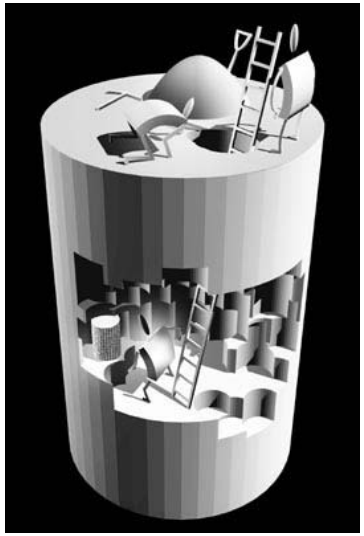
Tema 2

El proceso del descubrimiento de conocimiento a partir de bases de datos



Tema 2. El proceso de extracción de conocimiento a partir de bases de datos

Objetivos:

- Entender el objetivo del proceso de extracción de conocimiento a partir de bases de datos
- Conocer las distintas fases que componen este proceso
- Conocer las distintas tareas de Minería de Datos
- Conocer una taxonomía de técnicas que lo resuelven
- Conocer distintas características presentes en Minería de Datos

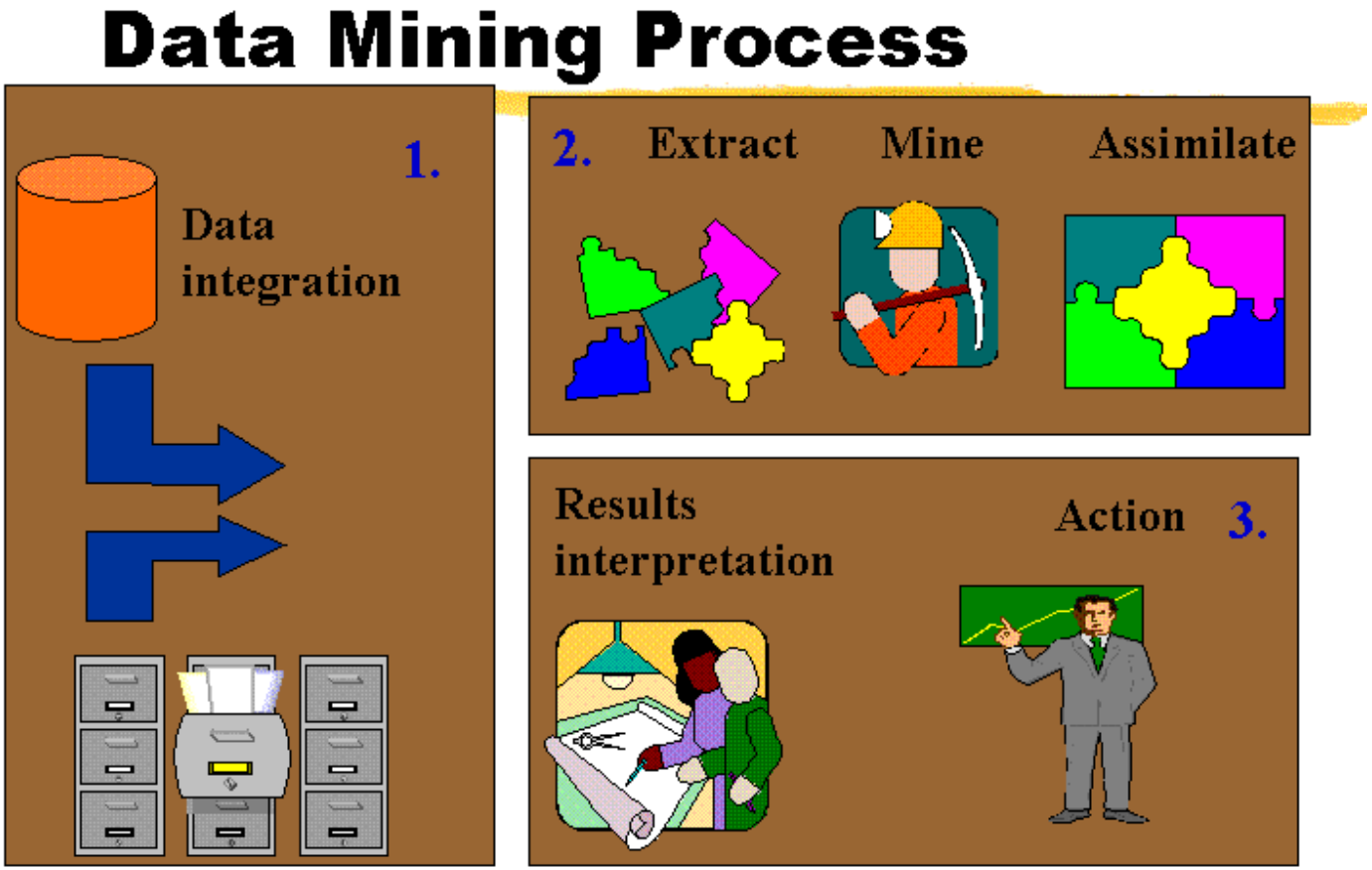
Tema 2. El proceso de extracción de conocimiento a partir de bases de datos

1. Introducción al KDD
2. Etapas en el proceso de KDD
3. Técnicas de Minería de Datos
4. Aspectos importantes en Minería de Datos

1. Introducción al KDD

- KDD = *Knowledge Discovery from Databases*
- El KDD es el proceso completo de extracción de conocimiento a partir de bases de datos
- El término se acuñó en 1989 para enfatizar que el conocimiento es el producto final de un proceso de descubrimiento guiado por los datos
- La Minería de Datos es sólo una etapa en el proceso de KDD
- Informalmente se asocia Minería de Datos con KDD

1. Introducción al KDD



1. Introducción al KDD

- KDD está enfocado al proceso global de descubrimiento del conocimiento a partir de bases de datos. Incluye:
  - cómo se almacenan y acceden los datos,
  - cómo se pueden escalar los algoritmos para trabajar con cantidades de datos enormes y seguir siendo eficientes,
  - cómo se pueden interpretar y visualizar los resultados, y
  - cómo modelar y dar soporte a la interacción hombre-máquina durante todo el proceso.
- KDD hace especial énfasis en la búsqueda de modelos/patrones comprensibles
- Es importante la robustez frente a grandes conjuntos de datos ruidosos

1. Introducción al KDD

El proceso de KDD contiene:

- El uso de la base de datos junto con cualquier operación de selección, preprocesamiento, (sub)muestreo y transformación de la misma
- Algoritmos para obtener patrones/modelos a partir de los datos (MD en sentido estricto)
- Evaluación del resultado de los algoritmos y selección de aquellos modelos que puedan considerarse conocimiento

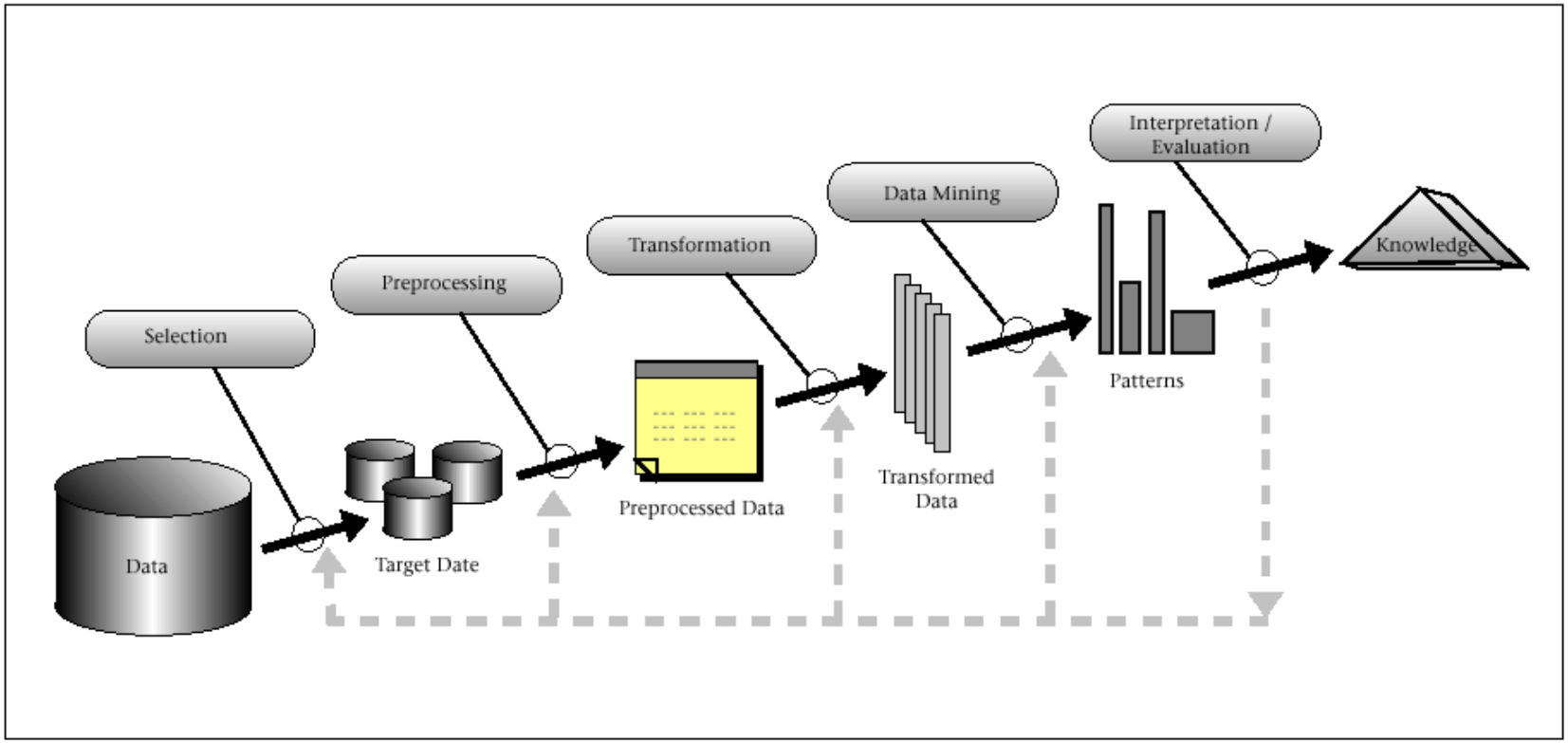
Es un proceso

- iterativo: a veces son necesarias varias iteraciones para extraer conocimiento de alta calidad, e
- interactivo: un experto en el dominio del problema debe ayudar en la preparación de los datos, validación del conocimiento extraído, etc.

