Instituto Politécnico Nacional
Escuela Superior de Cómputo
Secretaría Académica
Departamento de Ingeniería en Sistemas Computacionales

# Minería de datos (*Data Mining*)
# Árboles de decisión-1ª Parte



Profesora: *Dra. Fabiola Ocampo Botello*

# Clasificación

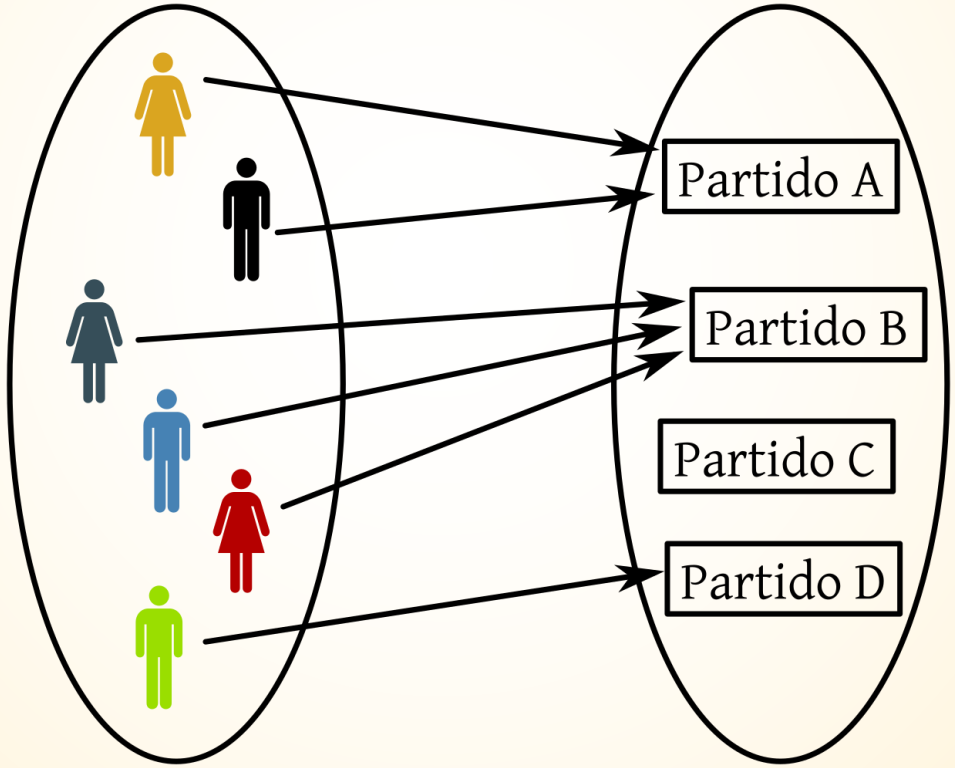
Han, Kamber & Pei (2012) establecen que la clasificación es una forma de analizar datos para generar modelos que describen importantes clases de datos. Estos modelos se llaman clasificadores, permiten predecir etiquetas de clases categóricas (discretas, desordenadas).

# Clasificación (*Continuación*)




Por ejemplo, se puede construir un modelo de clasificación para categorizar las solicitudes de préstamos bancarios como seguras o riesgosas. La clasificación tiene numerosas aplicaciones, incluida la detección de fraudes, el marketing de objetivos, la predicción del rendimiento, la fabricación y el diagnóstico médico (Han, Kamber & Pei, 2012).

**Definición de clasificación**: Es la tarea de aprendizaje que considera una función *f* que asocia cada conjunto de atributos *x* a una de las clases predefinidas y etiquetas en *y*. (Tan, Steinbach, Karpatne & Kumar, 2005).



Esta función f es conocida de manera informal como modelo de clasificación. Un modelo de clasificación es útil por las siguientes razones (Tan, Steinbach, Karpatne & Kumar, 2005).

Modelo descriptivo. Un modelo de clasificación puede servir como una herramienta para distinguir objetos de diferentes clases.

