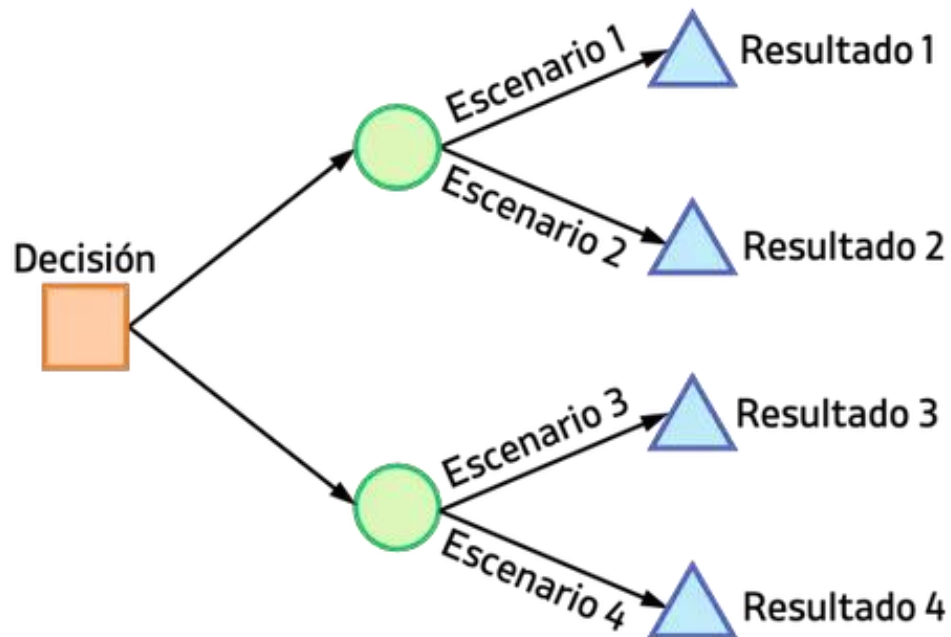




INSTITUTO TECNOLÓGICO BELTRÁN
Centro de Tecnología e Innovación

Árbol de decisión


Árbol de Decisiones



Alumno = Nicolas Mesquiatti

Materia = Aprendizaje Automático

Ejercicio

a) Construcción de un árbol de decisión Una empresa de telecomunicaciones quiere predecir si un cliente aceptará una oferta de plan de datos móviles. Para ello, se dispone de un conjunto de datos con información de 10 clientes, incluyendo los siguientes atributos: 

Edad (en años)

Nivel de uso mensual de datos (en GB)

Tiene línea fija (Sí / No)

Aceptó la oferta (Sí / No)

Datos del conjunto

ID	Edad	Uso de datos	Tiene línea fija	Aceptó oferta
1	24	2.5	No	No
2	38	6.0	Sí	Sí
3	29	3.0	No	No
4	45	8.0	Sí	Sí
5	52	7.5	Sí	Sí
6	33	4.0	No	No
7	41	5.5	Sí	Sí
8	27	2.0	No	No
9	36	6.5	Sí	Sí
10	31	3.5	No	No

Objetivo

Objetivo del ejercicio

1. Calcular la entropía del conjunto original.
2. Evaluar la ganancia de información para los atributos:
Edad (agrupada en rangos: Joven ≤ 30 , Adulto 31–50, Mayor > 50)

Tiene línea fija

Uso de datos (agrupado: Bajo ≤ 3 GB, Medio 3.1–6GB, Alto > 6 GB)

3. Construir el árbol de decisión paso a paso.

4. Concluir cuál es el mejor atributo para comenzar el árbol y cómo se puede usar para predecir si un cliente aceptará la oferta.

1) La Entropía de conjunto **original**

Calculamos las proporciones

Total **10 clientes**

Psi acepta 5 = $5/10 = 0,5$

Pno acepta 5 = $5/10 = 0,5$

Calculo $H(S) = -p(si) * \log_2(p(si)) - p(no) * \log_2(p(no))$

$$= -0,5 * (-1) - 0,5 * (-1) = 1.0000$$

Entropía del conjunto original 1.0, máxima incertidumbre ya que las clases están perfectamente balanceadas

2) Ganancia de información para los atributos

Entropía de cada grupo

a) **Edad**

Subgrupo 1= Joven ≤ 30 ()

3 ejemplos

- 0 “si” = $0/3 = 0$
- 3 “no” = $3/3 = 1$
- $\log_2(0) = 0$
- $\log_2(1) = 0$

Entropía del grupo = $-(0 * 0 + 1 * -0) = 0$

Subgrupo 2: Adulto 31–50

6 ejemplos

- 4 “si” = $4/6 = 0,667$
- 2 “no” = $2/6 = 0,333$
- $\text{Log}_2(0,667) = 0,585$
- $\text{Log}_2(0,333) = -1,585$

Entropía del grupo = $-0.667\log_2(0.667) - 0.333\log_2(0.333)$
= 0,918

Subgrupo 3: Mayor >50)

1 ejemplo

- 1 “si” = $1/1 = 1$
- 0 “no” = $0/1 = 0$
- $\text{Log}_2(1) = 0$
- $\text{Log}_2(0) = 0$

Entropía del grupo = $-(1 * 0 + 0 * -0) = 0$

Entropía ponderada del atributo **edad**

$H(\text{Edad}) = 3/10(0) + 6/10(0.918) + 1/10(0) = 0.5508$

Ganancia de información del atributo EDAD

La **ganancia de información** mide cuánto reduce la incertidumbre un atributo al dividir el conjunto de datos. Así que sí: **cuanto más grande el número, mejor para construir el árbol de decisión.**

Entropía del con original – Entropía ponderada del atributo edad

Ganancia = $H(\text{Original}) - H(\text{Edad}) = 1.000 - 0.5508 = 0.4492$

Tiene una ganancia útil de 44% si los otros dos atributos es menor a este número este sería el más útil, este la ayuda a reducir la incertidumbre, pero no tanto.

b) Tiene línea fija

Si = 5 ejemplos

No = 5 ejemplos

- 5 “si” = $5/5 = 1$
- 5 “no” = $5/5 = 1$
- $\text{Log}_2(1) = 0$
- $\text{Log}_2(1) = 0$

Entropía del grupo = $-(1 * 0 + 0 * -0) = 0$

Entropía Ponderada del atributo **Tiene línea fija**

$H(\text{TieneLineaFija}) = p(\text{si})5/10(0) + p(\text{no})5/10(0) = 0$

Ganancia de información del atributo TIENE LINEA FIJA

$\text{Ganancia} = H(\text{original}) - H(\text{TieneLineaFija}) = 1.000 - 0 = \mathbf{1.000}$

Esto significa que este atributo **divide perfectamente** el conjunto:

- Todos los que **tienen línea fija** aceptaron la oferta
- Todos los que **no tienen línea fija** la rechazaron

Lo cual podemos observar esto en el conjunto de datos, es el atributo ideal para la raíz del árbol de decisión ya que reduce la incertidumbre en un 100%, a diferencia de el atributo Edad que tenía una ganancia de 0,4492 este atributo se usaría en niveles inferiores del árbol.

c)Uso de datos

Entropía de cada grupo

Subgrupo 1= Bajo ≤3 GB

3 ejemplos

- 0 “si” = $0/3 = 0$
- 3 “no” = $3/3 = 1$
- $\text{Log}_2(0) = 0$
- $\text{Log}_2(1) = 0$

Entropía del grupo = $-(0 * 0 + 1 * -0) = 0$

Subgrupo 2: Medio 3.1–6 GB

4 ejemplos

- 2 “si” = $2/4 = 0,5$
- 2 “no” = $2/4 = 0,5$
- $\text{Log}_2(0,5) = -1$
- $\text{Log}_2(0,5) = -1$

Entropía del grupo = $-0,5 * (-1) - 0,5 * (-1) = 1.00$

Subgrupo 3: Alto >6 GB)

3 ejemplos

- 0 “si” = $0/3 = 0$
- 3 “no” = $3/3 = 1$
- $\text{Log}_2(0) = 0$
- $\text{Log}_2(1) = 0$

Entropía del grupo = $-(0 * 0 + 1 * -0) = 0$

Entropía Ponderada del atributo **Uso de datos**

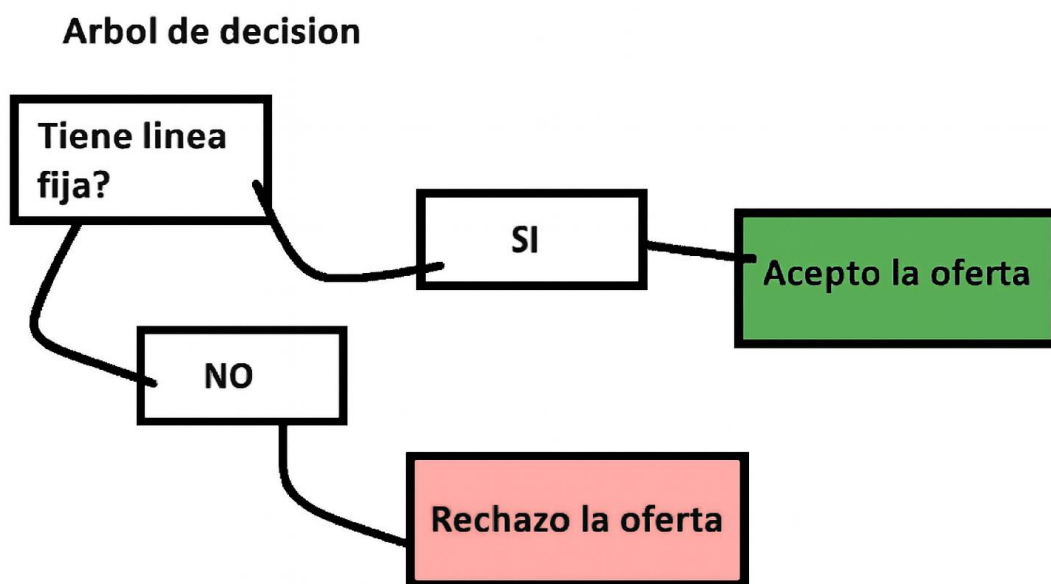
$H(\text{UsodeDatos}) = 3/10(0) + 4/10(1.0) + 3/10(0) = 0.4$

Ganancia de información del atributo Uso de datos

$$\text{Ganancia} = H(\text{Original}) - H(\text{UsodeDatos}) = 1.000 - 0.4 = 0.600$$

Este atributo reduce la incertidumbre un 60% no es tan perfecto como el atributo “Tiene línea fija” pero es mucho más útil de “Edad”

3) Construcción del árbol de decisión



Debido a que el atributo “Tiene línea fija” es el de mayor ganancia de información, se transforma en el nodo raíz, además separa perfectamente y no queda incertidumbre usando este atributo, por eso no hace falta utilizar los otros.

4) El mejor atributo para comenzar el árbol de decisión es **“Tiene línea fija”**, ya que presenta la **mayor ganancia de información (1.000)**. Esto significa que **divide perfectamente** a los clientes en dos grupos:

- Los que **tienen línea fija** → **aceptan** la oferta
- Los que **no tienen línea fija** → **rechazan** la oferta

¿Cómo se usa para predecir? Al evaluar si un nuevo cliente tiene línea fija, el árbol puede **predecir con certeza** su decisión:

- Si responde **“Sí”**, se le puede ofrecer el plan con alta probabilidad de aceptación.
- Si responde **“No”**, es muy probable que lo rechace, y se puede evitar insistir.

Esto es debido a que “Tiene línea fija” es perfecta para iniciar, si no era de 1.000 la ganancia de información y era de menos, aun así, sería el Nodo Raíz y despues le seguiría El Uso de datos que es el Segundo con mayor Ganancia, y por último El atributo de edad.