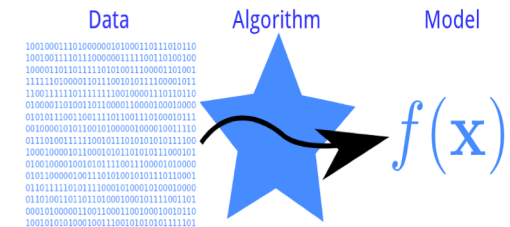


# Breve introducción al Aprendizaje Computacional

# Aprendizaje Computacional y Minería de Datos

## Aprendizaje Computacional (Machine learning)

Área de la Inteligencia Artificial cuyo objetivo es el desarrollo de algoritmos y técnicas que permitan a los computadores *aprender* de los datos.



## Minería de datos (Data Mining)

Proceso de extraer conocimiento útil y comprensible de grandes volúmenes de datos.



El conocimiento extraído debe:

- No ser trivial,
- Estar implícito en los datos,
- Ser previamente desconocido
- Ser potencialmente útil

# Terminología

- **Objetos:** casos a estudiar. Se les llama también registros, ejemplos, instancias.
- **Atributos:** propiedades de un objeto. Se le llama también variables, rasgos o características
- **Datos:** colección de objetos y sus atributos
- **Valores del atributo:** números o símbolos que pueden ser asignados a cada atributo
- Cuando el conjunto de datos incluye la *clase* a la que pertenece cada objeto se dice que los datos están **etiquetados**

4 atributos				
<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
				Clase

# Tipos de modelos (según su objetivo)

Según si se incluye la variable clase:

- Clasificación supervisada:  
Datos etiquetados (se incluye también la variable *clase*)
- Clasificación no supervisada:  
Datos no etiquetados (no se incluye la variable *clase*)

Según su objetivo:

- Modelos predictivos  
El objetivo es predecir los valores de la variable de interés (clase o variable respuesta) a partir de valores de otras variables.
- Modelos descriptivos  
El objetivo es describir el comportamiento de los datos de forma que sea interpretable

# Ejemplos

## Clasificación:

Predecir la nota de un estudiante según su actividad durante el curso

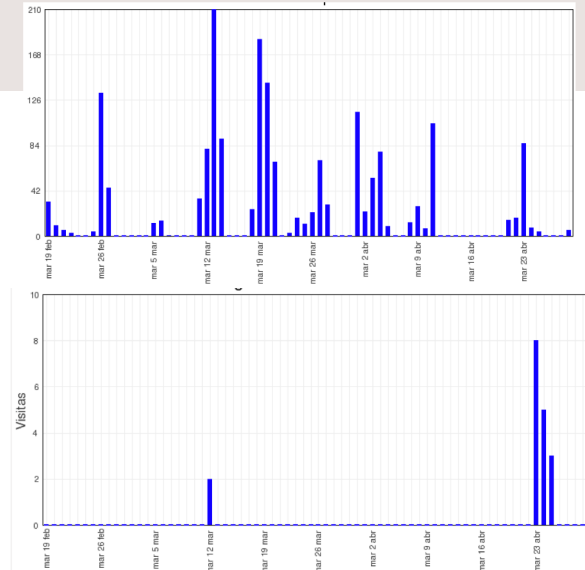
- Aprobado/suspenso (clasificación)
- Nota numérica (regresión)

## Reglas de asociación:

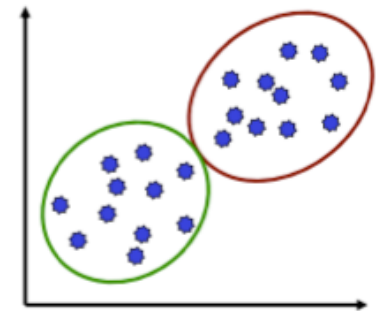
Cuando un cliente compra un producto, suele comprar (o mirar) otros también