Tema 2. El proceso de extracción de conocimiento a partir de bases de datos

1. Introducción al KDD
2. Etapas en el proceso de KDD
3. Técnicas de Minería de Datos
4. Aspectos importantes en Minería de Datos

2. Etapas en el proceso de KDD



2. Etapas en el proceso de KDD

1. Integración y recopilación: Comprensión del dominio de aplicación del problema, identificación de conocimiento a priori y creación del Datawarehouse
2. Selección de datos, limpieza y transformación
3. Selección de la técnica de MD y aplicación de algoritmos concretos de MD
4. Evaluación, interpretación y presentación de los resultados obtenidos
5. Difusión y utilización del nuevo conocimiento

Tema 2. El proceso de extracción de conocimiento a partir de bases de datos

1. Introducción al KDD

2. Etapas en el proceso de KDD
   - 2.1. Integración y recopilación
   - 2.2. Selección, limpieza y transformación
   - 2.3. Minería de Datos
   - 2.4. Evaluación, interpretación y presentación de resultados
   - 2.5. Difusión y uso del nuevo conocimiento

3. Técnicas de Minería de Datos

2.1. Integración y recopilación

- La familiarización con el dominio del problema y la obtención de conocimiento a priori disminuye el espacio de soluciones posibles

  → más eficiencia en el resto del proceso

- En problemas de KDD se suele trabajar con datos de diferentes departamentos de una entidad

  → es conveniente agrupar y unificar la información

- Unificación de la información en un Datawarehouse a partir de:
  - Información interna: distintas BBDD diseñadas para trabajo transaccional y de otro tipo (hojas de cálculo, informes,...)
  - Estudios publicados (demografía, catálogos, páginas, ...)
  - Otras bases de datos (compradas, industrias/empresas afines,...)

  El resto del proceso de KDD será más cómodo si la fuente de datos está unificada, es accesible y dedicada (desconectada del trabajo transaccional)

- El DW es conveniente para KDD aunque no imprescindible. A veces se trabaja directamente con la BD o con las BBDD en formatos heterogéneos

2.2. Selección, limpieza y transformación

- La calidad del conocimiento descubierto no depende sólo del algoritmo de DM sino de la calidad de los datos minados

- Objetivo general de esta fase: seleccionar el conjunto de datos adecuado para el resto del proceso de KDD

- Las tareas de esta etapa se agrupan en:
  - Limpieza de datos (*data cleaning*)
  - Transformación de los datos
  - Reducción de la dimensionalidad

2.2.1. Limpieza de datos: *data cleaning*

- Datos perdidos (*missing values*)
  - Pueden llevar a resultados poco precisos
  - Hay que analizar el motivo
    - Mal funcionamiento del dispositivo de recogida de datos
    - Cambios efectuados durante la recolección de datos
    - Datos que provienen de fuentes diversas
  - Soluciones: rellenarlos manualmente, ignorarlos, eliminar la fila/columna, usar un valor especial (p.e. *unknow)*, inferirlos usando técnicas estadísticas,...

- Datos anómalos (*outliers*)
  - Valores que no se ajustan al comportamiento general de los datos
  - Pueden ser erróneos o correctos pero diferentes a los demás
  - Primero hay que identificarlos, y después, en función del problema se tratarán como valores perdidos o se sacará información de ellos

- Inconsistencias: registros duplicados, datos inconsistentes, ...

  Normalmente ya tratado en la elaboración del DW

2.2.2. Transformación de los datos

- Construcción de atributos:

  construir nuevos atributos aplicando alguna operación a los atributos originales (agrupamiento, separación, fecha → enteros, convertir en números los valores categóricos…)

  - cuando los atributos no tienen mucho poder predictivo por sí solos,
  - cuando los patrones dependen de variaciones lineales de las variables globales

  En ocasiones => almacenar meta-información sobre la información realmente almacenada por cada campo

- Discretización:

  Pasar atributos continuos (o discretos con muchos valores) a casos discretos manejables o a categóricos

  - Hay diversas técnicas
  - Es imprescindible para muchos algoritmos de MD

2.2.3. Reducción de la dimensionalidad

- Reducción de casos / filas:
  - Puede hacer más eficiente el proceso de DM
  - Las técnicas utilizadas van desde la compresión al muestreo de los datos, pasando por la elección de representantes (*clustering*)

- Selección de variables *(feature selection):*

  Seleccionar el conjunto de atributos adecuado para la tarea específica a realizar
  - Se conoce también como proyección
  - Es uno de los pre-procesamientos más importantes
  - Técnicas utilizadas para esta tarea: estadísticas, basadas en búsqueda combinadas con métodos empíricos,...

2.4. Minería de datos

- Objetivo: Producir nuevo conocimiento que pueda utilizar el usuario
- ¿Cómo?

  Construyendo un modelo, basado en los datos recopilados, que sea una descripción de los patrones y relaciones entre los datos con los que se puedan hacer predicciones, entender mejor los datos o explicar situaciones pasadas
- Decisiones a tomar:
  - ¿Qué tipo de conocimiento buscamos?
    - Predictivo
    - Descriptivo
  - ¿Qué técnica es la más adecuada?
    - Clasificación
    - Regresión
    - Agrupamiento (clustering)
    - Asociaciones, …
  - ¿Qué tipo de modelo?
    - P.e. Clasificación: reglas, AANNs, árboles de decisión, etc.
  - ¿Es necesaria la incertidumbre en el modelo resultante? Certeza, probabilidad, lógica difusa,…
  - ¿Qué algoritmo es el más adecuado? P.e.: en clustering: duro, difuso, jerarquizado, k-means, iterativo, EM,…

2.5. Evaluación, interpretación y presentación de resultados

- La fase de MD puede producir varias hipótesis de modelos
- Es necesario establecer qué modelos son los más válidos
- Criterios: los patrones descubiertos deben ser
  - precisos,
  - comprensibles, e
  - interesantes (útiles, novedosos)
- Técnicas de evaluación: Al menos se divide el conjunto de datos en dos (entrenamiento y test)
  - Entrenamiento: Para extraer el conocimiento
  - Test: Para probar la validez del conocimiento extraído
  - Alternativas:
    - Validación simple
    - n-validación cruzada
    - *Bootstrapping*,…
- Medidas de evaluación de modelos: Dependen de la tarea:
  - Clasificación: precisión predictiva (%acierto)
  - Regresión: Error cuadrático medio
  - Agrupamiento: Medidas de cohesión y separación entre grupos
  - Reglas de asociación: cobertura, confianza…
- La interpretación de los mejores modelos (visualización, simplicidad, posibilidad de integración, ventajas colaterales,…) ayuda a la selección del modelo(s) final(es)

2.6. Difusión y utilización del nuevo conocimiento

Una vez construido y validado el modelo puede utilizarse:

- para recomendar acciones
- para aplicar el modelo a diferentes conjuntos de datos

En cualquier caso, es necesario:

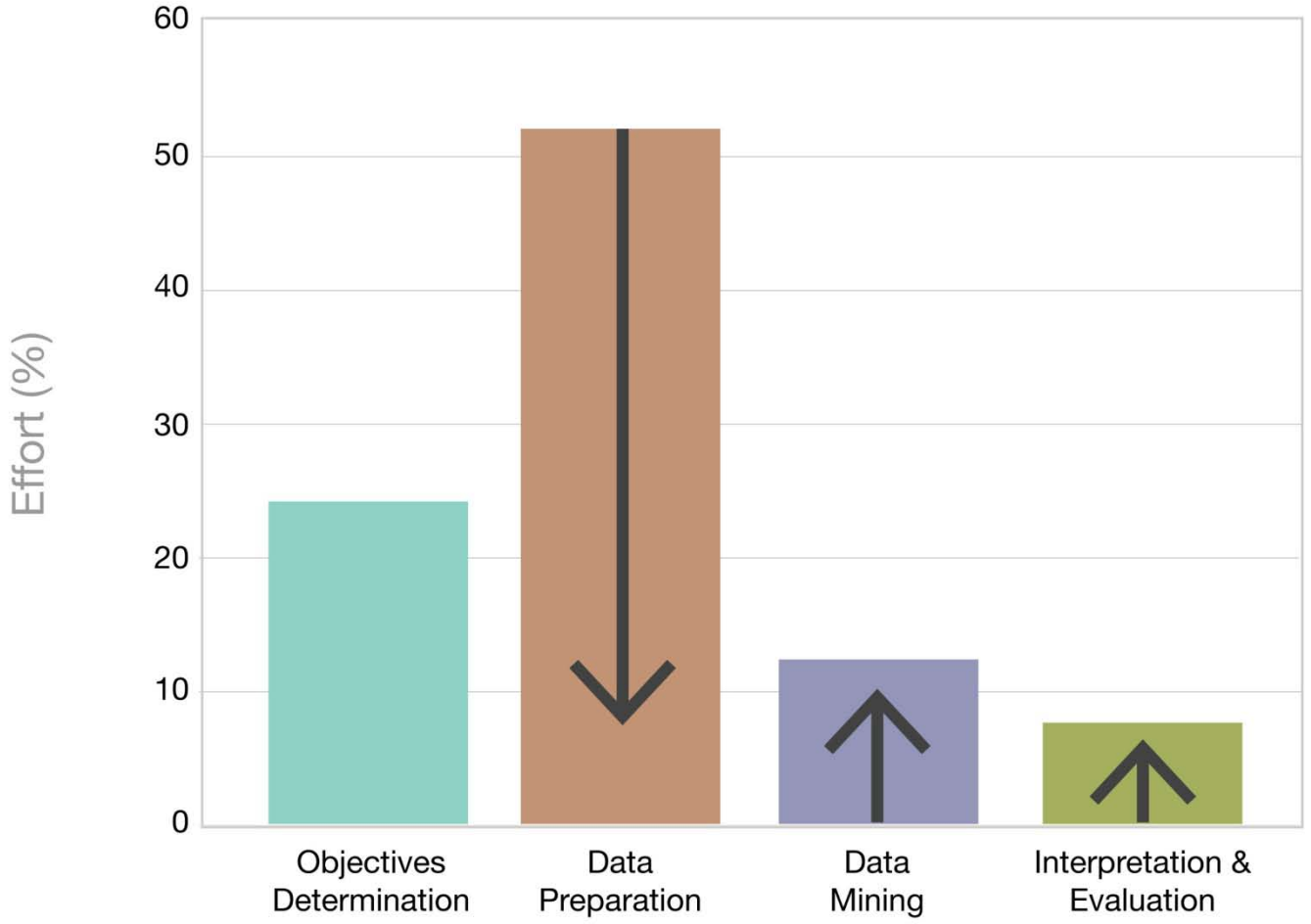
- Difusión: Elaboración de informes para su distribución
- Utilización del nuevo conocimiento de forma independiente
- Incorporación a sistemas ya existentes

