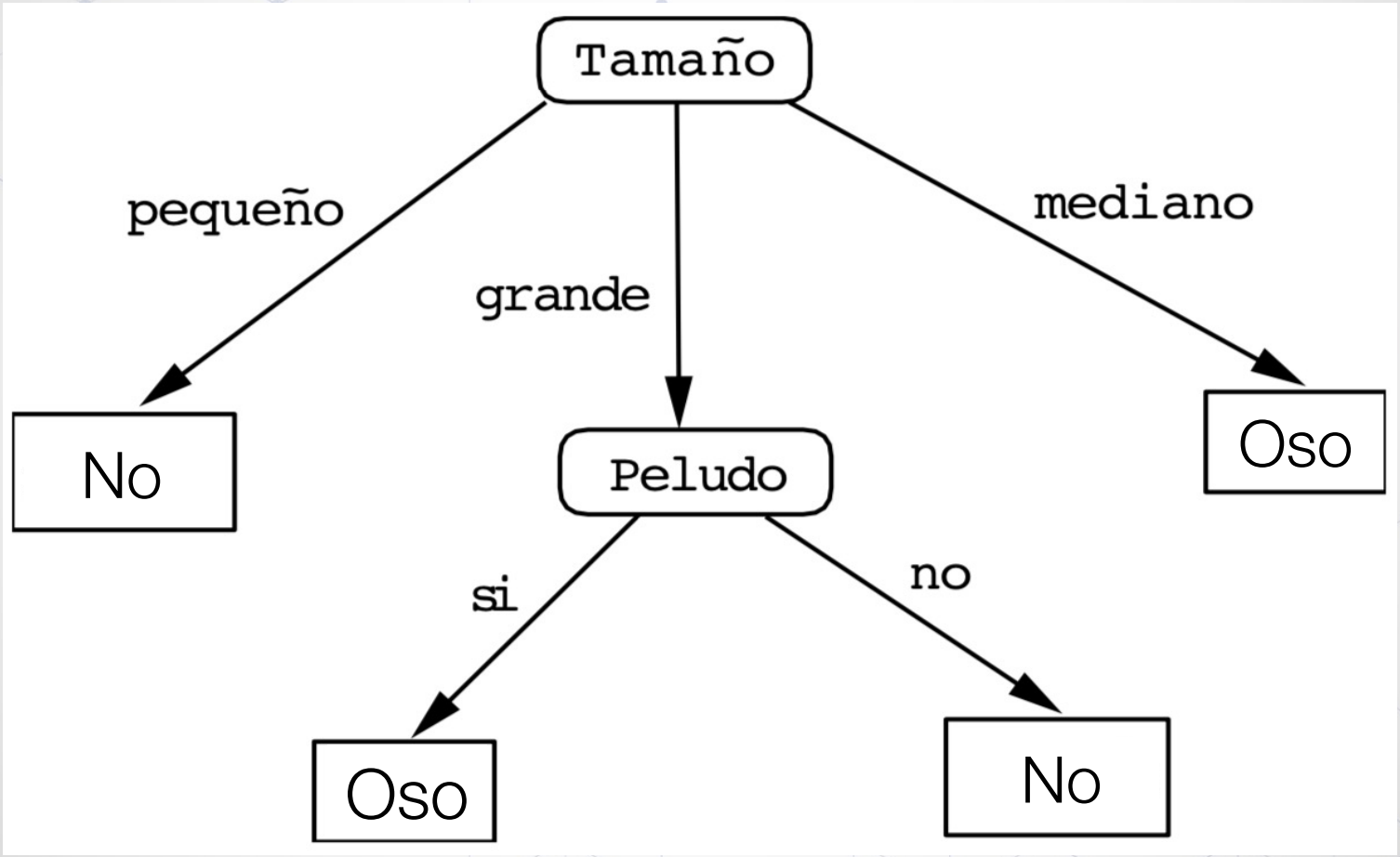
Árboles de Decisión – Decision Trees

Introducción

Imaginemos un problema clasificación

ID	Peludo	Edad	Tamaño	Clase
1	Si	Viejo	Grande	Oso
2	No	Joven	Grande	No
3	Si	Joven	Mediano	Oso
4	Si	Viejo	Pequeño	No
5	Si	Joven	Pequeño	No
6	Si	Joven	Grande	Oso
7	No	Joven	Pequeño	No
8	No	Viejo	Grande	No

Introducción



Intro

Brumotactillofobia

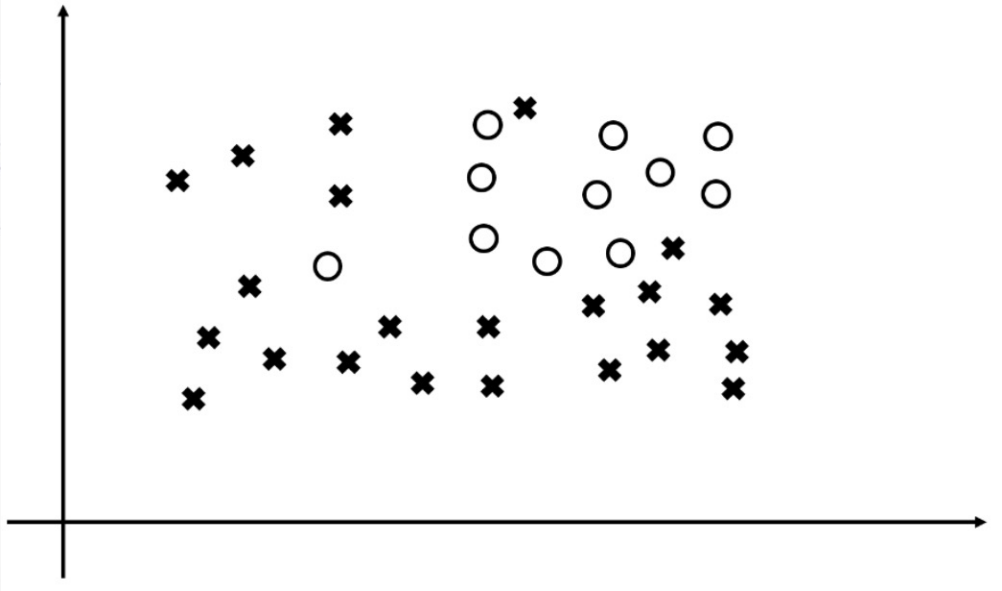
- Una versión leve del trastorno obsesivo-compulsivo, especialmente común en niños (y algunos adultos)
- necesidad de probar cada bocado de comida por separado, o
- necesidad de tener un plato que se vea **ordenado** y **organizado**, o
- querer saber exactamente qué están comiendo en cada bocado, o
- puede ser una cuestión de textura.

