ID3 ALGORITHM

ID3

- El algoritmo ID3 (Iterative Dichotomiser 3) es un algoritmo de aprendizaje automático utilizado para construir árboles de decisión, desarrollado por Ross Quinlan.
- Este algoritmo se basa en la estrategia "divide y vencerás" para clasificar objetos o situaciones, tomando como entrada un conjunto de ejemplos caracterizados por un conjunto de atributos, donde uno de ellos es el objetivo a clasificar, generalmente de tipo binario (sí/no, positivo/negativo).
- El proceso de construcción del árbol es recursivo, comenzando desde un nodo raíz y dividiendo los datos en subconjuntos según los valores de los atributos, con el objetivo de crear ramas cada vez más homogéneas en cuanto a la clase objetivo

Ejemplo

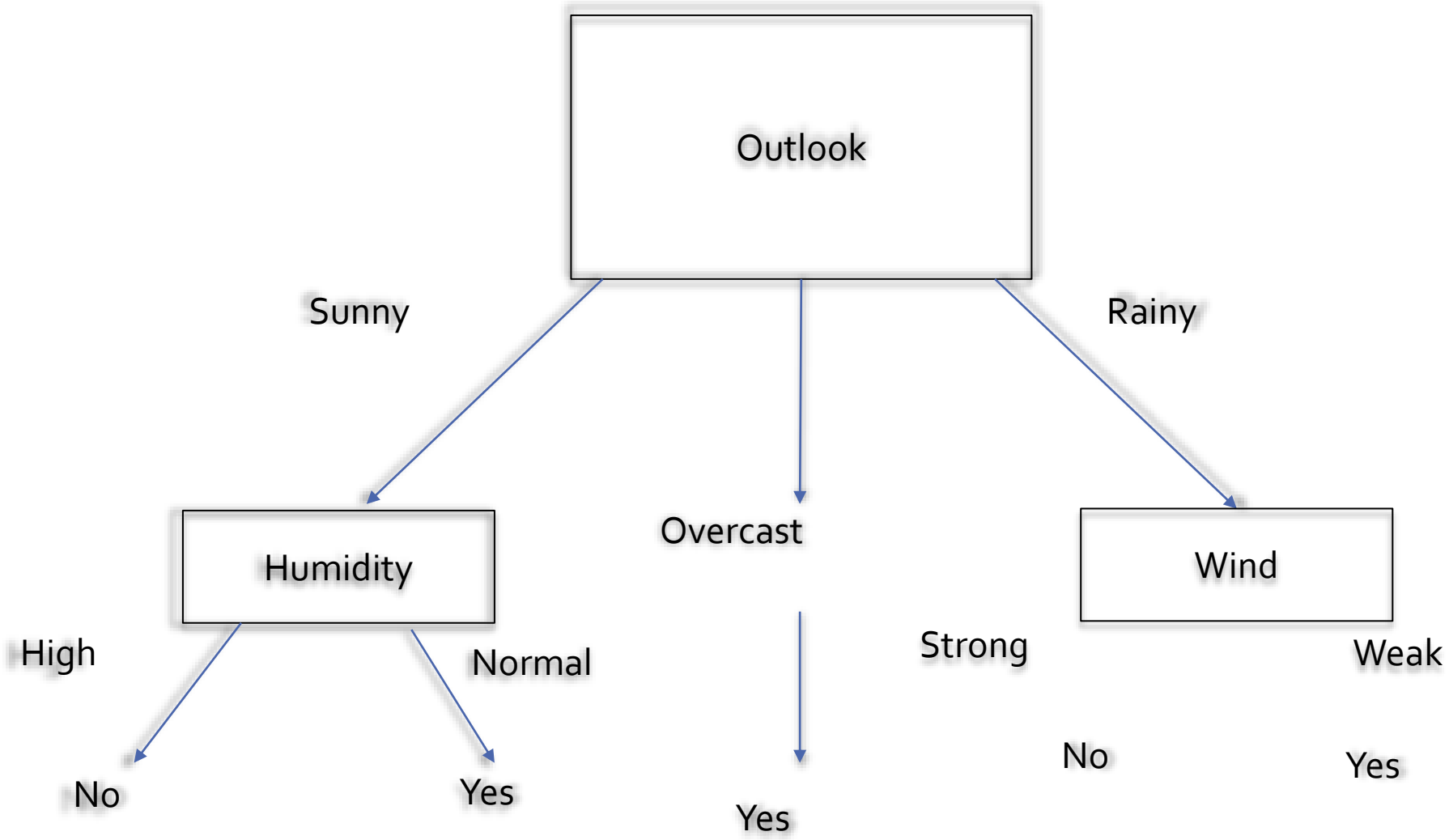
14 observaciones sobre el juego de tenis donde se decide si se juega o no dadas las condiciones meteorológicas de sol, temperatura, humedad, viento, en general tenemos los campos:

- OutLook
- Temperatura
- Humedad
- Viento
- Decisión (se juega o no: yes, no)

Datos

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Debemos obtener el siguiente árbol



Abstract

- ID3 builds a decision tree from a set of examples.
- Using this decision tree, future samples are classified.
- The example has several attributes and belongs to a class.
- The leaf nodes of the decision tree contain the class name whereas a non-leaf node is a decision node.
- The decision node is an attribute test with each branch being a possible value of the attribute.
- ID3 uses information gain to help it decide which attribute goes into a decision node.

Entropia

- La entropia es una medida de la aleatoriedad de la información.
- Si el ejemplo es completamente homogeneo la entropia es cero y Si la muestra se divide equitativamente, entonces tiene una entropía de uno.
- La entropia puede be calculated as:

$$Entropy(S) = \sum - p(I)\ log_2 p(I)$$

donde p(I) es la proporción de S que pertenece a la clase I.
Tenga en cuenta que S no es un atributo sino el conjunto de muestras completo.

Ganancia and Information gain

- La ganancia de información se puede calcular como :

$$Gain(S, A) = Entropy(S) - \Sigma ((|S_v| / |S|) * Entropy(S_v))$$

Donde:
Σ is each value v of all possible values of attribute A
S_v = subset of S for which attribute A has value v
$|S_v|$ = number of elements in S_v
|S| = number of elements in S

Queremos el árbol de decision para saber si se juega tenis usando datos de 14 dias

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Entropia

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

$$Entropy(S) = -p(Yes) \cdot log_2 p(Yes) - p(No) \cdot log_2 p(No)$$

$$Entropy(S) = -(9/14) \cdot \log_2(9/14) - (5/14) \cdot \log_2(5/14)$$

$$= \mathbf{0.940}$$

Obteniendo ganancia

A continuación por cada campo (Outlook, temp, humidity, wind) calculamos la ganancia.

Empezamos por ejemplo con el campo wind.

Factor Wind en la decision

- Observe el conjunto S de 14 examples en el cual uno de los atributos es velocidad del viento.
- Los valores de viento pueden ser: *Weak* or *Strong*.
- La clasificación de esos 14 ejemplos son 9 YES y 5 NO.
- Para el atributo Wind, hay 8 occurrencias de Wind = Weak y 6 occurrencias de Wind = Strong.
- Para Wind = Weak, 6 de los ejemplos son YES y 2 son NO.
- para Wind = Strong, 3 son YES y 3 son NO. Por lo tanto:

$$Gain(S, A) = Entropy(S) - \Sigma ((|S_v| / |S|) * Entropy(S_v))$$

$$Gain(S,Wind)=Entropy(S) - (8/14)*Entropy(S_{weak})-(6/14)*Entropy(S_{strong})$$
$$= 0.940 - (8/14)*0.811 - (6/14)*1.00$$
$$= 0.048$$

$$Entropy(S_{weak}) = - (6/8)*\log2(6/8) - (2/8)*\log2(2/8) = 0.811$$
$$Entropy(S_{strong}) = - (3/6)*\log2(3/6) - (3/6)*\log2(3/6) = 1.00$$

Para cada atributo, se calcula la ganacia y se toma la mas alta.

Weak wind factor

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
13	Overcast	Hot	Normal	Weak	Yes

A continuación dividimos la clase wind en subclases por ejemplo wind la dividimos en weak y strong, para las que calcularemos tambien entropia y ganacia

Hay 8 instancias para weak wind. Con 2 no en el campo Decision y 6 items son **yes**.

