Arboles de decisión

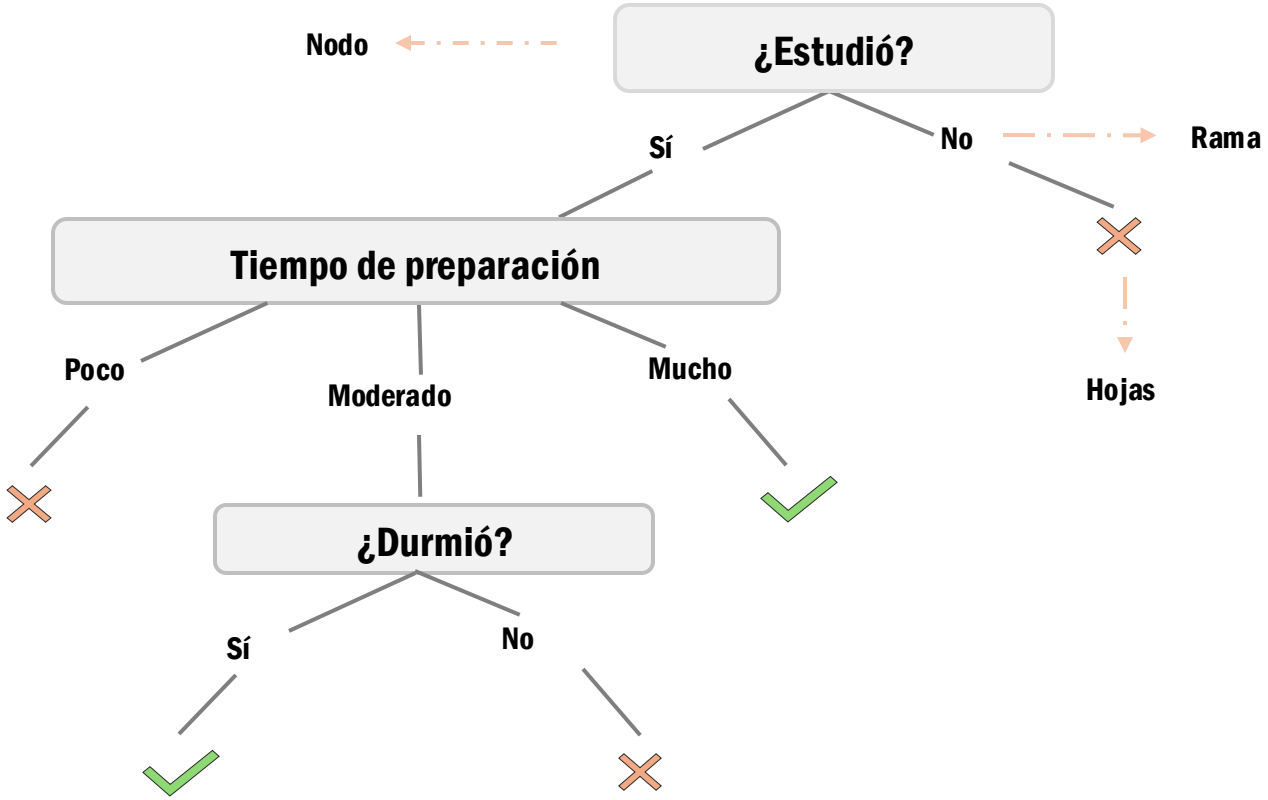
José Andrés Carvajal

Ms. inteligencia Artificial

Contenido

Definición

Un árbol de decisión es un modelo predictivo utilizado tanto para tareas de clasificación como de regresión. Se basa en una estructura en forma de árbol donde cada nodo interno representa una prueba o condición sobre un atributo de los datos, cada rama corresponde a un resultado posible de esa prueba, y las hojas finales contienen la predicción o decisión resultante.



