

Minería de Datos



Vallejos, Sofia

Contenido



- Introducción:
 - Inteligencia de negocios (Business Intelligence).
 - Componentes
 - Descubrimiento de conocimiento en bases de datos (KDD).
- Minería de Datos:
 - Perspectiva histórica.
 - Fases de un Proyecto.
 - Fuentes de datos.
 - Funciones de minería.
 - Modelos típicos de minería.
- Ejemplos:
 - Clustering.
 - Asociación.
 - Red neuronal como modelo predictivo.
- Web Mining.
- Conclusiones.

Inteligencia de Negocios



Hace referencia a un conjunto de productos y servicios para acceder a los datos, analizarlos y convertirlos en información.

“ Es un paraguas bajo el que se incluye un conjunto de conceptos y metodologías cuya misión consiste en mejorar el proceso de toma de decisiones en los negocios basándose en hechos y sistemas que trabajan con hechos.”

Howard Dresner
Gartner Group, 1989.

Inteligencia de Negocios Componentes



- Multidimensionalidad.
- Agentes.
- Data Warehouse.
- Data Mining.

Descubrimiento de Conocimiento en Bases de Datos



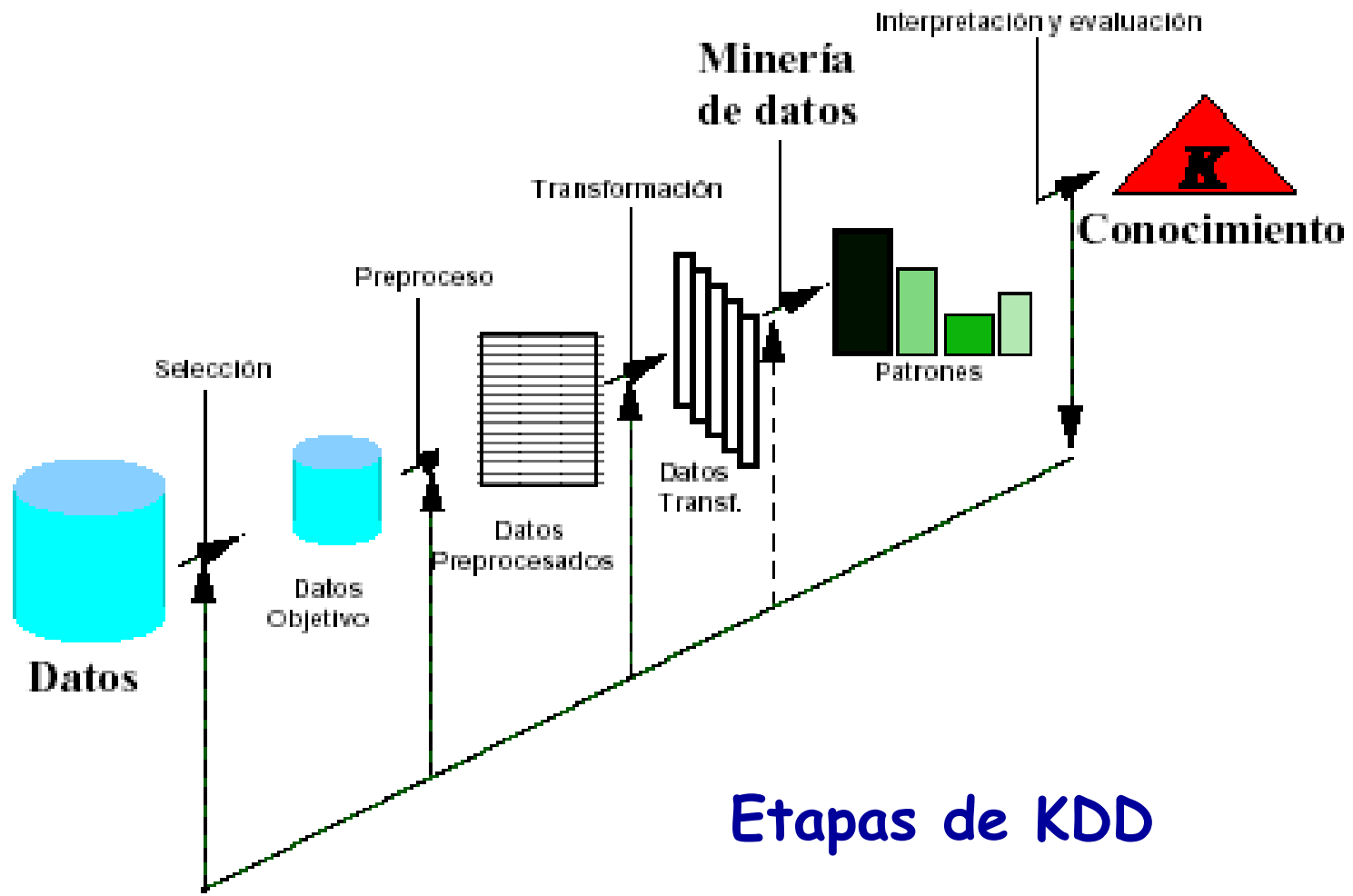
Es un proceso de extracción no trivial para identificar patrones que sean **válidos**, **novedosos**, potencialmente **útiles** y **entendibles**, a partir de los datos.

Su objetivo principal: procesar automáticamente grandes cantidades de datos para encontrar **conocimiento útil** para un usuario y **satisfacer sus metas**.

Descubrimiento de Conocimiento en Bases de Datos Jerarquía



Descubrimiento de Conocimiento en Bases de Datos



Qué es Minería de Datos



- Es el proceso de exploración y análisis - de manera automática o semiautomática - de los datos para obtener patrones significativos y reglas de negocio.
- Consideraciones:
 - Los patrones deben ser **significativos**.
 - Sin automatización es imposible mirar grandes cantidades de datos, pero se debe dar más énfasis a las etapas de **exploración y análisis**, que al modo de exploración.
 - Data Mining es un **proceso**.

Qué es Minería de Datos



➤ La MD puede ser dividida en:

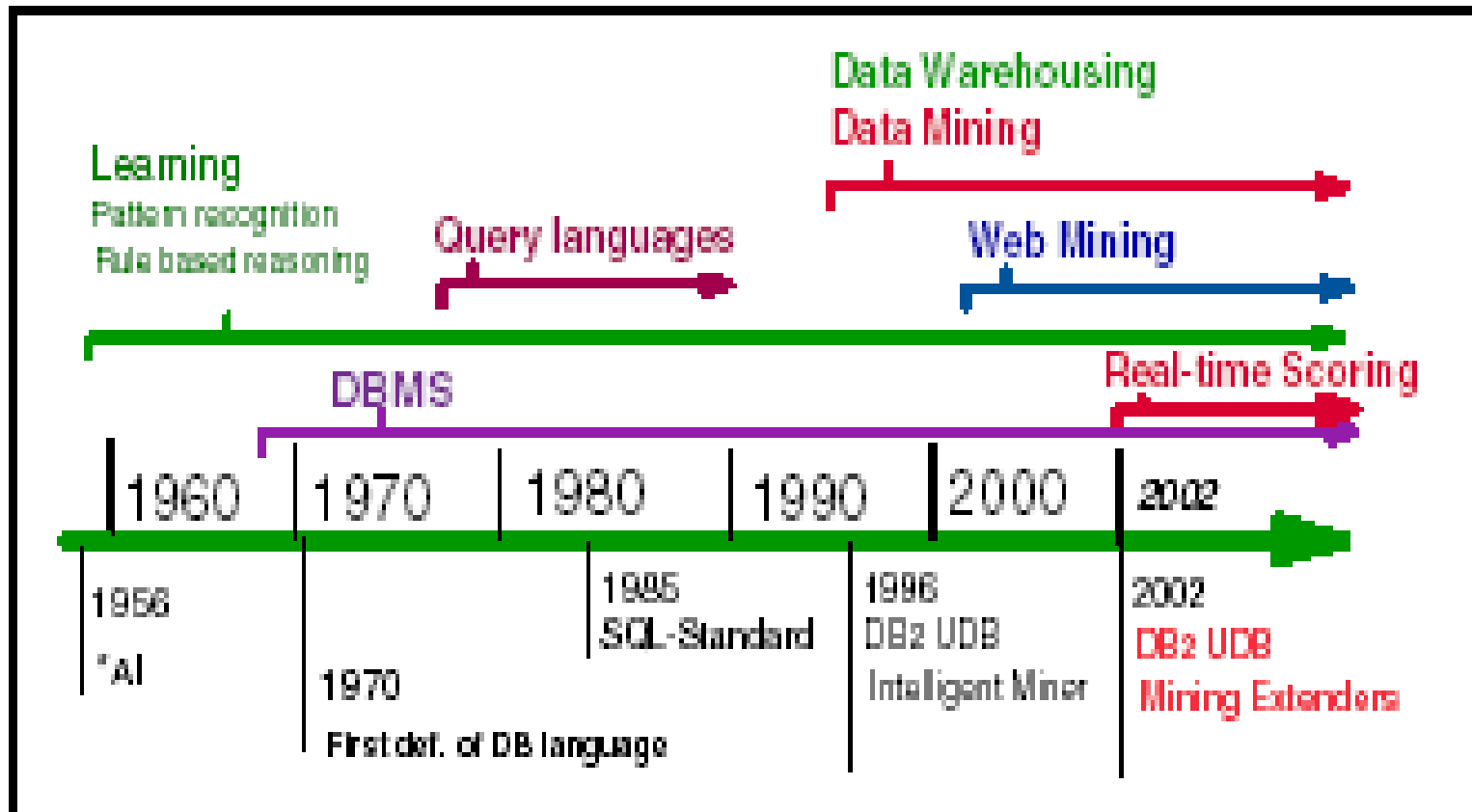
- Minería de datos predictiva (mdp): usa primordialmente técnicas estadísticas.
- Minería de datos para el descubrimiento de conocimiento (mddc): usa principalmente técnicas de inteligencia artificial.

