Census
Data Analysis & Data Mining

Data Mining

María del Rosario Bruera

IBM Scholars Program

Census
Data Analysis & Data Mining

Preguntas y respuestas

Preguntas:

- ¿Cuál es el valor de los clientes?
- ¿Cuáles son los clientes que tienen mayor probabilidad de desertar?
- ¿Cuáles son los productos que se venden en forma conjunta? ….

Respuestas:

- Están en los datos del usuario
- Se necesitan herramientas especiales para encontrarlas

Census
Data Analysis & Data Mining

Business Intelligence

- "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

Census
Data Analysis & Data Mining

B.I.: recursos y herramientas

- Fuentes de datos : warehouses, data marts, etc
- Herramientas de administración de datos
- Herramientas de extracción y consulta
- Herramientas de modelización (Data Mining)

Census
Data Analysis & Data Mining

¿Qué es Data Mining? (1997)

- **Data Mining** : es el proceso de exploración y análisis - de manera automática o semiautomática - de los datos para obtener patrones significativos y reglas de negocio.

  - Michael Berry, Gordon Linoff
    Data Mining for marketing,sales and customer support.
    Wiley. USA, 1997

Census
Data Analysis & Data Mining

Reflexiones (2000)

- .. nos gusta la noción de que los patrones deben ser significativos …
- .. Si hay algo que rechazamos es la frase "por medios automáticos o semiautomáticos", no porque no sea cierto - sin automatización es imposible minar grandes cantidades de datos - sino porque entendemos que se ha puesto demasiado énfasis en la automatización y no suficiente en las etapas de exploración y análisis
- .. Data Mining es un proceso

Michael Berry, Gordon Linoff
Mastering Data Mining. Wiley. USA 2000

Census
Data Analysis & Data Mining

Qué NO es Data Mining

- 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.

Census
Data Analysis & Data Mining

Pilares del proceso de Data Mining

- Datos
- Algoritmos y técnicas
- Prácticas de modelización.

Census

Data Analysis & Data Mining

Disciplinas que se integran

- Inteligencia Artificial
- Estadística
- Tecnologías de soporte de decisiones (OLTP)
- Tecnologías de hardware y software

Census

Data Analysis & Data Mining

Perspectiva histórica

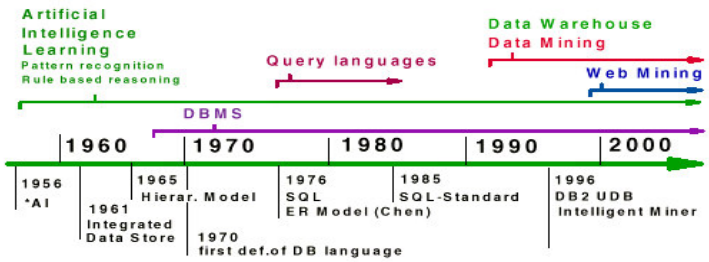


Figure 3-1 A historical view of data mining

Fuente: Mining your Own Business Data Using DB2 Intelligent Miner for Data

Census

Data Analysis & Data Mining

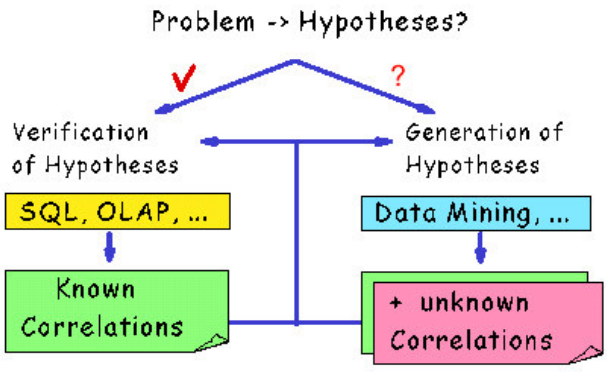


Figure 3-2 Standard and data mining approach on information detection

Fuente: Mining your Own Business Data Using DB2 Intelligent Miner for Data

Census

Data Analysis & Data Mining

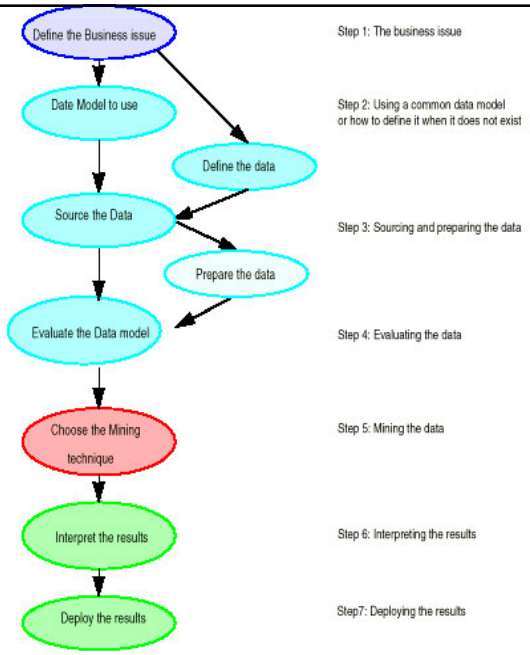
Etapas en el proceso de Data Mining

- Identificar el problema de negocio
- Transformar los datos en información
- Actuar a partir de los resultados
- Medir los resultados de las acciones

