Minería de Datos

Técnicas de Minería de datos – árboles de decisión

Análisis exploratorio de datos

- **Investigación preliminar de los datos con el objetivo de entender sus características.**
- **Serie de técnicas para investigar los datos con el objetivo de revelar tendencias, características, posibles errores, correlaciones.**

➡

- ☞ **Investigar las variables**
- ☞ **Visualizar las variables**
- ☞ **Examinar distribuciones**
- ☞ **Explorar relaciones entre conjuntos de variables**

Importante: no incluye modelos complejos ni predictivos.

Análisis exploratorio de datos

Según CRISP_DM:

Comprensión de los datos

Familiarización con los datos tomando en cuenta los objetivos del negocio

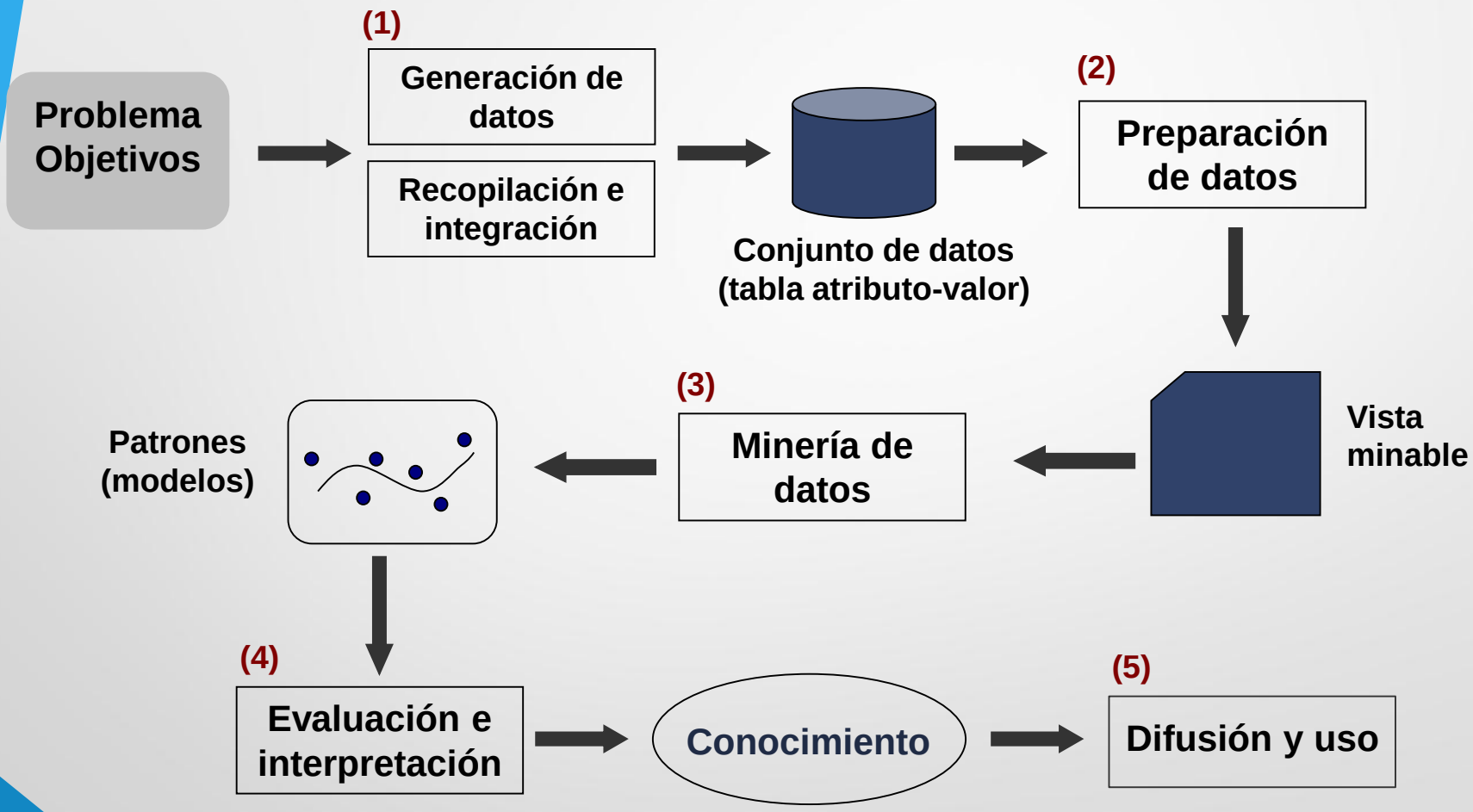
- ***Recopilación inicial de los datos***
- ***Descripción de los datos***
- ***Exploración de los datos***
- ***Verificación de la calidad de los datos***

