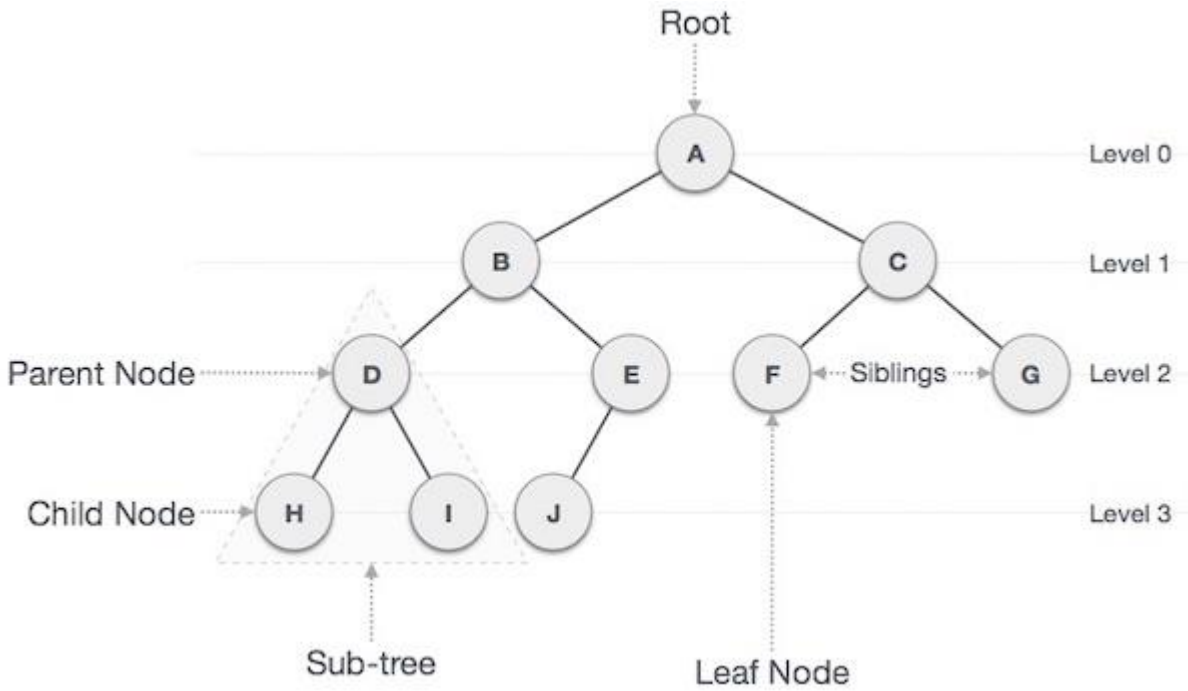
UNIDAD 6: ÁRBOLES DE DECISIÓN

Sesión 16-03-2022

Árbol: *estructura de datos abstracto que tiene una* ***jerarquía*** *y emula un árbol biológico.*

Se representa como un conjunto de ***nodos*** *unidos por* ***aristas****.*

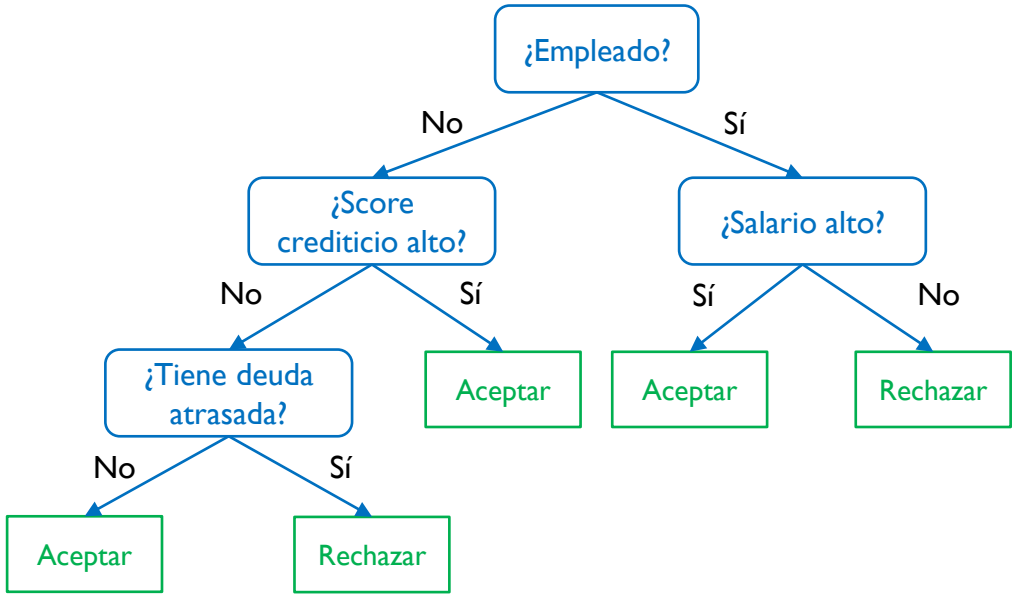


https://www.tutorialspoint.com/data_structures_algorithms/tree_data_structure.htm

¿QUÉ ES UN ÁRBOL?