Calculamos la entropia para el conjunto Weak

$$\text{Entropy}(S,\text{Wind=Weak}) = -(2/8) \cdot \log_2(2/8) - (6/8) \cdot \log_2(6/8)$$

$$= 0.811$$

Strong wind factor

Day	Outlook	Temp.	Humidity	Wind	Decision
2	Sunny	Hot	High	Strong	No
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
14	Rain	Mild	High	Strong	No

- Hy 6 instances para strong wind. Con 3 yes y 3 no en el campo decision.

$$\text{Entropy}(S,\text{Wind=Strong}) = -(3/6)\cdot\log_2(3/6) - (3/6)\cdot\log_2(3/6)$$

$$= \mathbf{1}$$

Hacemos esto con todos los campos

Aplicando un cálculo similar a las otras columnas, obtenemos:

Para todos los factores de decision

- Gain (S,wind)) = 0.048
- Gain(S, Outlook) = 0.246
- Gain(S, Temperature) = 0.029
- Gain(S, Humidity) = 0.151

Algoritmo

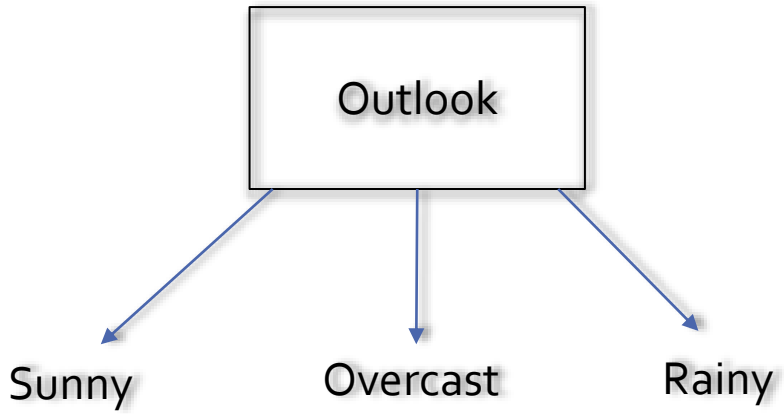
Dado un conjunto de ejemplos S categorizados en categorias c_i entonces:

1. Elijaa el nodo raiz como el atributo con la ganacia mas alta con relacion a S.
2. Para cada valor v que A pueda tomar dibuje una rama desde el nodo.
3. Para cada rama de A correspondiente al valor v, calcule S_v. entonces:
   3.1 si S_v es vacio, seleccione la categoría $c_{default}$ que contiene la mayor cantidad de ejemplos de S, y ponga esto como la categoría de nodo hoja que finaliza esa rama.
   3.2 Si S_v contiene solo ejemplos de una sola categoría c, entonces ponga c como la categoría de nodo hoja que finaliza esa rama.
   3.3 De lo contrario, elimine A del conjunto de atributos que se pueden colocar en los nodos.. Luego, coloque un nuevo nodo en el árbol de decisión, donde el nuevo atributo que se está probando en el nodo es el que tiene el puntaje más alto en ganancia de información en relación con S_v (note: no relativa a S). Este nuevo nodo inicia el ciclo nuevamente (desde 2), con S reemplazado por Sv en los cálculos y el árbol se construye iterativamente de esta manera. El algoritmo finaliza cuando se han agotado todos los atributos o el árbol de decisión clasifica perfectamente los ejemplos..

Aplicando el algoritmo

Paso 1: Como outlook tiene mayor ganacia lo ponemos como raiz.

Paso 2: como Outlook tiene los valores (suny, overcast, rainy), creamos esas ramas.