## Clustering:

División en grupos (clusters) con características similares (minimizando la distancia dentro de los clusters y maximizando la distancia entre los clusters)



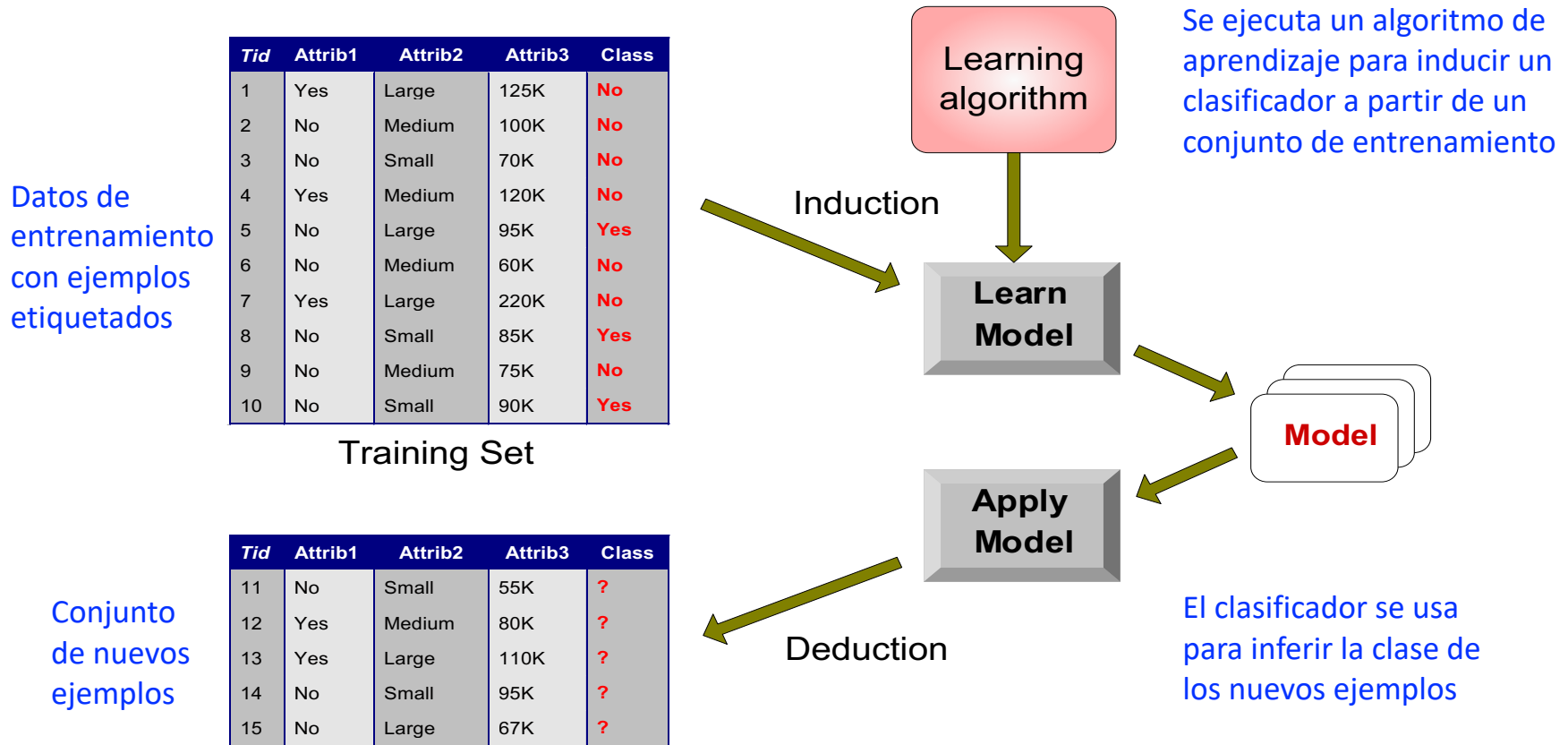
### ¿Qué otros productos compran los clientes