Qué no es Minería de Datos



- No es un **producto** que se compra enlatado sino una **disciplina** que debe ser dominada.
- No es una **solución instantánea** a los problemas de negocio.
- No es un **fin en sí mismo**, sino un **proceso** que ayuda a encontrar soluciones a problemas de negocio.

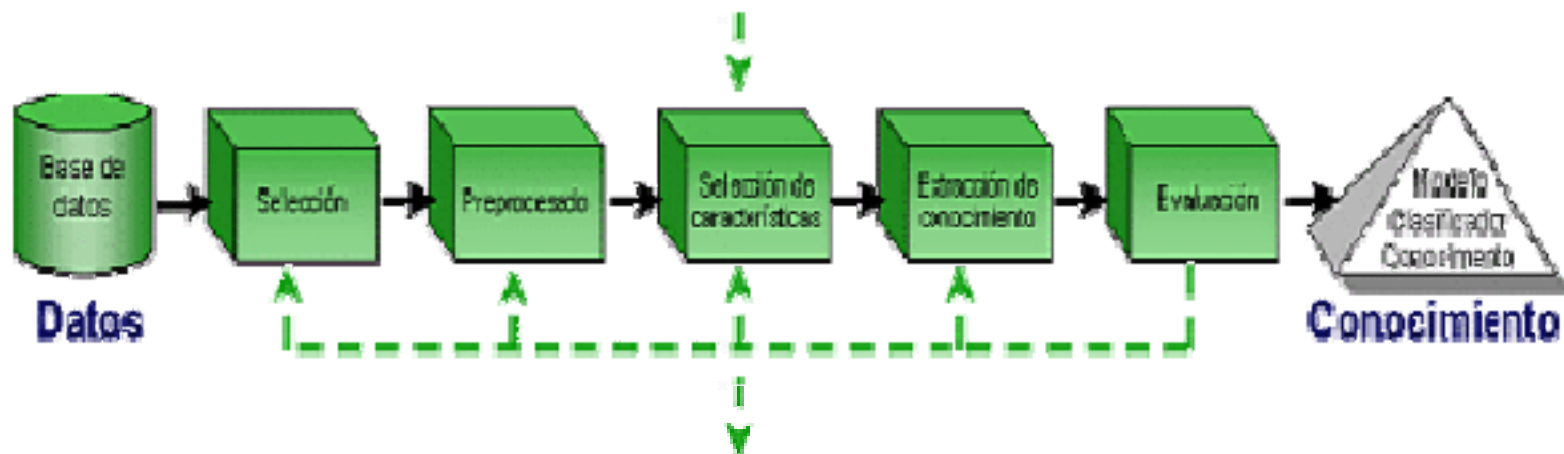
Minería de Datos: Perspectiva histórica



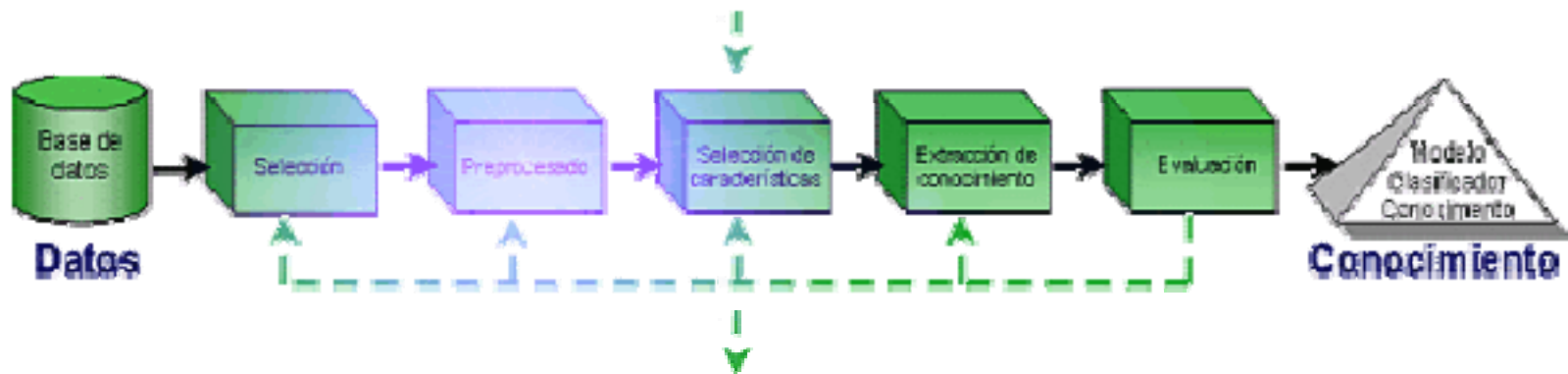
Fases de un Proyecto de Minería de Datos



- El proceso de minería de datos pasa por las siguientes fases:
- Filtrado de datos.
 - Selección de Variables.
 - Extracción de Conocimiento.
 - Interpretación y Evaluación.