Clasificador construido siguiendo el modelo de un árbol de decisión y presenta dos componentes: ***nodos de decisión*** *y* ***nodos hoja****.*

Ejemplo: *árbol de decisión para la aceptación de una solicitud de crédito.*



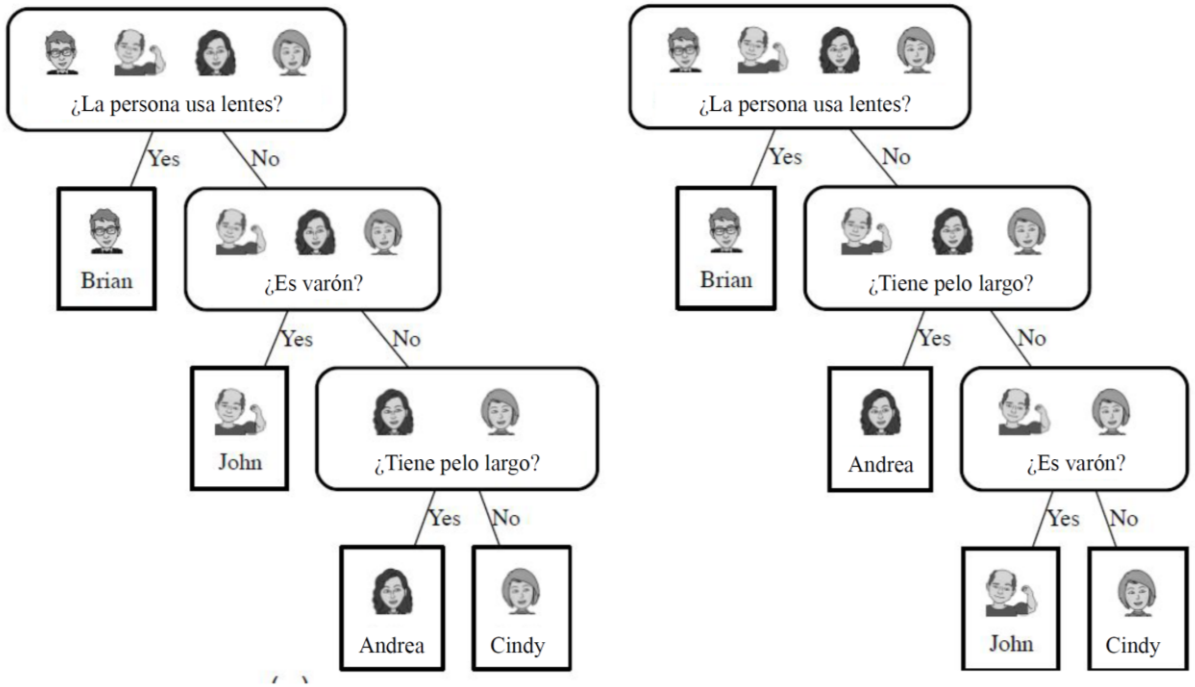
¿QUÉ ES UN ÁRBOL DE DECISIÓN?

¿CÓMO CONSTRUIMOS UN ÁRBOL? - INTUICIÓN

Ejercicio: *todos escojan un personaje. ¿Cuántas preguntas necesitan para adivinar el personaje de su compañero?*

Nombre	¿Lentes?	¿Pelo largo?	¿Hombre?
Brian	Sí	No	Sí
John	No	No	Sí
Andrea	No	Sí	No
Cindy	No	Sí	No

Caso 1: *Empezando con la pregunta ¿usa lentes?*

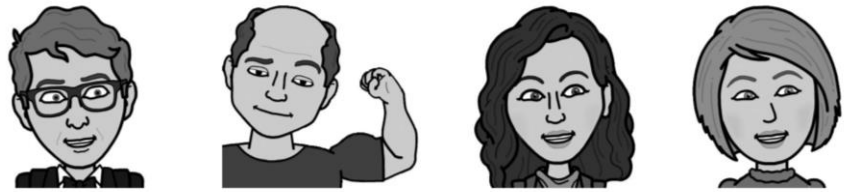


Beltrán, C. (2020). Material de clase "Aprendizaje de Máquina" del semestre 2020-2. PUCP.

Material para el curso "Fundamentos de Aprendizaje Automático" por Luis Muroya.

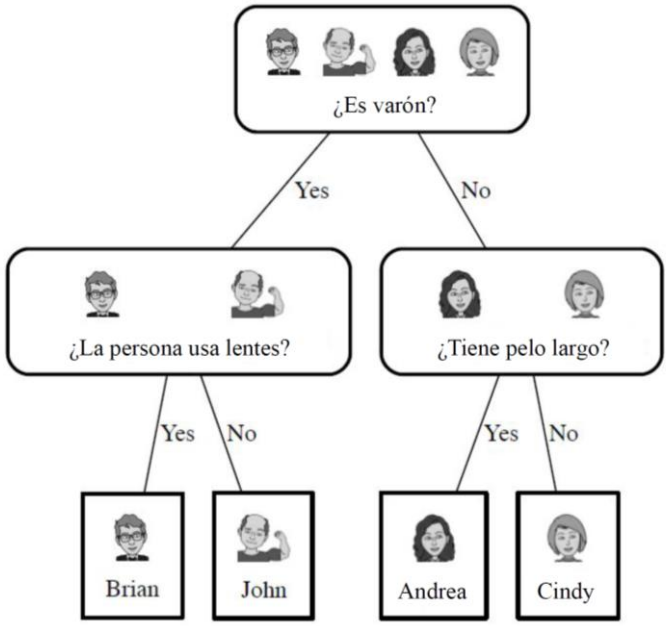
¿CÓMO CONSTRUIMOS UN ÁRBOL? - INTUICIÓN

Ejercicio: *todos escojan un personaje. ¿Cuántas preguntas necesitan para adivinar el personaje de su compañero?*