## Aprendizaje supervisado:

- Clasificadores basados en reglas
- Clasificadores basados en árboles de decisión

# Esquema general



# Clasificadores basados en reglas

El clasificador en este caso es un **conjunto de reglas** del tipo:

- Regla:  $(Condición) \rightarrow y$

donde:

- *condición* = conjunto de atributos
- *y* es la clase



# Ejemplo clasificador basado en reglas

## Variable clase

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
human	warm	yes	no	no	mammals
python	cold	no	no	no	reptiles
salmon	cold	no	no	yes	fishes
whale	warm	yes	no	yes	mammals
frog	cold	no	no	sometimes	amphibians
komodo	cold	no	no	no	reptiles
bat	warm	yes	yes	no	mammals
pigeon	warm	no	yes	no	birds
cat	warm	yes	no	no	mammals
leopard shark	cold	yes	no	yes	fishes
turtle	cold	no	no	sometimes	reptiles
penguin	warm	no	no	sometimes	birds
porcupine	warm	yes	no	no	mammals
eel	cold	no	no	yes	fishes
salamander	cold	no	no	sometimes	amphibians
gila monster	cold	no	no	no	reptiles
platypus	warm	no	no	no	mammals
owl	warm	no	yes	no	birds
dolphin	warm	yes	no	yes	mammals
eagle	warm	no	yes	no	birds

## Modelo aprendido:

R1: (Give Birth = no)  $\wedge$  (Can Fly = yes)  $\rightarrow$  Birds

R2: (Give Birth = no)  $\wedge$  (Live in Water = yes)  $\rightarrow$  Fishes

R3: (Give Birth = yes)  $\wedge$  (Blood Type = warm)  $\rightarrow$  Mammals

R4: (Give Birth = no)  $\wedge$  (Can Fly = no)  $\rightarrow$  Reptiles

R5: (Live in Water = sometimes)  $\rightarrow$  Amphibians

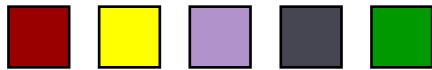
# Árboles de decisión

El clasificador en este caso será un **árbol de decisión**

## Problema de Clasificación:


$x_1, x_2$  - 2 variables predictoras

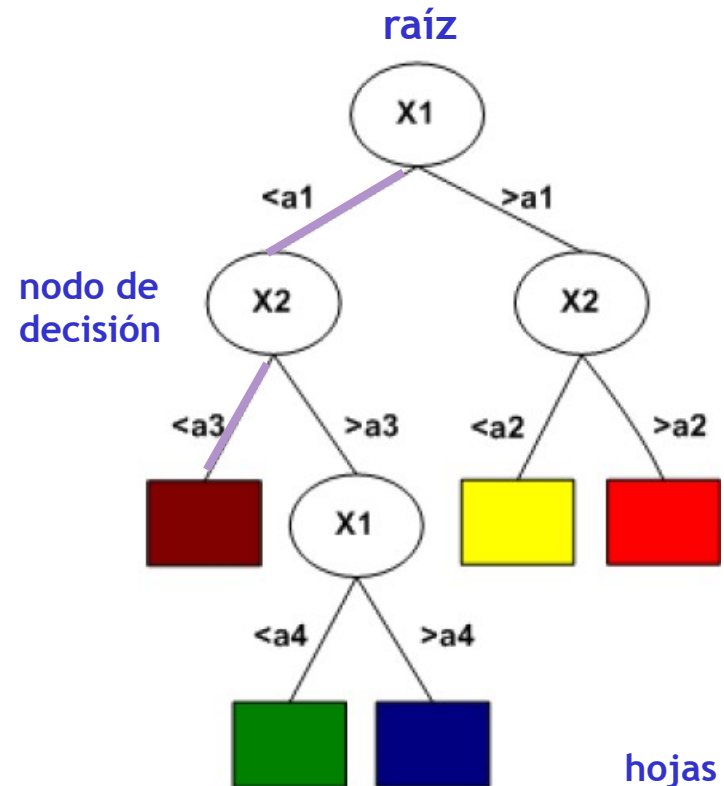
5 clases  $\in C$



$f: x_1 \times x_2 \rightarrow C$

- Cada nodo de decisión contiene un test
- Cada rama descendente corresponde a un valor posible del atributo
- Cada hoja está asociada a una clase
- Cada camino en el árbol (de la raíz a la hoja) corresponde a una regla de clasificación