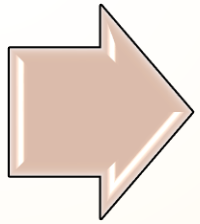
Modelo de predicción. Un modelo de clasificación puede servir para predecir la etiqueta de clase de un registro desconocido.

*Las técnicas de clasificación son más adecuadas para predecir o describir conjuntos de datos de categorías binarias o nominales. Son menos efectivas en categorías ordinales porque no consideran el orden jerárquico de los grupos.*

# Enfoque de la clasificación

La clasificación de datos es un proceso de dos pasos (Han, Kamber & Pei, 2012):

El primer paso (aprendizaje), se construye un modelo de clasificación



Un segundo paso (clasificación), en el cual el modelo se usa para predecir etiquetas de clase para otros datos.

# Primer paso. Aprendizaje.

(Han, Kamber & Pei, 2012)

## Este es el paso de aprendizaje (o fase de entrenamiento)

Un algoritmo de clasificación construye el clasificador analizando o "aprendiendo de" un conjunto de entrenamiento compuesto por tuplas de base de datos y sus etiquetas de clase asociadas.

Una tupla, X, está representada por un vector de atributos de n dimensiones,

$$X = \{x1, x2, \ldots, xn\},$$

que representa las n mediciones realizadas en la tupla que contiene n atributos de la base de datos, respectivamente, A1, A2, …, An.

Cada tupla, X, pertenece a una clase predefinida determinada por otro atributo de base de datos denominado atributo de etiqueta de clase.

El atributo de etiqueta de clase tiene un valor discreto y no está ordenado. Es categórico (o nominal) en el sentido de que cada valor sirve como categoría o clase.

# Primer paso. Aprendizaje (*continúa*).

(Han, Kamber & Pei, 2012)

- Debido a que se proporciona la etiqueta de clase de cada tupla de entrenamiento, este paso también se conoce como **aprendizaje supervisado**.
- Supervisado expresa, que la clasificación se "supervisa“, esto es, se le dice a qué clase pertenece cada tupla de entrenamiento.
- Contrasta con el **aprendizaje no supervisado (o agrupamiento**), en el que *no se conoce la etiqueta de clase de cada tupla de entrenamiento y es posible que no se conozca de antemano el número o conjunto de clases que se deben aprender*.

# Primer paso. Aprendizaje (*continúa*).

(Han, Kamber & Pei, 2012)

Este primer paso del proceso de clasificación también puede verse como el aprendizaje de un mapeo o función,

$$y = f(X)$$

que puede predecir la etiqueta de clase asociada *y* de una tupla X dada.

En esta vista, se desea aprender un mapeo o función que separa las clases de datos.

Normalmente, *este mapeo se representa en forma de reglas de clasificación, árboles de decisión o fórmulas matemáticas*.

Las reglas se pueden utilizar para categorizar futuras tuplas de datos, así como para proporcionar una visión más profunda del contenido de los datos.

# Segundo paso. Clasificación.



(Han, Kamber & Pei, 2012)

En el segundo paso se considera:

- El modelo se utiliza para la clasificación.
- Primero, se estima la precisión predictiva del clasificador.
- *Si se usara el conjunto de entrenamiento* para medir la precisión del clasificador, esta estimación probablemente sería optimista.
- Por lo tanto, se utiliza un conjunto de prueba, formado por tuplas de prueba y sus etiquetas de clase asociadas. Son independientes de las tuplas de entrenamiento, lo que significa que no se utilizaron para construir el clasificador.

# Segundo paso. Clasificación (*continuación*).

(Han, Kamber & Pei, 2012)

- La precisión de un clasificador en un conjunto de prueba dado es el porcentaje de tuplas de conjuntos de prueba que el clasificador clasifica correctamente.

- La etiqueta de clase asociada de cada tupla de prueba se compara con la predicción de clase del clasificador aprendido para esa tupla.