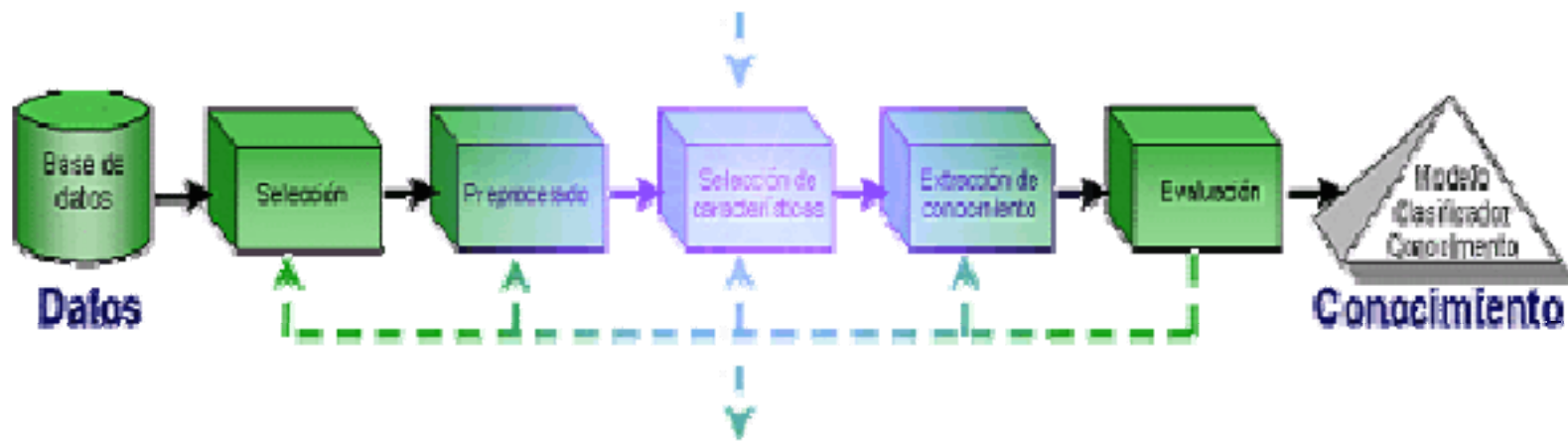


Fases de un Proyecto de DM: Filtrado de datos



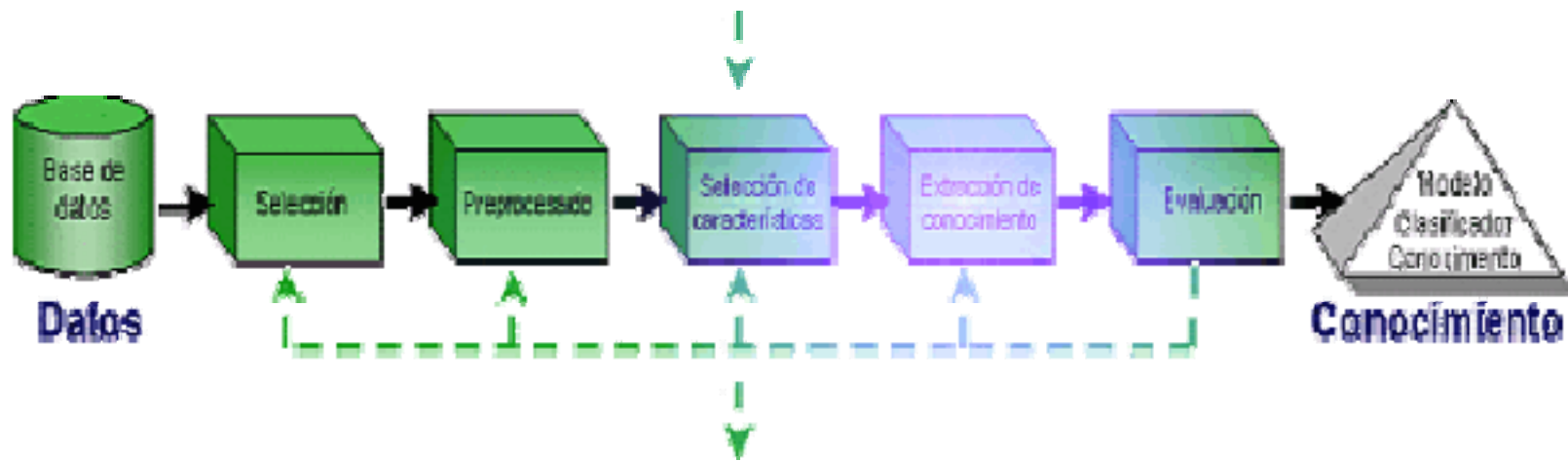
- Mediante el preprocesado, se filtran los datos
 - Se eliminan valores incorrectos, no válidos, desconocidos... según las necesidades y el algoritmo a usar).
 - Se obtienen muestras de los mismos (en busca de una mayor velocidad de respuesta del proceso).
 - Se reducen el número de valores posibles (mediante redondeo, clustering,...).

Fases de un Proyecto de DM: Selección de Variables



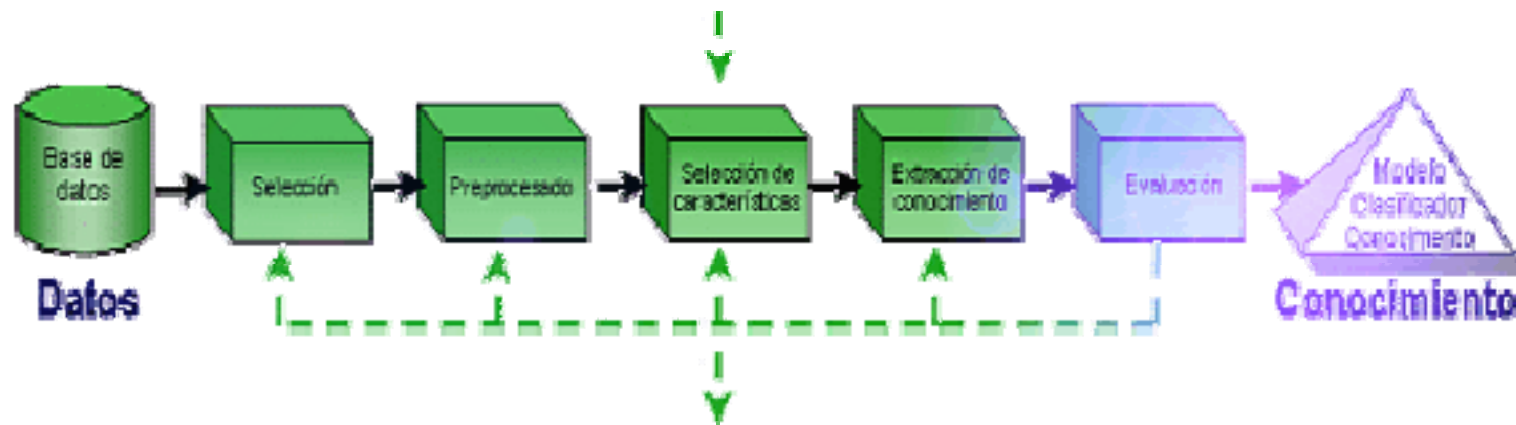
- Los métodos para la selección de características son básicamente dos:
 - Aquellos basados en la elección de los mejores atributos del problema.
 - Y aquellos que buscan variables independientes mediante tests de sensibilidad, algoritmos de distancia o heurísticos.

Fases de un Proyecto de DM: Extracción de Conocimiento



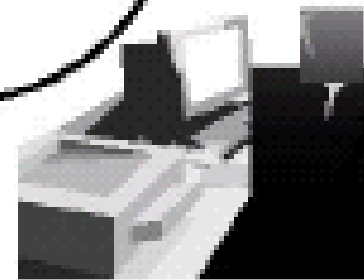
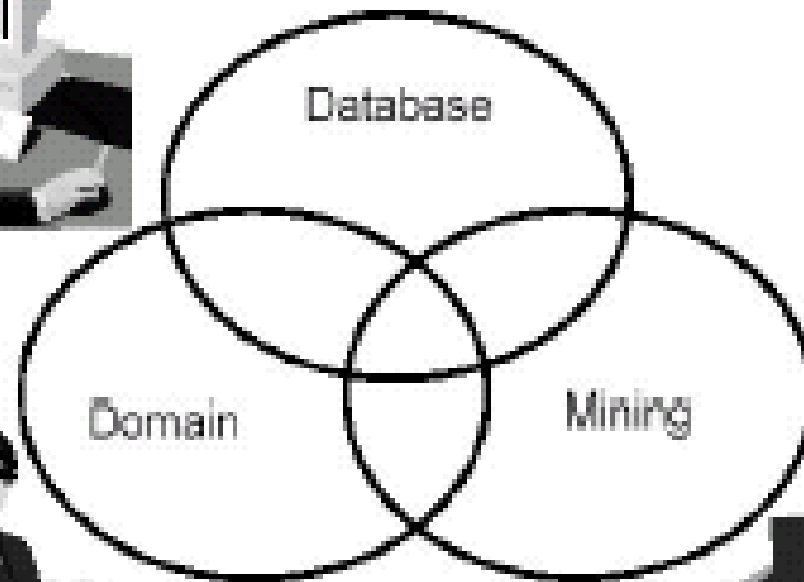
- Mediante una técnica de minería de datos:
 - Se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables.

