

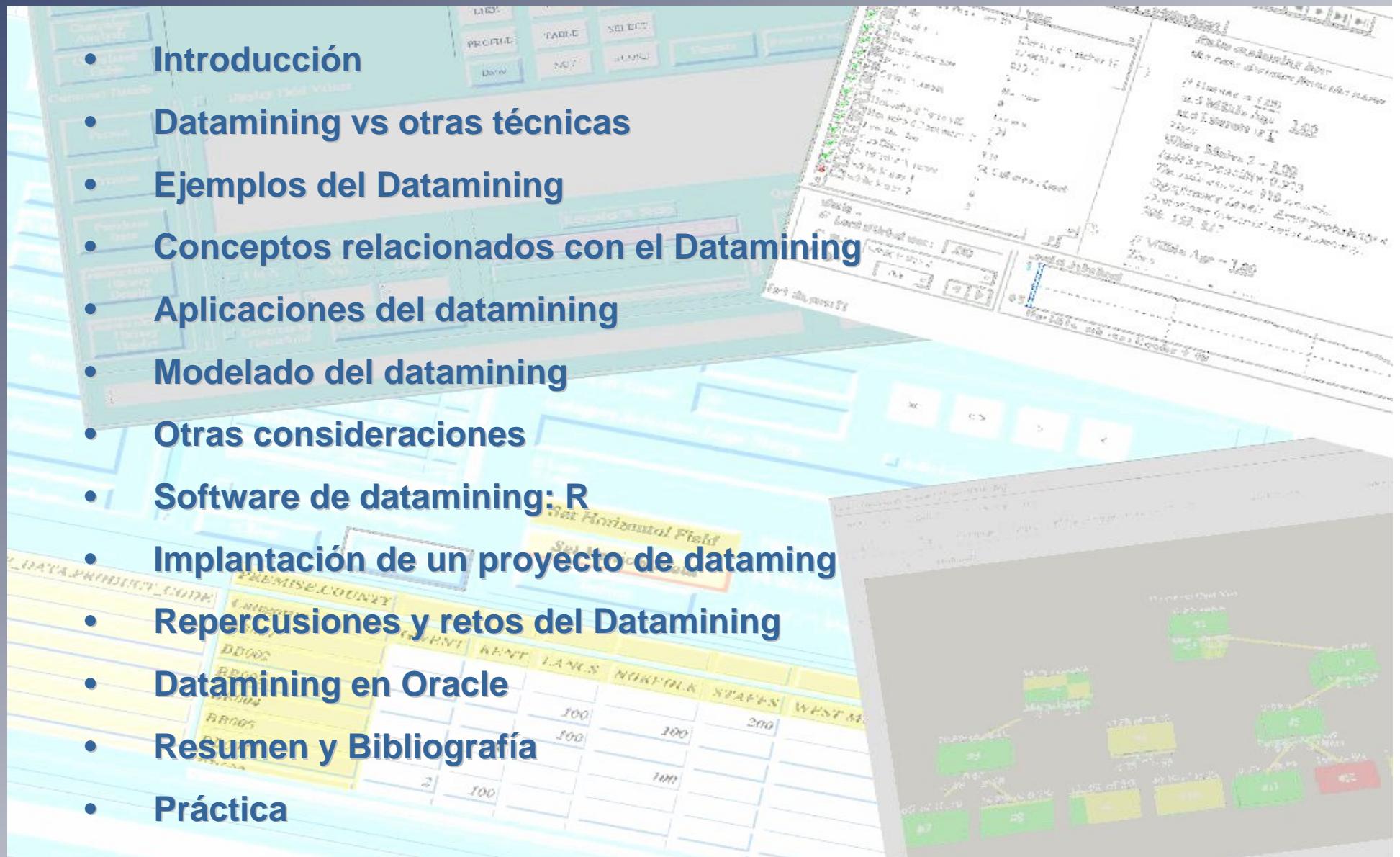
Celia Gutiérrez Cossío

2007

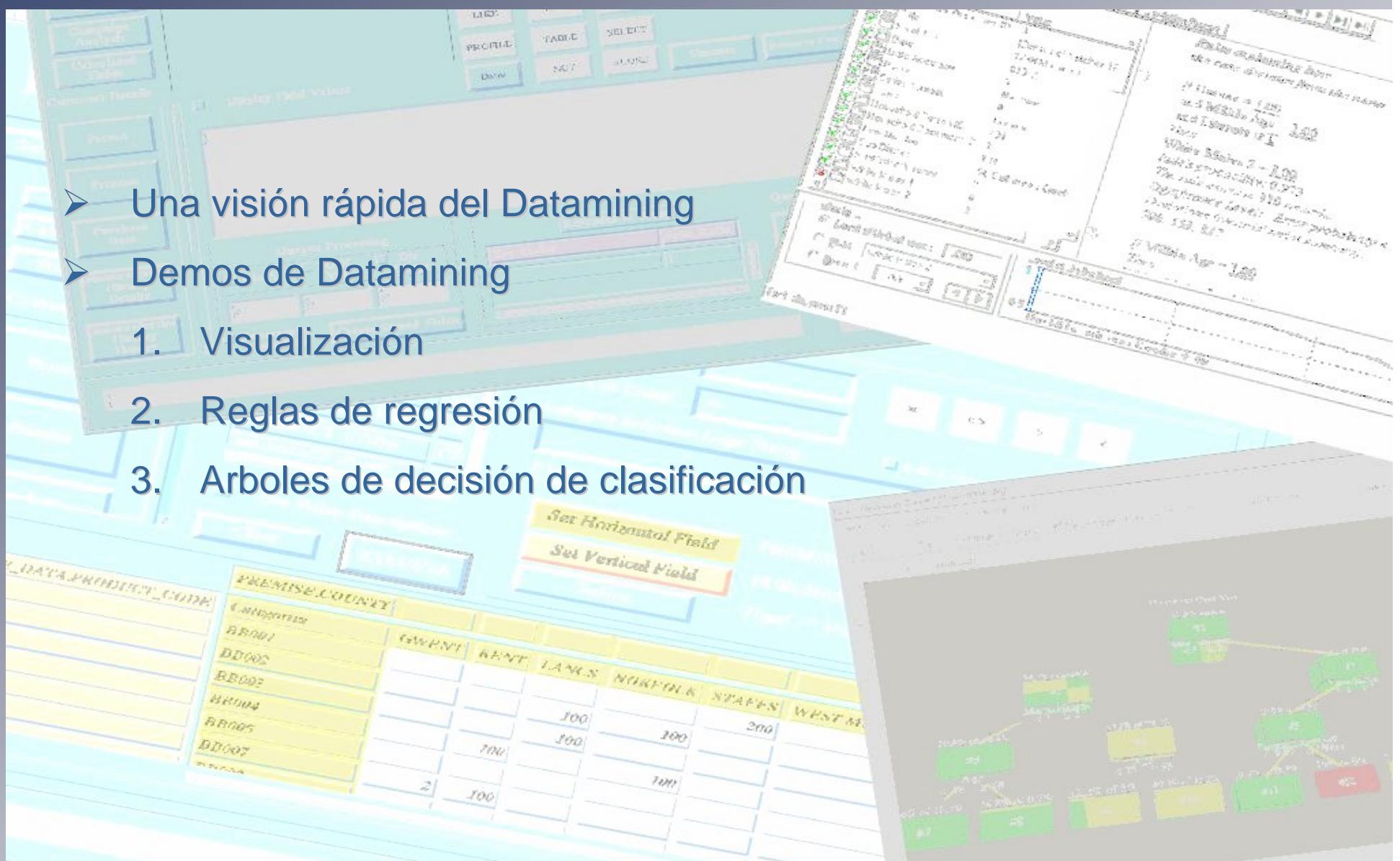
SISTEMAS INFORMATICOS I

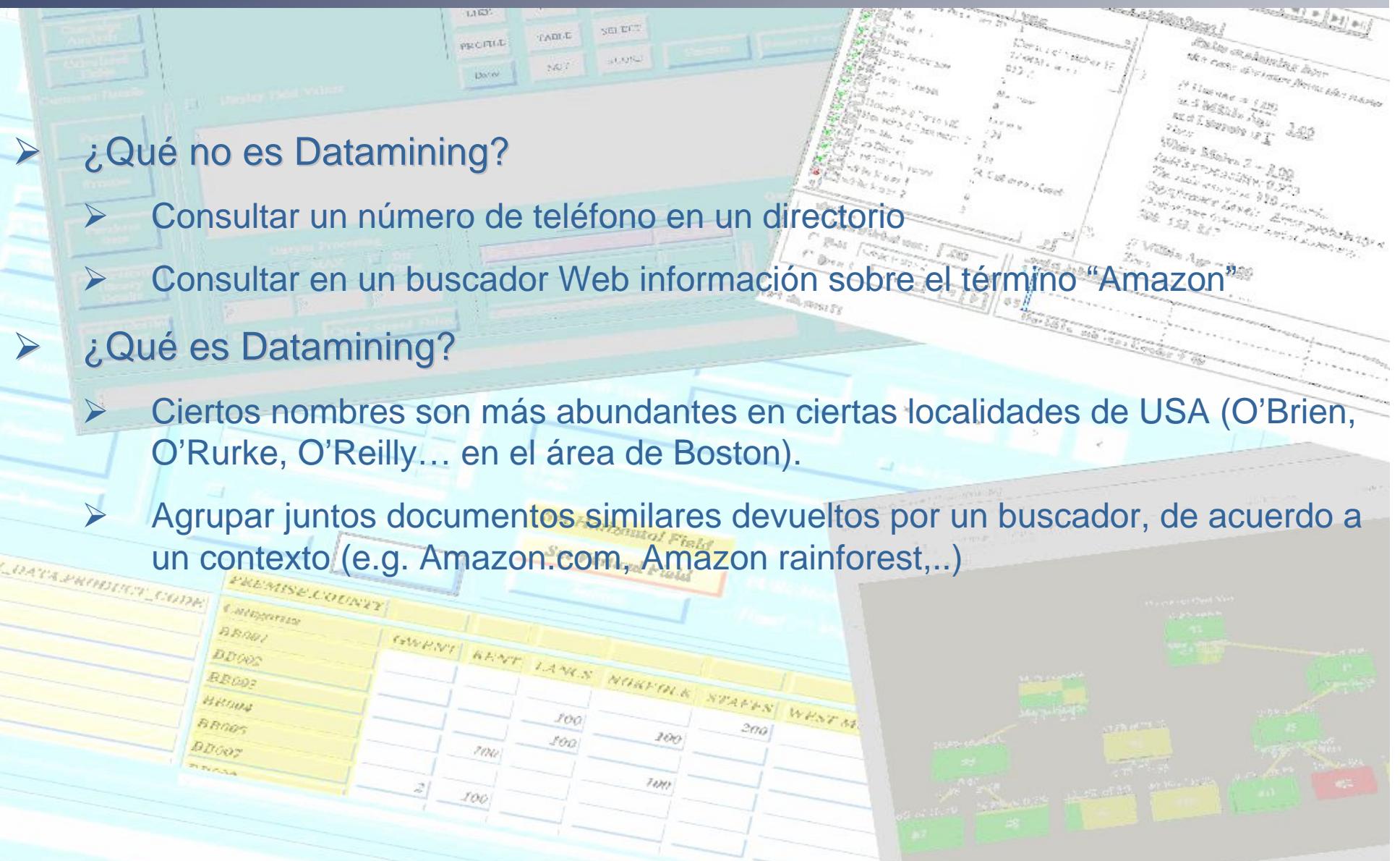
DATAMINING

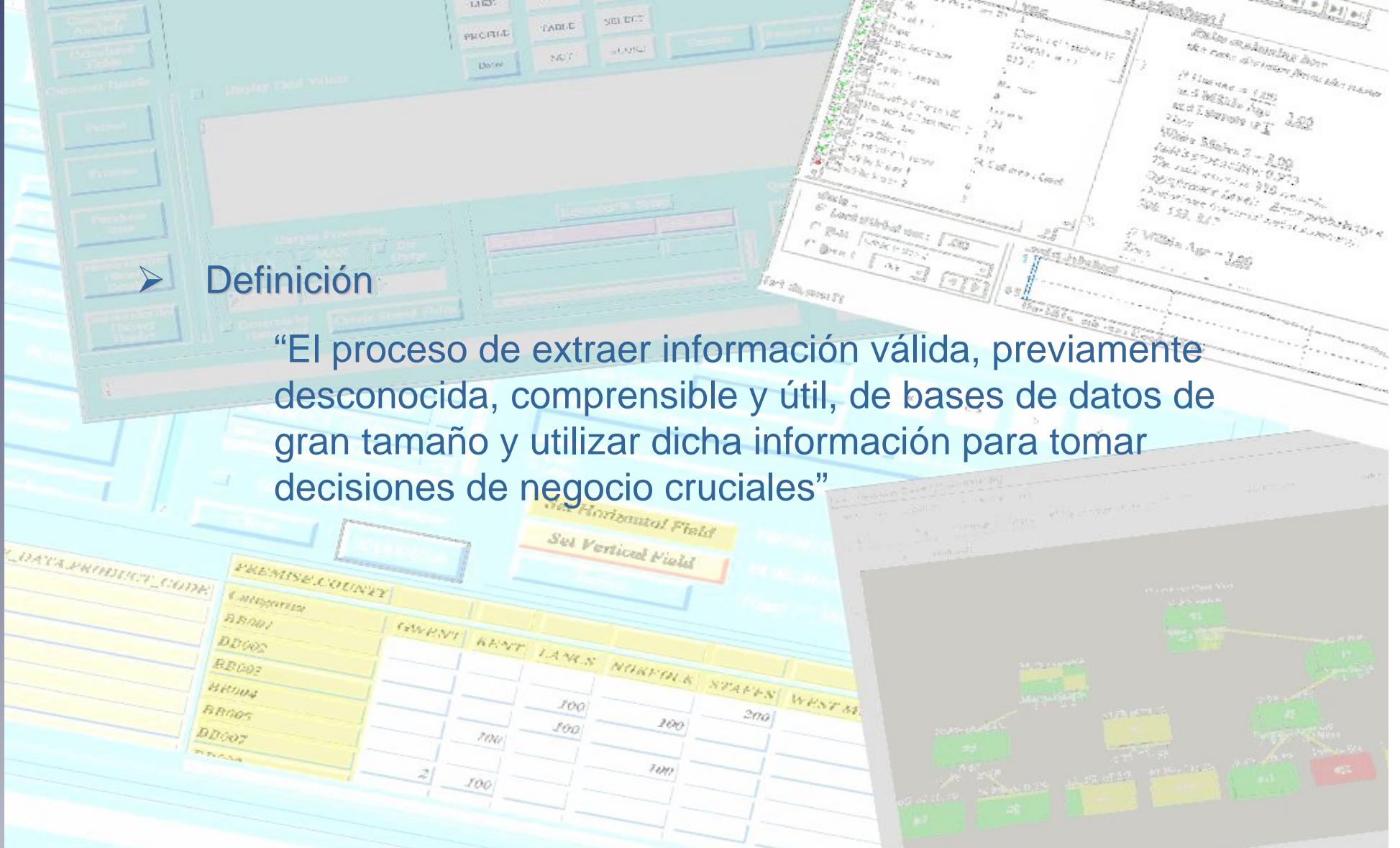
- **Introducción**
- **Datamining vs otras técnicas**
- **Ejemplos del Datamining**
- **Conceptos relacionados con el Datamining**
- **Aplicaciones del datamining**
- **Modelado del datamining**
- **Otras consideraciones**
- **Software de datamining: R**
- **Implantación de un proyecto de datamining**
- **Repercusiones y retos del Datamining**
- **Datamining en Oracle**
- **Resumen y Bibliografía**
- **Práctica**



- Una visión rápida del Datamining
- Demos de Datamining
 1. Visualización
 2. Reglas de regresión
 3. Arboles de decisión de clasificación

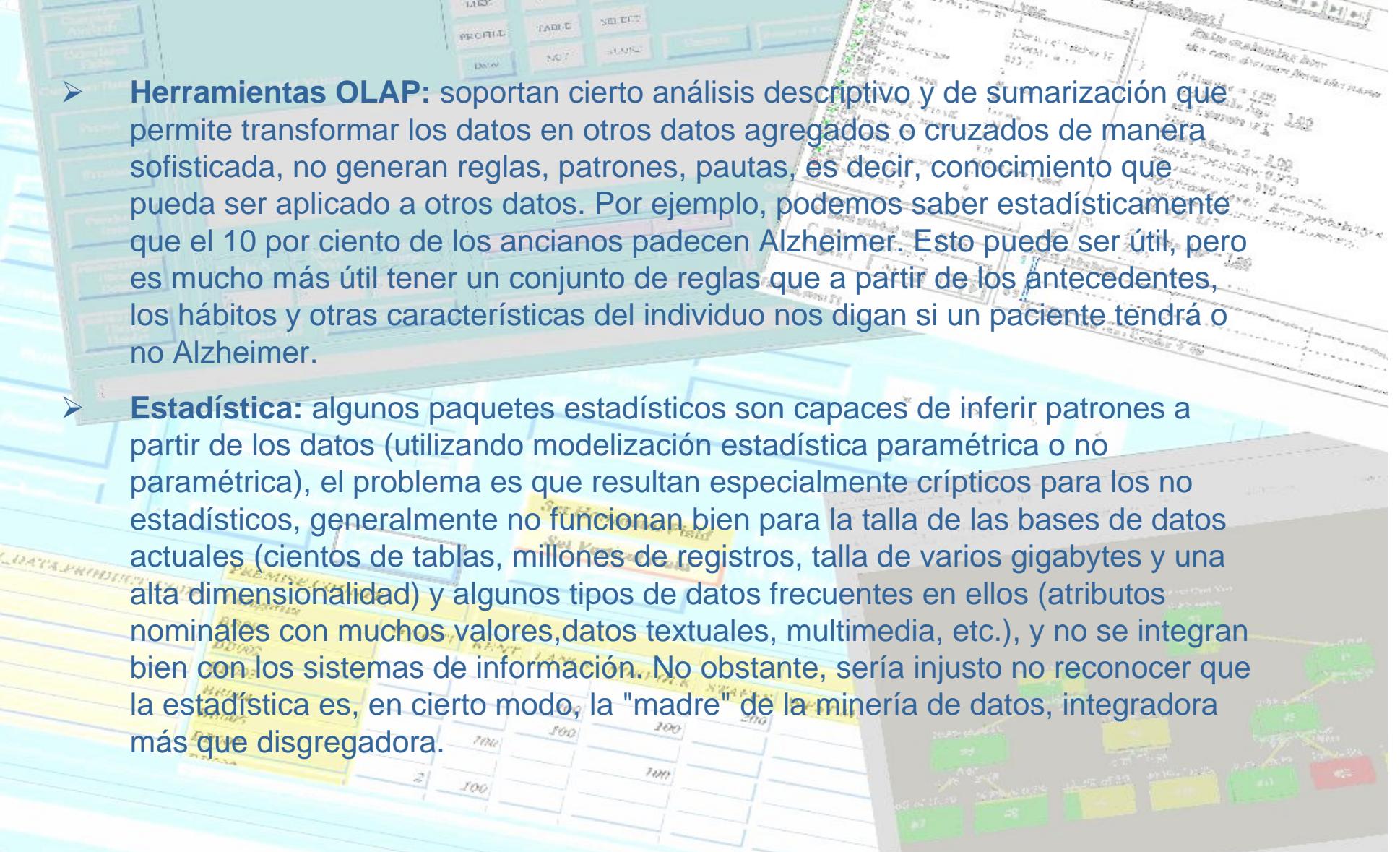


- 
- ¿Qué no es Datamining?
 - Consultar un número de teléfono en un directorio
 - Consultar en un buscador Web información sobre el término “Amazon”
 - ¿Qué es Datamining?
 - Ciertos nombres son más abundantes en ciertas localidades de USA (O'Brien, O'Rurke, O'Reilly... en el área de Boston).
 - Agrupar juntos documentos similares devueltos por un buscador, de acuerdo a un contexto (e.g. Amazon.com, Amazon rainforest,...)

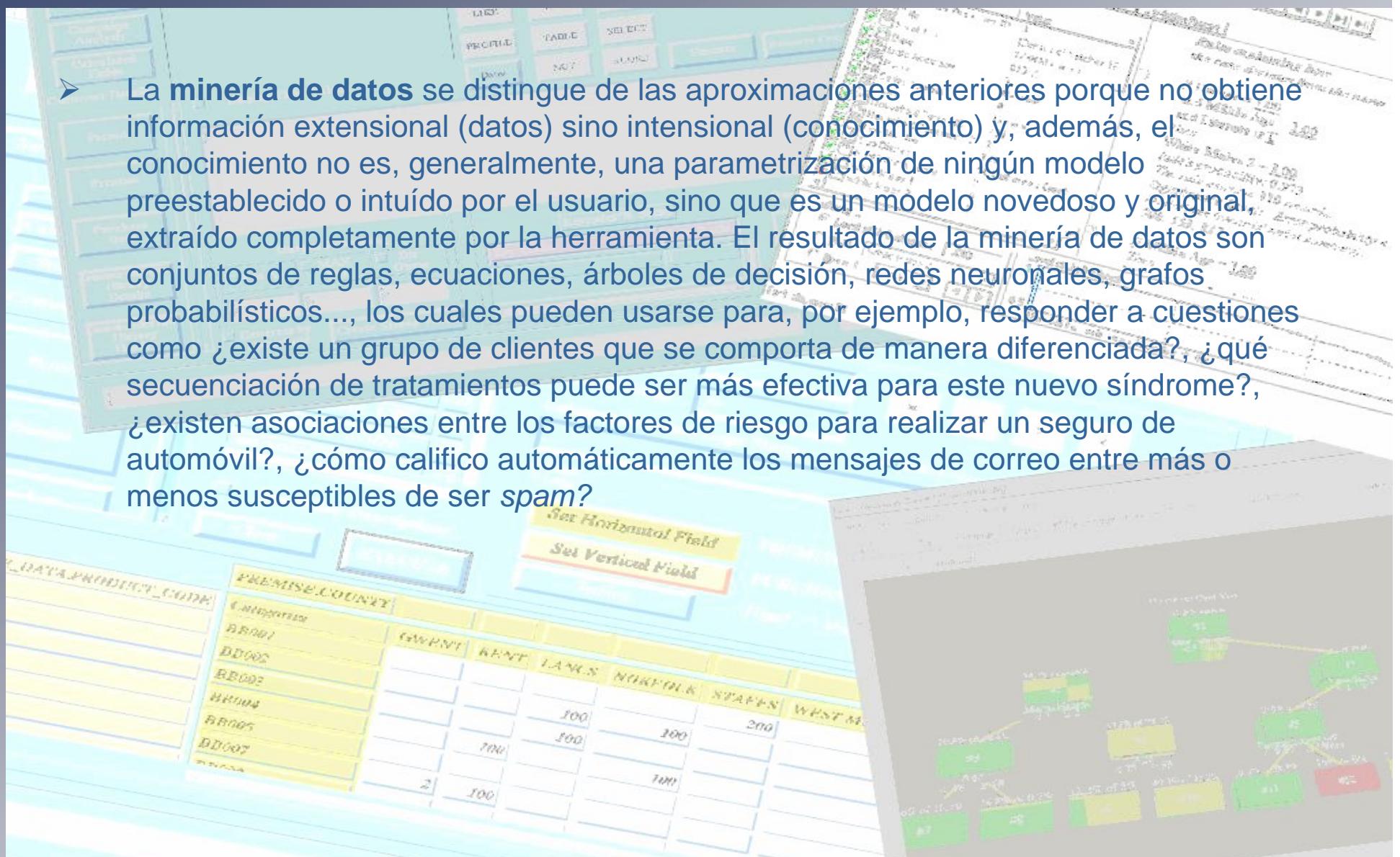


➤ Definición

“El proceso de extraer información válida, previamente desconocida, comprensible y útil, de bases de datos de gran tamaño y utilizar dicha información para tomar decisiones de negocio cruciales”

- 
- **Herramientas OLAP:** soportan cierto análisis descriptivo y de summarización que permite transformar los datos en otros datos agregados o cruzados de manera sofisticada, no generan reglas, patrones, pautas, es decir, conocimiento que pueda ser aplicado a otros datos. Por ejemplo, podemos saber estadísticamente que el 10 por ciento de los ancianos padecen Alzheimer. Esto puede ser útil, pero es mucho más útil tener un conjunto de reglas que a partir de los antecedentes, los hábitos y otras características del individuo nos digan si un paciente tendrá o no Alzheimer.
 - **Estadística:** algunos paquetes estadísticos son capaces de inferir patrones a partir de los datos (utilizando modelización estadística paramétrica o no paramétrica), el problema es que resultan especialmente crípticos para los no estadísticos, generalmente no funcionan bien para la talla de las bases de datos actuales (cientos de tablas, millones de registros, talla de varios gigabytes y una alta dimensionalidad) y algunos tipos de datos frecuentes en ellos (atributos nominales con muchos valores, datos textuales, multimedia, etc.), y no se integran bien con los sistemas de información. No obstante, sería injusto no reconocer que la estadística es, en cierto modo, la "madre" de la minería de datos, integradora más que disgregadora.

- La minería de datos se distingue de las aproximaciones anteriores porque no obtiene información extensional (datos) sino intensional (conocimiento) y, además, el conocimiento no es, generalmente, una parametrización de ningún modelo preestablecido o intuído por el usuario, sino que es un modelo novedoso y original, extraído completamente por la herramienta. El resultado de la minería de datos son conjuntos de reglas, ecuaciones, árboles de decisión, redes neuronales, grafos probabilísticos..., los cuales pueden usarse para, por ejemplo, responder a cuestiones como ¿existe un grupo de clientes que se comporta de manera diferenciada?, ¿qué secuenciación de tratamientos puede ser más efectiva para este nuevo síndrome?, ¿existen asociaciones entre los factores de riesgo para realizar un seguro de automóvil?, ¿cómo califico automáticamente los mensajes de correo entre más o menos susceptibles de ser spam?



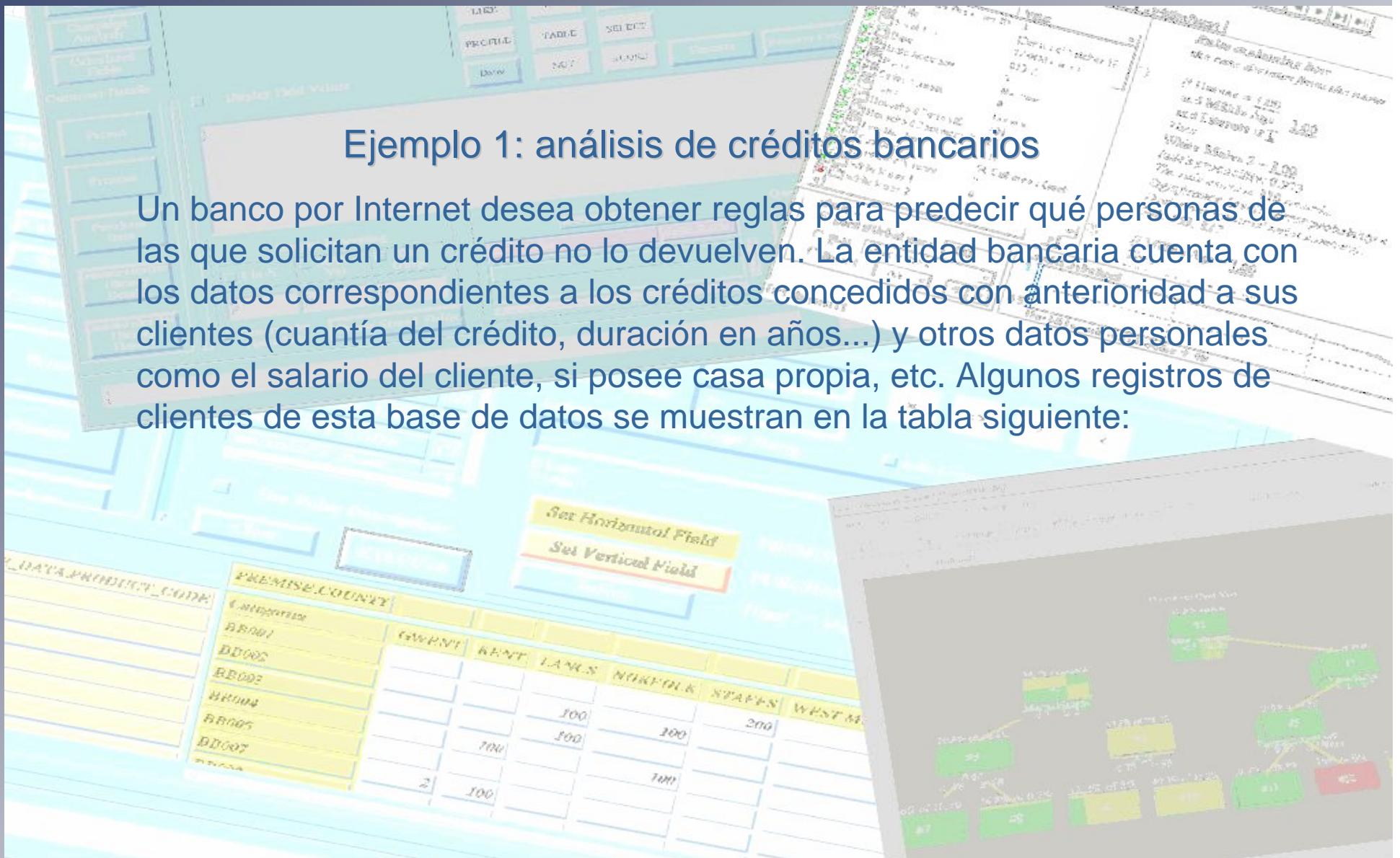
Ejemplos del Datamining

Celia Gutiérrez Cossío

2007

Ejemplo 1: análisis de créditos bancarios

Un banco por Internet desea obtener reglas para predecir qué personas de las que solicitan un crédito no lo devuelven. La entidad bancaria cuenta con los datos correspondientes a los créditos concedidos con anterioridad a sus clientes (cuantía del crédito, duración en años...) y otros datos personales como el salario del cliente, si posee casa propia, etc. Algunos registros de clientes de esta base de datos se muestran en la tabla siguiente:



Ejemplos del Datamining

Celia Gutiérrez Cossío
2007

IDC	D-crédito	C-crédito	Salario	Casa	Cuentas	Devuelve-
	(años)	(euros)	(euros)	propia	morosas	crédito
101	15	60.000	2.200	sí	2	no
102	2	30.000	3.500	sí	O	sí
103	9	9.000	1.700	sí	1	no
104	15	18.000	1.900	no	O	sí
105	10	24.000	2.100	no	O	no
oo.

A partir de éstos, las técnicas de minería de datos podrían sintetizar algunas reglas, como por ejemplo:

SI Cuentas-Morosas > 0 ENTONCES Devuelve-crédito = no

SI Cuentas-Morosas = 0 Y [(Salario > 2.500) O (D-crédito > 10)] ENTONCES Devuelve-crédito = sí

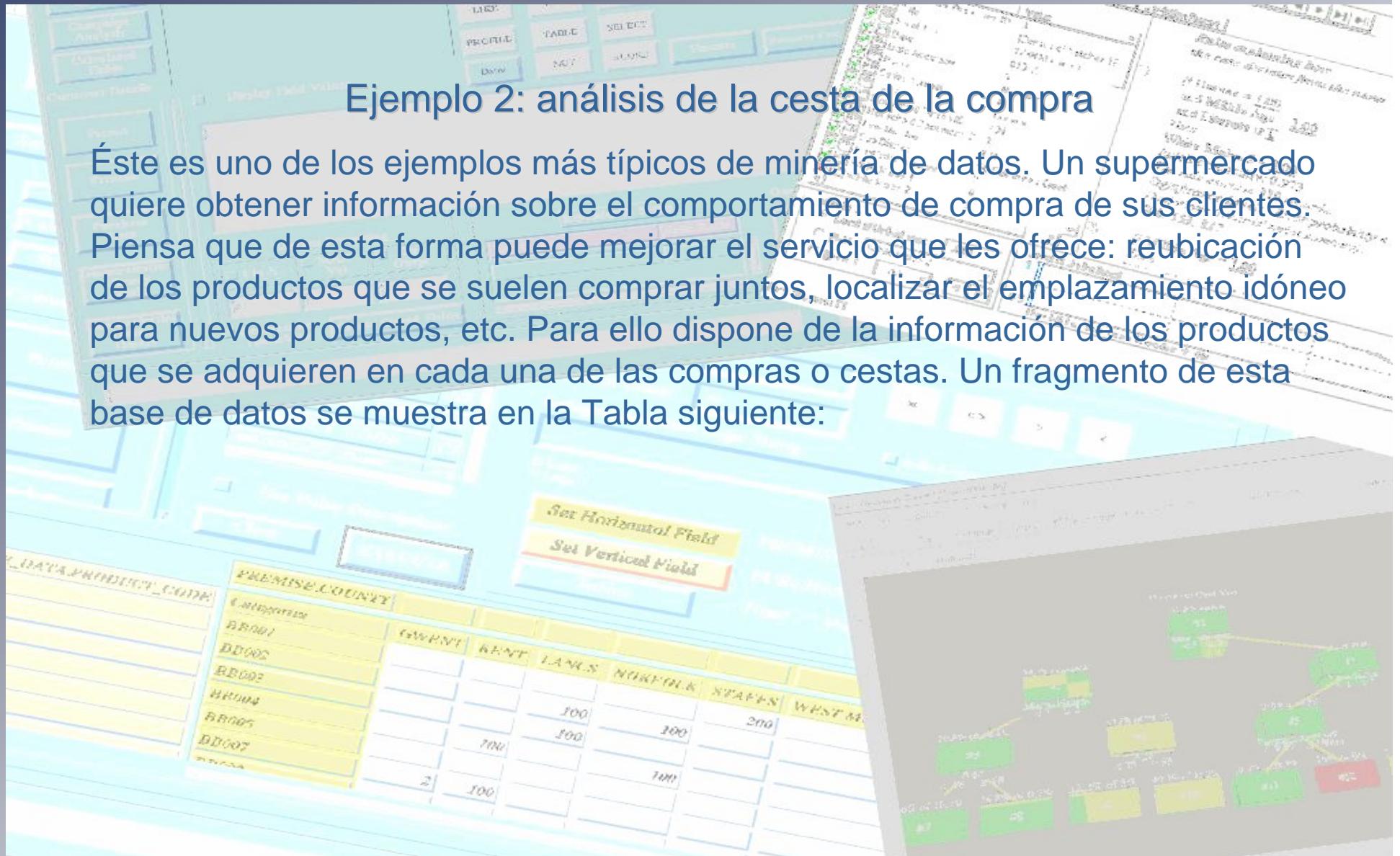
El banco podría entonces utilizar estas reglas para determinar las acciones a realizar en el trámite de los créditos: si se concede o no el crédito solicitado, si es necesario pedir avales especiales, etc.