Census

Data Analysis & Data Mining

The Mining Process



Fuente: Mining your Own Business Data Using DB2 Intelligent Miner for Data

Census

Data Analysis & Data Mining

IBM Business Partner

El Analista de Datos

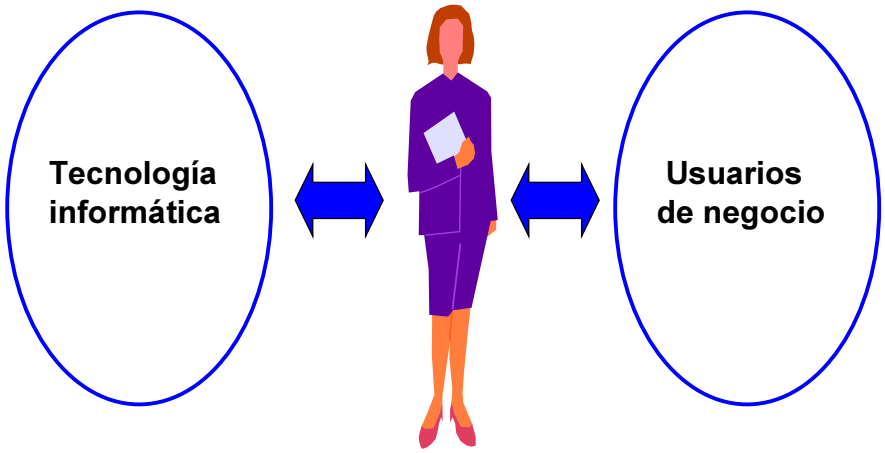
- Es el vínculo entre las áreas de tecnología informática y las áreas de negocios
- Traduce los requerimientos de información en preguntas apropiadas para su análisis con las herramientas de minería.
- Realimenta el Data Warehouse de la compañía con nuevos criterios de data cleaning y data validation.

Census

Data Analysis & Data Mining

IBM Business Partner

El Analista de Datos

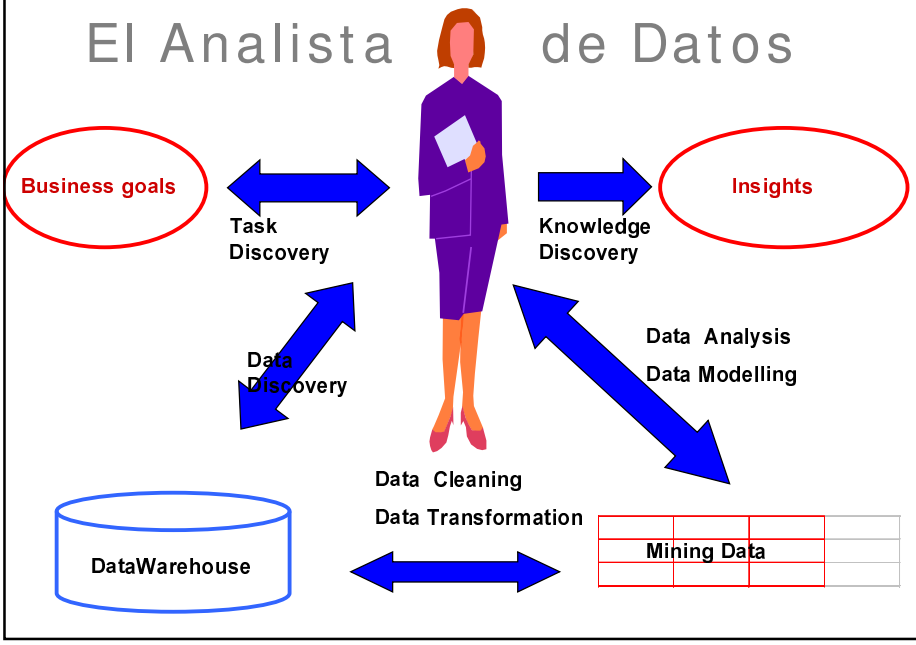


Census

Data Analysis & Data Mining

IBM Business Partner

El Analista de Datos



Census

Data Analysis & Data Mining

Habilidades requeridas

- Data manipulation (SQL)
- Conocimiento de las técnicas de minería y análisis exploratorio
- Habilidad de comunicación (interpretación) de los problemas de negocio
- Creatividad

Census

Data Analysis & Data Mining

Data Mining Team

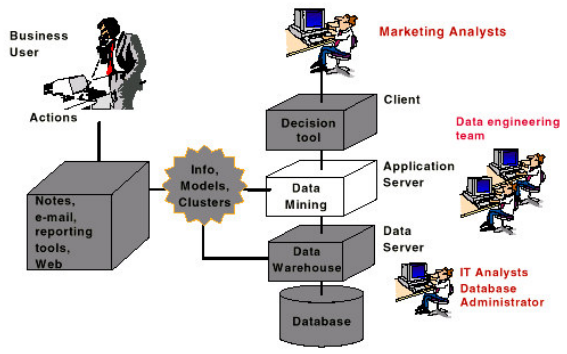


Figure 3-10 The data mining team

Fuente: Mining your Own Business Data Using DB2 Intelligent Miner for Data

Census

Data Analysis & Data Mining

Costos de proyecto

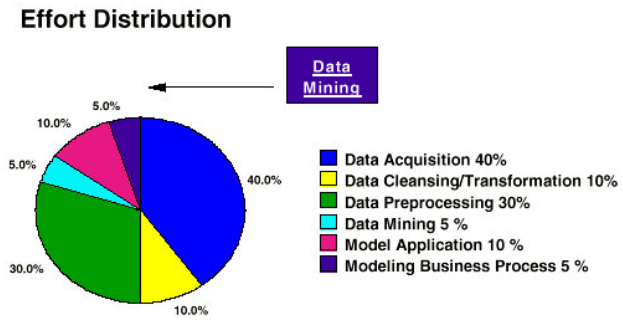


Figure 3-11 Effort distribution

Fuente: Mining your Own Business Data Using DB2 Intelligent Miner for Data

Census

Data Analysis & Data Mining

Origen de los datos

- Bases de Datos Relacionales
- Data Warehouses
- Data Marts and OLAP
- Otros formatos : Excel, archivos ASCII, encuestas, datos censales, etc.