Fases de un Proyecto de DM: Interpretación y Evaluación

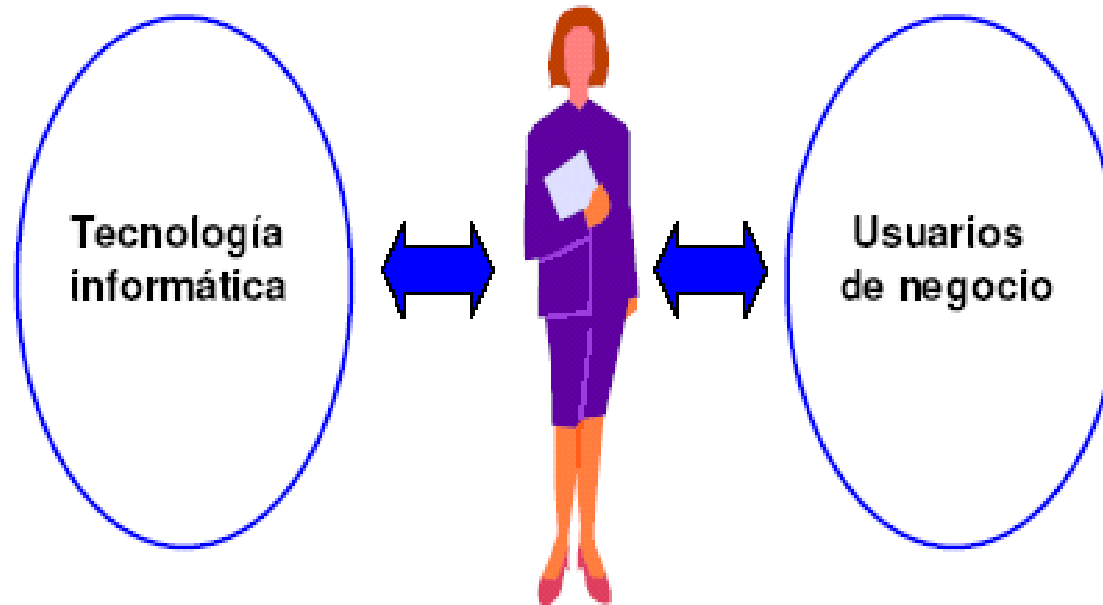


- Se debe proceder a su validación, comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias.
- Si ninguno de los modelos alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos modelos.

Integrantes del proyecto

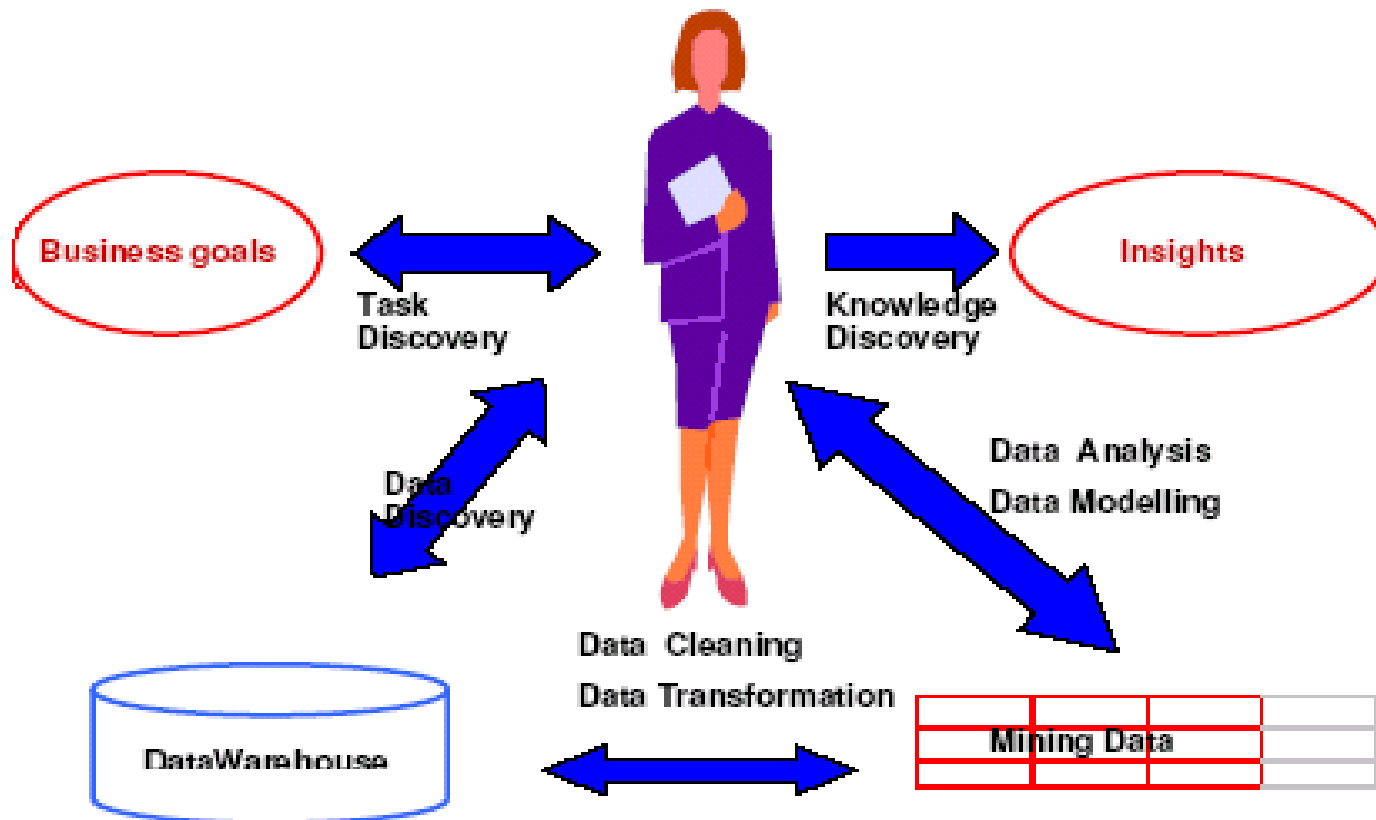


El analista de datos



- Es el vínculo entre las áreas de tecnología informática y las áreas de negocio.
- Habilidades requeridas:
 - Manipulación de datos (SQL).
 - Conocimiento de técnicas de minería y análisis exploratorio.
 - Habilidad de comunicación (interpretación) de los problemas de negocio.

El analista de datos



- Traduce los requerimientos de información en preguntas apropiadas para su análisis con las herramientas de minería.

Fuentes de Datos



➤ Tipos de fuentes:

- Transaccionales: Ej. operaciones realizadas con una tarjeta de crédito.
- Relacionales: Ej. estructura de productos que ofrece un banco.
- Demográficos: Ej. características del grupo familiar.

➤ Origen de datos:

- Bases de datos relacionales.
- DataWarehouses.
- Data Marts.
- Otros formatos: Excel, Access, encuestas, archivos planos.

Calidad de los Datos



- El éxito de las actividades de Data Mining se relaciona directamente con la CALIDAD de los datos.
- Muchas veces resulta necesario pre-procesar los datos, antes de derivarlos al modelo de análisis.
- El preproceso puede incluir transformaciones, reducciones o combinaciones de los datos.
- La semántica de los datos debe ayudar para seleccionar una conveniente representación, dado que influye directamente sobre la calidad del modelo.

Funciones de minería



- Utilizan técnicas matemáticas elaboradas para descubrir patrones ocultos en los datos. Ellas son:
 - Asociación.
 - Clasificación neuronal.
 - Clasificación en árbol.
 - Clustering demográfico.
 - Clustering neuronal.
 - Patrones secuenciales.
 - Secuencias semejantes.
 - Predicción neuronal.
 - Predicción - función base radial.

Modelos típicos de minería



- ✓ Clustering.
- ✓ Clasificación.
- ✓ Estimación.
- ✓ Predicción.
- ✓ Agrupamiento a partir de reglas de asociación.

Modelos típicos de minería:



Clustering

- Agrupar a los clientes según indicadores F (frecuencia), M (monto), etc en segmentos de comportamientos homogéneos.
- Resultado: Clientes **Buenos**, **Medios**, **Malos**.
- El 78% de la facturación se concentra en el cluster **Buenos**.
- Los clientes **Buenos** son casados, con hijos, trabajadores autónomos con ingreso superior a \$3000 pesos.

Modelos típicos de minería:

Clasificación y Estimación



- Clasificar un nuevo cliente - de acuerdo a su perfil sociodemográfico - como un cliente:
 - Bueno.
 - Medio.
 - Malo.
- Estimar el consumo de un determinado rubro de artículos de un grupo de clientes en el próximo trimestre.

Modelos típicos de minería: Predicción



- Predecir el abandono de un cliente:
 - Para una compañía de telefonía celular.
 - Para una AFJP.
 - Para una tarjeta de crédito.