Ejemplos del Datamining

Celia Gutiérrez Cossío
2007

Ejemplo 2: análisis de la cesta de la compra

Éste es uno de los ejemplos más típicos de minería de datos. Un supermercado quiere obtener información sobre el comportamiento de compra de sus clientes. Piensa que de esta forma puede mejorar el servicio que les ofrece: reubicación de los productos que se suelen comprar juntos, localizar el emplazamiento idóneo para nuevos productos, etc. Para ello dispone de la información de los productos que se adquieren en cada una de las compras o cestas. Un fragmento de esta base de datos se muestra en la Tabla siguiente:



Ejemplos del Datamining

Celia Gutiérrez Cossío
2007

Idcesta	Huevos	Aceite	Pañales	Vino	leche	Mantequilla	Salmón	lechugas
1	sí	no	no	sí	no	sí	sí	sí
2	no	sí	no	no	sí	no	no	sí
3	no	no	sí	no	sí	no	no	no
4	no	sí	sí	no	sí	no	no	no
5	sí	sí	no	no	no	sí	no	sí
6	sí	no	no	sí	sí	sí	sí	no
7	no	no	no	no	no	no	no	no
8	sí	sí	sí	sí	sí	sí	sí	no

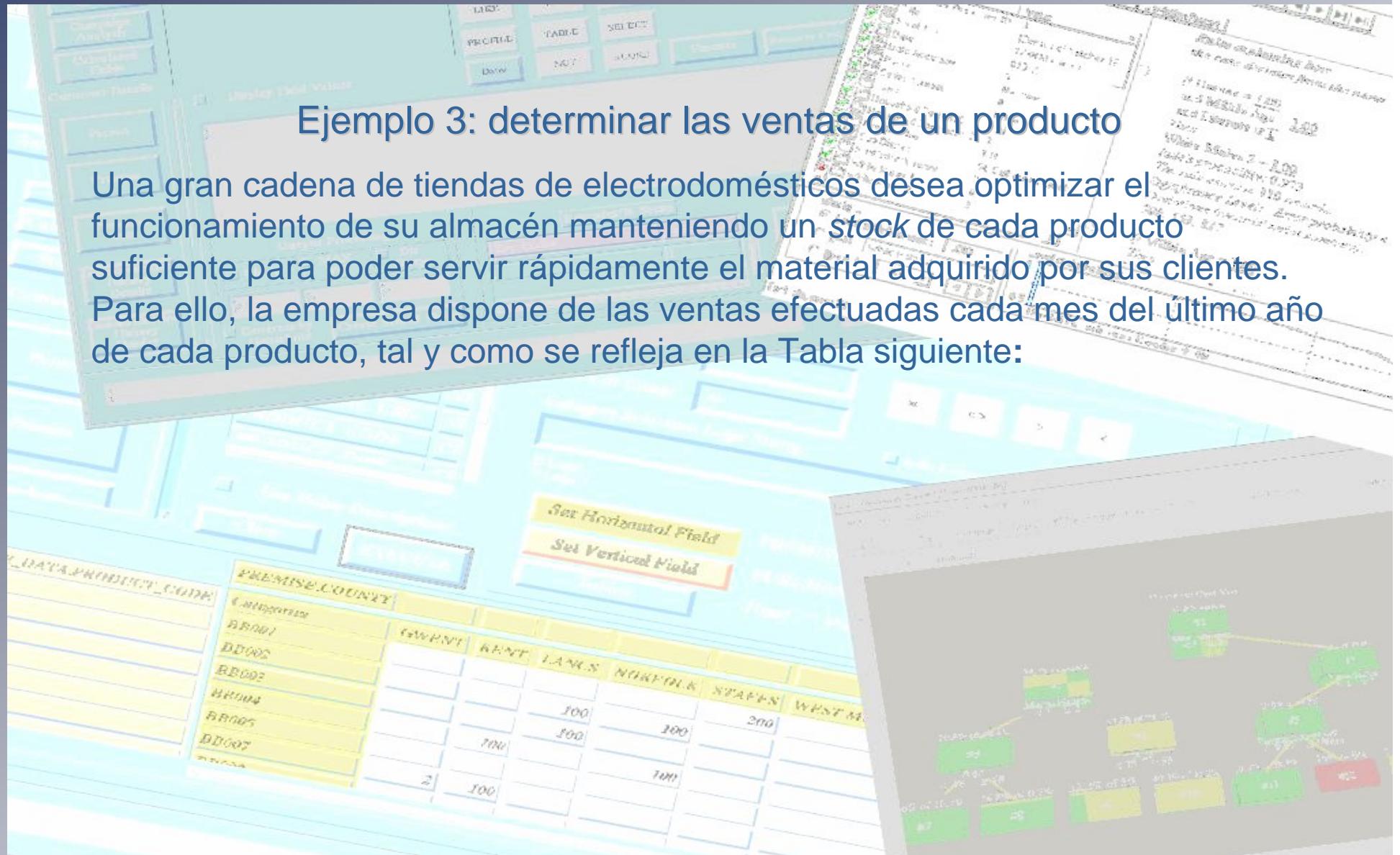
Analizando estos datos el supermercado podría encontrar, por ejemplo, que el 100 por cien de las veces que se compran pañales también se compra leche, que el 50 por ciento de las veces que se compran huevos también se compra aceite o que el 33 por ciento de las veces que se compra vino y salmón entonces se compran lechugas. También se puede analizar cuáles de estas asociaciones son frecuentes, porque una asociación muy estrecha entre dos productos puede ser poco frecuente y, por tanto, poco útil.

Ejemplos del Datamining

Celia Gutiérrez Cossío
2007

Ejemplo 3: determinar las ventas de un producto

Una gran cadena de tiendas de electrodomésticos desea optimizar el funcionamiento de su almacén manteniendo un stock de cada producto suficiente para poder servir rápidamente el material adquirido por sus clientes. Para ello, la empresa dispone de las ventas efectuadas cada mes del último año de cada producto, tal y como se refleja en la Tabla siguiente:



Ejemplos del Datamining

Celia Gutiérrez Cossío
2007

Producto	mes-12	...	mes-4	mes-3	mes-2	mes-1
televisor plano 30' Phlipis	20	...	52	14	139	74
vídeo-dvd-recorder Miesens	11	...	43	32	26	59
discman mp3 LJ	50	...	61	14	5	28
frigorífico no frost Jazzussi	3	...	21	27	1	49
microondas con grill Sanson	14	...	27	2	25	12
...

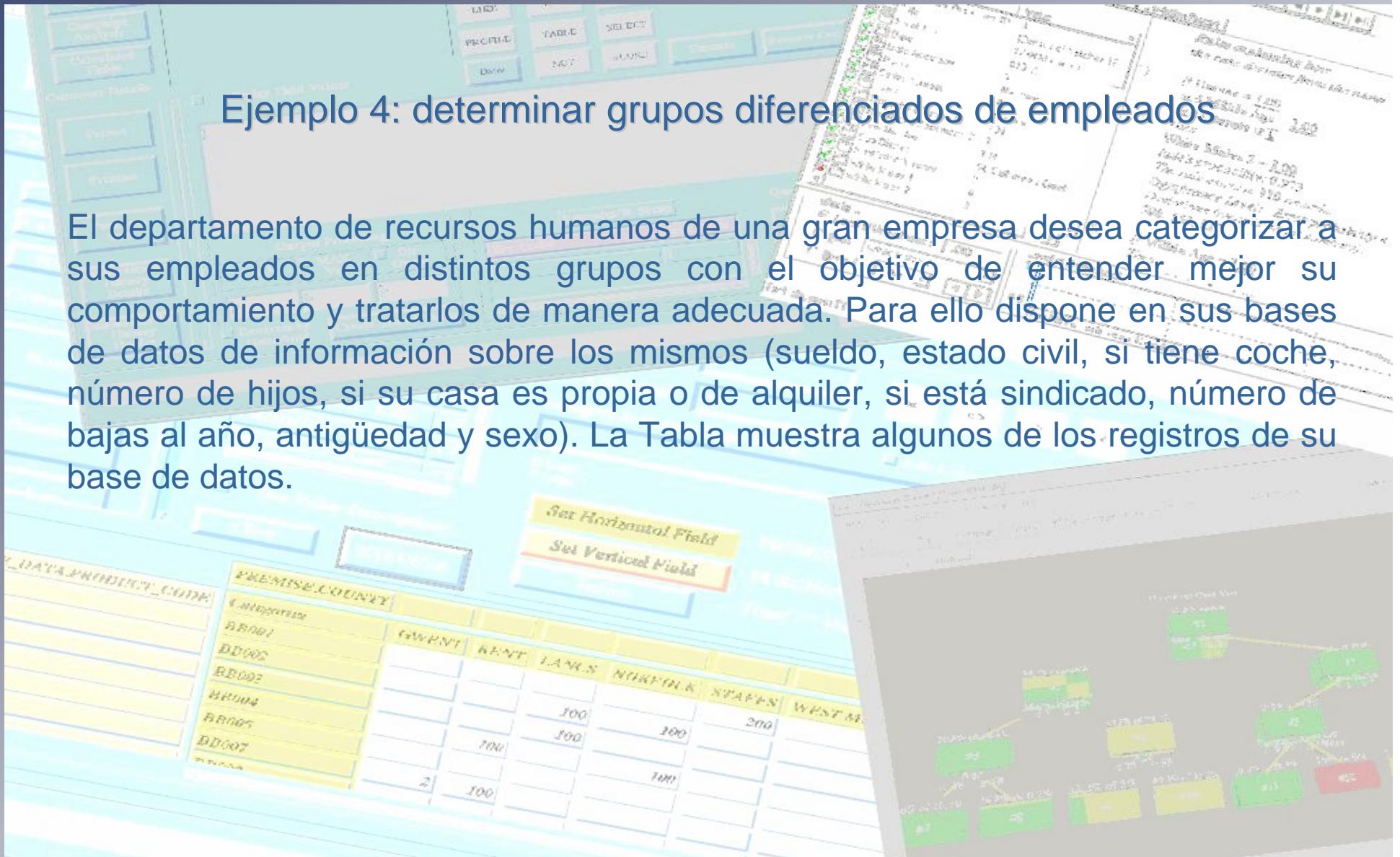
Esta información permite a la empresa generar un modelo para predecir cuáles van a ser las ventas de cada producto en el siguiente mes en función de las ventas realizadas en los meses anteriores, y efectuar así los pedidos necesarios a sus proveedores para disponer del stock necesario para hacer frente a esas ventas.

Ejemplos del Datamining

Celia Gutiérrez Cossío
2007

Ejemplo 4: determinar grupos diferenciados de empleados

El departamento de recursos humanos de una gran empresa desea categorizar a sus empleados en distintos grupos con el objetivo de entender mejor su comportamiento y tratarlos de manera adecuada. Para ello dispone en sus bases de datos de información sobre los mismos (sueldo, estado civil, si tiene coche, número de hijos, si su casa es propia o de alquiler, si está sindicado, número de bajas al año, antigüedad y sexo). La Tabla muestra algunos de los registros de su base de datos.



Ejemplos del Datamining

Celia Gutiérrez Cossío
2007

Id	Sueldo	Casado	Coche	Hijos	Alq/prop	Sindicado	Bajas/año	Antigüedad	Sexo
1	1.000	Sí	No	0	Alquiler	No	7	15	H
2	2.000	No	Sí	1	Alquiler	Sí	3	3	M
3	1.500	Sí	Sí	2	Prop	Sí	5	10	H
4	3.000	Sí	Sí	1	Alquiler	No	15	7	M
5	1.000	Sí	Sí	0	Prop	Sí	1	6	H
6	4.000	No	Sí	0	Alquiler	Sí	3	16	M
7	2.500	No	No	0	Alquiler	Sí	0	8	H
8	2.000	No	Sí	0	Prop	Sí	2	6	M
9	2.000	Sí	Sí	3	Prop	No	7	5	H
10	3.000	Sí	Sí	2	Prop	No	1	20	H
11	5.000	No	No	0	Alquiler	No	2	12	M
12	800	Sí	Sí	2	Prop	No	3	1	H
13	2.000	No	No	0	Alquiler	No	27	5	M
14	1.000	No	Sí	0	Alquiler	Sí	0	7	H
15	8 00	No	Sí	0	Alquiler	No	3	2	H
...

Un sistema de minería de datos podría obtener tres grupos con la siguiente descripción:

Ejemplos del Datamining

Celia Gutiérrez Cossío
2007

Grupo 1:

Sueldo: 1.535,2€
Casado: No -> 0,777
 Sí -> 0,223
Coche: No -> 0,82
 Sí -> 0,18
Hijos: 0,05
Alq/Prop: Alquiler -> 0,99
 Propia -> 0,01
Sindic.: No -> 0,8
 Sí -> 0,2
Bajas/Año: 8,3
Antigüedad: 8,7
Sexo: H -> 0,61
 M -> 0,39

Grupo 2:

Sueldo: 1.428,7€
Casado: No -> 0,98
 Sí -> 0,02
Coche: No -> 0,01
 Sí -> 0,99
Hijos: 0,3
Alq/Prop: Alquiler -> 0,75
 Propia -> 0,25
Sindic.: Sí -> 1,0
Bajas/Año: 2,3
Antigüedad: 8
Sexo: H -> 0,25
 M -> 0,75

Grupo 3:

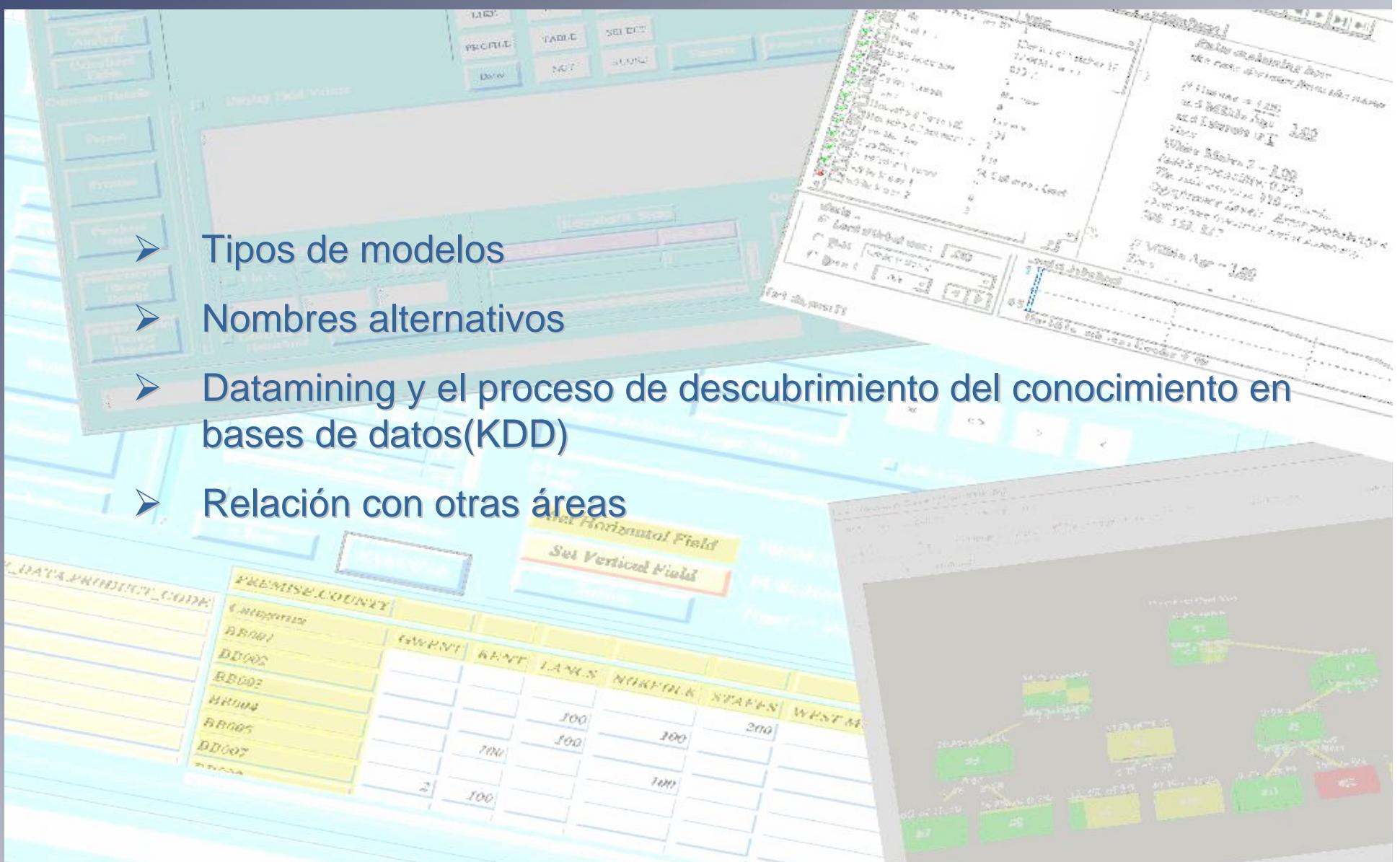
Sueldo: 1.233,8€
Casado: Sí -> 1,0
Coche: No -> 0,05
 Sí -> 0,95
Hijos: 2,3
Alq/Prop: Alquiler -> 0,17
 Propia -> 0,83
Sindic.: No -> 0,67
 Sí -> 0,33
Bajas/Año: 5,1
Antigüedad: 8,1
Sexo: H -> 0,83
 M -> 0,17



Conceptos relacionados con el Datamining

Celia Gutiérrez Cossío
2007

- Tipos de modelos
- Nombres alternativos
- Datamining y el proceso de descubrimiento del conocimiento en bases de datos(KDD)
- Relación con otras áreas



➤ Tipos de modelos:

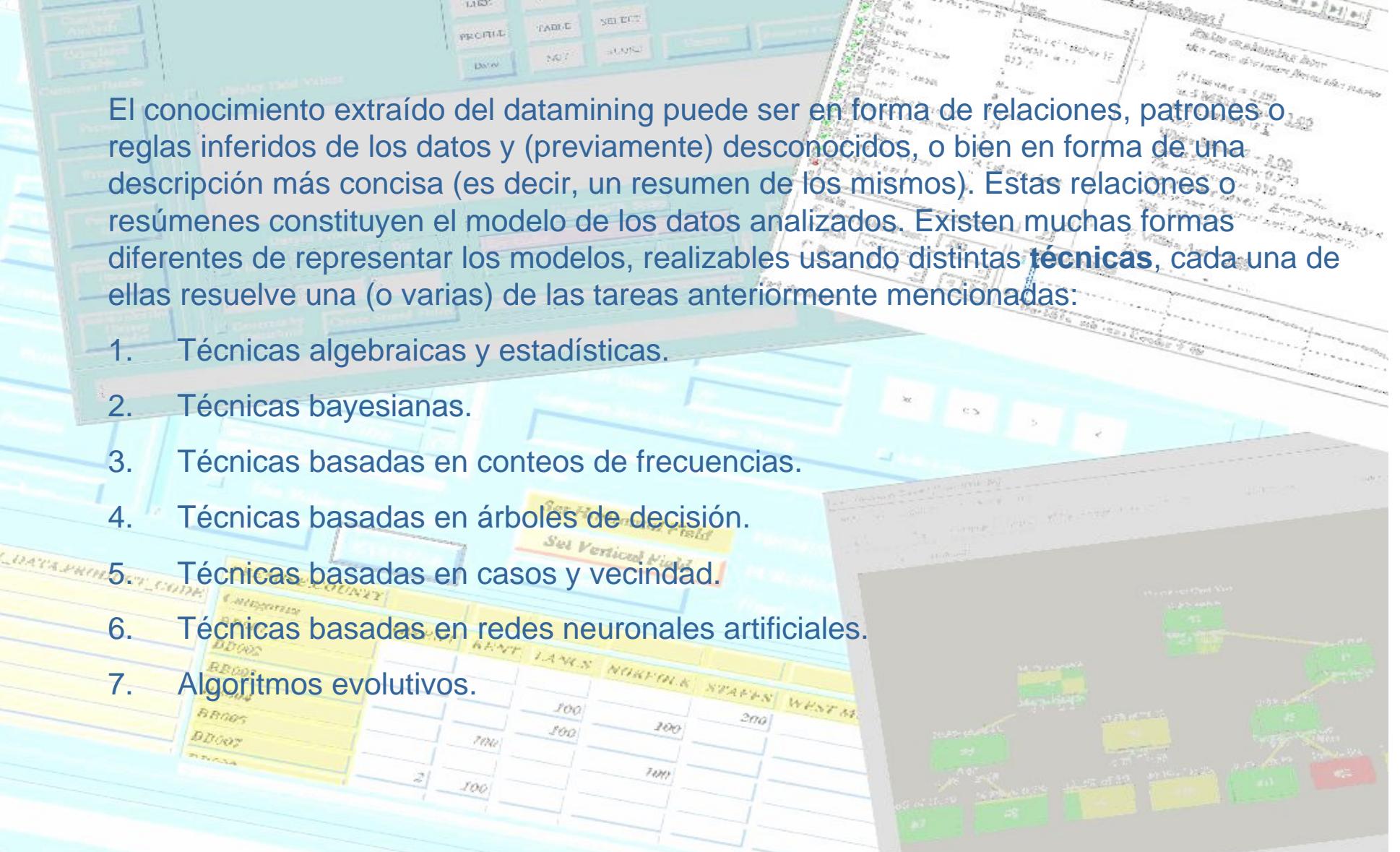
Las **tareas** de datamining se clasifican en:

1. Clasificación
2. Regresión
3. Agrupamiento
4. Correlaciones
5. Reglas de asociación
6. Reglas de asociación secuenciales



Conceptos relacionados con el Datamining

Celia Gutiérrez Cossío
2007



El conocimiento extraído del datamining puede ser en forma de relaciones, patrones o reglas inferidos de los datos y (previamente) desconocidos, o bien en forma de una descripción más concisa (es decir, un resumen de los mismos). Estas relaciones o resúmenes constituyen el modelo de los datos analizados. Existen muchas formas diferentes de representar los modelos, realizables usando distintas **técnicas**, cada una de ellas resuelve una (o varias) de las tareas anteriormente mencionadas:

1. Técnicas algebraicas y estadísticas.
2. Técnicas bayesianas.
3. Técnicas basadas en conteos de frecuencias.
4. Técnicas basadas en árboles de decisión.
5. Técnicas basadas en casos y vecindad.
6. Técnicas basadas en redes neuronales artificiales.
7. Algoritmos evolutivos.

Conceptos relacionados con el Datamining

Celia Gutiérrez Cossío

2007

Los tipos de modelos que se pueden implementar con técnicas de minería de datos son:

Modelos predictivos: pretenden estimar valores futuros o desconocidos de variables de interés, que denominamos *variables objetivo* o *dependientes*, usando otras variables o campos de la base de datos, a las que nos referiremos como *variables independientes* o *predictivas*. Por ejemplo, un modelo predictivo sería aquel que permite estimar la demanda de un nuevo producto en función del gasto en publicidad. Se incluyen la clasificación y la regresión. Los ejemplos 1 y 3 son modelos predictivos.

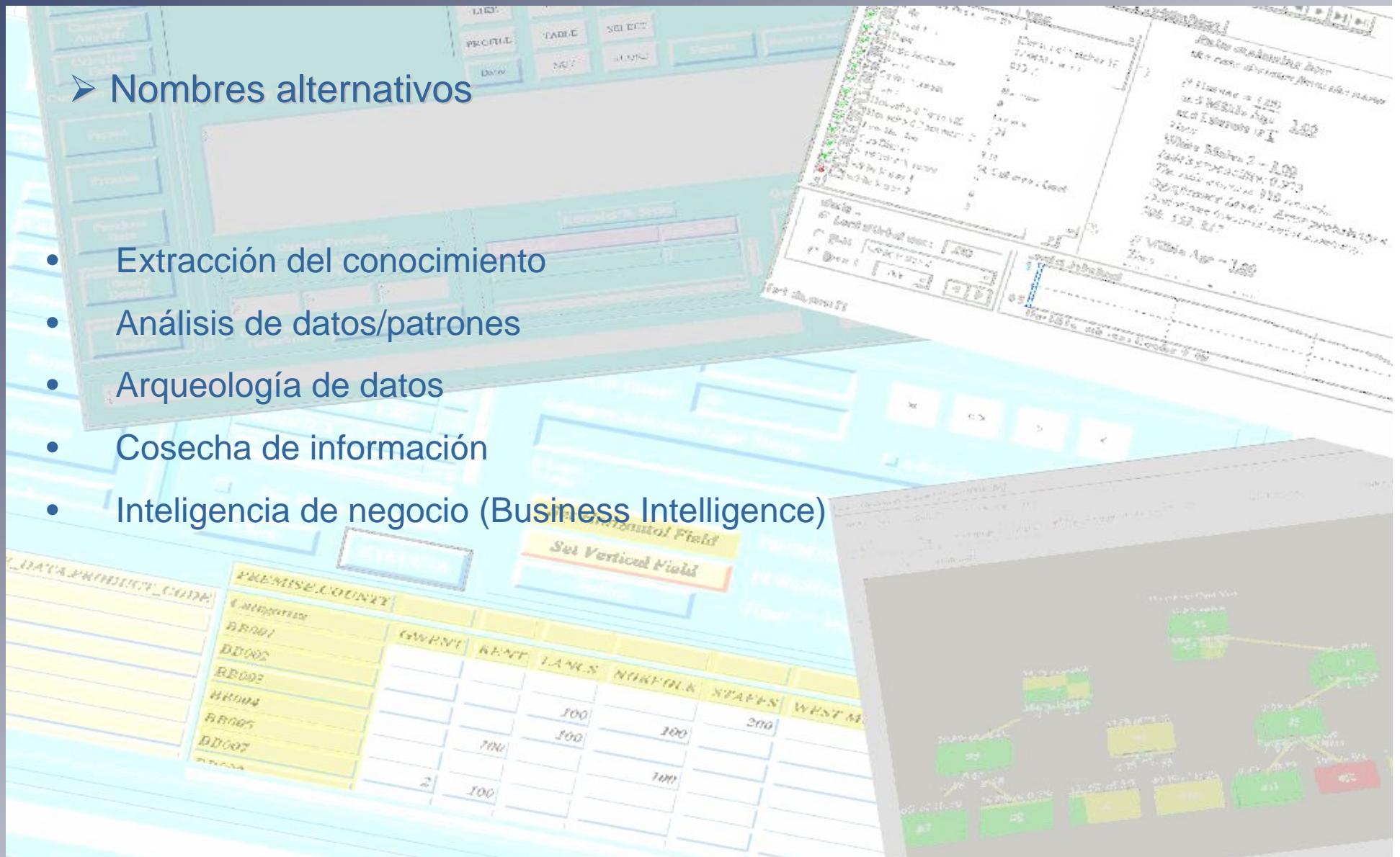
Modelos descriptivos: identifican patrones que explican o resumen los datos, es decir, sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos. Por ejemplo, una agencia de viaje desea identificar grupos de personas con unos mismos gustos, con el objeto de organizar diferentes ofertas para cada grupo y poder así remitirles esta información; para ello analiza los viajes que han realizado sus clientes e infiere un modelo descriptivo que caracteriza estos grupos. Se incluyen agrupamiento, las reglas de asociación y el análisis correlacional. Los ejemplos 2 y 4 son descriptivos.

Conceptos relacionados con el Datamining

Celia Gutiérrez Cossío
2007

➤ Nombres alternativos

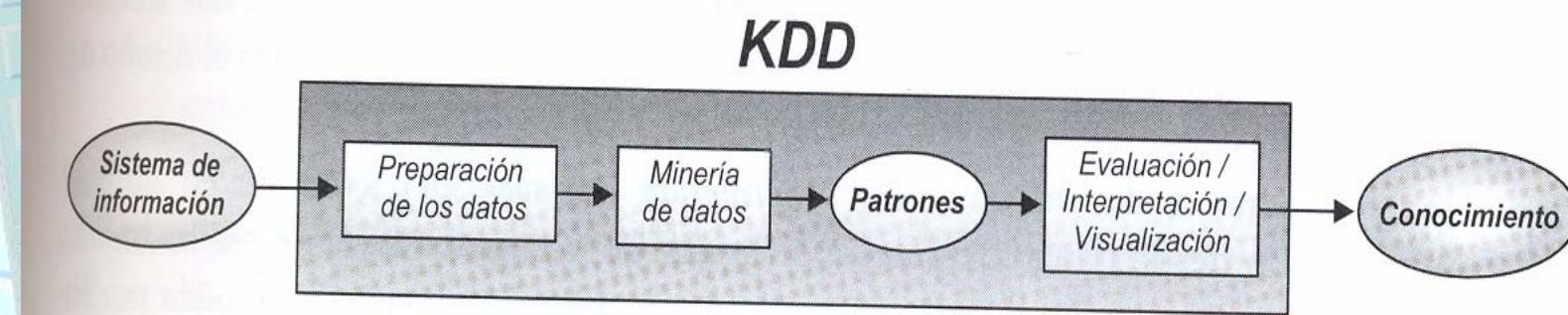
- Extracción del conocimiento
- Análisis de datos/patrones
- Arqueología de datos
- Cosecha de información
- Inteligencia de negocio (Business Intelligence)



Conceptos relacionados con el Datamining

Celia Gutiérrez Cossío
2007

- Datamining y el proceso de descubrimiento del conocimiento en bases de datos(KDD)

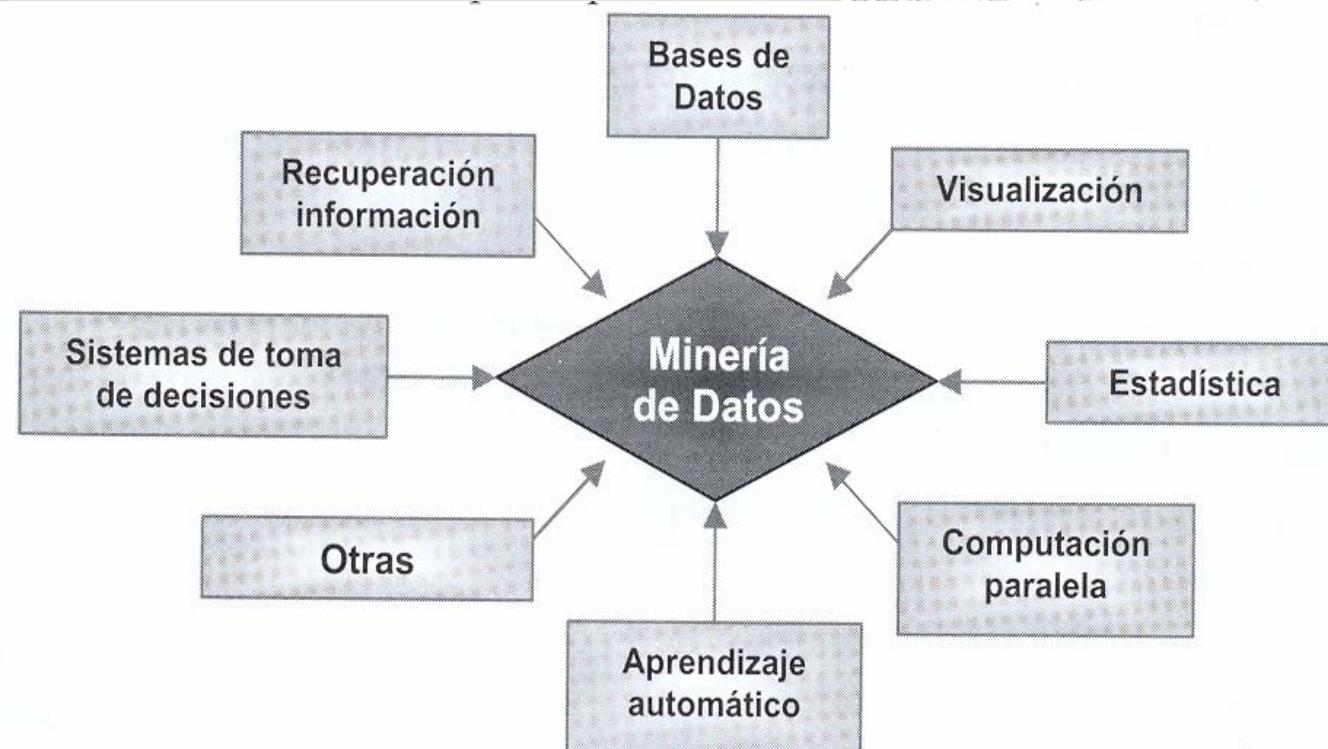


KDD y Datamining son términos habitualmente confundidos; en realidad el datamining está englobado en el proceso KDD, si bien es una parte fundamental, ya que extrae nuevos modelos y patrones de conocimiento.