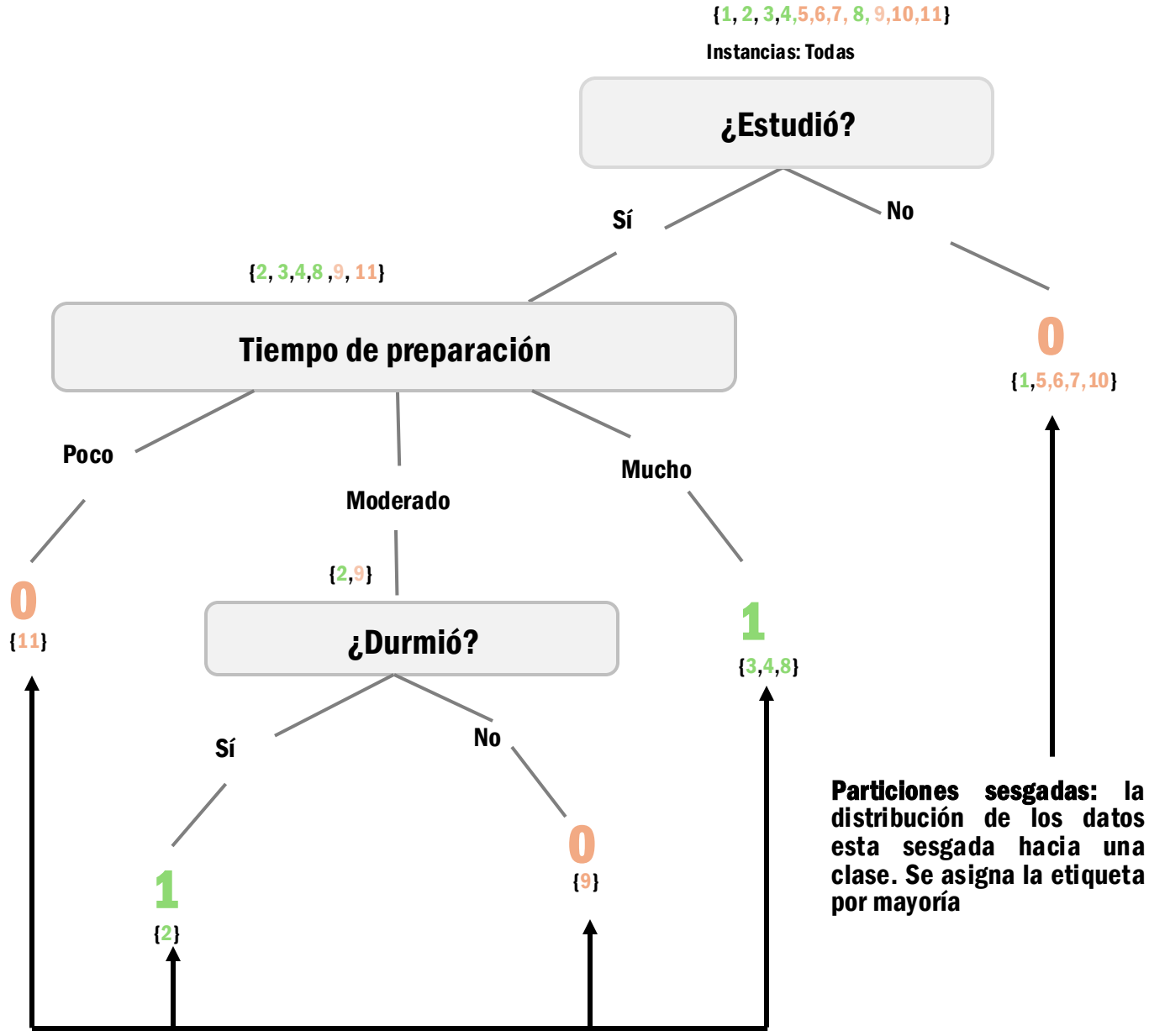
Reglas

1. Si estudió , el tiempo de preparación fue moderado y sí durmió, **entonces** aprueba el examen
2. No estudió, **entonces** no aprueba el examen

3. Si estudió y el tiempo de preparación fue poco **entonces,** no aprueba el examen

Ejemplo

id	Estudió	Tiempo Preparación	Durmió	Resultado
1	No	Poco	No	1
2	Sí	Moderado	Sí	1
3	Sí	Mucho	No	1
4	Sí	Mucho	Sí	1
5	No	Poco	Sí	0
6	No	Moderado	No	0
7	No	Poco	Sí	0
8	Sí	Mucho	No	1
9	Sí	Moderado	No	0
10	No	Poco	No	0
11	Sí	Poco	Sí	0



Particiones sesgadas: la distribución de los datos esta sesgada hacia una clase. Se asigna la etiqueta por mayoría

Particiones puras: solo hay datos de una clase

* En los nodos se conforman particiones disjuntas

Entrenamiento

El árbol aprende seleccionando el test que genera las particiones más puras en cada paso de construcción. Para ello, utiliza métricas que evalúan la pureza u homogeneidad de los subconjuntos resultantes en cada división.

