CIFF
Centro Internacional
de Formación Financiera

CIFF Trustees:

Santander

Universidad de Alcalá

Business Intelligence
&
Data Mining

Profesor:
Pedro Pasquau

Octubre - 2015

MASTER EN BUSINESS ANALYTICS & BIG DATA
2015– 2016

Business Analytics & DataMining

A.- Data Mining

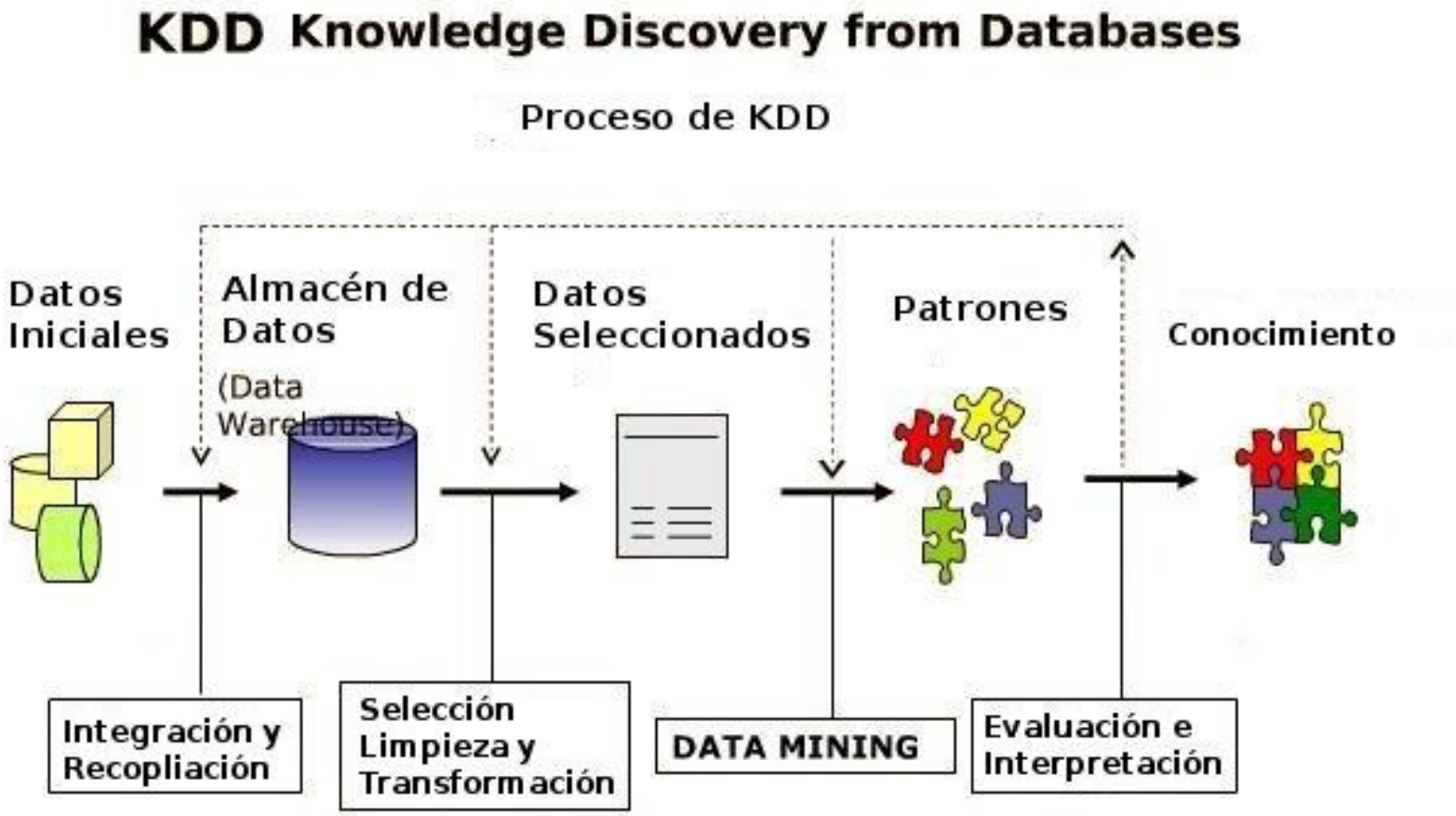
- Qué es Data Mining.
- Proceso Data Mining
- Algoritmos de Data Mining
- Herramienta de Data Mining: Introdución a WEKA