Modelos típicos de minería:

Asociación



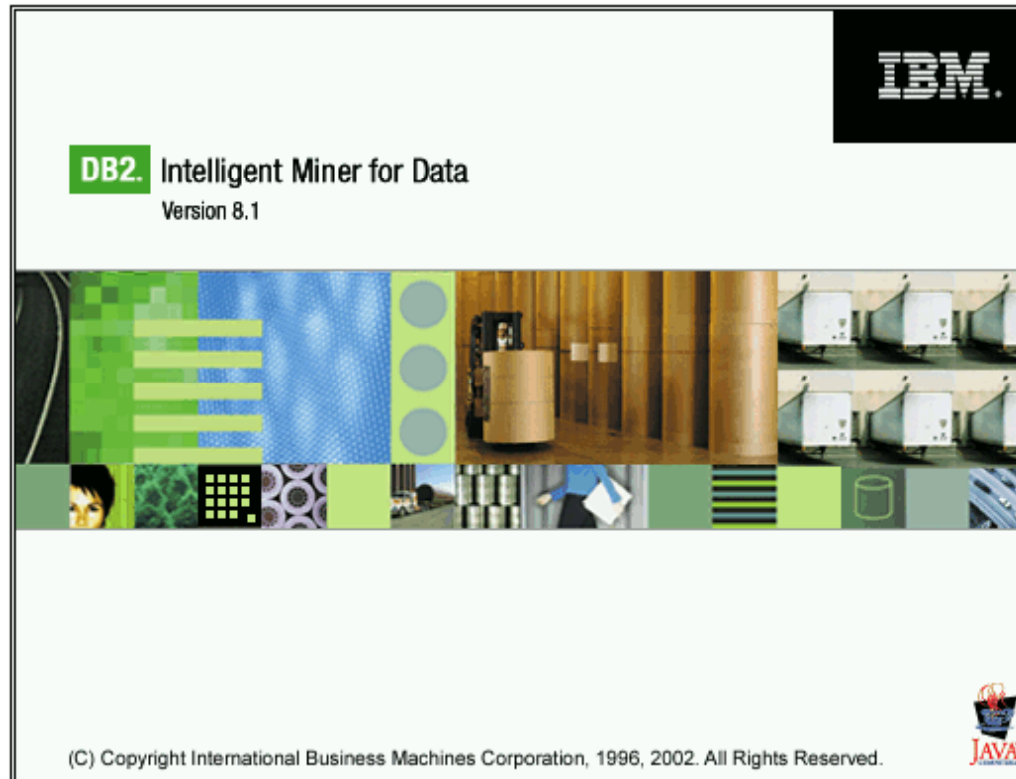
- Encontrar las reglas que determinan la interrelación entre productos para clientes de un banco. Por ejemplo:
 - “ Cuando un cliente se activa en Caja de Ahorros, el siguiente producto donde se activa es Préstamos Personales. Este patrón ocurre el 65 % de los casos. ”

Elección del modelo



- Principales objetivos del proceso de Data Mining:
 - Predicción.
 - Descripción.
- El método a utilizar depende de los objetivos perseguidos por el análisis pero también de la calidad y cantidad de los datos disponibles.

DB2-Intelligent Miner for Data



- ✓ Comprende un conjunto de funciones estadísticas, de proceso y de minería de datos.
- ✓ Ofrece herramientas de visualización.

Ejemplos con DB2 Intelligent Miner for Data



- ✓ Clustering.
- ✓ Asociación.
- ✓ Red neuronal como modelo predictivo.

Clustering



- Es la partición del conjunto de individuos en subconjuntos lo más homogéneos posibles.
- El objetivo es maximizar la **similitud** de individuos **del** cluster y maximizar las **diferencias entre** clusters.
- Se aplica para segmentación de bases de datos, identificación de tipos de clientes, etc.

Aportes del software de minería



- Determinar el número óptimo de clusters.
- Asignar a cada individuo a un único cluster.
- Evaluar el impacto de las variables en la formación del cluster.
- Comprender el "perfil" de cada cluster.

Ejemplo de Clustering



La gerencia comercial de un banco necesita identificar al segmento más valioso de clientes de una tarjeta de crédito para organizar sus gastos de promociones y campañas de marketing directo.

➤ Datos disponibles:

- Frecuencia de uso de la tarjeta.
- Saldo promedio mensual en \$.
- Posesión de tarjeta *Gold*.
- Monto promedio por cada transacción.
- Cantidad de servicios por débito automático.
- Datos sociodemográficos: sexo, edad, estado civil, ocupación, hijos.
- Fuente de datos: transacciones del último año, tabla de clientes.

Ejemplo de Clustering



➤ Preparación de los datos:

- Definir la unidad de análisis: ¿cuenta o tarjeta?
- Definir qué es una transacción: ¿cómo se consideran los ajustes?
- Describir las variables a incluir en el modelo.

➤ Tabla de datos:

avgtckt	cuenta	edad	estado_civil	frecu	gold	hijos	ocup	pesos	scios	sexo
10	" 1"	43	"2"	10	'1'	'0'	'2'	33511	4	'1'
1587	" 2"	43	"2"	10	'0'	'1'	'2'	30711	4	'1'
108	" 3"	45	"2"	4	'0'	'0'	'2'	24340	4	'2'
27	" 4"	45	"2"	8	'0'	'0'	'3'	24099	4	'1'
76	" 5"	46	"2"	3	'1'	'1'	'2'	21795	4	'1'
...