Intro

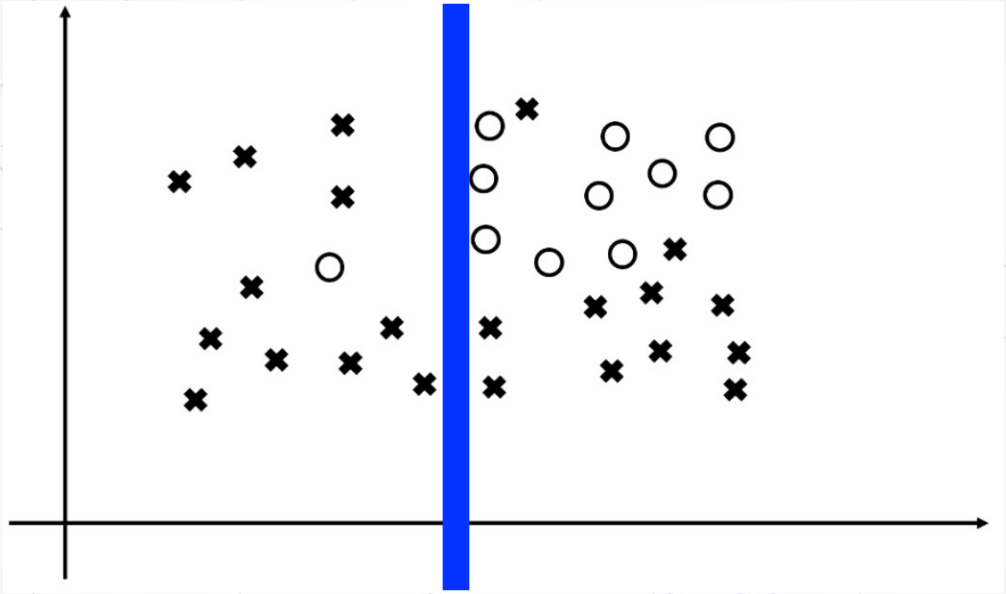
Brumotactillofobia

Try each component **alone** and **unadulterated**

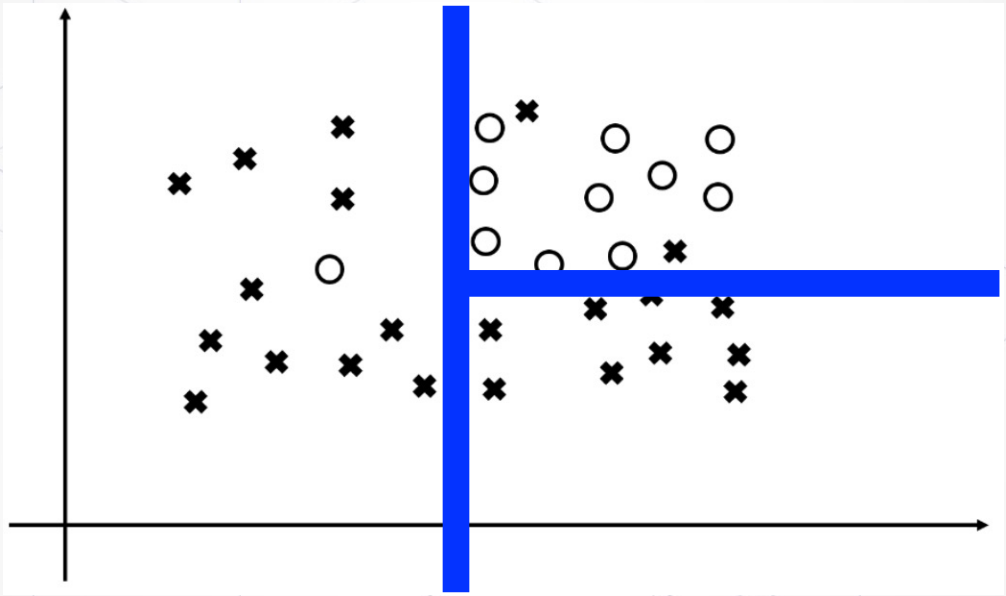
Intro



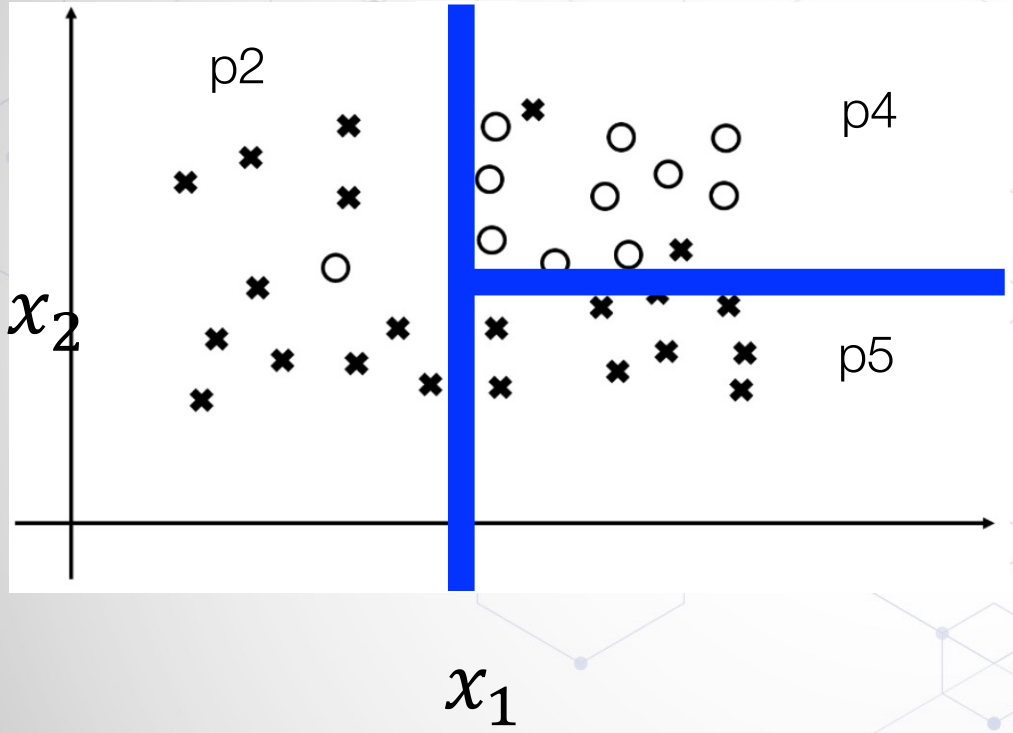
Intro

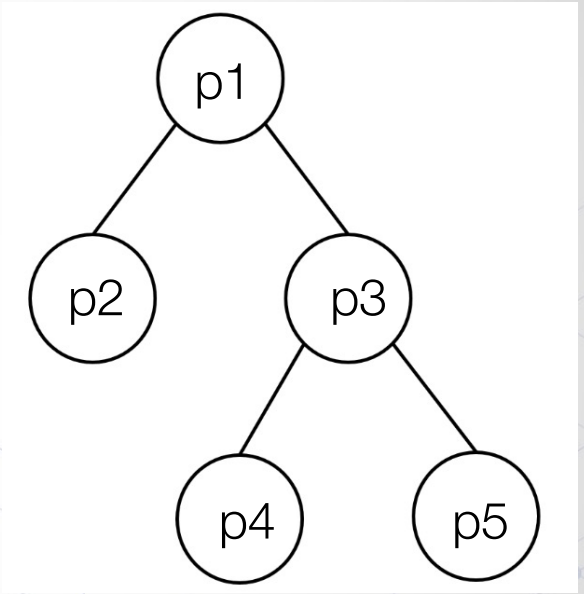


Intro

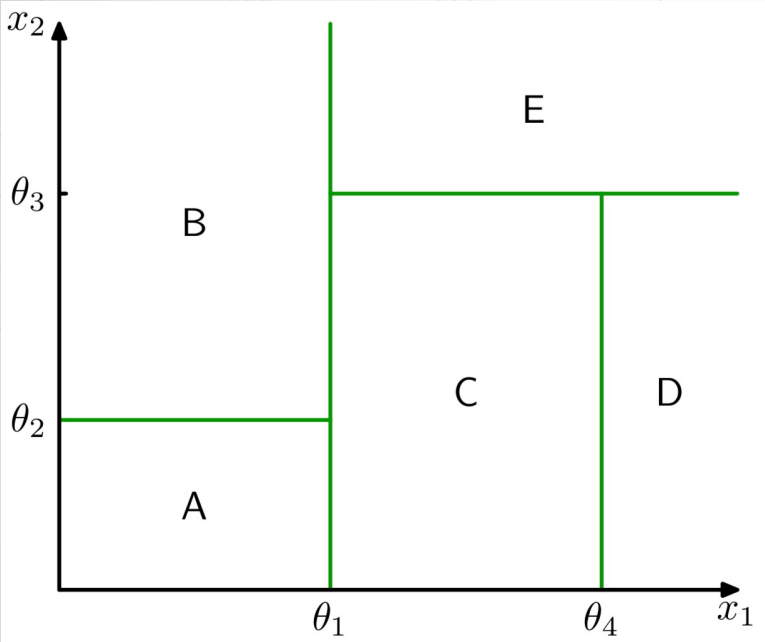


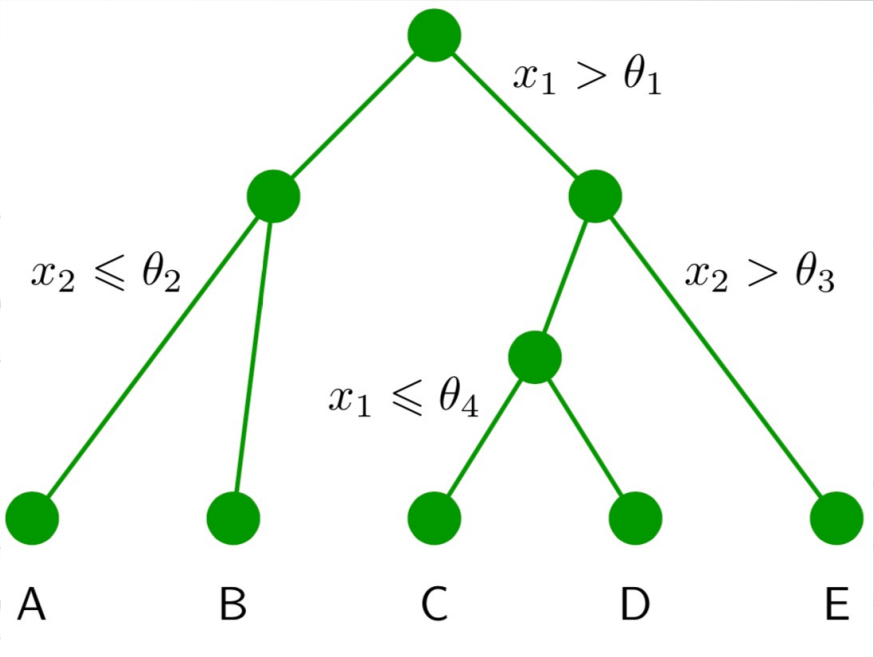
Intro



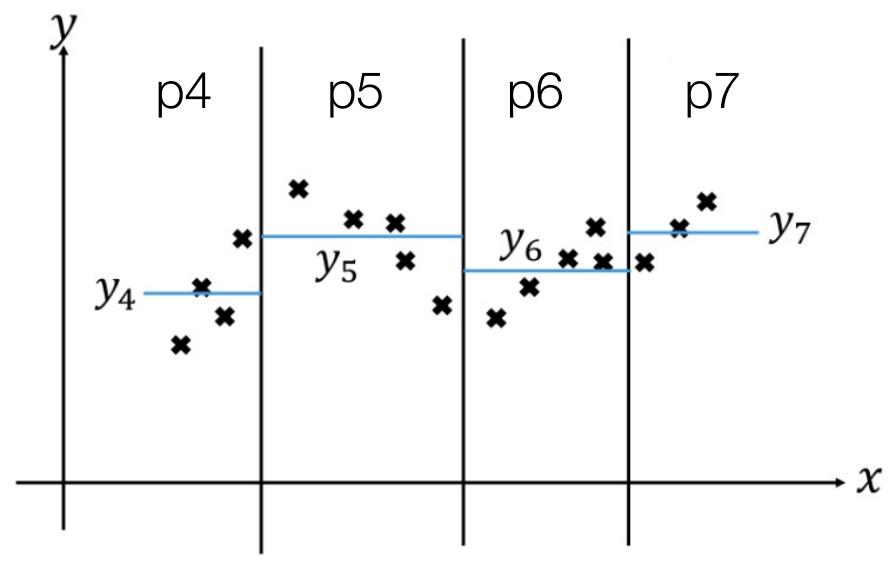


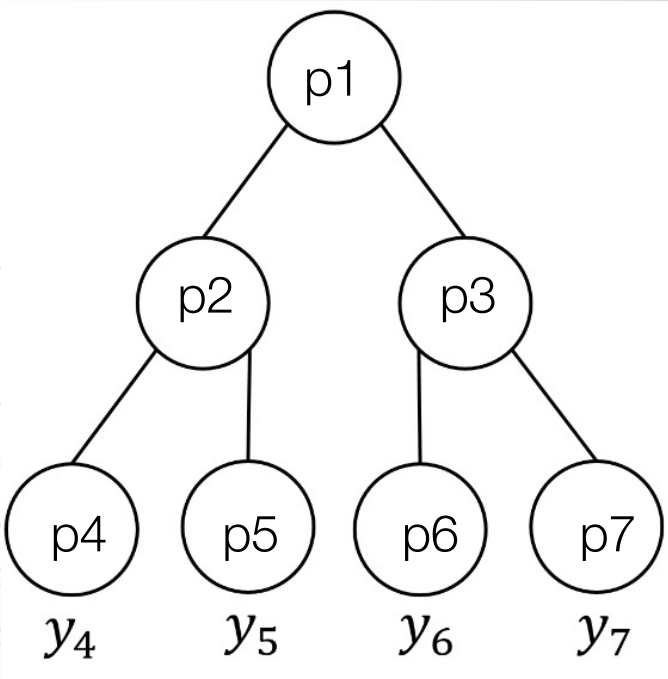
Clasificación





Regresión





Intro

