Escuela de Computación, Facultad de Ciencias, UCV

Minería de Datos

Técnicas de Minería de datos – árboles de decisión

Agenda:

- **Clasificación con árboles de decisión - Algoritmo de Hunt**
- **Aspectos de diseño**

Clasificación con árboles de decisión

¿Cómo construir un árbol de decisión?

- **Encontrar el árbol óptimo es poco factible debido al tamaño del espacio de búsqueda**
- **Se han desarrollado algoritmos eficientes que inducen árboles con una exactitud razonable en un tiempo razonable**
- **Estos algoritmos emplean una estrategia secuencial (*greedy*) realizando una serie de decisiones de optimización local acerca de cuál atributo utilizar para particionar los datos (espacio de entrada):**
- **Ejemplos:**

Algoritmo de Hunt

ID3, C4.5

CART

SLIQ

CHAID, entre otros

Base de muchos algoritmos de inducción

Algoritmo de Hunt

- **Construye un árbol de decisión de manera recursiva, particionando los registros (espacio de entrada) de forma sucesiva en conjuntos más puros.**
- **La pureza está determinada por la distribución de las clases en el nodo (registros que llegan a ese nodo).**
- **Dado:**

Un conjunto de entrenamiento $\longrightarrow$ $D = \{ x_i, y_i \}, \; i = 1..N$

donde $y \in \{y_1, y_2, \ldots, y_C\}$ **= etiquetas de clase**

Sea:

D_t **el conjunto de registros de entrenamiento asociado a un nodo t.**

Algoritmo de Hunt

Comienzo_procedimiento_general

1. **Si los registros de D_t son de la misma clase y_k entonces t es un nodo hoja etiquetado como y_k**

2. **Si D_t es un conjunto vacío entonces t es un nodo hoja etiquetado con la clase por defecto y_d**

3. **Si D_t contiene registros de más de un clase, entonces:**

    3.1. **Utilizar un test de atributo para dividir los datos en conjuntos más pequeños**
    3.2. **Crear un nodo hoja por cada resultado del test**
    3.3. **Basado en estos resultados, distribuir los registros de D_t a los nodos hijos**
    3.4. **Aplicar recursivamente *procedimiento_general* a cada nodo hijo**

Fin_Si

Fin_procedimiento_general

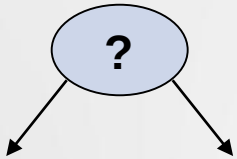
Algoritmo de Hunt

Ejemplo: **Conjunto de datos de una entidad bancaria**

Id	Casa propia	Estado civil	Ingreso anual (M)	Préstamo fallido
1	Si	Soltero	125	No
2	No	Casado	100	No
3	No	Soltero	70	No
4	Si	Casado	120	No
5	No	Divorciado	95	Si
6	No	Casado	60	No
7	Si	Divorciado	120	No
8	No	Soltero	85	Si
9	No	Casado	75	No
10	No	Soltero	90	Si

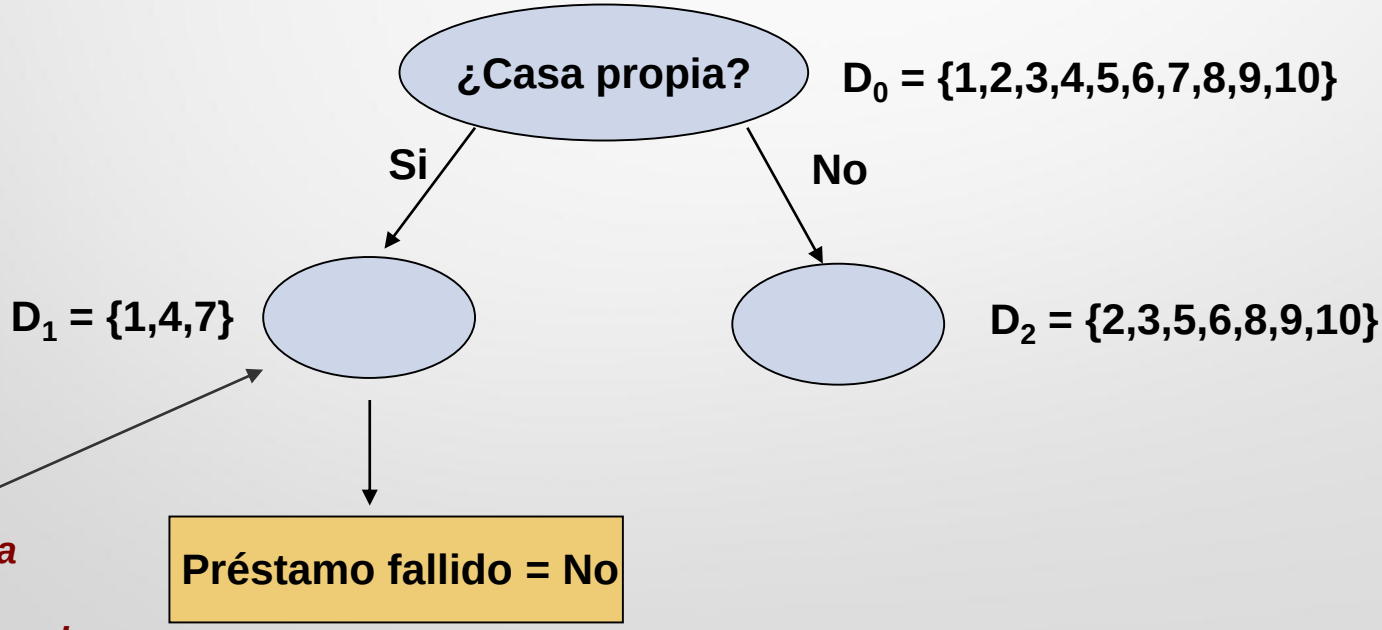
Algoritmo de Hunt

Iteración 0:



D_0 = todos los registros

Iteración 1:



D_0 = {1,2,3,4,5,6,7,8,9,10}

D_1 = {1,4,7}

D_2 = {2,3,5,6,8,9,10}

Préstamo fallido = No

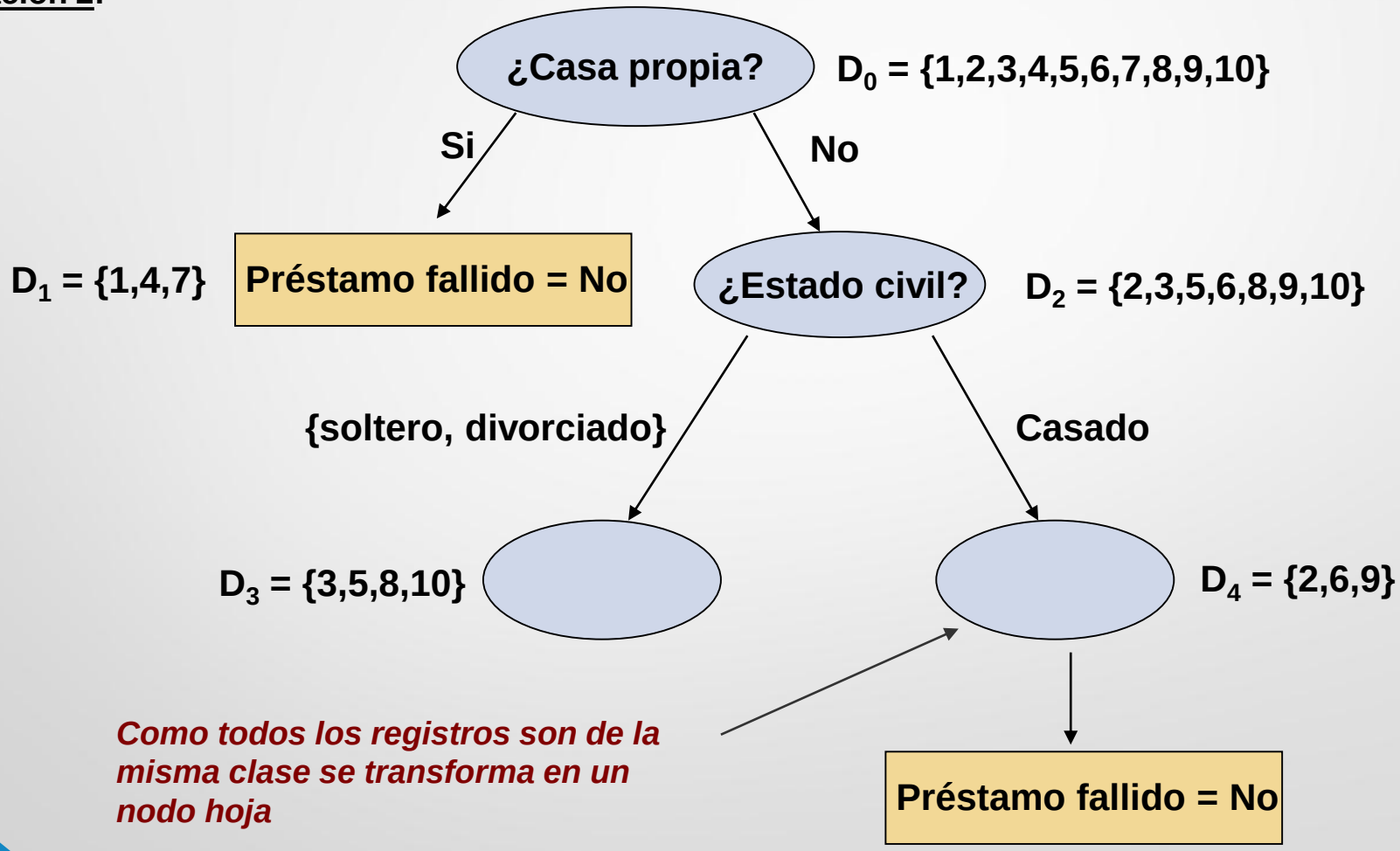
Como todos los registros son de la misma clase se transforma en un nodo hoja

Algoritmo de Hunt

Id	Casa propia	Estado civil	Ingreso anual (M)	Préstamo fallido
1	Si	Soltero	125	No
2	No	Casado	100	No
3	No	Soltero	70	No
4	Si	Casado	120	No
5	No	Divorciado	95	Si
6	No	Casado	60	No
7	Si	Divorciado	120	No
8	No	Soltero	85	Si
9	No	Casado	75	No
10	No	Soltero	90	Si