Business Analytics & DataMining

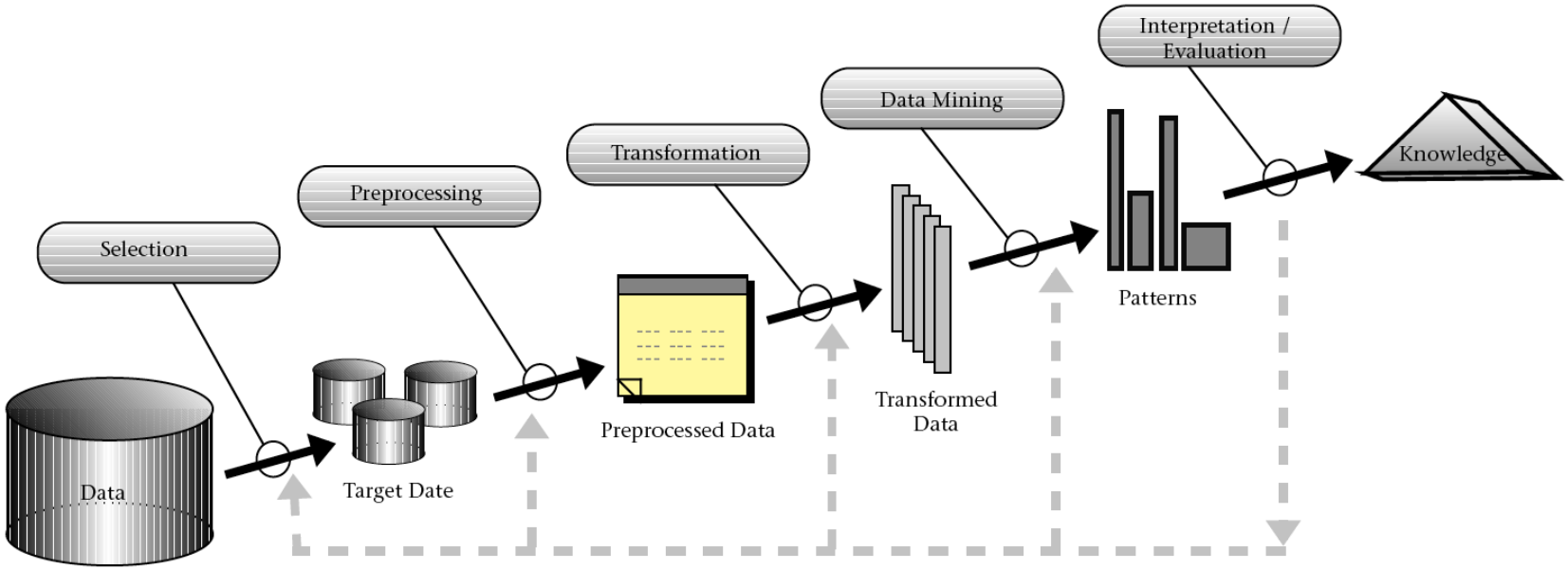
A.- Data Mining

- Qué es Data Mining.
- Proceso Data Mining
- Algoritmos de Data Mining
- Herramienta de Data Mining: Introdución a WEKA

Fases del KDD



Fases del KDD: DataMining



Fases del KDD: Recogida de Datos

IMP!! Las primeras fases del KDD (ETL) determinan que las fases sucesivas sean capaces de extraer conocimiento válido y útil a partir de la información original.

Generalmente, la información que se quiere investigar sobre un cierto dominio de la organización se encuentra:

- en bases de datos y otras **fuentes muy diversas**,
- tanto internas como **externas**.
- muchas de estas fuentes son las que se utilizan para el trabajo **transaccional**.

El análisis posterior será mucho más sencillo si la fuente es **unificada**, **accesible** (interna) y desconectada del trabajo **transaccional**.

Fases del KDD: Recogida de Datos

El proceso de minería de datos:

- Depende mucho de la fuente:
  - OLAP u OLTP-Transaccional.
  - Datawarehouse o esquema original.
  - ROLAP o MOLAP.

Fases del KDD: Recogida de Datos

El proceso de minería de datos:

- Depende también del tipo de usuario:
  - 'Minero o picapedrero' (o 'granjeros): se dedican fundamentalmente a realizar informes periódicos, ver la evolución de determinados parámetros, controlar valores anómalos, etc.
  - '**explorador**': encargados de encontrar nuevos patrones significativos utilizando técnicas de minería de datos.

Fases del KDD: Recogida de Datos

Recogida de Información Externa:

- Aparte de información interna de la organización, los almacenes de datos pueden recoger información externa:
  - Demografías (censo), páginas amarillas, psicografías (perfiles por zonas), uso de Internet, información de otras organizaciones.
  - Datos compartidos en una industria o área de negocio, organizaciones y colegios profesionales, catálogos, etc.
  - Datos resumidos de áreas geográficas, distribución de la competencia, evolución de la economía, información de calendarios y climatológicas, programaciones televisivas-deportivas, catástofres,..
  - Bases de datos externas compradas a otras compañías.

Recordar Conceptos Básicos

- Variable objetivo: variable sobre la que se quiere hacer una predicción o sobre la que se quiere modelar un comportamiento.
- Variables explicativas: resto de variables del modelo que se combinan para explicar el comportamiento de la variable objetivo.
- Variable Continua: Variable que puede tomar cualquier valor en un rango de números, por ejemplo el salario o las ventas de un producto.
- Variable Discreta Nominal: Variable que sólo puede tomar una serie de valores, sin relación de orden entre ellos, por ejemplo nombres de ciudad, departamentos de una empresa o el sexo.
- Variable Discreta Ordinal: Variable que sólo puede tomar una serie de valores pero entre los que existe una relación de orden, como por ejemplo la categoría laboral o la valoración en una encuesta

Fases: Selección, Limpieza y Transformación de Datos

Limpieza (data cleaning) y Selección de datos:

Se deben eliminar el mayor número posible de datos erróneos o inconsistentes (limpieza) e irrelevantes (criba).

Métodos estadísticos casi exclusivamente.

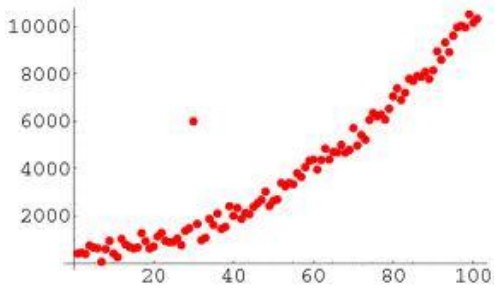
- **histogramas** (detección de datos anómalos).
- **selección** de datos (muestreo, ya sea verticalmente, eliminando atributos, u horizontalmente, eliminando tuplas).
- **redefinición** de atributos (agrupación o separación).

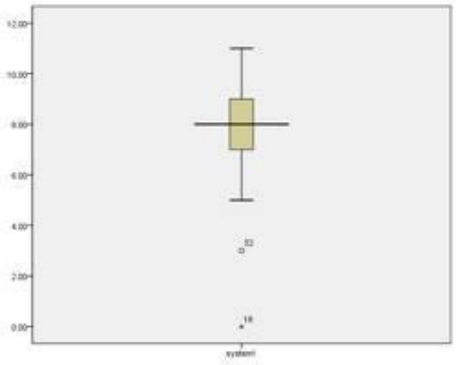
- Outlier: Elemento del conjunto de datos que es significativamente diferente a los otros datos de la colección, o un elemento que parece implicar un patrón que es inconsistente con el resto de los datos.

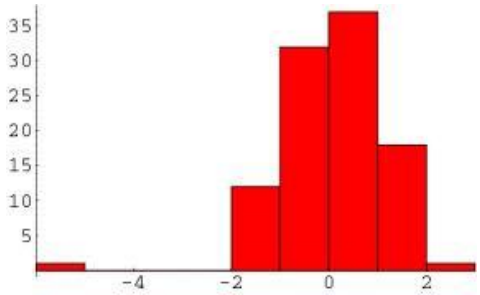
  De forma Matemática, se puede generalizar como:

  - El rango intercuartil (RI) puede calcularse restando Q1 (primer cuartil) del Q3 (tercer cuartil), esto es, RI = Q3 - Q1.
  - Se toma el RI y se multiplica por 1.5.
  - Cualquier elemento de datos será denominado una variable extraña cuando sea menor que Q1 - 1.5x(RI) o mayor que Q3 + 1.5x(RI).