Preparación de los datos

Se obtiene la vista minable

- ***Selección de los datos***
- ***Limpieza de datos***
- ***Construcción de datos***
- ***Integración de datos***
- ***Formateo de datos***

Proceso de minería de datos

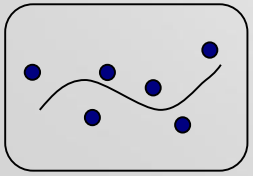


Técnicas de minería de datos

Vista minable

(3)

Minería de datos



Patrones (modelos)

- Determinar la tarea de minería de datos
  - Predictiva
    - Clasificación
    - Regresión
  - Descriptiva
    - Clustering
    - Asociación, entre otras
- Seleccionar el lenguaje de representación
- Seleccionar el algoritmo (técnica) que resuelva la tarea y que genera el tipo de patrón que se está buscando

¿Cuáles técnicas utilizar?

Técnicas de minería de datos

Algunas técnicas:

Técnicas	Tareas			
	Clasificación	Regresión	Agrupación	Asociación
Árboles de decisión	✓	✓		
Reglas de cobertura	✓			
Regresión lineal		✓		
K-medias			✓	
Apriori				✓
Redes bayesianas	✓			✓
K-vecinos	✓	✓		
Redes neuronales	✓	✓	✓	
SVM	✓	✓		
Algoritmos genéticos	✓	✓	✓	✓

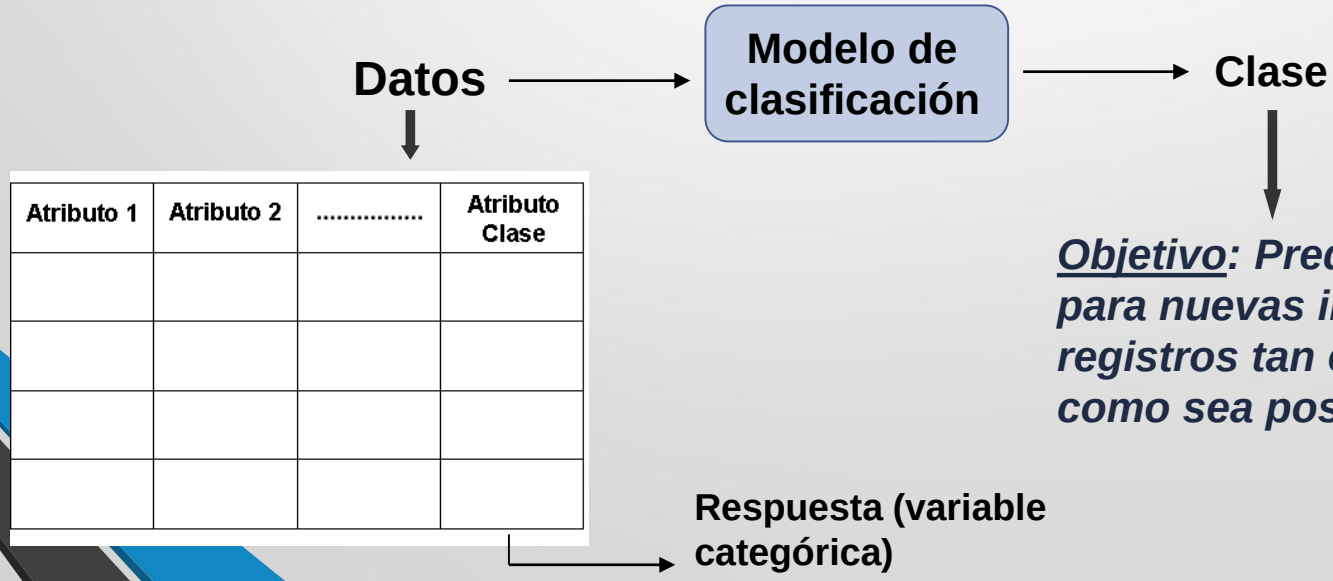
Técnicas de minería de datos - Clasificación

Clasificación:

- Dada una colección de registros o instancias (*conjunto de aprendizaje*)

Donde cada registro contiene un conjunto de atributos, uno de los cuales es la clase (cada instancia pertenece a una clase)

- Encontrar un modelo para el atributo clase como una función de los valores de los otros atributos.