Conceptos relacionados con el Datamining

Celia Gutiérrez Cossío
2007

➤ Relación con otras disciplinas



Conceptos relacionados con el Datamining

Celia Gutiérrez Cossío
2007

➤ Relación con otras disciplinas

Las técnicas tradicionales pueden no ser adecuadas por:

- La enorme cantidad de información => los algoritmos deben ser escalables para manejar datos masivos.
- La alta dimensionalidad de los datos.
- La naturaleza distribuída y heterogénea de los datos => datos espaciales, multimedia, series de tiempo, gráficos,



DATA_PRODUCT_CODE	PREMISE COUNTY	SWPNL	HWPNL	LANS	NWPNL	SWPNL	WEST M
DD001				100		200	
DD002				100		200	
DD003				100		200	
DD004				100		200	
DD005				100		200	
DD007				100		200	
DD008				100		200	

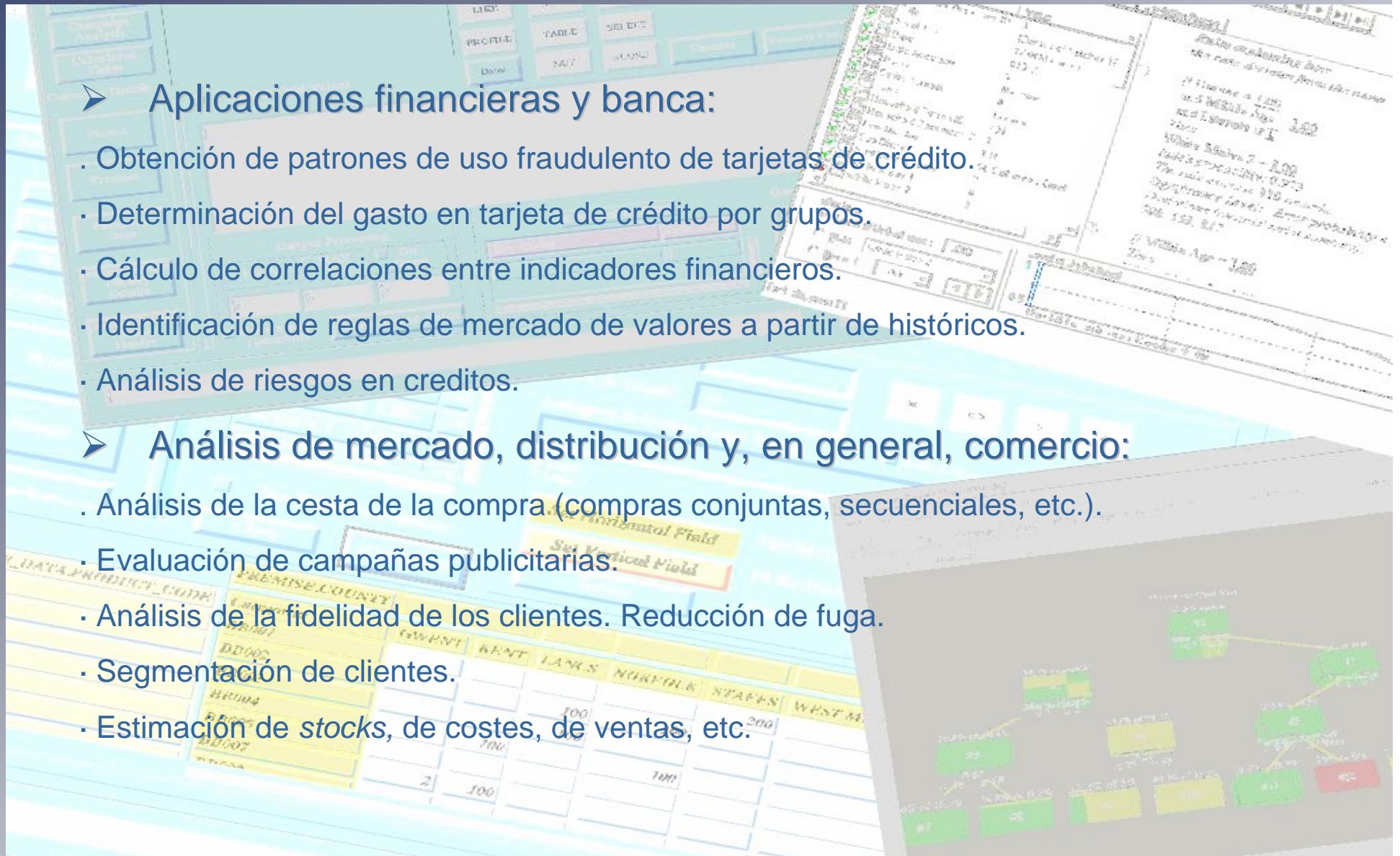


➤ Aplicaciones financieras y banca:

- Obtención de patrones de uso fraudulento de tarjetas de crédito.
- Determinación del gasto en tarjeta de crédito por grupos.
- Cálculo de correlaciones entre indicadores financieros.
- Identificación de reglas de mercado de valores a partir de históricos.
- Análisis de riesgos en créditos.

➤ Análisis de mercado, distribución y, en general, comercio:

- Análisis de la cesta de la compra (compras conjuntas, secuenciales, etc.).
- Evaluación de campañas publicitarias.
- Análisis de la fidelidad de los clientes. Reducción de fuga.
- Segmentación de clientes.
- Estimación de stocks, de costes, de ventas, etc.



➤ Seguros y salud privada:

- Determinación de los clientes que podrían ser potencialmente caros.
- Análisis de procedimientos médicos solicitados conjuntamente.
- Predicción de qué clientes contratan nuevas pólizas.
- Identificación de patrones de comportamiento para clientes con riesgo.
- Identificación de comportamiento fraudulento.
- Predicción de los clientes que podrían ampliar su póliza para incluir procedimientos extras (dentales, ópticos..)

➤ Educación:

- Selección o captación de estudiantes.
- Detección de abandonos y de fracaso.
- Estimación del tiempo de estancia en la institución.

➤ Procesos industriales:

- Extracción de modelos sobre comportamiento de compuestos.
- Detección de piezas con trabas. Modelos de calidad.
- Predicción de fallos y accidentes.
- Estimación de composiciones óptimas en mezclas.
- Extracción de modelos de coste.
- Extracción de modelos de producción.

➤ Medicina:

- Identificación de patologías. Diagnóstico de enfermedades.
- Detección de pacientes con riesgo de sufrir una patología concreta.
- Gestión hospitalaria y asistencial. Predicciones temporales de los centros asistenciales para el mejor uso de recursos, consultas, salas y habitaciones.
- Recomendación priorizada de fármacos para una misma patología.

The collage consists of four overlapping screenshots:

- Top Left:** A screenshot of a software interface showing a grid of small windows or tabs, likely a data mining tool's graphical user interface.
- Top Right:** A screenshot of a software interface displaying a large table of data with numerous columns and rows, with some cells highlighted in green.
- Bottom Left:** A screenshot of a software interface showing a grid of data with various fields labeled, such as "Set Horizontal Field" and "Set Vertical Field".
- Bottom Right:** A screenshot of a software interface showing a complex network graph with nodes and edges, representing a data structure or relationship map.

➤ Biología, bioingeniería y otras ciencias:

- Análisis de secuencias de genes.
- Análisis de secuencias de proteínas.
- Predecir si un compuesto químico causa cáncer.
- Clasificación de cuerpos celestes.
- Predicción de recorrido y distribución de inundaciones.
- Modelos de calidad de aguas, indicadores ecológicos.

➤ Telecomunicaciones:

- Establecimiento de patrones de llamadas.
- Modelos de carga en redes.
- Detección de fraude.

➤ Otras áreas

- Correo electrónico y agendas personales: clasificación y distribución automática de correo, detección de correo *spam*, gestión de avisos, análisis del empleo del tiempo.
- Recursos Humanos: selección de empleados.
- Web: análisis del comportamiento de los usuarios, detección de fraude en el comercio electrónico, análisis de los *logs* de un servidor web.
- Turismo: determinar las características socioeconómicas de los turistas en un determinado destino o paquete turístico, identificar patrones de reservas, etc.
- Tráfico: modelos de tráfico a partir de fuentes diversas: cámaras, GPS...
- Hacienda: detección de evasión fiscal.
- Policiales: identificación de posibles terroristas en un aeropuerto.
- Deportes: estudio de la influencia de jugadores y de cambios. Planificación de eventos.
- Política: diseño de campañas políticas, estudios de tendencias de grupos, etc.

Datamining Modelado del datamining

Celia Gutiérrez Cossío
2007

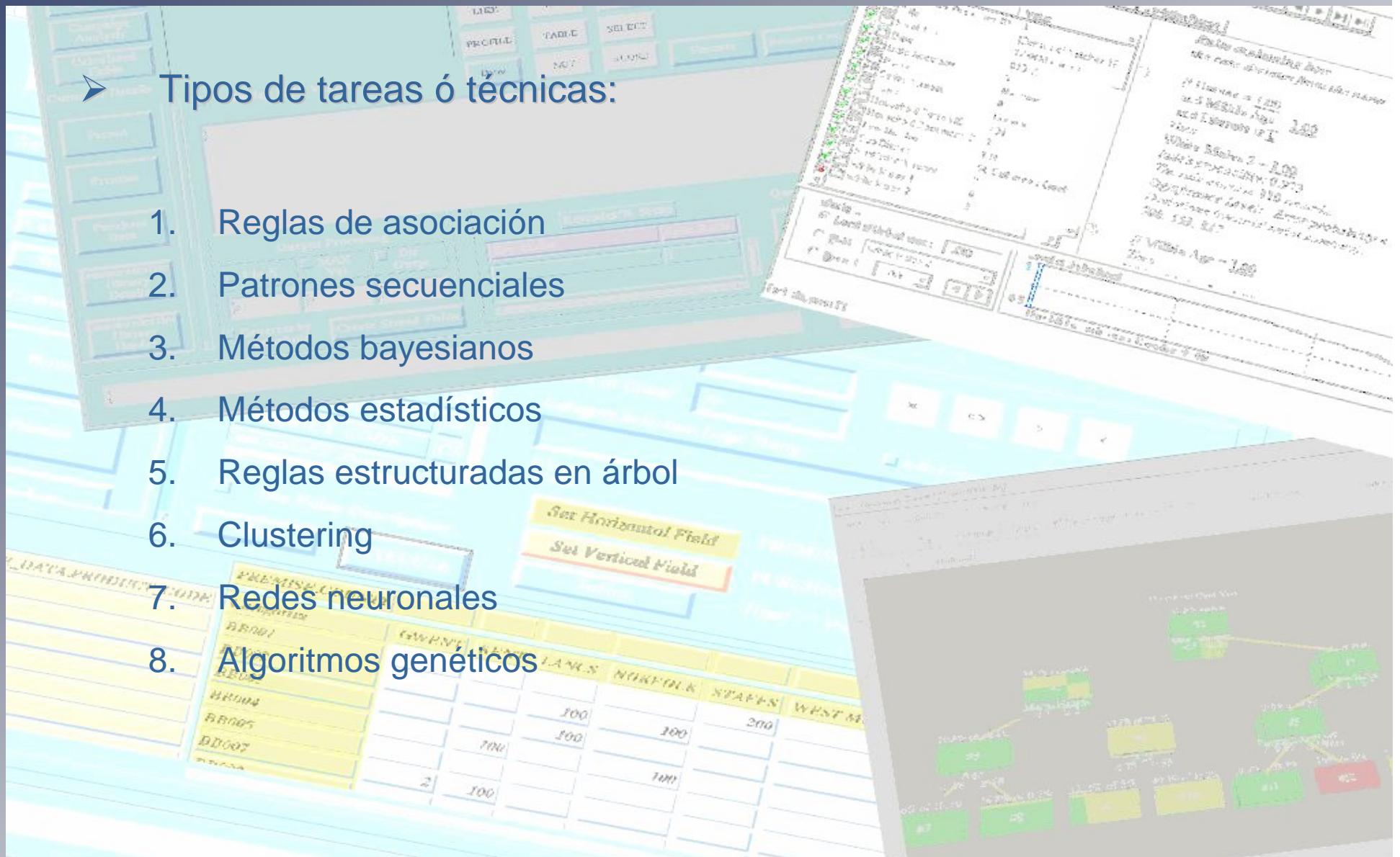
- Se trata de la construcción de un modelo basado en los datos recopilados
- El modelo es una descripción de los patrones y relaciones entre los datos, para entender mejor los datos ó explicar situaciones pasadas.
- Los pasos a dar son los siguientes:
 1. Determinar tarea de minería de datos más conveniente: clasificación,...
 2. Elegir técnica: árbol de decisión para implementar la clasificación,...
 3. Elegir algoritmo que implemente la técnica elegida: algoritmo BIRCH para generar árbol de decisión,...

Datamining Modelado del datamining

Celia Gutiérrez Cossío
2007

➤ Tipos de tareas ó técnicas:

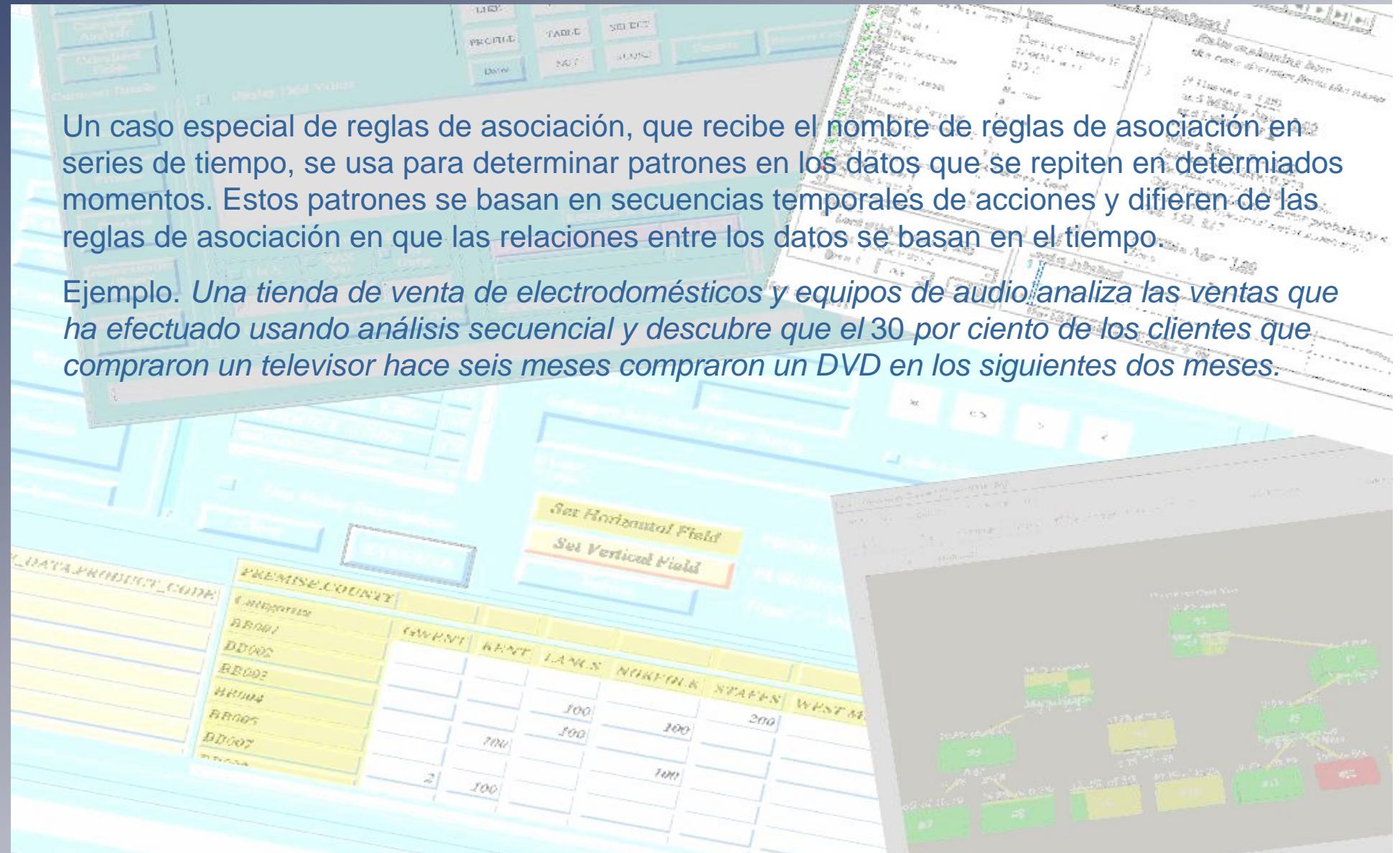
1. Reglas de asociación
2. Patrones secuenciales
3. Métodos bayesianos
4. Métodos estadísticos
5. Reglas estructuradas en árbol
6. Clustering
7. Redes neuronales
8. Algoritmos genéticos



➤ Concepto de regla de asociación:

Las reglas de asociación son una tarea descriptiva, que tiene como objetivo identificar relaciones no explícitas entre atributos categóricos. Pueden ser de muchas formas, aunque la formulación más común es del estilo "si el atributo X toma el valor *a* entonces el atributo Y toma el valor *b*". Las reglas de asociación no implican una relación causa-efecto, es decir, puede no existir una causa para que los datos estén asociados. Este tipo de tarea se utiliza frecuentemente en el análisis de la cesta de la compra, para identificar productos que son frecuentemente comprados juntos, información esta que puede usarse para ajustar los inventarios, para la organización física del almacén o en campañas publicitarias. Las reglas se evalúan usando dos parámetros: precisión (confianza) y soporte (cobertura)

Ejemplo. Una compañía de asistencia sanitaria desea analizar las peticiones de servicios médicos solicitados por sus asegurados. Cada petición contiene información sobre las pruebas médicas que fueron realizadas al paciente durante una visita. Toda esta información se almacena en una base de datos en la que cada petición es un registro cuyos atributos expresan si se realiza o no cada una de las posibles pruebas médicas que pueden ser realizadas a un paciente. Mediante reglas de asociación, un sistema encontraría aquellas pruebas médicas que frecuentemente se realizan juntas, por ejemplo que un 70 por ciento de las veces que se pide un análisis de orina también se solicita uno de sangre, y esto ocurre en dos de cada diez pacientes. La precisión de esta regla es del 70 por ciento y el soporte del 20 por ciento.



Un caso especial de reglas de asociación, que recibe el nombre de reglas de asociación en series de tiempo, se usa para determinar patrones en los datos que se repiten en determinados momentos. Estos patrones se basan en secuencias temporales de acciones y difieren de las reglas de asociación en que las relaciones entre los datos se basan en el tiempo.

Ejemplo. Una tienda de venta de electrodomésticos y equipos de audio analiza las ventas que ha efectuado usando análisis secuencial y descubre que el 30 por ciento de los clientes que compraron un televisor hace seis meses compraron un DVD en los siguientes dos meses.

➤ Soporte y confianza:

Función del datamining: encontrar todas las reglas posibles que superen un umbral de soporte y confianza especificados por el usuario.

1. **Soporte:** el soporte de un conjunto de items es el porcentaje de transacciones que contienen todos esos items. El soporte para la regla de la forma $LHS \Rightarrow RHS$, donde LHS y RHS son conjuntos de items, es el soporte del conjunto de items formado por $LHS \cup RHS$.
2. **Confianza:** la confianza de una regla $LHS \Rightarrow RHS$, es el porcentaje de transacciones de LHS que también contienen RHS . La confianza de una regla es el indicativo de la fuerza de la regla.

Ejemplo: dada la tabla siguiente de cesta de compra, encontrar cual es el soporte y la confianza de la regla $\{\text{pen}\} \Rightarrow \{\text{ink}\}$. (Se leería: “Si un bolígrafo es comprado en una transacción, la tinta también será comprada en dicha transacción”).

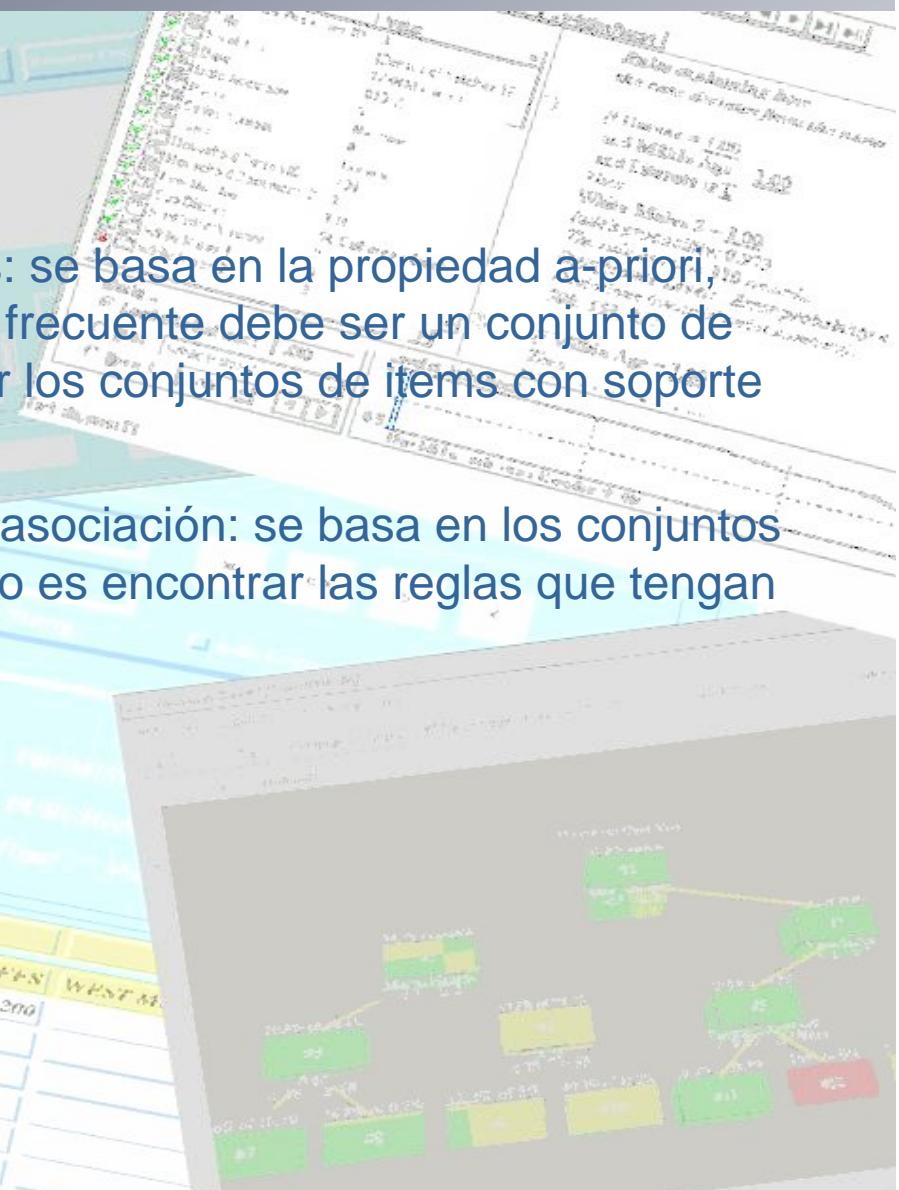
transid	custid	date	item	qty
111	201	5/1/99	pen	2
111	201	5/1/99	ink	1
111	201	5/1/99	milk	3
111	201	5/1/99	juice	6
112	105	6/3/99	pen	1
112	105	6/3/99	ink	1
112	105	6/3/99	milk	1
113	106	5/10/99	pen	1
113	106	5/10/99	milk	1
114	201	6/1/99	pen	2
114	201	6/1/99	ink	2
114	201	6/1/99	juice	4

- El soporte de la regla es el soporte del conjunto de datos {pen, ink} => 75%
- La confianza de la regla es el 75%.

➤ Algoritmo a-priori:

1. Cálculo del conjunto de items frecuentes: se basa en la propiedad a-priori, “Todo subconjunto de un conjunto items frecuente debe ser un conjunto de items frecuente”. El objetivo es encontrar los conjuntos de items con soporte mínimo.
2. Cálculo de la confianza de las reglas de asociación: se basa en los conjuntos calculados en el paso anterior. El objetivo es encontrar las reglas que tengan una mínima confianza y soporte.

ITEMSET	FREQUENCY	Set Horizontal Field						Set Vertical Field					
		1	2	3	4	5	6	7	8	9	10	11	12
BB001	100												
BB002	100												
BB003	100												
BB004	100												
BB005	100												
BB007	100												
BB008	100												



- **Algoritmo de extracción de los conjuntos de items**

```
foreach item,
    Check if it is a frequent itemset //appears in > minsup transactions
    k=1
repeat           // Iterative, level-wise identification of frequent itemsets
    foreach new frequent itemset Ik with k items                  // Level k + 1
        generate all itemsets Ik+1 with k + 1 items, Ik ⊂ Ik+1
    Scan all transactions once and check if the k + 1-itemsets are frequent
    k=k+1
until no new frequent itemsets are identified
```

Nota: los parámetros de entrada al algoritmo son los items y minsup.

- **Creación de reglas a partir de los conjuntos de items frecuentes**

Se trata de generar todas las reglas posibles de la forma LHS => RHS, sabiendo LHS, RHS, LHS U RHS fueron clasificados como conjuntos de items frecuentes en el paso previo. Por ello, se sugiere almacenar los valores de soporte en algún tipo de estructura de almacenamiento.

Dado el conjunto de items frecuente, x, cuyo soporte s_x fue calculado en el paso previo, la confianza de la regla, LHS => RHS, es s_x / s_{LHS} , donde x=LHS U RHS.

Una vez calculada la confianza de la regla, se puede examinar si supera minconf.

- **Ejercicio:**

Dada la tabla:

transid	custid	date	item	qty
111	201	5/1/99	pen	2
111	201	5/1/99	ink	1
111	201	5/1/91	milk	3
111	201	5/1/99	juice	6
112	105	6/3/99	pen	1
112	105	6/3/99	ink	1
112	105	6/3/99	milk	1
113	106	5/10/99	pen	1
113	106	5/10/99	milk	1
114	201	6/1/99	pen	2
114	201	6/1/99	ink	2
114	201	6/1/99	juice	4

1. Obtener conjuntos de items frecuentes para minsup=90 y reglas de asociación para minconf=90.
2. Obtener conjuntos de items frecuentes para minsup=10 y reglas de asociación para minconf=90.

➤ Reglas de asociación y jerarquías de categorías:

- Puede haber una jerarquía de categoría dentro del conjunto de items => una transacción contiene para cada uno de sus items todos los ancestros del item en la jerarquía.

Ejemplo: con la siguiente jerarquía



transid	custid	date	item	qty
111	201	5/1/99	stationery	3
111	201	5/1/99	beverage	9
112	105	6/3/99	stationery	2
112	105	6/3/99	beverage	1
113	106	5/1/99	stationery	1
113	106	5/1/99	beverage	1
114	201	5/15/99	stationery	4
114	201	5/15/99	beverage	4

- La jerarquía nos permite detectar relaciones entre items a niveles distintos de la jerarquía. El soporte de un conjunto de items puede aumentar solo si un item es reemplazado por uno de sus ancentros en la jerárquía => valor añadido a reglas de asociación.

Ejemplo: si el soporte del conjunto de items {ink, juice} es 50%, y reemplazamos juice por beverage, el soporte del itemset {ink, beverage} aumenta al 75%.

- Se puede usar el algoritmo para calcular conjuntos de items frecuentes sobre la base de datos aumentada.



Reglas de asociación negativas:

- El problema de descubrir una asociación negativa es más arduo que el de descubrir una asociación positiva => ausencia de combinaciones de elementos.
- “El 60 por ciento de los clientes que compran patatas fritas no compran agua mineral”. (Aquí, el 60 por ciento hace referencia a la confianza de la regla de asociación negativa.)
- Problema: encontrar sólo aquellas reglas negativas que sean *interesantes* => casos en los que dos conjuntos específicos de elementos aparecen muy raramente en la misma transacción y por tanto los porcentajes “negativos” son altos:
 1. Para un inventario total de 10.000 elementos, la probabilidad de que se compren juntos dos elementos dados es $1/10.000 * 1/10.000 = 10^{-8}$. Si el soporte real de que ambos elementos aparezcan juntos es cero, este no supone una desviación significativa respecto a lo esperado y por lo tanto, no resulta una asociación (negativa) de interés.
 2. Conclusión: dada la inmensa cantidad de datos, las posibilidades de combinaciones crecen exponencialmente, y por tanto, si aparecen poco, la desviación entre predicho y real es poco significativa; sin embargo, si se hace a nivel de jerarquías, se pueden encontrar desviaciones interesantes.

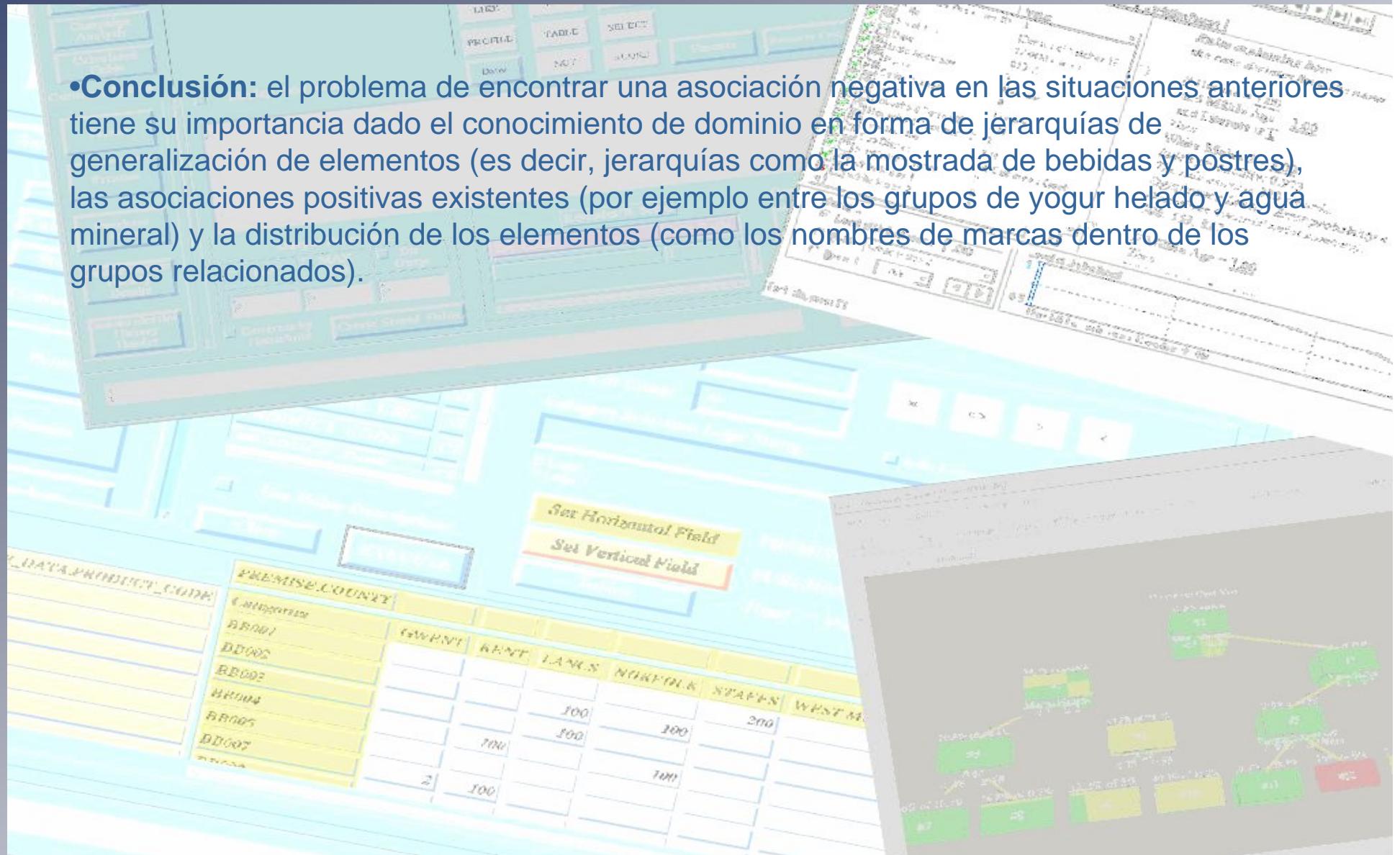
•Ejemplo: Supongamos que se ha identificado una fuerte asociación positiva entre refrescos y patatas fritas. Sería interesante si por ejemplo encontrásemos un gran soporte para el hecho de que cuando los clientes compran patatas fritas Days, estos compran predominantemente refrescos Topsy pero no Joke ni Wakeup. (Si compran Topsy junto a Days, ¿por qué no compran Wakeup o Joke, si tambien son refrescos?)



•Ejemplo: supongamos que la distribución entre las marcas de yogur helado Reduce y Healthy es 80-20 y entre las marcas de agua Plain y Clear es 60-40. Esto daría una probabilidad conjunta de que se comprase yogur helado Reduce junto con agua mineral Plain del 48 por ciento, entre las transacciones que contienen yogur helado y agua mineral. Sin embargo, si se observa que este soporte es tan solo del 20 por ciento, ello indicaría que existe una asociación negativa significativa entre el yogur Reduce y el agua mineral Plain; lo cual, podría ser interesante.



- Conclusión:** el problema de encontrar una asociación negativa en las situaciones anteriores tiene su importancia dado el conocimiento de dominio en forma de jerarquías de generalización de elementos (es decir, jerarquías como la mostrada de bebidas y postres), las asociaciones positivas existentes (por ejemplo entre los grupos de yogur helado y agua mineral) y la distribución de los elementos (como los nombres de marcas dentro de los grupos relacionados).