→ comprobar con el conocimiento ya utilizado para evitar inconsistencias y posibles conflictos

La monitorización del sistema en acción dará lugar a nuevos casos que realimentarán el ciclo del KDD

Las conclusiones iniciales pueden variar, invalidando el modelo adquirido

2. Etapas en el proceso de KDD



Tema 2. El proceso de extracción de conocimiento a partir de bases de datos

1. Introducción al KDD

2. Etapas en el proceso de KDD

3. Técnicas de Minería de Datos
   3.1. Visión sistemática de los algoritmos de MD
   3.2. Taxonomía de los algoritmos de MD

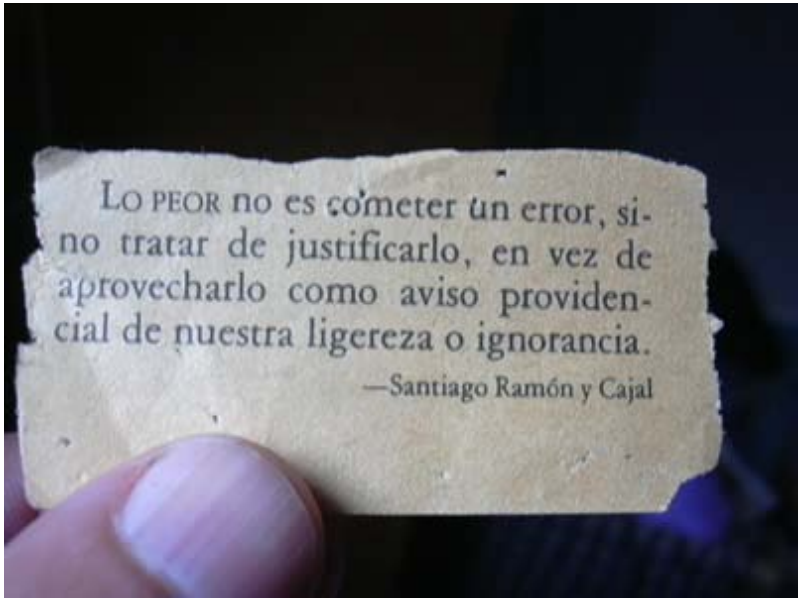
4. Aspectos importantes en Minería de Datos

3. Técnicas de Minería de Datos

La MD es una forma de aprender del pasado para tomar mejores decisiones en el futuro

3. Técnicas de Minería de Datos. Errores comunes

- Aprender de cosas que no son ciertas
  - Patrones que no representan ninguna regla subyacente
  - Datos que no reflejan lo relevante
  - Datos con un nivel de detalle erróneo
- Aprender cosas ciertas, pero inútiles
  - Aprender información ya conocida
  - Aprender cosas que no se pueden utilizar



3. Técnicas de Minería de Datos. Tendencias y objetivos

Tendencias en MD:

- Comprobación de hipótesis
- MD supervisada
- MD no supervisada

Objetivos en MD:

- Predicción
- Descripción

3. Técnicas de Minería de Datos. Componentes

Un algoritmo de MD es un procedimiento bien definido que toma datos como entrada y produce modelos o patrones como salida

Un algoritmo de MD se puede especificar mediante la definición de cinco componentes:

- Tarea
- Estructuración del modelo/patrón
- Función objetivo
- Método de búsqueda/optimización
- Técnica de manejo de los datos

3.1. Visión sistemática de los algoritmos de Minería de Datos

- Tarea: Identificar el tipo de problema a abordar con el algoritmo de MD (clasificación, visualización, clustering,...)

- Estructura: Describir el modelo a aprender, es decir, cuál será el patrón o modelo que intentaremos descubrir para que represente a los datos (árboles, reglas, ecuaciones, gráficos,...)

- Función objetivo: Criterio a optimizar durante el proceso de MD. Medirá la bondad de los modelos encontrados respecto a los datos

  Puede estar basado únicamente en bondad de ajuste o por el contrario puede intentar capturar generalización

3.1. Visión sistemática de los algoritmos de Minería de Datos

- Método de búsqueda/optimización. Es el tipo de método que se usará en el intento de que el patrón obtenido optimice la función objetivo (métodos voraces, heurísticos, probabilísticos,...).

  Dependiendo de si la estructura es fija o no, será aprendizaje estructural o paramétrico.

- Técnica de manejo de los datos. Técnica a utilizar para el almacenamiento, indexado y recuperación de los datos.

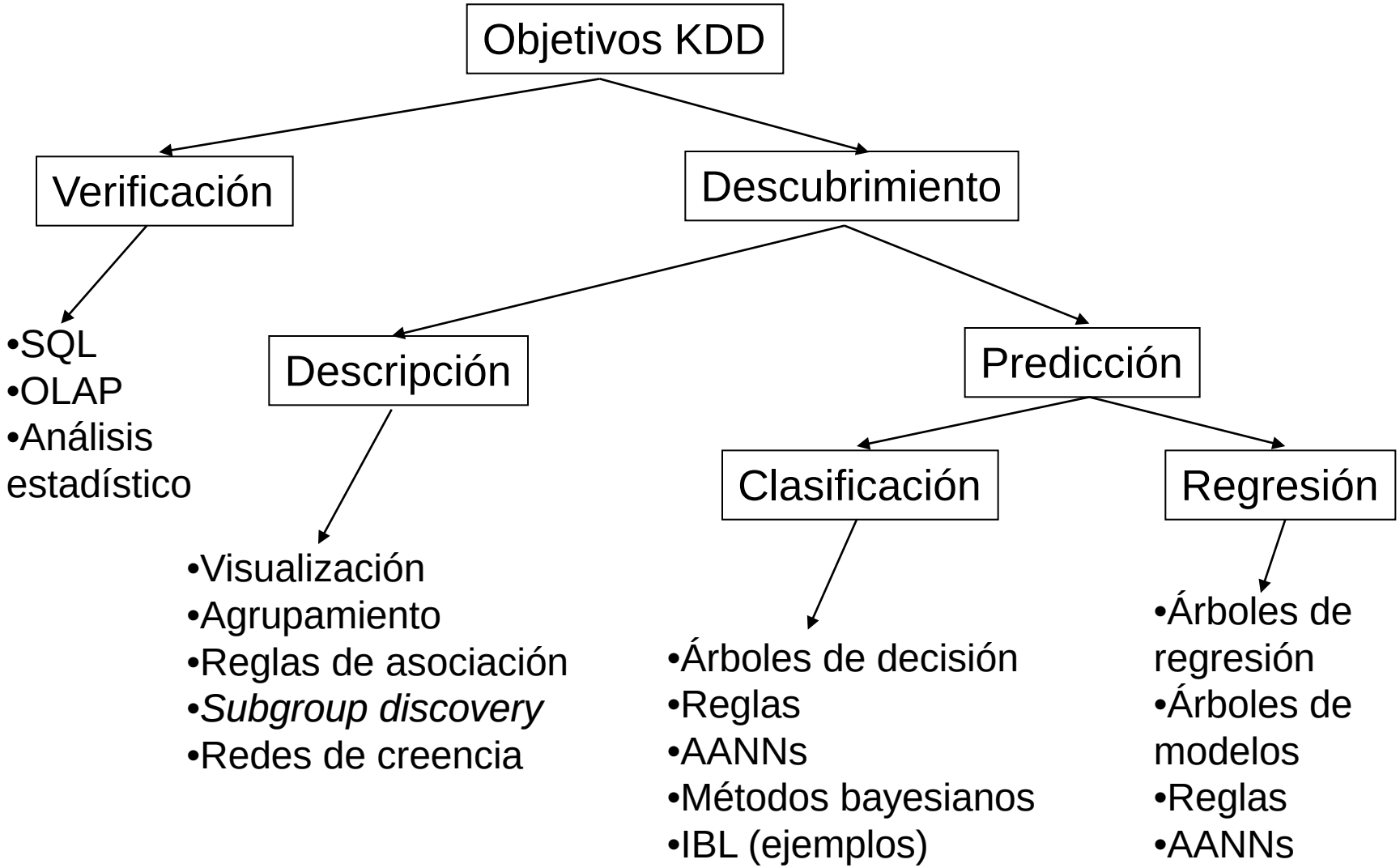
  La mayoría de los métodos de aprendizaje automatizado obvian este paso porque asumen que el volumen de datos es lo suficientemente pequeño para estar en memoria principal.

  Con grandes volúmenes de datos, este paso es muy importante.