- Si la precisión del clasificador se considera aceptable, el clasificador se puede utilizar para clasificar tuplas de datos futuras para las que no se conoce la etiqueta de clase. (Estos datos también se denominan en la literatura sobre aprendizaje automático (*machine learning*) como datos "desconocidos" o "no vistos anteriormente").

# Clasificación

La clasificación se refiere a la tarea de asignar objetos a una de varias categorías predefinidas.

La entrada de datos para la clasificación se compone de una serie de registros, donde cada registro representa una instancia y se caracteriza por ser una tupla (x, *y*) donde x es el conjunto de atributos y *y* es un atributo especial, la etiqueta de la clase.

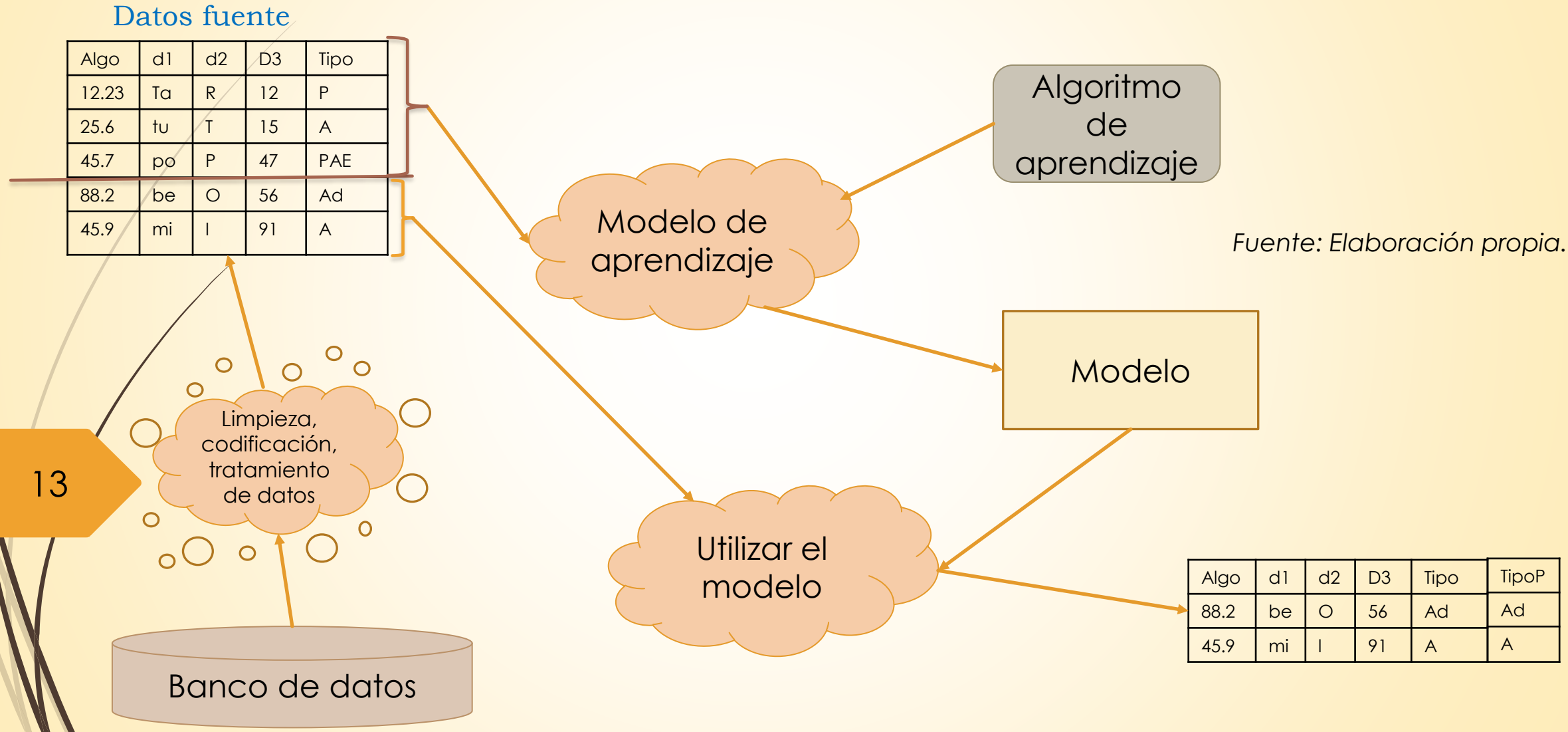


Por ejemplo, suponga que se tiene la clase persona* de la Escuela: profesor, alumnos, paae y administradores. El conjunto de atributos (x) contiene los datos identificados de las personas y la variable *y* es de tipo discreta que representa las diversas clases o categorías que puede tener *x*.

* Imagine una clase disjunta, total. En el que todos las subclases tienen los mismos atributos.

Los árboles de decisión son una de las técnicas de clasificación.

Proceso para construir un modelo de clasificación




*Fuente: Elaboración propia.*

Rokach, L. & Maimon, O. (2015) establecen que la meta de un algoritmo de clasificación se puede definir formalmente como:

Dado un conjunto de entrenamiento S con atributos de entrada

$$A = \{a1, a2, \ldots, an\}$$

y un atributo nominal *y* y una distribución desconocida D, la meta es inducir un clasificador óptimo con el mínimo error de generalización.

Notación:

| | |
|---|---|
| DT | Representa el inductor del árbol de decisión. |
| DT(S) | Representa un árbol de clasificación que se generó al ejecutar DT sobre el conjunto de datos S. |
| $DT(S)(x_q)$ | Es la predicción de $x_q$ usando DT(S). |