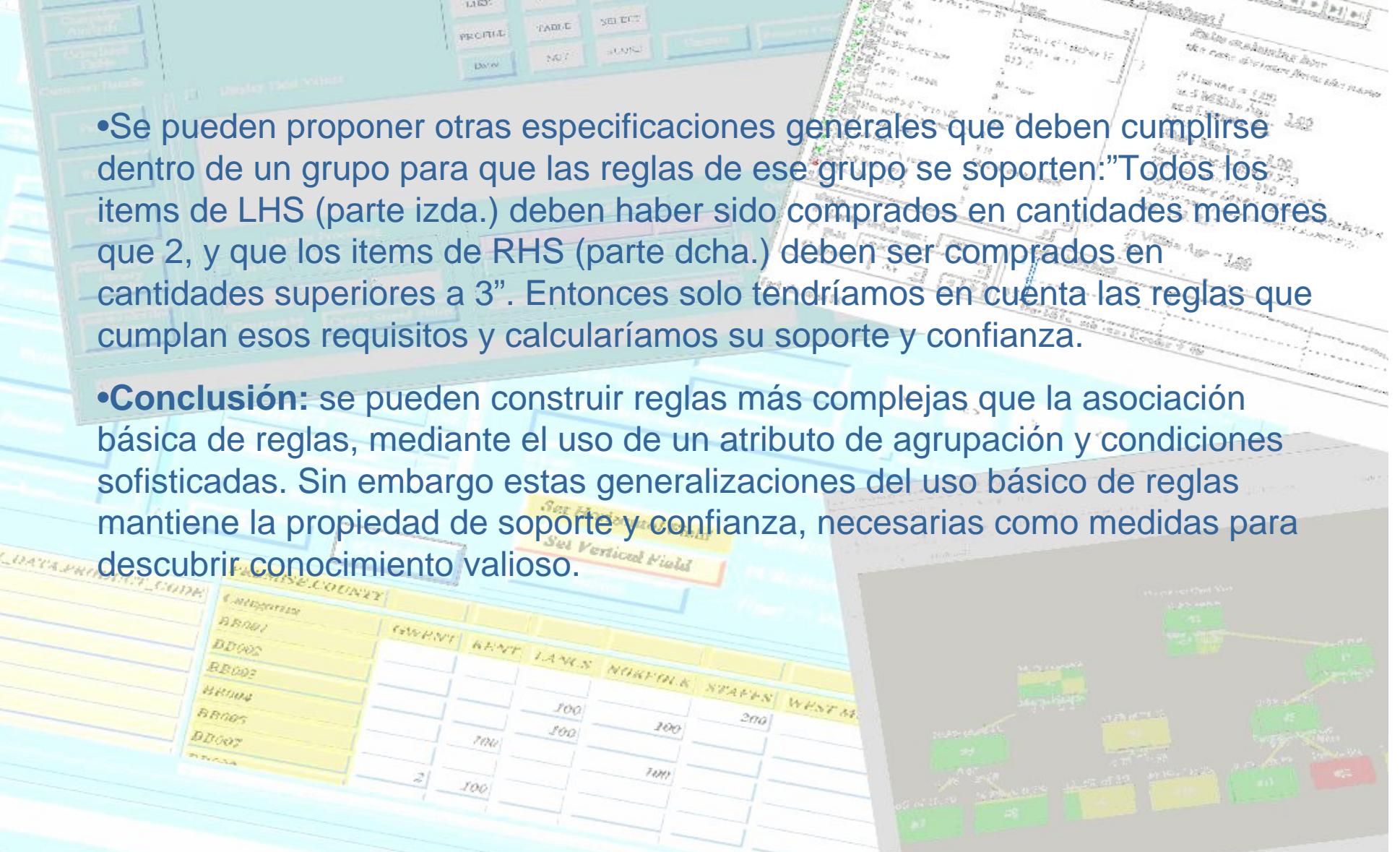


➤ Reglas de asociación generalizadas:

- Las reglas de asociación no solo se aplican a la cesta de la compra para descubrir asociaciones entre ítems: el concepto de estas reglas es más general.
- Para la tabla inferior, la regla $\{pen\} \Rightarrow \{milk\}$ tiene un soporte y confianza del 100%. (teniendo en cuenta que la tabla esta clasificada por cliente).

transid	custid	date	item	qty
112	105	6/3/99	pen	1
112	105	6/3/99	ink	1
112	105	6/3/99	milk	1
113	106	5/1/99	pen	1
113	106	5/1/99	milk	1
114	201	5/15/99	pen	2
114	201	5/15/99	ink	2
114	201	5/15/99	juice	4
111	201	5/1/99	pen	2
111	201	5/1/99	ink	1
111	201	5/1/99	milk	3
111	201	5/1/99	juice	6

- Se pueden extraer patrones entre ítems que descubran relaciones temporales: “En el día en que se compra un lápiz (pen), es probable que también se compre leche (milk)”.
- Si se usa el campo fecha como un atributo de agrupación, se está usando un problema más general llamado **análisis de cesta de compra en series de tiempo**. En dichos análisis el usuario especifica una colección de **calendarios**, que es cualquier grupo de fechas: “Todos los domingos del año 1999, todos los primeros días de mes,...”. Una regla se sostiene si se sostiene en todos los días de dicho calendario => se pueden calcular reglas de asociación sobre un conjunto de tuplas cuyo valor del campo fecha esté dentro del calendario especificado.
- Pueden existir reglas sin la suficiente confianza y soporte con respecto a la base de datos completa, pero si con respecto al subconjunto de tuplas que caen dentro del calendario => relaciones temporales interesantes.
- **Ejemplo:** con la tabla anterior y el calendario todos los días primeros de mes, la asociación {pen} => {ink} tiene soporte y confianza del 100%, mientras que con respecto a la base de datos completa solo tiene un 75%.

- 
- Se pueden proponer otras especificaciones generales que deben cumplirse dentro de un grupo para que las reglas de ese grupo se soporten: "Todos los items de LHS (parte izda.) deben haber sido comprados en cantidades menores que 2, y que los items de RHS (parte dcha.) deben ser comprados en cantidades superiores a 3". Entonces solo tendríamos en cuenta las reglas que cumplan esos requisitos y calcularíamos su soporte y confianza.
 - **Conclusión:** se pueden construir reglas más complejas que la asociación básica de reglas, mediante el uso de un atributo de agrupación y condiciones sofisticadas. Sin embargo estas generalizaciones del uso básico de reglas mantiene la propiedad de soporte y confianza, necesarias como medidas para descubrir conocimiento valioso.

➤ Uso de reglas de asociación para Predicción:

- Aunque el uso de las reglas de asociación en casos de predicción esta extendido, hay que realizar un conocimiento del dominio para justificarlas.

Ejemplo: $\{pen\} \Rightarrow \{ink\}$

La confianza asociada a esta regla es la probabilidad condicionada de que se compre tinta (ink) si se compra bolígrafo (pen), sobre toda la base de datos => medida descriptiva.

Para futuras promociones, se pueden aplicar descuentos solo en bolígrafos para aumentar su venta, y además aumentar la venta de tinta => indicador de futuras transacciones, ya que existe un vínculo causal entre la compra de bolígrafos y de tinta.

- Pero pueden existir reglas con alto porcentaje de soporte y confianza, para las que no existe ningún vínculo causal.

Ejemplo: supongamos que la compra de bolígrafos va unida a la de lápices (por una tendencia de los compradores a adquirir a la vez instrumentos de escribir). La regla $\{pencil\} \Rightarrow \{ink\}$ tiene la misma confianza y soporte que $\{pen\} \Rightarrow \{ink\}$. Pero no existe ningún vínculo causal entre lápices y tinta. Si se rebajan los lápices no se tiene por qué comprarse más tinta.

- **Conclusión:** aunque las reglas de asociación no indican relaciones causales entre LHS y RHS, proporcionan un punto de comienzo adecuado para identificar dichas relaciones.

➤ Concepto de patrones secuenciales:

Se basa en el concepto de secuencia de conjuntos de elementos. Asumimos que las transacciones, como las de la cesta de la compra, se ordenan por tiempo de compra. Este ordenamiento da lugar a una secuencia de conjuntos de elementos.

Ejemplo: {leche, pan, zumo}, {pan, huevos}, {galletas, leche, café} pueden constituir una secuencia de conjuntos de elementos basada en tres visitas del mismo cliente al establecimiento.

El **soporte** para una secuencia S de conjuntos de elementos lo constituye el porcentaje de veces que S es subsecuencia del conjunto de secuencias dado U.

El **problema** a la hora de identificar los patrones secuenciales, por lo tanto, es el de encontrar todas las subsecuencias a partir de los conjuntos de secuencias dados, que tengan un mínimo de soporte para el usuario.

Conclusión: la secuencia S₁, S₂, S₃,... constituye un previsor de la siguiente situación: es probable que un comprador que adquiere un conjunto de elementos S₁, también compre un conjunto de elementos S₂ y luego S₃, y así sucesivamente. Este resultado se basa en la frecuencia (soporte) de esta secuencia en el pasado.

➤ Ejemplo de secuencias válidas :

ID-Cliente	ID-Tiempo	Items
1	15/03/2003	{23,56}
1	17/03/2003	{42,13}
1	18/03/2003	{45,33}
2	12/03/2003	{12,13}
2	18/03/2003	{23,34,5,8}

ID-Cliente	Secuencia
1	<{23 56}{42 13}{45 33}>
2	<{12 13}{23 34 5 8}>

Secuencia	Subsecuencia	Válido
<{23,56}{42,13}{45,33}>	<{23}{42,13}>	✓
<{23,56}{42,13}{45,33}>	<{42}{45}>	✓
<{23,56}{42,13}{45,33}>	<{45}{42}>	✗
<{23,56}{42,13}{45,33}>	<{42,13}>	✓
<{42,33}>	<{42}{33}>	✗



Implementación de patrones secuenciales:

Como las reglas de asociación, los patrones secuenciales son sentencias sobre grupos de tuplas en la base de datos en curso. Computacionalmente, los algoritmos que encuentran patrones secuenciales frecuentes se parecen a los algoritmos que encuentran conjuntos de ítems frecuentes: Se identifican iterativamente secuencias cada vez más largas con un soporte mínimo, de una manera similar a la identificación iterativa de conjuntos de ítems frecuentes.

El algoritmo más conocido para extraer patrones secuenciales sobre la información residente en una base de datos, que tenga un mínimo de confianza, es el denominado A-priori All:

1. Ordenación: el IDCliente como clave primaria, y el ID-Tiempo como clave secundaria.
2. Se construye una secuencia de conjuntos de ítems por cada cliente.
3. Selección de conjuntos de ítems: con una mínima cobertura, respecto a cada cliente. Transformación y renombramiento: se le asigna a cada conjunto de ítems frecuentes un identificador.
4. Cada secuencia se transforma de manera que sólo contenga sus ítems frecuentes. A continuación se renombra cada conjunto por su identificador.
5. Construcción de secuencias frecuentes: con el criterio de cobertura.
6. Selección de secuencias máximas: filtra el conjunto de secuencias frecuentes de manera que no haya subsecuencias partiendo desde las secuencias de mayor tamaño.

Modelado del datamining: patrones secuenciales

Celia Gutiérrez Cossío
2007

Ejemplo de ejecución de A-priori All

ID-Cliente	ID-Tiempo	Items
1	15/03/2003	{30}
1	17/03/2003	{90}
2	8/03/2003	{10,20}
2	12/03/2003	{30}
2	18/03/2003	{40,60,70}
3	18/03/2003	{30,50,70}
4	13/03/2003	{30}
4	15/03/2003	{40,70}
4	17/03/2003	{90}
5	14/03/2003	{90}

ID-Cliente	Secuencia
1	<{30}{90}>
2	<{10 20}{30}{40 60 70}>
3	<{30 50 70}>
4	<{30}{40 70}{90}>
5	<{90}>

Conjuntos de items	Identificador
{30}	1
{40}	2
{70}	3
{40 70}	4
{90}	5

Cobertura = 2 clientes (40%)

Modelado del datamining: patrones secuenciales

Celia Gutiérrez Cossío
2007

ID-Cliente	Secuencia	Secuencia transformada	Sec, transformada y renombrada
1	<{30}{90}>	<{{30}} {{90}}>	<{1}{5}>
2	<{10 20}{30}{40 60 70}>	<{{30}} {{40}}, {{70}}, {{40, 70}}>	<{1}{2,3,4}>
3	<{30 50 70}>	<{{30}}, {{70}}>	<{1,3}>
4	<{30}{40 70}{90}>	<{{30}} {{40}, {70}}, {{40 70}} {{90}}>	<{1}{2,3,4}{5}>
5	<{90}>	<{{90}}>	<{5}>

Tamaño 1		Tamaño 2		Tamaño 3		Tamaño 4	
Secuencia	Soporte	Secuencia	Soporte	Secuencia	Soporte	Secuencia	Soporte
<1>	4	<1 2>	2	<1 2 3>	2	<1 2 3 4>	2
<2>	2	<1 3>	3	<1 2 4>	2		
<3>	3	<1 4>	2	<1 3 4>	2		
<4>	2	<1 5>	2	<2 3 4>	2		
<5>	3	<2 3>	2				
		<2 4>	2				
		<3 4>	2				

Secuencia	Secuencia Original	Cobertura
<1 2 3 4>	<{30} {40} {70} {40 70}>	2
<1 5>	<{30} {90}>	2

Ejercicio:

Encontrar todos los patrones secuenciales con $\text{minsup}=80\%$.

<i>transid</i>	<i>custid</i>	<i>date</i>	<i>item</i>	<i>qty</i>
111	201	5/1/99	pen	2
111	201	5/1/99	ink	1
111	201	5/1/99	milk	3
111	201	5/1/99	juice	6
112	105	6/3/99	pen	1
112	105	6/3/99	ink	1
112	105	6/3/99	milk	1
113	106	5/10/99	pen	1
113	106	5/10/99	milk	1
114	201	6/1/99	pen	2
114	201	6/1/99	ink	2
114	201	6/1/99	juice	4



Uno de los problemas más frecuentes a los que se enfrentan las técnicas de minería de datos es cómo trabajar con incertidumbre => métodos bayesianos.

➤ **Modelo descriptivo y predictivo:**

1. Descriptivo: para descubrir relaciones de independencia y/o relevancia entre las variables que constituyen las redes bayesianas. Establecen relaciones mucho más ricas que las reglas de asociación o patrones secuenciales.
2. Predictivo: para clasificar.

➤ **Ventajas de los métodos bayesianos:**

1. Método práctico para realizar inferencias a partir de los datos, induciendo modelos probabilísticos que después serán usados para razonar (formular hipótesis) sobre los nuevos valores observados.
2. Además permiten calcular la probabilidad asociada a cada una de las hipótesis posibles.

Ejemplo:

Por ejemplo, se trata de recomendar si se debe ó no invertir en bolsa para dos productos P1 y P2, a partir de unos datos de entrada. Se puede obtener una salida afirmativa para los dos productos, pero un método que maneja probabilidades, como el bayesiano, puede obtener los siguientes resultados:

- Para P1: Si con probabilidad 0,9; No con probabilidad 0,1.
- Para P2: Si con probabilidad 0,52;No con probabilidad 0,48.

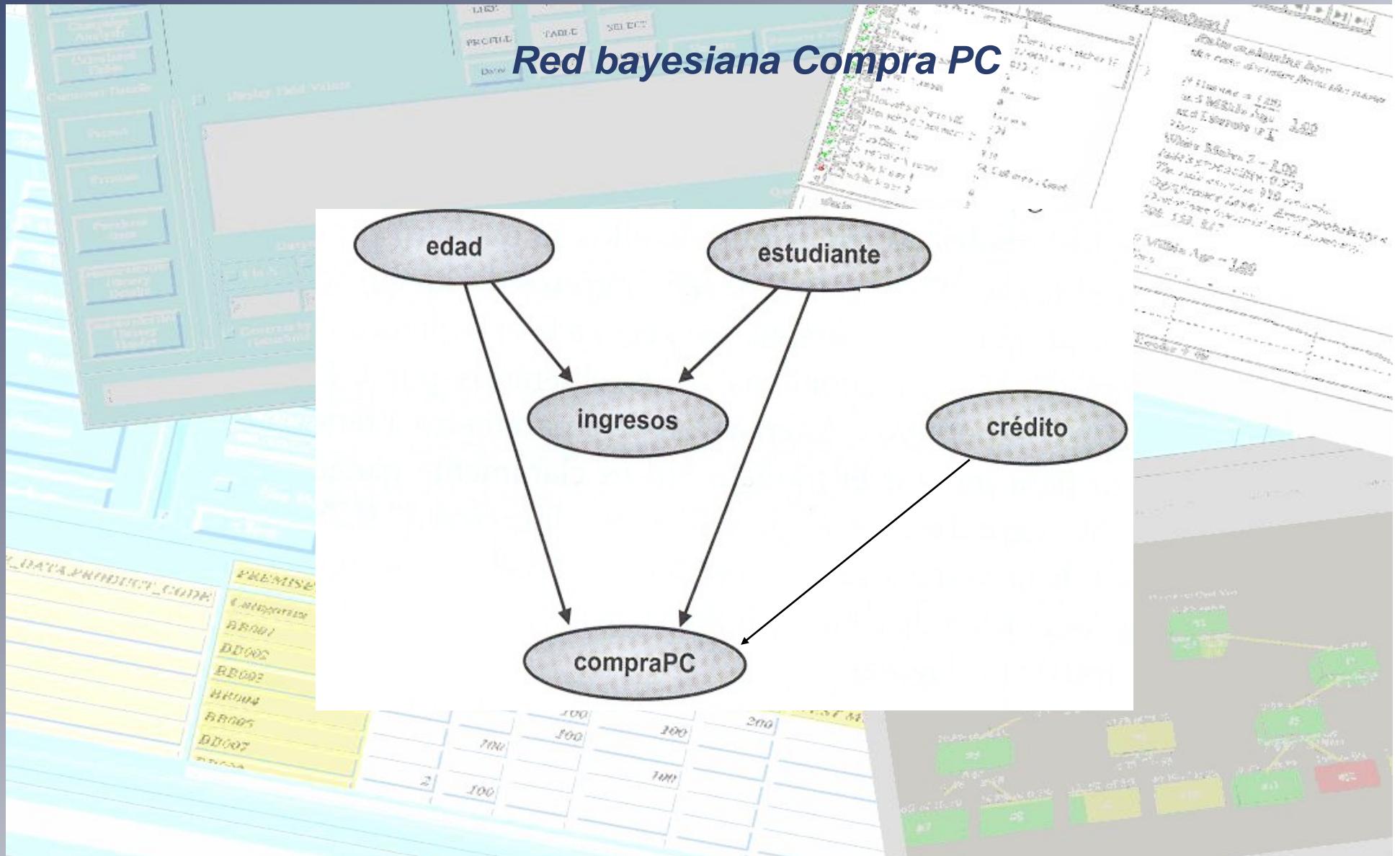
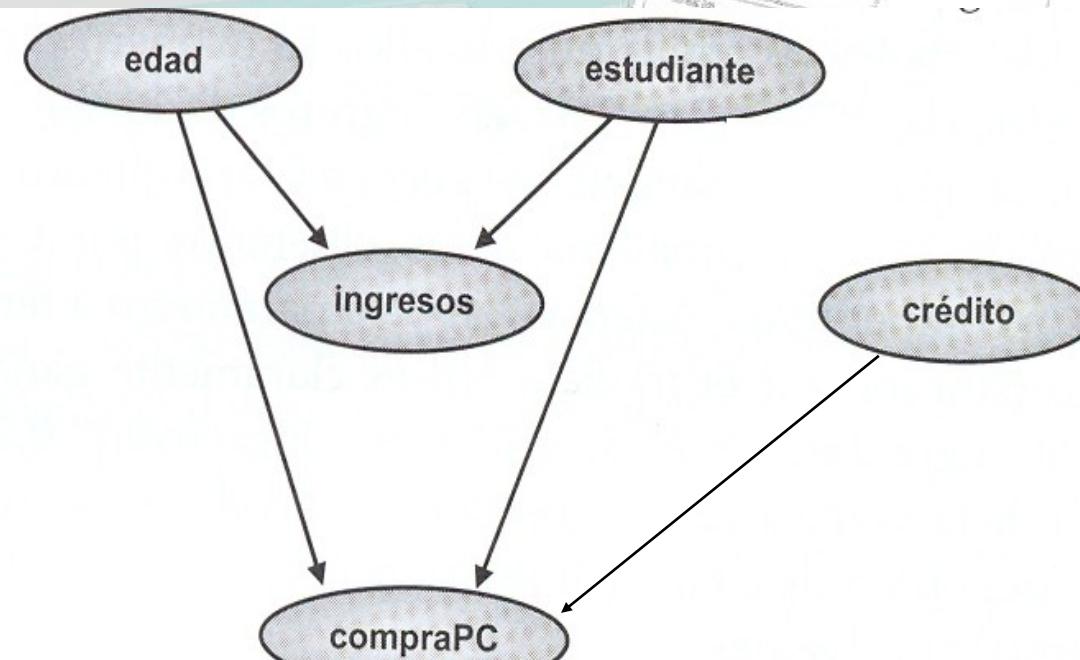
➤ Desventajas

Frente a estas ventajas, y al hecho indiscutible de constituir uno de los enfoques teóricamente más sólidos y elegantes, se encuentran con la desventaja del alto coste computacional. No obstante, la aparición de las Redes Bayesianas permiten reducir el coste computacional sin perder expresividad en el modelo probabilístico. Estas redes representan el conocimiento mediante un grafo dirigido acíclico, y muestra las relaciones de dependencia e independencia entre las variables. Además expresan de forma numérica la fuerza de las relaciones entre variables, en forma de probabilidad.

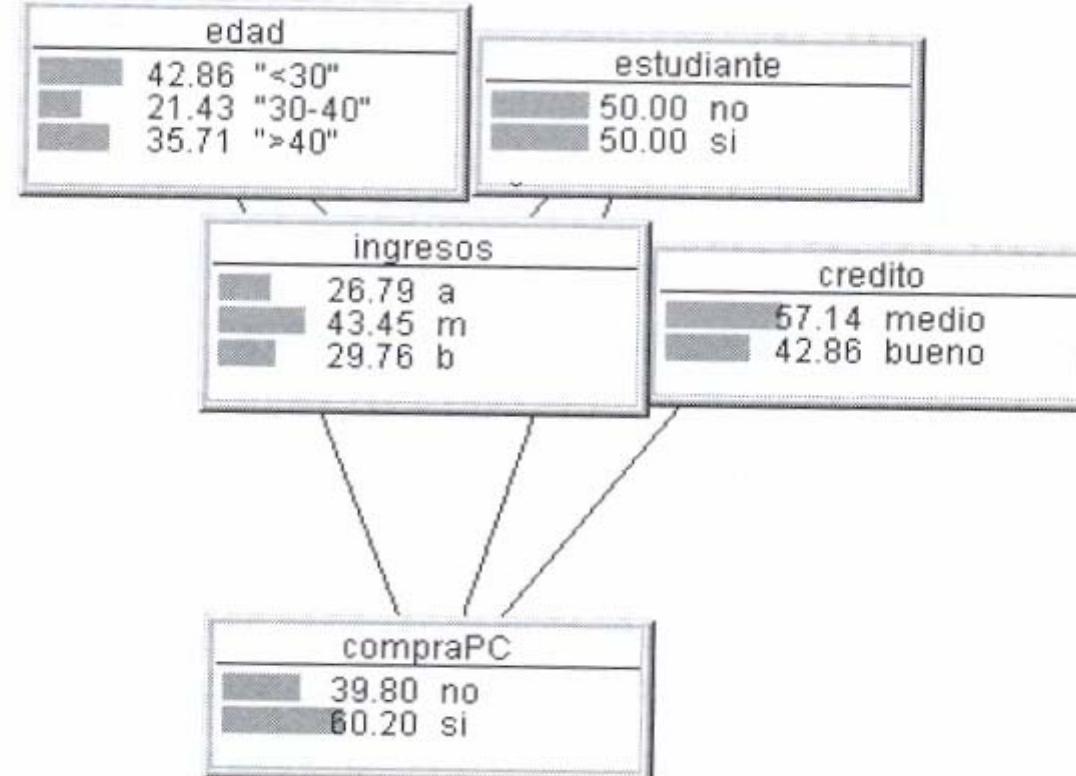
Modelado del datamining: métodos bayesianos

Celia Gutiérrez Cossío
2007

Red bayesiana Compra PC



Red bayesiana Compra PC (con slots de probabilidades)



Uno de los métodos más utilizados es el Naive Bayes, basado en la regla de Bayes.

Funciona bien con bases de datos reales, sobretodo cuando se combina con otros procedimientos de selección de atributos que sirven para eliminar la redundancia.

➤ Regla de Bayes:

Si establecemos una hipótesis H sustentada para una evidencia E , entonces:

$$p(H | E) = \frac{p(E | H).p(H)}{p(E)}$$

donde $p(A)$ representa la probabilidad del suceso A , usando la notación $p(A / B)$ para denotar la probabilidad del suceso A condicionada al suceso B .

Ejemplo:

Una compañía de seguros dispone de los siguientes datos sobre sus clientes, clasificados en buenos y malos clientes (Tabla siguiente):

Datamining

Modelado del datamining: métodos bayesianos

Celia Gutiérrez Cossío
2007

#Instancia	edad	hijos	practica_deporte	salario	buen_cliente
1	joven	sí	no	alto	sí
2	joven	no	no	medio	no
3	joven	sí	sí	medio	no
4	joven	sí	no	bajo	sí
5	mayor	sí	no	bajo	sí
6	mayor	no	sí	medio	sí
7	joven	no	sí	medio	sí
8	joven	sí	sí	alto	sí
9	mayor	sí	no	medio	sí
10	mayor	no	no	bajo	no

Y quiere meter un nuevo cliente, del cual tiene que calcular si es buen cliente:

edad	hijos	practica_deporte	salario	buen_cliente
mayor	no	no	medio	?

La hipótesis H es que *buen-cliente* sea *sí* (o, alternativamente, *no*). La evidencia E es una combinación de los valores de los atributos *edad*, *hijos*, *practica_deporte* y *salario* del dato nuevo, por lo que su probabilidad se obtiene multiplicando las probabilidades de estos valores. Es decir:

$$p(\text{Sí} | E) = [p(\text{edad}_E | \text{sí}) \cdot p(\text{hijos}_E | \text{sí}) \cdot p(\text{practica_deporte}_E | \text{sí}) \cdot p(\text{salario}_E | \text{sí})] \cdot p(\text{sí}) / p(E)$$

El término $p(\text{edad}_E | \text{sí})$ se calcula dividiendo el número de instancias de la Tabla que tienen el valor *mayor* en el atributo *edad* (de los que el *buen-cliente* es *sí*) dividido por el número de instancias cuyo valor del atributo *buen-cliente* es *sí*, es decir, $p(\text{edad}_E | \text{sí}) = p(\text{mayor}_E | \text{sí}) = 3/7$. De igual forma obtenemos el resto de probabilidades condicionadas en el numerador de la ecuación anterior. El término $p(\text{sí})$ se calcula como el número de instancias de la Tabla cuyo valor del atributo *buen-cliente* es *sí* dividido por el número total de instancias, es decir $p(\text{Sí}) = 7/10$. Por último, el denominador $p(E)$ desaparece normalizando. Sustituyendo todos estos valores se obtiene que la probabilidad de que se asigne el valor *sí* al atributo *buen-cliente* del dato E es

$$p(\text{sí} | E) = 3/7 \cdot 2/7 \cdot 4/7 \cdot 3/7 \cdot 7/10 = 0,0210$$

Procediendo de igual forma para la clase *no* resulta $p(\text{no}|E) = 0,0296$, por lo que se asignará el valor *no* al atributo *buen-cliente* del dato E .



Reglas de clasificación:

La clasificación es quizá la tarea más utilizada en datamining. Es una tarea predictiva.

En ella, cada instancia (o registro de la base de datos) pertenece a una clase, la cual se indica mediante el valor de un atributo que llamamos la clase de la instancia. Este atributo puede tomar diferentes valores discretos, cada uno de los cuales corresponde a una clase. El resto de los atributos de la instancia (los relevantes a la clase) se utilizan para predecir la clase. El objetivo es predecir la clase de nuevas instancias de las que se desconoce la clase. Más concretamente, el objetivo del algoritmo es maximizar la razón de precisión de la clasificación de las nuevas instancias, la cual se calcula como el cociente entre las predicciones correctas y el número total de predicciones (correctas e incorrectas).

Ejemplo. Consideremos un oftalmólogo que desea disponer de un sistema que le sirva para determinar la conveniencia o no de recomendar la cirugía ocular a sus pacientes.

Para ello dispone de una base de datos de sus antiguos pacientes clasificados en operados satisfactoriamente o no en función del tipo de problema que padecían (miopía y su grado, o astigmatismo) y de su edad. El modelo encontrado se utiliza para clasificar nuevos pacientes, es decir, para decidir si es conveniente operarlos o no.

Las reglas de clasificación siguen la **forma**:

$(\text{var}_1 \text{ en rango}_1) \text{ y } \dots (\text{var}_n \text{ en rango}_n) \Rightarrow \text{Objeto } O \text{ pertenece a la clase } C_1$

Las variables $\text{var}_1 \dots \text{var}_n$ constituyen los atributos del objeto O y compondrían las columnas de una relación con una tupla por objeto, y cada tupla pertenece a una clase. Sería posible establecer un conjunto de consultas SQL que convirtieran los datos obtenidos en instancias de las clases, una vez que estas se hayan definido. El problema del datamining es descubrir las clases así como las condiciones que definen dichas clases.

Se usa:

- Conjunto de entrenamiento (*Training set*) => para construir el modelo
- Conjunto de prueba (*Test set*) => para validar

La **diferencia** con las reglas de asociación es que las variables de reglas de clasificación toman valores de un dominio discreto ó continuo a diferencia de los conjuntos de elementos, que se constituyen a partir de un conjunto de elementos predefinido en las reglas de asociación. Además una regla de asociación corresponde a un conjunto de transacciones (registros de entrada), pero una regla de clasificación nos dice cómo situar cada registro en una clase.

➤ Reglas de regresión:

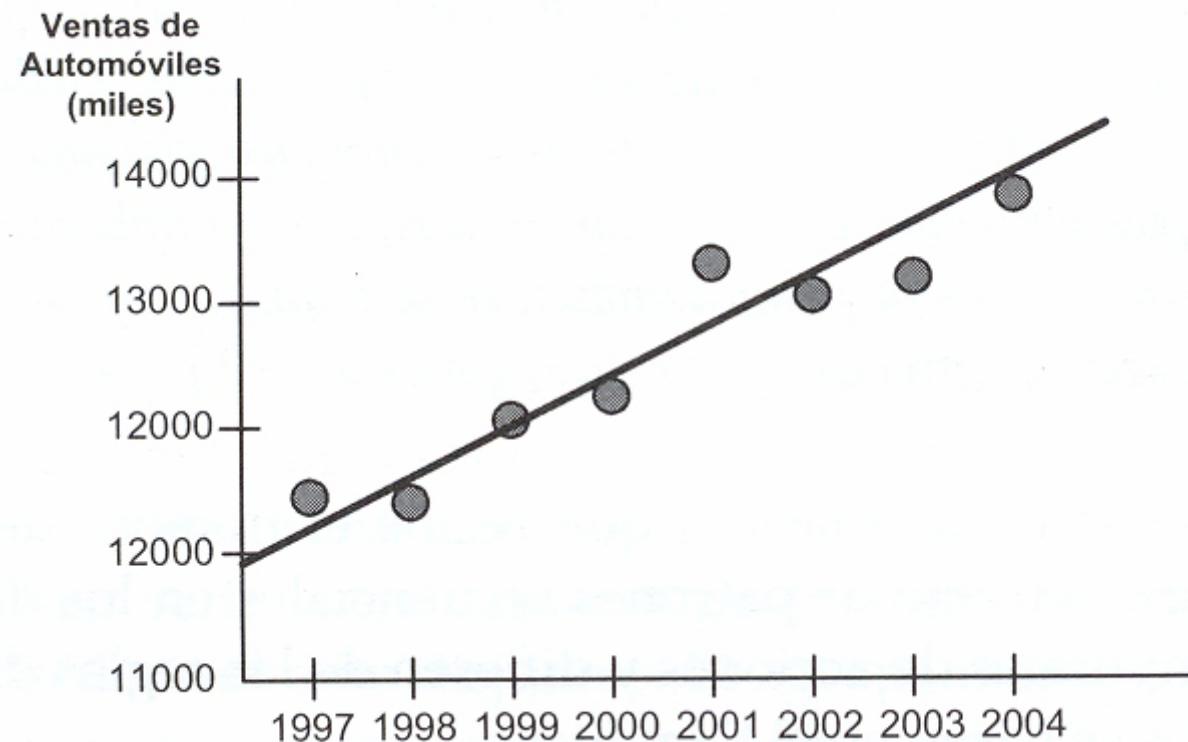
La regresión es también una tarea predictiva que consiste en aprender una función real que asigna a cada instancia un valor real. Ésta es la principal diferencia respecto a la clasificación; el valor a predecir es numérico. El objetivo en este caso es minimizar el error (generalmente el error cuadrático medio) entre el valor predicho y el valor real. El análisis de regresión es una herramienta muy común en el análisis de datos en muchos dominios de investigación: el descubrimiento de la función para predecir el valor de la variable objetivo es equivalente a una operación de datamining.

Ejemplo. Un empresario quiere conocer cuál es el costo de un nuevo contrato basándose en los datos correspondientes a contratos anteriores. Para ello usa una fórmula de regresión lineal, ajustando con los datos pasados la función lineal y usándola para predecir el costo en el futuro.

En general, la fórmula para una regresión lineal es $y=C_0 + C_1 X_1 + \dots + C_n X_n$ donde X_i son los atributos predictores e y la salida (la variable dependiente). Si los atributos son modificados en la función de regresión por alguna otra función (cuadrados, inversa, logarítmicos, combinaciones de variables...), es decir $y=C_0 + f_1(X_1) + \dots + f_n(X_n)$ la regresión se dice no lineal. Se pueden incorporar variantes locales o transformaciones en las variables predictoras y en la salida, permitiendo flexibilizar este tipo de técnicas.

En el siguiente gráfico se muestra un modelo de regresión lineal:

Ejemplo de visualización de regresión lineal



➤ Soporte y confianza para las reglas de clasificación y de regresión.

Al igual que en las reglas de asociación, tambien se pueden definir estos dos parámetros para reglas de clasificación y de regresión.

La formulación general de las reglas es la siguiente:

$$P_1(X_1) \wedge \dots \wedge P_n(X_n) \Rightarrow Y = c$$

donde $P_i(X_i)$ son predicados que involucran la variable.

Ejemplo:

$\text{Age} \geq 16 \wedge \text{age} \leq 25 \wedge \text{cartype} \in \{\text{sports}, \text{truck}\} \Rightarrow \text{highrisk} = \text{true}$

Soporte de la regla $C_1 \Rightarrow C_2$ es el soporte de la condición $C_1 \wedge C_2$, es decir, el porcentaje de todas las tuplas que satisfacen las dos condiciones.

Confianza de una regla $C_1 \Rightarrow C_2$, es el porcentaje de tuplas que satisfacen la condición C_2 entre todas las que satisfacen C_1 .

➤ Aplicaciones de las reglas de clasificación

- Clasificación de resultados de experimentos científicos, donde el tipo de objeto a reconocer depende de las medidas tomadas.
- Predicciones de mailings, donde la respuesta dada por un cliente dado a una promoción es función de su nivel de rentas y edad.
- Predicción de riesgo en una compañía aseguradora, donde el cliente puede ser clasificado como arriesgado, dependiendo de su edad, profesión y tipo de coche.

➤ Aplicaciones de las reglas de regresión

- Pronóstico financiero, donde el precio futuro del café es función de las lluvias caídas en Colombia hace un mes.
- Pronóstico médico, donde la probabilidad de que un tumor sea canceroso es función de atributos medidos en el tumor.

➤ Presentación de una herramienta que realiza modelos de regresión.