3.1. Visión sistemática de los algoritmos de Minería de Datos

	ID3	**RNs-Backprop.**	**A priori**
Tarea	Clasificación	Regresión/clasificación	Descubrimiento de reglas
Estructura	Árbol de decisión	Red Neuronal	Reglas de asociación
Función objetivo	Ganancia de información	Error cuadrático	Soporte / confianza
Método de búsqueda	Voraz	Gradiente descendiente	Primero-mejor + poda
Manejo de los datos			Lecturas secuenciales

3.2. Taxonomía de técnicas de Minería de Datos



3.2.1. Clasificación

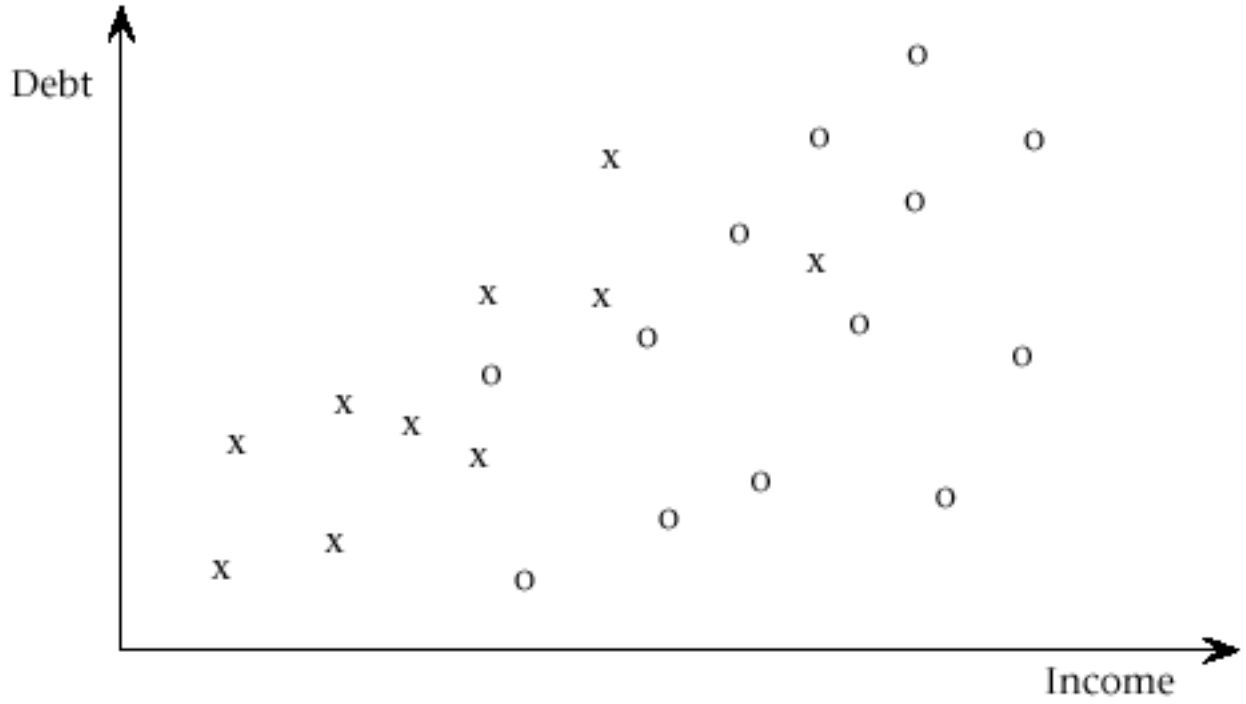
- Objetivo: establecer una correspondencia entre datos y grupos predefinidos o clases
- ¿Cómo? A partir de los valores que toman ciertos atributos o variables predictivas se extraerá información que permita predecir la pertenencia de un objeto nuevo (del que se desconozca la clase) a una de las posibles clases
- Aprendizaje supervisado
- Ejemplos:
  - Sistema de ayuda a la decisión para asignación de créditos
  - Sistema de ayuda a la decisión de cirugía ocular
  - Sistema de reconocimiento facial en un sistema de seguridad de un aeropuerto

3.2.1. Clasificación

- Clasificación. Algunas técnicas:
  - Umbrales
  - Clasificadores lineales
  - Árboles de clasificación
  - Clasificadores basados en reglas
  - Clasificadores bayesianos
  - Redes Neuronales (AANNs)
  - Métodos basados en ejemplos (IBL)

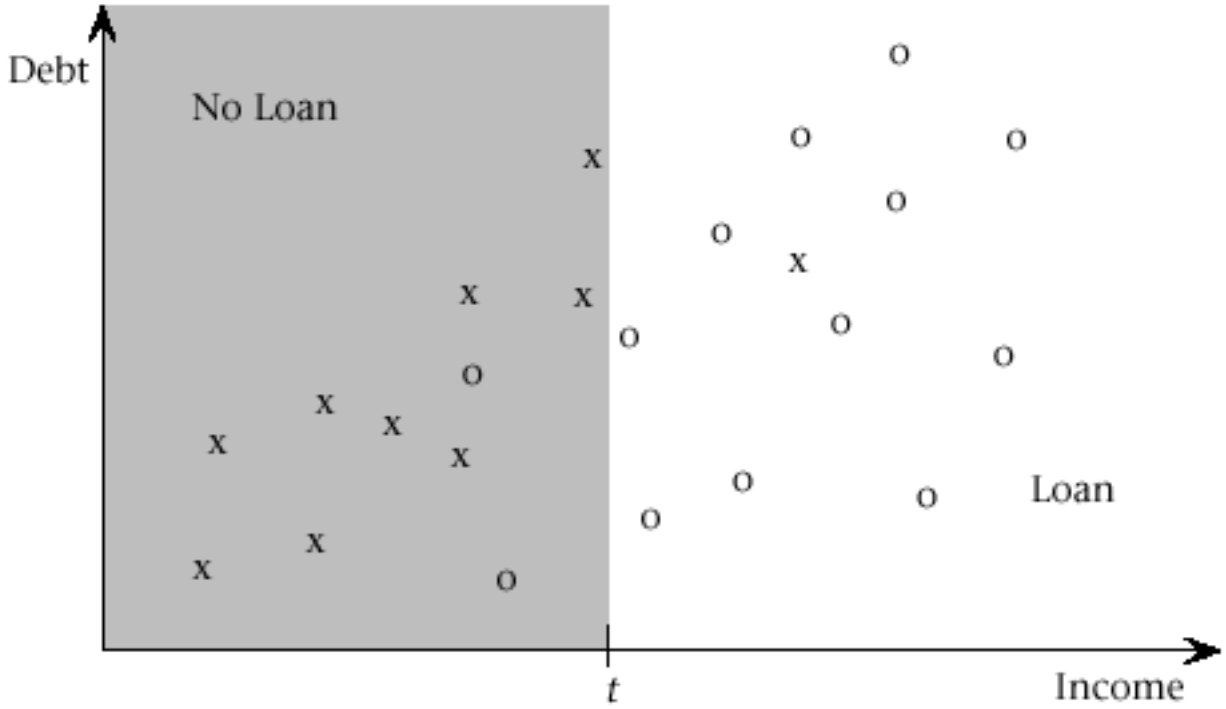
3.2.1. Clasificación

Ejemplo 1: Problema con dos atributos (deuda e ingresos) y dos clases: prestar (o) y no prestar (x)