- El aprendizaje de árboles de decisión es sencillo, fácil de implementar y poderoso
- Un árbol recibe un objeto o situación descrita por un conjunto de atributos y regresa una decisión
- Cada nodo interno corresponde a una prueba en el valor de uno de los atributos y las ramas están etiquetadas con los posibles valores de la prueba
- Cada hoja especifica el valor de la clase

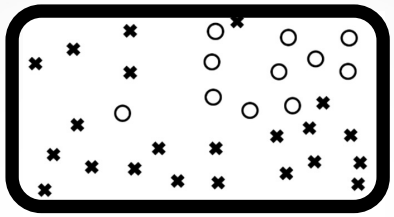
Idea

- Probar primero el atributo más “importante”
- Esto particiona los ejemplos y cada subconjunto es un nuevo problema con menos ejemplos
- Este proceso recursivo tiene 4 posibles resultados:

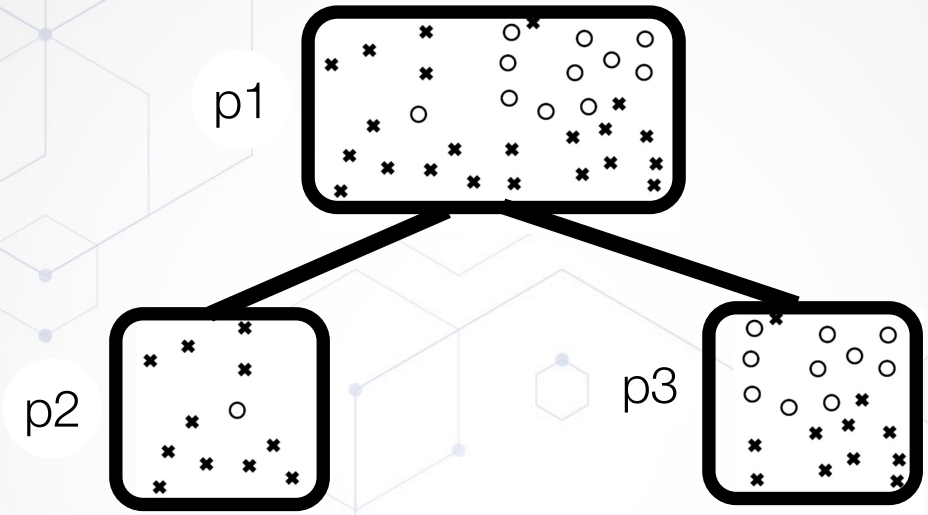
1. Si existen ejemplos positivos y negativos, escoge el “mejor atributo”
2. Si todos los ejemplos son positivos (o negativos), termina y regresa True (o False)
3. No quedan ejemplos, regresa un default con base en la clasificación mayoritaria de su nodo padre
4. No hay más atributos, pero seguimos con ejemplos positivos y negativos. Posible solución: toma la clase mayoritaria