Nombre	¿Lentes?	¿Pelo largo?	¿Hombre?
Brian	Sí	No	Sí
John	No	No	Sí
Andrea	No	Sí	No
Cindy	No	Sí	No

Caso 2: *Empezando con la pregunta ¿es varón?*



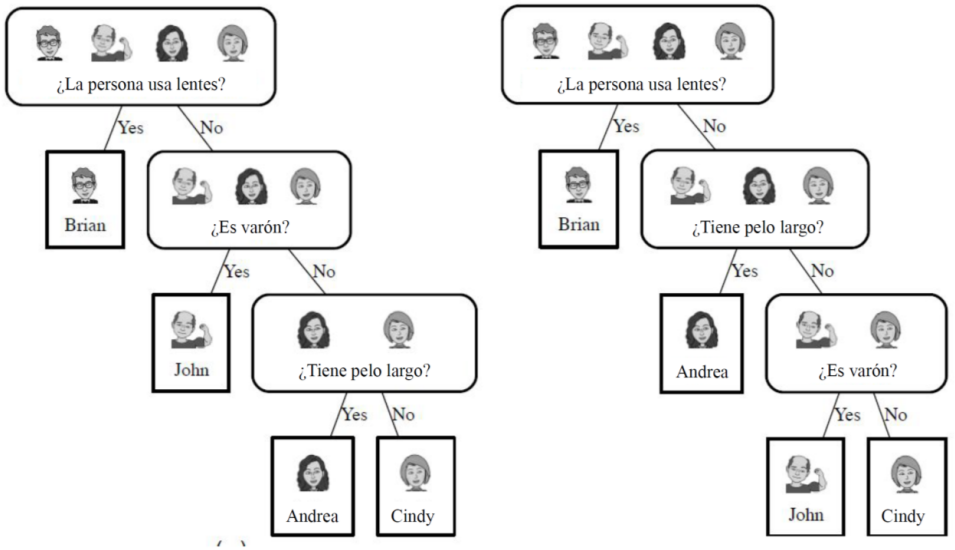
Beltrán, C. (2020). Material de clase "Aprendizaje de Máquina" del semestre 2020-2. PUCP.

Material para el curso "Fundamentos de Aprendizaje Automático" por Luis Muroya.

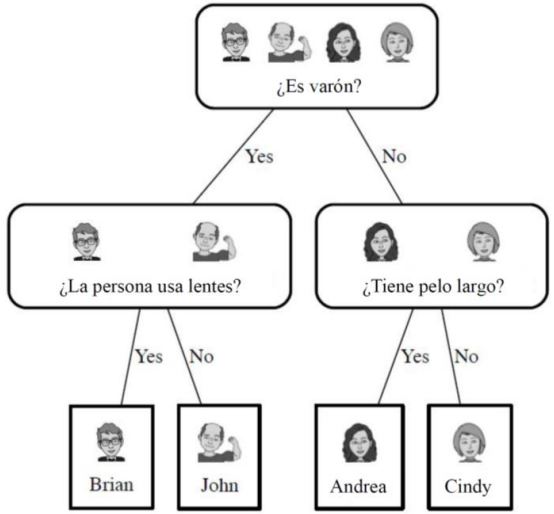
¿CÓMO CONSTRUIMOS UN ÁRBOL? - INTUICIÓN

Ejercicio: *comparemos los árboles resultantes. ¿Qué árbol resulta más eficiente?*

Caso 1: *Empezando con la pregunta ¿usa lentes?*



Caso 2: *Empezando con la pregunta ¿es varón?*



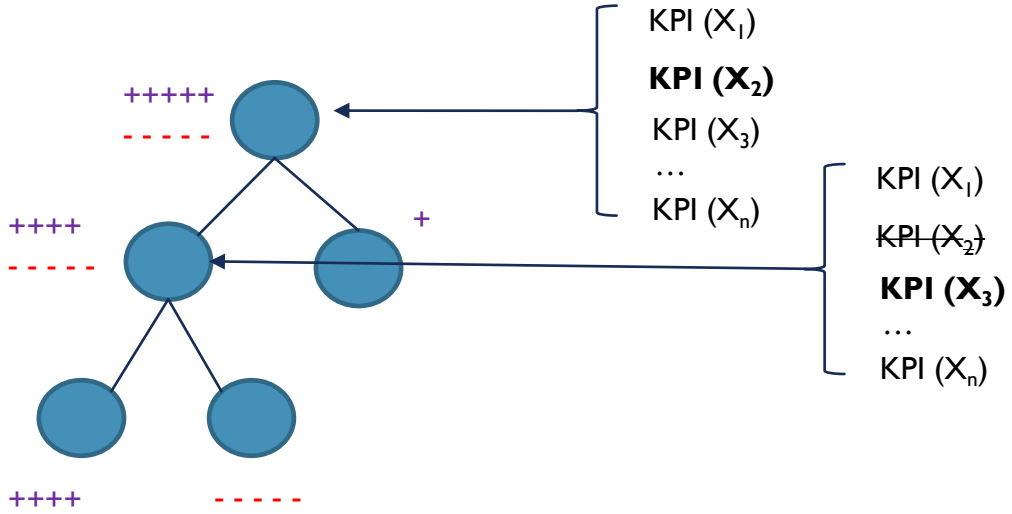
La idea para construir un árbol consiste en identificar las características más informativas para responder a una pregunta considerando cómo es dividido el dominio luego de c/respuesta y la semejanza de cada respuesta.