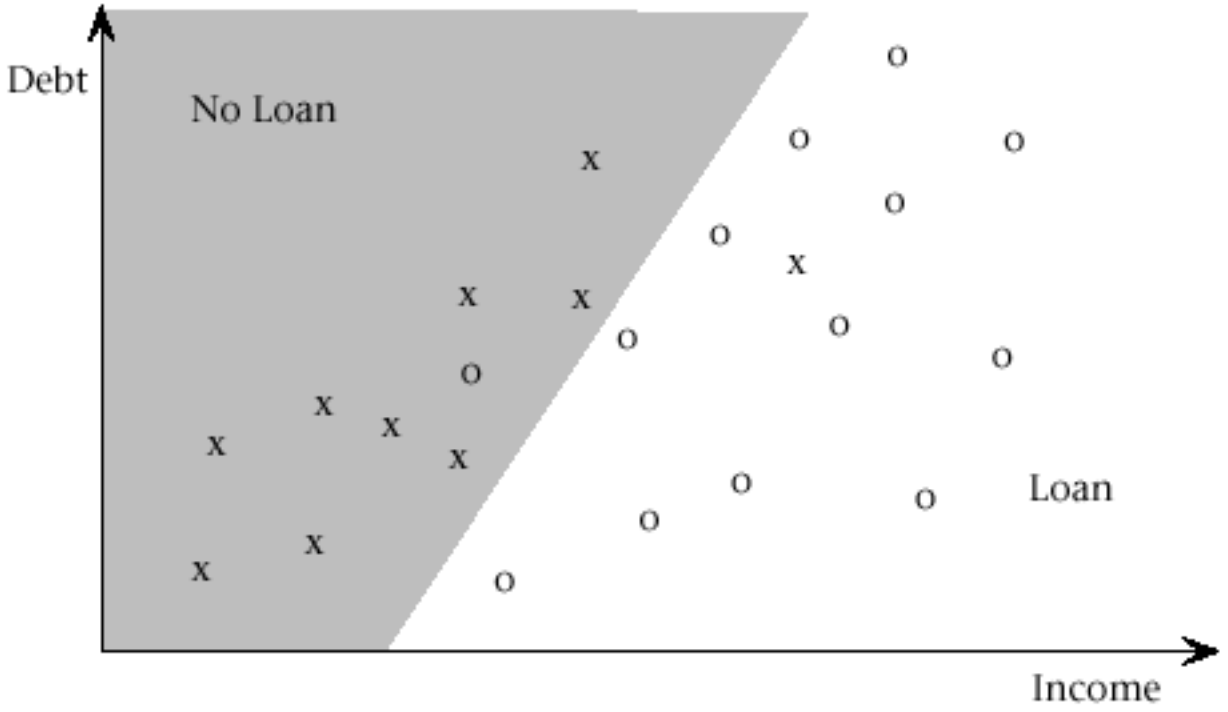
3.2.1. Clasificación

Clasificación utilizando un umbral → 3 errores



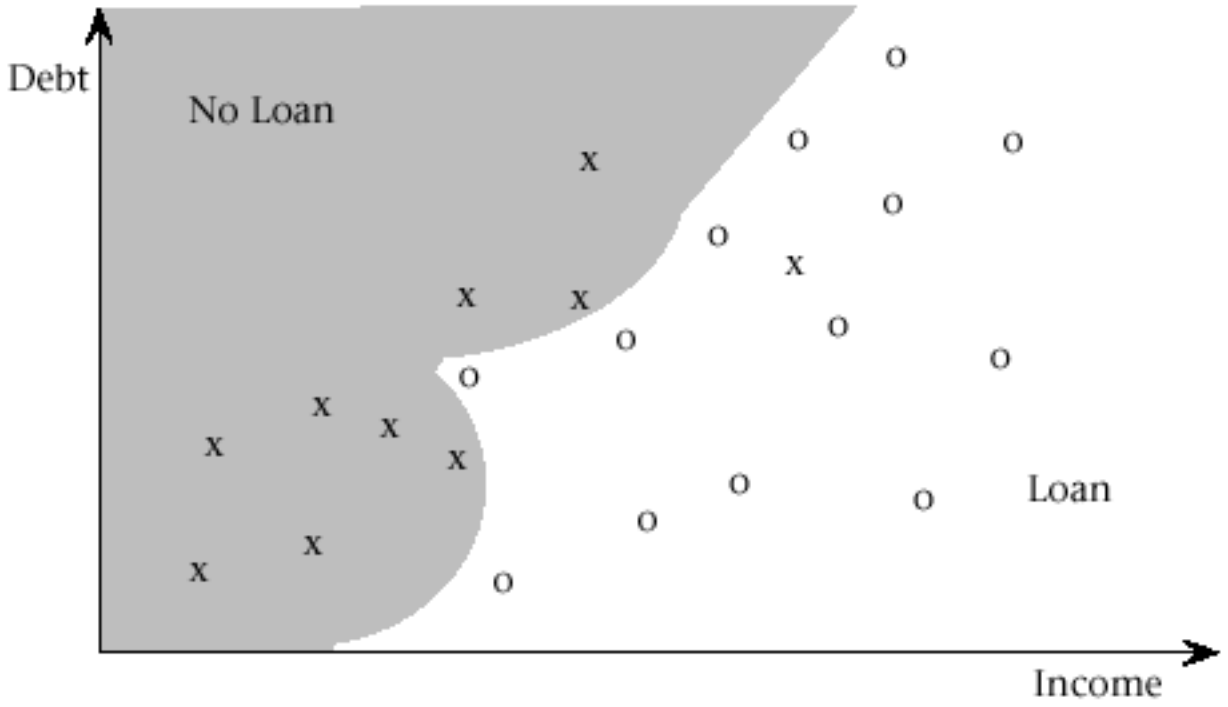
3.2.1. Clasificación

Clasificación lineal → 2 errores



3.2.1. Clasificación

Clasificación no lineal con una red neuronal multicapa → 1 error



3.2.1. Clasificación

Clasificación basada en ejemplos con k-vecinos más próximos → 0 errores

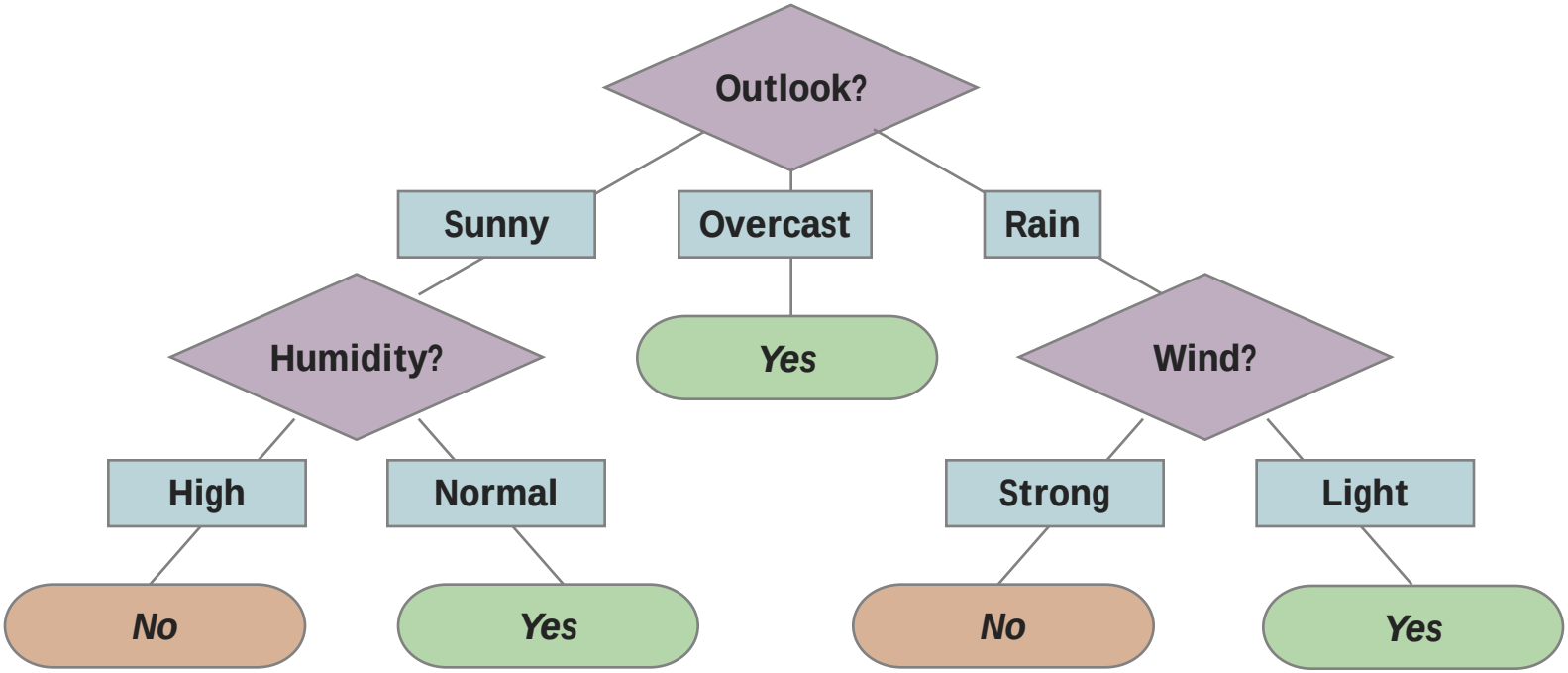
3.2.1. Clasificación

Ejemplo 2

Day	Outlook	Temperature	Humidity	Wind	*Play Tennis?*
1	Sunny	Hot	High	Light	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Light	Yes
4	Rain	Mild	High	Light	Yes
5	Rain	Cool	Normal	Light	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Light	No
9	Sunny	Cool	Normal	Light	Yes
10	Rain	Mild	Normal	Light	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Light	Yes
14	Rain	Mild	High	Strong	No

3.2.1. Clasificación

Árboles de clasificación: ID3

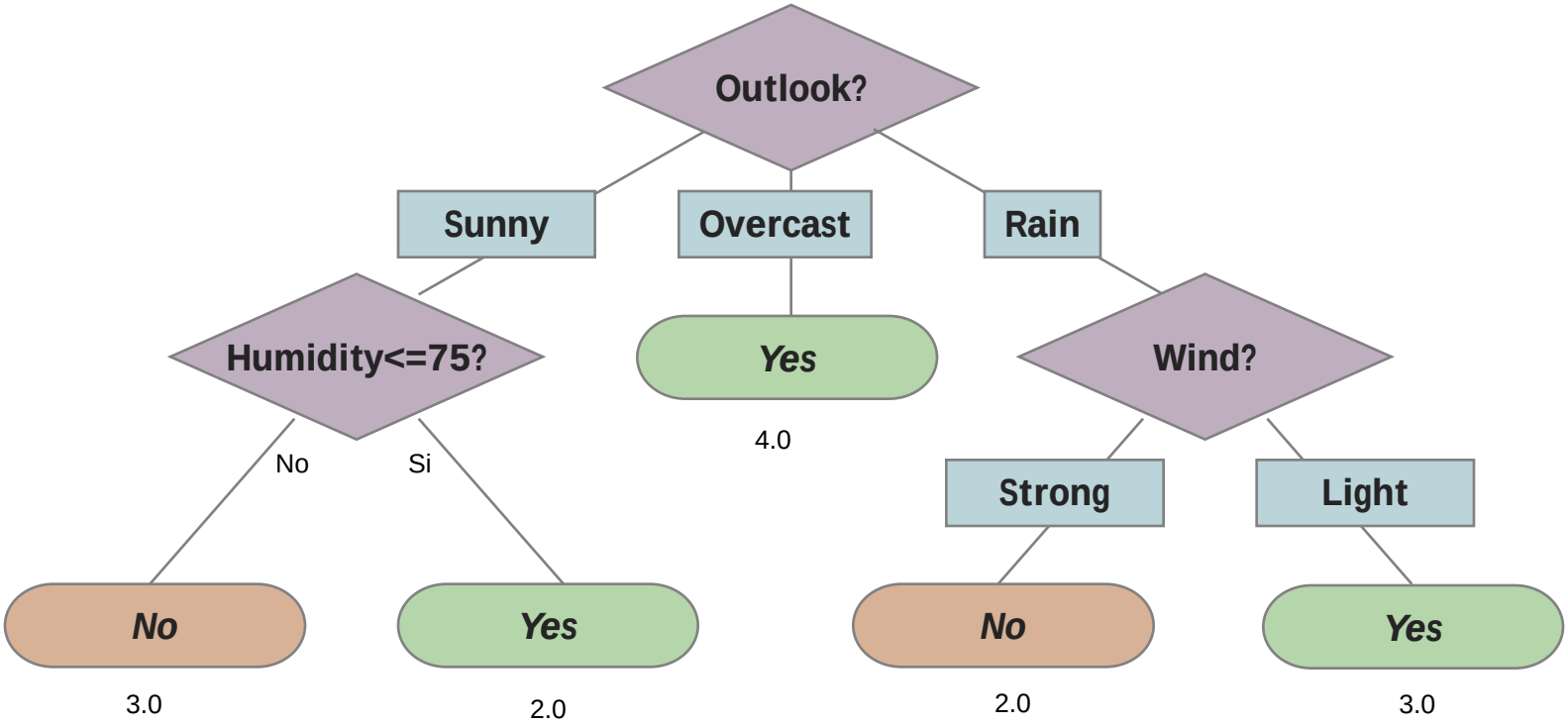


3.2.1. Clasificación

Day	Outlook	Temperature	Humidity	Wind	*Play Tennis?*
1	Sunny	85	85	Light	No
2	Sunny	80	90	Strong	No
3	Overcast	83	86	Light	Yes
4	Rain	70	96	Light	Yes
5	Rain	68	80	Light	Yes
6	Rain	65	70	Strong	No
7	Overcast	64	65	Strong	Yes
8	Sunny	72	95	Light	No
9	Sunny	69	70	Light	Yes
10	Rain	75	80	Light	Yes
11	Sunny	75	70	Strong	Yes
12	Overcast	72	90	Strong	Yes
13	Overcast	81	75	Light	Yes
14	Rain	71	91	Strong	No