IF  $x_1 < a_1$  and  $x_2 < a_3$  THEN 

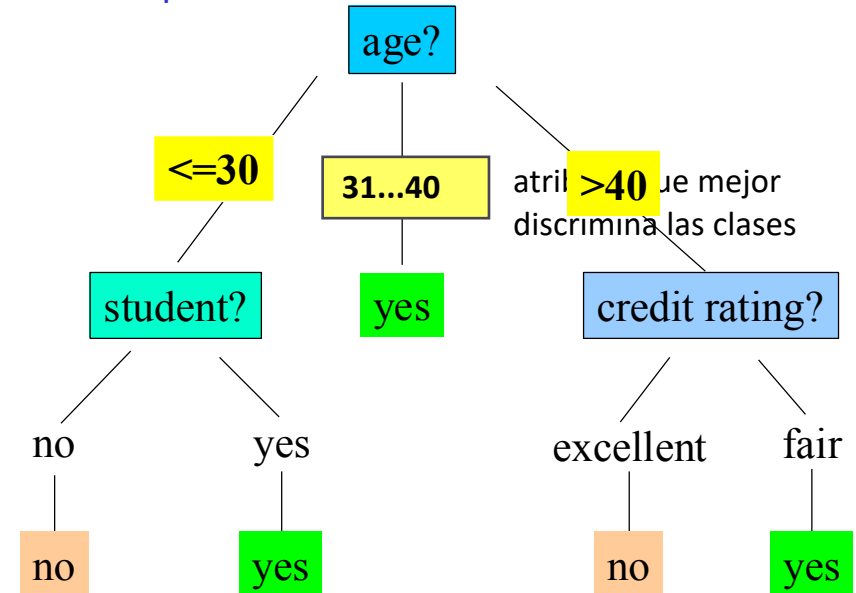


# Ejemplo de clasificador (árbol de decisión)

Variable clase

age	income	student	credit_rating	buys_comp.
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

Modelo aprendido



Nuevos ejemplos

age	income	student	credit_rating	buys_comp.
<=30	fair	no	fair	?
>40	high	no	excellent	?
31...40	low	no	fair	?

➔ Modelo aprendido ➔

buys_comp.
no
no
yes

# Aprendizaje no supervisado:

- Reglas de asociación
- Clustering

# Reglas de asociación

Dado un conjunto de datos, el objetivo es encontrar reglas que descubran hechos/valores que ocurren simultáneamente