Beltrán, C. (2020). Material de clase "Aprendizaje de Máquina" del semestre 2020-2. PUCP.

Material para el curso "Fundamentos de Aprendizaje Automático" por Luis Muroya.

EL ALGORITMO ID-3

El algoritmo ID3 (Interactive Dichotomizer Version 3) prueba cada característica para determinar cuál es la más útil como raíz para un árbol. Para ello, se compara el KPI (se verá luego) obtenido por cada variable y se escoge aquella que lo maximice.



+ = Clase 1

- = Clase 2

Se selecciona como raíz la variable que maximiza el KPI.

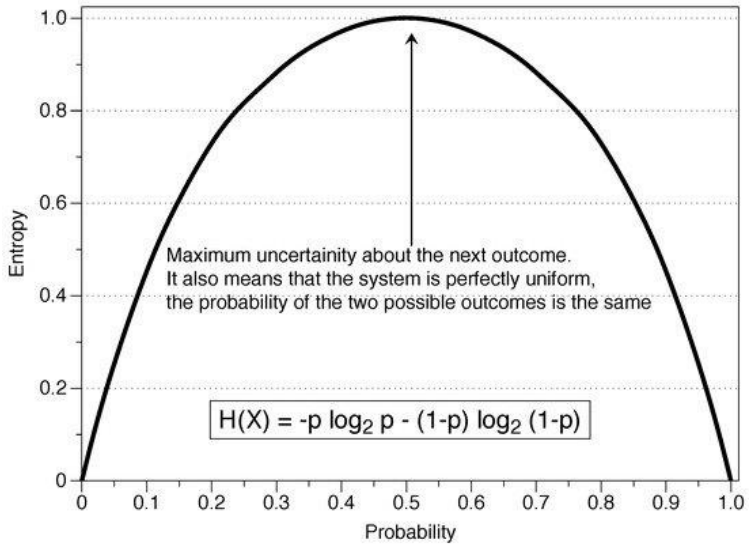
Una vez seleccionada la raíz, se repite el algoritmo para cada hijo (que sea nodo de decisión), considerando que éste es la raíz del árbol formado por toda su descendencia.

El algoritmo se repite hasta que solo queden subconjuntos puros (nodos hoja).

ENTROPÍA Y GANANCIA DE INFORMACIÓN

Entropía

La entropía mide el grado de desorden o incerteza de un sistema.



Delgado-Bonal, A.; Marshak, A. (2019) Approximate Entropy and Sample Entropy: A Comprehensive Tutorial. https://www.mdpi.com/1099-4300/21/6/541/htm

$$Entropia\ (S) = -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$

Ganancia de Información

La ganancia es una medida de cuánto podemos reducir la incerteza.

$$\boldsymbol{Ganancia(S, A)} = E(S) - E(S|A)$$

E(S): entropía del sistema (antes de partición)
E(S|A): entropía condicional del sistema dada la partición con la variable **A**.

$$\boldsymbol{E(S|A)} = \sum_{v \in Val(A)} \left(\frac{|S_v|}{|S|} \right) E(S_v)$$

Val(A): todos los posibles valores del atributo **A**.

S_v: subconjunto de S en el cual el atributo A tiene el valor v,

*En el algoritmo ID3 se escoge la variable V que **maximice** la ganancia de información.*

EJEMPLO ID3: PLAY TENNIS

Se tiene un dataset con información meteorológica de días en que se tenían programadas partidas de tenis. El objetivo es predecir si dadas las condiciones, se jugará o no una partida.

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

9

5

1. Calculamos la entropía del sistema.

$$E\ (S) = -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$

$$E\ (S) = -\left(\frac{9}{14}\right) \log_2 \left(\frac{9}{14}\right) - \left(\frac{5}{14}\right) \log_2 \left(\frac{5}{14}\right) = 0.94$$

2. Seleccionamos una variable y calculamos las entropías condicionales. Por ejemplo, humedad.

Valor: humidity	+	-	T	E(S)
High	3	4	7	-(3/7)$\log_2$(3/7) – (4/7)$\log_2$(4/7)=0.985
Normal	6	1	7	-(6/7) $\log_2$(6/7) – (1/7)$\log_2$(1/7)=0.592

$$E\ (S|A) = \left(\frac{7}{14}\right) 0.985 + \left(\frac{7}{14}\right) 0.592 = 0.79$$

3. Calculamos la ganancia para la variable en estudio.

$$G(S, A) = E(S) - E(S|A) = 0.94 - 0.79 = 0.15$$

Beltrán, C. (2020). Material de clase "Aprendizaje de Máquina" del semestre 2020-2. PUCP.