3.2.1. Clasificación

Árboles de clasificación: C4.5



3.2.1. Clasificación

Reglas derivadas del árbol obtenido por C4.5:

Rule 1: IF outlook = overcast
THEN class Yes [70.7%]

Rule 2: IF outlook = rainy AND windy = false
THEN class Yes [63.0%]

Rule 3: IF outlook = sunny AND humidity > 75
THEN class No [63.0%]

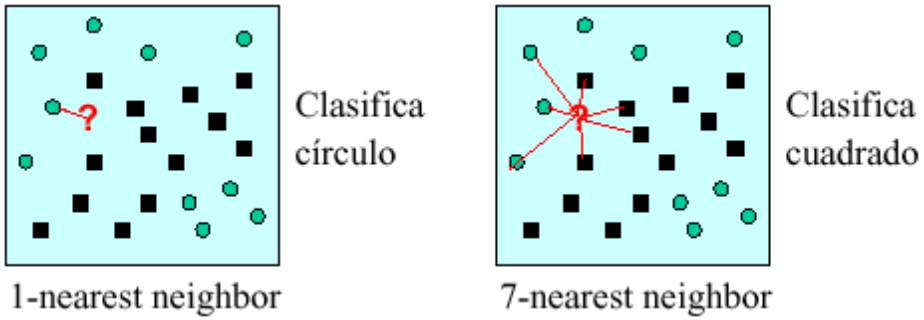
Rule 4: IF outlook = rainy AND windy = true
THEN class No [50.0%]

Default class: yes

3.2.1. Clasificación

Métodos basados en ejemplos

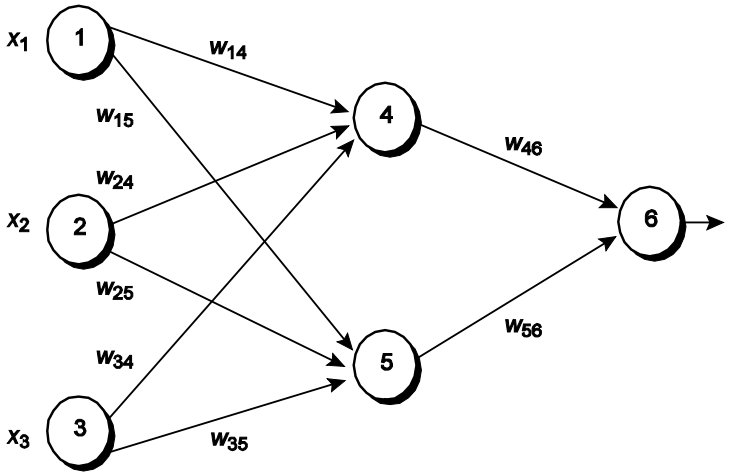
- La propia base de datos (ejemplos) constituye el modelo
- La clase de un nuevo ejemplo se obtiene de la clase de ejemplos similares
- Ejemplos:
  - K-vecino más cercano
  - Razonamiento basado en casos

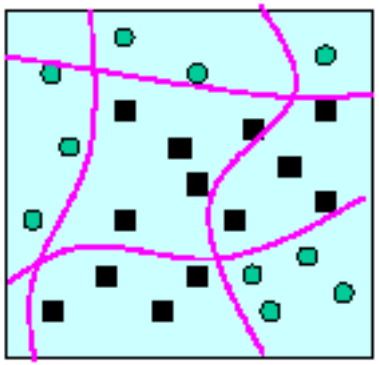


3.2.1. Clasificación

Redes neuronales:

- Perceptrón (sin capas ocultas) → clasificadores lineales
- Redes neuronales multicapa → permiten particiones no lineales





3.2.2. Regresión

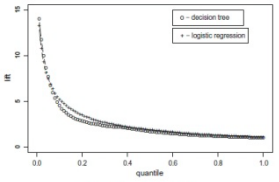


Fig. 1. Lift curves for train sample.

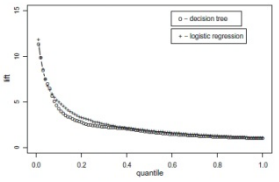


Fig. 2. Lift curves for calibration sample.

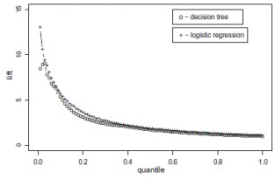


Fig. 3. Lift curves for test sample.

- Se utiliza para designar de forma general el problema de predecir una variable de tipo continuo
- Se trata de aproximar el valor numérico de dicha variable conociendo el resto de atributos
- Implica el aprendizaje de una función que establece la correspondencia entre los datos y el valor a predecir
- En ocasiones se fija el tipo de función (lineal, logística,…) y se determina la función del tipo que mejor se adapta a los datos
- Ejemplos:
  - Estimación de bio-masa en un bosque a partir de fotografías vía satélite
  - Estimación de probabilidades de supervivencia ante determinados diagnósticos o tratamientos
  - Estimación de la demanda de determinados productos en función de clientes, fechas,...

3.2.2. Regresion

		Memoria RAM		Caché	Canales		Rendimiento
Vendor	MCYT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
Dec	133	1000	12000	9	3	12	54
Dec	133	1000	8000	9	3	12	41
Dg	700	384	8000	0	1	1	34
Dg	700	256	2000	0	1	1	19
Hp	90	256	1000	0	3	10	18
Hp	105	256	2000	0	3	10	20
Ibm	57	4000	24000	64	12	16	171
Ibm	26	16000	32000	64	16	24	361
.....							
Ncr	56	2000	8000	0	1	8	41
Nixdorf	200	1000	2000	0	1	2	21
Siemens	240	512	2000	8	1	5	22
Siemens	105	2000	4000	8	3	8	31

3.2.2. Regresion

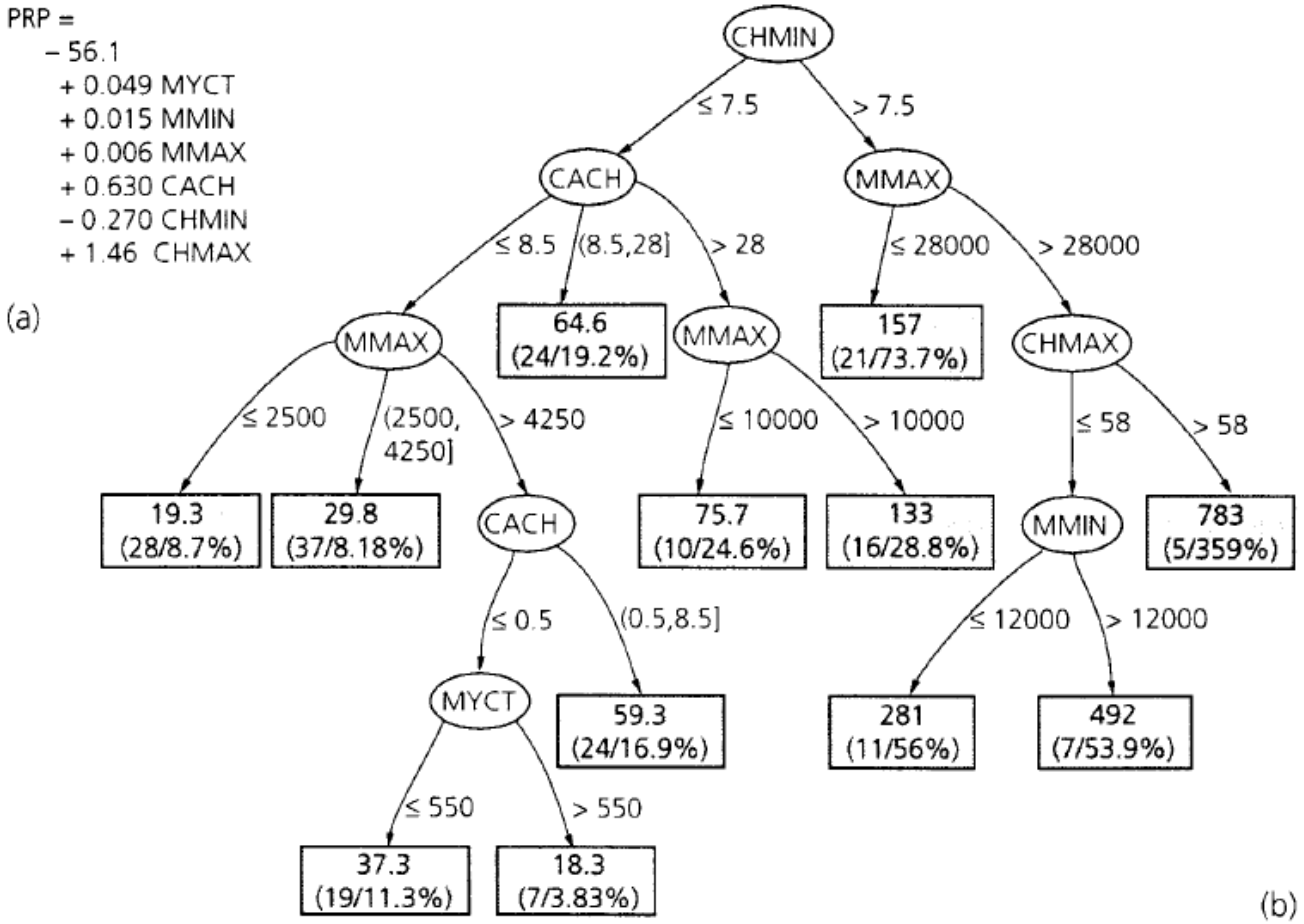
- Regresión lineal
  - Objetivo: Obtener una expresión que prediga la cantidad numérica

PRP = -55.0 + 0.0489·MCYT + 0.0153·MMIN + 0.0056·MMAX + 0.6410·CACH – 0.2700·CHMIN + 1.480·CHMAX

- Regresión con árboles
  - Un árbol de regresión es un árbol de decisión cuyas hojas predicen una cantidad numérica
    - El valor de predicción es la media de los ejemplos que llegan a cada hoja
  - Un árbol de modelos es un árbol de regresión que contiene una expresión de regresión lineal en cada hoja
    - Permite aproximar funciones continuas