Census
Data Analysis & Data Mining

Tipos de fuentes de datos

- Transaccionales: ej. las operaciones realizadas con tarjeta de crédito
- Relacionales : ej. la estructura de los productos que ofrece el Banco
- Demográficos: ej. características del grupo familiar

Census
Data Analysis & Data Mining

La forma de los datos para Data Mining

- Se organizan en forma de una tabla plana compuesta por filas y columnas.
- Las Filas : unidad de análisis.Por ejemplo: una cuenta, un ticket
- Las columnas : los atributos de cada unidad de análisis.Por ejemplo : frecuencia de uso de la tarjeta de crédito

Census

Data Analysis & Data Mining

Características de las tablas de datos para Data Mining

- **Todos los datos deben estar en una sola tabla o "vista" de la Base de Datos**
- **Cada fila debe corresponder a una instancia relevante al negocio**
- **Las Columnas sin variabilidad deben ser ignoradas**
- **Las Columnas con valores únicos para cada caso deben ser ignoradas (Nro de cuenta)**

Census

Data Analysis & Data Mining

La calidad de los datos

- **El éxito de las actividades de** Data Mining **se relaciona directamente con la** CALIDAD **de los datos.**
- **Se debe identificar los datos faltantes** "missings" **o fuera de rango** "outliers".

Census

Data Analysis & Data Mining

La calidad de los datos

- Muchas veces resulta necesario pre-procesar los datos antes de derivarlos al modelo de análisis. El pre-procesamiento puede incluir transformaciones, reducciones o combinaciones de los datos.
- La semántica de los datos debe ayudar para la selección de una conveniente representación y las bondades de la representación elegida gravitan directamente sobre la calidad del modelo y de los resultados posteriores.

Census

Data Analysis & Data Mining

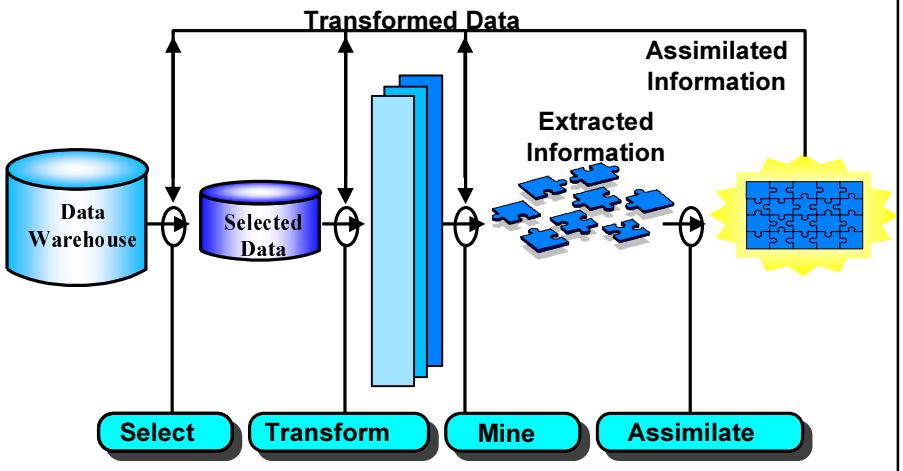
Problemas con los datos

- Demasiados datos
  - datos corruptos o con ruido
  - datos redundantes (requieren factorización)
  - datos irrelevantes
  - excesiva cantidad de datos (muestreo)
- Pocos datos
  - atributos perdidos (missings)
  - valores perdidos
  - poca cantidad de datos
- Datos fracturados
  - datos incompatibles
  - múltiples fuentes de datos

Census

Data Analysis & Data Mining

Preparación de los datos



Census

Data Analysis & Data Mining

Data Warehouse

- Data Warehouse is a subject-oriented, integrated, time-variant, non volatile collection of data in support of management decisions

  Bill Inmon (1990)

- A copy of transaction data specifically structured for query and analysis.

  Ralph Kimball

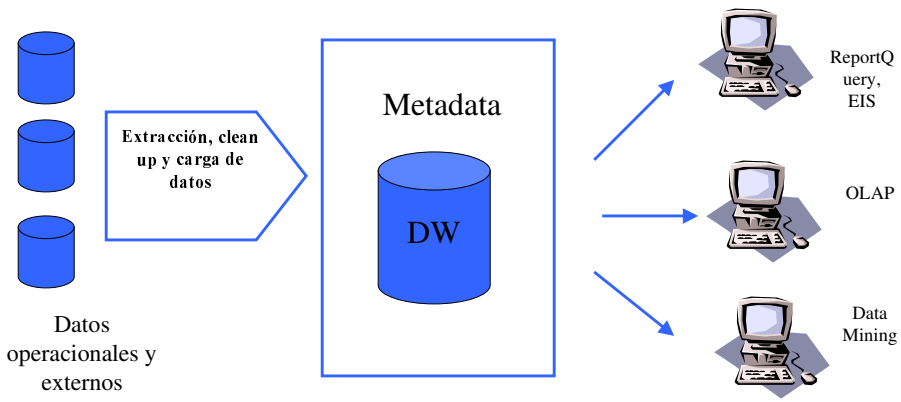
Census
Data Analysis & Data Mining

Data Marts

- Técnicamente es un subconjunto del DW orientado a una finalidad específica de negocio : marketing, finanzas, producción, etc
- El término se utiliza también para identificar soluciones alternativas a un DW corporativo más reducidas y de menor costo y tiempo de implantación.