Objetivo: Predecir la clase para nuevas instancias o registros tan exactamente como sea posible.

Técnicas de minería de datos - Clasificación

Ejemplos de aplicaciones:

- **Determinar la categoría de un Servicio Web.**
- **Determinar las fallas de un servicio de telefonía móvil a partir de los registros de los usuarios**
- **Identificación de reglas del mercado de valores a partir de datos históricos**
- **Identificar los pacientes con riesgo de sufrir una patología concreta, a partir de los registros médicos**
- **Determinar la capacidad de pagos de crédito de una compañía a partir de variables financieras.**
- **Determinar la aptitud física de la tierra para un determinado cultivo.**

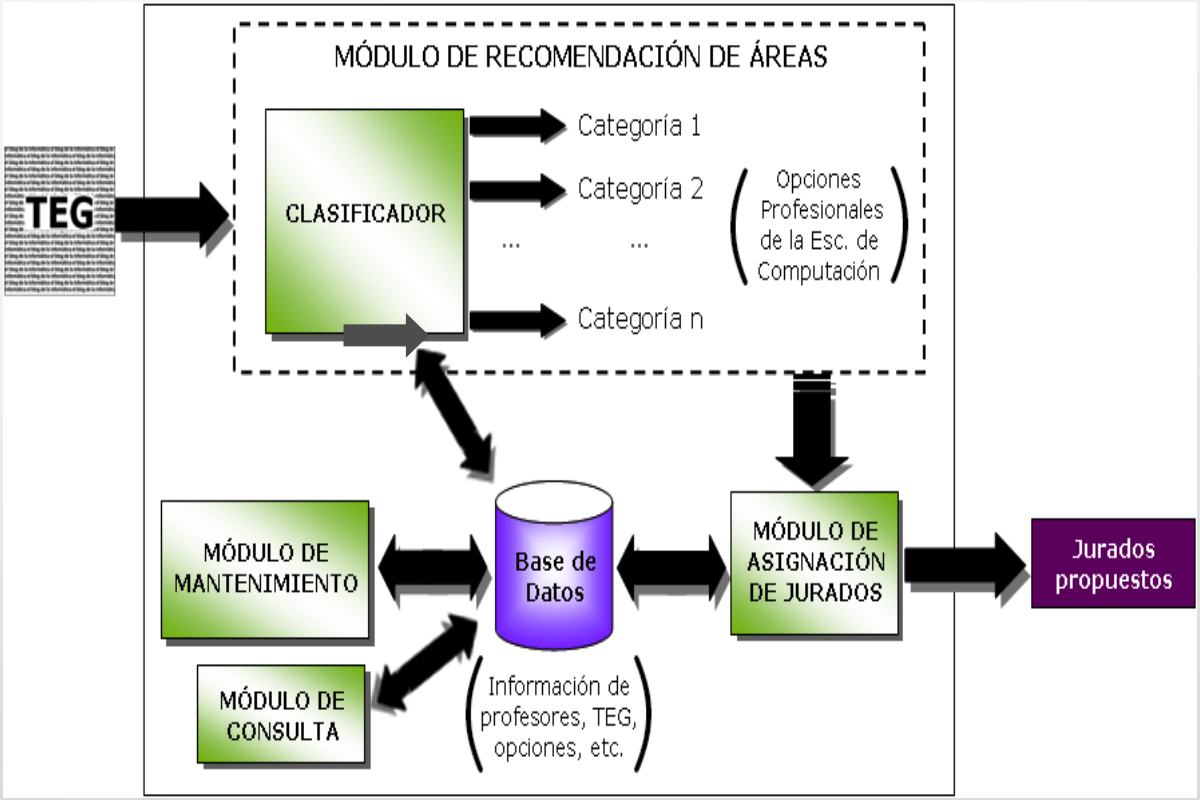
Técnicas de minería de datos - Clasificación

- **Categorizar los documentos según su contenido temático**

Documentos de TEG (título, resumen, palabras claves. opción)