La evaluación del desempeño de un modelo de clasificación considera dos aspectos:

1. La cantidad de registros previstos por el modelo de forma adecuada.
2. La cantidad de registros previstos por el modelo de forma inadecuada.

Lo anterior se presenta en una **matriz de confusión**.

Ejemplo de una matriz de confusión:

| | | Clase prevista | |
|---|---|---|---|
| | | Clase 1 | Clase 0 |
| Clase actual | Clase 1 | f11 | f10 |
| | Clase 0 | f01 | f00 |



$$\text{Exactitud } (Accuracy) = \frac{Número\ de\ predicciones\ correctas}{Número\ total\ de\ predicciones} \qquad \text{Accuracy} = \frac{f11+f00}{f11+f10+f01+f00}$$

$$\text{Tasa de erros } (Error\ rate) = \frac{Número\ de\ predicciones\ incorrectas}{Número\ total\ de\ predicciones} \qquad \text{Error rate} = \frac{f10+f01}{f11+f10+f01+f00}$$

# Árboles de decisión

Han, Kamber & Pei (2012) establecen que **Un árbol de decisión** es una estructura de árbol similar a un diagrama de flujo, donde cada nodo interno (nodo no hoja) denota una prueba en un atributo, cada rama representa un resultado de la prueba y cada nodo hoja (o nodo terminal) tiene una etiqueta de clase.

Los nodos internos se indican con rectángulos y los nodos de hoja se indican con óvalos. Algunos algoritmos de árboles de decisión producen solo árboles binarios (donde cada nodo interno se ramifica exactamente a otros dos nodos), mientras que otros pueden producir árboles no binarios.

## Referencias bibliográficas


Han, Jiawei; Kamber, Micheline & Pei, Jian. (2012). *Data Mining: concepts and techniques*. Third edition. Morgan Kaufman Series.

Rokach, L. & Maimon, O. (2015). Data Mining with decision trees. Theory and Applications. Second Edition. World Scientific Publishing Co. Pte. Ltd.

Tan Pang-Ning, Steinbach Michael, Karpatne Anuj, Kumar Vipin. (2005). Introduction to data mining. Second Edition. Pearson