Ejemplo

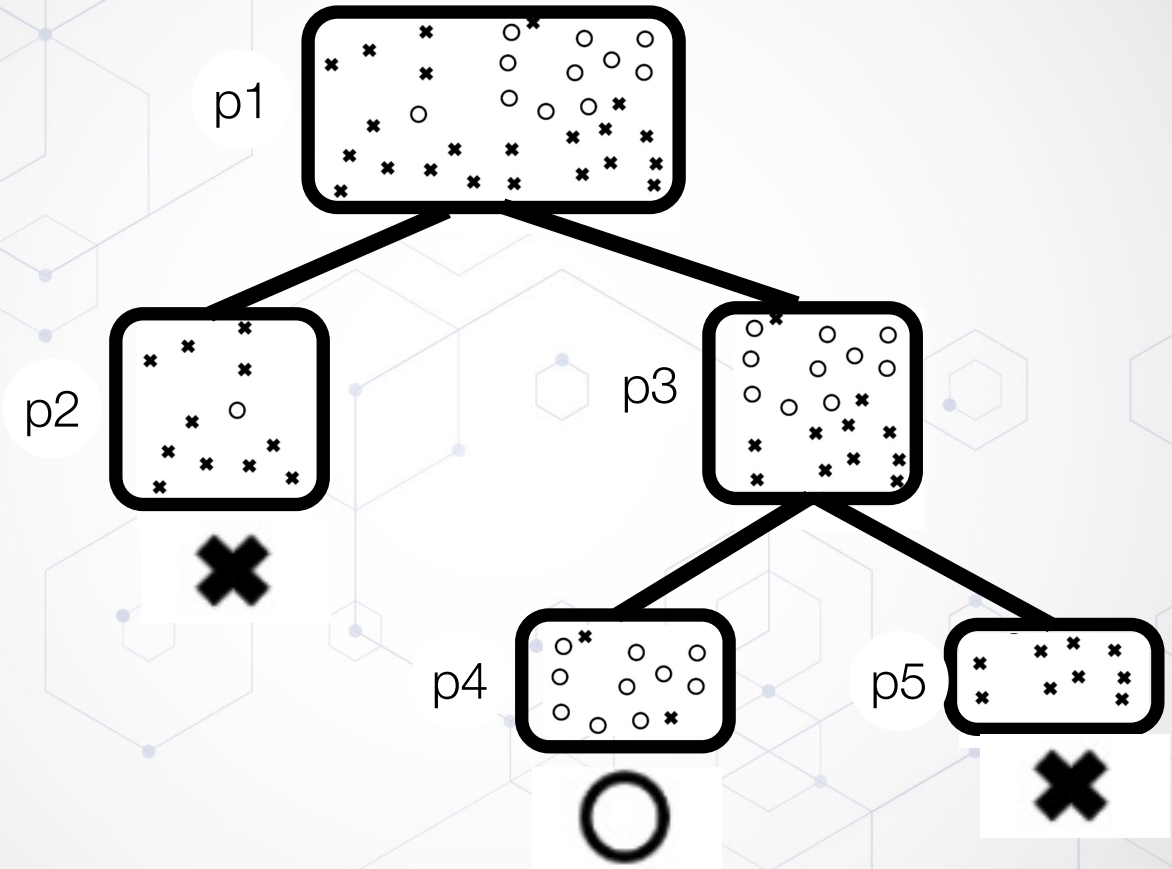
p1



Ejemplo



Ejemplo



Ejemplo

ID	Peludo	Edad	Tamaño	Clase
1	Si	Viejo	Grande	Oso
2	No	Joven	Grande	No
3	Si	Joven	Mediano	Oso
4	Si	Viejo	Pequeño	No
5	Si	Joven	Pequeño	No
6	Si	Joven	Grande	Oso
7	No	Joven	Pequeño	No
8	No	Viejo	Grande	No

Ejemplo

ID	Peludo	Edad	Tamaño	Clase
1	Si	Viejo	Grande	Oso
2	No	Joven	Grande	No
3	Si	Joven	Mediano	Oso
4	Si	Viejo	Pequeño	No
5	Si	Joven	Pequeño	No
6	Si	Joven	Grande	Oso
7	No	Joven	Pequeño	No
8	No	Viejo	Grande	No

Peludo

Si

ID	Edad	Tamaño	Clase
1	Viejo	Grande	Oso
3	Joven	Mediano	Oso
4	Viejo	Pequeño	No
5	Joven	Pequeño	No
6	Joven	Grande	Oso

No

ID	Edad	Tamaño	Clase
2	Joven	Grande	No
7	Joven	Pequeño	No
8	Viejo	Grande	No

Edad

Joven

ID	Tamaño	Clase
3	Mediano	Oso
5	Pequeño	No
6	Grande	Oso

Viejo

ID	Tamaño	Clase
1	Grande	Oso
4	Pequeño	No

Tamaño

Pequeño

ID	Clase
5	No

Mediano

ID	Clase
3	Oso

Grande

ID	Clase
6	Oso

Pequeño

ID	Clase
4	No

Grande

ID	Clase
1	Oso

Ejemplo

ID	Peludo	Edad	Tamaño	Clase
1	Si	Viejo	Grande	Oso
2	No	Joven	Grande	No
3	Si	Joven	Mediano	Oso
4	Si	Viejo	Pequeño	No
5	Si	Joven	Pequeño	No
6	Si	Joven	Grande	Oso
7	No	Joven	Pequeño	No
8	No	Viejo	Grande	No

Tamaño

Pequeño

ID	Peludo	Edad	Clase
4	Si	Viejo	No
5	Si	Joven	No
7	No	Joven	No

Grande

ID	Peludo	Edad	Clase
1	Si	Viejo	Oso
2	No	Joven	No
6	Si	Joven	Oso
8	No	Viejo	No

Mediano

ID	Peludo	Edad	Clase
3	Si	Joven	Oso

Peludo

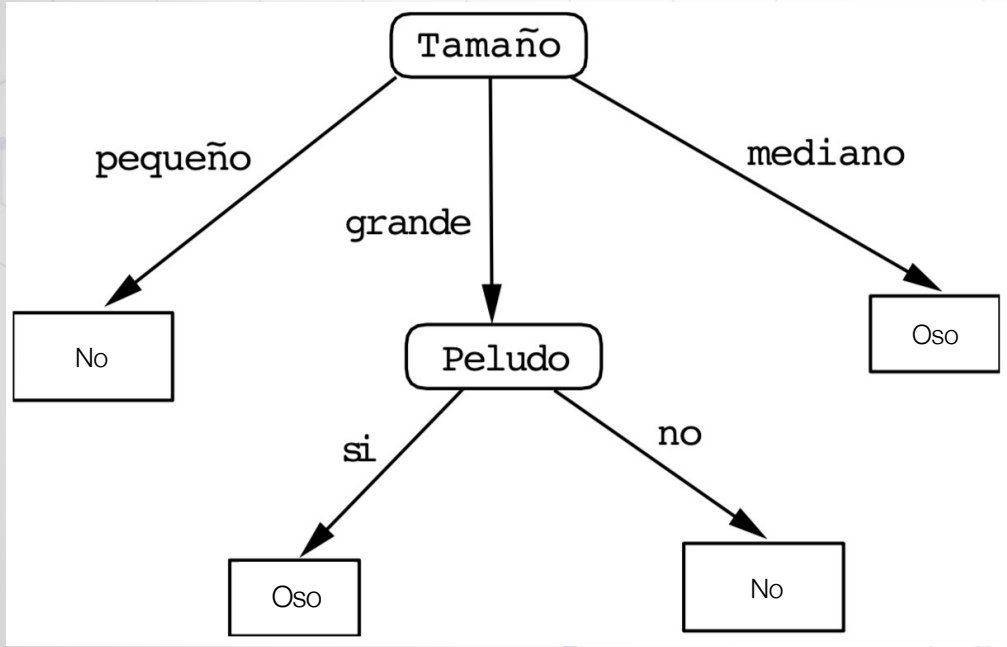
Si

ID	Edad	Clase
1	Viejo	Oso
6	Joven	Oso

No

ID	Edad	Clase
2	Joven	No
8	Viejo	No

Ejemplo



- Reglas
- IF … then

Algoritmos

- ID3
- C4.5
- CART
- Random Forest

¿Cómo seleccionar el “mejor” atributo?