Minería de datos

Modelo de clasificación de textos

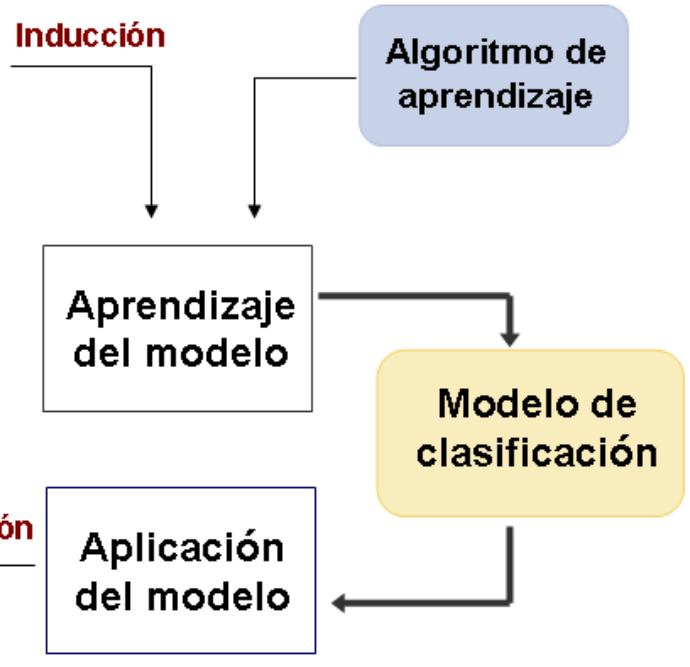


Técnicas de minería de datos - Clasificación

Enfoque general para resolver un problema de clasificación:

Datos (conjunto de aprendizaje)

ID	Atributo1	Atributo2	Atributo3	Clase
1	SI	Largo	125	A
2	NO	Medio	100	B
3	NO	Corto	70	B
⋮				
N	SI	Medio	85	A



Nuevas instancias

ID	Atributo1	Atributo2	Atributo3	Clase
N+1	NO	Corto	55	?
N+2	SI	Medio	80	?
N+3	NO	Largo	110	?

Técnicas de minería de datos - Clasificación

¿Cómo evaluar un clasificador?

Importante: Determinar el rendimiento del modelo

- **La evaluación del rendimiento de un clasificador se basa en contar los registros o instancias predichos correcta e incorrectamente**

- **Estos números pueden ser tabulados en una tabla conocida como *matriz de confusión.***

Técnicas de minería de datos - Clasificación

Ejemplo para dos clases:

		Clase predicha	
		Clase = 1	Clase = 0
Clase actual	Clase = 1	F_{11}	F_{10}
	Clase = 0	F_{01}	F_{00}

Donde F_{ij} = Número de instancias de la clase i predichas como clase j

➡ **No. de predicciones correctas = $F_{11} + F_{00}$**
No. de predicciones incorrectas = $F_{10} + F_{01}$

Técnicas de minería de datos - Clasificación

Aunque la matriz de confusión suministra la información necesaria para determinar el rendimiento de un clasificador, resumir esta información en único número es más conveniente para comparar el rendimiento de diferentes modelos.

➡ ***Métrica de rendimiento o índice (medida de evaluación)***

$$\text{Exactitud} = \frac{\text{No. de predicciones correctas}}{\text{No. total de predicciones}} = \frac{F_{11} + F_{00}}{F_{11} + F_{00} + F_{10} + F_{01}}$$

$$\text{Error} = \frac{\text{No. de predicciones incorrectas}}{\text{No. total de predicciones}} = \frac{F_{10} + F_{10}}{F_{11} + F_{00} + F_{10} + F_{01}}$$

Los algoritmos de clasificación buscan modelos que alcancen la más alta exactitud o el más bajo error

Técnicas de minería de datos - Clasificación

Recordar:

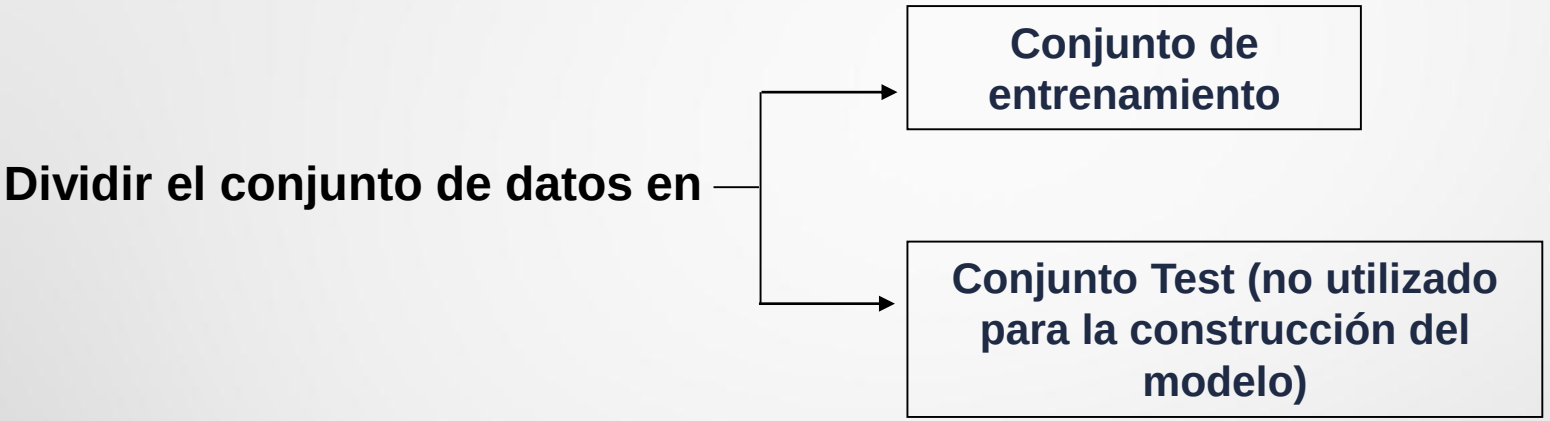
Población

Muestra = conjunto de aprendizaje

Que permita hacer predicciones, dar respuestas, explicaciones, VÁLIDAS, para la población

Técnicas de minería de datos - Clasificación

El rendimiento sobre el conjunto de datos no es un buen indicador



- **Para la evaluación será necesario entonces:**

☞ **Una medida de evaluación (error de clasificación, exactitud predictiva, entre otras)**

☞ **Una técnica de evaluación: Para obtener un estimado de la medida de evaluación sobre datos no vistos durante el aprendizaje (= *capacidad de generalización del modelo*)**

Técnicas de minería de datos - Clasificación

- **Técnicas de evaluación.**

a) Selección aleatoria:

- **Se divide el conjunto de datos D en dos conjuntos de manera aleatoria: uno para el entrenamiento (70 -80 %), y otro para el test (20-30%).**
- **Se construye un modelo utilizando el conjunto de entrenamiento y se determina el valor de la medida de evaluación sobre el conjunto test.**
- **Este proceso se repite varias veces. Luego se promedia sobre todos los valores.**
- **Funciona bien para muestras grandes**

b) Validación cruzada:

- **Se divide el conjunto de datos D en k particiones disjuntas (Di, i =1.. k) de manera aleatoria.**

  Desde i = 1 hasta k

  - Entrenar con D – D_i

  - Determinar el valor de la medida de evaluación sobre D_i (= me_i)

  Fin

- **Determinar el valor promedio: ME_{Final} =**
- **Actualmente el más utilizado**

$$\frac{1}{K}\sum_{i=1}^{k} me_i$$

Técnicas de minería de datos - Clasificación

¿Cómo determinar un modelo de clasificación?

Algunas técnicas:

- **Métodos basados en árboles de decisión**
- **Métodos basados en reglas de cobertura**
- **Métodos basados en vecindad (k-vecinos)**
- **Redes bayesianas**
- **Redes neuronales**
- **Máquinas de soporte vectorial**
- **Métodos estadísticos (discriminantes lineales)**
- **Otros**

Clasificación con árboles de decisión

¿Cómo resolver un problema de clasificación?

- **Se puede realizar en secuencia una serie de preguntas.**
- **Cada vez que se obtiene una respuesta, se hace la siguiente pregunta y así sucesivamente hasta alcanzar la etiqueta de clase.**
- **Las preguntas con sus respuestas van particionando el espacio de entrada**

La serie de preguntas y sus posibles respuestas pueden organizarse en la forma de un *árbol de decisión*

Estructura jerárquica consistente de nodos y arcos dirigidos

Clasificación con árboles de decisión

En el árbol se pueden distinguir 3 tipos de nodos:

Nodo raíz: No tiene arcos de entrada y puede o no tener arcos de salida.

Nodos internos: con un arco de entrada y varios arcos de salida

Estos nodos están asociados a condiciones o tests de atributos que se utilizan para separar registros que tienen características similares

Nodos hoja: Tiene un solo arco de entrada y ninguno de salida

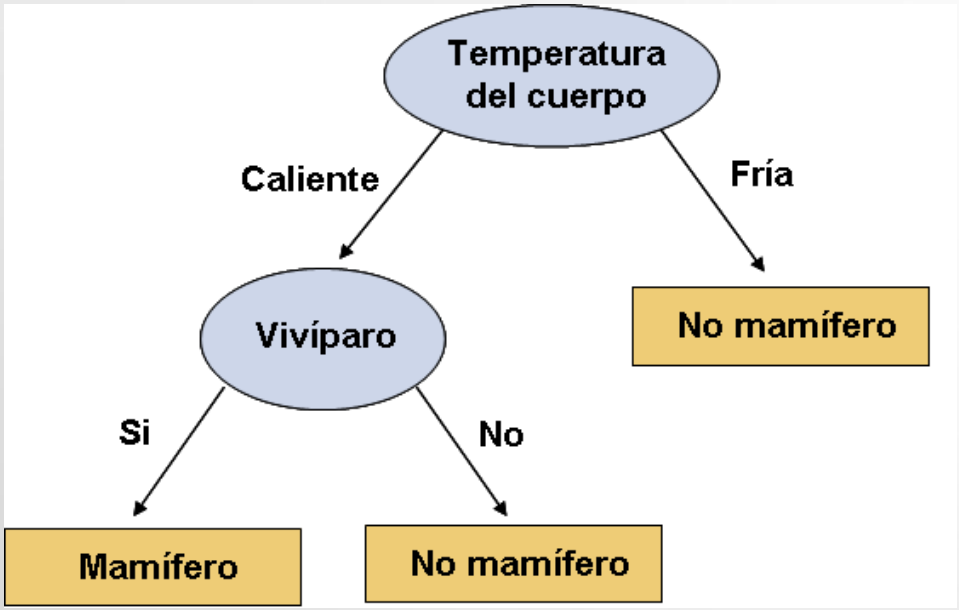
A estos nodos se les asigna una etiqueta de clase

¿Cómo clasificar un registro?

- **Empezando desde el nodo raíz, se aplica el test al registro y se sigue la rama adecuada dependiendo del resultado**
- **Esto puede conducir a un nodo interno (en cuyo caso se aplica otro test) o a un nodo hoja (se asigna al registro la etiqueta de clase del nodo)**

Clasificación con árboles de decisión

***Ejemplo*: clasificación de mamíferos**



Modelo descriptivo que sumariza los datos (herramienta explicativa).
Modelo predictivo que puede predecir a clase de ejemplos desconocidos (herramienta de predicción)

Para un nueva instancia:

Nombre	Temperatura del cuerpo	Cubierta de piel	Vivíparo	Acuática	Aérea	Patas	Hiberna	CLASE
Flamingo	Caliente	Plumas	No	No	Si	Si	No	?