Algoritmo de Hunt

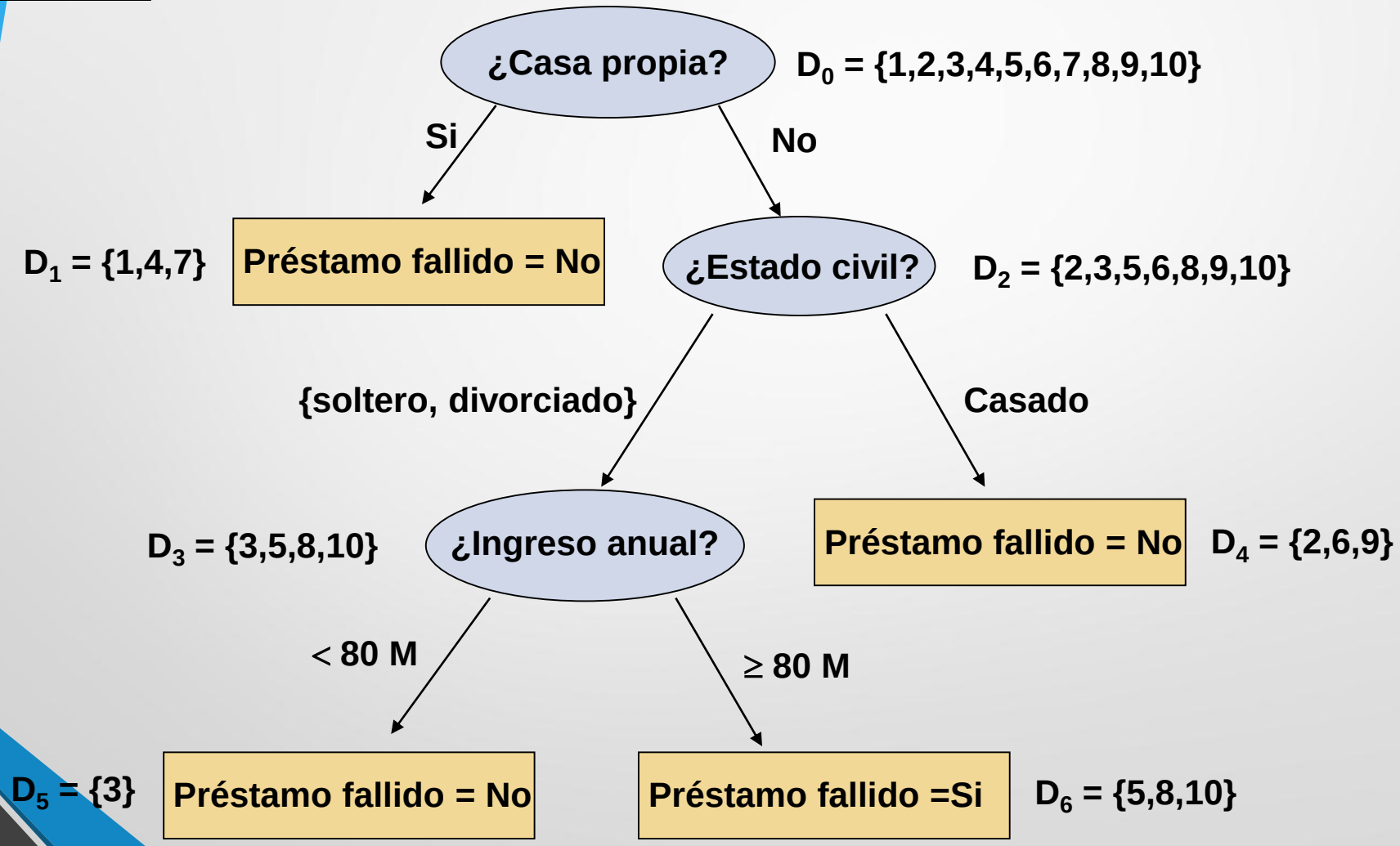
Iteración 2:



Como todos los registros son de la misma clase se transforma en un nodo hoja

Algoritmo de Hunt

Iteración 3:



Aspectos de diseño

Estrategia para la inducción de árboles de decisión:

➡ **Dividir los registros (espacio de entrada) utilizando un test de atributo que optimiza un cierto criterio**

Aspectos de diseño:

- **¿Cómo dividir los registros de entrenamiento?**
  - ☞ ***¿Cómo especificar el test para diferentes tipos de atributos?***
  - ☞ ***¿Cómo determinar el mejor test?***
- **¿Cuándo parar el proceso de división?**
  - ☞ ***Condición de parada***

Aspectos de diseño

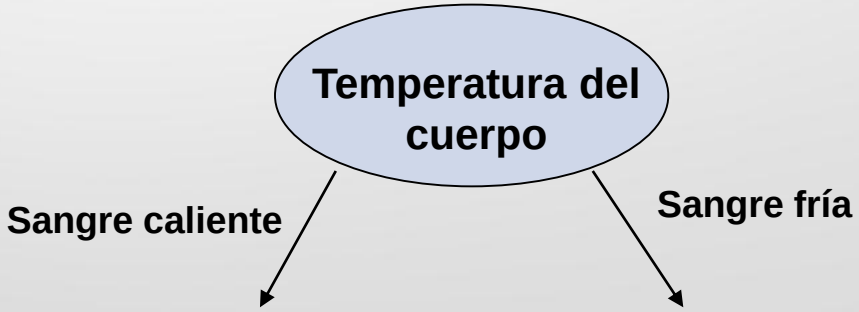
☞ **¿Cómo especificar el test para diferentes tipos de atributos?**

En cada paso recursivo, el algoritmo debe seleccionar un test de atributo para dividir los registros en subconjuntos más pequeños
¿Cómo especificar este test?

Posibilidades:

a) Atributos binarios

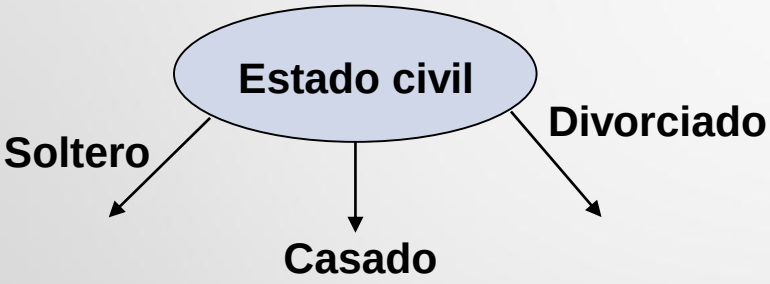
Test de dos vías ➡
(división binaria)



b) Atributos Nominales

Como pueden asumir varios valores, el test puede expresarse de dos formas

Test multivía
(división multivía)



Una vía para cada resultado

Test de dos vías
(división binaria)

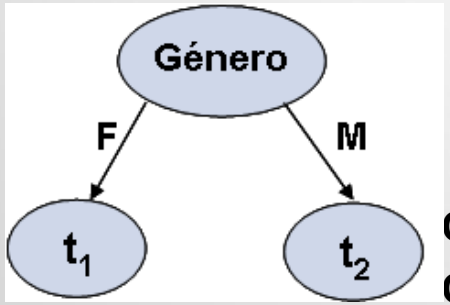
Si k valores, al algoritmo debe considerar $2^{k-1} - 1$ posibilidades