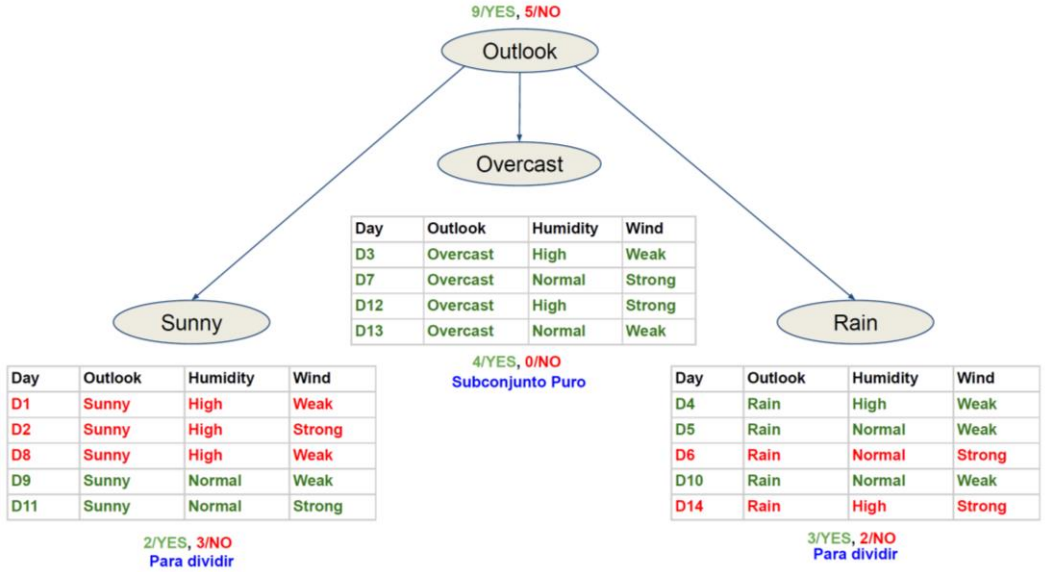
Material para el curso "Fundamentos de Aprendizaje Automático" por Luis Muroya.

EJEMPLO ID3: PLAY TENNIS

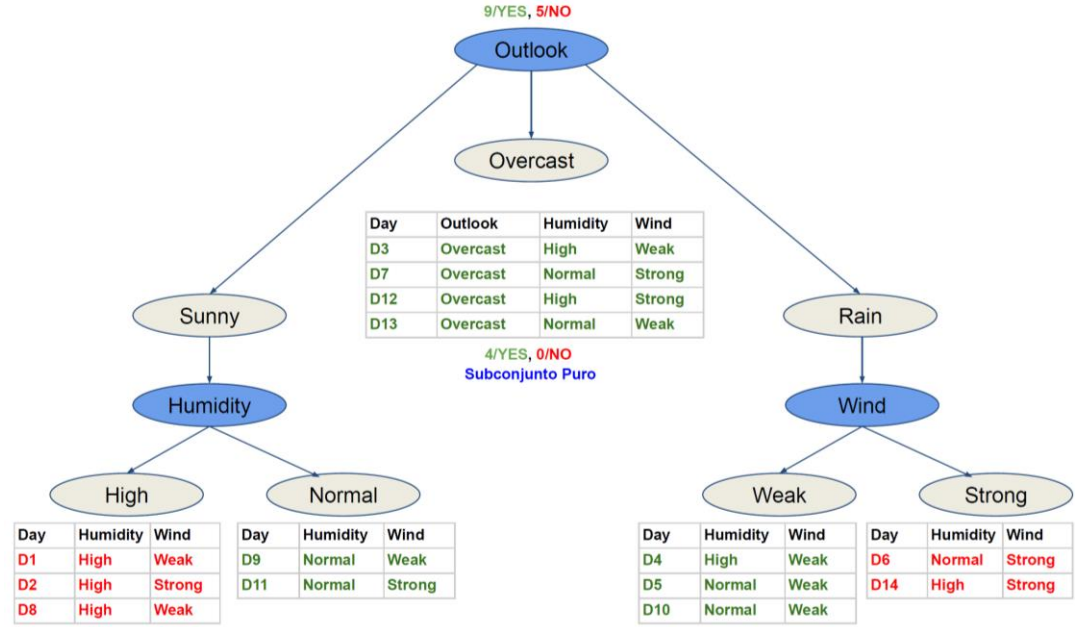
4. Repetimos los pasos 2 y 3 para cada variable.

Variable	Ganancia
Outlook	**0.25**
Humidity	0.15
Wind	0.05
Temperature	0.03

Elegimos ***Outlook*** *como variable raíz*



5. Como el árbol resultante aun tiene subconjuntos impuros, se repite el proceso. En estos casos, ya no se toma el dataset completo, sino aquellos registros que cumplen con la condición de la rama. Por ejemplo, para la rama de más a la izquierda se toman los 5 registros en los que Outlook="Sunny". A continuación, el árbol final resultante.