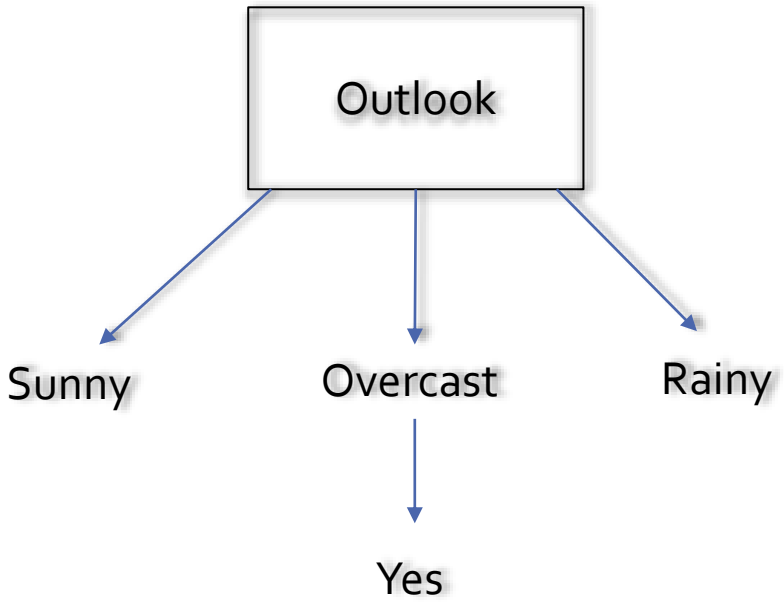


Overcast outlook on decision

Continuando con el algoritmo observe que para el valor overcast en la rama Outlook, La decisión de jugyes siempre será sí incluso aunque el panorama esté nublado.

Day	Outlook	Temp.	Humidity	Wind	Decision
3	Overcast	Hot	High	Weak	Yes
7	Overcast	Cool	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes

Aplicando 3.2, de la rama overcast
Ponemos una hoja yes y ahí termina esa rama



Quedan 2 ramas por anlizar

- Sunny
- Rainy

Analicemos Sunny

Sunny outlook on decision

Tenemos 5 instancias para sunny outlook. Decision tiene la probabilidad

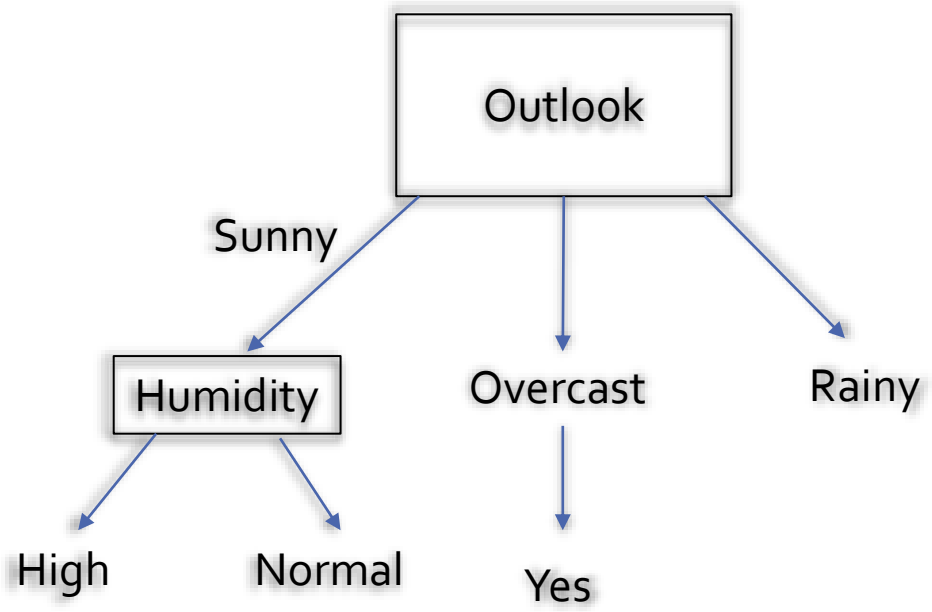
3/5 porcentaje no,

2/5 porcentaje yes

- Gain(Outlook=Sunny|Temperature) = **0.570**
- Gain(Outlook=Sunny|Humidity) = **0.970**
- Gain(Outlook=Sunny|Wind) = **0.019**

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Como Humidity tiene la mayor ganancia será el inicio de la rama, se agregan 2 ramas (Higth, Normal).