  * Por comodidad consideraremos outlier los valores que se encuentren por debajo del 5% y por encima del 95%







Fases: Selección, Limpieza y Transformación de Datos

Acciones ante datos anómalos (outliers):

- **ignorar**: algunos algoritmos son robustos a datos anómalos (p.ej. árboles)
- **filtrar** (eliminar o reemplazar) **la columna**: solución extrema, pero a veces existe otra columna dependiente con datos de mayor calidad.

Alternativa→Preferible a eliminar la columna es reemplazarla por una columna discreta diciendo si el valor era normal u outlier (por encima o por debajo).

- **filtrar la fila**: puede sesgar los datos, porque muchas veces las causas de un dato erróneo están relacionadas con casos o tipos especiales.
- **reemplazar** el valor: por el valor '**nulo**' si el algoritmo lo trata bien o por máximos o mínimos, dependiendo por donde es el outlier, o por medias. A veces se puede *predecir* a partir de otros datos, utilizando cualquier técnica de ML.
- **discretizar**: transformar un valor continuo en uno discreto (p.ej. muy alto, alto, medio, bajo, muy bajo) hace que los outliers caigan en 'muy alto' o 'muy bajo' sin mayores problemas.

Fases: Selección, Limpieza y Transformación de Datos

Acciones ante datos que faltan (missing values):

- **ignorar**: algunos algoritmos son robustos a datos faltantes (p.ej. árboles).
- **filtrar** (eliminar o reemplazar) **la columna**: solución extrema, pero a veces existe otra columna dependiente con datos de mayor calidad.

Alternativa→Preferible a eliminar la columna es reemplazarla por una columna booleana diciendo si el valor existía o no.

- **filtrar la fila**: claramente sesga los datos, porque muchas veces las causas de un dato faltante están relacionadas con casos o tipos especiales.
- **reemplazar el valor**: por medias. A veces se puede *predecir* a partir de otros datos, utilizando cualquier técnica de ML.
- **segmentar:** se segmentan las tuplas por los valores que tienen disponibles. Se obtienen modelos diferentes para cada segmento y luego se combinan.

Fases: Selección, Limpieza y Transformación de Datos

Razones sobre datos faltantes (missing values):

A veces es importante examinar las razones y actuar en consecuencia:

- algunos valores expresan características relevantes subyacenyes: p.ej. la falta de teléfono puede representar en muchos casos un deseo de que no se moleste a la persona en cuestión, o un cambio de domicilio reciente.
- valores no existentes: muchos valores faltantes existen en la realidad, pero otros no. P.ej. el cliente que se acaba de dar de alta no tiene consumo medio de los últimos 12 meses.
- datos incompletos: si los datos vienen de fuentes diferentes, al combinarlos se suele hacer la unión y no la intersección de campos, con lo que muchos datos faltantes representan que esas tuplas vienen de una/s fuente/s diferente/s al resto.

Fases: Selección, Limpieza y Transformación de Datos

Transformación del Esquema:

- Esquema Original:
  - Ventajas: Las R.I. se mantienen (no hay que reaprenderlas, no despistan)
  - Inconvenientes: Muchas técnicas no se pueden utilizar.
- Tabla Universal: *Cualquier Esquema Relacional se puede convertir (en una correspondencia 1 a 1) a una tabla universal.*
  - Ventajas: Modelos de aprendizaje más simples (proposicionales).
  - Desventajas: Muchísima Redundancia (tamaños ingentes). La información del esquema se pierde.
- Desnormalizado Tipo Estrella o Copo de Nieve (***datamarts***):
  - Ventajas: Se pueden buscar reglas sobre información sumarizada y si resultan factibles se pueden comprobar con la información detallada.
    Con operadores propios: *Roll-up*, *Drill-down*, *Slicing and Dicing*.
  - Desventajas: Orientadas a extraer un tipo de información.

Fases: Selección, Limpieza y Transformación de Datos

Transformación de los Campos:

- Numerización / Etiquetado
  - Ventajas: Se reduce espacio. Ej: apellido ⇒ entero. Se pueden utilizar técnicas más simples.
  - Desventajas: Se necesita meta-información para distinguir los datos inicialmente no numéricos (la cantidad no es relevante) de los inicialmente numéricos (la cantidad es relevante: precios, unidades, etc.)
    A veces se puede "sesgar" el modelo.
- Discretización:
  - Ventajas: Se reduce espacio. Ej. 0..10 ⇒ (pequeño, mediano, grande). Se pueden utilizar árboles de decisión y construir reglas discretas.
  - Desventajas: Una mala discretización puede invalidar los resultados.

Fases del KDD: La Minería de Datos

Patrones a descubrir:

- Una vez recogidos los datos de interés, el usuario puede decidir qué tipo de patrón quiere descubrir.
- El tipo de conocimiento que se desea extraer va a marcar claramente la *técnica* de minería de datos a utilizar.
- Según como sea la búsqueda del conocimiento se puede distinguir entre:
  - *Directed data mining*: se sabe claramente lo que se busca, generalmente predecir unos ciertos datos o clases.
  - *Undirected data mining*: no se sabe lo que se busca, se trabaja con los datos

Fases del KDD: Evaluación y Validación

La fase anterior produce una o más hipótesis de **modelos**.

Para seleccionar y validar estos modelos es necesario el uso de **criterios de evaluación de hipótesis**.

Por ejemplo:

1ª Fase: Comprobación de la precisión del modelo en un **banco de ejemplos independiente** del que se ha utilizado para aprender el modelo. Se puede elegir el mejor modelo.

2ª Fase: Se puede realizar una **experiencia piloto** con ese modelo. Por ejemplo, si el modelo encontrado se quería utilizar para predecir la respuesta de los clientes a un nuevo producto, se puede enviar un mailing a un subconjunto de clientes y evaluar la *fiabilidad del modelo*.