Census
Data Analysis & Data Mining

Arquitectura del Datawarehouse



Census
Data Analysis & Data Mining

Herramientas de explotación del DW

- Herramientas de visualización
- Reporting
- OLAP
- Data Mining

Census
Data Analysis & Data Mining

OLAP

- On Line Analytical Processing
- Permiten la elaboración de vistas multidimensionales del DW para optimizar performance
- Están soportadas por motores de administración del DW que admiten la construcción de estos "cubos"

Census
Data Analysis & Data Mining

OLAP

- Herramientas útiles y poderosas para acceder a Bases de Datos y Data Warehouses y obtener "reportes" de información.
- La tecnología OLAP commplementa las actividades de Data Mining y supera las posibilidades del SQL

Census
Data Analysis & Data Mining

Data Mining y OLAP

- Las herramientas de reporting, OLAP y consulta responden efectivamente para la construcción de modelos descriptivos y retrospectivos para confirmar o rechazar hipótesis previas del usuario

Census
Data Analysis & Data Mining

Data Mining y OLAP

- Las herramientas de Data Mining permiten encontrar patrones no evidentes en los grandes volúmenes de información del DW y proponer modelos predictivos

Census
Data Analysis & Data Mining

Qué es la Estadística

- Es la disciplina que extrae información general a partir de datos específicos.
- Es el estudio de la estabilidad en la variación
- Es el arte de examinar, sumarizar y extraer conclusiones a partir de los datos.

Census
Data Analysis & Data Mining

Data Mining y Estadística

- Los métodos estadísticos son el corazón de muchas de las técnicas de minería de datos.
- Originalmente muchas de estas técnicas fueron diseñadas con propósitos confirmatorios.
- La estadística exploratoria aparece en los 70 con los aportes de J.Tuckey

Census
Data Analysis & Data Mining

Data Mining y Estadística

- En la Minería de Datos no se hacen supuestos a priori sobre la naturaleza de las variables y de las relaciones entre ellas (normalidad, linealidad, etc).
- Los algoritmos estadísticos se adaptan - para Minería de Datos - al procesamiento de grandes volúmenes de datos

Census
Data Analysis & Data Mining

Data Mining e IA

- La Inteligencia Artificial se integra a la Minería de Datos a partir de las redes neuronales artificiales
- Se utilizan para construir modelos predictivos no lineales que aprenden a través de entrenamiento y que se asimilan a los modelos de redes de neuronas biológicas.

Census
Data Analysis & Data Mining

Redes neuronales

- Las redes neuronales son adecuadas para problemas de tipo predictivo.
- Un problema apropiado para una red neuronal tiene tres características:
  - Se comprenden claramente los INPUTS
  - Se comprende claramente el OUTPUT
  - Existen ejemplos (experiencia) suficientes para entrenar a la red

Census
Data Analysis & Data Mining

Los modelos neuronales

- La red neuronal no produce reglas explícitas que describan el modelo
- Un modelo neuronal es tan bueno como lo es el set de datos usado para entrenar la red
- El modelo es estático y debe ser explícitamente actualizado agregando ejemplos recientes y re-entrenando la red para asegurar su vigencia y utilidad

Census
Data Analysis & Data Mining

Los modelos neuronales

- Con modelos neuronales se puede atacar una gran variedad de problemas y producir buenos resultados aún en dominios complejos con variables continuas y categóricas
- Son apropiados para tareas de clasificación y predicción cuando los resultados del modelo son más importantes que comprender cómo funciona el modelo.

Census
Data Analysis & Data Mining

Customer Relationship Management

- Es el proceso que administra la relación entre la compañía y sus clientes
- Para que resulte exitoso resulta necesario identificar los patrones de consumo y comportamiento de los clientes

Census
Data Analysis & Data Mining

Data Mining - CRM

- Data Mining se utiliza para sistematizar los procesos de búsqueda de los predictores de comportamiento de los clientes en las etapas de diseño de campañas
- También se aplica para la medición de los resultados de la campaña y la realimentación del CRM

Census
Data Analysis & Data Mining

Problemas típicos de Data Mining

- Clasificación
- Estimación
- Predicción
- Agrupamiento a partir de reglas de asociación
- Clustering
- Descripción y visualización, etc

Census
Data Analysis & Data Mining

Problema de Clustering

Agrupar a los clientes según sus indicadores R(Recency), F(Frecuencia), M (Monto), etc en segmentos de comportamiento homogéneo.

Resultado : Clientes Heavy, Medium, Light, etc

El 78% de la facturación se concentra en el cluster Heavy (27% de los clientes).

Los clientes Heavy son casados, con hijos, trabajadores autónomos con un ingreso superior a $3800.

Census

Data Analysis & Data Mining

Problema de Clasificación

Clasificar un nuevo cliente - de acuerdo a su perfil sociodemográfico - como potencial cliente Heavy, Medium, Light.

Census

Data Analysis & Data Mining

Problema de Estimación

Estimar el consumo de un determinado rubro de artículos de un grupo clientes en el próximo trimestre.

Estimar el LTV (Life Time Value) potencial de un nuevo cliente

Census

Data Analysis & Data Mining

Problema de Predicción

Predecir el abandono de un cliente (churning, atrittion)

. Para una compañía de telefonía celular

. Para una AFJP

. Para una tarjeta de crédito

Census

Data Analysis & Data Mining

Problema de Asociación

Encontrar las reglas que determinan el cross - traffic entre productos para los clientes de un Banco. Por ejemplo:

"Cuando un cliente se activa en Caja de Ahorros , el siguiente producto en donde se activa es Préstamos personales . Este patrón ocurre en el 65% de los casos."

Census
Data Analysis & Data Mining

Problema de visualización

Representar mediante un software de geolocalización (GIS) la distribución de los clientes en la zona de influencia de las sucursales de un comercio.

Census
Data Analysis & Data Mining

Problemas usuales

- Caracterización de perfiles de clientes para definir acciones de Up selling y Cross selling
- Tracking de campañas y predicción de respuesta / no respuesta
- Canasta de consumo de tarjetas de crédito y prevención de fraudes
- Modelos de predicción de abandono

Census
Data Analysis & Data Mining

Problemas usuales

- Programas de millaje y fidelización
- Consolidación de Bases de Datos propias con fuentes externas
- Web mining y análisis de tráfico y uso de recursos de e-business
- Definición de marcos muestrales para investigaciones de mercado y encuestas de customer satisfaction.

Census
Data Analysis & Data Mining

La elección del modelo para Data Mining

- 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

Census

Data Analysis & Data Mining

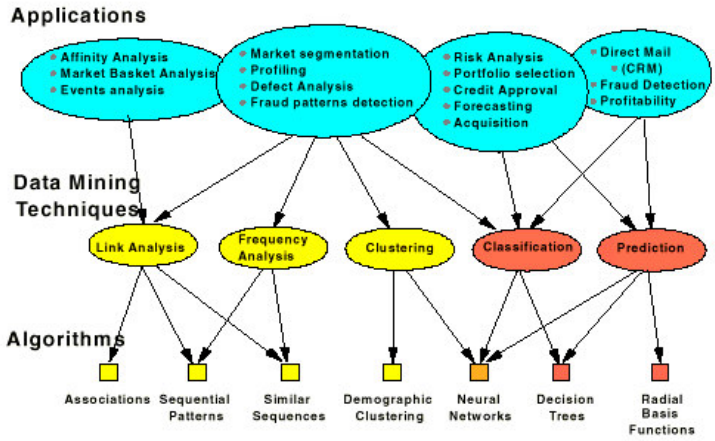


Figure 3-3 From applications to algorithms

Fuente: Mining your Own Business Data Using DB2 Intelligent Miner for Data

Census

Data Analysis & Data Mining

Cómo seleccionar una potencial aplicación de DM

Consideraciones prácticas:

- **Potencial impacto significativo (Relación costo - beneficio).**
- **No hay otra alternativa**
- **Existe soporte institucional**
- **No existen impedimentos legales de uso de la información**

Census
Data Analysis & Data Mining

IBM Business Partner

Cómo seleccionar una potencial aplicación de DM

Consideraciones técnicas:

- Disponibilidad suficiente de datos
- Relevancia de atributos
- Bajos niveles de ruido en los datos
- Precisar el nivel de confianza para los resultados
- Conocimiento anterior existente

Census
Data Analysis & Data Mining

IBM Business Partner

La evaluación de los modelos

- Cuán ajustado es el modelo?
- Es correcta su descripción de los datos observados?
- Cuanta confianza se puede tener en sus predicciones?
- Cuán comprensible es el modelo?

Census
Data Analysis & Data Mining

Las medidas

- La concordancia de un modelo predictivo con la realidad se mide con relación a la tasa de error, es decir, el porcentaje de casos clasificados o cuya predicción fue incorrecta.
- Para ello se dispone de datos de validación y testing sobre los que debe aplicarse periódicamente el modelo a modo de control.

Census
Data Analysis & Data Mining

Las medidas

- En el caso de los modelos descriptivos, una buena regla es la que proporciona la información más comprensible con la menor "longitud" de expresión de la regla.
- En definitiva, la medida más importante de efectividad es el retorno de la inversión

Census
Data Analysis & Data Mining

Un proyecto exitoso

- Un único project leader
- Un equipo multidisciplinario integrado por personas de las áreas de IT y de negocio
- Las unidades de negocio están involucradas desde el comienzo
- El área de IT está involucrada desde el comienzo
- Un pequeño proyecto piloto que muestre las ventajas de Data Mining

Census
Data Analysis & Data Mining

Las nuevas tecnologías

Census

Data Analysis & Data Mining

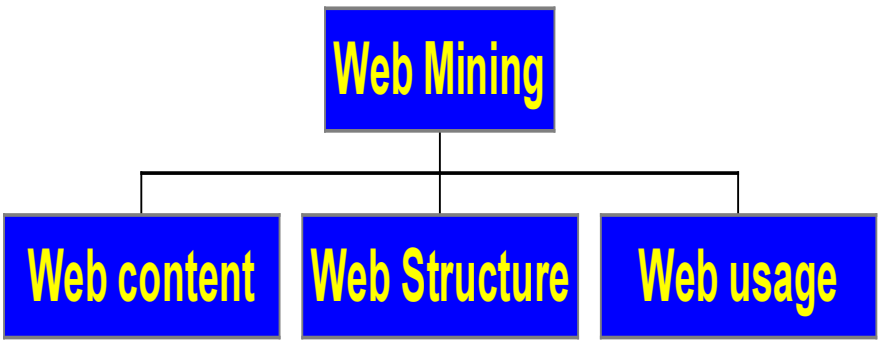
Web Mining

- Es el descubrimiento de patrones significativos a partir del análisis de la estructura, contenidos y uso de la Web

Census

Data Analysis & Data Mining

Web Mining Taxonomy



Census
Data Analysis & Data Mining

Resultados Web mining

- El 60% de los visitantes que acceden a www.ibm.com/redbooks acceden a:

  www.ibm.com/software/data/iminer/fordata/

- Entry and Exit points

Census
Data Analysis & Data Mining

Resultados Web mining

- Link analysis y patrones secuenciales de enlaces de páginas
- Segmentación de clientes de e-commerce
- Canasta de productos
- etc, etc, etc

Census
Data Analysis & Data Mining

Text Mining

- Son nuevas herramientas destinadas a extraer información de documentos "no estructurados", organizarlos, segmentarlos, indexarlos.

Census
Data Analysis & Data Mining

Problemas de Text Mining

- Direccionamiento automático de emails según su contenido
- Clasificación automática de documentos de una intranet
- Búsqueda de información en documentos de distintos idiomas simultáneamente.

Census
Data Analysis & Data Mining

Problemas de Text Mining

- Análisis de contenidos de páginas Web
- Organización de servicios de búsqueda en la Web
- Extracción de conceptos de síntesis en documentos referidos al mismo asunto.

Census
Data Analysis & Data Mining

Conclusiones

Census
Data Analysis & Data Mining

Para qué Minería de Datos

- La Minería de Datos es una herramienta eficaz para dar respuesta preguntas complejas de Inteligencia de Negocios
- Las herramientas disponibles permiten automatizar parte de la tarea de encontrar los patrones de comportamiento ocultos en los datos
- Pero ….

Census
Data Analysis & Data Mining

Qué no puede automatizarse (todavía)

- La elección de los problemas de negocio candidatos para tareas de Data Mining
- La identificación y recolección de los datos que contienen la información buscada
- El masajeo y tratamiento de los datos que posibilita la búsqueda de patrones
- El diseño y cálculo de variables derivadas

Census
Data Analysis & Data Mining

Qué no puede automatizarse (todavía)

- El **plan de acciones** que apoyándose en los resultados del modelo produzca el ROI
- La **medición del éxito** de las acciones realizadas a partir de los resultados proporcionados por Data Mining

Census
Data Analysis & Data Mining

Conclusiones

- Convierta a Data Mining en una parte de su proyecto de negocio.
- Incluya a Data Mining en la "cultura" de su organización.

Census

Data Analysis & Data Mining

Ejemplos con DB2 Intelligent Miner for Data

Census

Data Analysis & Data Mining

Técnicas utilizadas

- Clustering (segmentación)
- Canasta de productos
- Arbol de decisión
- Red neuronal como modelo predictivo

Census
Data Analysis & Data Mining

¿Qué es "clustering"?

- Es la partición del conjunto de individuos en subconjuntos lo más homogéneos posibles.
- El objetivo es maximizar la similitud de los individuos del cluster y maximizar las diferencias entre clusters.

Census
Data Analysis & Data Mining

Aplicaciones de la técnica

- Segmentación de la base de datos
- Detección de fraudes
- Detección de defectos

Census
Data Analysis & Data Mining

Objetivos

- 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

Census
Data Analysis & Data Mining

Medidas de similaridad

- Variables categóricas (escalas nominales y ordinales) : son similares si son iguales.
- Variables numéricas (escalas métricas) : el algoritmo determina su diferencia expresada en unidades de desviaciones standard.

Census

Data Analysis & Data Mining

Ejemplo similaridad

Nombre	Sexo	Est. Civil	Lugar	Similaridad
Juan	M	C	Cap.Fed	0.33
Maria	F	C	GBA	0.33
No evaluado	Diferente	Igual	Diferente	

Census

Data Analysis & Data Mining

Criterio Condorcet

- Es una medida de similaridad que varía entre 0 y 1
- Vale 0 : los individuos están ubicados aleatoriamente en los clusters
- Vale 1 : Todos los individuos de los clusters son idénticos y no hay individuos con esas características fuera de cada cluster.
- Condorcet mínimo usual = 0.65

Census

Data Analysis & Data Mining

El problema

Se trata de segmentar la Base de Datos de los clientes de una tarjeta de crédito a partir de sus indicadores de consumo para identificar al segmento de mayor valor.

Census

Data Analysis & Data Mining

Los datos disponibles

- **A partir de la Base de Datos de transacciones del último año de los clientes se obtienen como variables:**
  - **Frecuencia de uso de la tarjeta** : calculada como media de días entre transacciones.
  - **Saldo promedio mensual de transacciones en $$**
  - **Monto promedio por transacción**
  - **Cantidad de servicios por débito automático**
  - **Datos sociodemográficos : sexo, edad, estado civil, ocupación, hijos**

Census
Data Analysis & Data Mining

La preparación de datos

- Definir la unidad de análisis : ¿cuenta o tarjeta?
- Definir qué es una transacción, ej.:¿cómo se consideran los ajustes (montos negativos)?
- Definir variables derivadas : en la frecuencia ¿cómo intervienen los débitos automáticos?

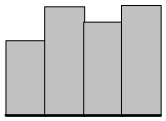
Census
Data Analysis & Data Mining

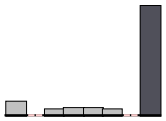
La preparación de datos

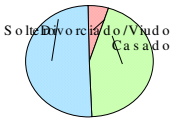
- Describir las variables a incluir en el modelo para:
  - Calcular medidas de posición y dispersión
  - Identificar distribuciones asimétricas
  - Identificar missings
  - Identificar valores incorrectos o fuera de rango
  - Identificar outliers

Census
Data Analysis & Data Mining

Descriptivos generales

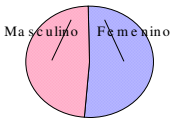

scios


edad

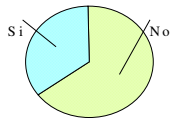


estado_civil

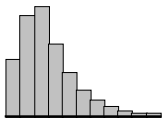
ocup

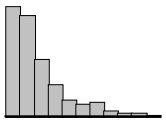


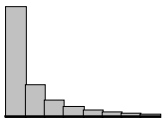
sexo



hijos


avgtckt

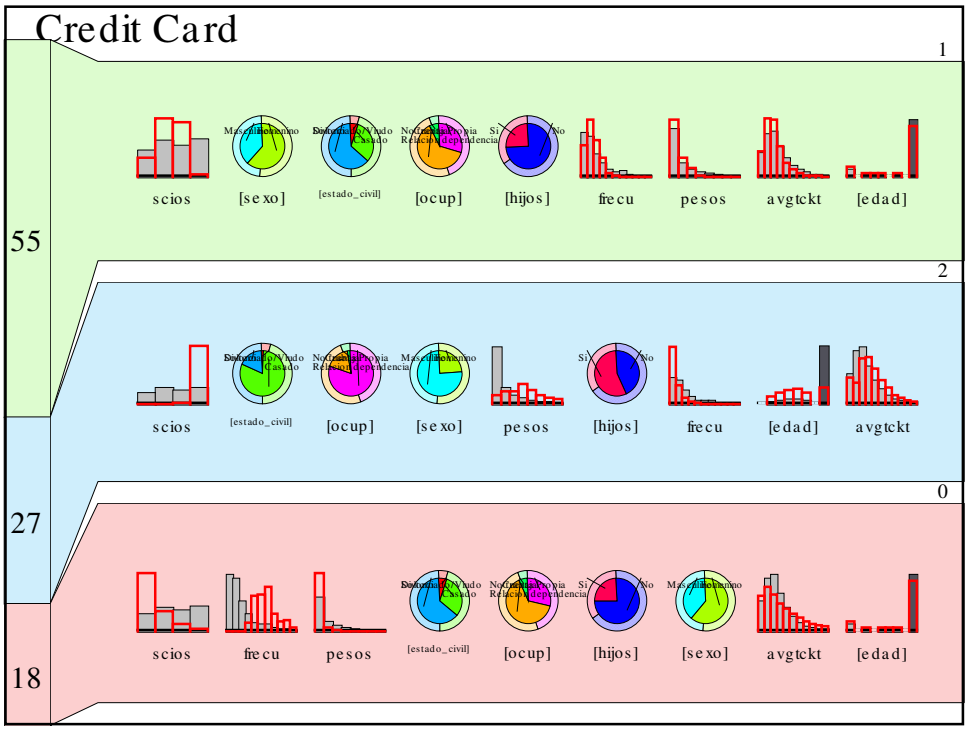

frecu

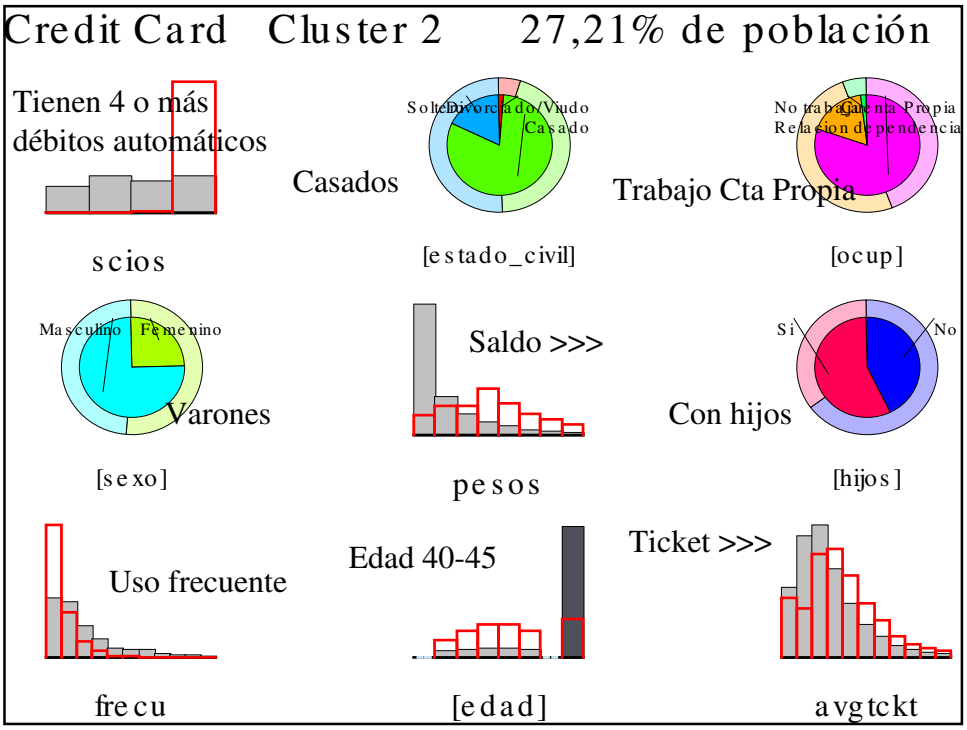

pesos

Census
Data Analysis & Data Mining

Criterios de segmentación

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

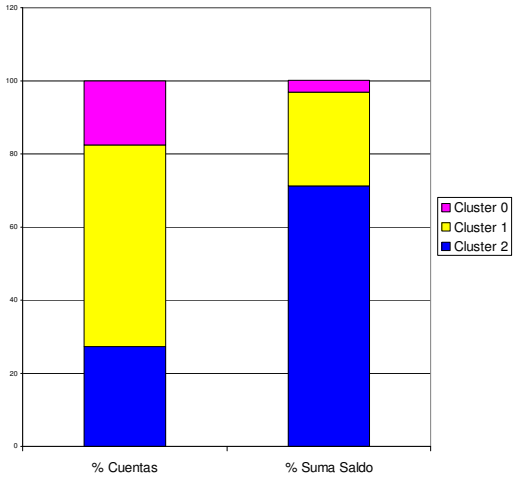

Credit Card
1
55
scios
[sexo]
[estado_civil]
[ocup]
[hijos]
frecu
pesos
avgtckt
[edad]
2
27
scios
[estado_civil]
[ocup]
[sexo]
pesos
[hijos]
frecu
[edad]
avgtckt
0
18
scios
frecu
pesos
[estado_civil]
[ocup]
[hijos]
[sexo]
avgtckt
[edad]


Credit Card Cluster 2 27,21% de población
Tienen 4 o más débitos automáticos
scios
Casados
Soltero/Divorciado/Viudo
Casado
[estado_civil]
Trabajo Cta Propia
No trabaja/Jubilado
Propia
Relacion de dependencia
[ocup]
Masculino
Femenino
Varones
[sexo]
Saldo >>>
pesos
Con hijos
Si
No
[hijos]
Uso frecuente
frecu
Edad 40-45
[edad]
Ticket >>>
avgtckt

Census

Data Analysis & Data Mining

Pareto



Census

Data Analysis & Data Mining

Arboles de decisión

- Son técnicas que se utilizan con finalidad predictiva y de clasificación.
- Se obtiene como resultado "reglas" que explican el comportamiento de una variable (TARGET) con relación a otras (PREDICTORAS).
- En este ejemplo se utilizan para "explicar" los clusters.