Entropía

$$E(t) = -\sum_{i=1}^{C} \mathbb{P}(i|t) \log_2 \mathbb{P}(i|t)$$

Nivel de desorden o incertidumbre en un conjunto de datos.

Entropía = 0: Nodo es puro

Entropía baja = 1: mezcla uniforme de clases

Índice de GINI

$$GINI(t) = 1 - \sum_{i=1}^{C} \mathbb{P}(i|t)^2$$

Probabilidad de que un elemento, seleccionado al azar, sea clasificado incorrectamente si se asigna según la distribución del nodo

GINI = 0: Nodo es puro

GINI = 0.5: mezcla uniforme de clases

Error esperado

$$Error(t) = 1 - \max_i \mathbb{P}(i|t)$$

Proporción de elementos mal clasificados si se elige la clase mayoritaria en un nodo

Error esperado = 0: Nodo es puro

Error esperado = 0.5 : mezcla uniforme de clases

* $\mathbb{P}(i|t)$:= Fracción de registros pertenecientes a la clase i en el nodo t

¿Cómo funcionan estas métricas?

Nodo	# Datos
Clase 1	0
Clase 2	6

$$\boldsymbol{E(t)} = -\left(\frac{0}{6}\right)\log_2\left(\frac{0}{6}\right) - \left(\frac{6}{6}\right)\log_2\left(\frac{6}{6}\right) = 0$$

$$\boldsymbol{GINI(t)} = 1 - \left(\frac{0}{6}\right)^2 - \left(\frac{6}{6}\right)^2 = 0$$

$$\boldsymbol{Error(t)} = 1 - \max\left\{\left(\frac{0}{6}\right), \left(\frac{1}{6}\right)\right\} = 0$$

Sí la distribución de las clases en un nodo esta totalmente sesgada, es una partición pura y todas las medidas dan un valor de cero

---

Nodo	# Datos
Clase 1	1
Clase 2	5

$$\boldsymbol{E(t)} = -\left(\frac{1}{6}\right)\log_2\left(\frac{1}{6}\right) - \left(\frac{5}{6}\right)\log_2\left(\frac{5}{6}\right) = 0$$

$$\boldsymbol{GINI(t)} = 1 - \left(\frac{1}{6}\right)^2 - \left(\frac{5}{6}\right)^2 = 0.27$$

$$\boldsymbol{Error(t)} = 1 - \max\left\{\left(\frac{1}{6}\right), \left(\frac{5}{6}\right)\right\} = 0.167$$

Sí la distribución de las clases en un nodo esta sesgada, todas las medidas dan un valor menor a uno

¿Cómo funcionan estas métricas?

Nodo	# Datos
Clase 1	2
Clase 2	4

$$\boldsymbol{E(t)} = -\left(\frac{2}{6}\right)\log_2\left(\frac{2}{6}\right) - \left(\frac{4}{6}\right)\log_2\left(\frac{4}{6}\right) = 0.91$$

$$\boldsymbol{GINI(t)} = 1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2 = 0.44$$

$$\boldsymbol{Error(t)} = 1 - \max\left\{\left(\frac{2}{6}\right), \left(\frac{4}{6}\right)\right\} = 0.33$$

A medida que hay mas incertidumbre en los nodos (son menos puros) el valor de las métricas aumenta.

Nodo	# Datos
Clase 1	3
Clase 2	3

$$\boldsymbol{E(t)} = -\left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right) - \left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right) = 1$$

$$\boldsymbol{GINI(t)} = 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = 0.5$$

$$\boldsymbol{Error(t)} = 1 - \max\left\{\left(\frac{3}{6}\right), \left(\frac{3}{6}\right)\right\} = 0.5$$

Sí la distribución de las clases es homogénea entonces todas las medidas alcanzan su valor máximo.

¡Entrenemos un Árbol!

id	Perspectiva	Temperatura	Humedad	Viento	Futbol
1	Soleado	Caliente	Alta	Débil	No
2	Soleado	Caliente	Alta	Fuerte	No
3	Nublado	Caliente	Alta	Débil	Sí
4	Lluvioso	Leve	Alta	Débil	Sí
5	Lluvioso	Frio	Normal	Débil	Sí
6	Lluvioso	Frio	Normal	Fuerte	No
7	Nublado	Frio	Normal	Fuerte	Sí
8	Soleado	Leve	Alta	Débil	No
9	Soleado	Frio	Normal	Débil	Sí
10	Lluvioso	Leve	Normal	Débil	Sí
11	Soleado	Leve	Normal	Fuerte	Sí
12	Nublado	Leve	Alta	Fuerte	Sí
13	Nublado	Caliente	Normal	Débil	Sí
14	Lluvioso	Leve	Alta	Fuerte	No

1. Encontrar la entropía de la etiqueta

$$\boldsymbol{E(t_0)} = -\left(\frac{5}{14}\right)\log_2\left(\frac{5}{14}\right) - \left(\frac{9}{14}\right)\log_2\left(\frac{9}{14}\right)$$

$$\boldsymbol{E(t_0)} = 0.94$$

¡Entrenemos un Árbol!

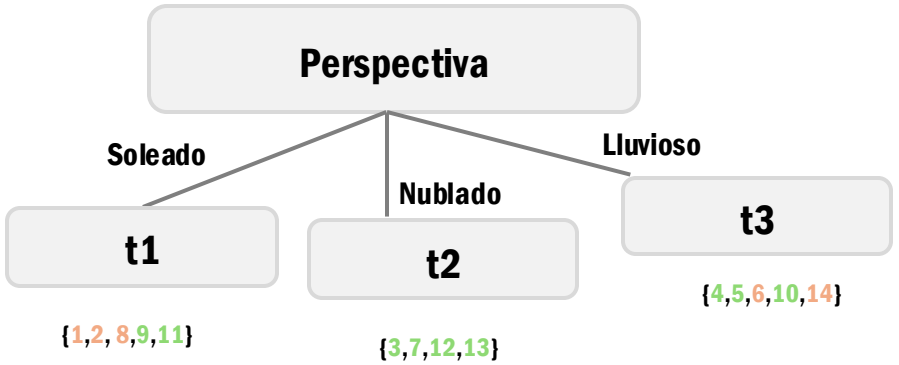
Para determinar el rendimiento del test de atributo se compara el grado de impureza del nodo padre, con el grado de impureza de los nodos hijos. Esto se hace con la ganancia de información que se calcula de la siguiente manera

$$\Delta Info = I(Nodo_{Padre}) - \sum_{j=1}^{k} \frac{N_j}{N} I(Nodo_{Hijo})$$

I: medida de impureza del nodo
N: Número de registros en el nodo padre
k: Número de valores de atributo (divisiones)
Nj: Número de registros asociados al nodo hijo

2. Encontrar la ganancia de información de cada variable

2.1 Se calcula la ganancia de información con la variable Perspectiva



$$E(t_1) = -\left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) - \left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right) = 0.97$$

$$E(t_2) = -\left(\frac{0}{4}\right)\log_2\left(\frac{0}{4}\right) - \left(\frac{4}{4}\right)\log_2\left(\frac{4}{4}\right) = 0$$

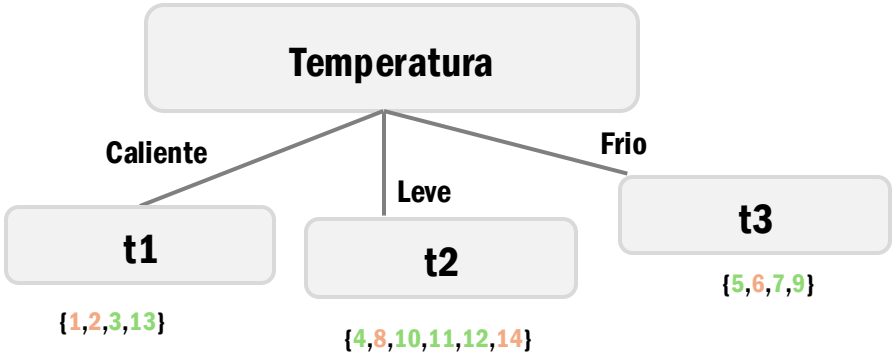
$$E(t_3) = -\left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) - \left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right) = 0.97$$

¡Entrenemos un Árbol!

$$\sum_{j=1}^{k} \frac{N_j}{N} I(Nodo_{Hijo}) = \frac{5}{14} * 0.97 + \frac{4}{14} * 0 + \frac{5}{14} * 0.97 = 0.69$$

$$\Delta Info = 0.94 - 0.69 = 0.25$$

2.2 Se calcula la ganancia de información con la variable Temperatura



$$\boldsymbol{E(t_1)} = -\left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right) - \left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right) = 1$$

$$\boldsymbol{E(t_2)} = -\left(\frac{4}{6}\right)\log_2\left(\frac{4}{6}\right) - \left(\frac{2}{6}\right)\log_2\left(\frac{2}{6}\right) = 0.92$$

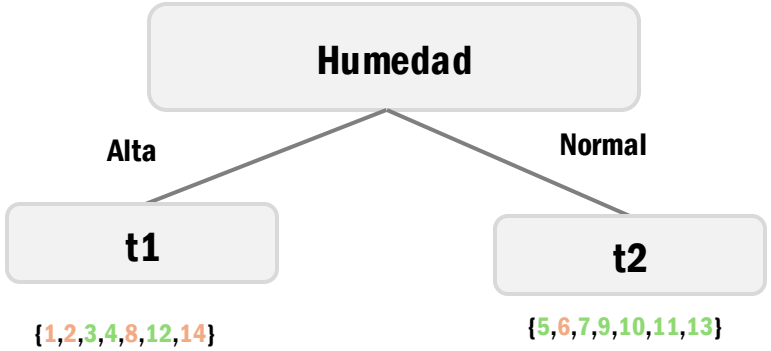
$$\boldsymbol{E(t_3)} = -\left(\frac{3}{4}\right)\log_2\left(\frac{3}{4}\right) - \left(\frac{1}{4}\right)\log_2\left(\frac{1}{4}\right) = 0.81$$

$$\sum_{j=1}^{k} \frac{N_j}{N} I(Nodo_{Hijo}) = \frac{4}{14} * 1 + \frac{6}{14} * 0.92 + \frac{4}{14} * 0.81 = 0.911$$

$$\Delta Info = 0.94 - 0.911 = 0.029$$

¡Entrenemos un Árbol!

2.3 Se calcula la ganancia de información con la variable Humedad



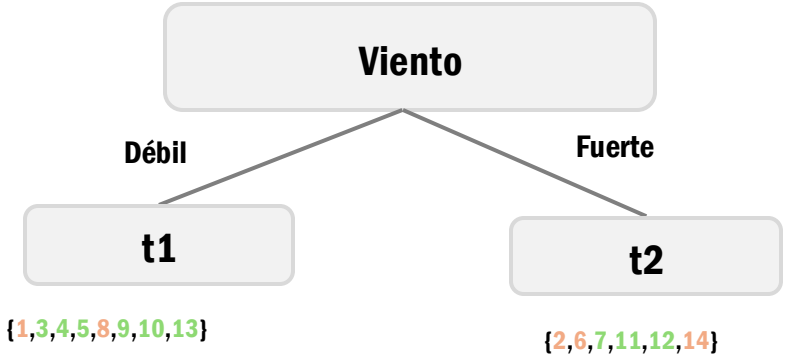
$$E(t_1) = -\left(\frac{3}{7}\right)\log_2\left(\frac{3}{7}\right) - \left(\frac{4}{7}\right)\log_2\left(\frac{4}{7}\right) = 0.99$$

$$E(t_2) = -\left(\frac{1}{7}\right)\log_2\left(\frac{1}{7}\right) - \left(\frac{6}{7}\right)\log_2\left(\frac{6}{7}\right) = 0.59$$

$$\sum_{j=1}^{k}\frac{N_j}{N} I(Nodo_{Hijo}) = \frac{7}{14} * 0.99 + \frac{7}{14} * 0.59 = 0.79$$

$$\Delta Info = 0.94 - 0.79 = 0.15$$

2.4 Se calcula la ganancia de información con la variable Viento



$$E(t_1) = -\left(\frac{6}{8}\right)\log_2\left(\frac{6}{8}\right) - \left(\frac{2}{8}\right)\log_2\left(\frac{2}{8}\right) = 0.81$$

$$E(t_2) = -\left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right) - \left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right) = 1$$

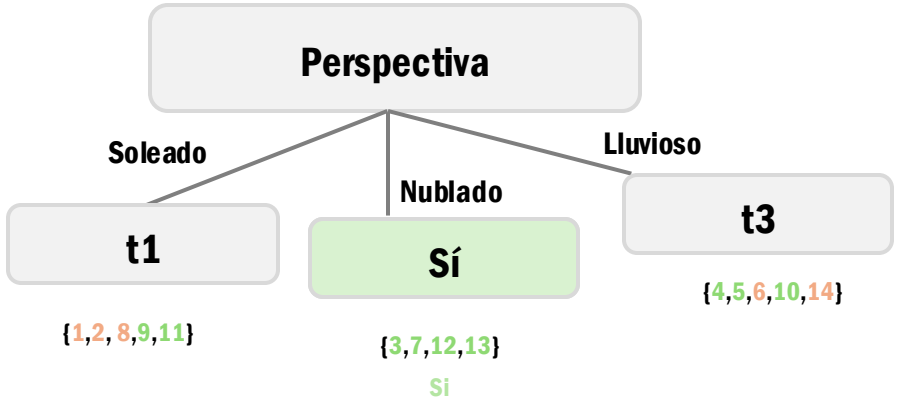
$$\sum_{j=1}^{k}\frac{N_j}{N} I(Nodo_{Hijo}) = \frac{8}{14} * 0.81 + \frac{6}{14} * 1 = 0.89$$

$$\Delta Info = 0.94 - 0.89 = 0.05$$

¡Entrenemos un Árbol!

2.5 Se selecciona la característica que tenga la mayor ganancia de entropía

Característica	Gancia de Entropía
Perspectiva	0.25
Temperatura	0.029
Humedad	0.15
Viento	0.05



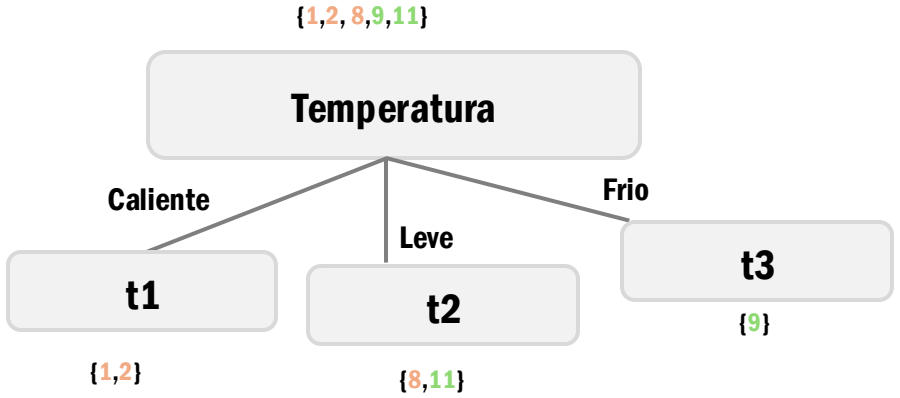
3. Encontrar la ganancia de información de cada nodo con las variables faltantes

¡Entrenemos un Árbol!

3.1 Se calcula la entropía parental

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

3.2 Se calcula la ganancia de información con la variable Temperatura bajo la categoría soleado



$$E(t_1) = -\left(\frac{2}{2}\right)\log_2\left(\frac{2}{2}\right) - \left(\frac{0}{2}\right)\log_2\left(\frac{0}{2}\right) = 0$$

$$E(t_2) = -\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) = 1$$

$$E(t_3) = -\left(\frac{1}{1}\right)\log_2\left(\frac{1}{1}\right) - \left(\frac{0}{1}\right)\log_2\left(\frac{0}{1}\right) = 0$$

$$\sum_{j=1}^{k}\frac{N_j}{N}I(Nodo_{Hijo}) = \frac{2}{5}*0 + \frac{2}{5}*1 + \frac{1}{5}*0 = 0.4$$

$$\Delta Info = 0.97 - 0.4 = 0.571$$

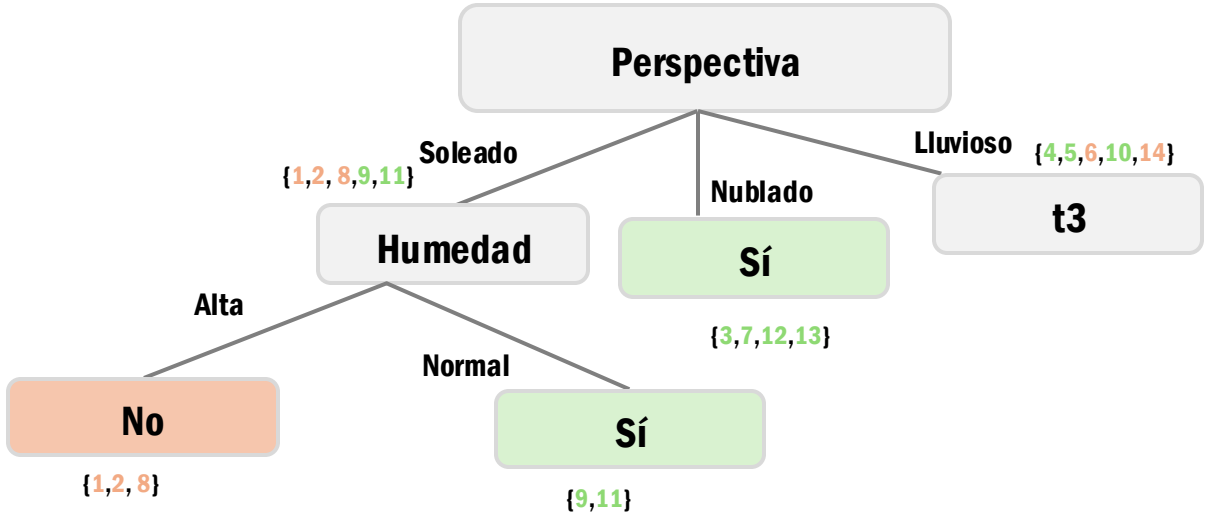
¡Entrenemos un Árbol!

3.3 Se calcula la ganancia de información con la variable Humedad bajo la categoría soleado

3.4 Se calcula la ganancia de información con la variable Viento bajo la categoría soleado

3.5 Se selecciona la característica que tenga la mayor ganancia de entropía

Característica	Ganancia de Entropía
Temperatura	0.571
Humedad	0.971
Viento	0.020



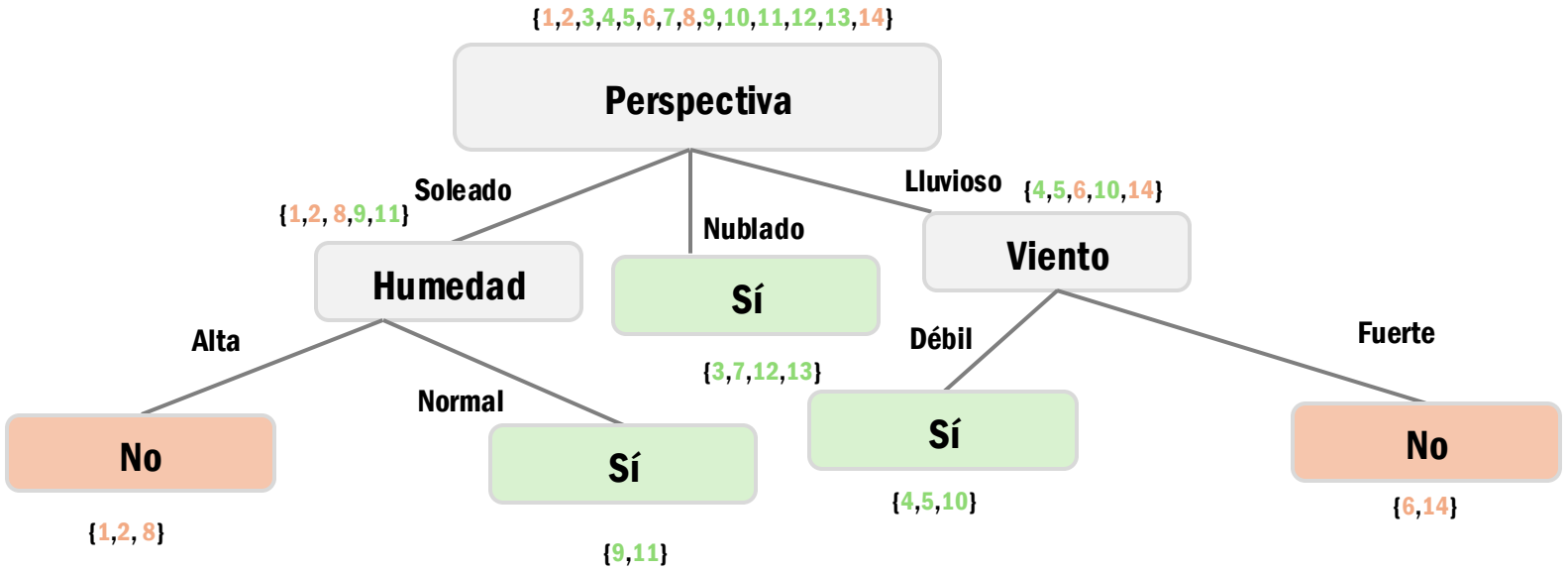
¡Entrenemos un Árbol!