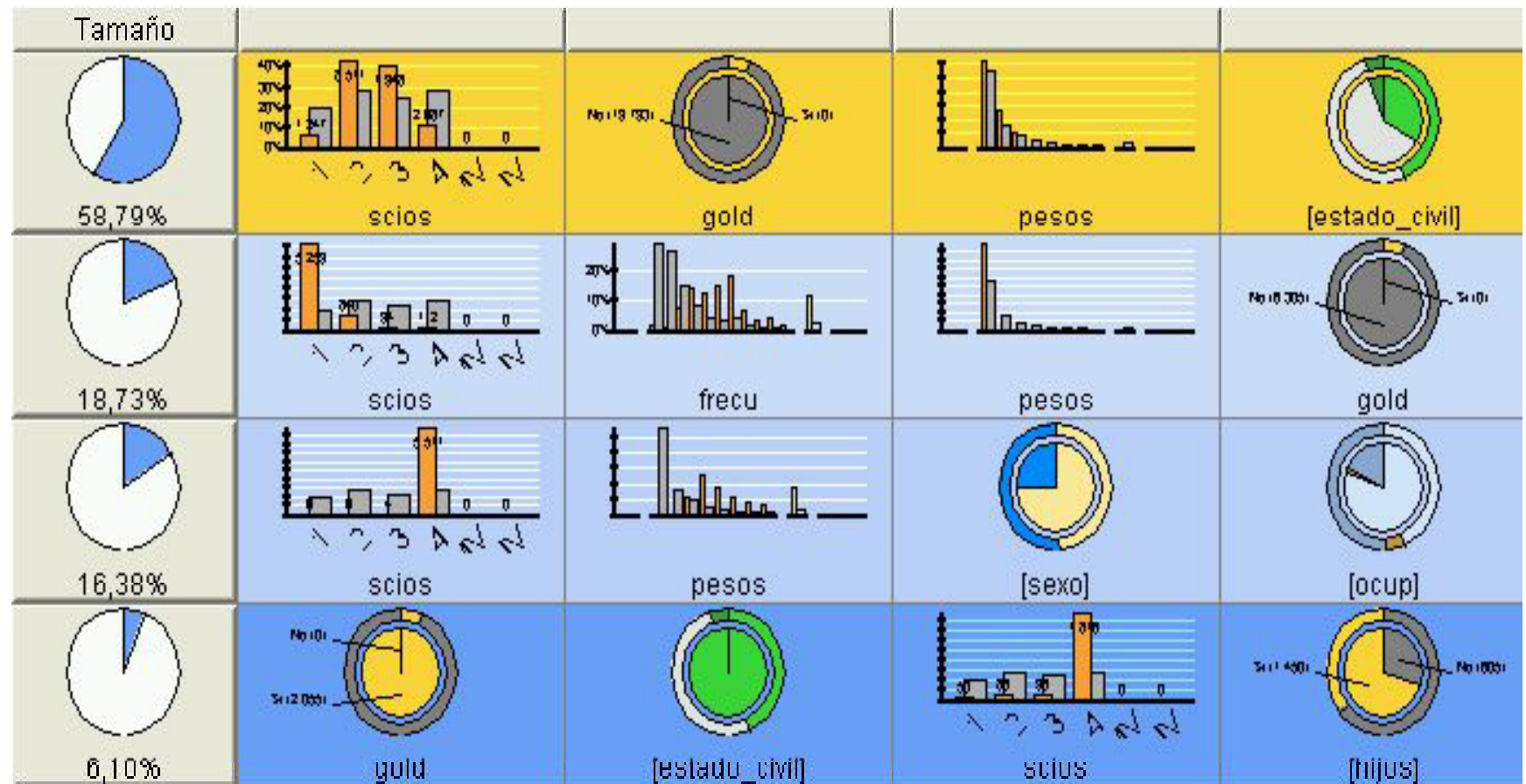
Ejemplo de Clustering



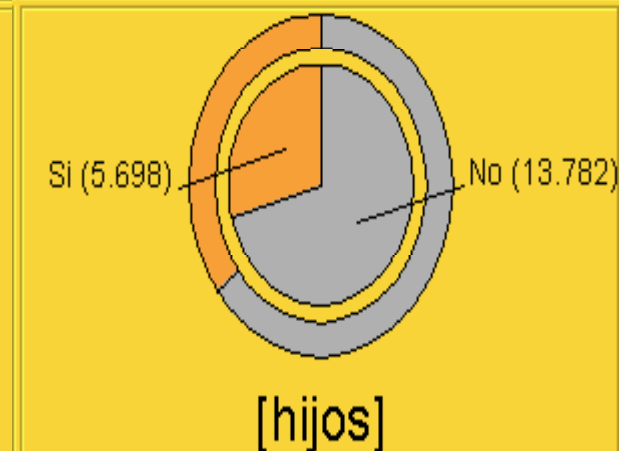
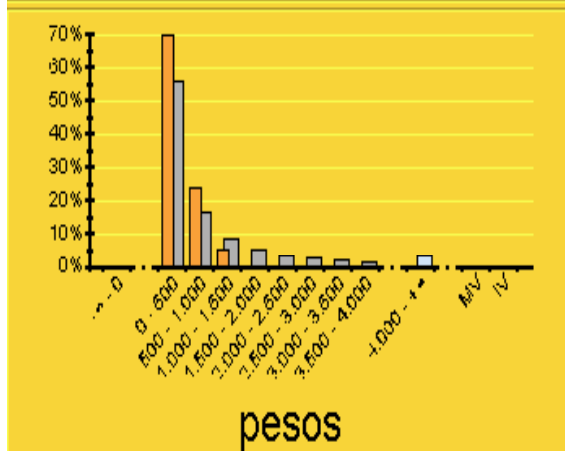
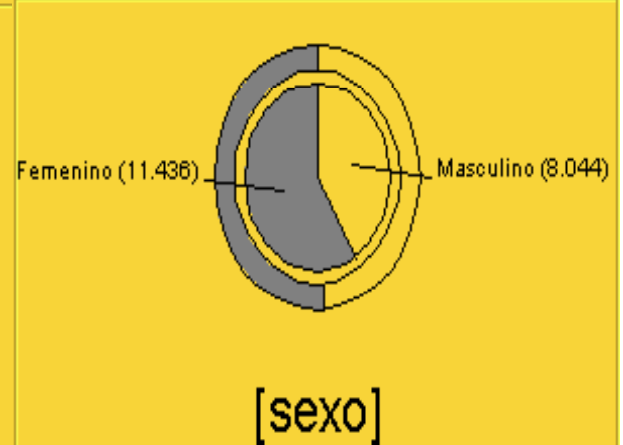
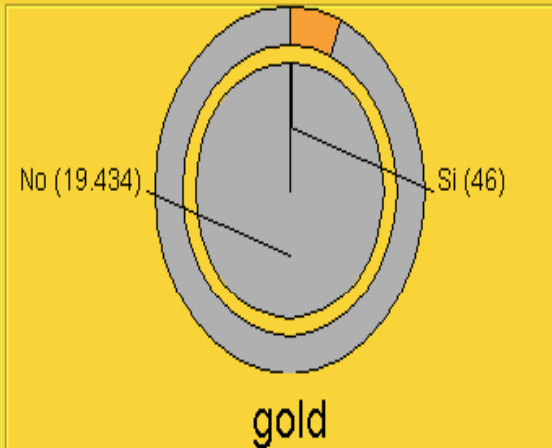
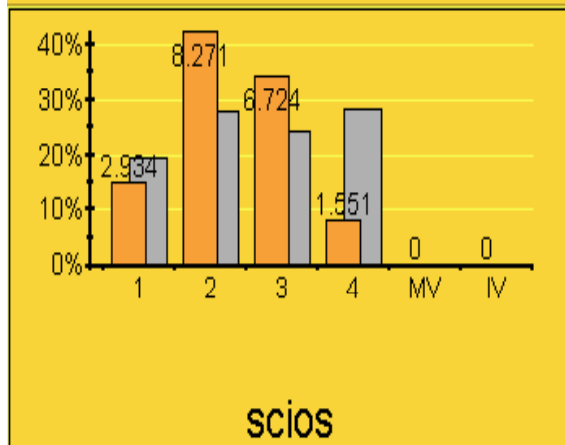
- Medida de calidad del modelo:
 - Criterio de Condorcet: asume un valor entre 0 y 1.

- Criterios de segmentación:
 - Se toman como variables activas las que corresponden al comportamiento de consumo.
 - Se toman como variables suplementarias los atributos sociodemográficos.

Solución de 4 clusters



Buenos clientes sin tarjeta Gold



Asociación



- Análisis de la canasta de mercado:
 - Objetivo: generar reglas del tipo:
SI condición ENTONCES resultado
 - Ejemplo:
SI producto A y producto C ENTONCES producto B

- ¿Cuán buena es una regla?. Medidas que la califican:
 - Soporte.
 - Confianza.
 - Mejora.

Ejemplo de Asociación



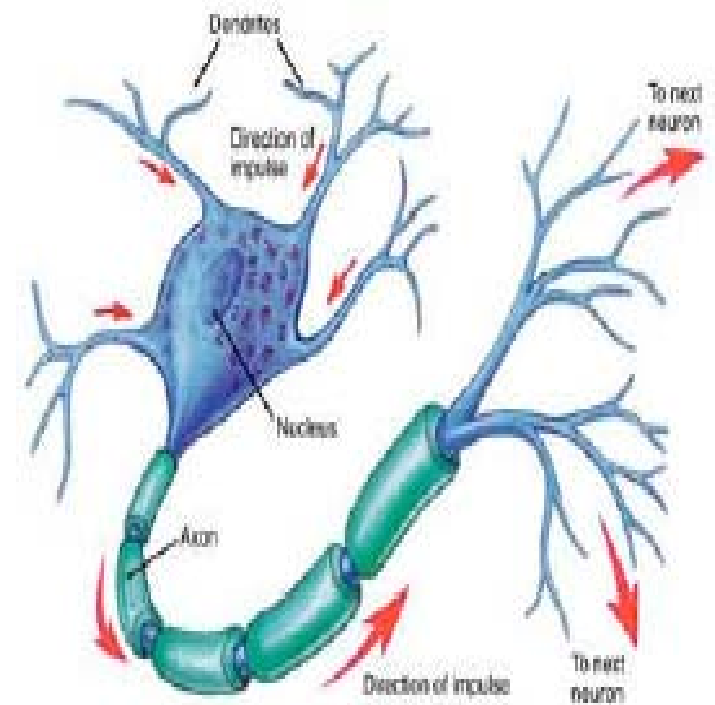
- El dueño de una pizzería vende 3 gustos de pizzas: pepperoni, queso y hongos, y quiere armar "combos" con las combinaciones más convenientes.
- Parte de un conjunto de 2000 tickets con los correspondientes items (gusto de pizza) incluido en cada uno.