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

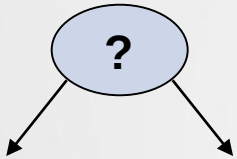
Algoritmo de Hunt

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

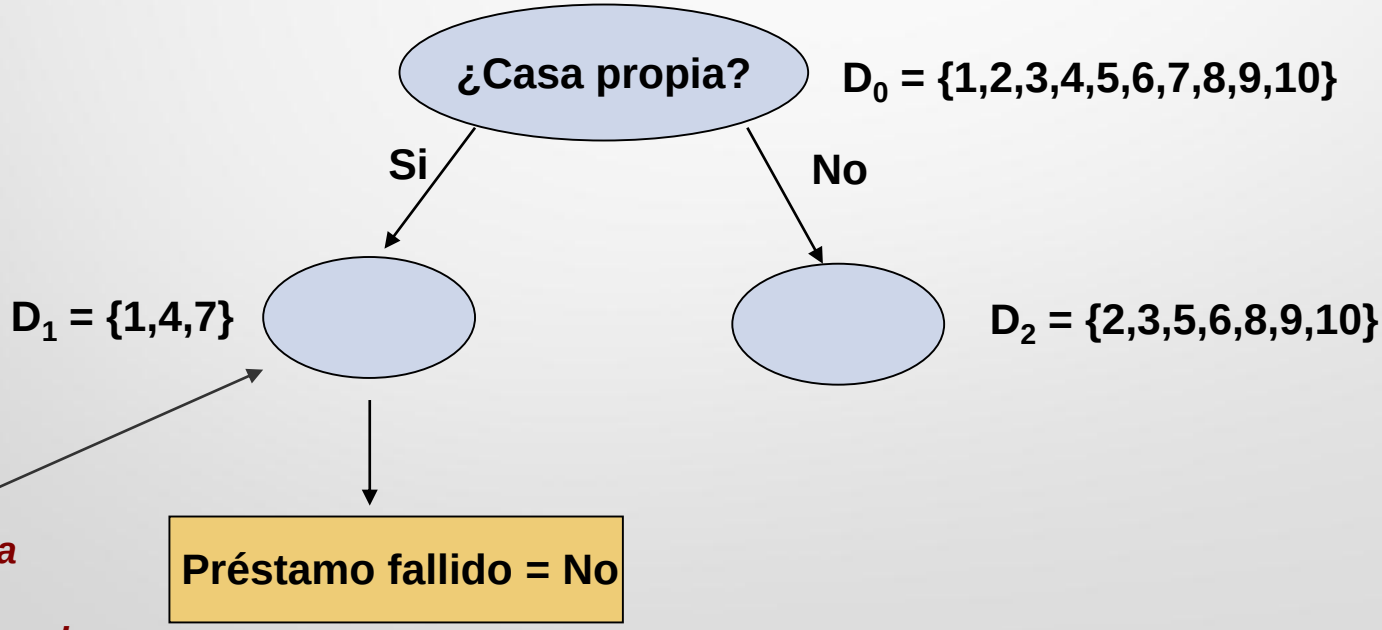
Algoritmo de Hunt

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

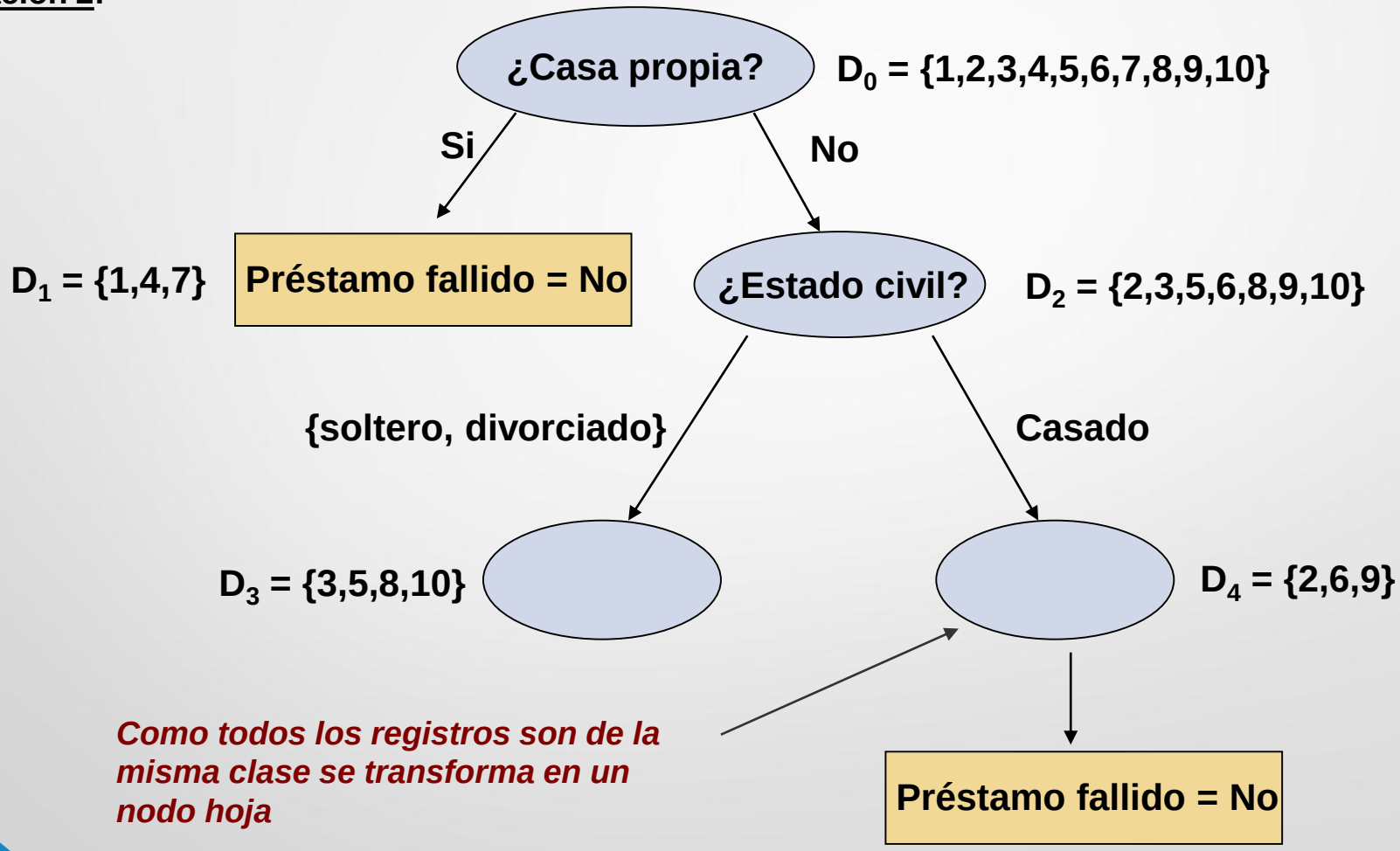
Como todos los registros son de la misma clase se transforma en un nodo hoja