Fases del KDD: Interpretación y Difusión

El despliegue del modelo a veces a veces es trivial pero otras veces requiere un proceso de implementación o interpretación:

- El modelo puede requerir **implementación** (p.ej. tiempo real detección de tarjetas fraudulentas).
- El modelo es descriptivo y requiere **interpretación** (p.ej. una caracterización de zonas geográficas según la distribución de los productos vendidos).
- El modelo puede tener muchos usuarios y necesita **difusión**: el modelo puede requerir ser expresado de una manera comprensible para ser distribuido en la organización (p.ej. las cervezas y los productos congelados se compran frecuentemente en conjunto $\Rightarrow$ ponerlos en estantes distantes).

Fases del KDD: Actualización y Monitorización

Los procesos derivan en un **mantenimiento**:

- Actualización: Un modelo válido puede dejar de serlo: cambio de contexto (económicos, competencia, fuentes de datos, etc.).

- Monitorización: Consiste en ir revalidando el modelo con cierta frecuencia sobre nuevos datos, con el objetivo de detectar si el modelo requiere una actualización.

Producen **realimentaciones** en el proceso KDD.

Tipología de Técnicas de Minería de Datos

Las técnicas de minería de datos crean modelos que son **predictivos y/o descriptivos**.

Un modelo predictivo responde preguntas sobre datos futuros.

- ¿Cuáles serán las ventas el año próximo?
- ¿Es esta transacción fraudulenta?
- ¿Qué tipo de seguro es más probable que contrate el cliente X?

Un modelo descriptivo proporciona información sobre las relaciones entre los datos y sus características. Genera información del tipo:

- Los clientes que compran pañales suelen comprar cerveza.
- El tabaco y el alcohol son los factores más importantes en la enfermedad Y.
- Los clientes sin televisión y con bicicleta tienen características muy diferenciadas del resto.

Tipología de Técnicas de Minería de Datos

Ejemplo de Modelo **Predictivo**:

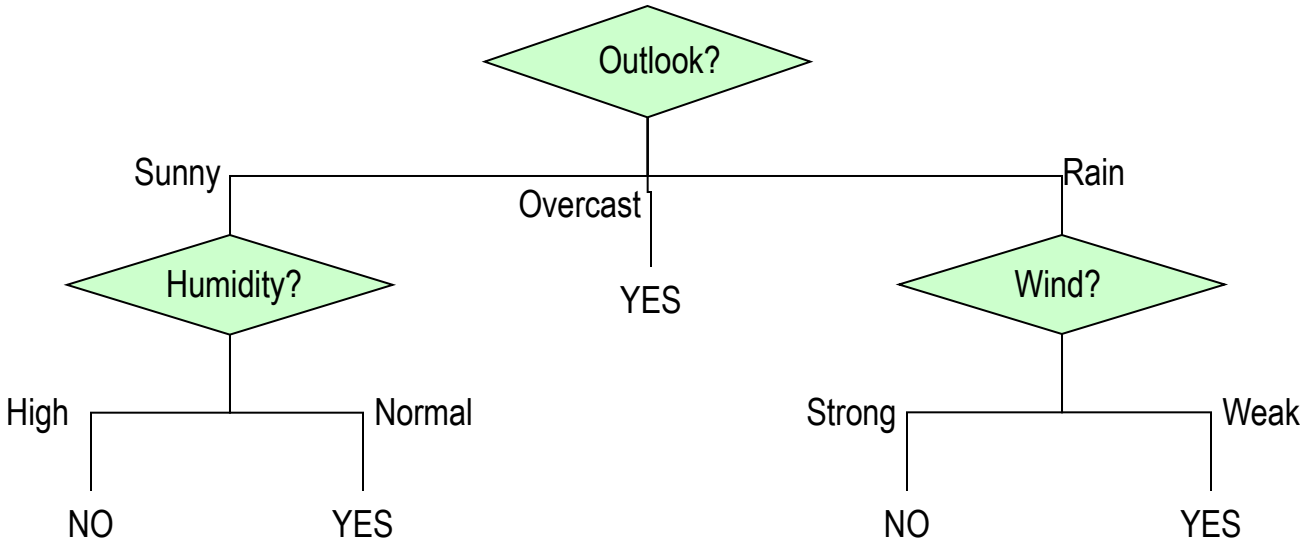
- Queremos saber si jugar o no jugar esta tarde al tenis.
- Hemos recogido datos de experiencias anteriores:

Example	Sky	Temperature	Humidity	Wind	PlayTennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Tipología de Técnicas de Minería de Datos

Ejemplo de Modelo Predictivo:

- Pasamos estos ejemplos a un algoritmo de aprendizaje de **árboles de decisión**, señalando el atributo "PlayTennis" como la clase (output).
- El resultado del algoritmo es el siguiente modelo:



- Ahora podemos utilizar este modelo para predecir si esta tarde jugamos o no al tenis. P.ej., la instancia:

  (Outlook = sunny, Temperature = hot, Humidity = high, Wind = strong)

  es NO.

Tipología de Técnicas de Minería de Datos

Ejemplo de Modelo **Descriptivo**:

- Queremos categorizar nuestros empleados.
- Tenemos estos datos de los empleados:

#Ej	Sueldo	Casado	Coche	Hijos	Alq/Prop	Sindic.	Bajas/Año	Antigüedad	Sexo
1	10000	Sí	No	0	Alquiler	No	7	15	H
2	20000	No	Sí	1	Alquiler	Sí	3	3	M
3	15000	Sí	Sí	2	Prop	Sí	5	10	H
4	30000	Sí	Sí	1	Alquiler	No	15	7	M
5	10000	Sí	Sí	0	Prop	Sí	1	6	H
6	40000	No	Sí	0	Alquiler	Sí	3	16	M
7	25000	No	No	0	Alquiler	Sí	0	8	H
8	20000	No	Sí	0	Prop	Sí	2	6	M
9	20000	Sí	Sí	3	Prop	No	7	5	H
10	30000	Sí	Sí	2	Prop	No	1	20	H
11	50000	No	No	0	Alquiler	No	2	12	M
12	8000	Sí	Sí	2	Prop	No	3	1	H
13	20000	No	No	0	Alquiler	No	27	5	M
14	10000	No	Sí	0	Alquiler	Sí	0	7	H
15	8000	No	Sí	0	Alquiler	No	3	2	H