Hongos	Pepperoni	Queso	Cantidad
Si	Si	Si	100
Si	Si	No	400
Si	No	Si	300
Si	No	No	100
No	Si	Si	200
No	Si	No	150
No	No	Si	200
No	No	No	550
TOTAL			2000

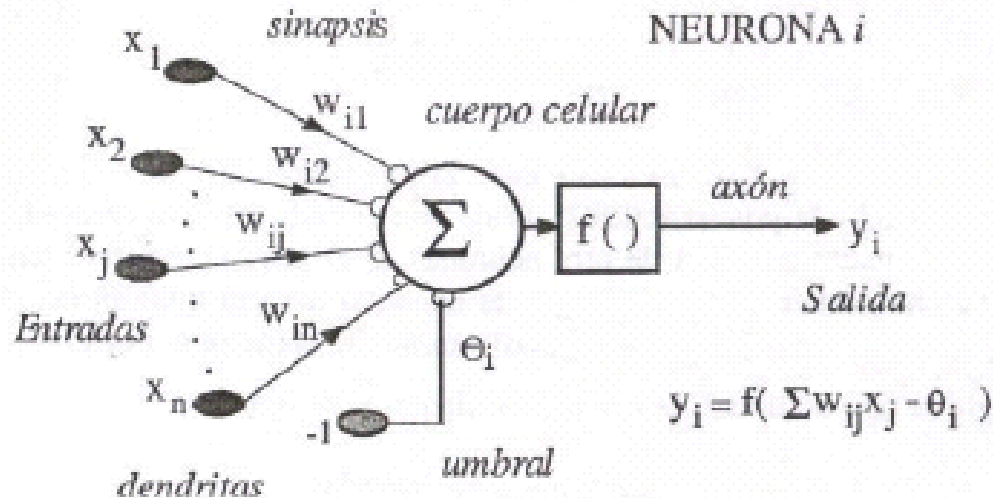
Red neuronal



- La Inteligencia Artificial trabaja con modelos conexionistas.
- El modelo conexionista imita el sistema más complejo conocido hasta el momento: el **cerebro**.
- El cerebro está formado por millones de células llamadas **neuronas**.
- Estas neuronas son unos procesadores de información muy sencillos con un canal de entrada de información (dendrita), un órgano de cómputo (soma) y un canal de salida de información (axón).



La neurona artificial



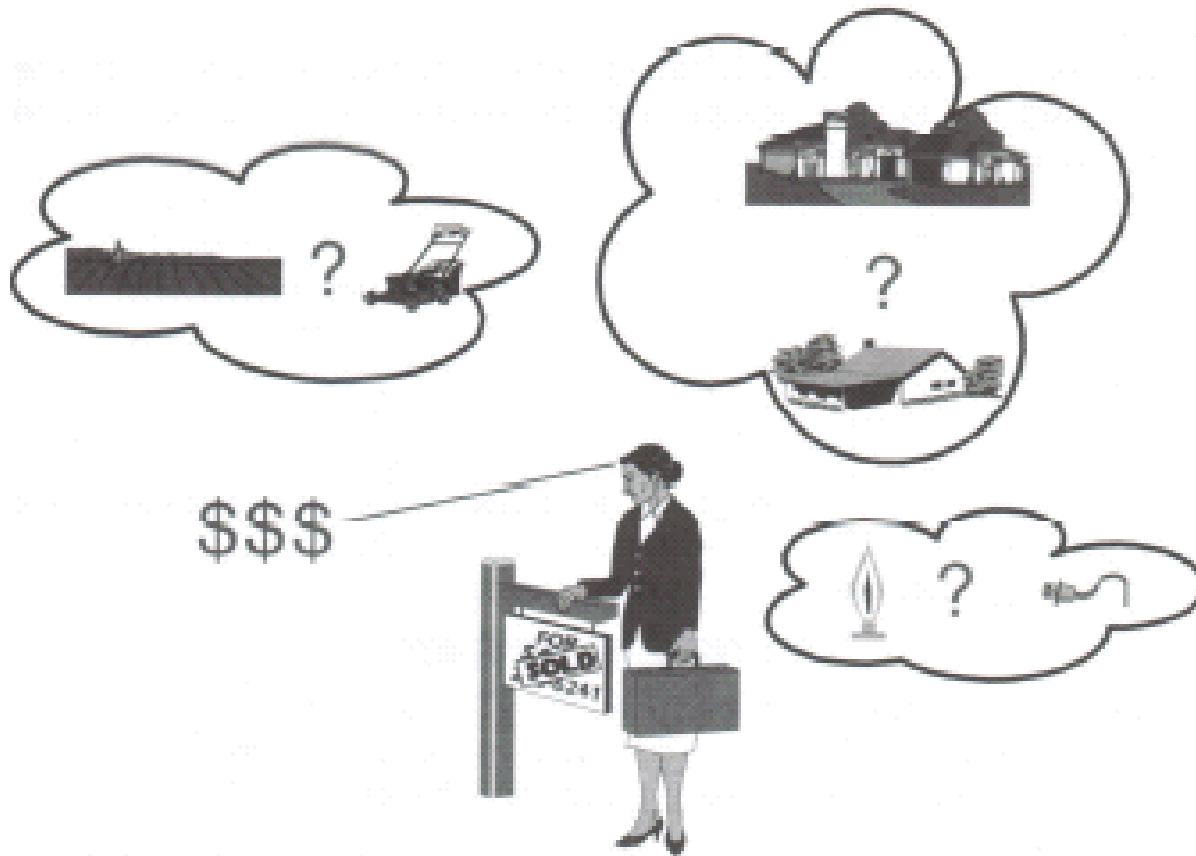
➤ Elementos:

- Entradas.
- Pesos sinápticos.
- Reglas de propagación.
- Función de activación.