Cube

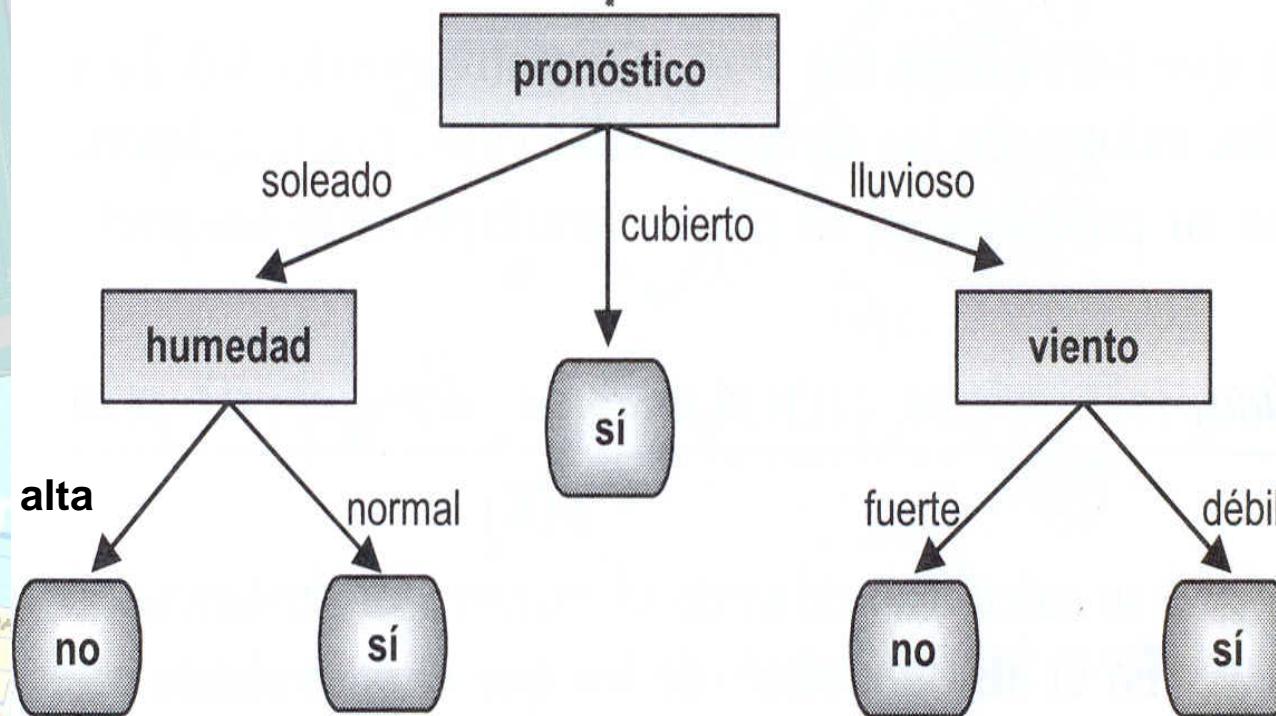
Los árboles de decisión son una serie de decisiones o condiciones organizadas en forma jerárquica, a modo de árbol. Son muy útiles para encontrar estructuras en espacios de alta dimensionalidad y en problemas que mezclen datos categóricos y numéricos. Esta técnica se usa en tareas de clasificación, agrupamiento y regresión. Los árboles de decisión usados para predecir variables categóricas reciben el nombre de árboles de clasificación; cuando los árboles de decisión se usan para predecir variables continuas se llaman árboles de regresión.

➤ Ejemplo:

Se trata de un conjunto de datos ficticios que muestra las condiciones climatológicas (pronóstico, humedad y viento) adecuadas para jugar un cierto deporte (por ejemplo, tenis en Wimbledon). Los datos de los que disponemos son los siguientes:

#instancia	pronóstico	humedad	viento	jugar
1	soleado	alta	débil	No
2	cubierto	alta	débil	Sí
3	lluvioso	alta	débil	Sí
4	lluvioso	normal	fuerte	No
5	soleado	normal	débil	Sí
...

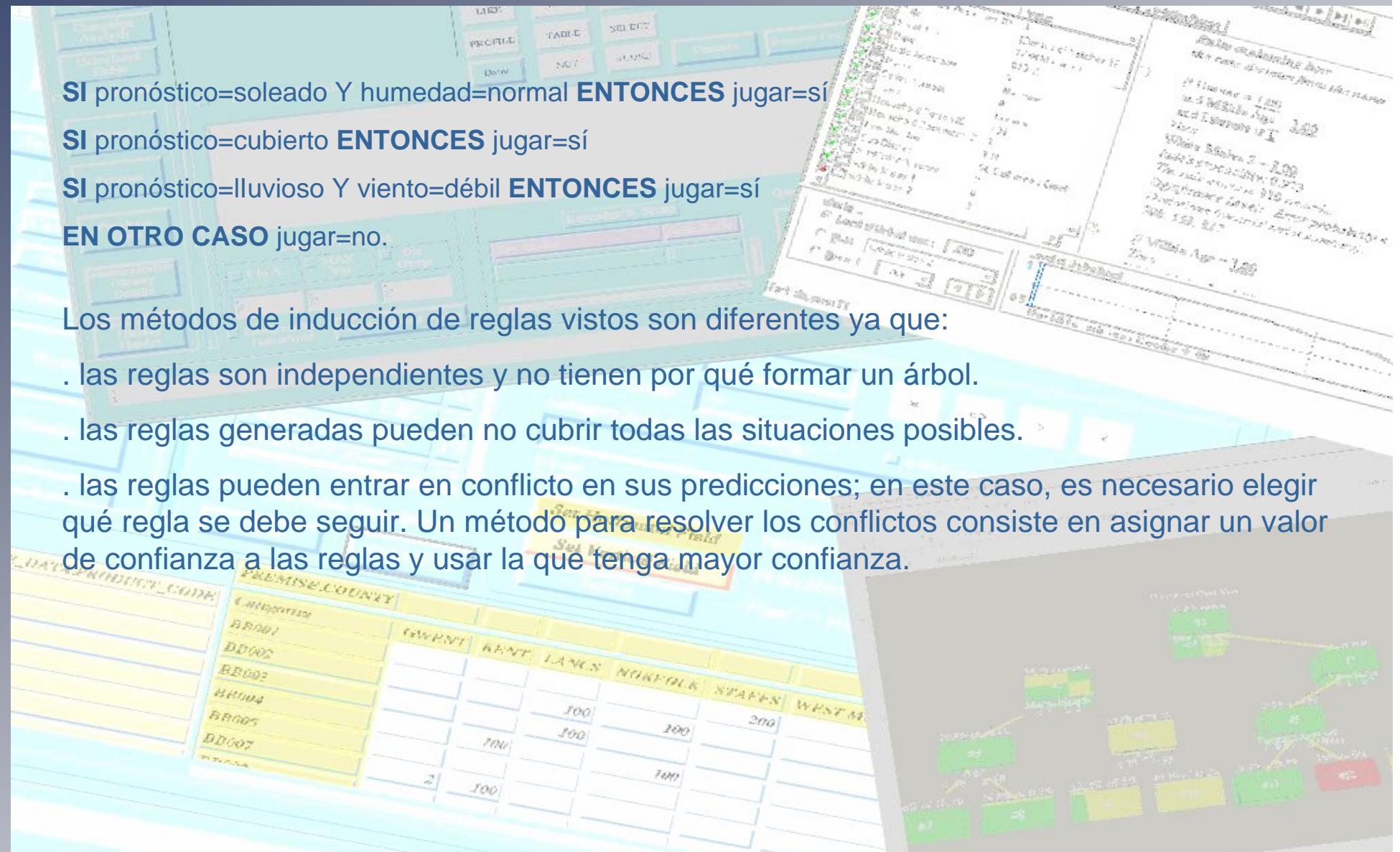
Usando un algoritmo de aprendizaje de árboles de decisión podríamos obtener el árbol que se muestra:



Los árboles de decisión siguen una aproximación "divide y vencerás" para partir el espacio del problema en subconjuntos. Encima del nodo raíz del árbol tenemos el problema a resolver. En nuestro ejemplo, se trata de decidir si *jugar* o no. Los nodos internos (nodos de decisión) corresponden a particiones sobre atributos particulares, como por ejemplo *pronóstico*, y los arcos que emanan de un nodo corresponden a los posibles valores del atributo considerado en ese nodo (por ejemplo, *soleado*, *cubierto* o *lluvioso*). Cada arco conduce a otro nodo de decisión o a un nodo hoja. Los nodos hoja representan la predicción (o clase) del problema para todas aquellas instancias que alcanzan esa hoja. Para clasificar una instancia desconocida, se recorre el árbol de arriba hacia abajo de acuerdo a los valores de los atributos probados en cada nodo y, cuando se llega a una hoja, la instancia se clasifica con la clase indicada por esa hoja.

➤ Reglas del árbol

Los árboles de decisión pueden considerarse una forma de aprendizaje de reglas, ya que cada rama del árbol puede interpretarse como una regla, donde los nodos internos en el camino desde la raíz a las hojas definen los términos de la conjunción que constituye el antecedente de la regla, y la clase asignada en la hoja es el consecuente. En la siguiente figura se muestra el conjunto de reglas que corresponde al árbol, en la que se han agrupado en una regla por defecto ("EN OTRO CASO") todas las ramas del árbol cuya hoja asigna la clase *no*.



➤ Construcción del árbol de decisión

En el nodo raíz, se examina la base de datos y se computa el criterio de división que se considera mejor localmente. Entonces se divide la base de datos en dos partes, una partición para el hijo de la izquierda y otra para el árbol de la derecha.

El criterio de división en un nodo se consigue mediante una aplicación de un método de selección de división, que es un algoritmo que toma como entrada una relación (ó parte) y obtiene el criterio de división localmente mejor.

Ejemplo: examina los atributos cartype y age, selecciona uno de ellos como atributo de división y entonces selecciona los predicados de división.

age	cartype	highrisk
23	Sedan	false
30	Sports	false
36	Sedan	false
25	Truck	true
30	Sedan	false
23	Truck	true
30	Truck	false
25	Sports	true
18	Sedan	false



Algoritmo de construcción de árbol de decisión

Input: node n , partition D , split selection method S

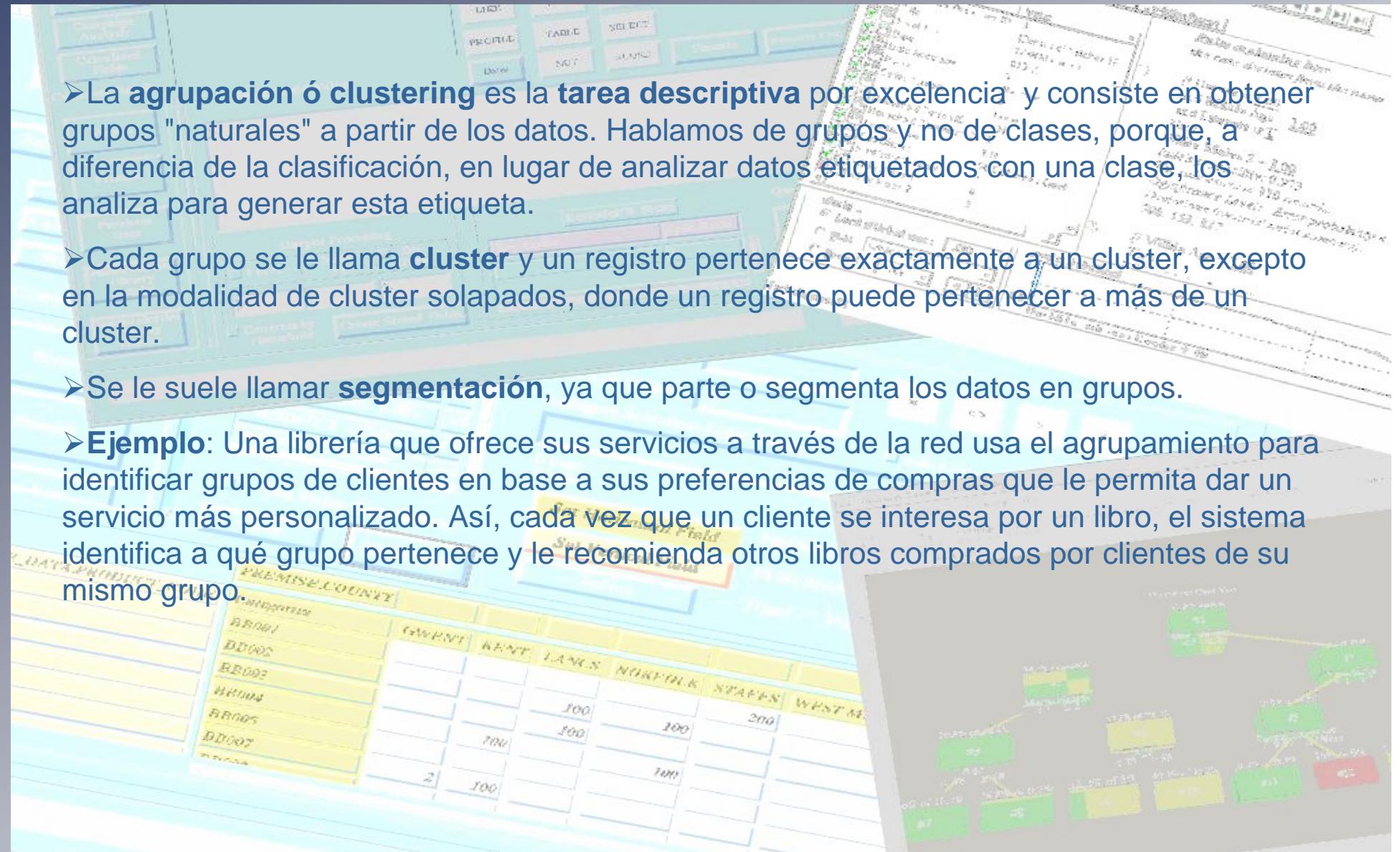
Output: decision tree for D rooted at node n

Top-Down Decision Tree Induction Schema:

BuildTree(Node n , data partition D , split selection method S)

- (1) Apply S to D to find the splitting criterion
- (2) **if** (a good splitting criterion is found)
- (3) Create two children nodes n_1 and n_2 of n
- (4) Partition D into D_1 and D_2
- (5) BuildTree(n_1, D_1, S)
- (6) BuildTree(n_2, D_2, S)
- (7) **endif**



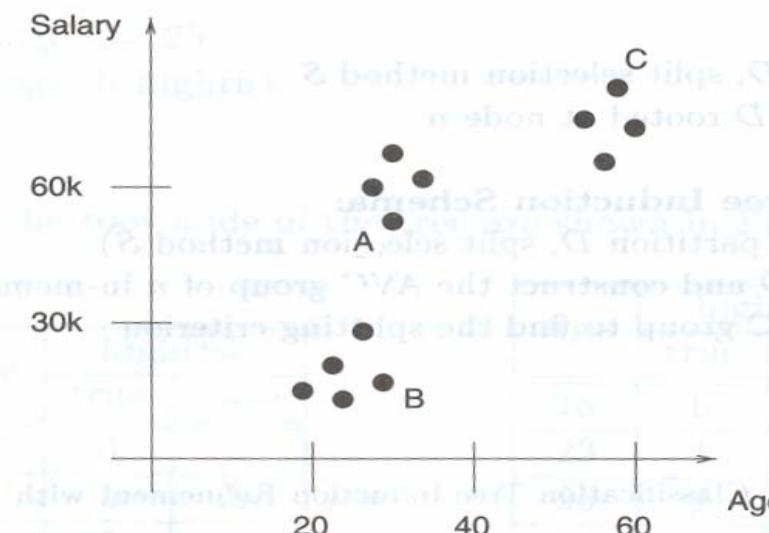


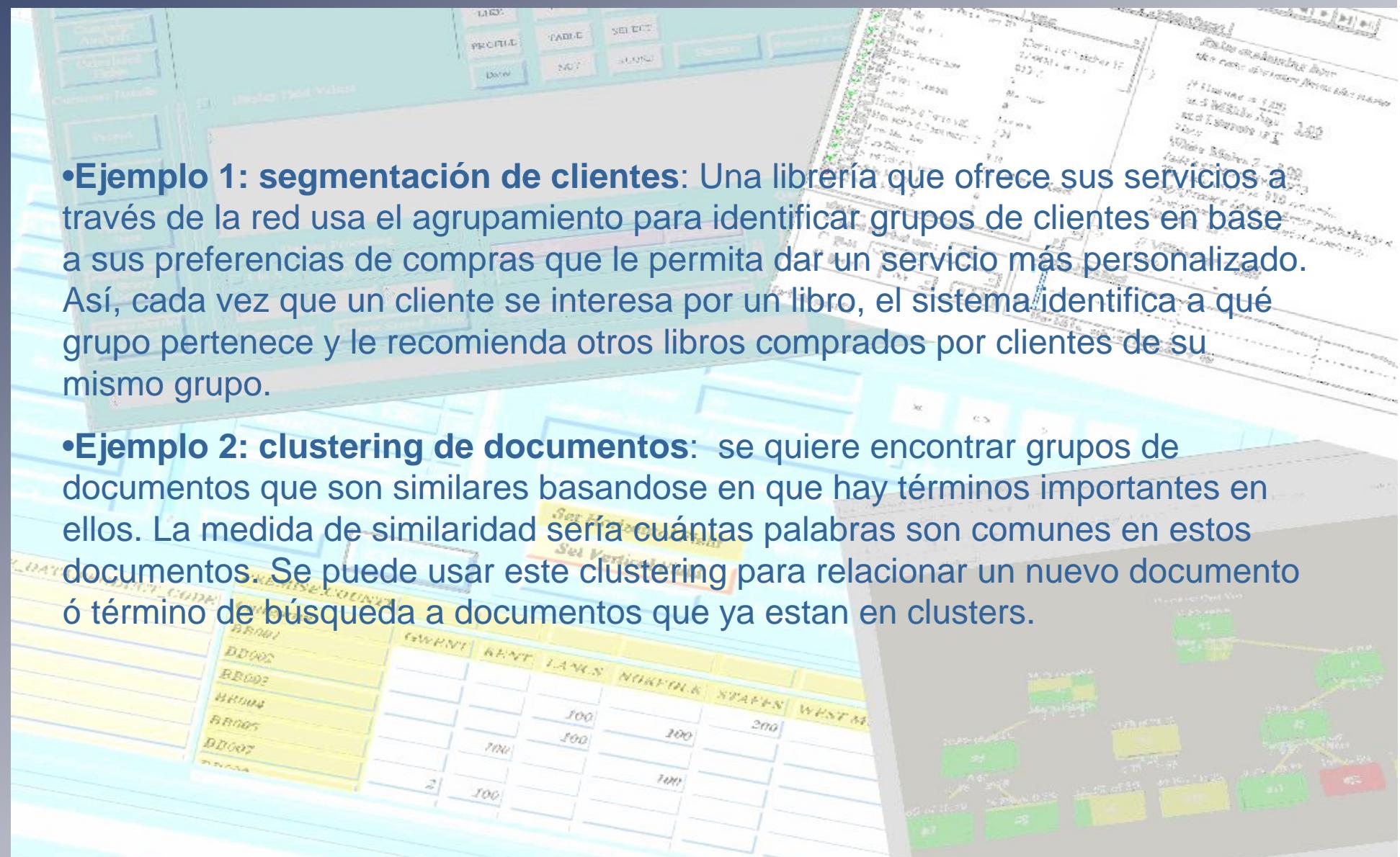
Modelado del datamining: Clustering

Celia Gutiérrez Cossío
2007

- La agrupación ó clustering es la tarea descriptiva por excelencia y consiste en obtener grupos "naturales" a partir de los datos. Hablamos de grupos y no de clases, porque, a diferencia de la clasificación, en lugar de analizar datos etiquetados con una clase, los analiza para generar esta etiqueta.
- Cada grupo se le llama **cluster** y un registro pertenece exactamente a un cluster, excepto en la modalidad de cluster solapados, donde un registro puede pertenecer a más de un cluster.
- Se le suele llamar **segmentación**, ya que parte o segmenta los datos en grupos.

➤ Los datos son agrupados basándose en el principio de maximizar la **similaridad** entre los elementos de un grupo minimizando la similitud entre los distintos grupos. Es decir, se forman grupos tales que los objetos de un mismo grupo son muy similares entre sí y, al mismo tiempo, son muy diferentes a los objetos de otro grupo. Para calcular la similaridad entre registros se usa una **función de distancia**: toma como entradas dos registros y saca como salida un valor que devuelve su similitud. La noción de similaridad varía entre aplicaciones. En el gráfico inferior se pueden observar 3 grupos de registros similares en cuanto a las variables salario y edad: clientes jóvenes con salarios altos, clientes jóvenes con salarios bajos, clientes mayores con salarios altos.





- **Ejemplo 1: segmentación de clientes:** Una librería que ofrece sus servicios a través de la red usa el agrupamiento para identificar grupos de clientes en base a sus preferencias de compras que le permita dar un servicio más personalizado. Así, cada vez que un cliente se interesa por un libro, el sistema identifica a qué grupo pertenece y le recomienda otros libros comprados por clientes de su mismo grupo.
- **Ejemplo 2: clustering de documentos:** se quiere encontrar grupos de documentos que son similares basandose en que hay términos importantes en ellos. La medida de similaridad sería cuántas palabras son comunes en estos documentos. Se puede usar este clustering para relacionar un nuevo documento ó término de búsqueda a documentos que ya estan en clusters.

➤ Está muy relacionado con la **sumarización**, que algunos autores consideran una tarea en sí misma, en la que cada grupo formado se considera como un resumen de los elementos que lo forman para así describir de una manera concisa los datos. Esta es la forma que tienen las salidas de muchos algoritmos de clustering. El tipo de representación summarizada depende fuertemente del tipo y forma de los clusters que el algoritmo computa.

Ejemplo: para la gráfica anterior se puede computar cada cluster mediante su centro C, y su radio R, calculados en función de los valores de los registros que componen el cluster.

$$C = \sum_{i=1}^n r_i / n$$

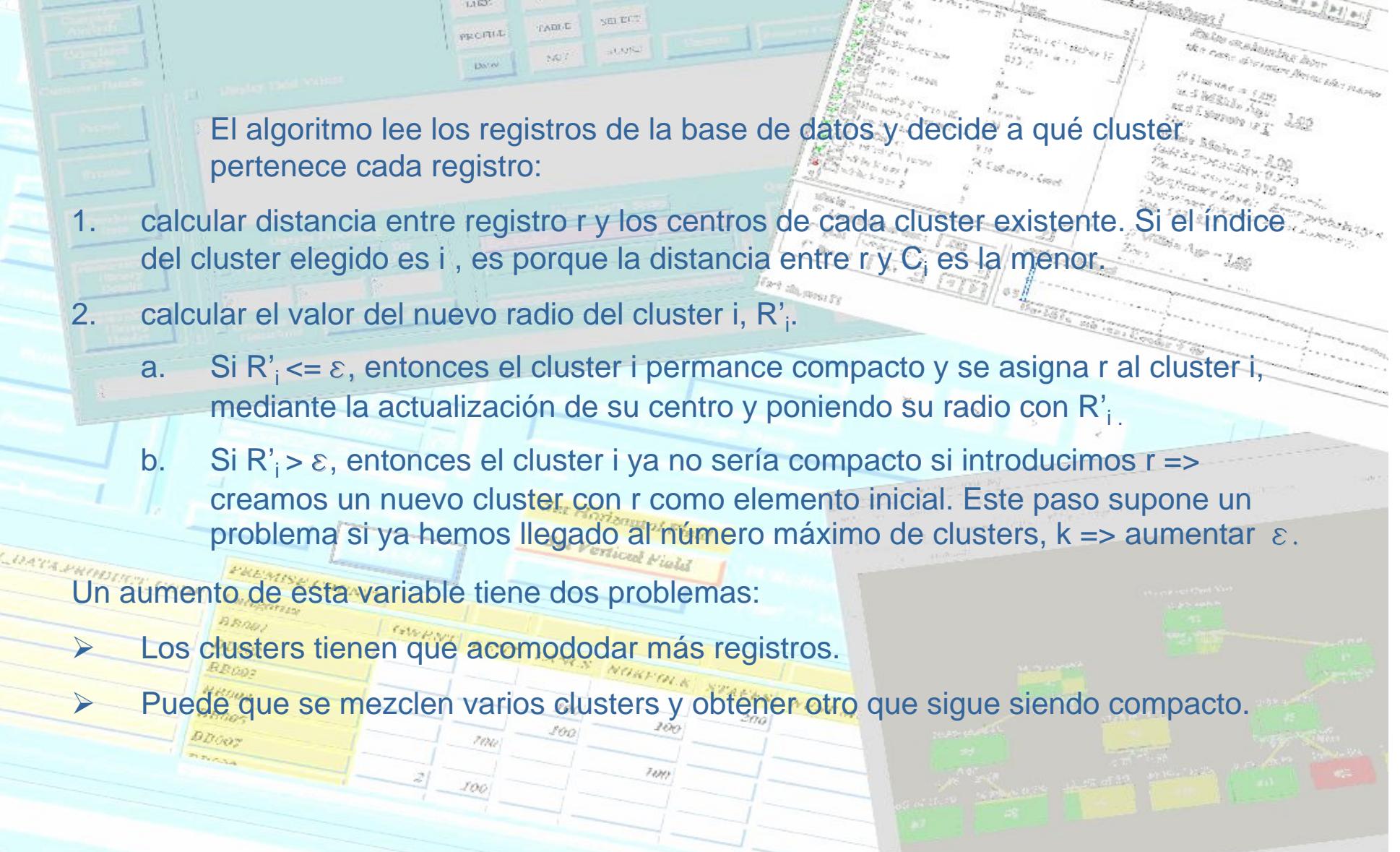
$$R = \sum_{i=1}^n \text{distancia}(r_i, C) / n$$

➤ Tipos de algoritmos:

1. Un algoritmo de clustering particional divide los datos en k grupos, de tal manera que son evaluados mediante un criterio que define si la calidad de clustering es óptima. El número de clusters, k, es un parámetro cuyo valor es definido por el usuario.
2. Un algoritmo de clustering jerárquico genera una secuencia de particiones en los registros: comenzando con una partición en la cual cada cluster tiene un solo registro, el algoritmo mezcla dos particiones en cada paso hasta que una sola partición permanece al final.

➤ Un algoritmo de clustering : BIRCH

El usuario debe dar valores a dos parámetros: k (número máximo de clusters), ε (radio máximo de los clusters). Se dice que un cluster es compacto si su radio es menor que ε . BIRCH siempre mantiene k o menos resúmenes de clusters (C_i, R_i) en memoria, siendo el primer componente el centro del cluster i, y el segundo componente el radio del cluster i.



El algoritmo lee los registros de la base de datos y decide a qué cluster pertenece cada registro:

1. calcular distancia entre registro r y los centros de cada cluster existente. Si el índice del cluster elegido es i , es porque la distancia entre r y C_i es la menor.
2. calcular el valor del nuevo radio del cluster i , R'_i .
 - a. Si $R'_i \leq \varepsilon$, entonces el cluster i permanece compacto y se asigna r al cluster i , mediante la actualización de su centro y poniendo su radio con R'_i .
 - b. Si $R'_i > \varepsilon$, entonces el cluster i ya no sería compacto si introducimos $r \Rightarrow$ creamos un nuevo cluster con r como elemento inicial. Este paso supone un problema si ya hemos llegado al número máximo de clusters, $k \Rightarrow$ aumentar ε .

Un aumento de esta variable tiene dos problemas:

- Los clusters tienen que acomodar más registros.
- Puede que se mezclen varios clusters y obtener otro que sigue siendo compacto.

➤ Ejercicio:

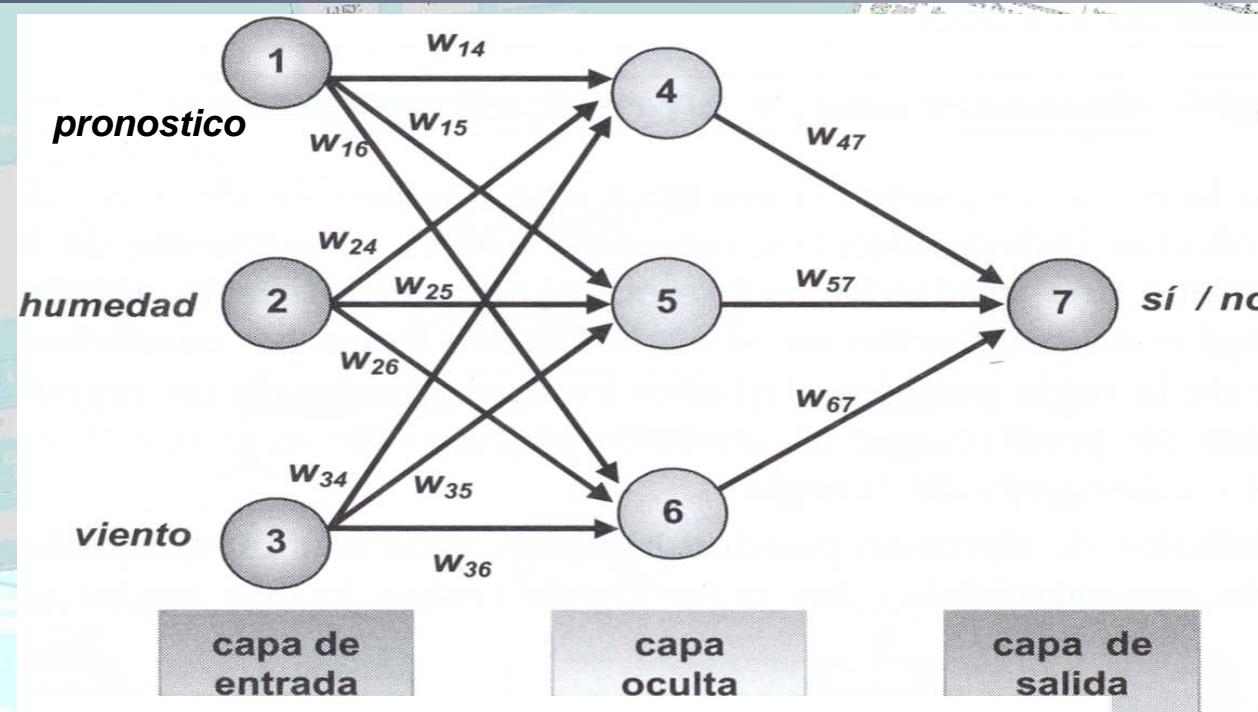
Dados los registros: (7,55), (21,202), (25, 220), (12, 73), (8, 61), (22, 249):

1. Asumiendo que pertenecen al mismo cluster, calcular su centro y su radio.
2. Asumiendo que los 3 primeros registros pertenecen a un cluster y los 3 últimos pertenecen a otro cluster, calcular centro y radio de estos clusters.
3. ¿Cuál de las dos maneras de hacer clustering es mejor y por qué?

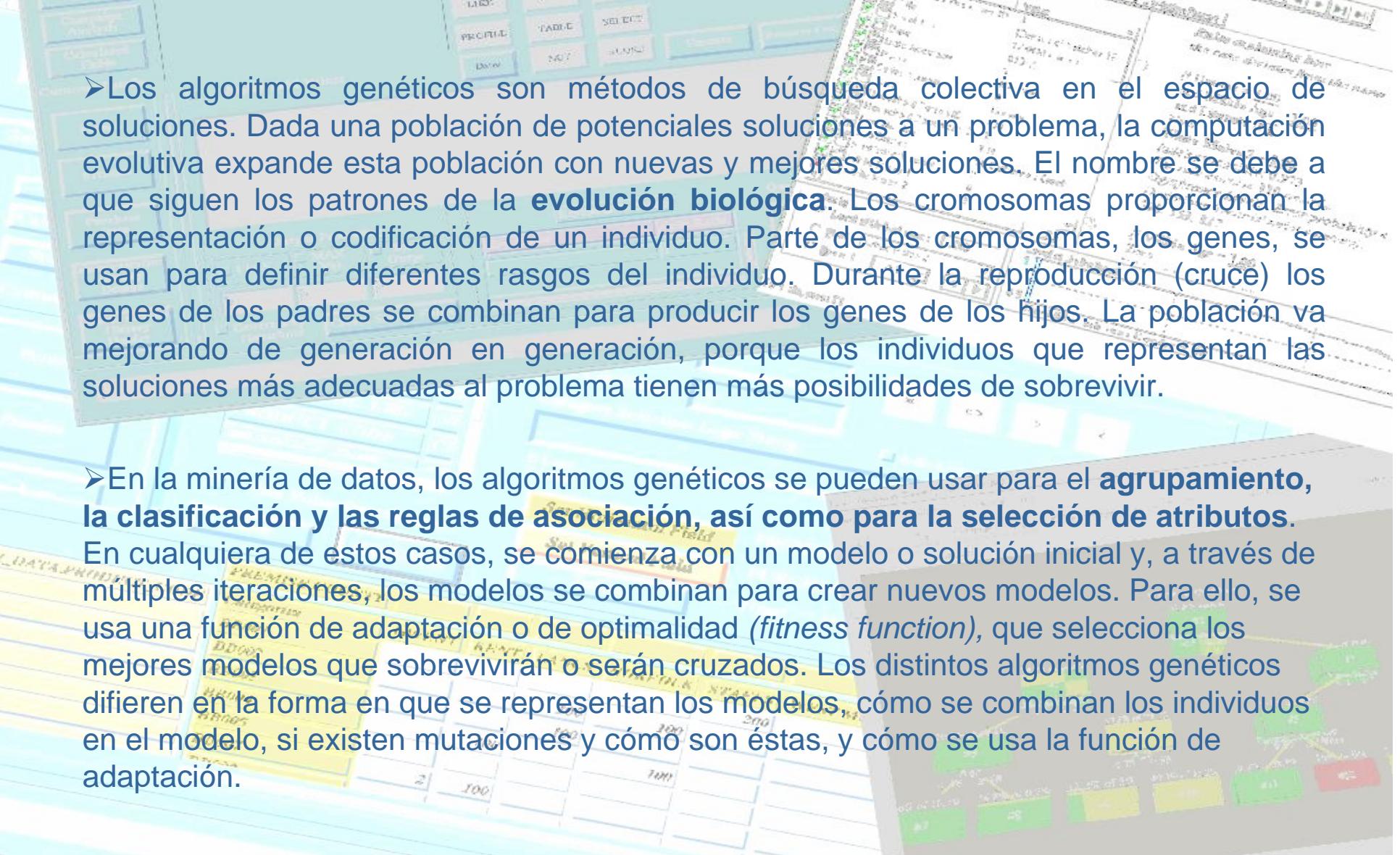
- Paradigma de computación muy **potente** que permite modelizar problemas complejos en los que puede haber interacciones no lineales entre variables.
- Como los árboles de decisión, las redes neuronales pueden usarse en problemas de **clasificación, de regresión y de agrupamiento**.
- Las redes neuronales trabajan directamente con **datos numéricos**. Para usarlas con datos nominales éstos deben numerizarse primero.
- Una red neuronal puede verse como un **grafo** dirigido con muchos nodos (elementos del proceso) y arcos entre ellos (sus interconexiones). Cada uno de estos elementos funciona independientemente de los demás, usando datos locales (la entrada y la salida del nodo) para dirigir su procesamiento. Se componen de una capa de entrada con las variables independientes, una capa de salida con valores para variable(s) a calcular. Cada arco contiene un peso y cada nodo tiene una función de activación sobre los datos que entran en él. Los pesos tienen valores que deben estimarse por un método de entrenamiento, lo cual exige bastantes datos (para su entrenamiento).
- **Ejemplo:** red neuronal para decidir si se juega determinado deporte.