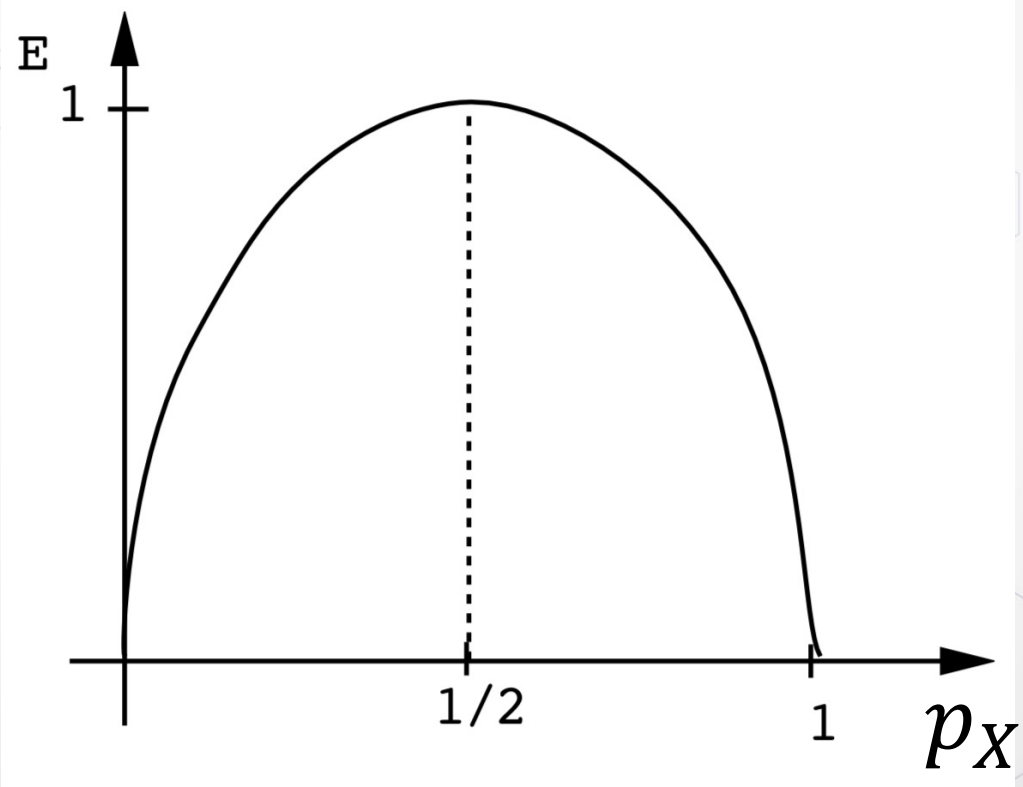
Ideas iniciales

- Entropia
  - Caracteriza la “impureza” de un conjunto de observaciones
- Asumiendo una muestra S con observaciones pertenecientes a dos clases:
  - Positivos: p
  - Negativos: n

$$Entropia(S) = -\frac{p}{p+n} log_2 \frac{p}{p+n} - \frac{n}{p+n} log_2 \frac{n}{p+n}$$

Entropia

$$Entropia(S) = -p_A log_2(p_A) - p_B log_2(p_B)$$



Ejemplo - ¿Jugar o no tenis?

ID	Ambiente	Temp.	Humedad	Viento	Clase
1	soleado	alta	alta	no	N
2	soleado	alta	alta	si	N
3	nublado	alta	alta	no	P
4	lluvioso	media	alta	no	P
5	lluvioso	baja	normal	no	P
6	lluvioso	baja	normal	si	N
7	nublado	baja	normal	si	P
8	soleado	media	alta	no	N
9	soleado	baja	normal	no	P
10	lluvioso	media	normal	no	P
11	soleado	media	normal	si	P
12	nublado	media	alta	si	P
13	nublado	alta	normal	no	P
14	lluvioso	media	alta	si	N

Ejemplo

ID	Ambiente	Temp.	Humedad	Viento	Clase
1	soleado	alta	alta	no	N
2	soleado	alta	alta	si	N
3	nublado	alta	alta	no	P
4	lluvioso	media	alta	no	P
5	lluvioso	baja	normal	no	P
6	lluvioso	baja	normal	si	N
7	nublado	baja	normal	si	P
8	soleado	media	alta	no	N
9	soleado	baja	normal	no	P
10	lluvioso	media	normal	no	P
11	soleado	media	normal	si	P
12	nublado	media	alta	si	P
13	nublado	alta	normal	no	P
14	lluvioso	media	alta	si	N

$$Entropia(S) = -\frac{p}{p+n} log_2 \frac{p}{p+n} - \frac{n}{p+n} log_2 \frac{p}{p+n}$$

$$Entropia(S) = -\frac{9}{9+5} log_2 \frac{9}{9+5} - \frac{5}{9+5} log_2 \frac{5}{9+5} = 0.940$$

Ejemplo

ID	Ambiente	Temp.	Humedad	Viento	Clase
1	soleado	alta	alta	no	N
2	soleado	alta	alta	si	N
3	nublado	alta	alta	no	P
4	lluvioso	media	alta	no	P
5	lluvioso	baja	normal	no	P
6	lluvioso	baja	normal	si	N
7	nublado	baja	normal	si	P
8	soleado	media	alta	no	N
9	soleado	baja	normal	no	P
10	lluvioso	media	normal	no	P
11	soleado	media	normal	si	P
12	nublado	media	alta	si	P
13	nublado	alta	normal	no	P
14	lluvioso	media	alta	si	N

- Efectividad de un atributo para clasificar las muestras
- **Ganancia de Información**: Reducción esperada en la entropía al usar dicho atributo para particionar la muestra

$$Ganancia(A) = Entropia(S) - \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} Entropia(S_v)$$

Ejemplo

ID	Ambiente	Temp.	Humedad	Viento	Clase
1	soleado	alta	alta	no	N
2	soleado	alta	alta	si	N
3	nublado	alta	alta	no	P
4	lluvioso	media	alta	no	P
5	lluvioso	baja	normal	no	P
6	lluvioso	baja	normal	si	N
7	nublado	baja	normal	si	P
8	soleado	media	alta	no	N
9	soleado	baja	normal	no	P
10	lluvioso	media	normal	no	P
11	soleado	media	normal	si	P
12	nublado	media	alta	si	P
13	nublado	alta	normal	no	P
14	lluvioso	media	alta	si	N

- Cada atributo/característica **A**, divide a los ejemplos en subconjuntos **E1**, **E2**, . . . , **Ev** de acuerdo a los **v** valores del atributo
- Cada subconjunto **Ei** tiene **pi** ejemplos positivos y **ni** ejemplos negativos, por lo que para cada rama necesitamos calcular su entropía

$$Entropia(i) = -\frac{p_i}{p_i + n_i} log_2 \frac{p_i}{p_i + n_i} - \frac{n_i}{p_i + n_i} log_2 \frac{p_i}{p_i + n_i}$$