Aspectos de diseño

☞ **¿Cómo determinar el mejor test?**

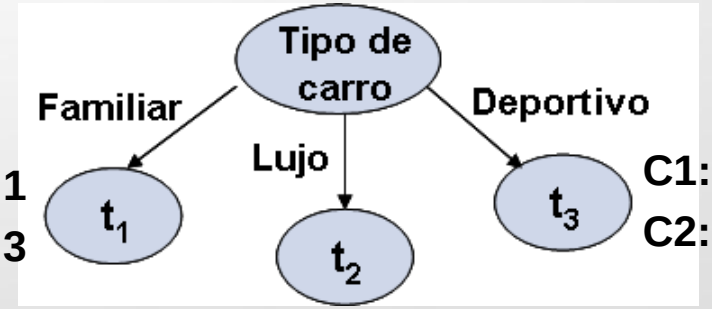
- **Hay muchas medidas que pueden utilizarse para determinar la mejor manera de dividir los registros.**
- **Estas medidas son definidas en función de la distribución de clases (pureza) en el nodo, antes y después de la división**

Ejemplo: en un problema de dos clases (C1,C2) la distribución antes de la división es (0.5, 0.5)



C1: 6
C2: 4

C1: 4
C2: 6



C1: 1
C2: 3

C1: 8
C2: 0

C1: 1, C2: 7

Distribución de clases en los nodos hijos ➡ T1: (0.6, 0.4)
T2: (0.4, 0.6)

Distribución de clases en los nodos hijos ➡ T1: (0.25, 0.75)
T2: (0.125, 0.875)
T3: (1, 0)

Aspectos de diseño

- **A menor el grado de impureza, más sesgada será la distribución de clases.**

  - **Un nodo con una distribución (1, 0)** → **Impureza = 0 (mínima)**
  - **Un nodo con una distribución (0.5, 0.5)** → **Impureza máxima**

Algunas medidas de impureza son:

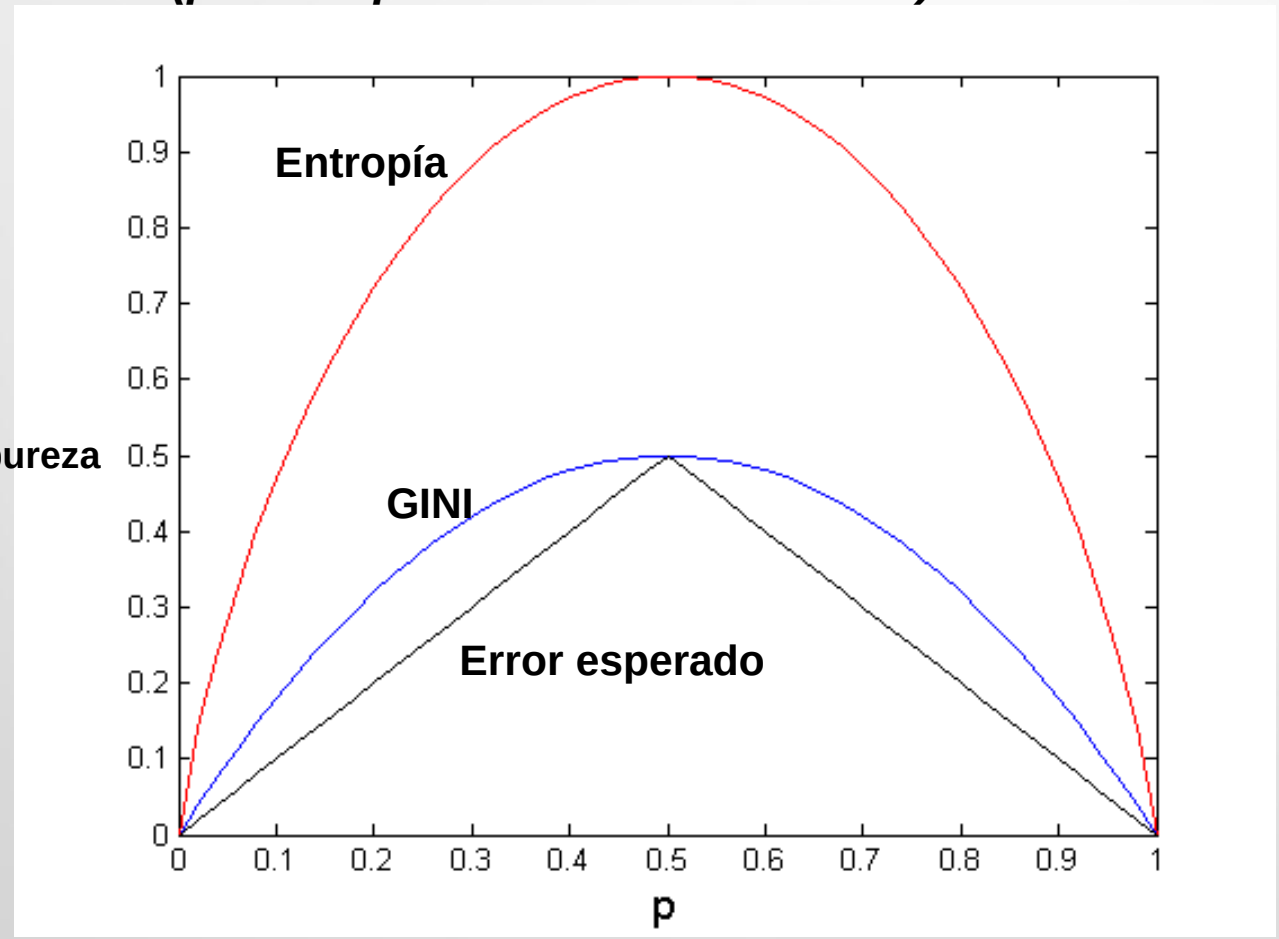
Si p(i|t) = fracción de registros pertenecientes a la clase i en el nodo t

- **Entropía** $= -\sum_{i=1}^{C} p(i|t)\log_2 p(i|t)$
- **Error esperado** $= 1 - \max_i [\, p(i|t) \,]$
- **GINI** $= 1 - \sum_{i=1}^{C} [\, p(i|t) \,]^2$

Donde C = número de clases

Aspectos de diseño

Comparación (para un problema de dos clases):



Las tres medidas obtienen su valor máximo cuando la distribución de clases es uniforme

Aspectos de diseño

Ejemplo:

NODO 1	Cantidad
C1	0
C2	6

Entropía = $- (0/6) \log_2 (0/6) - (6/6) \log_2 (6/6) = 0$

GINI = $1 - (0/6)^2 - (6/6)^2 = 0$

Error esperado = $1 - \max [(0/6), (6/6)] = 0$

NODO 2	Cantidad
C1	1
C2	5

Entropía = $- (1/6) \log_2 (1/6) - (5/6) \log_2 (5/6) = 0.650$

GINI = $1 - (1/6)^2 - (5/6)^2 = 0.278$

Error esperado = $1 - \max [(1/6), (5/6)] = 0.167$

NODO 2	Cantidad
C1	3
C2	3

Entropía = $- (3/6) \log_2 (3/6) - (3/6) \log_2 (3/6) = 1$

GINI = $1 - (3/6)^2 - (3/6)^2 = 0.5$

Error esperado = $1 - \max [(3/6), (3/6)] = 0.5$

Según estos resultados el NODO 1 tiene el menor grado de impureza

Aspectos de diseño

¿Cómo usar las medidas?

- **Para determinar el rendimiento de un test de atributo, se compara el grado de impureza del nodo padre (*antes de la división*), con el grado de impureza de los nodos hijos (*después de la división*).**
- **A más grande la diferencia, mejor el test**
- **La ganancia (Δ) es un criterio que puede utilizarse para determinar la bondad de una división:**

$$\Delta = I(\text{NodoPadre}) - \sum_{j=1}^{K} \frac{N_j}{N} I(\text{NodoHijo}_j)$$

Donde:

I = medida de impureza del nodo
N = número de registros en el nodo padre
K = número de valores del atributo (divisiones)
N_j = número de registros asociados al nodo hijo

Aspectos de diseño

- **Los algoritmos de inducción de árboles de decisión seleccionan el test que proporciona la mayor ganancia**
- **Como I(NodoPadre) es el mismo para todos los posibles test de atributo, maximizar la ganancia es equivalente a minimizar el promedio ponderado de la medida de impureza de los nodos hijos.**

Ejemplo:

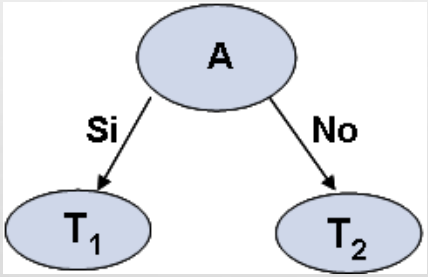
C1: 6
C2: 6 → **GINI = 0.5**

Hay dos atributos (posibles test)

A = { SI, NO}

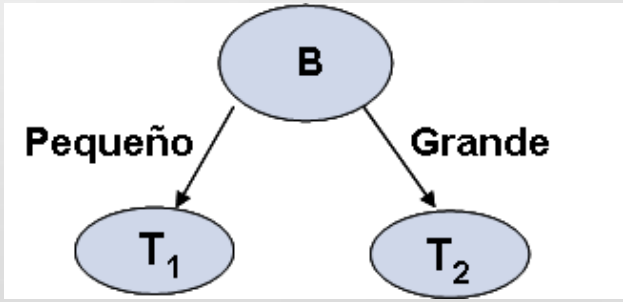
B = {Pequeño, Grande}

¿Cuál escoger? ➡



C1: 4
C2: 3

C1: 2
C2: 3



C1: 1
C2: 4

C1: 5
C2: 3

Aspectos de diseño

Para el atributo A:

GINI (T1) = $1 - (4/7)^2 - (3/7)^2 = 0.4898$

GINI (T2) = $1 - (2/5)^2 - (3/5)^2 = 0.4800$

Promedio ponderado del índice GINI = $(7/12)\ (0.4898) + (5/12)\ (0.4800) = 0.4860$

Para el atributo B:

GINI (T1) = $1 - (1/5)^2 - (4/5)^2 = 0.3200$

GINI (T2) = $1 - (5/7)^2 - (2/7)^2 = 0.4082$

Promedio ponderado del índice GINI = $(5/12)\ (0.3200) + (7/12)\ (0.4082) = 0.3715$

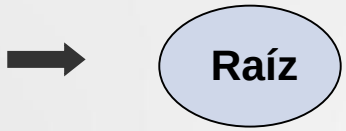
➡ ***Como el atributo B tiene el índice más pequeño, se prefiere sobre el atributo A***

Aspectos de diseño

Ejemplo: determinar el árbol de decisión para el siguiente conjunto de datos utilizando GINI como criterio de evaluación

ID	PREVISIÓN	TEMPERATURA	HUMEDAD	VENTOSO	JUGAR
1	Soleado	Caliente	Alta	No	No
2	Soleado	Caliente	Alta	Si	No
3	Nublado	Caliente	Alta	No	Si
4	Lluvioso	Suave	Alta	No	Si
5	Lluvioso	Fría	Normal	No	Si
6	Lluvioso	Fría	Normal	Si	No
7	Nublado	Fría	Normal	Si	Si
8	Soleado	Suave	Alta	No	No
9	Soleado	Fría	Normal	No	Si
10	Lluvioso	Suave	Normal	No	Si
11	Soleado	Suave	Normal	Si	Si
12	Nublado	Suave	Alta	Si	Si
13	Nublado	Caliente	Normal	No	Si
14	Lluvioso	Suave	Alta	Si	No

Aspectos de diseño

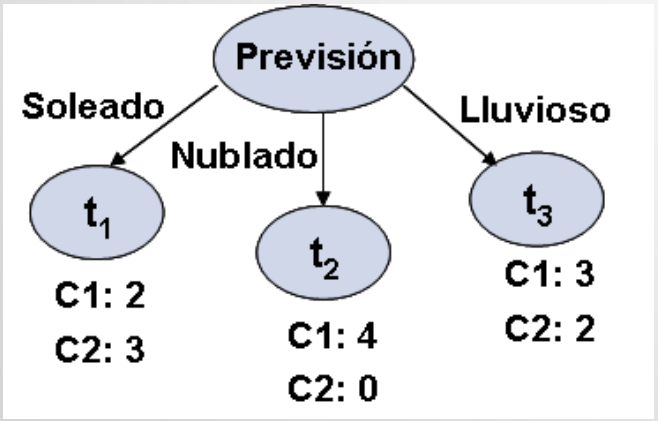


SI = C1: 9

NO = C2: 5

1. Determinar el test en el nodo raíz.

Posibilidades:



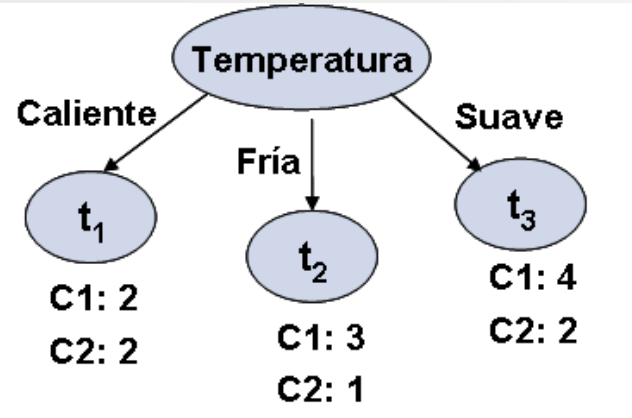
GINI (Soleado) = 1 - $(2/5)^2$ – $(3/5)^2$ = 0.4800

GINI (Nublado) = 1 - $(4/4)^2$ – $(0/4)^2$ = 0.0000

GINI (Lluvioso) = 1 - $(3/5)^2$ – $(2/5)^2$ = 0.4800

GINI (Previsión) = (5/14)(0.4800) + (4/14)(0.0000) + (5/14)(0.4800) = 0.3429

Aspectos de diseño

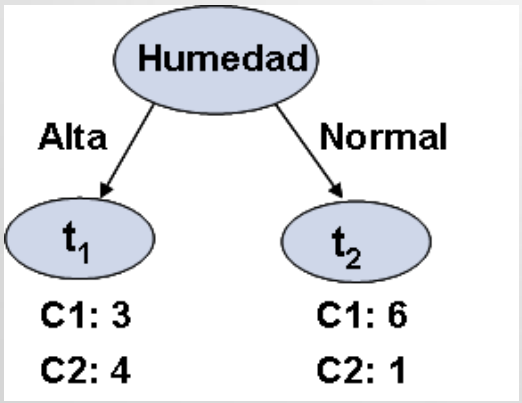


GINI (Caliente) = 1 - $(2/4)^2$ – $(2/4)^2$ = 0.5000

GINI (Fría) = 1 - $(3/4)^2$ – $(1/4)^2$ = 0.3750

GINI (Suave) = 1 - $(4/6)^2$ – $(2/6)^2$ = 0.4444

GINI (Temperatura) = (4/14)(0.5000) + (4/14)(0.3750) + (6/14)(0.4444) = 0.4405

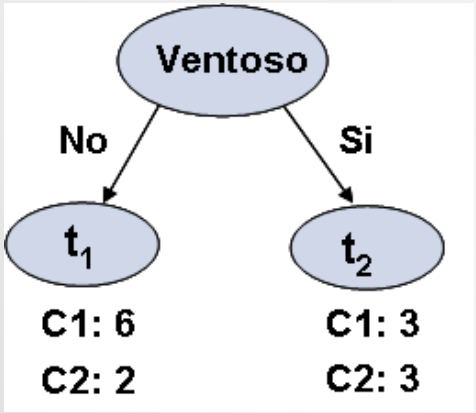


GINI (Alta) = 1 - $(3/7)^2$ – $(4/7)^2$ = 0.4898

GINI (Normal) = 1 - $(6/7)^2$ – $(1/7)^2$ = 0.2449

GINI (Humedad) = (7/14)(0.4898) + (7/14)(0.2449) = 0.3674

Aspectos de diseño



GINI (No) = 1 - $(6/8)^2$ – $(2/5)^2$ = 0.3750

GINI (Si) = 1 - $(3/6)^2$ – $(3/6)^2$ = 0.5000

GINI (Ventoso) = (8/14)(0.3750) + (6/14)(0.5000) = 0.4286

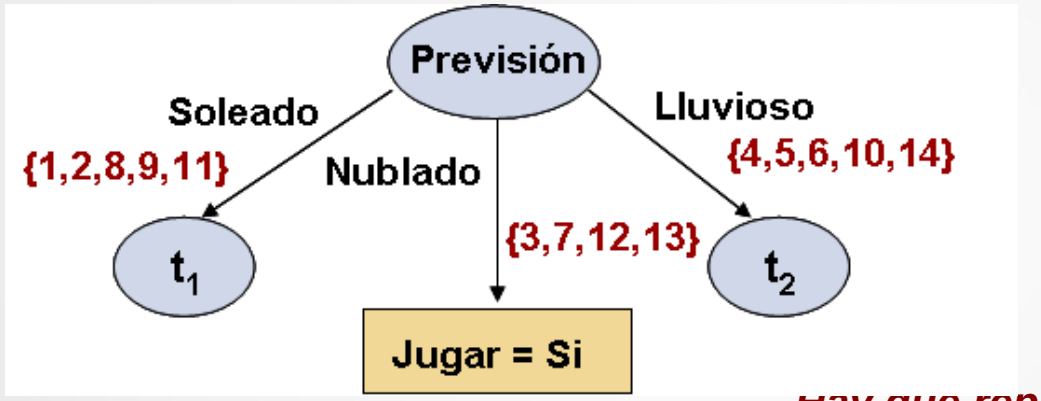
GINI (Previsión) = 0.3429 ← ***Seleccionado***