Tipología de Técnicas de Minería de Datos

Ejemplo de Modelo Descriptivo:

- Pasamos estos ejemplos a un algoritmo de clustering K-meams.
- Se crean tres clusters, con la siguiente descripción:

cluster 1: 5 examples
Sueldo : 22600
Casado : No -> 0.8
Sí -> 0.2
Coche : No -> 0.8
Sí -> 0.2
Hijos : 0
Alq/Prop : Alquiler -> 1.0
Sindic. : No -> 0.8
Sí -> 0.2
Bajas/Año : 8
Antigüedad : 8
Sexo : H -> 0.6
M -> 0.4

cluster 2: 4 examples
Sueldo : 22500
Casado : No -> 1.0
Coche : Sí -> 1.0
Hijos : 0
Alq/Prop : Alquiler -> 0.75
Prop -> 0.25
Sindic. : Sí -> 1.0
Bajas/Año : 2
Antigüedad : 8
Sexo : H -> 0.25
M -> 0.75

cluster 3: 6 examples
Sueldo : 18833
Casado : Sí -> 1.0
Coche : Sí -> 1.0
Hijos : 2
Alq/Prop : Alquiler -> 0.17
Prop -> 0.83
Sindic. : No -> 0.67
Sí -> 0.33
Bajas/Año : 5
Antigüedad : 8
Sexo : H -> 0.83
M -> 0.17

Tipología de Técnicas de Minería de Datos

Ejemplo de Modelo Descriptivo:

cluster 1: 5 examples
Sueldo : 22600
Casado : No -> 0.8
Sí -> 0.2
Coche : No -> 0.8
Sí -> 0.2
Hijos : 0
Alq/Prop : Alquiler -> 1.0
Sindic. : No -> 0.8
Sí -> 0.2
Bajas/Año : 8
Antigüedad : 8
Sexo : H -> 0.6
M -> 0.4

cluster 2: 4 examples
Sueldo : 22500
Casado : No -> 1.0
Coche : Sí -> 1.0
Hijos : 0
Alq/Prop : Alquiler -> 0.75
Prop -> 0.25
Sindic. : Sí -> 1.0
Bajas/Año : 2
Antigüedad : 8
Sexo : H -> 0.25
M -> 0.75

cluster 3: 6 examples
Sueldo : 18833
Casado : Sí -> 1.0
Coche : Sí -> 1.0
Hijos : 2
Alq/Prop : Alquiler -> 0.17
Prop -> 0.83
Sindic. : No -> 0.67
Sí -> 0.33
Bajas/Año : 5
Antigüedad : 8
Sexo : H -> 0.83
M -> 0.17

- GRUPO 1: Sin hijos y de alquiler. Poco sindicados. Muchas bajas.
- GRUPO 2: Sin hijos y con coche. Muy sindicados. Pocas bajas. Normalmente de alquiler y mujeres.
- GRUPO 3: Con hijos, casados y con coche. Propietarios. Poco sindicados. Hombres.

Tipología de Técnicas de Minería de Datos

Tipos de conocimiento:

- **Asociaciones**: Una asociación entre dos atributos ocurre cuando la frecuencia de que se den dos valores determinados de cada uno conjuntamente es relativamente alta.
    - Ejemplo, en un supermercado se analiza si los pañales y los potitos de bebé se compran conjuntamente.
- **Dependencias**: Una dependencia funcional (aproximada o absoluta) es un patrón en el que se establece que uno o más atributos determinan el valor de otro. Ojo! Existen muchas dependencias nada interesantes (causalidades inversas).
    - Ejemplo: que un paciente haya sido ingresado en maternidad determina su sexo.

La búsqueda de asociaciones y dependencias se conoce a veces como análisis exploratorio.

Tipología de Técnicas de Minería de Datos

Tipos de conocimiento (cont.):

- **Clasificación**: Una clasificación se puede ver como el esclarecimiento de una dependencia, en la que el atributo dependiente puede tomar un valor entre varias clases, ya conocidas.
    - Ejemplo: se sabe (por un estudio de dependencias) que los atributos edad, número de miopías y astigmatismo han determinado los pacientes para los que su operación de cirugía ocular ha sido satisfactoria.
        - Podemos intentar determinar las reglas exactas que clasifican un caso como positivo o negativo a partir de esos atributos.
- **Agrupamiento / Segmentación**: El agrupamiento (o clustering) es la detección de grupos de individuos. Se diferencia de la clasificación en el que no se conocen ni las clases ni su número (aprendizaje no supervisado), con lo que el objetivo es determinar grupos o racimos (clusters) diferenciados del resto.