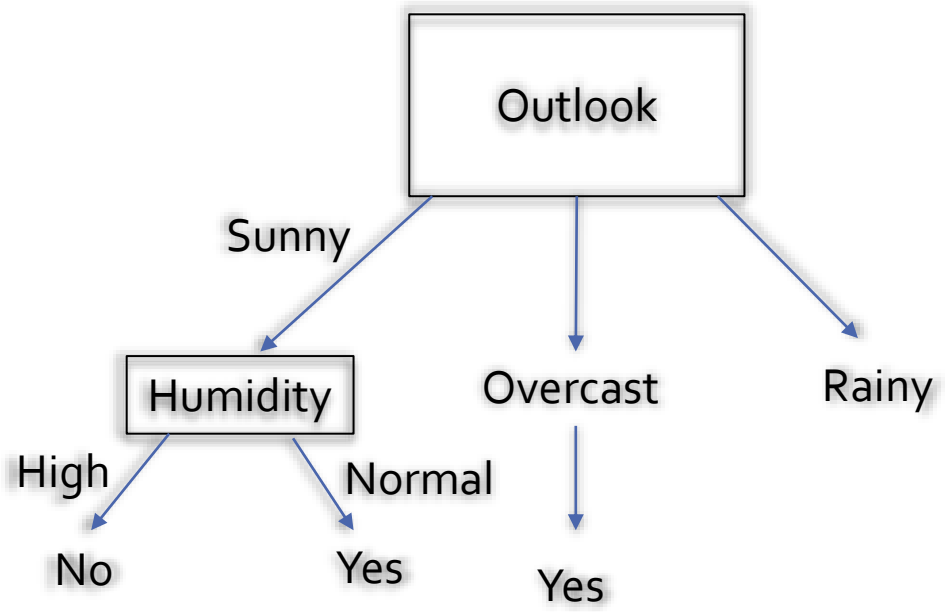


Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No

Day	Outlook	Temp.	Humidity	Wind	Decision
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

- Decision will always be **no** when humidity is **high**.
- Decision will always be **yes** when humidity is **normal**.

Como Las ramas high y normal solo tiene un valor cada una

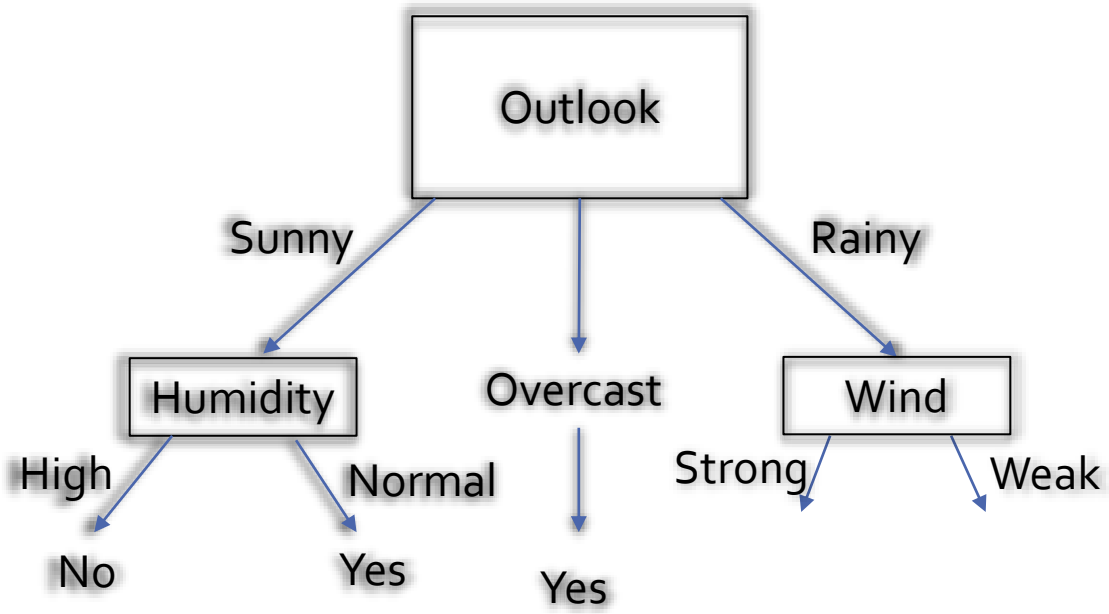

Outlook
Sunny
Humidity
Overcast
Rainy
High
Normal
No
Yes
Yes

Rain outlook on decision

Information gain for Rain outlook are:

- Gain(Outlook=Rain | Temperature) = 0.02
- Gain(Outlook=Rain | Humidity) = 0.02
- Gain(Outlook=Rain | Wind) = 0.971