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:



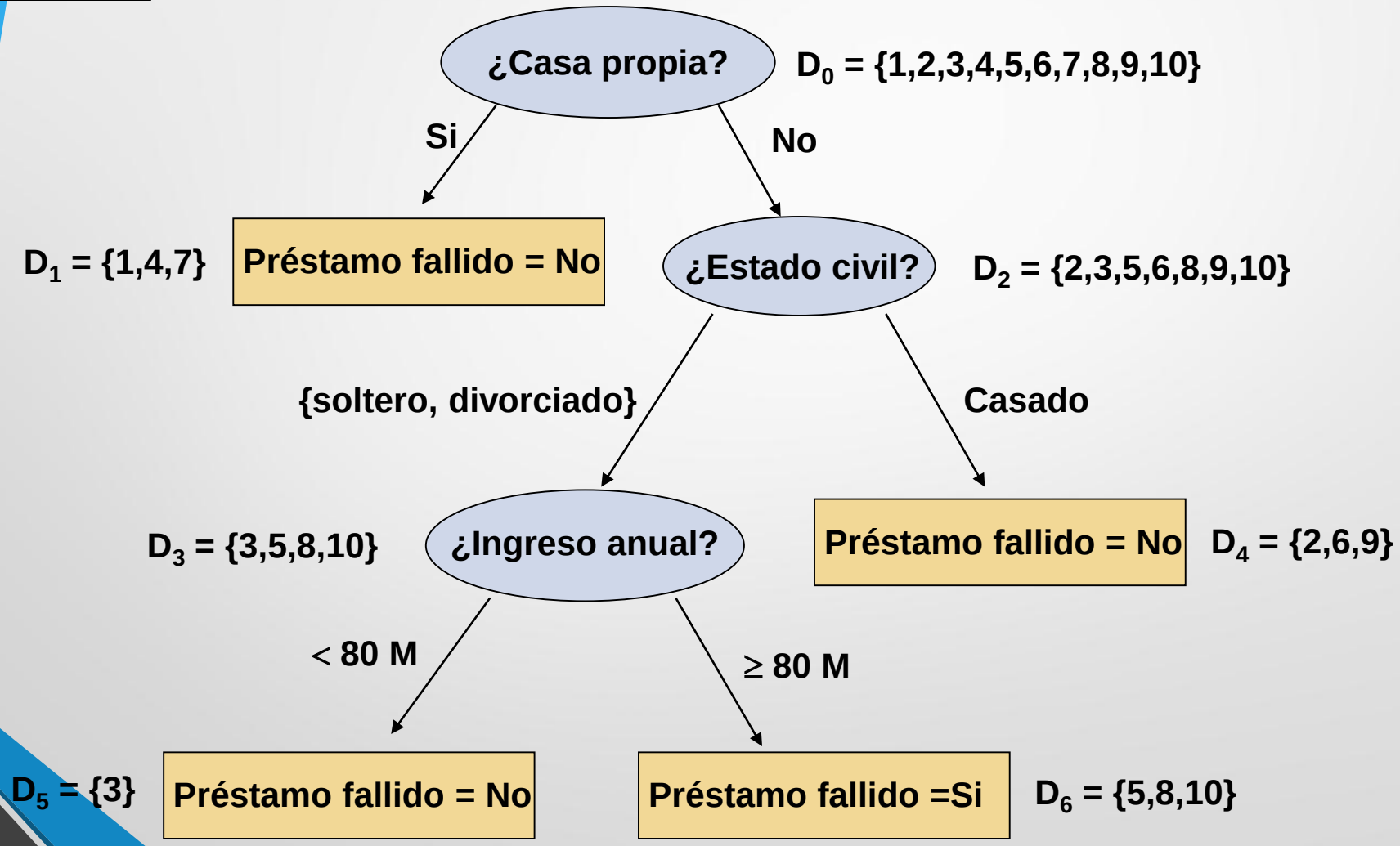
Como todos los registros son de la misma clase se transforma en un nodo hoja

Algoritmo de Hunt

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 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***