Ejemplo

ID	Ambiente	Temp.	Humedad	Viento	Clase
1	soleado	alta	alta	no	N
2	soleado	alta	alta	si	N
3	nublado	alta	alta	no	P
4	lluvioso	media	alta	no	P
5	lluvioso	baja	normal	no	P
6	lluvioso	baja	normal	si	N
7	nublado	baja	normal	si	P
8	soleado	media	alta	no	N
9	soleado	baja	normal	no	P
10	lluvioso	media	normal	no	P
11	soleado	media	normal	si	P
12	nublado	media	alta	si	P
13	nublado	alta	normal	no	P
14	lluvioso	media	alta	si	N

- Ambiente:
  - soleado: p1 = 2, n1 = 3, Entropia(soleado) = 0.971
  - nublado: p2 = 4,n2 = 0, Entropia(nublado) = 0
  - lluvioso: p3 = 3, n3 = 2, Entropia(lluvioso) = 0.971

$$Ganancia(A) = 0.940 - (\frac{5}{14} * 0.971 + \frac{4}{14} * 0 + \frac{5}{14} * 0.971) = 0.940 - 0.694 = 0.246$$

Ejemplo

ID	Ambiente	Temp.	Humedad	Viento	Clase
1	soleado	alta	alta	no	N
2	soleado	alta	alta	si	N
3	nublado	alta	alta	no	P
4	lluvioso	media	alta	no	P
5	lluvioso	baja	normal	no	P
6	lluvioso	baja	normal	si	N
7	nublado	baja	normal	si	P
8	soleado	media	alta	no	N
9	soleado	baja	normal	no	P
10	lluvioso	media	normal	no	P
11	soleado	media	normal	si	P
12	nublado	media	alta	si	P
13	nublado	alta	normal	no	P
14	lluvioso	media	alta	si	N

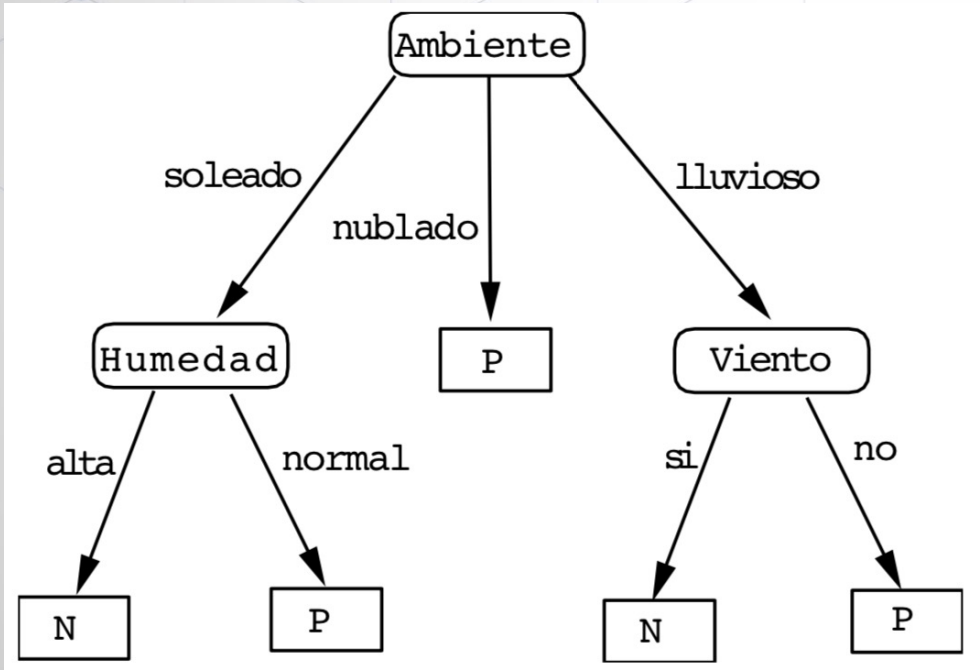
- Humedad:
  - alta: p1 = 3,n1 = 4, E(alta) = 0.985
  - normal: p2 = 6, n2 = 1, E(normal) = 0.592
  - Entropía (Humedad) = 0.789
- Viento:
  - no: p1 = 6, n1 = 2, E(no) = 0.811
  - si: p2 = 3,n2 = 3, E(p2,n2) = 1.0
  - Entropía(Viento) = 0.892
- Temperatura
  - alta: p1 = 2, n1= 2, E(alta) = 1.0
  - media: p2= 4, n2=2, E(media)= 0.918
  - baja: p3= 3, n3= 1, E(baja) = 0.811
  - Entropía(Temperatura) = 0.9111

Ejemplo - Ganacia de Información

ID	Ambiente	Temp.	Humedad	Viento	Clase
1	soleado	alta	alta	no	N
2	soleado	alta	alta	si	N
3	nublado	alta	alta	no	P
4	lluvioso	media	alta	no	P
5	lluvioso	baja	normal	no	P
6	lluvioso	baja	normal	si	N
7	nublado	baja	normal	si	P
8	soleado	media	alta	no	N
9	soleado	baja	normal	no	P
10	lluvioso	media	normal	no	P
11	soleado	media	normal	si	P
12	nublado	media	alta	si	P
13	nublado	alta	normal	no	P
14	lluvioso	media	alta	si	N

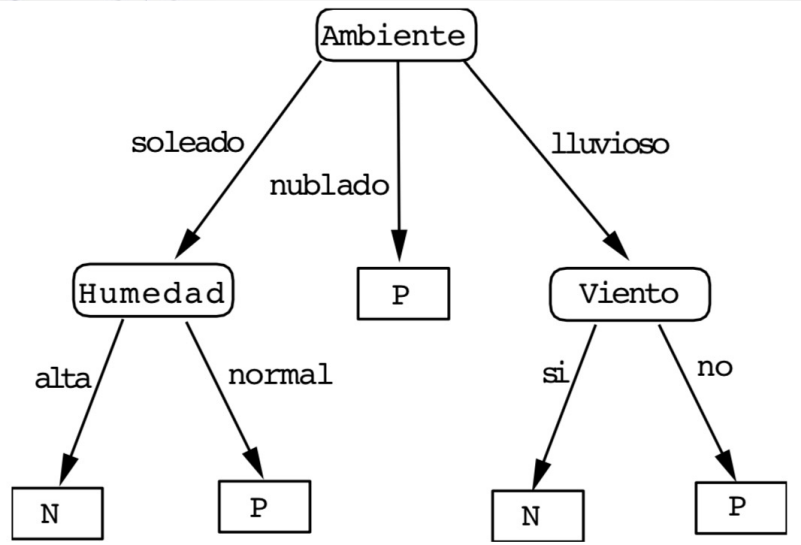
- Ganancia(Ambiente) = 0.246*
- Ganancia(Humedad) = 0.151
- Ganancia(Viento) = 0.048
- Ganancia(Temperatura) = 0.029