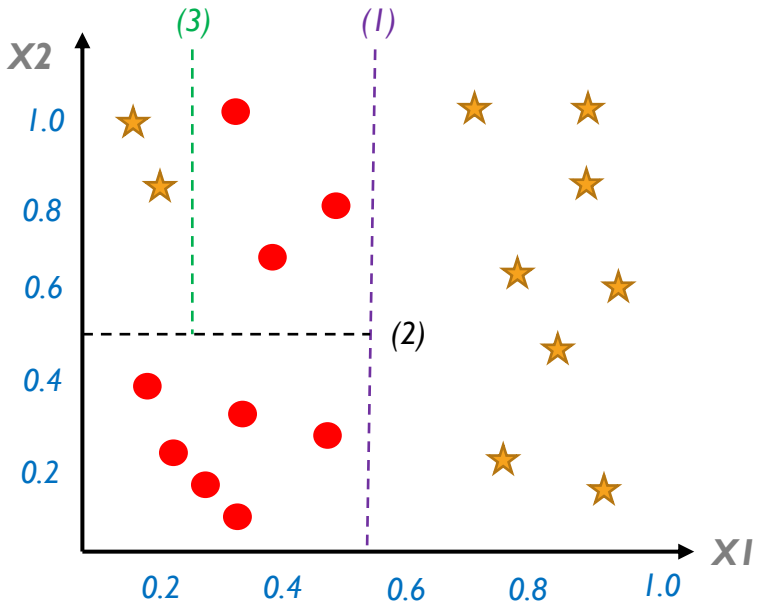


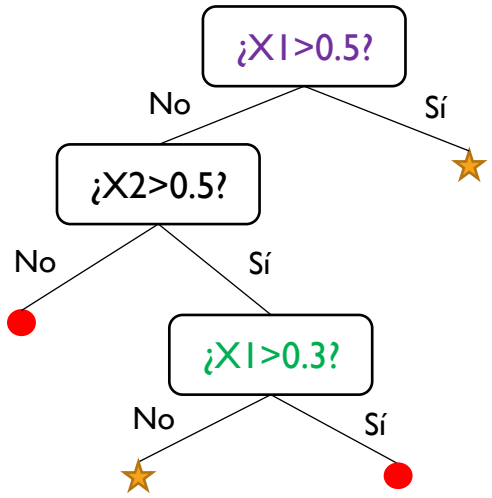
A diferencia del algoritmo ID3, el algoritmo CART genera un árbol binario de decisión, que es más eficiente computacionalmente.

Este algoritmo puede ser empleado tanto para problemas de clasificación como de regresión.

A diferencia del algoritmo ID3, este requiere que todas las variables sean numéricas.

En cada iteración, se busca la variable predictiva y el punto de corte adecuado para reducir la impureza, medida con una métrica llamada Índice de Impureza de Gini. El algoritmo se repite hasta que todos los nodos estén limpios de impurezas.



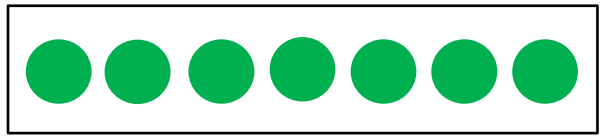


ALGORITMO CART

El índice de Gini indica el grado de ***impureza*** *de un conjunto: un conjunto puro tendrá un índice 0; mientras un conjunto con todos sus elementos diferentes tendrá un índice de 1.*

La fórmula del Índice de Impureza de Gini, cuando se tienen J clases, es la siguiente:

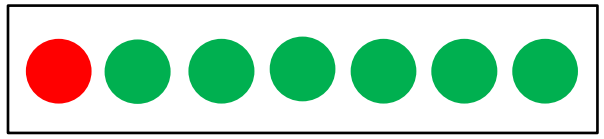
$$\boldsymbol{IG(p)} = 1 - \sum_{i=1}^{J} p_i^2$$



$p = \frac{0}{7} = 0.00$

$p = \frac{7}{7} = 1.00$

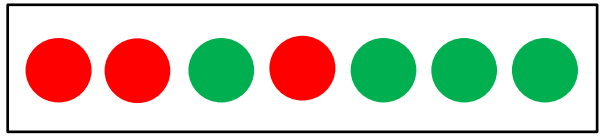
$$\boldsymbol{IG(p)} = 1 - (0^2 + 1^2) = \mathbf{0}$$



$p = \frac{1}{7} = 0.14$

$p = \frac{6}{7} = 0.86$

$$\boldsymbol{IG(p)} = 1 - (0.14^2 + 0.86^2) = \mathbf{0.24}$$



$p = \frac{3}{7} = 0.43$

$p = \frac{4}{7} = 0.57$

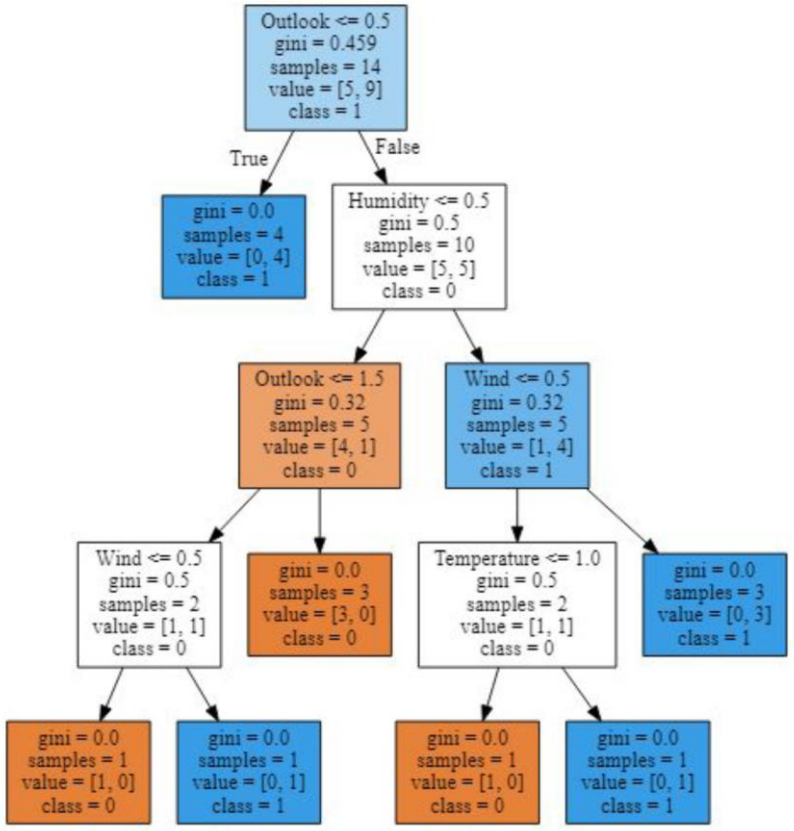
$$\boldsymbol{IG(p)} = 1 - (0.43^2 + 0.57^2) = \mathbf{0.49}$$

ÍNDICE DE IMPUREZA DE GINI

EJEMPLO: PLAY TENNIS

Se puede aplicar el algoritmo CART sobre el conjunto de datos que vimos en ID3.

Day	Outlook	Temperature	Humidity	Wind	Play
D1	2	1	0	1	No
D2	2	1	0	0	No
D3	0	1	0	1	Yes
D4	1	2	0	1	Yes
D5	1	0	1	1	Yes
D6	1	0	1	0	No
D7	0	0	1	0	Yes
D8	2	2	0	1	No
D9	2	0	1	1	Yes
D10	1	2	1	1	Yes
D11	2	2	1	0	Yes
D12	0	2	0	0	Yes
D13	0	1	1	1	Yes
D14	1	2	0	0	No

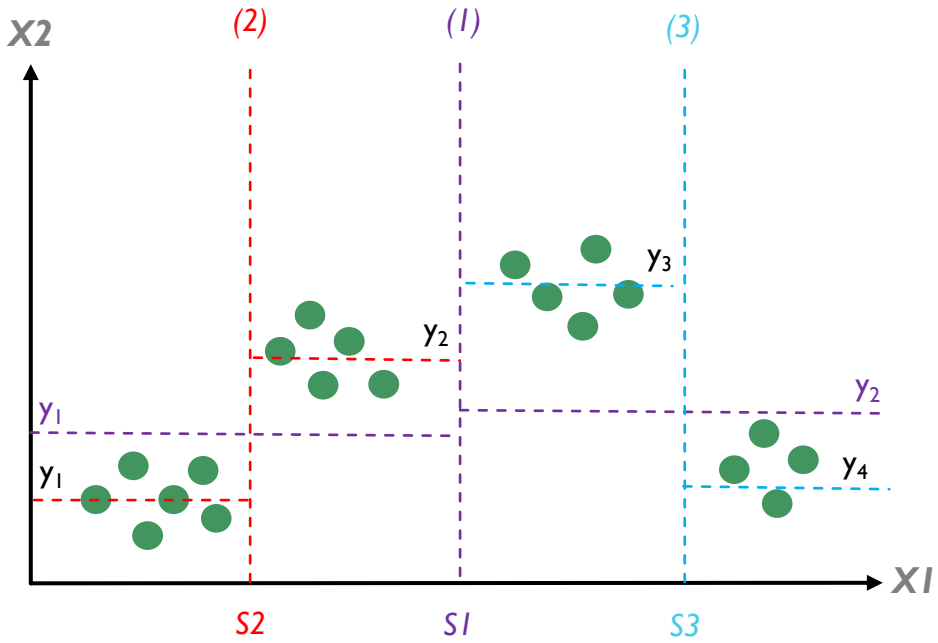


Se puede apreciar que las variables categóricas fueron codificadas usando un encoder por etiquetas (rev. Und. 2).

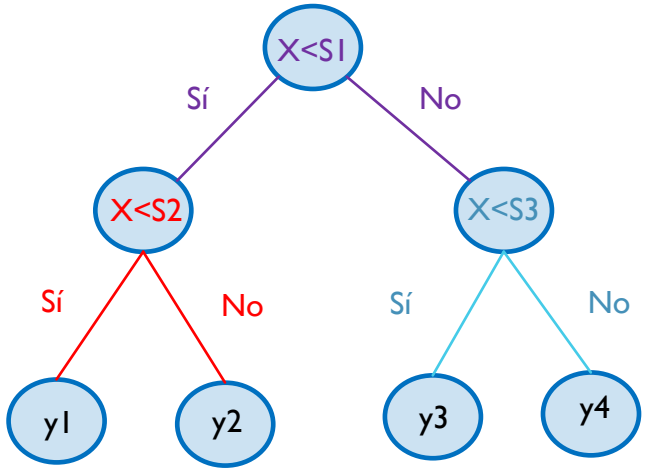
Beltrán, C. (2020). Material de clase "Aprendizaje de Máquina" del semestre 2020-2. PUCP.

Material para el curso "Fundamentos de Aprendizaje Automático" por Luis Muroya.

ÁRBOLES DE REGRESIÓN



Nota: las letras en color NEGRO representan los valores finales que serán predichos por el árbol resultante.



Al igual que en el caso de un árbol de clasificación, se evalúan todas las variables disponibles. Por cada una, se generan puntos de corte y se evalúa una métrica en cada uno. Finalmente, se calcula el valor de la métrica total de la variable y se escoge aquella que la minimice.

Para un árbol de regresión, dicha métrica es el error cuadrático medio:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

El árbol seguirá creciendo hasta que se prediga c/u de los puntos, reduciendo al máximo el MSE.

Beltrán, C. (2020). Material de clase "Aprendizaje de Máquina" del semestre 2020-2. PUCP.

Material para el curso "Fundamentos de Aprendizaje Automático" por Luis Muroya.

EJEMPLO: PLAY TENNIS

Paso previo: codificar las variables categóricas (CART no funciona con datos no numéricos)

	Outlook	Temperatu	Humidity	Wind	Hours Played
0	2	1	0	1	26
1	2	1	0	0	30
2	0	1	0	1	48
3	1	2	0	1	46
4	1	0	1	1	62
5	1	0	1	0	23
6	0	0	1	0	43
7	2	2	0	1	36
8	2	0	1	1	38
9	1	2	1	1	48
10	2	2	1	0	48
11	0	2	0	0	62
12	0	1	1	1	44
13	1	2	0	0	30

Para X=Outlook

Dom(X) = [0, 1, 2]

Puntos de corte: 0.5 y 1.5

Para x<0.5:

$$Media = \frac{48 + 43 + 62 + 44}{4} = 49.25$$

$$MSE = \frac{(49.25 - 48)^2 + \cdots + (49.25 - 44)^2}{4} = 57.68$$

Para x>0.5

$$Media = 38.7$$

$$MSE = 133.61$$

$$\boldsymbol{MSE_{PC=0.5} = 57.68 + 133.61 = 191.29}$$

Repitiendo para las otras variables, se obtiene:

Variable	Split	MSE
Outlook	0.5	191.3
	1.5	433.7
Temperature	0.5	305.0
	1.5	247.0
Humidity	0.5	261.3
Wind	0.5	271.6

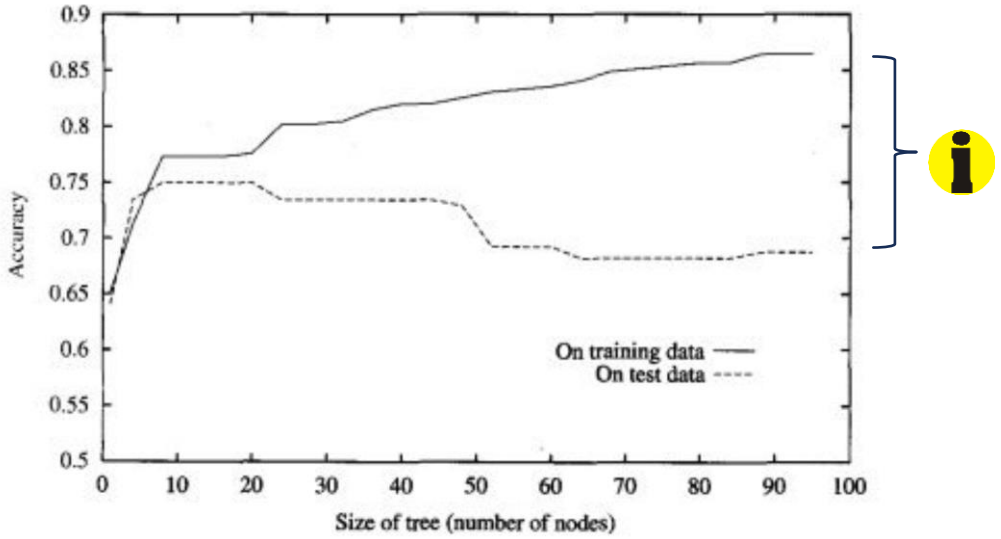
Por tanto, habría que escoger la variable Outlook con punto de corte en 0.5 para el primer Split.

Beltrán, C. (2020). Material de clase "Aprendizaje de Máquina" del semestre 2020-2. PUCP.

Material para el curso "Fundamentos de Aprendizaje Automático" por Luis Muroya.

OVERFITTING Y ÁRBOLES

En ambos algoritmos vemos que el árbol crece hasta tener nodos puros, lo que hace que el problema del ***overfitting*** *sea muy* ***recurrente*** *con ellos.*



Beltrán, C. (2020). Material de clase "Aprendizaje de Máquina" del semestre 2020-2. PUCP.

¿Cómo aplicamos técnicas de regularización en árboles?

A través de la ***poda o prunning****. Esto significa* ***detener*** *el crecimiento del árbol en un punto de su entrenamiento. Esto puede hacerse de dos maneras.*

- ***Pre-prunning:*** *se fija de antemano el máximo número de niveles que puede tener el árbol.*

  — *Efectividad* — *Costoso*

- ***Post-prunning:*** *se deja crecer el árbol hasta el nivel 1, 2, 3 … N. En cada uno se evalúa el resultado del árbol sobre un conjunto de validación.*

  \+ *Efectividad* \+ *Costoso*

FIN DE SESIÓN