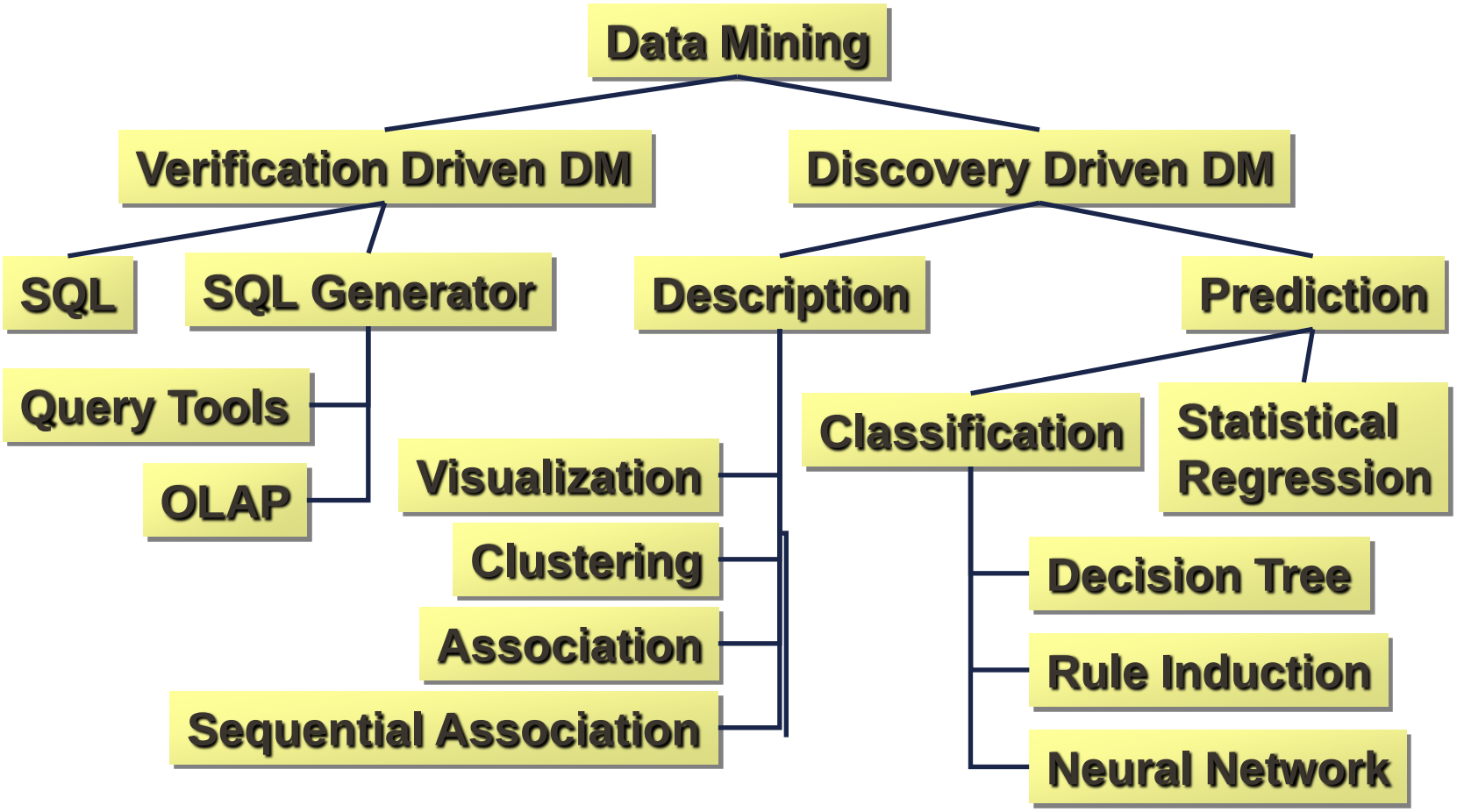
Tipología de Técnicas de Minería de Datos

Tipos de conocimiento (cont.):

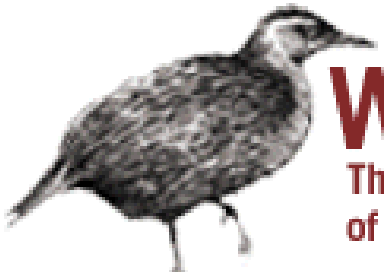
- **Tendencias/Regresión**: El objetivo es predecir los valores de una variable **continua** a partir de la evolución sobre otra variable continua, generalmente el tiempo.
    - Ejemplo, se intenta predecir el número de clientes o pacientes, los ingresos, llamadas, ganancias, costes, etc. a partir de los resultados de semanas, meses o años anteriores.

- **Reglas Generales**: patrones no se ajustan a los tipos anteriores. Recientemente los sistemas incorporan capacidad para establecer otros patrones más generales.

Taxonomía Técnicas de Minería de Datos.



RAPID|MINER



WEKA
The University
of Waikato

EM
EnterpriseMiner

KNIME

Sistemas

http://www.stratebi.es/todobi/ago10/Algoritmos-Herramientas_Data_Mining.pdf

Visualización

Las técnicas de visualización de datos se utilizan fundamentalmente con dos objetivos:

- aprovechar la gran capacidad humana de extraer patrones a partir de imágenes.
- ayudar al usuario a comprender más rápidamente patrones descubiertos automáticamente por un sistema de KDD.

Visualización

Estos dos objetivos marcan dos momentos diferentes del uso de la visualización de los datos (no excluyentes):

- visualización *previa* (tb. Visual Data Mining [Wong 1999]): se utiliza para entender mejor los datos y sugerir posibles patrones o qué tipo de herramienta de KDD utilizar.

- visualización *posterior* al proceso de minería de datos: se utiliza para mostrar los patrones y entenderlos mejor.

Visualización

También marcan dos tipos de usuarios diferentes de las técnicas:

- La visualización ***previa*** se utiliza frecuentemente por picapedreros, para ver tendencias y resúmenes de los datos, y por exploradores, para ver 'filones' que investigar.

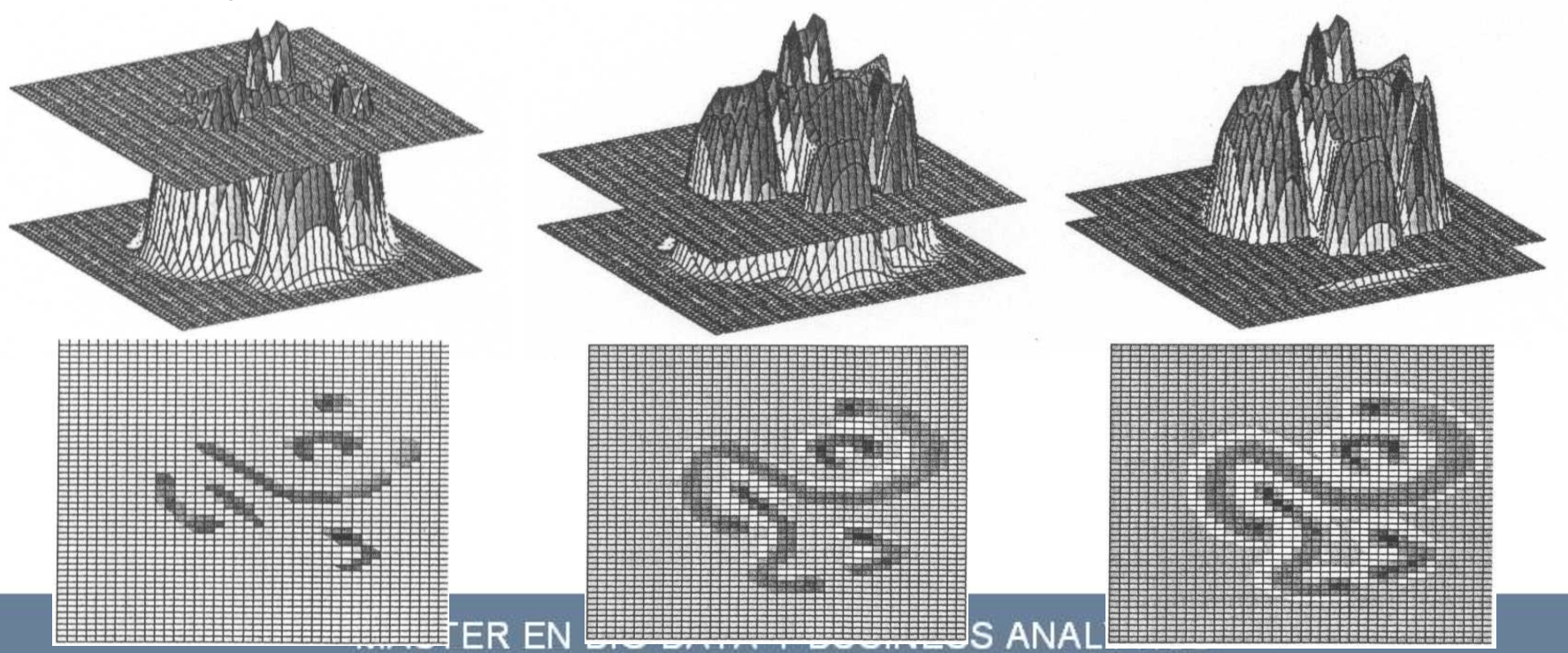
- La visualización ***posterior*** se utiliza frecuentemente para validar y mostrar a los expertos los resultados del DM.