GINI (Temperatura) = 0.4405

GINI (Humedad) = 0.3674

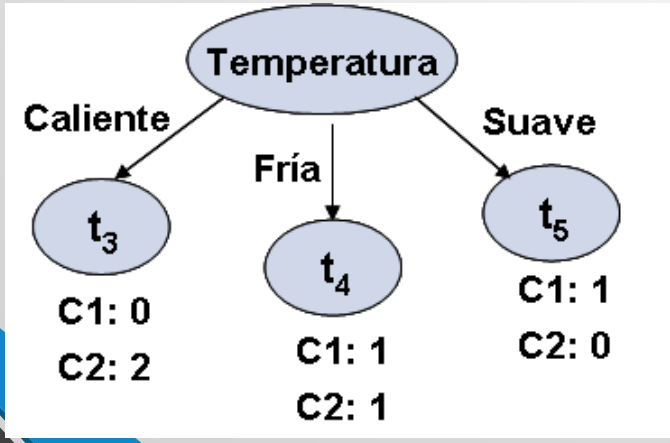
GINI (Ventoso) = 0.4286

Aspectos de diseño



Hay que repetir para los nodos t_1 y t_2

2. Determinar el test en el nodo t_1:



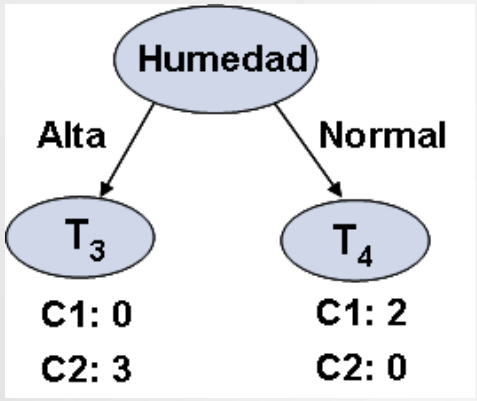
GINI (Caliente) = 1 - $(0/2)^2$ – $(2/2)^2$ = 0.0000

GINI (Fría) = 1 - $(1/2)^2$ – $(1/2)^2$ = 0.5000

GINI (Suave) = 1 - $(1/2)^2$ – $(0/2)^2$ = 0.0000

GINI (Temperatura) = (2/5)(0.0000) + (2/5)(0.5000) + (1/5)(0.0000) = 0.2000

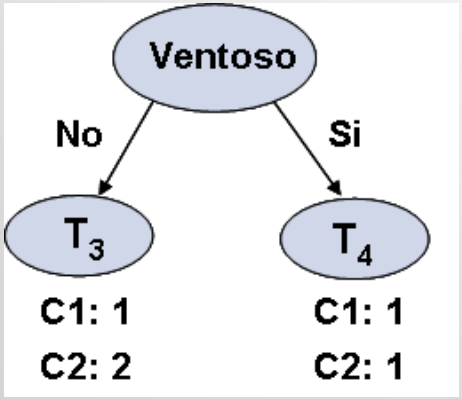
Aspectos de diseño



GINI (Alta) = 1 - $(0/3)^2$ – $(3/3)^2$ = 0.0000

GINI (Normal) = 1 - $(2/2)^2$ – $(0/2)^2$ = 0.0000

GINI (Humedad) = (3/5)(0.0000) + (2/5)(0.0000) = 0.0000



GINI (No) = 1 - $(1/3)^2$ – $(2/3)^2$ = 0.4444

GINI (Si) = 1 - $(1/2)^2$ – $(1/2)^2$ = 0.5000

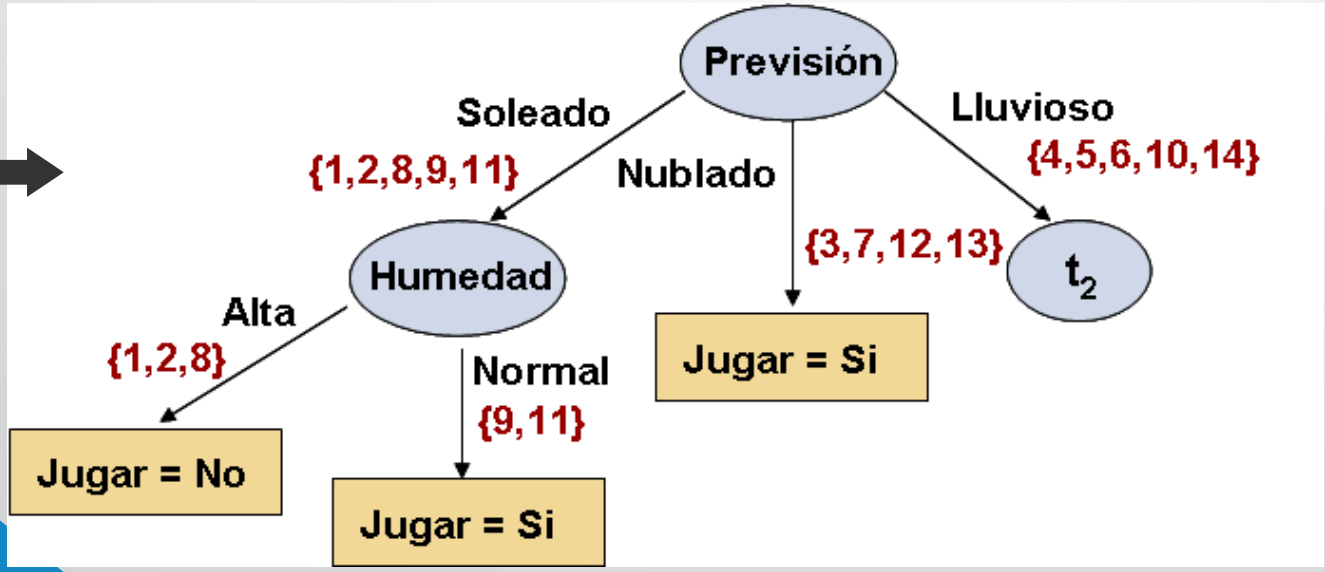
GINI (Ventoso) = (3/5)(0.4444) + (2/5)(0.5000) = 0.4666