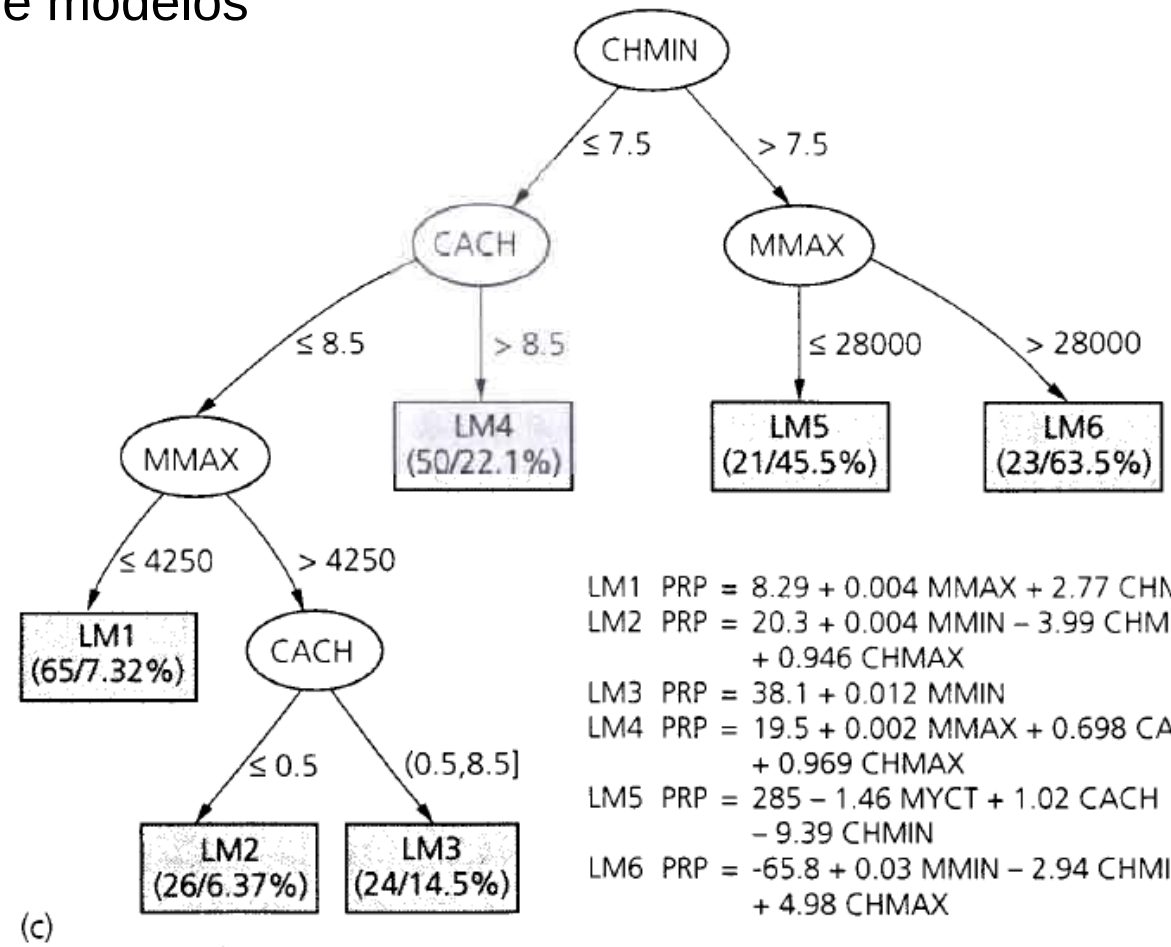
3.2.2. Regresión

Árbol de regresión



3.2.2. Regresión

Árbol de modelos

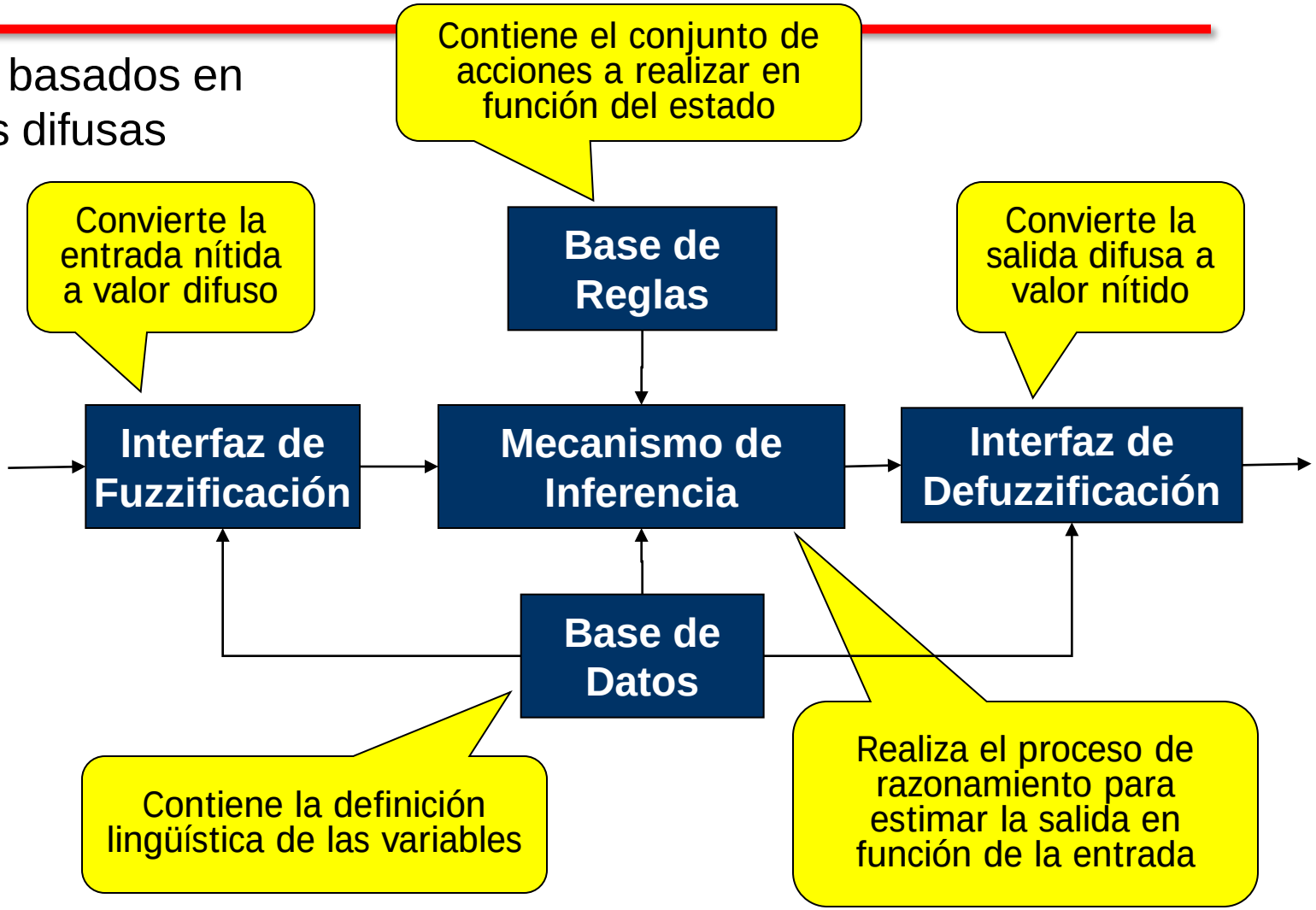


(c)

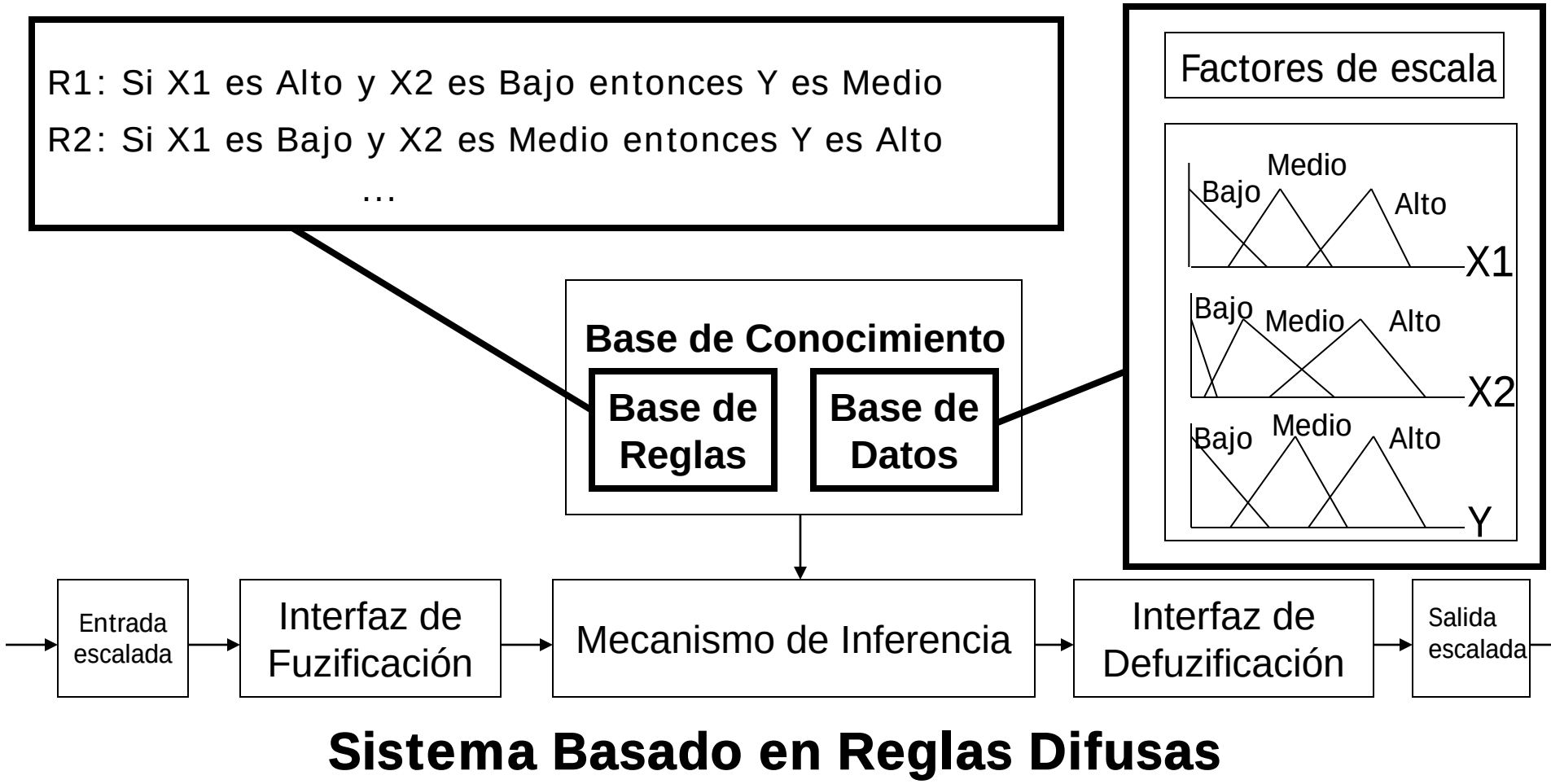
LM1 PRP = 8.29 + 0.004 MMAX + 2.77 CHMIN
LM2 PRP = 20.3 + 0.004 MMIN – 3.99 CHMIN
+ 0.946 CHMAX
LM3 PRP = 38.1 + 0.012 MMIN
LM4 PRP = 19.5 + 0.002 MMAX + 0.698 CACH
+ 0.969 CHMAX
LM5 PRP = 285 – 1.46 MYCT + 1.02 CACH
– 9.39 CHMIN
LM6 PRP = -65.8 + 0.03 MMIN – 2.94 CHMIN
+ 4.98 CHMAX