Ejemplo de red neuronal



Valuación de propiedades



Ejemplo de red neuronal

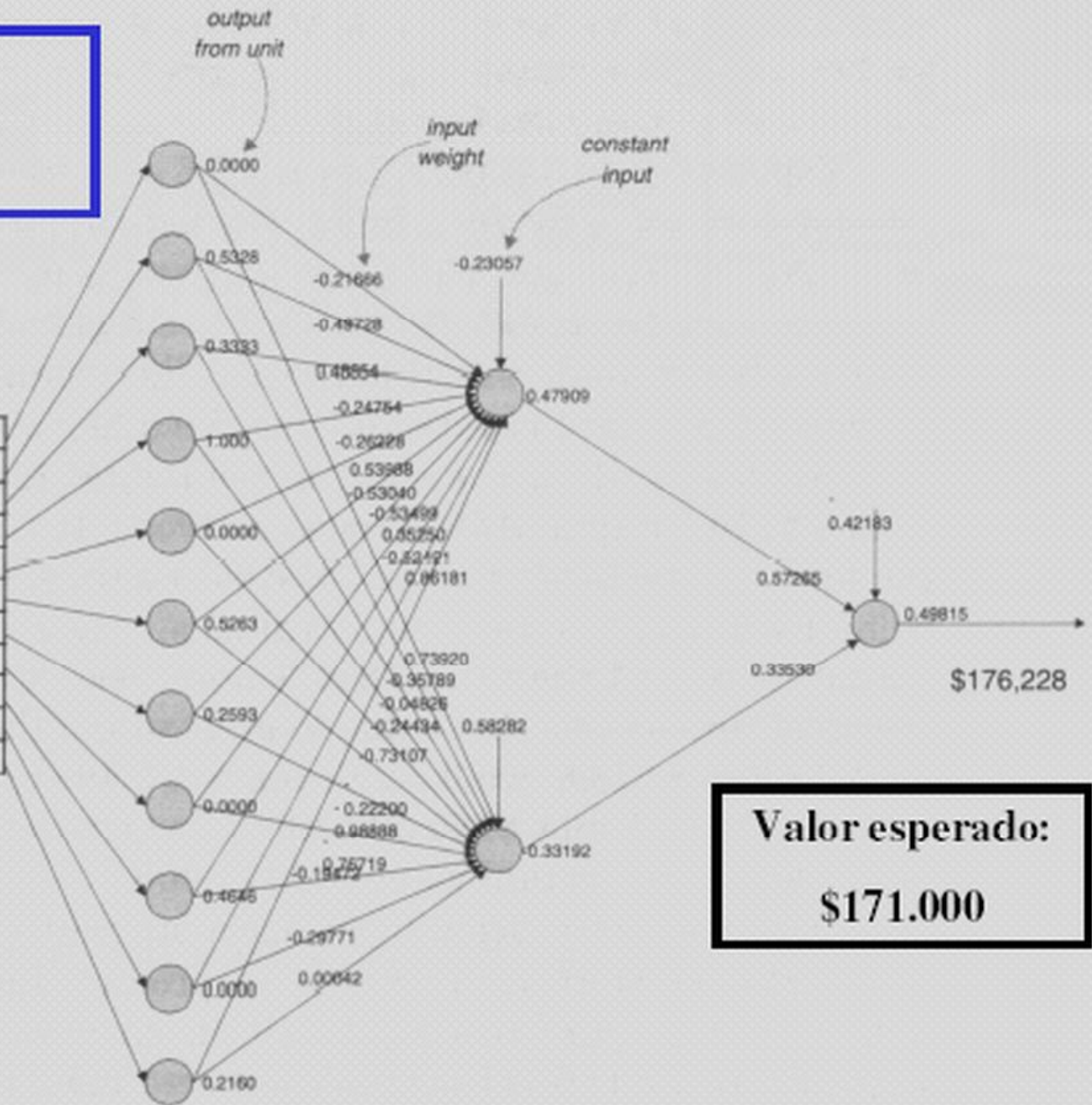


Datos

Feature	Description	Values	Training set Example	Massaged Value
Sales_price	Sales price	\$103,000 - \$250,000	\$171,000	0.4626
Num_Apts	Number of dwelling units	1-3	1	0.0000
Year_built	Year built	1850-1986	1923	0.5328
Plumbing	Number of plumbing fixtures	5-17	9	0.3333
Heating	Heating system type	coded A or B	B	1.0000
Basement_Garage	Basement garage (number of cars)	0-2	0	0.0000
Attached_Garage	Attached garage area	0-228	120	0.5263
Living_Area	Total living area	714-4185	1614	0.2593
Deck_Area	Deck Area	0-738	0	0.0000
Porch_Area	Porch area	0-452	210	0.4616
Recroom_Area	Recreation room area	0-672	0	0.0000
Basement_Area	Basement area	0-810	175	0.2160

Una iteración de la red

Num_Apartments	1	0.0000
Year_Built	1923	0.5328
Plumbing_Fixtures	9	0.3333
Heating_Type	B	1.0000
Basement_Garage	0	0.0000
Attached_Garage	120	0.5263
Living_Area	1614	0.2593
Deck_Area	0	0.0000
Porch_Area	210	0.4646
Recroom_Area	0	0.0000
Basement_Area	175	0.2160



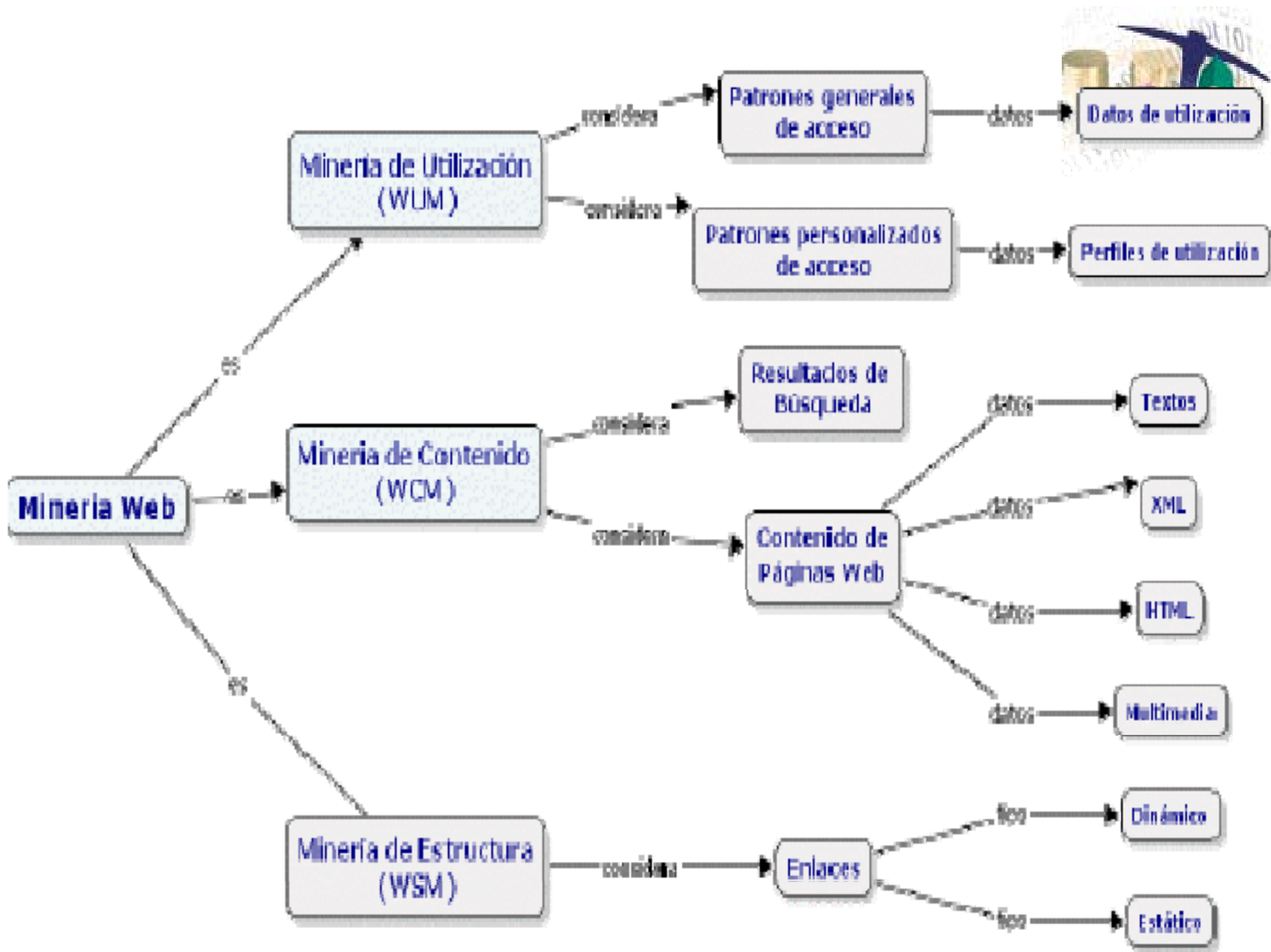
Valor esperado:
\$171.000

Extensiones de la DM



➤ Web mining

- Web content mining (minería de contenido web). Es el proceso que consiste en la extracción de conocimiento del contenido de documentos o sus descripciones.
- Web structure mining (minería de estructura web). Es el proceso de inferir conocimiento de la organización del WWW y la estructura de sus ligas.
- Web usage mining (minería de uso web). Es el proceso de extracción de modelos interesantes usando los logs de los accesos al web.



Conclusiones Ventajas



- La Minería de Datos es una herramienta eficaz para dar respuestas a preguntas complejas de Inteligencia de Negocios.
- Las herramientas disponibles permiten automatizar gran parte de la tarea de encontrar los patrones de comportamiento ocultos en los datos.
- Es una buena manera de convertir datos en información, y esta a su vez en conocimiento, para la correcta toma de decisiones.

Conclusiones Ventajas



- Ahorra grandes cantidades de dinero a una empresa y abre nuevas oportunidades de negocios.
- proporciona poder de decisión a los usuarios del negocio, y es capaz de medir las acciones y resultados de la mejor forma.

Conclusiones Desventajas



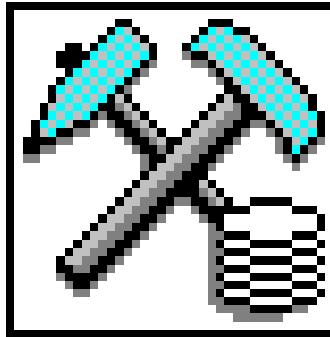
- Que los productos a comercializar son significativamente costosos.
- Que se requiera de experiencia para utilizar herramientas de tecnología.
- Que sea fácil de hallar patrones equívocos triviales o no interesantes.
- La Privacidad.

Referencias



- Building Data Mining Applications for CRM. A. Berson, S. Shmit, K. Thearling. Mc Graw Hill, 2000.
- Data Mining with Neuronal Networks. Joseph Bigus. Mc Graw Hill, 1996.
- Principles of Data Mining. D. Hand, H. Manilla, P. Smyth. The MIT Press. USA, 2000.
- U. Fayyad, G. Grinstein, A. Wierse. Data Mining and Knowledge Discovery. M. Kaufmann, Harcourt Intl., USA, 2001.

Fin de la Presentación



Muchas Gracias !!