3.7 Se calcula la ganancia de información con la variable Humedad bajo la categoría lluvioso

3.8 Se calcula la ganancia de información con la variable Viento bajo la categoría lluvioso

3.9 Se calcula la ganancia de información con la variable Temperatura bajo la categoría lluvioso

3.10 Se selecciona la característica que tenga la mayor ganancia de entropía



Condiciones de parada

Pureza del nodo

Parar la expansión del nodo cuando todos los registros pertenezcan a la misma clase

Ganancia de información mínima

El árbol se detiene si la mejora en la ganancia de información es menor que un umbral predefinido.

Tamaño mínimo del nodo

Si un nodo tiene menos de $\mathrm{n}_{\min}$ instancias, no se divide.

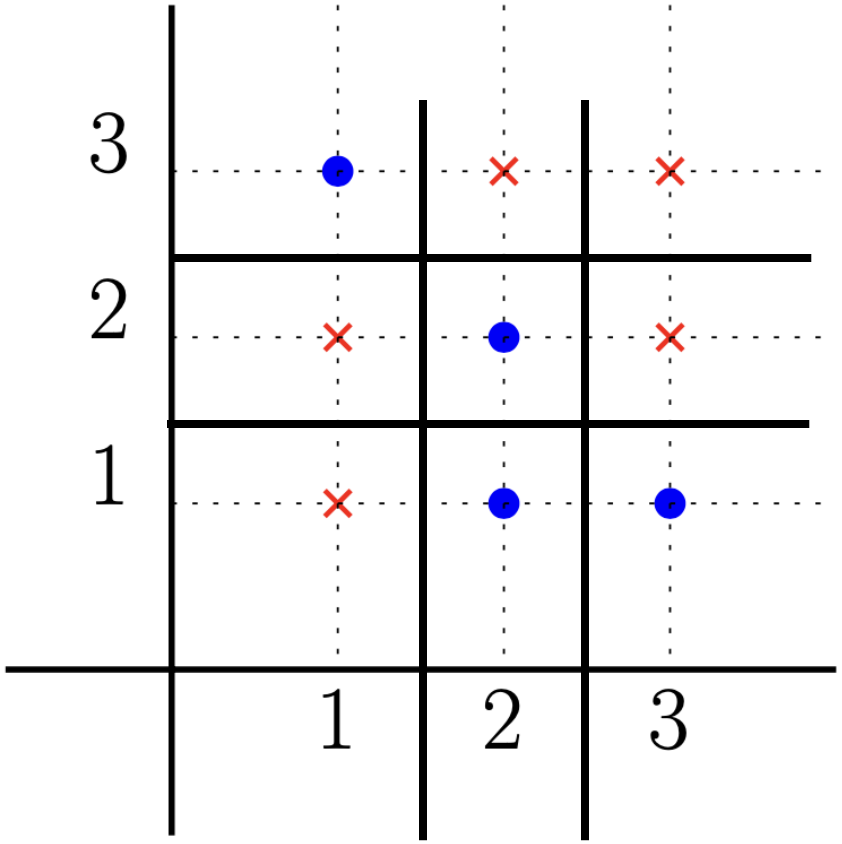
Número máximo de hojas

Se puede definir un límite máximo para el número de hojas en el árbol..

Profundidad máxima

Un árbol de profundidad máxima d no tendrá ningún camino raíz –hoja de más de d niveles

Variables numéricas



$$GINI(t) = 1 - \sum_{i=1}^{C} \mathbb{P}(i|t)^2$$

$$G = \sum_{j=1}^{k} \frac{N_j}{N} I(Nodo_{Hijo})$$

X	t	P
x1	1.5	0.48
x1	2.5	0.48
x2	1.5	0.44
x2	2.5	0.48

Bibliografía

Quinlan, J.R. (1983). Introduction to decision trees. Machine Learning. 1:81-106. Springer.

Quinlan, J. R. (1993). C4.5: programs for machine learning. Morgan Kaufmann Publishers,Inc.

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984). Classification and Regression Trees. Chapman and Hall/CRC.

Kubat M (2017). An Introduction to Machine Learning. Second Edition. 6:113 - 130

Gracias

Jose Andres Carvajal
joseandresb1@hotmail.com
Regional AI Specialist - LATAM