3.2.2. Regresión

Sistemas basados en reglas difusas



3.2.2. Regresión



Sistema Basado en Reglas Difusas

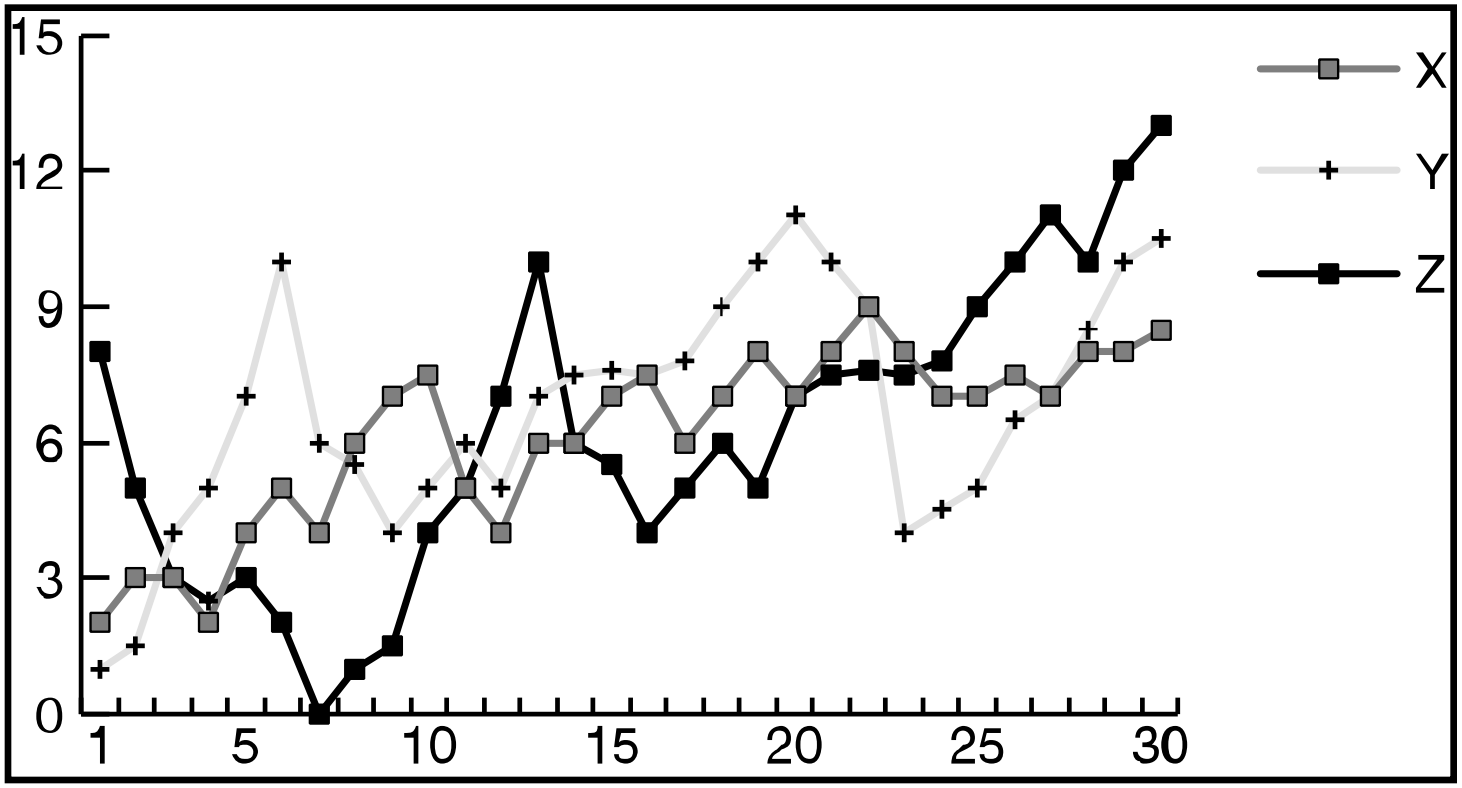
3.2.3. Análisis de series temporales

- Objetivo: Observar la variación del valor de un atributo en el tiempo

- Normalmente los valores que se analizan, están distribuidos en el tiempo

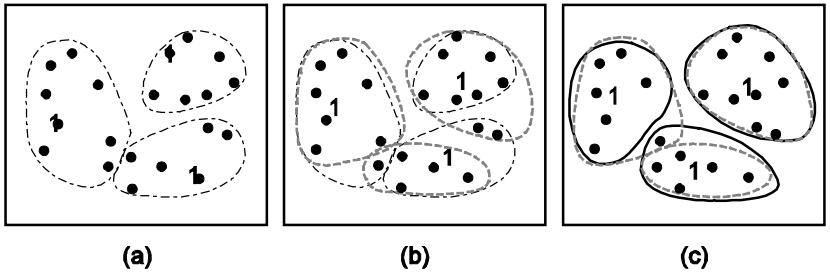
- Se suelen visualizar
  - Permite utilizar medidas de distancia para determinar la similitud entre diferentes series temporales
  - Permite determinar el comportamiento
  - Permite predecir comportamiento

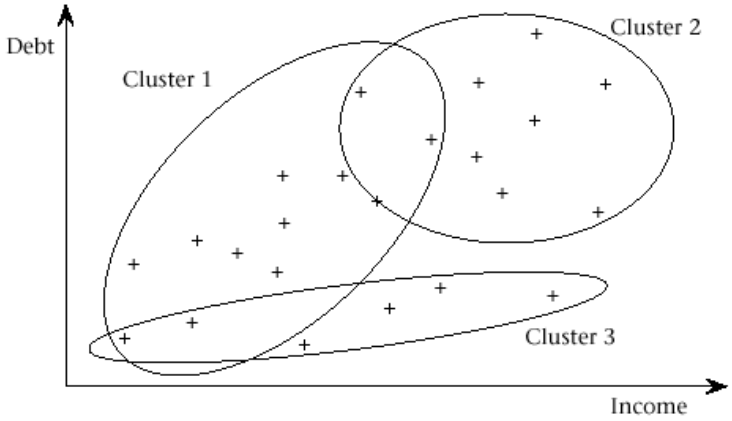
3.2.3. Análisis de series temporales



3.2.4. Agrupamiento (*clustering*)

- Similar a la clasificación excepto que los grupos no están definidos
- Técnica que nos permite agrupar objetos similares en grupos *(clusters)*
- Aprendizaje no supervisado si no se conoce el número de clusters, supervisado en otro caso
- Ejemplos: clustering duro con k-means y con solapamientos





3.2.5. Reglas de asociación

- Objetivo:

  Descubrir relaciones desconocidas en los datos

- Regla de asociación = modelo que identifica tipos de asociaciones específicas en los datos

- Ejemplos: análisis de cestas de mercados

3.2.6. Descubrimiento de secuencias

- Objetivo:

  Determinar patrones secuenciales en los datos

- Estos patrones son asociaciones en los datos pero con una relación en el tiempo

- Ejemplo: Descubrimiento de secuencias en el análisis de un Web log para determinar como acceden los usuarios a determinadas páginas

Tema 2. El proceso de extracción de conocimiento a partir de bases de datos

1. Introducción al KDD

2. Etapas en el proceso de KDD

3. Técnicas de Minería de Datos
   3.1. Visión sistemática de los algoritmos de MD
   3.2. Taxonomía de los algoritmos de MD

4. Aspectos importantes en Minería de Datos

4. Aspectos importantes en Minería de Datos

Cuestiones relacionadas con DM:

1. Interacción humana: Necesidad de interfaces con expertos en el dominio y expertos técnicos

2. Sobreajuste (*overfitting*): Cuando el modelo no se ajusta a datos futuros
   - Puede estar provocado por suposiciones erróneas sobre los datos o por un tamaño pequeño de la BD

3. Datos anómalos: Muy frecuentes en grandes BBDD

4. Interpretación de los resultados: pueden requerir expertos

5. Visualización de resultados

6. Grandes conjuntos de datos: BBDD masivas crean problemas en algunos algoritmos de DM

4. Aspectos importantes en Minería de Datos

7. Alta dimensionalidad (*dimensionality curse*). Solución: Aplicar técnicas de reducción de la dimensionalidad
8. Datos multimedia
9. Datos perdidos
10. Datos irrelevantes
11. Datos con ruido: Algunos valores de variables pueden ser inválidos o incorrectos. Deben corregirse antes de la aplicación de algoritmos de DM
12. Datos cambiantes: Muchos algoritmos de DM asumen datos estáticos
13. Integración: La integración de algoritmos de DM dentro de sistemas tradicionales de BBDD es un objetivo deseable
14. Aplicación: Determinar la intención de uso de los datos obtenidos es un reto