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

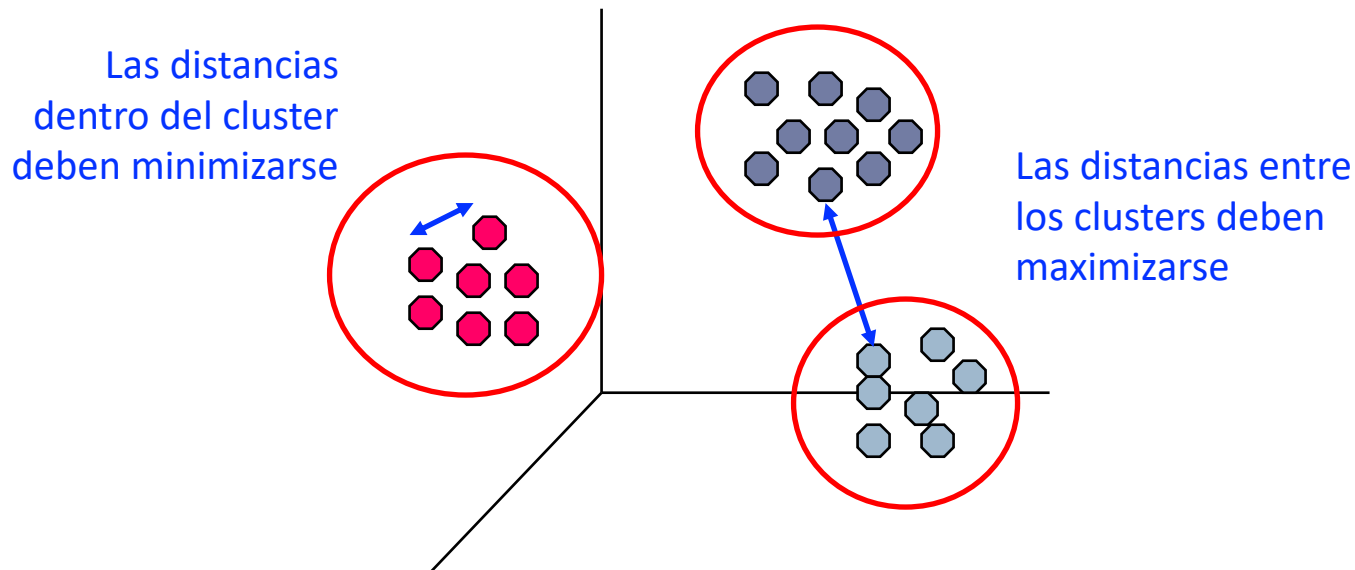
Por ejemplo, las reglas

[Milk} → Bread

[Milk, Diaper} → Beer

# Clustering

Dados un conjunto de datos, el objetivo es agruparlos en *clusters* de modo que haya una alta variabilidad entre los grupos y una alta similitud dentro de cada grupo



# Medidas de calidad

# Medidas de calidad en clasificación

Tasa de aciertos (Accuracy): % de instancias bien clasificadas

Tasa de error (Error rate): % de instancias mal clasificadas

Matriz de confusión:

- Cada columna de la matriz representa el número de predicciones en cada clase
- Cada fila representa a las instancias en la clase real

		Predicted class		
		Cat	Dog	Rabbit
Actual class	Cat	5	3	0
	Dog	2	3	1
	Rabbit	0	2	11



# Medidas de calidad para reglas de asociación

Dado un conjunto de N datos y una regla  $X \rightarrow Y$ , se denomina:

Soporte (s): proporción de ejemplos que contienen el antecedente y el consecuente de la regla con respecto al total.

$$s = \frac{\text{frecuencia}(X,Y)}{N}$$

Confianza (c): proporción de ejemplos que contienen el antecedente y el consecuente de la regla con respecto al total de ejemplos que contienen el antecedente.

$$c = \frac{\text{frecuencia}(X,Y)}{\text{frecuencia}(X)}$$

La confianza mide la fiabilidad de la regla, mientras que el soporte mide cuántos ejemplos permite clasificar

# Medidas de calidad para reglas de asociación

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

$$s = \frac{\text{Frq}(\text{Milk}, \text{Diaper}, \text{Beer})}{|N|} = \frac{2}{5} = 0.4$$

$$c = \frac{\text{Frq}(\text{Milk}, \text{Diaper}, \text{Beer})}{\text{Frq}(\text{Milk}, \text{Diaper})} = \frac{2}{3} = 0.67$$

{Milk, Diaper} → {Beer} (s=0.4, c=0.67)

{Milk, Beer} → {Diaper} (s=0.4, c=1.0)

{Diaper, Beer} → {Milk} (s=0.4, c=0.67)

{Beer} → {Milk, Diaper} (s=0.4, c=0.67)

{Diaper} → {Milk, Beer} (s=0.4, c=0.5)

{Milk} → {Diaper, Beer} (s=0.4, c=0.5)