Modelado del datamining: redes neuronales

Celia Gutiérrez Cossío
2007



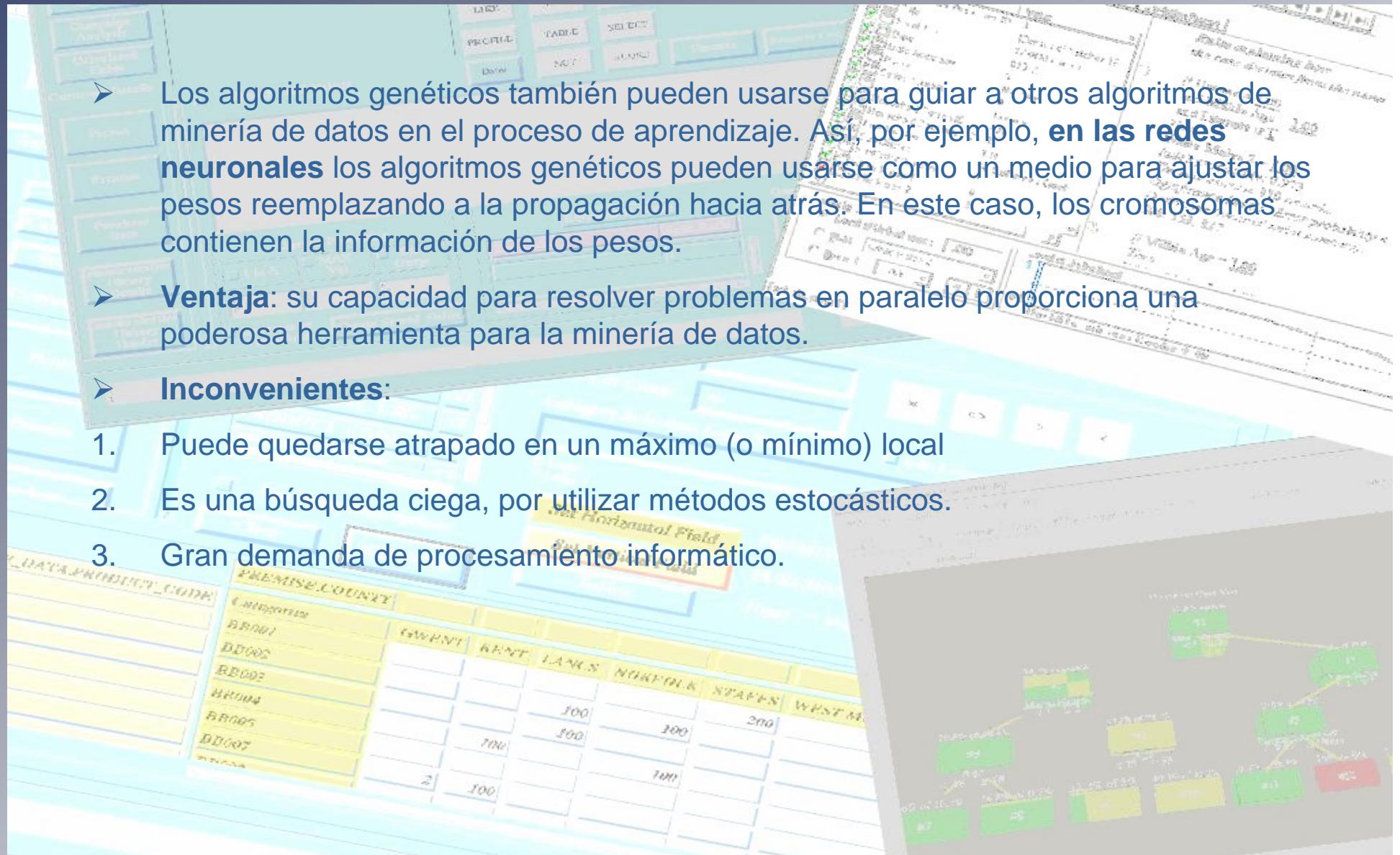
- Durante el proceso, las funciones y los pesos actúan sobre las entradas de los nodos.
- Dada la tupla de entrada (pronóstico, humedad, viento) con los valores de los 3 atributos de entrada: la salida del nodo 1 sería $f_1(\text{pronóstico})$, la del nodo 2 sería $f_2(\text{humedad})$ y la del 3 sería $f_3(\text{viento})$.
- Similarmente la salida del nodo 4 sería $f_4(w_{14}f_1(\text{pronóstico})+w_{24}f_2(\text{humedad})+w_{34}f_3(\text{viento}))$.

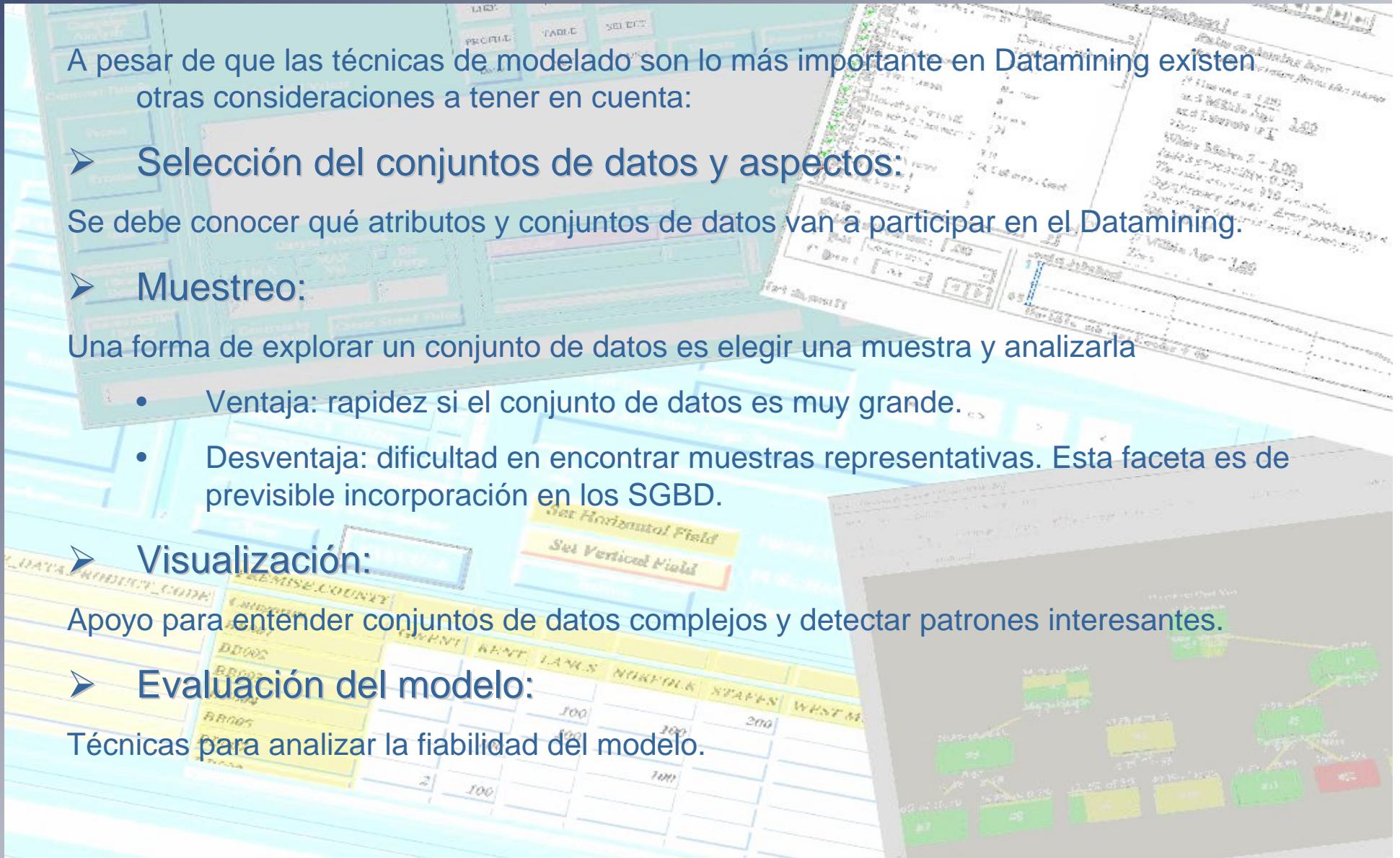


➤ Los algoritmos genéticos son métodos de búsqueda colectiva en el espacio de soluciones. Dada una población de potenciales soluciones a un problema, la computación evolutiva expande esta población con nuevas y mejores soluciones. El nombre se debe a que siguen los patrones de la **evolución biológica**. Los cromosomas proporcionan la representación o codificación de un individuo. Parte de los cromosomas, los genes, se usan para definir diferentes rasgos del individuo. Durante la reproducción (cruce) los genes de los padres se combinan para producir los genes de los hijos. La población va mejorando de generación en generación, porque los individuos que representan las soluciones más adecuadas al problema tienen más posibilidades de sobrevivir.

➤ En la minería de datos, los algoritmos genéticos se pueden usar para el **agrupamiento, la clasificación y las reglas de asociación, así como para la selección de atributos**. En cualquiera de estos casos, se comienza con un modelo o solución inicial y, a través de múltiples iteraciones, los modelos se combinan para crear nuevos modelos. Para ello, se usa una función de adaptación o de optimalidad (*fitness function*), que selecciona los mejores modelos que sobrevivirán o serán cruzados. Los distintos algoritmos genéticos difieren en la forma en que se representan los modelos, cómo se combinan los individuos en el modelo, si existen mutaciones y cómo son éstas, y cómo se usa la función de adaptación.

- Los algoritmos genéticos también pueden usarse para guiar a otros algoritmos de minería de datos en el proceso de aprendizaje. Así, por ejemplo, **en las redes neuronales** los algoritmos genéticos pueden usarse como un medio para ajustar los pesos reemplazando a la propagación hacia atrás. En este caso, los cromosomas contienen la información de los pesos.
- **Ventaja:** su capacidad para resolver problemas en paralelo proporciona una poderosa herramienta para la minería de datos.
- **Inconvenientes:**
 1. Puede quedarse atrapado en un máximo (o mínimo) local
 2. Es una búsqueda ciega, por utilizar métodos estocásticos.
 3. Gran demanda de procesamiento informático.





A pesar de que las técnicas de modelado son lo más importante en Datamining existen otras consideraciones a tener en cuenta:

➤ **Selección del conjuntos de datos y aspectos:**

Se debe conocer qué atributos y conjuntos de datos van a participar en el Datamining.

➤ **Muestreo:**

Una forma de explorar un conjunto de datos es elegir una muestra y analizarla

- Ventaja: rapidez si el conjunto de datos es muy grande.
- Desventaja: dificultad en encontrar muestras representativas. Esta faceta es de previsible incorporación en los SGBD.

➤ **Visualización:**

Apoyo para entender conjuntos de datos complejos y detectar patrones interesantes.

➤ **Evaluación del modelo:**

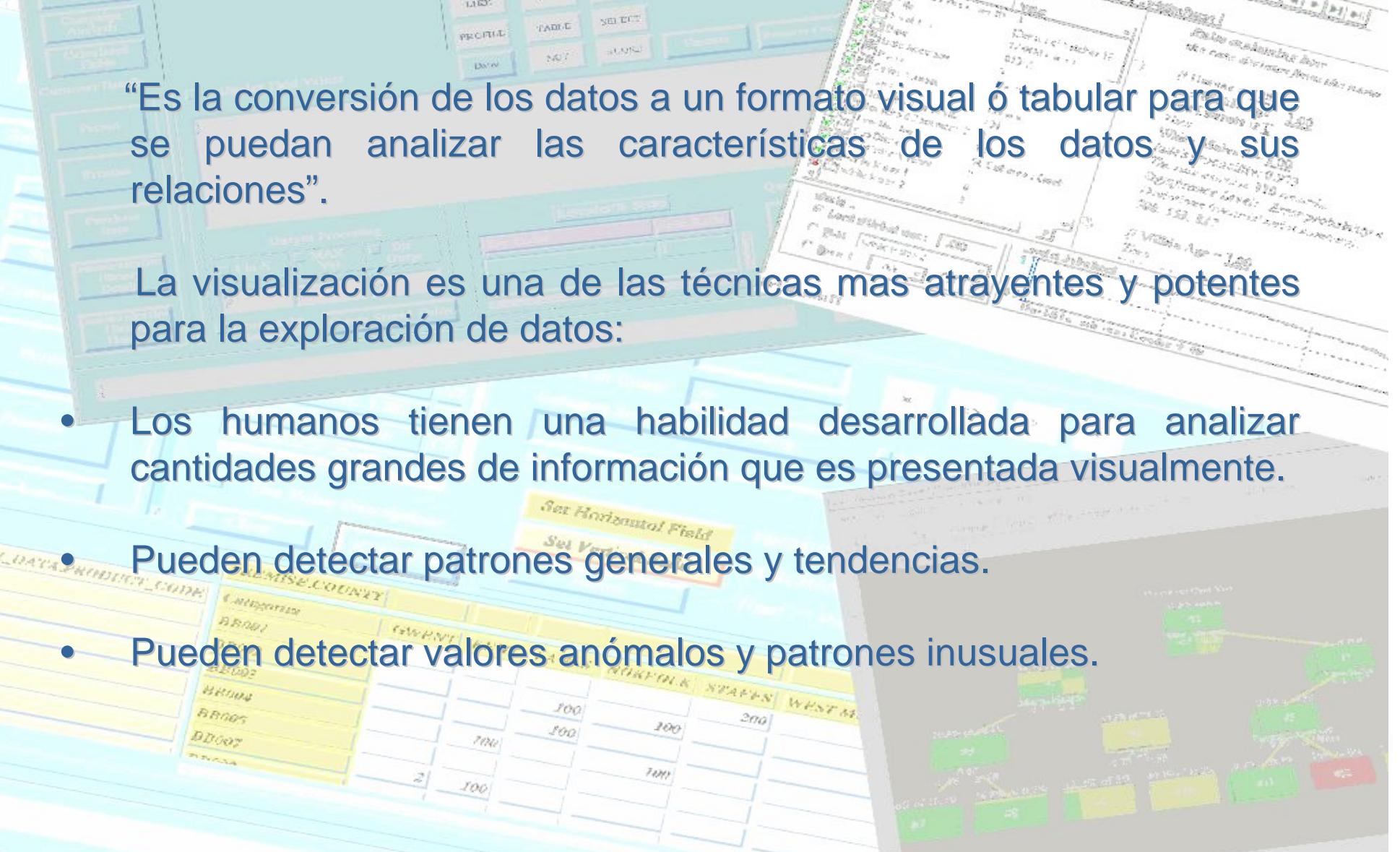
Técnicas para analizar la fiabilidad del modelo.

➤ Técnicas de muestreo:

- Muestreo al azar: existe una probabilidad igual de seleccionar cualquier elemento.
- Muestreo sin reemplazo: cuando un ítem es seleccionado, se quita de la población.
- Muestreo con reemplazo: cuando un ítem es seleccionado, no se quita de la población y puede ser seleccionado de nuevo.
- Muestreo estratificado: dividir los datos en varias particiones; para cada una de ellas obtener muestreo al azar.

Otras consideraciones: visualización

Celia Gutiérrez Cossío
2007



“Es la conversión de los datos a un formato visual ó tabular para que se puedan analizar las características de los datos y sus relaciones”.

La visualización es una de las técnicas mas atrayentes y potentes para la exploración de datos:

- Los humanos tienen una habilidad desarrollada para analizar cantidades grandes de información que es presentada visualmente.
- Pueden detectar patrones generales y tendencias.
- Pueden detectar valores anómalos y patrones inusuales.

Otras consideraciones: visualización

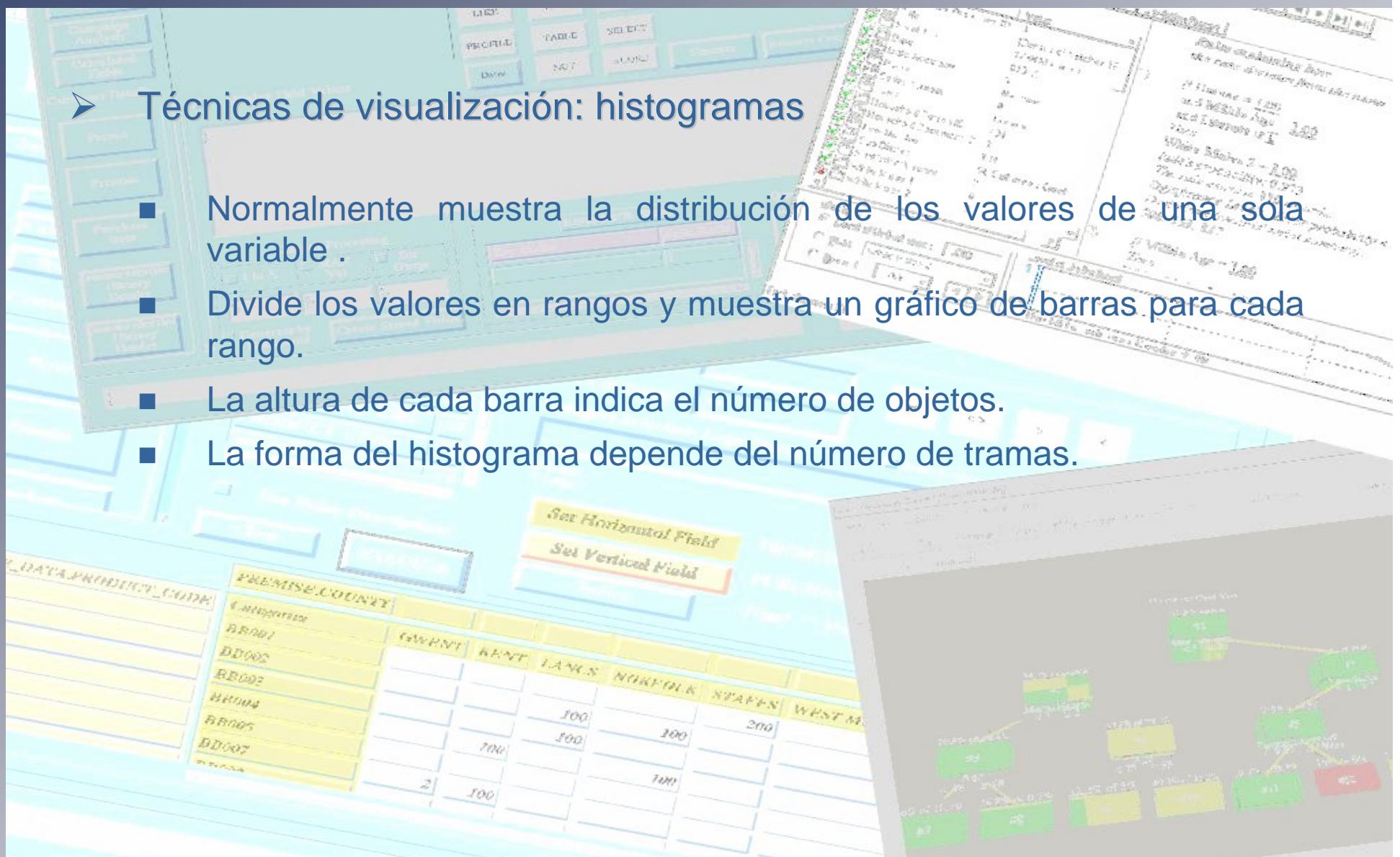
Celia Gutiérrez Cossío
2007

Se usan varias técnicas de representación de datos:

- “Es el mapeo de la información a un formato visual”.
- Los objetos de datos, sus atributos, y las relaciones entre objetos de datos se traducen a elementos gráficos como puntos, líneas, formas y colores.
- Ejemplo:
 - Los objetos se suelen representar como puntos.
 - Los valores de atributos se pueden representar como la posición de los puntos ó las características de los puntos (colores, coordenadas, ...).
 - Si se usa la posición, entonces las relaciones de los puntos (i.e. si forman grupo ó si es un valor anómalo), se pueden percibir fácilmente.

➤ Técnicas de visualización: histogramas

- Normalmente muestra la distribución de los valores de una sola variable .
- Divide los valores en rangos y muestra un gráfico de barras para cada rango.
- La altura de cada barra indica el número de objetos.
- La forma del histograma depende del número de tramas.

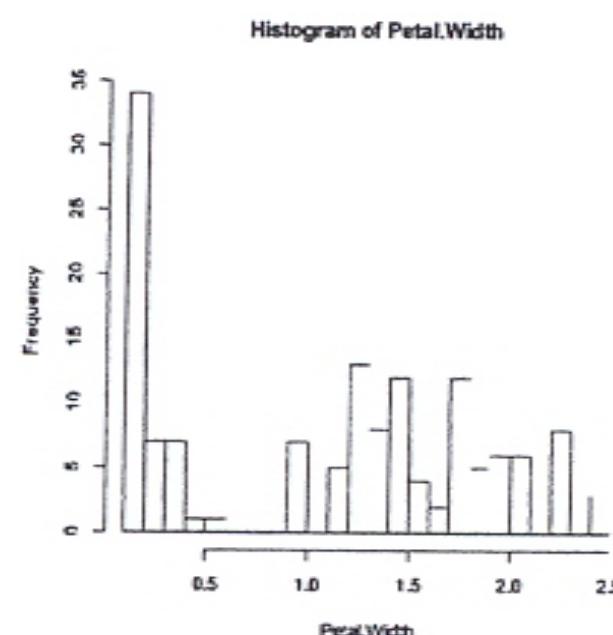
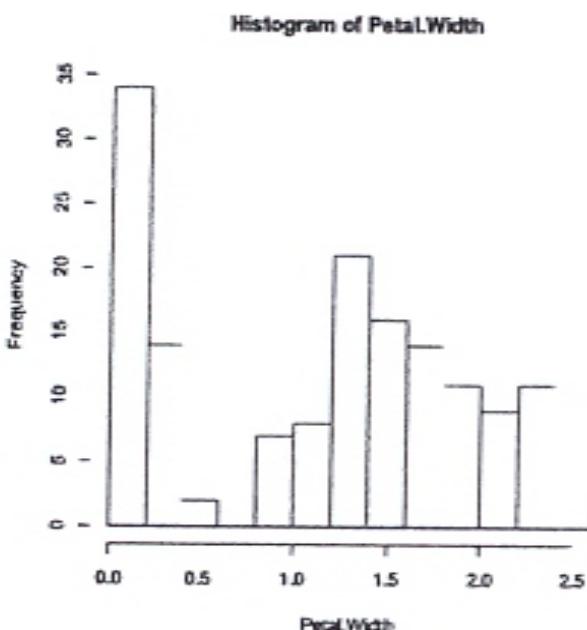


Otras consideraciones: visualización

Celia Gutiérrez Cossío
2007

- Técnicas de visualización: histogramas

Ejemplo: Petal.Width

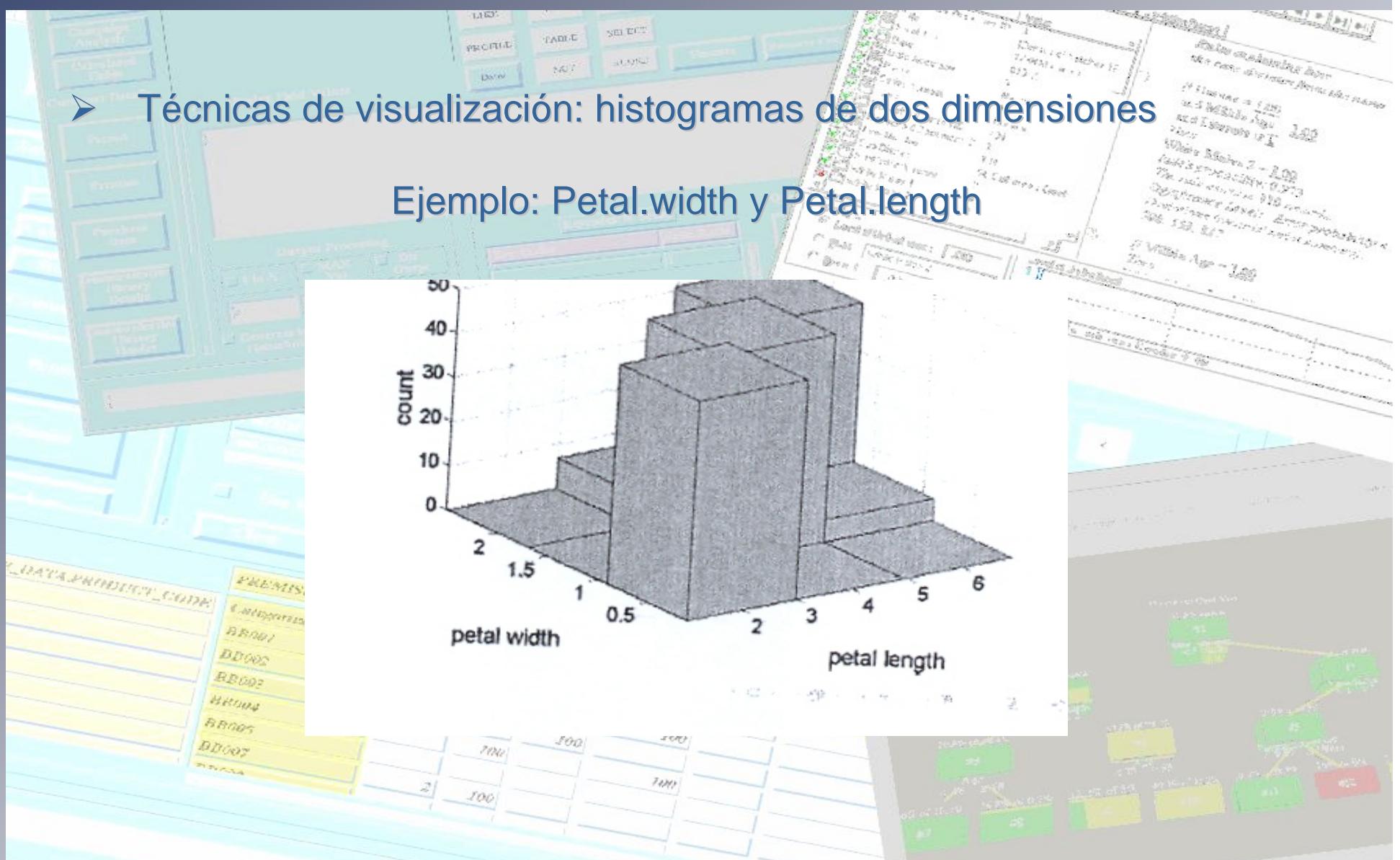
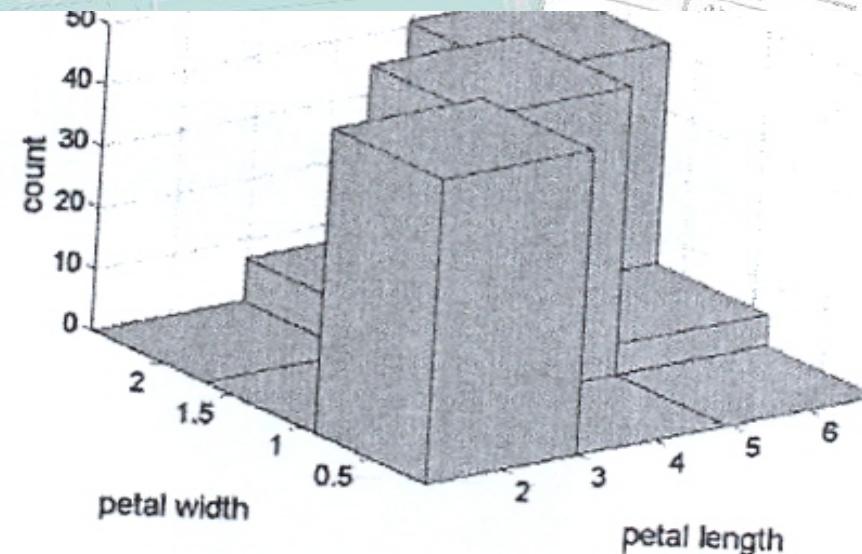


Otras consideraciones: visualización

Celia Gutiérrez Cossío
2007

- Técnicas de visualización: histogramas de dos dimensiones

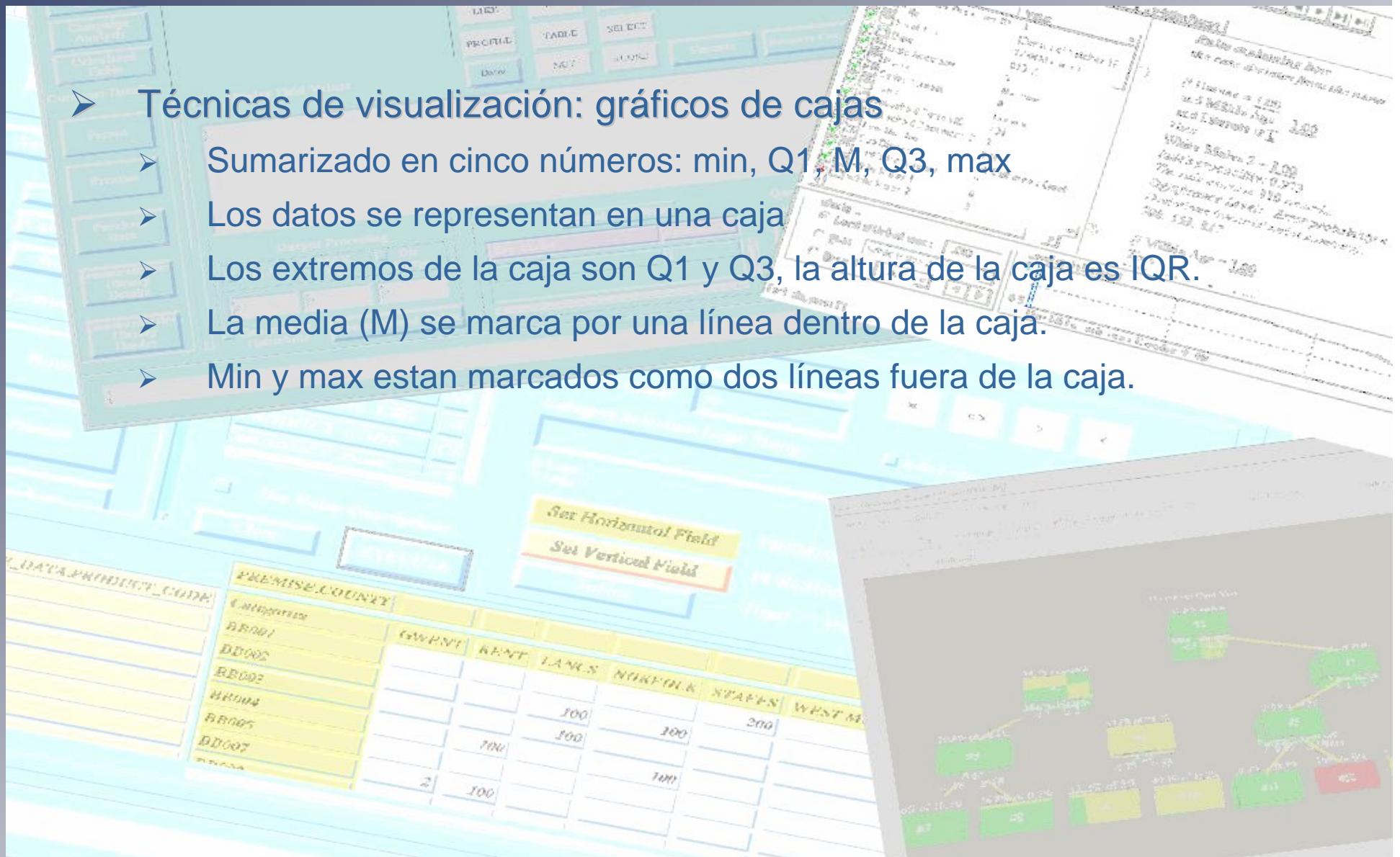
Ejemplo: Petal.width y Petal.length



Otras consideraciones: visualización

Celia Gutiérrez Cossío
2007

- Técnicas de visualización: gráficos de cajas
 - Sumarizado en cinco números: min, Q1, M, Q3, max
 - Los datos se representan en una caja
 - Los extremos de la caja son Q1 y Q3, la altura de la caja es IQR.
 - La media (M) se marca por una línea dentro de la caja.
 - Min y max están marcados como dos líneas fuera de la caja.