Aspectos de diseño

GINI (Temperatura) = 0.2000

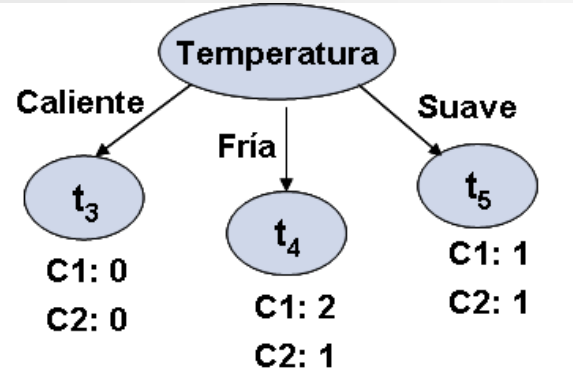
GINI (Humedad) = 0.0000 ← ***Seleccionado***

GINI (Ventoso) = 0.4666



Aspectos de diseño

3. Determinar el test en el nodo t_2:

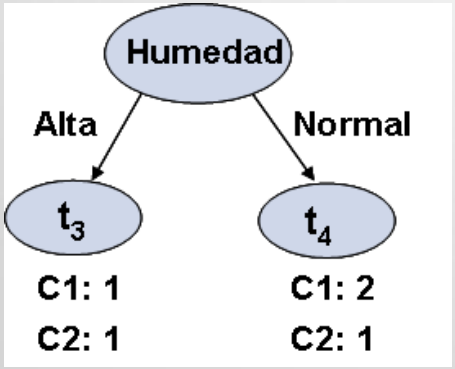


GINI (Caliente) = 1

GINI (Fría) = 1 - $(2/3)^2$ – $(1/3)^2$ = 0.4444

GINI (Suave) = 1 - $(1/2)^2$ – $(1/2)^2$ = 0.5000

GINI (Temperatura) = (0)(1) + (3/5)(0.4444) + (2/5)(0.5000) = 4666

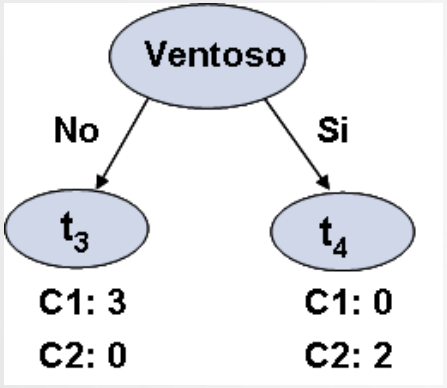


GINI (Alta) = 1 - $(1/2)^2$ – $(1/2)^2$ = 0.5000

GINI (Normal) = 1 - $(2/3)^2$ – $(1/3)^2$ = 0.4444

GINI (Humedad) = (2/5)(0.5000) + (3/5)(0.4444) = 0.4666

Aspectos de diseño



GINI (No) = 1 - $(3/3)^2$ – $(0/3)^2$ = 0.0000

GINI (Si) = 1 - $(0/2)^2$ – $(2/2)^2$ = 0.0000

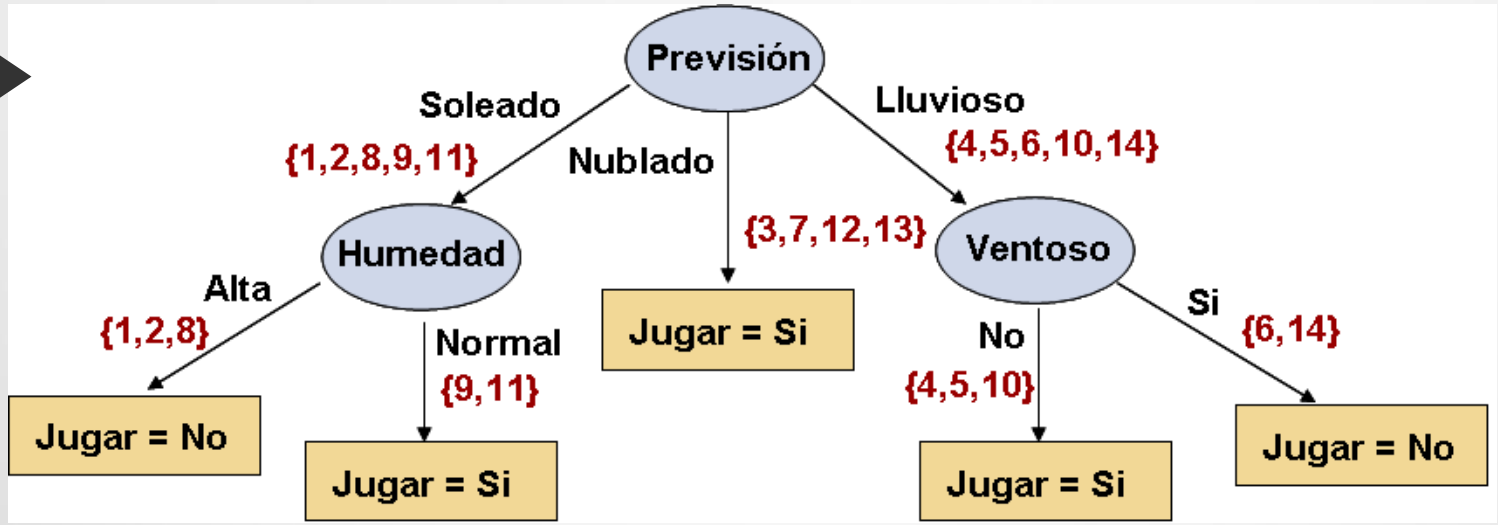
GINI (Ventoso) = (3/5)(0.0000) + (2/5)(0.0000) = 0.0000

GINI (Temperatura) = 0.4666

GINI (Humedad) = 0.4666

GINI (Ventoso) = 0.0000 ← ***Seleccionado***

Aspectos de diseño



Reglas:

1. **Si (Previsión es Soleado) y (Humedad es Normal) entonces Jugar es Si**
2. **Si (Previsión es Nublado) entonces Jugar es Si**
3. **Si (Previsión es Lluvioso) y (Ventoso es No) entonces Jugar es Si**
4. **Si (Previsión es Soleado) y (Humedad es Alta) entonces Jugar es No**
5. **Si (Previsión es Lluvioso) y (Ventoso es Si) entonces Jugar es No**