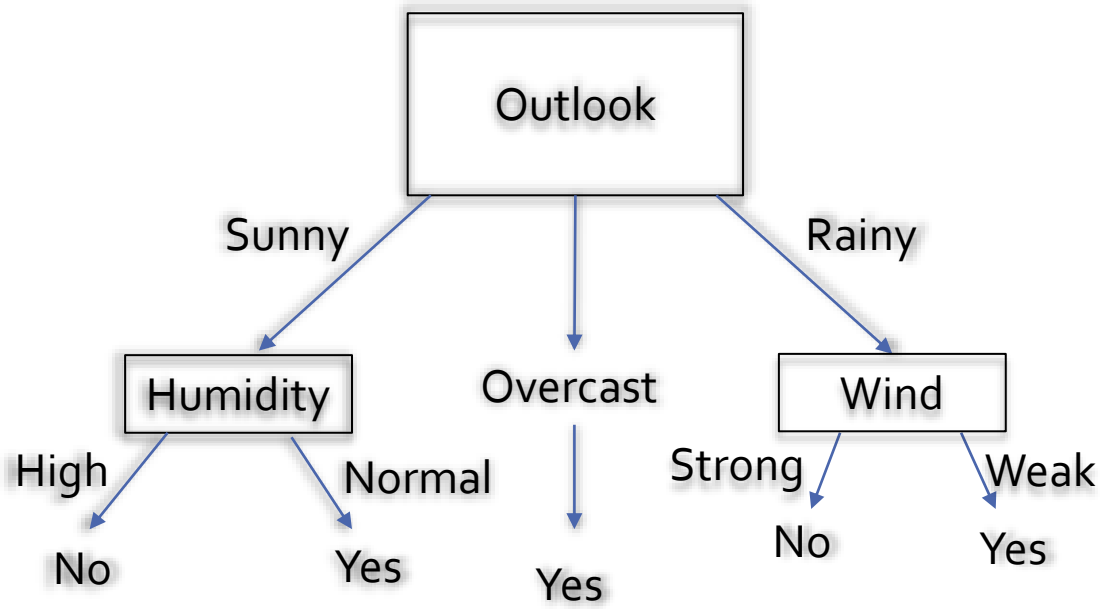
Day	Outlook	Temp.	Humidity	Wind	Decision
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No


Outlook
Sunny
Rainy
Humidity
Overcast
Wind
High
Normal
Strong
Weak
No
Yes
Yes

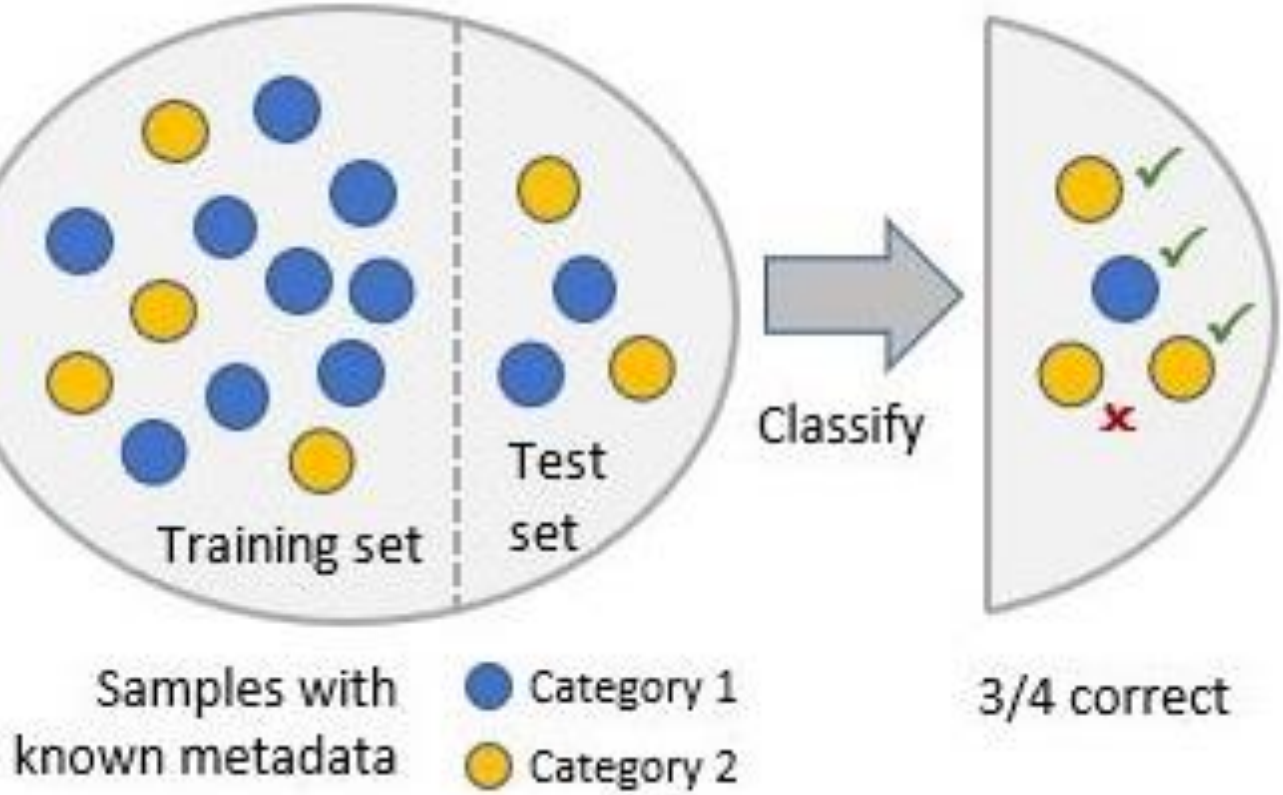
Day	Outlook	Temp.	Humidity	Wind	Decision
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes

Day	Outlook	Temp.	Humidity	Wind	Decision
6	Rain	Cool	Normal	Strong	No
14	Rain	Mild	High	Strong	No

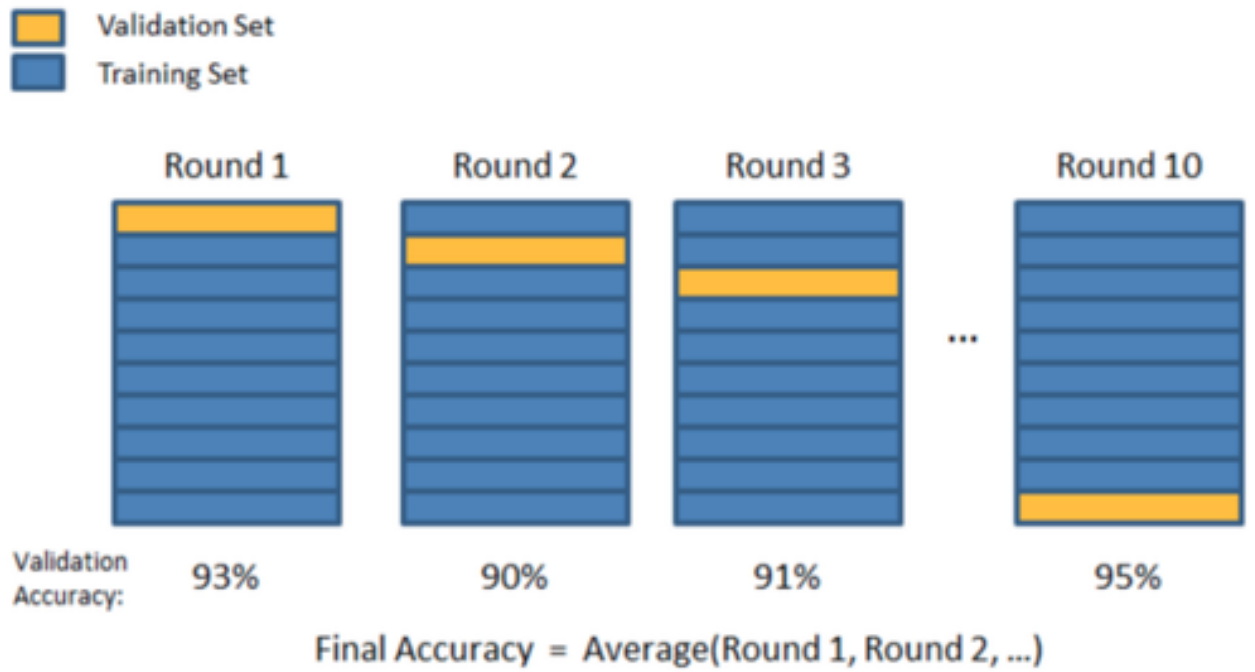
- Decision will always be **yes** if wind were **weak** and outlook were **rain**.
- Decision will always be **no** if wind were **strong** and outlook were **rain**.


Outlook
Sunny
Rainy
Humidity
Overcast
Wind
High
Normal
Strong
Weak
No
Yes
Yes
No
Yes

Crossvalidation



10 fold crossvalidation



Actividad

Crear un árbol de decisión usando el algoritmo ID3 para los datos de la flor iris:

https://archive.ics.uci.edu/dataset/53/iris