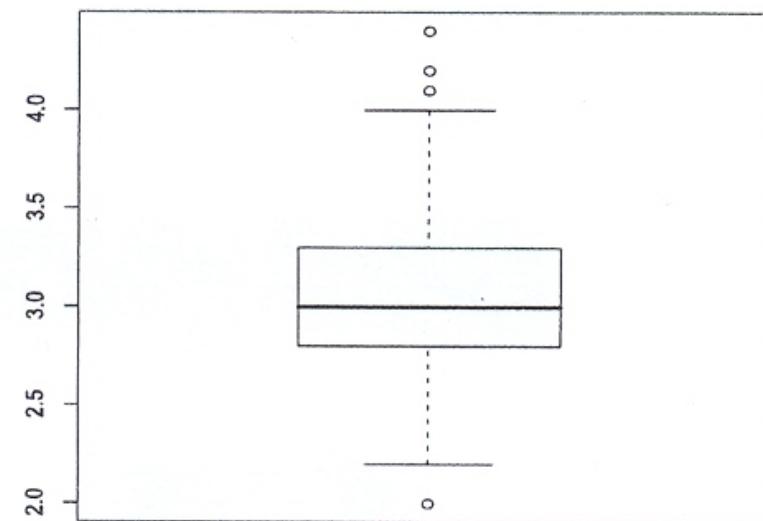


Otras consideraciones: visualización

Celia Gutiérrez Cossío
2007

- Técnicas de visualización: gráficos de cajas

Ejemplo: Iris

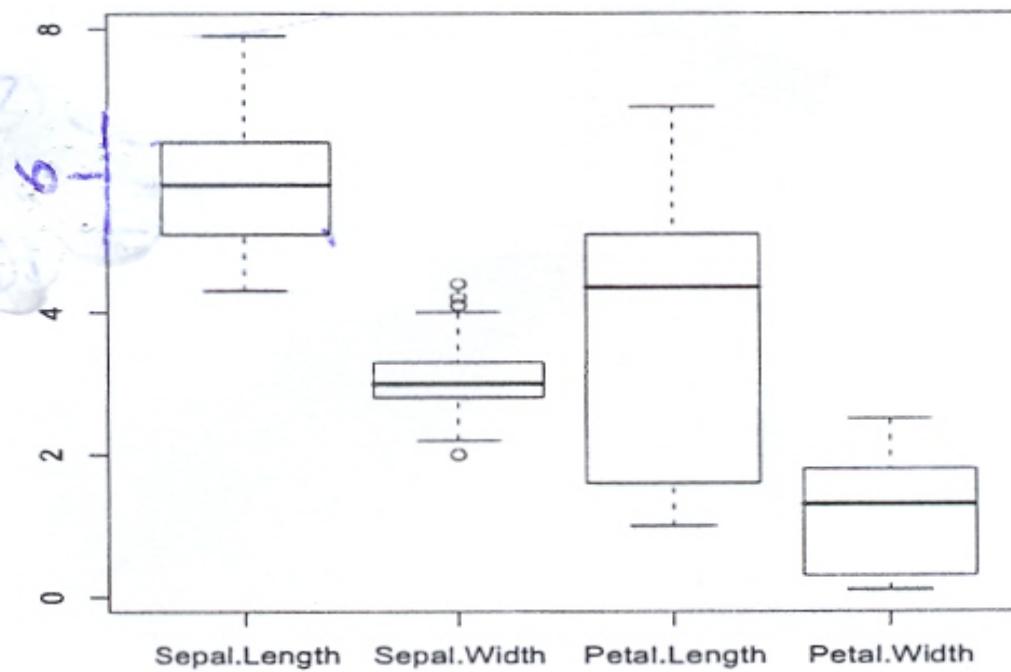


Otras consideraciones: visualización

Celia Gutiérrez Cossío
2007

- Técnicas de visualización: gráficos de cajas

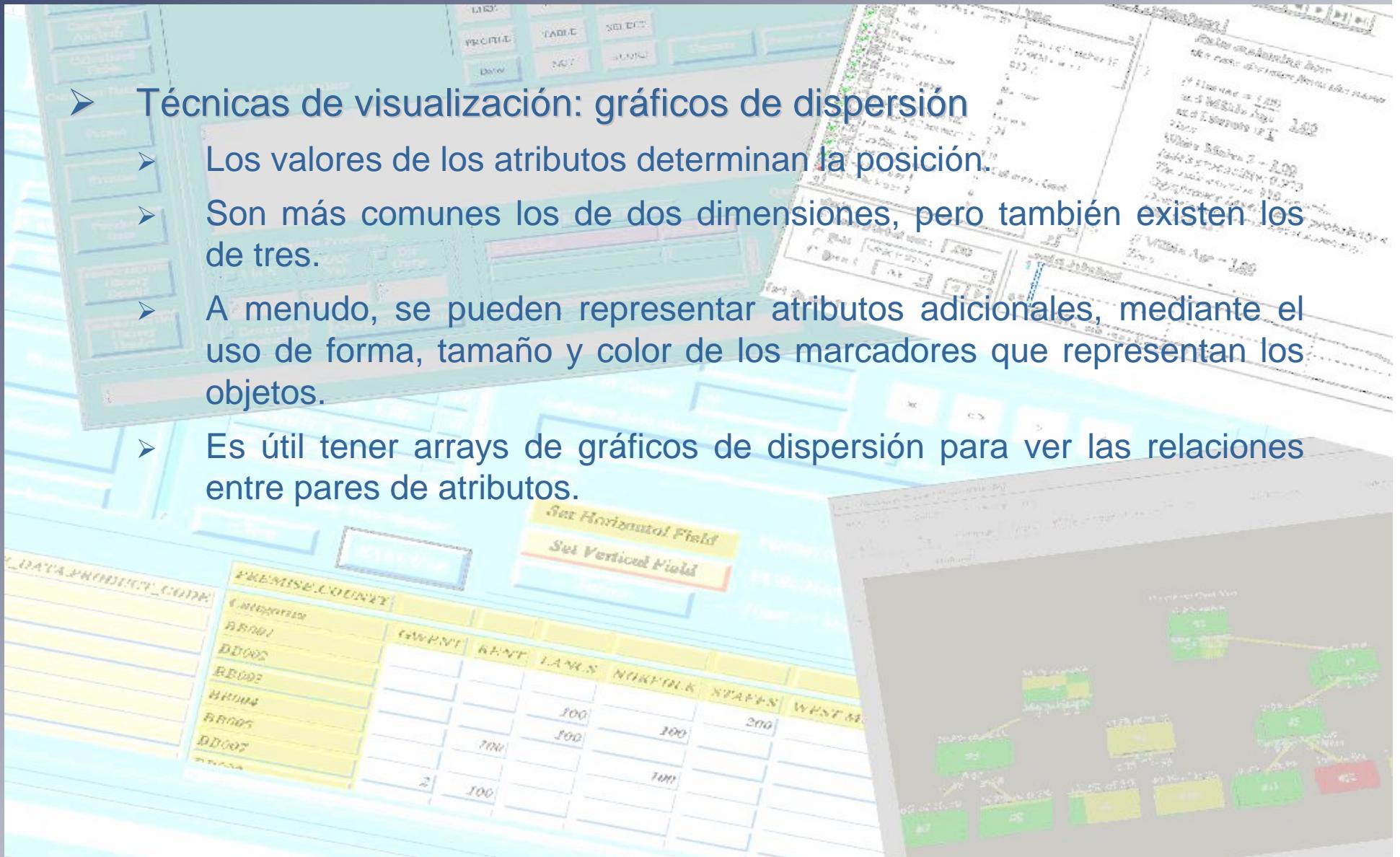
Ejemplo: Iris



Otras consideraciones: visualización

Celia Gutiérrez Cossío
2007

- Técnicas de visualización: gráficos de dispersión
 - Los valores de los atributos determinan la posición.
 - Son más comunes los de dos dimensiones, pero también existen los de tres.
 - A menudo, se pueden representar atributos adicionales, mediante el uso de forma, tamaño y color de los marcadores que representan los objetos.
 - Es útil tener arrays de gráficos de dispersión para ver las relaciones entre pares de atributos.

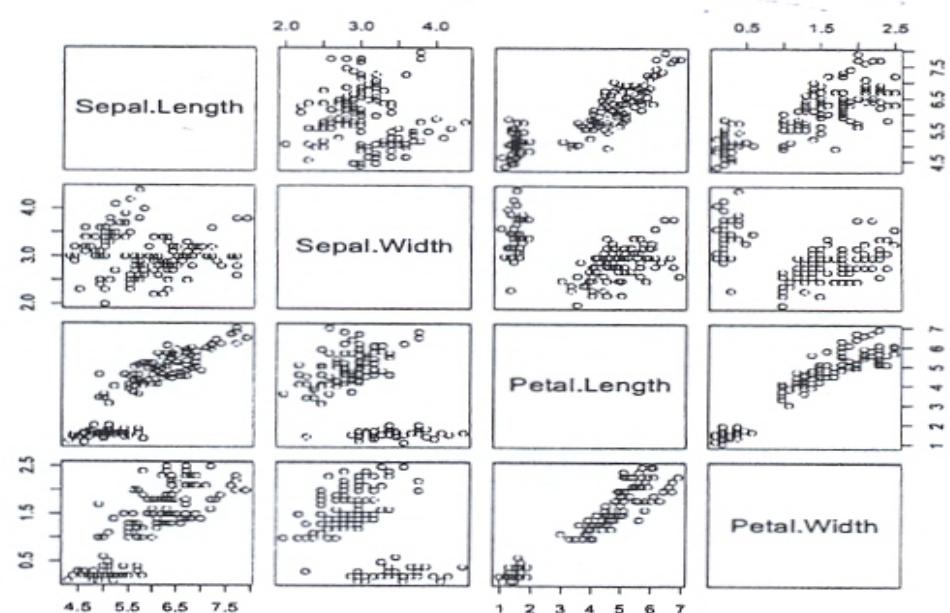


Otras consideraciones: visualización

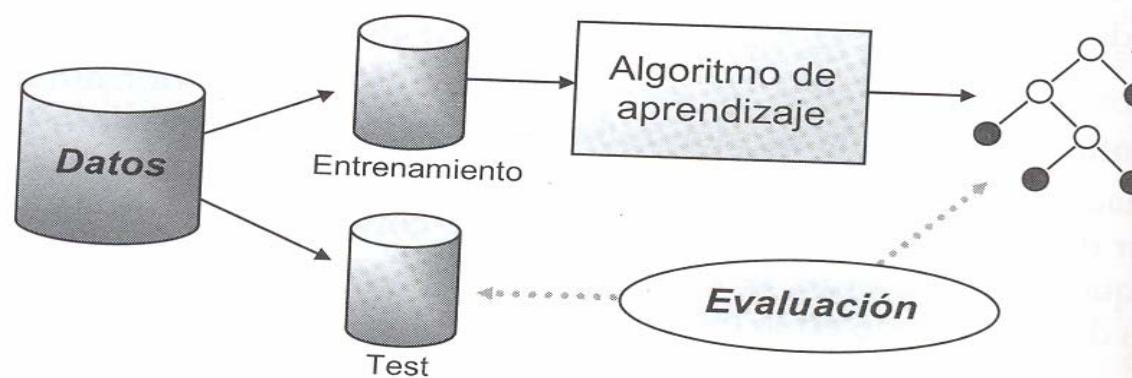
Celia Gutiérrez Cossío
2007

➤ Técnicas de visualización: gráficos de dispersión

Example Iris



- Técnicas de evaluación: matriz de confusión
- Se usa en métodos de clasificación.
 - Indica cómo se distribuyen los errores.
 - Se basa en dos conjuntos: training set (modelo de entrenamiento) y test set (modelo de evaluación):



- Problema: los resultados dependen de cómo sean ambos conjuntos, que se componen de manera aleatoria. Además, si disponemos de pocos datos, y reservamos datos para el test set, el resultado se puede falsear todavía más.

- Técnicas de evaluación: matriz de confusión
 - Matriz de confusión: se distribuyen en filas las clases estimadas y por columnas las clases reales, y se colocan en los cruces el número de muestras que coinciden en esa clase real y estimada.

Estimado	Real		
	Salida	Observación	UCI
Salida	71	3	1
Observación	8	7	1
UCI	4	2	3

- Técnicas de evaluación: matriz de confusión
 - Precisión de un clasificador: nº aciertos(en la diagonal) / nº casos
Para el ejemplo anterior: $81/100 = 81\%$
 - Matriz de costes: se asocia un coste a cada elemento de la matriz de confusión en la que el caso estimado y el caso real no coinciden.

Estimado	Real		
	Salida	Observación	UCI
Salida	0 €	5.000 €	500.000 €
Observación	300 €	0 €	50.000 €
UCI	800 €	500 €	0 €

- Técnicas de evaluación: matriz de confusión
 - Coste total del clasificador:

$$\text{Coste} = \sum_{1 \leq i \leq n, 1 \leq j \leq n} C_{i,j} \cdot M_{i,j}$$

Donde C expresa la matriz de coste y M la matriz de confusión.

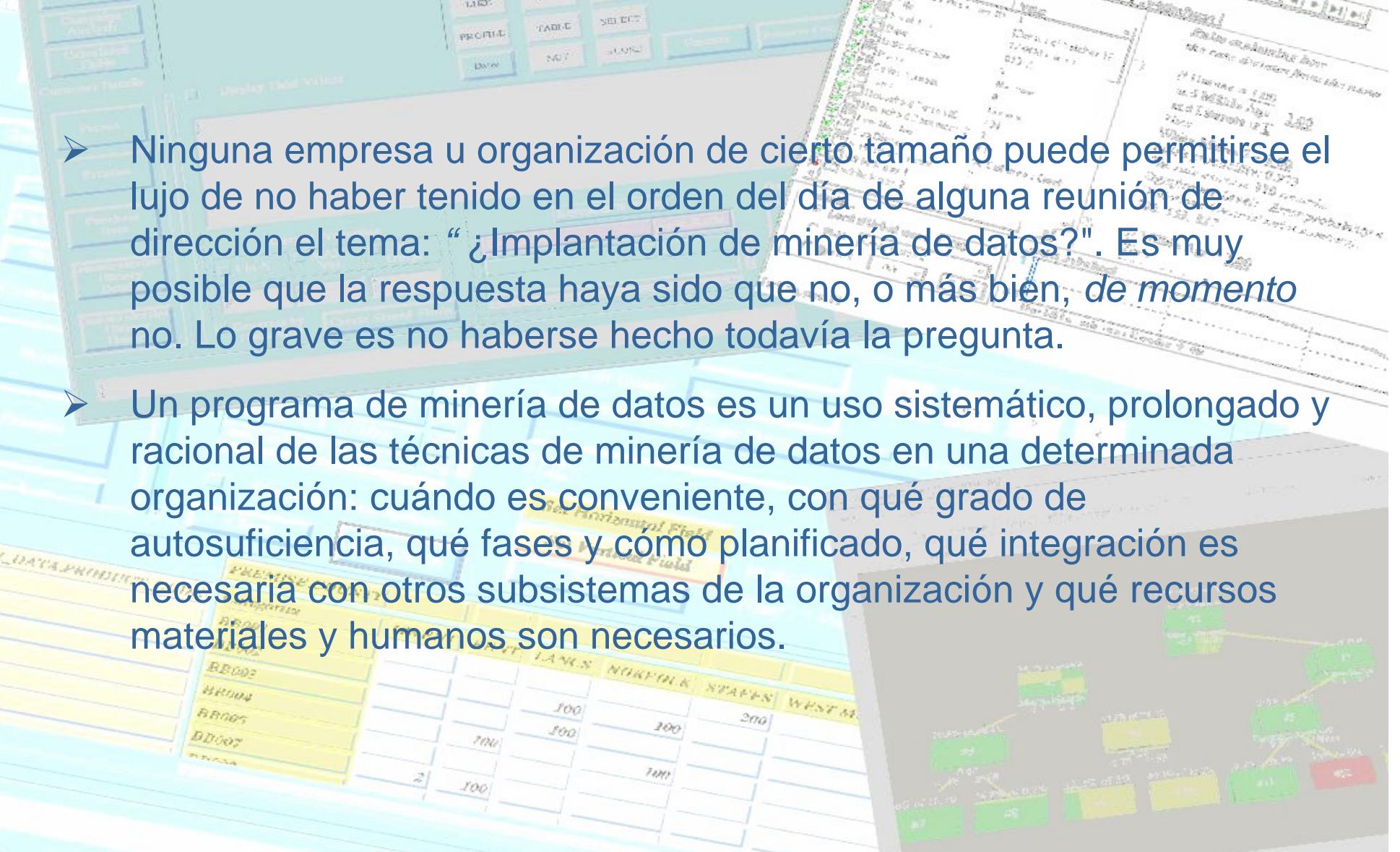
Para el caso anterior Coste=571.600 euros.

- Problema: saber cuales son los costes. P.e.: ¿Cuál es el coste que supone tener a un individuo en la UCI cuando en realidad no debería haber tenido tratamiento?

- R está muy en la línea con los métodos de modelado vistos, aunque refleja la investigación y los intereses docentes de un número reducido de colaboradores:
 - Además de la clase base R, se consideran otros paquetes como e1071, mclust, nnet, rpart, tree, klaR y muchos otros, que implementan una técnica de minería de datos.
 - Como R es un entorno programable, es fácil de implementar otros métodos.
 - Visualización de minería de datos: Base R tiene muchas de las técnicas mencionadas y VR tiene la mayor parte del resto.
- El mayor defecto es la inducción de reglas, especialmente la de reglas de asociación. Eso con bastante certeza no es accidental, sino que R carece de métodos usados por los estadísticos para manejar grandes conjuntos de variables categóricas.
- Práctica para relacionarse con R

Implantación de un proyecto de datamining

Celia Gutiérrez Cossío
2007

- 
- Ninguna empresa u organización de cierto tamaño puede permitirse el lujo de no haber tenido en el orden del día de alguna reunión de dirección el tema: “¿Implantación de minería de datos?”. Es muy posible que la respuesta haya sido que no, o más bien, *de momento* no. Lo grave es no haberse hecho todavía la pregunta.
 - Un programa de minería de datos es un uso sistemático, prolongado y racional de las técnicas de minería de datos en una determinada organización: cuándo es conveniente, con qué grado de autosuficiencia, qué fases y cómo planificado, qué integración es necesaria con otros subsistemas de la organización y qué recursos materiales y humanos son necesarios.

- La decisión de implantar un programa de minería de datos y el diseño de un plan del mismo deben preceder a cualquiera de las fases que se han visto en capítulos anteriores.
- De hecho, establecer cuál es el contexto del negocio, los objetivos del mismo y plasmarlos en objetivos de minería de datos, es previo a pararnos a pensar en recopilar y preparar los datos, realizar los modelos, evaluados y utilizados. Sin embargo, es muy difícil realizar un plan de minería de datos (o entender cómo se puede hacer uno) sin conocer la tecnología.
- De alguna manera este capítulo marca la diferencia entre la tecnología y la ingeniería: conocer la tecnología no asegura el éxito si no se sabe cómo aplicada en un contexto concreto y teniendo en cuenta unas limitaciones de costes, de recursos (tanto materiales como humanos), de plazos y demás aspectos que hacen de la ingeniería un híbrido entre la tecnología y la metodología. Este capítulo desarrolla cierta metodología más general sobre organización, gestión y planificación de proyectos de minería de datos.



Claves para el éxito en la implantación de un Datamining:

1. Especificar los problemas y objetivos de negocio=>qué datos van a ser necesarios y podrán surgir los objetivos y tareas de minería de datos.
2. Una buena especificación de problemas concretos y específicos de minería de datos: trasladar correctamente los objetivos de negocio a los objetivos concretos de minería de datos.
3. La integración del resto de programas de la organización.
4. La calidad de datos (ya sea por un programa previo o por limpieza)
5. El uso de herramientas integradas y de entornos amigables es otro factor destacable.
6. La necesidad de un equipo heterogéneo de personal formado no sólo en minería de datos, sino también en estadística, bases de datos y el área de negocio.
7. Una evaluación de los modelos más holista (teniendo en cuenta costes, comprensibilidad, relevancia, etc...) y un despliegue de los mismos a todos los niveles (personal de toma de decisiones, resto de usuarios, aplicaciones e incluso la propia base de datos transaccional).

➤ ¿Cuándo empezar? Necesidades y objetivos de negocio:

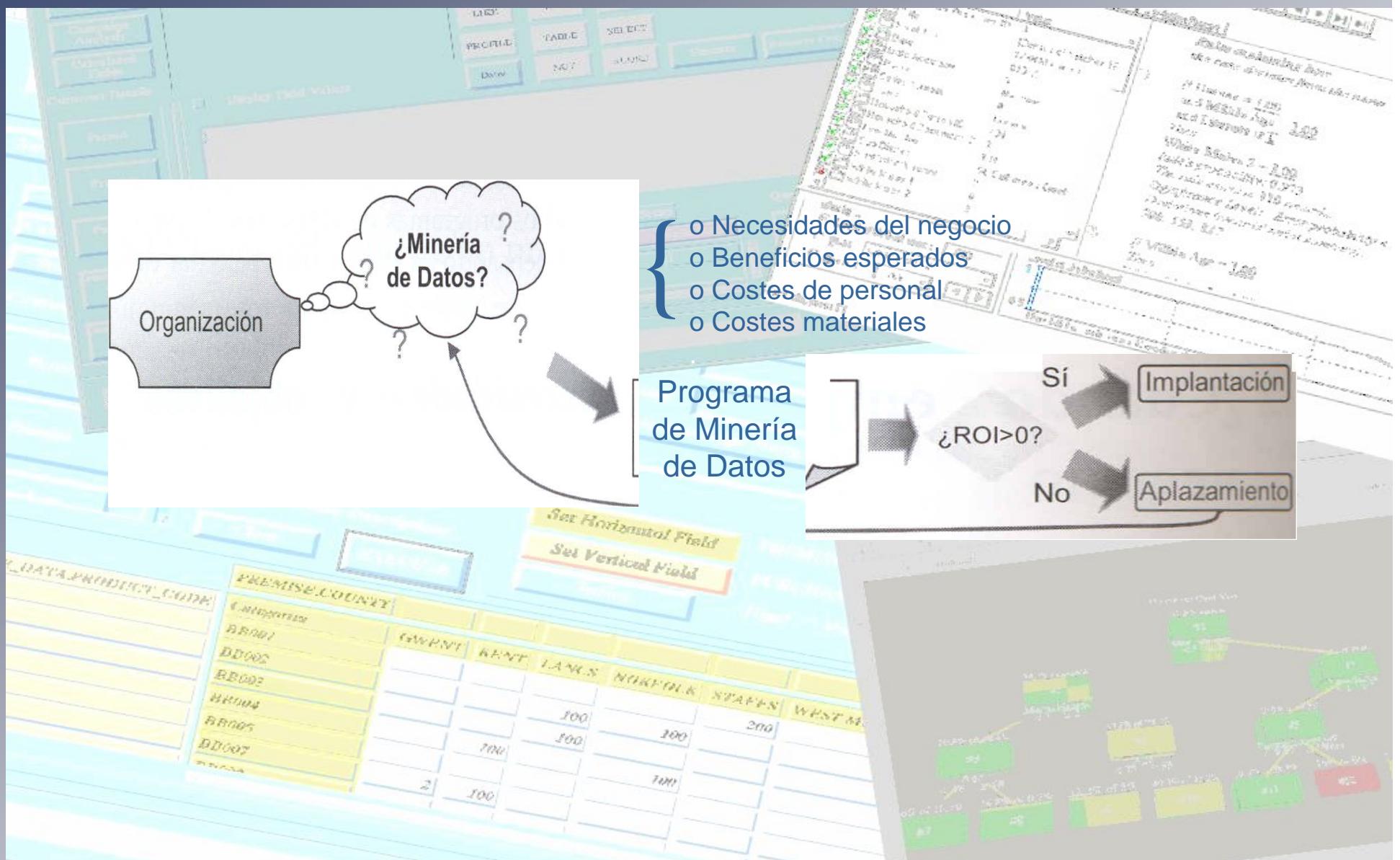
Los consultores, asesores y comerciales *productivos*, así como las revistas y expertos *especializados*, se las ingenian, especialmente en momentos de crisis, para ofrecernos productos que ya tenemos (con nombres y precios más "modernos") o vendernos productos que no necesitamos, pero que acabamos necesitando.

No obstante, aunque la minería de datos *funciona en general*, no funciona igual de bien en todos los ámbitos y, evidentemente, existirán organizaciones para las que será mayor el esfuerzo que el beneficio obtenido. En muchos casos no es una cuestión de tamaño o incluso de la rama de negocio, sino más bien dependerá de que se tomen decisiones importantes diariamente sobre un entorno cambiante y que exista una cierta tradición de informatización y de gestión de datos en la organización. ¿Qué necesito saber, por tanto, para tomar la decisión de si debo o no implantar minería de datos? En primer lugar, diseñar y tener entre las manos un programa de minería de datos.

Con un (esquema de) programa de minería de datos podemos estimar cuál va a ser el rendimiento o ROI (*Return On Investment*), evaluando los beneficios y evaluando de la manera más ajustada posible los costes, ya que implementar minería de datos puede requerir una inversión considerable en formación, herramientas y personal. El concepto de "beneficio" o de "productividad" es muy variable dependiendo de la organización.

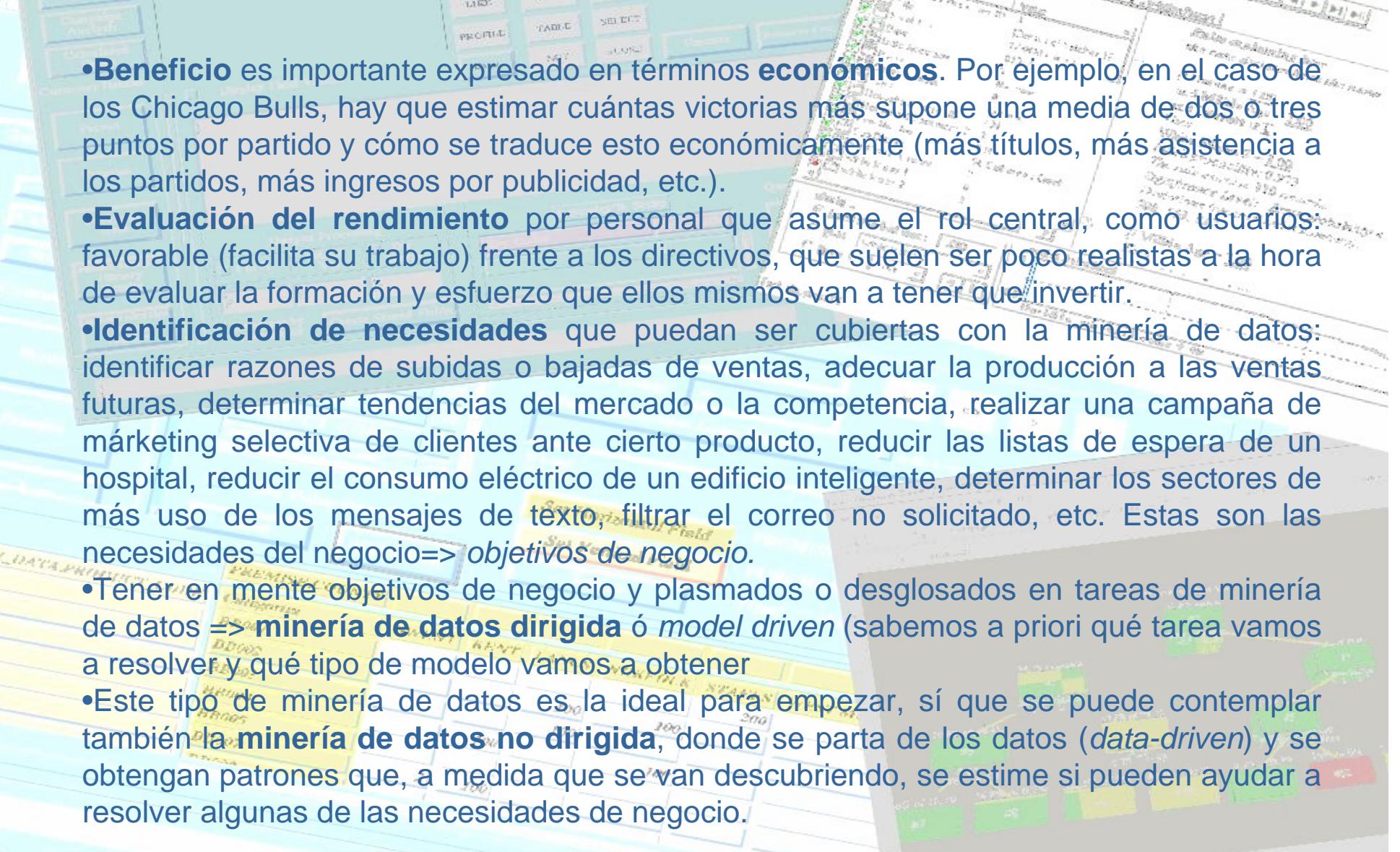
Implantación de un proyecto de datamining

Celia Gutiérrez Cossío
2007



Implantación de un proyecto de datamining

Celia Gutiérrez Cossío
2007

- 
- **Beneficio** es importante expresado en términos **económicos**. Por ejemplo, en el caso de los Chicago Bulls, hay que estimar cuántas victorias más supone una media de dos o tres puntos por partido y cómo se traduce esto económicoamente (más títulos, más asistencia a los partidos, más ingresos por publicidad, etc.).
 - **Evaluación del rendimiento** por personal que asume el rol central, como usuarios favorable (facilita su trabajo) frente a los directivos, que suelen ser poco realistas a la hora de evaluar la formación y esfuerzo que ellos mismos van a tener que invertir.
 - **Identificación de necesidades** que puedan ser cubiertas con la minería de datos: identificar razones de subidas o bajadas de ventas, adecuar la producción a las ventas futuras, determinar tendencias del mercado o la competencia, realizar una campaña de marketing selectiva de clientes ante cierto producto, reducir las listas de espera de un hospital, reducir el consumo eléctrico de un edificio inteligente, determinar los sectores de más uso de los mensajes de texto, filtrar el correo no solicitado, etc. Estas son las necesidades del negocio=> *objetivos de negocio*.
 - Tener en mente objetivos de negocio y plasmados o desglosados en tareas de minería de datos => **minería de datos dirigida** ó *model driven* (sabemos a priori qué tarea vamos a resolver y qué tipo de modelo vamos a obtener)
 - Este tipo de minería de datos es la ideal para empezar, sí que se puede contemplar también la **minería de datos no dirigida**, donde se parte de los datos (*data-driven*) y se obtengan patrones que, a medida que se van descubriendo, se estime si pueden ayudar a resolver algunas de las necesidades de negocio.

➤ ¿Subcontratar? Grados de autosuficiencia de un programa datamining

1. Mediante la compra de las puntuaciones (scores) o predicciones:

Alguna empresa o consultora externa tiene buenos modelos de negocio, obtenidos a partir de grandes bases de datos (de otras organizaciones similares). Le proporcionamos preguntas y ellos nos responden utilizando esos modelos. Por ejemplo, podríamos proporcionarles un listado de clientes y un producto, y la consultora nos puede devolver el listado acompañado de una probabilidad de compra de cada cliente del producto en cuestión (una puntuación, al fin al cabo, de cada cliente). Este modelo se utiliza frecuentemente en la detección de fraudes; por ejemplo una tienda envía información a un centro externo que determina si la operación parece correcta o fraudulenta.

2. Mediante la compra de los modelos:

Parece lógico que si alguien vende los modelos ya hechos se va a ahorrar tiempo y dinero. Además, parece que va a ser mucho más eficiente. Por ejemplo, podemos comprar un modelo de fuga de los clientes y utilizarlo cuando deseemos. Los modelos se pueden comprar con mantenimiento, en paquetes y con actualizaciones periódicas, con lo que se evita el problema de que se queden obsoletos. Los sistemas expertos o las bases de conocimiento que se compran a entidades externas siguen esta filosofía.

Implantación de un proyecto de datamining

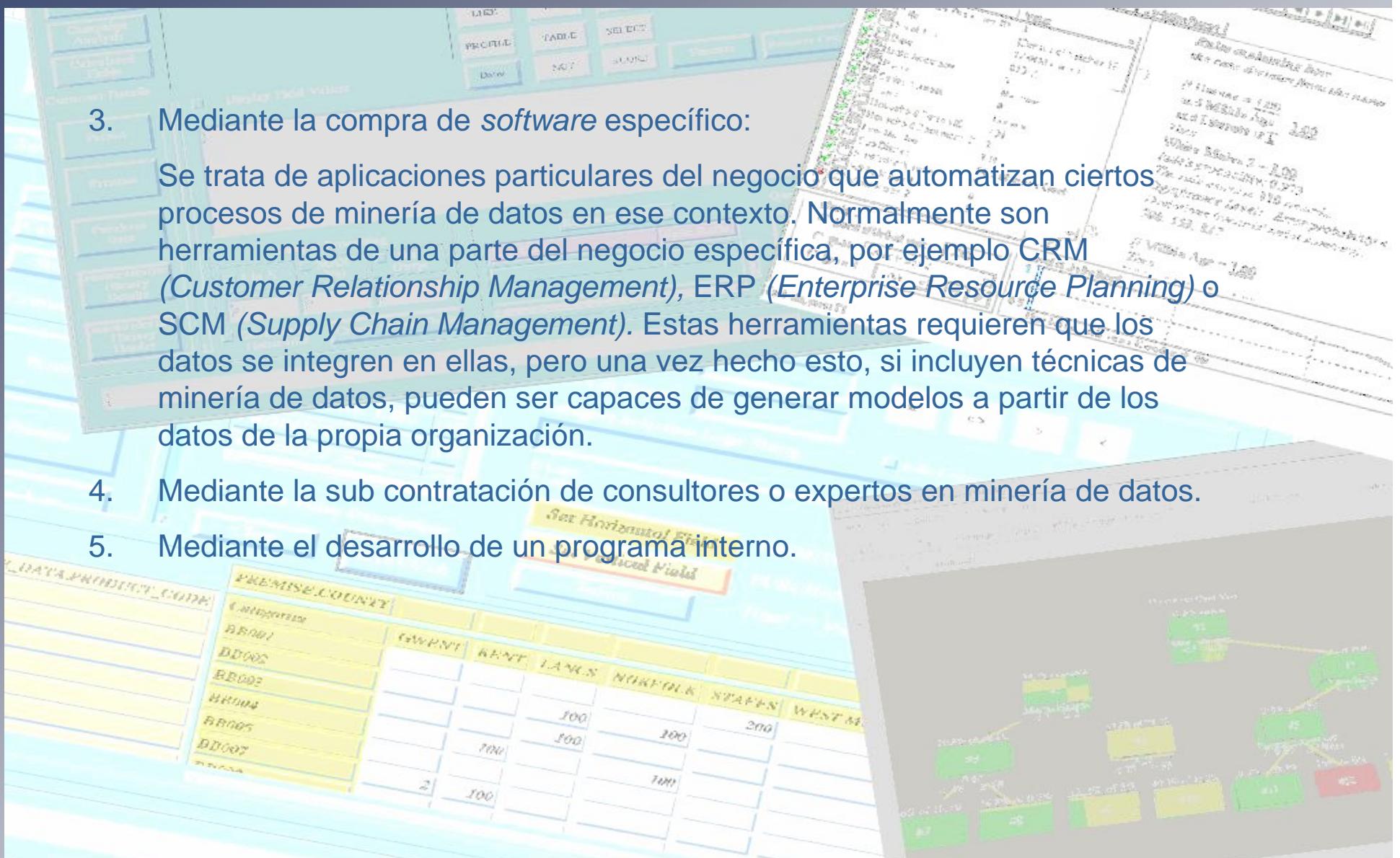
Celia Gutiérrez Cossío
2007

3. Mediante la compra de software específico:

Se trata de aplicaciones particulares del negocio que automatizan ciertos procesos de minería de datos en ese contexto. Normalmente son herramientas de una parte del negocio específica, por ejemplo CRM (*Customer Relationship Management*), ERP (*Enterprise Resource Planning*) o SCM (*Supply Chain Management*). Estas herramientas requieren que los datos se integren en ellas, pero una vez hecho esto, si incluyen técnicas de minería de datos, pueden ser capaces de generar modelos a partir de los datos de la propia organización.