Visualización

Visualización Previa:

Ejemplo: segmentación mediante funciones de densidad, generalmente representadas tridimensionalmente.

Los seres humanos ven claramente los segmentos (clusters) que aparecen con distintos parámetros



Visualización

Visualización Previa:

Mayor problema: **dimensionalidad > 3.**

Objetivo: conseguir proyectar las dimensiones en una representación en 2 (ó 3 simuladas) dimensiones.

Solución:

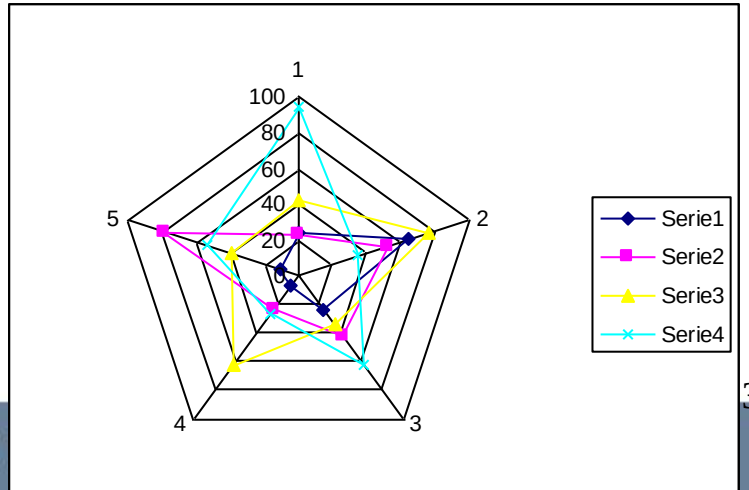
Uso de proyecciones geométricas:

Visualización

Visualización Previa: Proyecciones geométricas:

- técnica de visualización de coordenadas paralelas [Inselberg & Dimsdale 1990]. Se mapea el espacio k-dimensional en dos dimensiones mediante el uso de *k* ejes de ordenadas (escalados linealmente) por uno de abscisas. Cada punto en el espacio *k*-dimensional se hace corresponder con una línea poligonal (polígono abierto), donde cada vértice de la línea poligonal intersecta los *k* ejes en el valor para la dimensión.
    - Cuando hay pocos datos cada línea se dibuja de un color.
    - Cuando hay muchos datos se utiliza una tercera dimensión para los casos.

- técnica radial (igual que la anterior pero los ejes se ponen circularmente) →



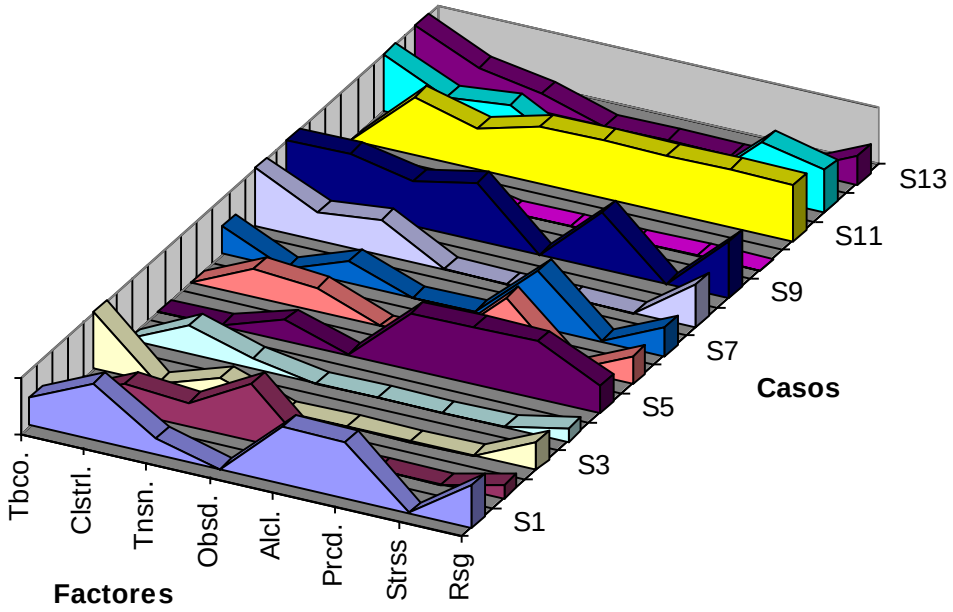
Visualización

Visualización Previa: Ejemplo:dimensionalidad…

Dados ciertos atributos de pacientes (tabaquismo, colesterol, tensión, obesidad, alcoholismo, precedentes, estrés) y su riesgo (muy bajo, bajo, medio, alto, muy alto) de enfermedades coronarias:

Visualización

Tbco.	Clstrl	Tnsn	Obsd	Alcl	Prcd	Strs	Rsg
Med	Alto	8	No	Sí	Sí	No	Alto
Bajo	Med	9	Sí	No	No	No	Bajo
Alto	Bajo	8,5	No	No	No	No	Med
Bajo	Med	7	No	No	No	No	Bajo
Bajo	Bajo	8,5	No	Sí	Sí	Sí	Med
Bajo	Med	9	No	No	Sí	No	Med
Med	Bajo	9	No	No	Sí	No	Med
Alto	Med	11	No	No	No	No	Alto
Alto	Alto	13	Sí	No	Sí	No	M.A.
Bajo	Bajo	7	No	No	No	No	M.B.
Bajo	Alto	12	Sí	Sí	Sí	Sí	M.A.
Alto	Med	11	No	No	No	Sí	Alto
Alto	Med	8	No	No	No	No	Med



Representación por coordenadas paralelas:

El mayor problema de estas representaciones (y de otras muchas) es que no acomodan bien las variables discretas.

Visualización

Visualización Posterior:

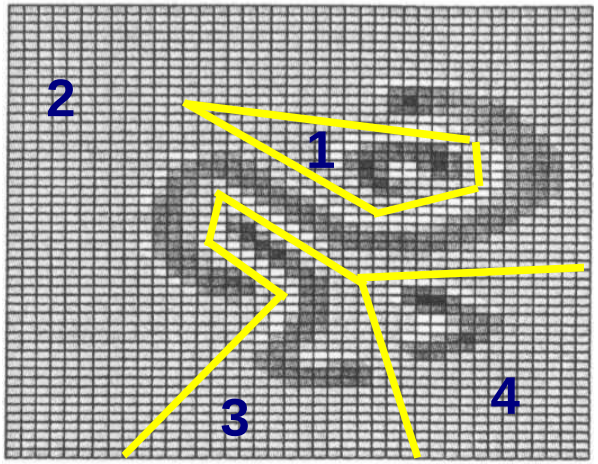
Se utiliza para mostrar los patrones y entenderlos mejor.

- Un árbol de decisión es un ejemplo de visualización posterior.
- Otros gráficos de visualización posterior de patrones:
  - muestran una determinada segmentación de los datos, una asociación, una determinada clasificación.
  - utilizan para ello gráficos de visualización *previa* en los que además se señala el patrón.
  - permiten evaluar gráficamente la calidad del modelo.

Visualización

Visualización Posterior:

EJEMPLO: se muestra una segmentación lineal para el corte del ejemplo anterior:

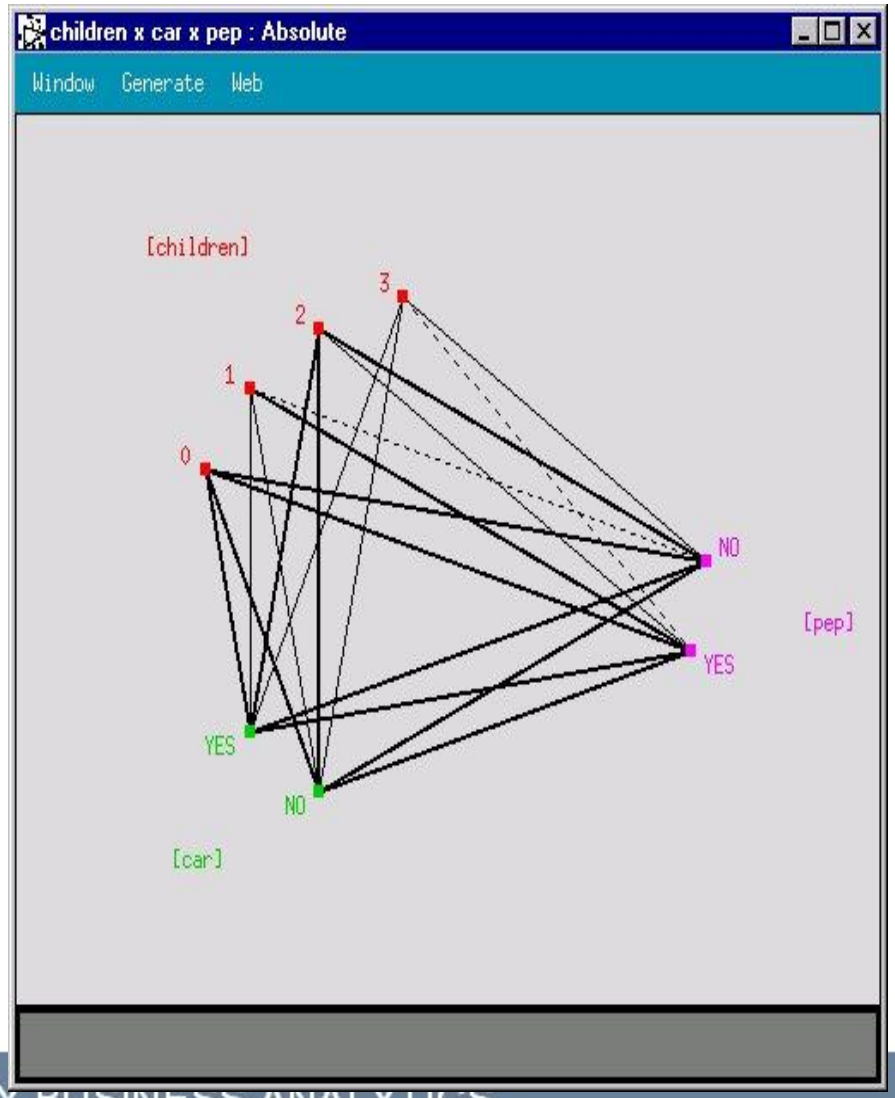


Visualización

Visualización Posterior:

EJEMPLO:

se muestra el grado de asociación según la línea que conecta los valores (continua gruesa, continua, discontinua o inexistente):



Visualización

Visualización Posterior:

EJEMPLO:

representación de ganancias acumulativas de un árbol de decisión:

lift º = arcsen No/Total

El árbol óptimo sería así: