

IIC2433 - Minería de datos

Árbol de decisión
Hernán Valdivieso
hfvaldivieso@uc.cl

Árboles de decisión

- Técnica de Clasificación supervisada.
- Nodos internos del árbol representan atributos.
- Cada nodo realiza un test basado en los valores del atributo al cual representa.

¿Qué haremos?

- Revisar conceptos claves como entropía y ganancia.
- Dado un dataset que contiene datos categóricos y **numéricos**, construir nuestro árbol de decisión a mano.

Descripción del *dataset*

Cada fila es una **solicitud** realizada por un alumno para **inscribir un curso** de forma excepcional.

- **creditos**: indica la cantidad de créditos que tiene inscrito el alumno.
- **otra_solicitud**: Indica si el alumno solicitó otro curso más para inscribir de forma excepcional.
- **consecuencias**: indica que puede pasar si no le dan el curso solicitado.
- **aceptado**: indica si la solicitud del alumno fue aceptada o no.

creditos	otra_solicitud	consecuencias	aceptado
10	Si	Ninguna	No
10	Si	Se atrasa un semestre	Si
10	No	Se atrasa un semestre	Si
30	Si	Se atrasa la licenciatura	No
30	No	Se atrasa la licenciatura	Si
40	No	Ninguna	No
40	No	Se atrasa un semestre	Si
50	Si	Se atrasa la licenciatura	No
50	Si	Ninguna	No
50	No	Ninguna	No
60	Si	Se atrasa la licenciatura	No

Árbol de decisión - Sólo 1 hoja

División: No tiene
Dist: Si (4) - No (7)
Entropía: XXXX
Respuesta: **No**

¿Cómo calcular la entropía?

$$H(columna) = - \sum_{c=1}^{\#Clases} p_c \cdot \log_2(p_c)$$

p_c Es la proporción de la clase c en la columna indicada.

Para este caso, la columna indicada es la clase objetivo (**aceptada**)

¿Cómo calcular la entropía?

División: No tiene
Dist: Si (4) - No (7)
Entropía: 0.95
Respuesta: **No**

$$H(columna) = - \sum_{c=1}^{\#Clases} p_c \cdot \log_2(p_c)$$

$$H(aceptado) = -(p_{si} \cdot \log_2(p_{si}) + p_{no} \cdot \log_2(p_{no}))$$

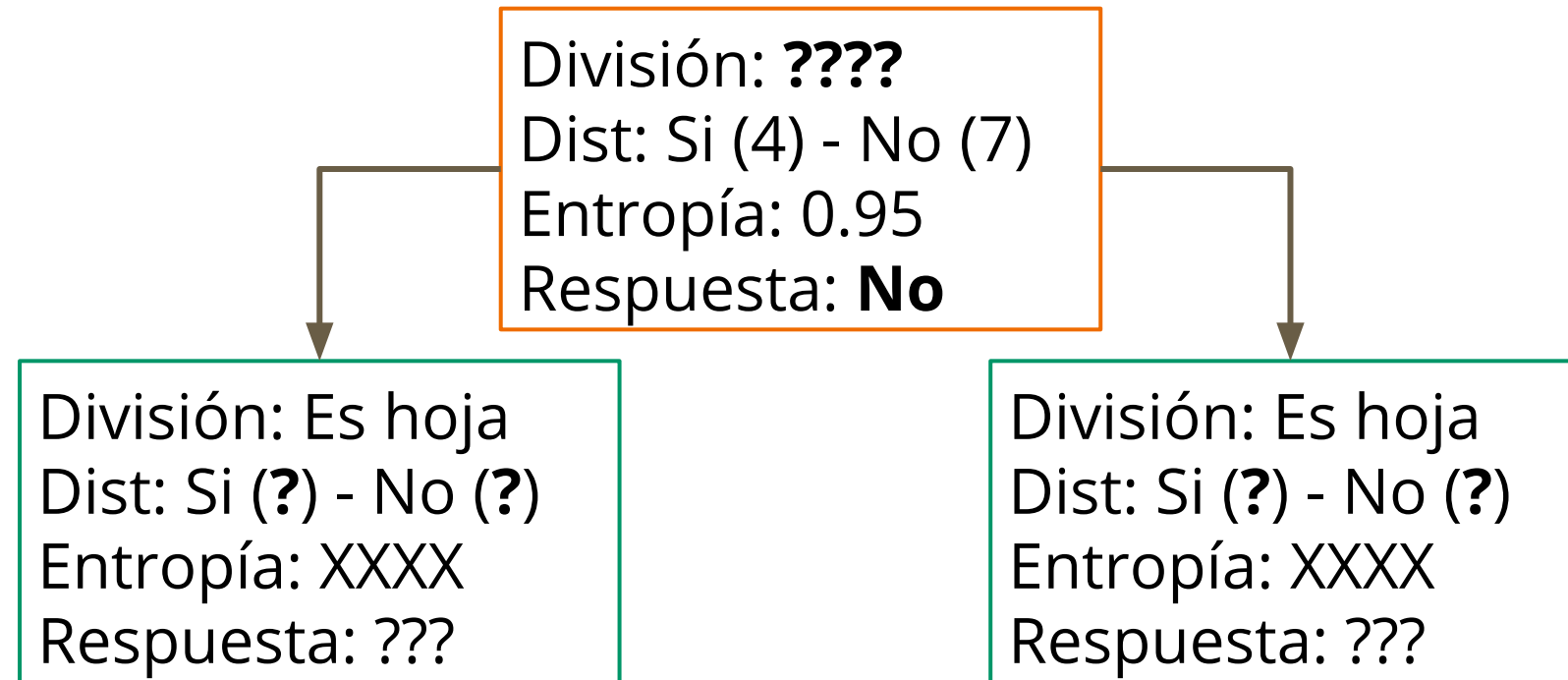
$$H(aceptado) = -(\frac{4}{11} \cdot \log_2(\frac{4}{11}) + \frac{7}{11} \cdot \log_2(\frac{7}{11}))$$

H(aceptado) es aproximadamente **0.95**

Disclaimer: $\log_2 \frac{x}{y} == \log_2(\frac{x}{y})$

creditos	otra_solicitud	consecuencias	aceptado
10	Si	Ninguna	No
10	Si	Se atrasa un semestre	Si
10	No	Se atrasa un semestre	Si
30	Si	Se atrasa la licenciatura	No
30	No	Se atrasa la licenciatura	Si
40	No	Ninguna	No
40	No	Se atrasa un semestre	Si
50	Si	Se atrasa la licenciatura	No
50	Si	Ninguna	No
50	No	Ninguna	No
60	Si	Se atrasa la licenciatura	No

¿Cómo hacemos una división?



Podemos hacer una división por cualquiera de las columnas:

- créditos
- otra_solicitud
- consecuencias

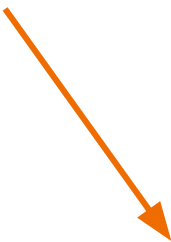
¡Tenemos que ver cuál columna otorga mayor ganancia!

¿Cómo calcular la ganancia?


$$Ganancia(atributo) = H(nodo\ padre) - \sum_{c=1}^{\#Clases} \frac{\#filas\ con\ atributo = c}{\#filas} \cdot H(clase\ objetivo | atributo = c)$$

¿Cómo calcular la ganancia?

Entropía del nuevo dataset
formado por sólo escoger las filas
donde el atributo tiene valor c



$$Ganancia(atributo) = H(nodo\ padre) - \sum_{c=1}^{\#Clases} \frac{\#filas\ con\ atributo = c}{\#filas} \cdot H(clase\ objetivo | atributo = c)$$



Entropía del nodo al
cual le aplicaremos la
división.

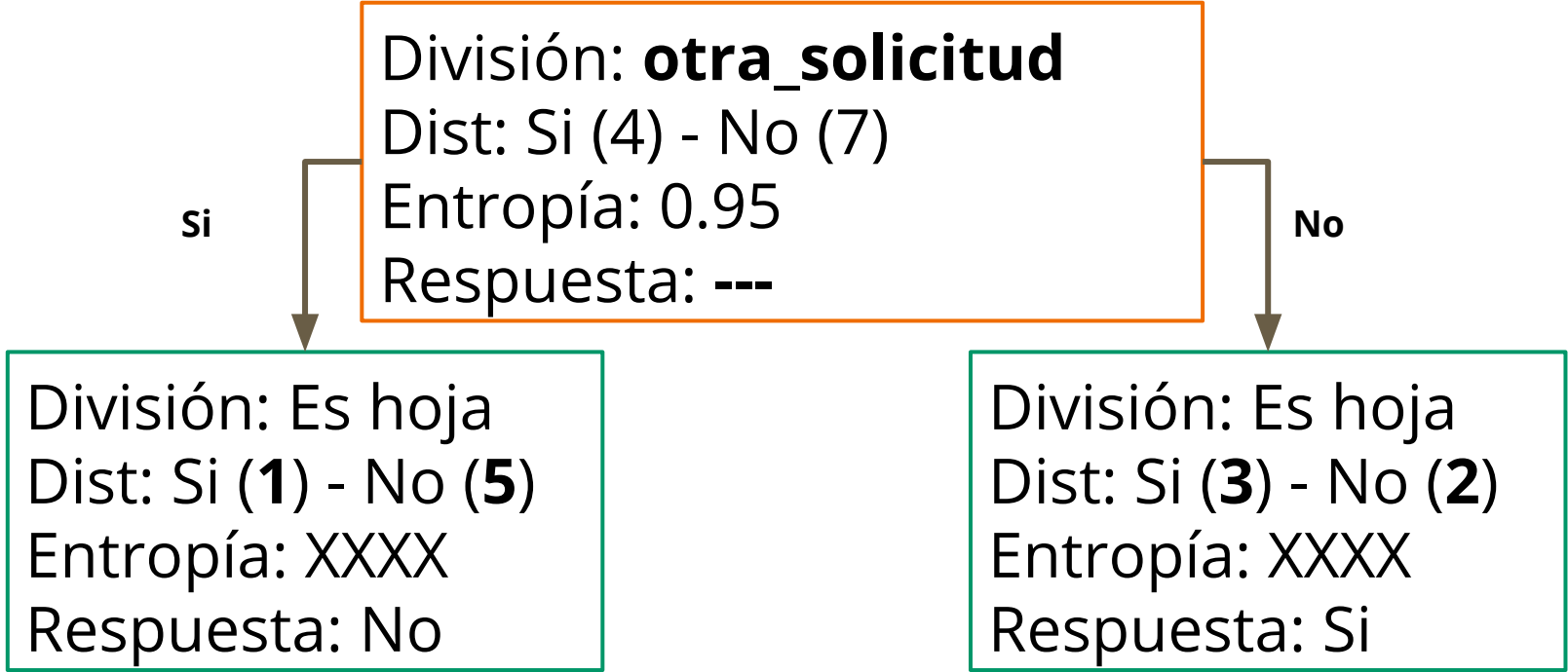


Proporción del atributo
con valor c

Calcular ganancia con otra_solicitud

creditos	otra_solicitud	consecuencias	aceptado
10	Si	Ninguna	No
10	Si	Se atrasa un semestre	Si
30	Si	Se atrasa la licenciatura	No
50	Si	Se atrasa la licenciatura	No
50	Si	Ninguna	No
60	Si	Se atrasa la licenciatura	No

creditos	otra_solicitud	consecuencias	aceptado
10	No	Se atrasa un semestre	Si
30	No	Se atrasa la licenciatura	Si
40	No	Ninguna	No
40	No	Se atrasa un semestre	Si
50	No	Ninguna	No



1. Se genera una posible división.
2. Se obtienen los nuevos *dataset* basados en dicha división.
3. Se calcula la entropía de los nuevos 2 nodos.
4. Se calcula la ganancia otorgada gracias a esa división.

Calcular entropía con otra_solicitud

creditos	otra_solicitud	consecuencias	aceptado
10	Si	Ninguna	No
10	Si	Se atrasa un semestre	Si
30	Si	Se atrasa la licenciatura	No
50	Si	Se atrasa la licenciatura	No
50	Si	Ninguna	No
60	Si	Se atrasa la licenciatura	No

$$H(\text{aceptado}|\text{otra_solicitud} = \text{Si}) = -(p_{si} \cdot \log_2(p_{si}) + p_{no} \cdot \log_2(p_{no}))$$

$$H(\text{aceptado}|\text{otra_solicitud} = \text{Si}) = -(\frac{1}{6} \cdot \log_2(\frac{1}{6}) + \frac{5}{6} \cdot \log_2(\frac{5}{6}))$$

$$H(\text{aceptado}|\text{otra_solicitud} = \text{Si}) = 0.65$$

creditos	otra_solicitud	consecuencias	aceptado
10	No	Se atrasa un semestre	Si
30	No	Se atrasa la licenciatura	Si
40	No	Ninguna	No
40	No	Se atrasa un semestre	Si
50	No	Ninguna	No

$$H(\text{aceptado}|\text{otra_solicitud} = \text{No}) = -(p_{si} \cdot \log_2(p_{si}) + p_{no} \cdot \log_2(p_{no}))$$

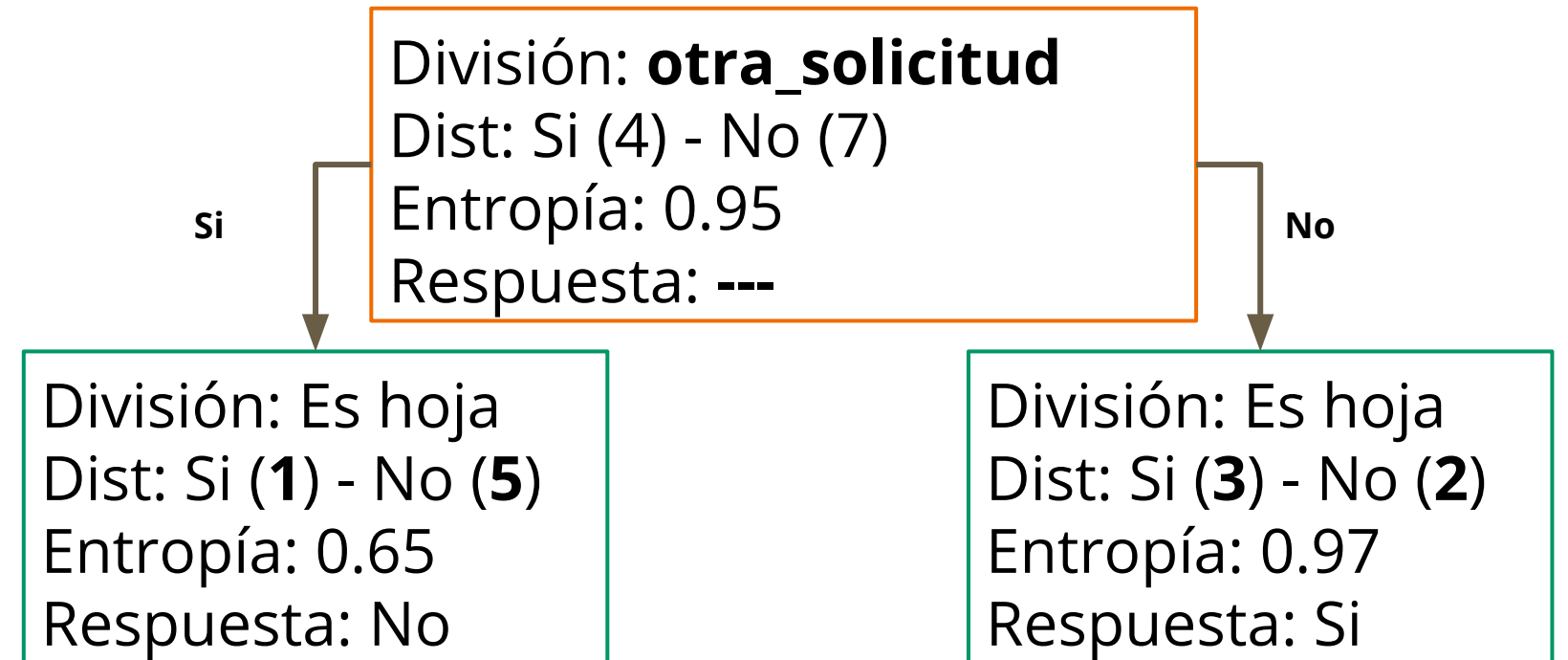
$$H(\text{aceptado}|\text{otra_solicitud} = \text{No}) = -(\frac{3}{5} \cdot \log_2(\frac{2}{5}) + \frac{5}{6} \cdot \log_2(\frac{5}{6}))$$

$$H(\text{aceptado}|\text{otra_solicitud} = \text{No}) = 0.97$$

Calcular ganancia con otra_solicitud

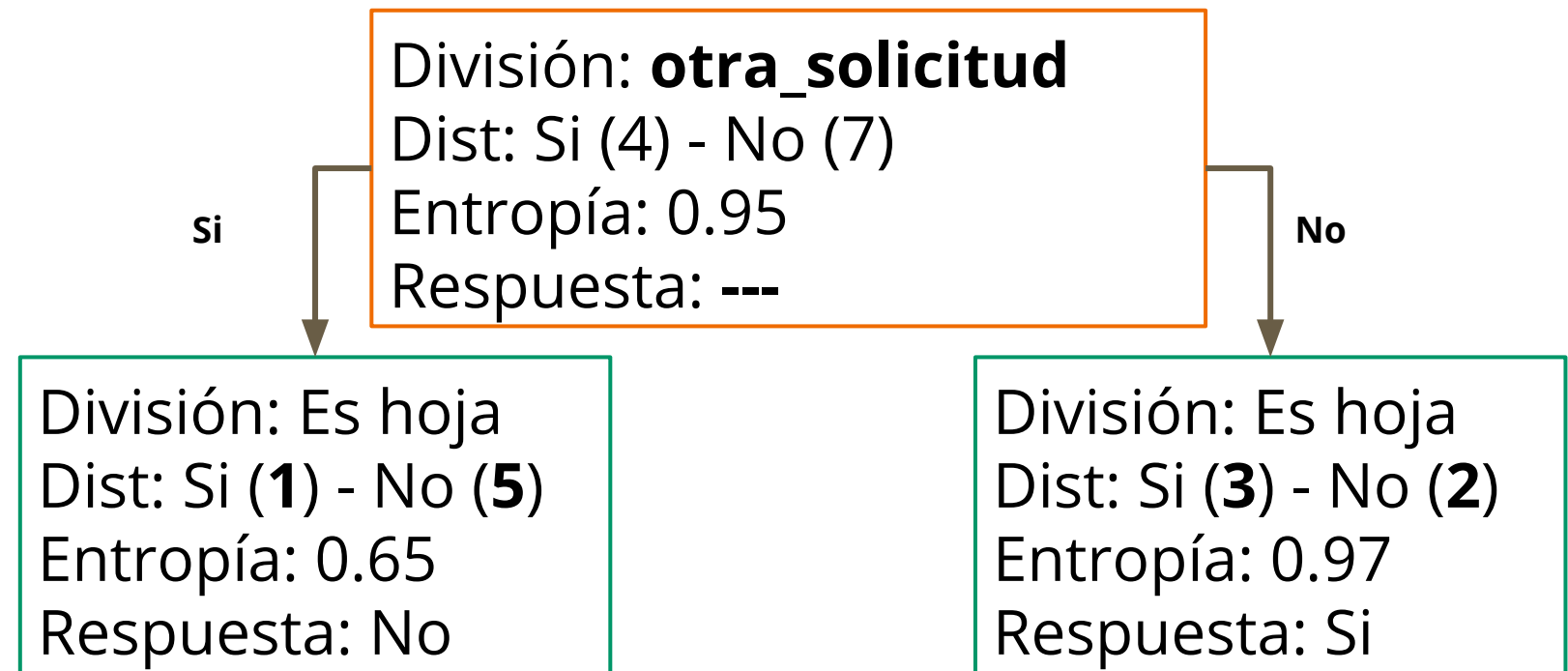
creditos	otra_solicitud	consecuencias	aceptado
10	Si	Ninguna	No
10	Si	Se atrasa un semestre	Si
30	Si	Se atrasa la licenciatura	No
50	Si	Se atrasa la licenciatura	No
50	Si	Ninguna	No
60	Si	Se atrasa la licenciatura	No

creditos	otra_solicitud	consecuencias	aceptado
10	No	Se atrasa un semestre	Si
30	No	Se atrasa la licenciatura	Si
40	No	Ninguna	No
40	No	Se atrasa un semestre	Si
50	No	Ninguna	No



1. Se genera una posible división.
2. Se obtienen los nuevos dataset basados en dicha división.
3. Se calcula la entropía de los nuevos 2 nodos.
4. Se calcula la ganancia otorgada gracias a esa división.

Calcular ganancia con otra_solicitud

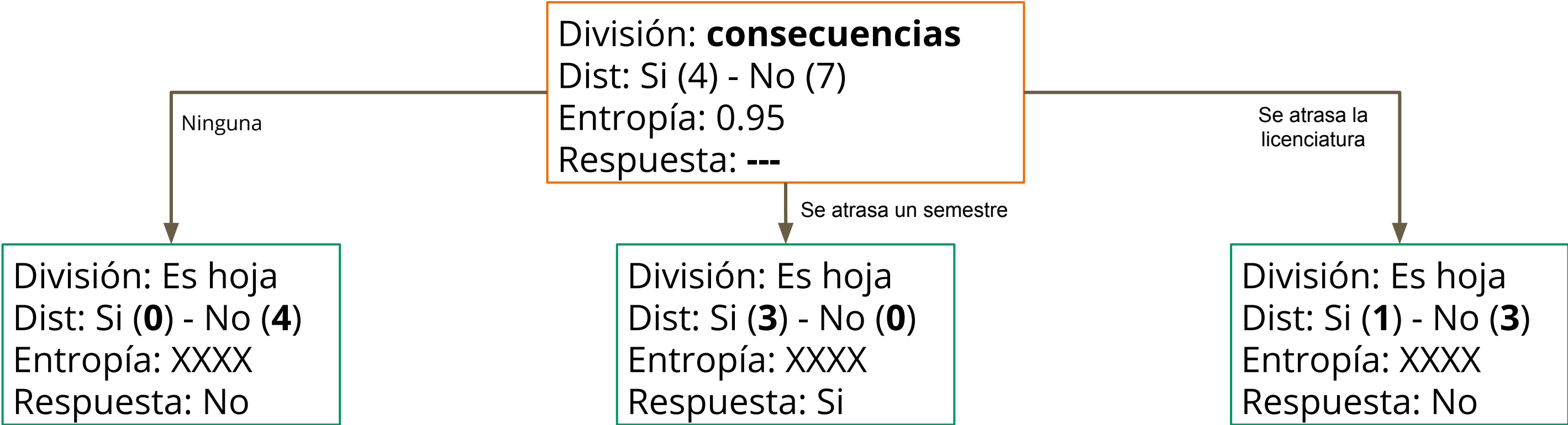


$$Ganancia(atributo) = H(nodo\ padre) - \sum_{c=1}^{\#Clases} \frac{\#filas\ con\ atributo = c}{\#filas} \cdot H(clase\ objetivo | atributo = c)$$

$$0.95 - \left(\frac{\#filas\ otra_solicitud=Si}{\#filas} \cdot H(aceptado | otra_solicitud = Si) + \frac{\#filas\ otra_solicitud=No}{\#filas} \cdot H(aceptado | otra_solicitud = No) \right)$$

$$0.95 - \left(\frac{6}{11} \cdot 0.65 + \frac{5}{11} \cdot 0.97 \right) = 0.15$$

Calcular ganancia con consecuencias

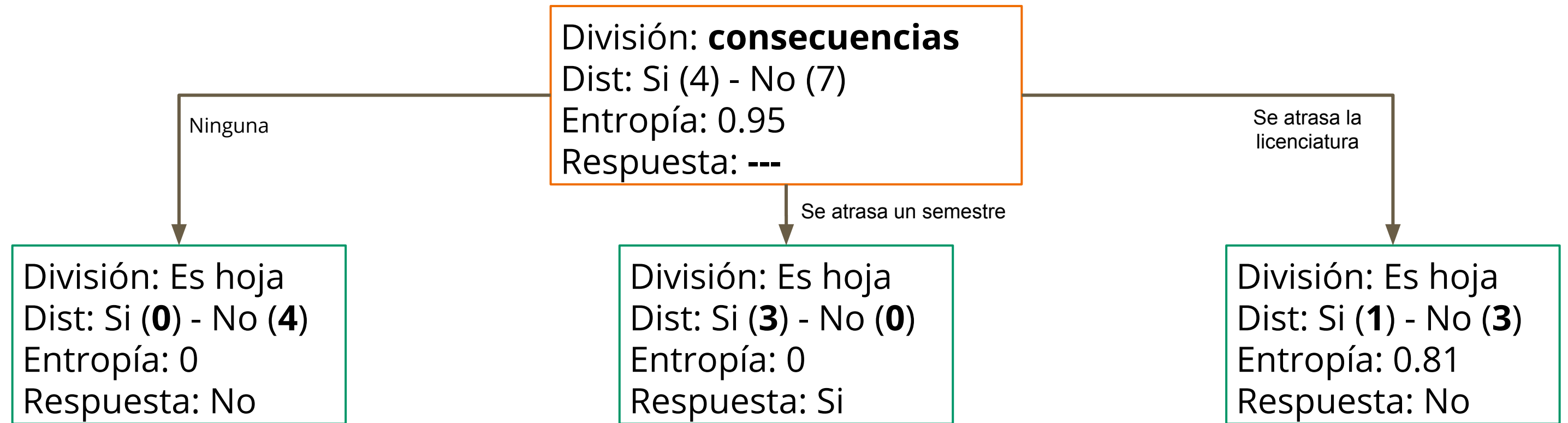


consecuencias	aceptado
Ninguna	No
Ninguna	No
Ninguna	No
Ninguna	No

consecuencias	aceptado
Se atrasa un semestre	Si
Se atrasa un semestre	Si
Se atrasa un semestre	Si

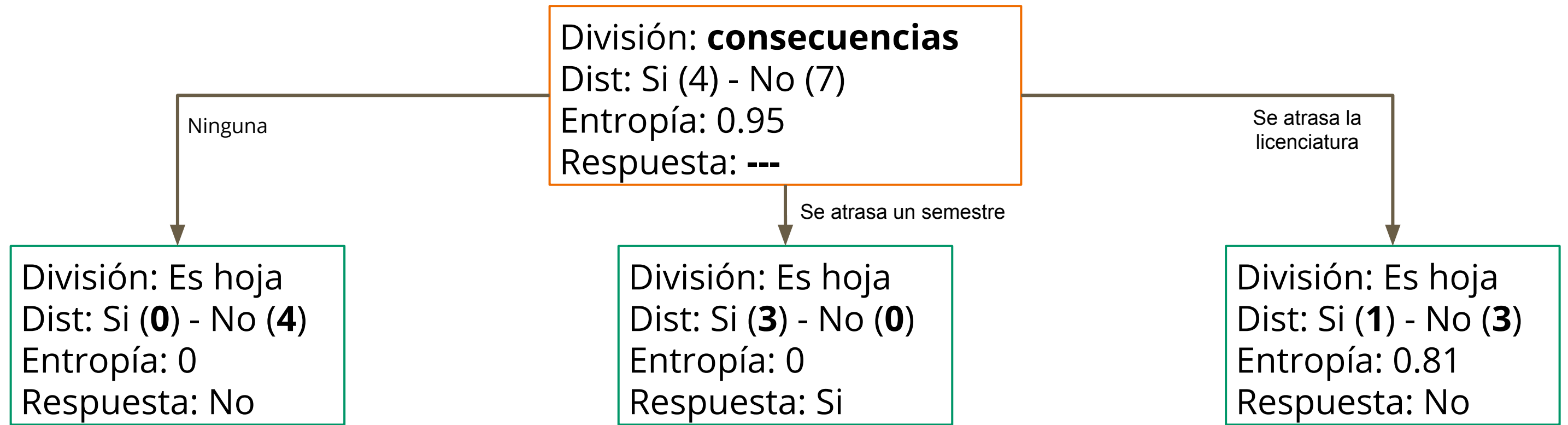
consecuencias	aceptado
Se atrasa la licenciatura	No
Se atrasa la licenciatura	Si
Se atrasa la licenciatura	No
Se atrasa la licenciatura	No

Calcular ganancia con consecuencias



$$-\left(\frac{0}{4} \cdot \log_2\left(\frac{0}{4}\right) + \frac{4}{4} \cdot \log_2\frac{4}{4}\right) = 0 \quad -\left(\frac{3}{3} \cdot \log_2\left(\frac{3}{3}\right) + \frac{0}{3} \cdot \log_2\frac{0}{3}\right) = 0 \quad -\left(\frac{1}{4} \cdot \log_2\left(\frac{1}{4}\right) + \frac{3}{4} \cdot \log_2\frac{3}{4}\right) = 0.81$$

Calcular ganancia con consecuencias



$$-\left(\frac{0}{4} \cdot \log_2\left(\frac{0}{4}\right) + \frac{4}{4} \cdot \log_2\frac{4}{4}\right) = 0 \quad -\left(\frac{3}{3} \cdot \log_2\left(\frac{3}{3}\right) + \frac{0}{3} \cdot \log_2\frac{0}{3}\right) = 0 \quad -\left(\frac{1}{4} \cdot \log_2\left(\frac{1}{4}\right) + \frac{3}{4} \cdot \log_2\frac{3}{4}\right) = 0.81$$

$$Ganancia(atributo) = H(nodo\ padre) - \sum_{c=1}^{\#Clases} \frac{\#filas\ con\ atributo = c}{\#filas} \cdot H(clase\ objetivo|atributo = c)$$

$$0.95 - \left(\frac{4}{11} \cdot 0 + \frac{3}{11} \cdot 0 + \frac{4}{11} \cdot 0.81\right) = 0.66$$

Calcular ganancia con credits

Pero... ¡No es un dato categórico! ¿cómo calculamos la ganancia?

Calcular ganancia con credits

Pero... ¡No es un dato categórico! ¿cómo calculamos la ganancia?

1. Tenemos que pasarlo a una forma categórica.

Calcular ganancia con credits

Pero... ¡No es un dato categórico! ¿cómo calculamos la ganancia?

1. Tenemos que pasarlo a una forma categórica. Para eso, se defina alguna **división** tales como:

- **>= 50**: Se generan 2 grupos, aquellos con 50 o más créditos y aquellos con menos de 50 créditos.
- **== 30**: Se generan 2 grupos, aquellos con justo 30 créditos y aquellos que tienen menos o más.

Calcular ganancia con credits

Pero... ¡No es un dato categórico! ¿cómo calculamos la ganancia?

1. Tenemos que pasarlo a una forma categórica. Para eso, se defina alguna **división** tales como:

- **≥ 50** : Se generan 2 grupos, aquellos con 50 o más créditos y aquellos con menos de 50 créditos.
- **$= 30$** : Se generan 2 grupos, aquellos con justo 30 créditos y aquellos que tienen menos o más.
- También se pueden hacer 3 grupos con la siguiente división:
 - ≤ 20
 - > 20 and ≤ 40
 - > 40 .

Calcular ganancia con credits

Pero... ¡No es un dato categórico! ¿cómo calculamos la ganancia?

1. Tenemos que pasarlo a una forma categórica. Para eso, se defina alguna **división** tales como:

- **>= 50**: Se generan 2 grupos, aquellos con 50 o más créditos y aquellos con menos de 50 créditos.
- **= 30**: Se generan 2 grupos, aquellos con justo 30 créditos y aquellos que tienen menos o más.
- También se pueden hacer 3 grupos con la siguiente división:
 - ≤ 20
 - > 20 and ≤ 40
 - > 40 .

La división debe generar 2 o más conjuntos cuya intersección sea vacía. Hacer una división como créditos ≤ 20 VS créditos ≤ 40 hará que una dato con 20 créditos esté en ambos conjunto y eso no es posible.

Calcular ganancia con credits

Pero... ¡No es un dato categórico! ¿cómo calculamos la ganancia?

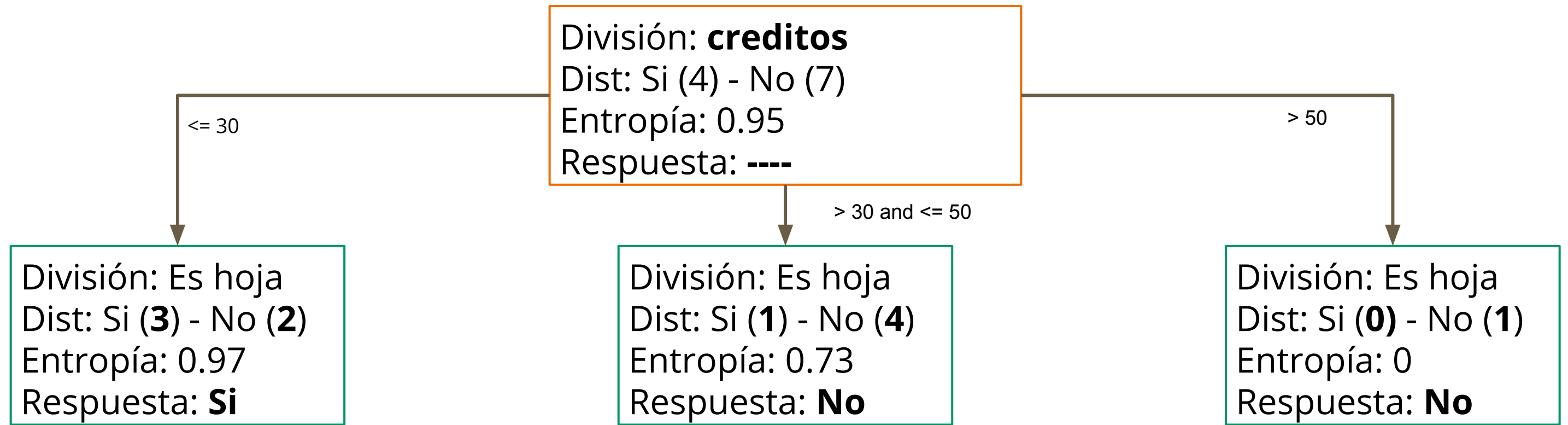
1. Tenemos que pasarlo a una forma categórica. Para eso, se defina alguna **división** tales como:

- **≥ 50** : Se generan 2 grupos, aquellos con 50 o más créditos y aquellos con menos de 50 créditos.
- **$= 30$** : Se generan 2 grupos, aquellos con justo 30 créditos y aquellos que tienen menos o más.
- También se pueden hacer 3 grupos con la siguiente división:
 - ≤ 20
 - > 20 and ≤ 40
 - > 40 .

La división debe generar 2 o más conjuntos cuya intersección sea vacía. Hacer una división como $\text{créditos} \leq 20 \mid \text{créditos} \leq 40$ hará que una dato con 20 créditos esté en ambos conjunto y eso no es posible.

2. Calcular la ganancia como si fuera un dato categórico.

Calcular ganancia con credits



$$-\left(\frac{3}{5} \cdot \log_2\left(\frac{3}{5}\right) + \frac{2}{5} \cdot \log_2\left(\frac{2}{5}\right)\right) = 0.97$$

$$-\left(\frac{1}{5} \cdot \log_2\left(\frac{1}{5}\right) + \frac{4}{5} \cdot \log_2\left(\frac{4}{5}\right)\right) = 0.73$$

$$-\left(\frac{1}{1} \cdot \log_2\left(\frac{1}{1}\right) + \frac{0}{1} \cdot \log_2\left(\frac{0}{1}\right)\right) = 0$$

$$Ganancia(atributo) = H(nodo\ padre) - \sum_{c=1}^{\#Clases} \frac{\#filas\ con\ atributo = c}{\#filas} \cdot H(clase\ objetivo|atributo = c)$$

$$0.95 - \left(\frac{5}{11} \cdot 0.97 + \frac{5}{11} \cdot 0.73 + \frac{1}{11} \cdot 0\right) = 0.18$$

¿Con qué división me quedo?

Probamos 3 divisiones:

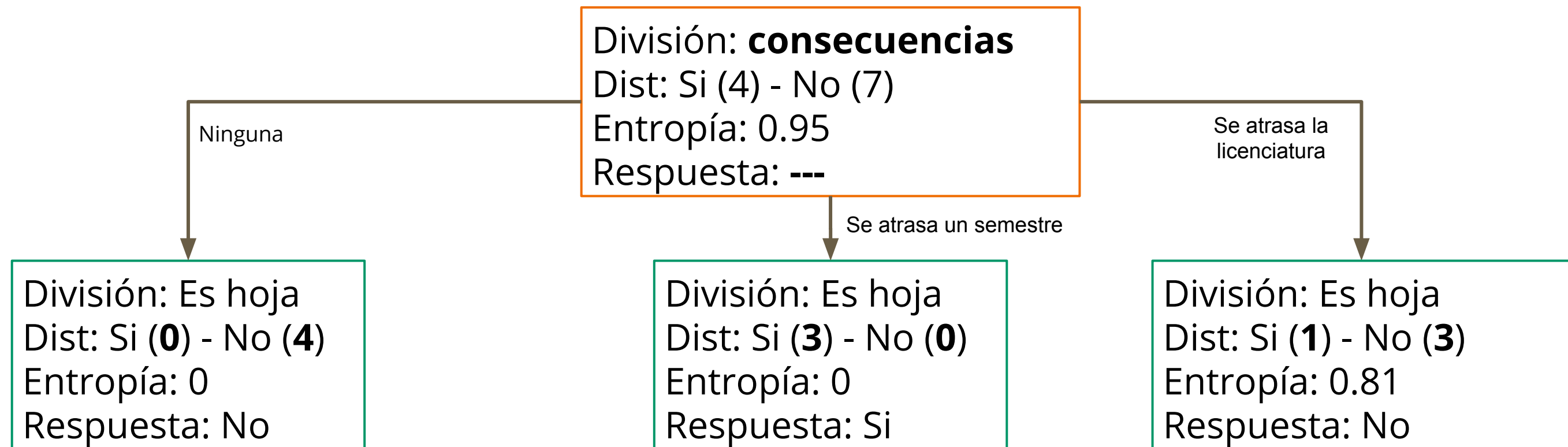
1. Dividir por la columna `otra_solicitud`. Ganancia = 0.15
2. Dividir por la columna `consecuencias`. Ganancia = 0.66
3. Dividir por la columna `creditos` con 3 categorías posibles. Ganancia = 0.18

¿Con qué división me quedo?

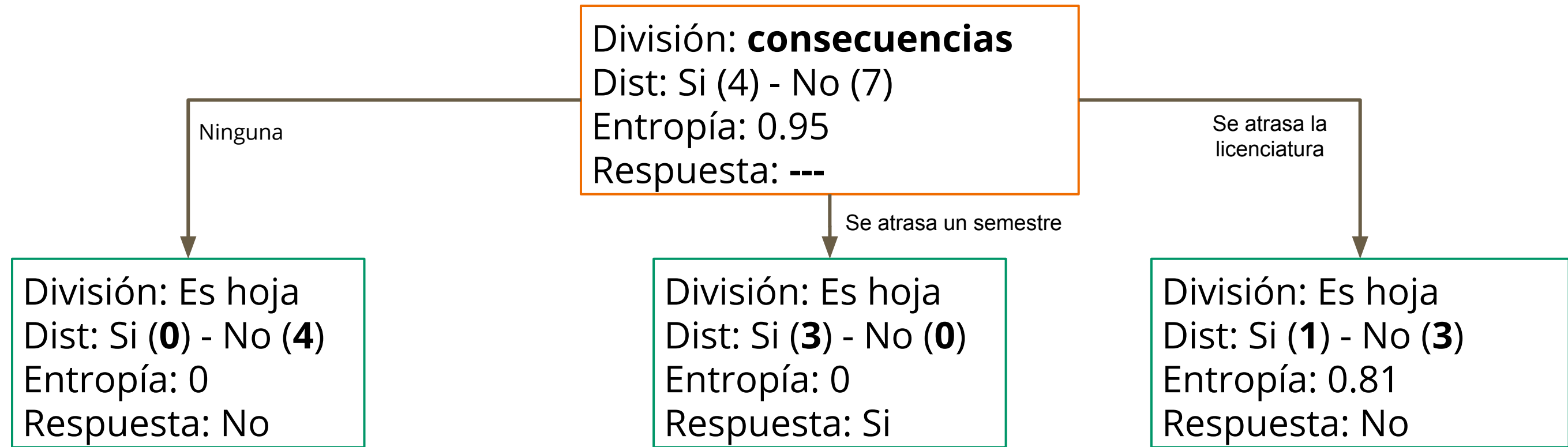
Probamos 3 divisiones:

1. Dividir por la columna otra_solicitud. Ganancia = 0.15
2. **Dividir por la columna consecuencias. Ganancia = 0.66**
3. Dividir por la columna credits con 3 categorías posibles. Ganancia = 0.18

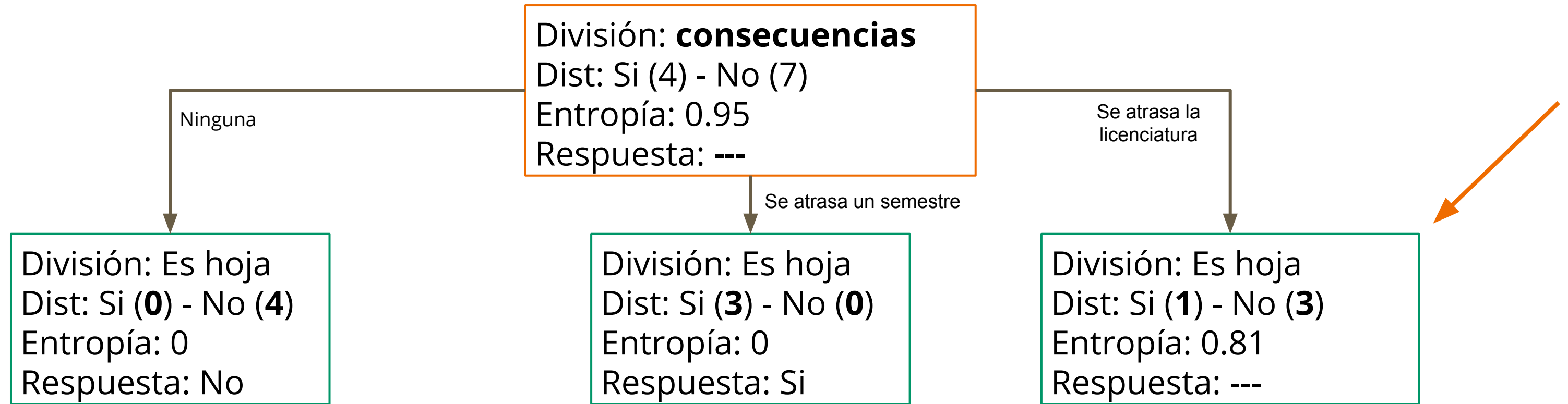
Nos quedamos con la que da **mayor ganancia**.



¿Que sigue?



¿Que sigue?



Todavía se puede dividir otro nodo.

Nos mudamos a dicho nodo, observamos sólo los datos que corresponden a dicha división y volvemos a empezar.

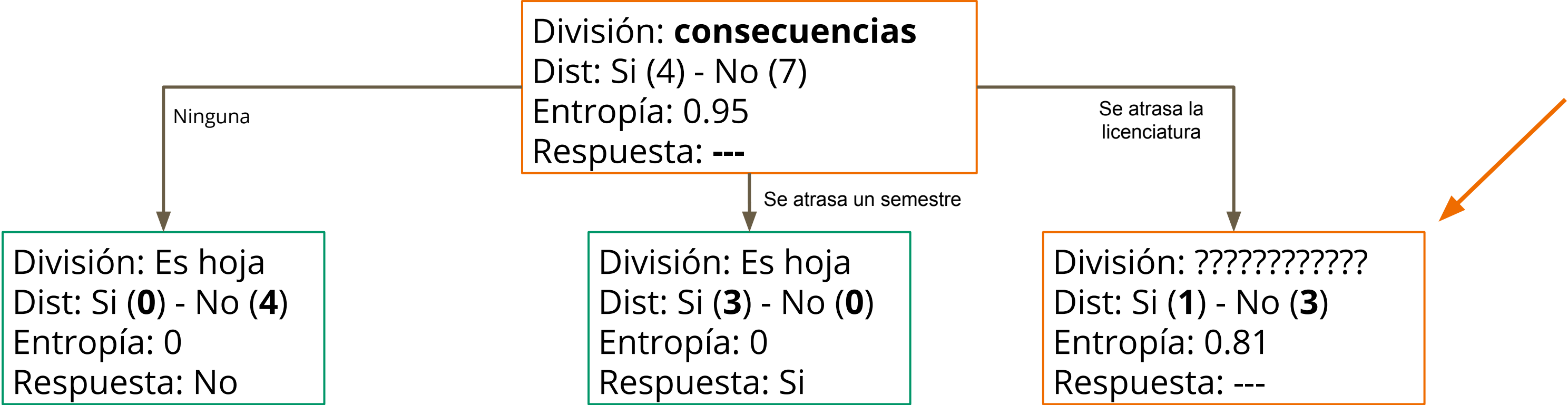
¿Que sigue?

creditos	otra_solicitud	consecuencias	aceptado
10	Si	Ninguna	No
10	Si	Se atrasa un semestre	Si
10	No	Se atrasa un semestre	Si
30	Si	Se atrasa la licenciatura	No
30	No	Se atrasa la licenciatura	Si
40	No	Ninguna	No
40	No	Se atrasa un semestre	Si
50	Si	Se atrasa la licenciatura	No
50	Si	Ninguna	No
50	No	Ninguna	No
60	Si	Se atrasa la licenciatura	No



creditos	otra_solicitud	consecuencias	aceptado
30	Si	Se atrasa la licenciatura	No
30	No	Se atrasa la licenciatura	Si
50	Si	Se atrasa la licenciatura	No
60	Si	Se atrasa la licenciatura	No

¿Que sigue?



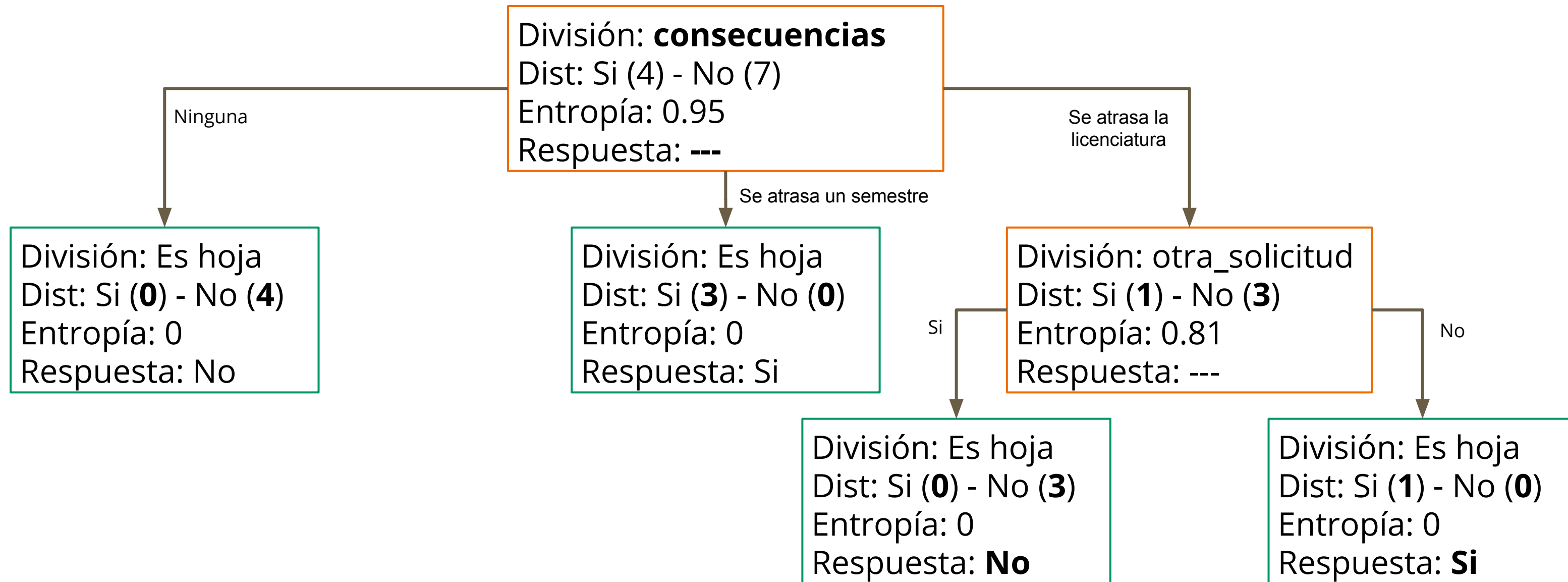
Todavía se puede dividir otro nodo.

Nos mudamos a dicho nodo, observamos sólo los datos que corresponden a dicha división y volvemos a empezar

Entre las columnas disponibles, ¿cuál nos da mayor ganancia?

creditos	otra_solicitud	consecuencias	aceptado
30	Si	Se atrasa la licenciatura	No
30	No	Se atrasa la licenciatura	Si
50	Si	Se atrasa la licenciatura	No
60	Si	Se atrasa la licenciatura	No

¿Que sigue? - Resultado final



Preguntas frecuentes

¿Qué pasa si un nodo hoja no tiene entropía 0?

> Se responde con la clase más presente o buscar otra métrica/criterio para tomar la decisión.



División: Es hoja
Dist: Si (1) - No (4)
Entropía: 0.73
Respuesta: **No**

¿Qué pasa si un nodo hoja tiene entropía 1? (empate de clases)

> Se puede setear (hardcodear) para que diga una clase específica o buscar otra métrica/criterio para tomar la decisión.



División: Es hoja
Dist: Si (3) - No (3)
Entropía: 1
Respuesta: ???

1. Otra métrica puede ser [Information ratio gain](#).
2. En ambos casos deben asegurar **determinismo**, es decir, ante N datos de entrada **exactamente iguales**, se debe responder siempre la misma clase.

Algunas dudas hechas en clases

¿Cuando categorizo, se ocupan las mismas divisiones en los siguientes nodos?

> Depende, puedes utilizarlas, pero lo 100% recomendado es volver a calcular divisiones. Si divides, por ejemplo, por la mediana. En cada división debes volver a calcular la mediana según los datos que ve el nodo porque **se podría hacer una segunda división por dicho atributo numérico.**

Entre una categorización que genera 2 nodos y otra de 3 nodos. ¿Cambia la ganancia?

> Si esa categorización cambia la distribución de la clase objetivo. **Si cambiará la ganancia.**

¿Cómo controlar la profundidad en términos de código?

> Si **no hay más atributos con varianza > 0** , no se puede dividir. Si la **entropía es 0**, no es recomendado porque no generan ganancia. Pueden **definir hiper-parámetros** como profundidad máxima o cantidad mínima de datos para decirle al árbol que no siga dividiendo algún nodo.

Extra: ¿Otras formas de categorizar?

> Puedes dividir por la mediana, por el promedio, si es igual a la moda, por deciles, por quintiles, etc.



PONTIFICIA
UNIVERSIDAD
CATÓLICA
DE CHILE

IIC2433 - Minería de datos

—

Árbol de decisión
Hernán Valdivieso
hfvaldivieso@uc.cl

—