4. Mediante la sub contratación de consultores o expertos en minería de datos.

5. Mediante el desarrollo de un programa interno.



Implantación de un proyecto de datamining

Celia Gutiérrez Cossío
2007

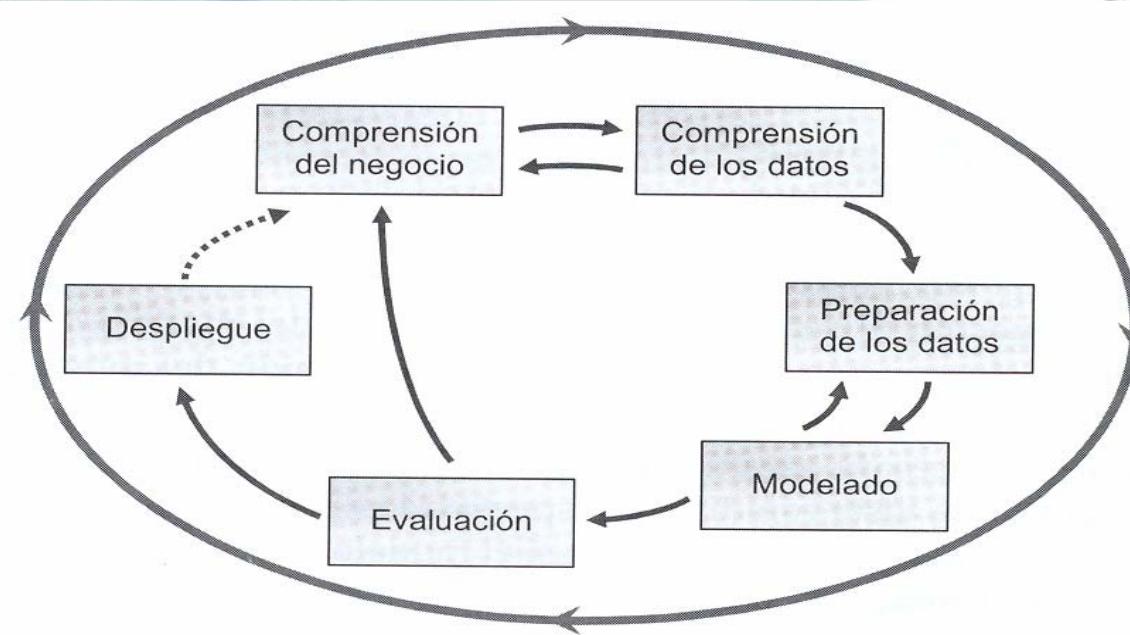
Desde la primera a la quinta solución se aumenta en coste inicial y complejidad pero también aumentan en flexibilidad, privacidad y potencialidad y, a largo plazo, en rentabilidad.

Además, existe un criticismo importante acerca de la contratación a una firma externa o a una consultora de aspectos de minería de datos. Si la "inteligencia de negocio" se relega a una empresa externa, podemos tener como resultado que el conocimiento extraído y las ventajas competitivas pueden, de alguna manera, ser "reutilizadas" por la empresa contratada para la competencia (pasando a suministrar puntuaciones o modelos nuestros a otras empresas sin nosotros saberlo). En teoría esto debería impedirse legalmente en los contratos, pero es muy difícil impedir no usar.

Una táctica bastante inteligente es subcontratar inicialmente y aprender de los consultores para, posteriormente, ir independizándose formando un equipo propio, es decir, ir avanzando en el nivel de autosuficiencia. Esta solución tiene el riesgo de que normalmente la empresa subcontratada intenta seguir siendo necesaria, un fenómeno que es muy usual con las consultoras: al resolver el problema resulta que ha aparecido otro que sólo ellos saben resolver.

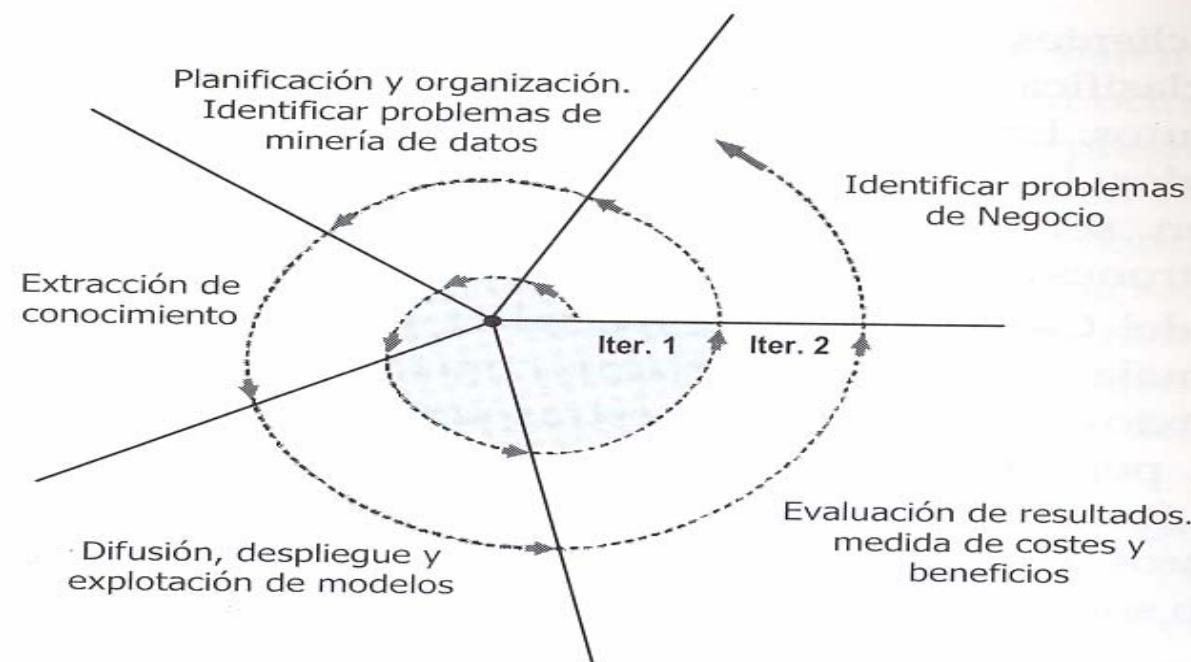
Una solución menos arriesgada es pedir asesoramiento externo para la creación del grupo de minería de datos o para el arranque, pero sin que todo sea realizado externamente. Además, en este caso, se puede recurrir a expertos en universidades u otros centros de investigación que, por regla general, darán un servicio más desinteresado y desearán que la organización sea autosuficiente cuanto antes, al contrario de una consultora profesional, que, en cierto modo, basa su negocio en que la organización no sea autosuficiente.

- El modelo y guía de referencia CRISP-DM
 1. CRISP-DM: consorcio de empresas cuyo estandar ha tenido una difusión amplísima debido a la independencia de la plataforma.
 2. El estandar 1.0 incluye un modelo de referencia y una guía para implantar un proyecto de datamining. Es útil como referencia incluso para adaptarlo a las necesidades de cada organización.



➤ Ciclo de vida en espiral

1. La primera implantación debe marcarse unos objetivos concretos y beneficios manifiestos.
2. Las dificultades de esta etapa serán superadas gracias a una elección clara de los problemas y de las herramientas para solucionarlos.
3. Del primer proyecto y de lo aprendido nos podemos plantear objetivos más ambiciosos, incluso un datamining no dirigido.
4. ¿Cuánto tiempo debe durar el primer ciclo?. No más de 6 meses.



➤ Impacto social de la minería de datos

1. Término muy **popular** en los últimos años:

Cada vez son más los usuarios, las aplicaciones, las investigaciones y los desarrollos relacionados con ella, y crecen los sistemas *software* que afirman ser productos de minería de datos.

2. Tecnología ampliamente **reconocida** por compañías de todo tipo:

El uso de información aprendida desde los datos es necesario para mantener la competitividad en todos los entornos empresariales, así como optimizar las decisiones de las instituciones públicas para dar un mejor servicio a los ciudadanos. Los almacenes de datos, que han hecho posible el almacenamiento de grandes volúmenes de datos en un mismo repositorio, junto con el incremento en potencia de la computación, son las causas de que las empresas de hoy en día busquen herramientas y tecnologías capaces de extraer información útil de los datos.

3. Nivel **real** de implantación y arraigo de la minería de datos en la sociedad:

Es subjetivo y muy difícil de determinar, y lógicamente depende de países y zonas geográficas. Grandes empresas y organizaciones, así como los gobiernos, están cambiando la perspectiva de un análisis de datos más tradicional, más descriptivo y confirmatorio, a una minería de datos más orientada al sistema de información, a la búsqueda de modelos y patrones, y, sobre todo, una visión más en el fin (sacar partido de la información que nos rodea) que en el medio (este medio se encuentra, justamente, en la "minería de datos"). Si bien las grandes organizaciones incorporan o incluso sustituyen terminologías, métodos y metodologías en los departamentos de estadística e investigación operativa, prospectiva y similar, para las pequeñas y medianas empresas la minería de datos representa su primera oportunidad de entrar en el mundo del análisis de la información. No sólo existen herramientas de minería de datos cada vez más completas y accesibles (o incluso gratuitas), sino que los sistemas de gestión de bases de datos están incluyendo primitivas, lenguajes e incluso entornos de minería de datos. Hasta las aplicaciones ofimáticas (hojas de cálculo) comienzan a incluir este tipo de herramientas.

4. Difusión mediante la web:

Cada vez más compañías llevan a cabo sus actividades a través de la web, sobre todo la relación con los clientes o entre proveedores. Los datos recogidos (patrones de compra y de navegación) pueden proporcionar mucha más información sobre los clientes tanto individualmente como en grupo, y esta información puede ayudar a las empresas a ofrecer unos servicios más personalizados y adaptados a las características de los clientes.

Servir a las necesidades de los clientes puede significar un ahorro económico sustancial para las empresas (por ejemplo, evitando hacer campañas publicitarias generales), y un beneficio para los clientes que comprueban satisfechos cómo les ofrecen productos en los que están interesados y no pierden su tiempo en digerir ofertas por las que no están interesados.

5. Otros ámbitos:

- Medicina, bioinformática.
- Seguridad o la lucha antiterrorista, campo en el que se está trabajando profusamente.

6. A nivel individual:

Por ejemplo, la mayoría de navegadores de última generación incluyen métodos de aprendizaje automático (generalmente bayesianos) para clasificar el correo electrónico y detectar los mensajes *spam*.

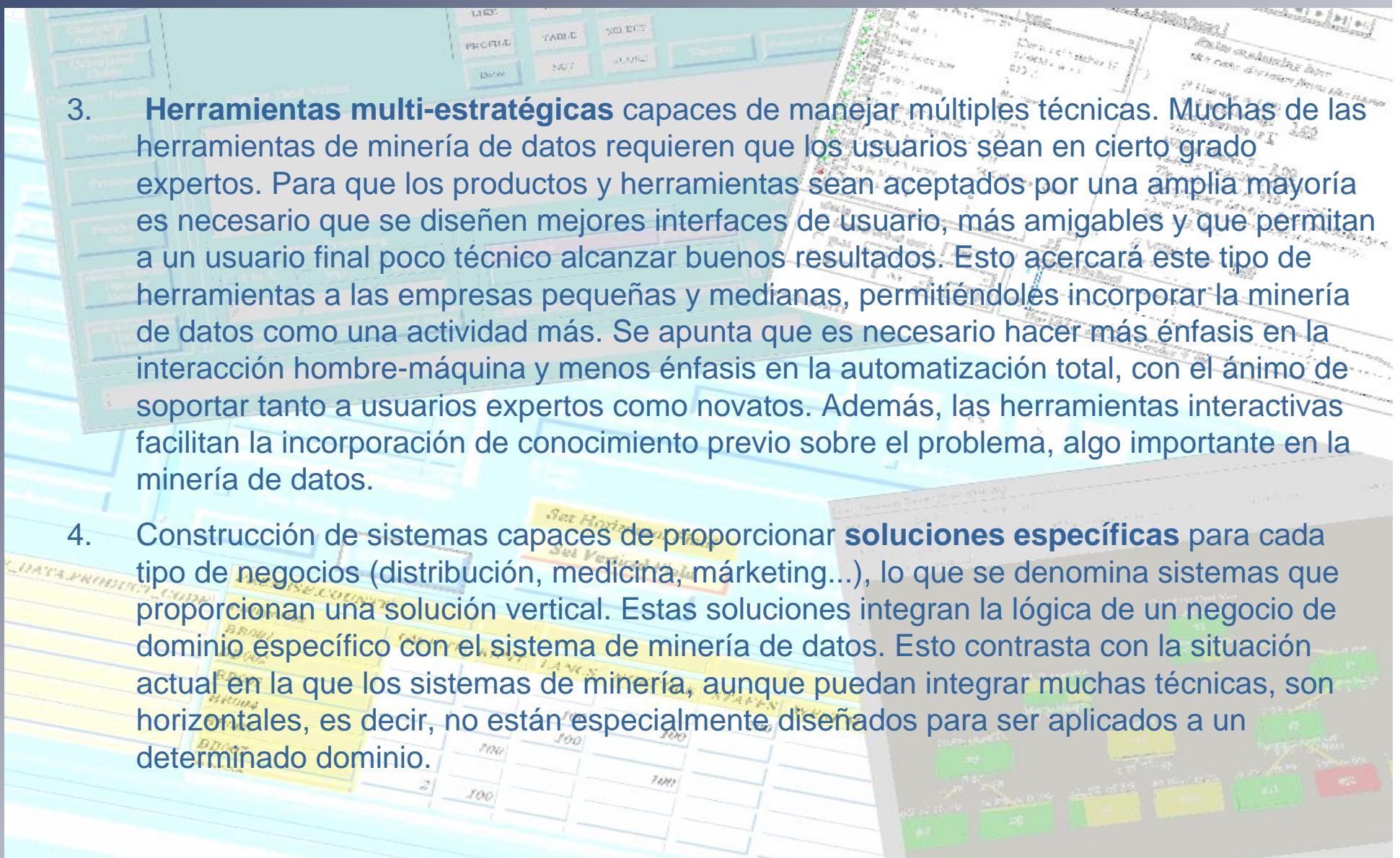
➤ Tendencias futuras

La minería de datos es el resultado de la integración de múltiples técnicas => los retos que se plantean han de resolverse por avances en estas disciplinas pero, fundamentalmente, por la combinación de estas disciplinas.

1. La materia prima de la minería de datos son datos => disponer de buenos datos es clave para esta disciplina ya que la **calidad del conocimiento** extraído depende tanto o más de los datos usados que de la técnica empleada. Muchos de los datos que se recopilan son imprecisos, incompletos o inciertos => Técnicas ETL del datawarehouse.
2. Trabajar de forma eficiente y efectiva con **grandes bases de datos**: los conjuntos de datos masivos y con una alta dimensionalidad crean espacios de búsqueda combinatoriamente explosivos e incrementan la probabilidad de que el algoritmo de minería de datos requiera un tiempo excesivo y además encuentre patrones no válidos. La escalabilidad de las técnicas requiere un trabajo considerable tanto en los fundamentos teóricos como en las pruebas con conjuntos de datos cada vez mayores. Una buena gestión del procesamiento entre memoria y disco, el uso de índices específicos para la minería de datos y de compactación, puede ser crucial para obtener esta eficiencia.

Repercusiones y retos del Datamining

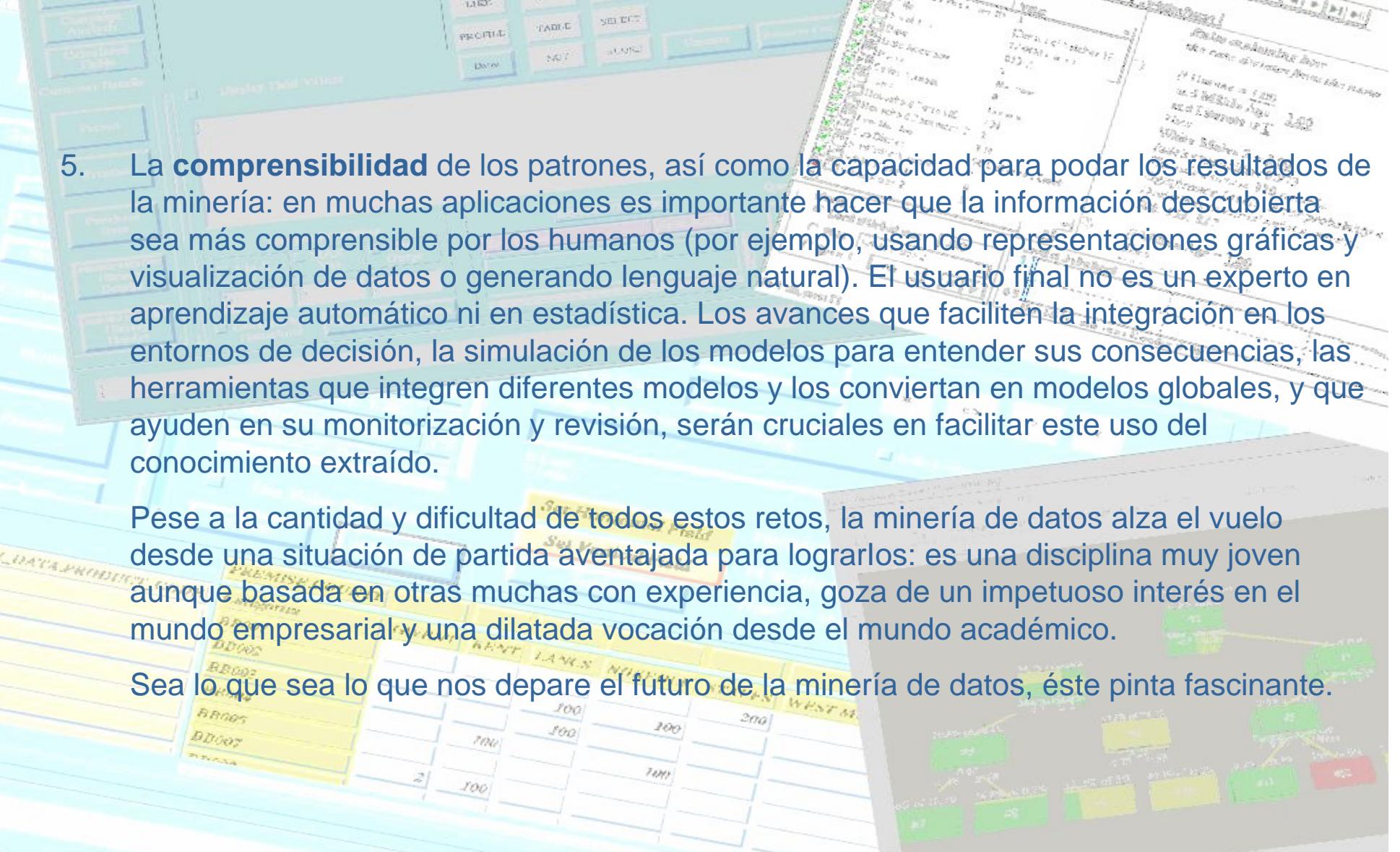
Celia Gutiérrez Cossío
2007



3. **Herramientas multi-estratégicas** capaces de manejar múltiples técnicas. Muchas de las herramientas de minería de datos requieren que los usuarios sean en cierto grado expertos. Para que los productos y herramientas sean aceptados por una amplia mayoría es necesario que se diseñen mejores interfaces de usuario, más amigables y que permitan a un usuario final poco técnico alcanzar buenos resultados. Esto acercará este tipo de herramientas a las empresas pequeñas y medianas, permitiéndoles incorporar la minería de datos como una actividad más. Se apunta que es necesario hacer más énfasis en la interacción hombre-máquina y menos énfasis en la automatización total, con el ánimo de soportar tanto a usuarios expertos como novatos. Además, las herramientas interactivas facilitan la incorporación de conocimiento previo sobre el problema, algo importante en la minería de datos.
4. Construcción de sistemas capaces de proporcionar **soluciones específicas** para cada tipo de negocios (distribución, medicina, marketing...), lo que se denomina sistemas que proporcionan una solución vertical. Estas soluciones integran la lógica de un negocio de dominio específico con el sistema de minería de datos. Esto contrasta con la situación actual en la que los sistemas de minería, aunque puedan integrar muchas técnicas, son horizontales, es decir, no están especialmente diseñados para ser aplicados a un determinado dominio.

Repercusiones y retos del Datamining

Celia Gutiérrez Cossío
2007

- 
5. La **comprendibilidad** de los patrones, así como la capacidad para podar los resultados de la minería: en muchas aplicaciones es importante hacer que la información descubierta sea más comprensible por los humanos (por ejemplo, usando representaciones gráficas y visualización de datos o generando lenguaje natural). El usuario final no es un experto en aprendizaje automático ni en estadística. Los avances que faciliten la integración en los entornos de decisión, la simulación de los modelos para entender sus consecuencias, las herramientas que integren diferentes modelos y los conviertan en modelos globales, y que ayuden en su monitorización y revisión, serán cruciales en facilitar este uso del conocimiento extraído.

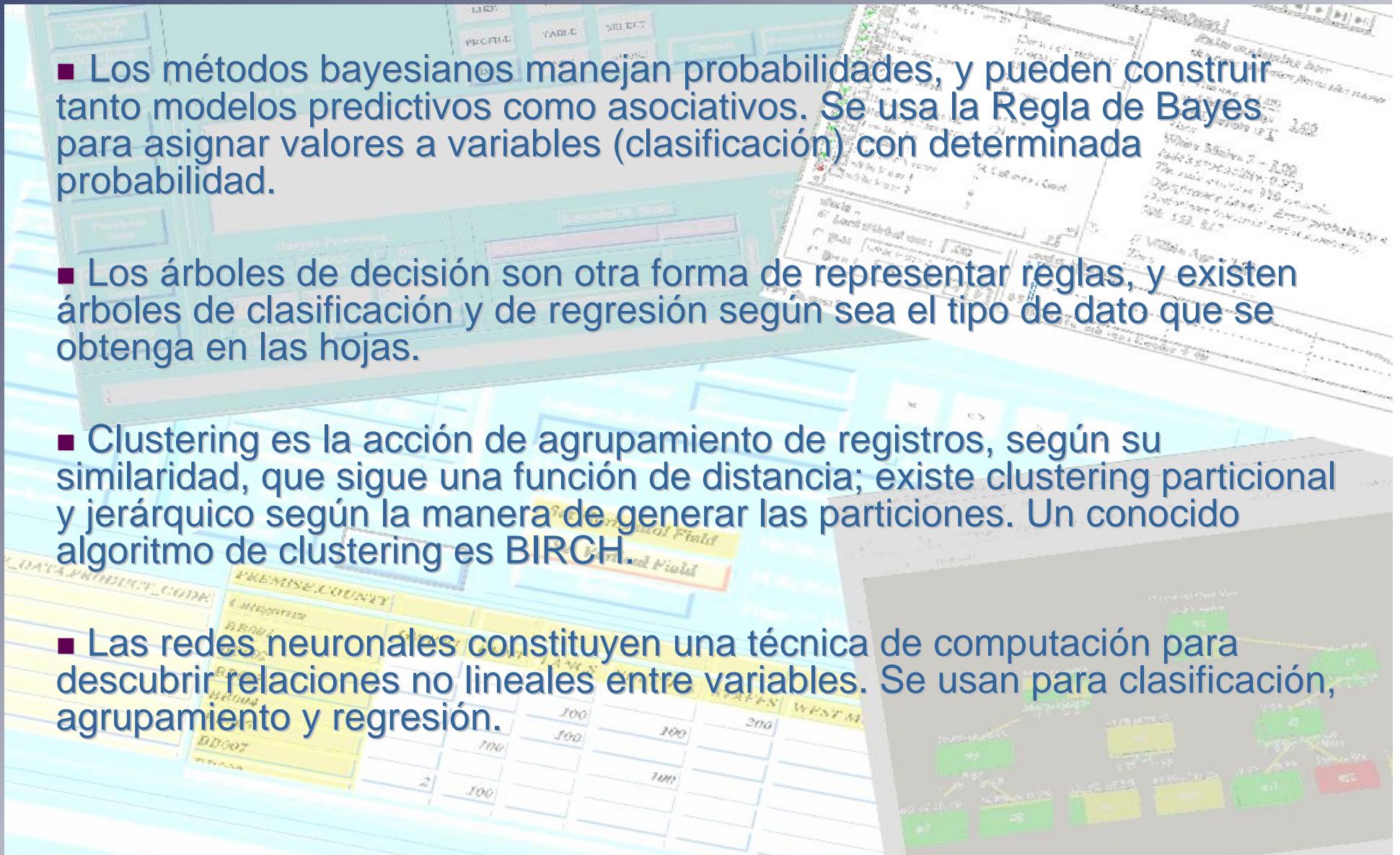
Pese a la cantidad y dificultad de todos estos retos, la minería de datos alza el vuelo desde una situación de partida aventajada para lograrlos: es una disciplina muy joven aunque basada en otras muchas con experiencia, goza de un impetuoso interés en el mundo empresarial y una dilatada vocación desde el mundo académico.

Sea lo que sea lo que nos depare el futuro de la minería de datos, éste pinta fascinante.

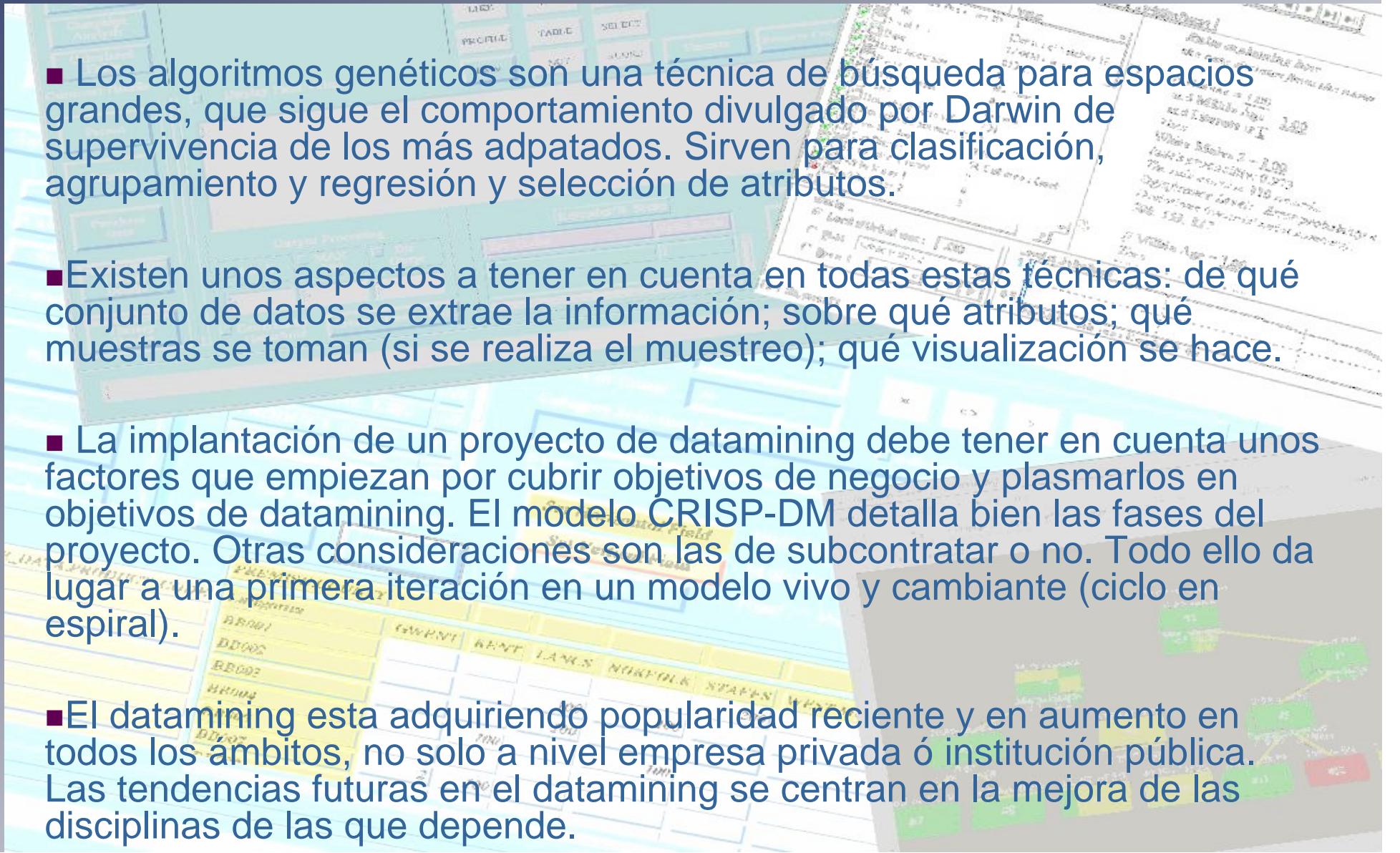
- El producto de minería de datos de Oracle se llama **Oracle Data Mining (ODM)**.
- Esta integrado, junto a Oracle OLAP dentro de la base de datos **Oracle9i**.
- Contiene **API de Java**:
 1. Modelado potente y escalable que permite hacer predicciones en tiempo real.
 2. Integración completa con aplicaciones e-Business.
- **Múltiples algoritmos**: Bayes, árboles, agrupamientos y reglas de asociación.
- **Fases del entorno de ODM**:
 1. Preparación de datos: creación de nuevas tablas o vistas => más rápido que transferirlos hacia la utilidad de minería de datos externa.
 2. Construcción del modelo: el soporte de múltiples algoritmos de predicción cubre la resolución de amplia variedad de problemas. Dentro de la base de datos => más rápido.
 3. Evaluación del modelo: modelos almacenados en la base de datos => evaluación, generación de informes, análisis posterior.
 4. Puntuación: en tiempo real y por lotes => clasificación ó probabilidad de que se produzca un resultado específico.

- La técnica del datamining ó minería de datos trata la información para descubrir patrones ó modelos desconocidos que ayudan a la toma de decisiones.
- Es fuertemente dependiente de los almacenes de datos, y por tanto, el diseño de estos últimos depende (entre otras cosas) de su uso en un datamining. El éxito de un datamining depende mucho del datawarehouse del que proceden sus datos.
- Tiene ventajas sobre otras herramientas de toma de decisiones en que aporta conocimiento desconocido.
- Existen modelos predictivos y descriptivos según su función. Para cada uno de ellos se engloban diversas técnicas de datamining.
- El proceso de datamining se engloba dentro del KDD, ó descubrimiento del conocimiento de bases de datos, cuyo objetivo final es la obtención de conocimiento; para ello se relaciona con otras disciplinas.

- Su aplicación es multidisciplinar, y abarca desde áreas directamente lucrativas (como Seguros) hasta otras como la Medicina ó deportes, pasando por el correo electrónico.
- La importancia del datamining radica en los múltiples modelos que puede generar basados en variadas técnicas.
- Las reglas de asociación tienen variantes (como las aplicadas a jerarquías ó de selección de un atributo de agrupamiento) y se basan en la obtención de una relación del tipo: "Si A toma el valor a, entonces B toma el valor b", con unos umbrales que la garanticen, llamados soporte y confianza. Sin embargo, dichas relaciones no tienen por qué ser causales. Existen diversos algoritmos para aplicación de estas reglas y sus variaciones, el más importante de ellos es el a-priori.
- Los patrones secuenciales indican comportamientos similares que se han desarrollado repetidamente; para el caso de compra de artículos, se puede explicar como series de artículos que se adquieren con un mínimo de soporte. Los algoritmos que implementan esto se parecen a los anteriores, el más conocido es el a-priori all.



- Los métodos bayesianos manejan probabilidades, y pueden construir tanto modelos predictivos como asociativos. Se usa la Regla de Bayes para asignar valores a variables (clasificación) con determinada probabilidad.
- Los árboles de decisión son otra forma de representar reglas, y existen árboles de clasificación y de regresión según sea el tipo de dato que se obtenga en las hojas.
- Clustering es la acción de agrupamiento de registros, según su similaridad, que sigue una función de distancia; existe clustering particional y jerárquico según la manera de generar las particiones. Un conocido algoritmo de clustering es BIRCH.
- Las redes neuronales constituyen una técnica de computación para descubrir relaciones no lineales entre variables. Se usan para clasificación, agrupamiento y regresión.

- 
- Los algoritmos genéticos son una técnica de búsqueda para espacios grandes, que sigue el comportamiento divulgado por Darwin de supervivencia de los más adaptados. Sirven para clasificación, agrupamiento y regresión y selección de atributos.
 - Existen unos aspectos a tener en cuenta en todas estas técnicas: de qué conjunto de datos se extrae la información; sobre qué atributos; qué muestras se toman (si se realiza el muestreo); qué visualización se hace.
 - La implantación de un proyecto de datamining debe tener en cuenta unos factores que empiezan por cubrir objetivos de negocio y plasmarlos en objetivos de datamining. El modelo CRISP-DM detalla bien las fases del proyecto. Otras consideraciones son las de subcontratar o no. Todo ello da lugar a una primera iteración en un modelo vivo y cambiante (ciclo en espiral).
 - El datamining está adquiriendo popularidad reciente y en aumento en todos los ámbitos, no solo a nivel empresa privada ó institución pública. Las tendencias futuras en el datamining se centran en la mejora de las disciplinas de las que depende.

- Oracle Data Mining es el producto de datamining de Oracle. Esta integrado en la base de datos Oracle9i, y presenta otras ventajas como API de Java y múltiples técnicas de modelado.



- “Database Management Systems”, R. Ramakrishnan (tema 24), Ed. Mc Graw Hill.
- “Sistemas de Bases de Datos”, T. Connolly, C. Begg (tema 34), Ed. Pearson-Addison Wesley.
- “Introducción a la Minería de Datos”, J. Hernandez, M.J. Quintana, C. Ferri (temas 1,2,9,22,23), Ed. Pearson-Addison Wesley.
- “Fundamentos de Sistemas de Bases de Datos”, R. A. Elmasri, S. B. Navathe (tema 26.1), Ed. Addison-Wesley.

- Ejercicio 1 (4 puntos): algoritmo a-priori
 1. Obtener conjuntos de ítems frecuentes con **mínimo de soporte**. El **soporte debe ser un parámetro del algoritmo** (2 puntos)
 2. Derivado de lo anterior, obtener reglas de asociación con un **mínimo de confianza**. La **confianza debe ser un parámetro del algoritmo**. (2 puntos)
Probarlo con el ejercicio de clase.

- Ejercicio 2 (6 puntos): algoritmo a-priori-all
Implementar el algoritmo a-priori all para extraer patrones secuenciales. El **soporte ó cobertura debe ser un parámetro del algoritmo**. (1 punto cada apartado del algoritmo).
Probarlo con el ejercicio de clase.

- Formato de la práctica
No hay que hacer ninguna presentación, sino que con la herramientas disponibles, se tiene que hacer una demostración al resto de la clase de cómo se han realizado los requisitos y de que realmente funcionan.
- Valor de la práctica en el segundo parcial: 50%
Hace falta un 4 para continuar en la evaluación continua.
- Fecha entrega de la práctica: antes de la exposición del 16-05-2007
- Fecha de presentación de la práctica: 16-05-2007, 17-05-2007