Ejemplo



ID	Ambiente	Temp.	Humedad	Viento	Clase
1	soleado	alta	alta	no	N
2	soleado	alta	alta	si	N
3	nublado	alta	alta	no	P
4	lluvioso	media	alta	no	P
5	lluvioso	baja	normal	no	P
6	lluvioso	baja	normal	si	N
7	nublado	baja	normal	si	P
8	soleado	media	alta	no	N
9	soleado	baja	normal	no	P
10	lluvioso	media	normal	no	P
11	soleado	media	normal	si	P
12	nublado	media	alta	si	P
13	nublado	alta	normal	no	P
14	lluvioso	media	alta	si	N

Ejemplo



- ¿Podemos jugar el sábado con **ambiente** soleado, **temperatura** alta, **humedad** alta y con **viento**?
- ¿Podemos jugar el domingo con **ambiente** lluvioso, **temperatura** normal, **humedad** alta y sin **viento**?

Código

CART

Classification and regression tree

CART

- Índice Gini

$$Gini(t) = 1 - \sum_{j=1}^{m} (p(j|t))^2$$

- $p(j|t)$ es la frecuencia relativa de la clase j en t
- Se debe calcular el índice en cada rama del atributo tomando en cuenta su proporción de ejemplos
- Dividiendo en v ramas

$$Gini_A = \sum_{i=1}^{v} \frac{n_i}{n} Gini(v)$$

n_i son los ejemplos de la rama y n los del nodo

CART

Ambiente	P	N	Total por rama
Soleado	2	3	5
Nublado	4	0	4
Lluvioso	3	2	5

$$Gini(t) = 1 - \sum_{j=1}^{m} (p(j|t))^2$$

$$Gini_A = \sum_{i=1}^{v} \frac{n_i}{n} Gini(v)$$

- Gini(Ambiente=Soleado) = $1 - (2/5)^2 - (3/5)^2 = 1 - 0.16 - 0.36 = 0.48$
- Gini(Ambiente=Nublado) = $1 - (4/4)^2 - (0/4)^2 = 0$
- Gini(Ambiente=Lluvioso) = $1 - (3/5)^2 - (2/5)^2 = 1 - 0.36 - 0.16 = 0.48$
- Gini para ambiente
- Gini(Ambiente) = (5/14) x 0.48 + (4/14) x 0 + (5/14) x 0.48 = 0.171 + 0 + 0.171 = 0.342

CART

Atributo	Índice Gini
Ambiente	0.342
Temperatura	0.439
Humedad	0.367
Viento	0.428

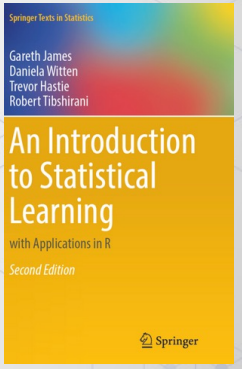
ID	Ambiente	Temp.	Humedad	Viento	Clase
1	soleado	alta	alta	no	N
2	soleado	alta	alta	si	N
3	nublado	alta	alta	no	P
4	lluvioso	media	alta	no	P
5	lluvioso	baja	normal	no	P
6	lluvioso	baja	normal	si	N
7	nublado	baja	normal	si	P
8	soleado	media	alta	no	N
9	soleado	baja	normal	no	P
10	lluvioso	media	normal	no	P
11	soleado	media	normal	si	P
12	nublado	media	alta	si	P
13	nublado	alta	normal	no	P
14	lluvioso	media	alta	si	N

- Ambiente -> Menor

Código

Ventajas

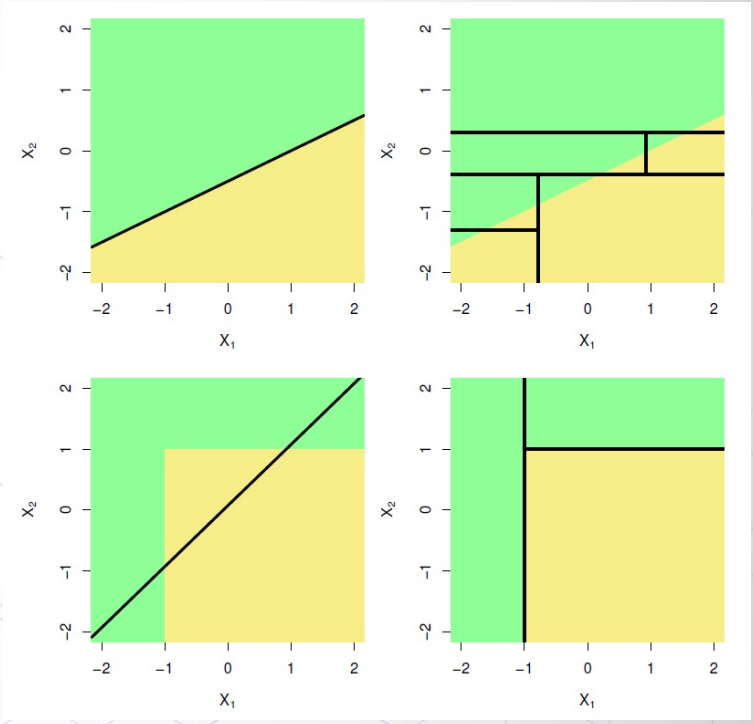
- Los árboles son muy fáciles de explicar a las personas
  - De hecho, ¡son incluso más fáciles de explicar que la regresión lineal!
- Algunas personas creen que los árboles de decisión reflejan más de cerca la toma de decisiones humanas
- Los árboles pueden mostrarse gráficamente y se interpretan fácilmente incluso por un no experto (especialmente si son pequeños)
- Los árboles pueden manejar fácilmente predictores cualitativos sin necesidad de crear variables ficticias



Desventajas

- Los árboles pueden sobreajustar y no generalizar bien
- No son robustos: un pequeño cambio en los datos puede provocar un gran cambio en el árbol final estimado

Árboles vs modelos lineales



Código

EngineeringX

Founded by the Royal Academy of Engineering
and Lloyd's Register Foundation

GRACIAS

https://hubiq.mx/

HUBIQRO HUBIQ HUBIQRO