Census

Data Analysis & Data Mining

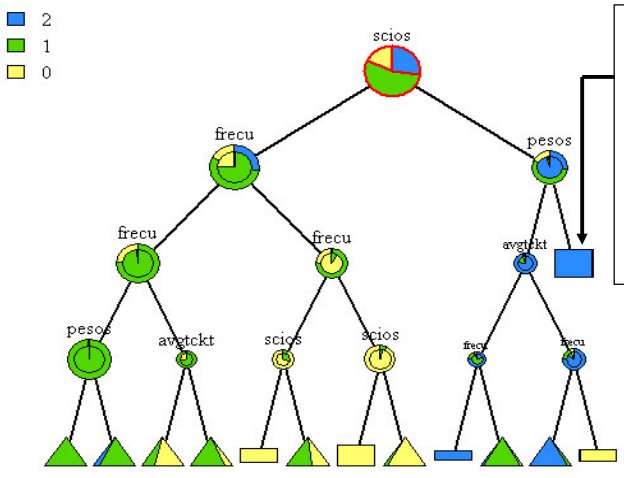
Algoritmos

- CHAID (Chi Squared Automatic Detection).
- C&RT : Classification and Regression Tree.
- C4.5, Quest y otros
- Intelligent Miner utiliza una variante de C&RT

Census

Data Analysis & Data Mining

Arbol de comportamiento

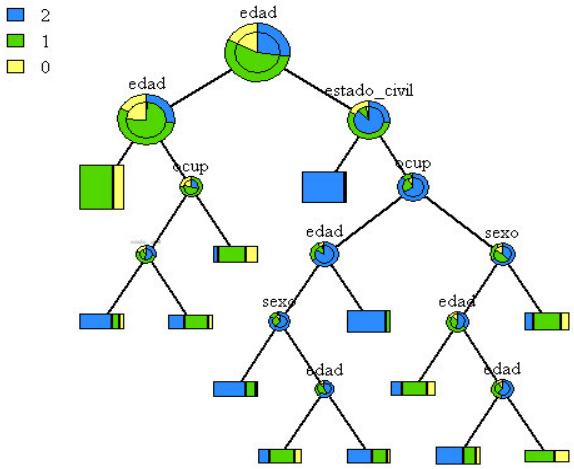


Si tiene 4 o más débitos automáticos y un saldo > $ 727 entonces su probabilidad de pertenecer al cluster 2 es del 99%

Census

Data Analysis & Data Mining

Arbol sociodemográfico



Census

Data Analysis & Data Mining

Market Basket Analysis

- El problema: Se trata de encontrar las reglas de asociación que organizan los pedidos de "toppings" extra de una pizzería, a partir del análisis de un conjunto de 2000 tickets de venta.

Census
Data Analysis & Data Mining

La tabla de Data Mining

- Id ticket
- Código de producto:
  - 1 Hongos
  - 2 Pepperoni
  - 3 Queso
  - 4 Cerveza
  - 5 Gaseosa
  - 6 Otra bebida

Census
Data Analysis & Data Mining

Propósito de MBA

- Generar reglas del tipo:
  - IF (SI) condición ENTONCES (THEN) resultado
- Ejemplo:
  - Si producto A y producto C ENTONCES producto B

Census
Data Analysis & Data Mining

Tipos de reglas

- **Utiles / aplicables** : reglas que contienen buena calidad de información que pueden traducirse en acciones de negocio.
- **Triviales** : reglas ya conocidas en el negocio por su frecuente ocurrencia
- **Inexplicables** : curiosidades arbitrarias sin aplicación práctica

Census
Data Analysis & Data Mining

Problemas del MBA

- La existencia de muchos items en el set de análisis complica exponencialmente el tiempo de cálculo
- Resulta necesario definir criterios para seleccionar las mejores reglas
- Es importante la construcción de una taxonomía de productos

Census
Data Analysis & Data Mining

¿Cuán buena es una regla?

Medidas que califican a una regla:

- Soporte
- Confianza
- Lift (Improvement)

Census
Data Analysis & Data Mining

Soporte

Es la cantidad (%) de transacciones en donde se encuentra la regla.

- Ej: "Si A entonces B" está presente en 4000 de 10000 transacciones.
- Soporte (A/B): 40%

Census
Data Analysis & Data Mining

Confianza

- **Cantidad (%) de transacciones que contienen la regla referida a la cantidad de transacciones que contienen la cláusula condicional**
  - Ej : Para el caso anterior, si A está presente en 6000 transacciones (60%)
  - Confianza (A/B) = 40% / 60% = 66%

Census
Data Analysis & Data Mining

Mejora (Improvement)

- **Capacidad predictiva de la regla:**
  - Mejora = p(A/B) / p(A) * p(B)
  - Ej :

p(A/B) = 40% ; p(A) = 60%; p(B) = 30%

Improv (A/B) = 40% / (60% * 30%) = 2.22

Mayor a 1 : la regla tiene valor predictivo

Census

Data Analysis & Data Mining

Ejemplo de cálculo

Census

Data Analysis & Data Mining

Datos básicos

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

Census

Data Analysis & Data Mining

Reglas

			Cantidad	Soporte	Confianza	Mejora
Hongos			900	0.45		
Pepperoni			850	0.43		
Queso			800	0.40		
Hongos --> Pepperoni			500	0.25	0.56	1.31
Hongos --> Queso			400	0.20	0.47	1.18
Queso --> Pepperoni			300	0.15	0.38	0.88
Hongos + Pepperoni --> Queso			100	0.05	0.20	0.80
Hongos + Queso --> Pepperoni			100	0.05	0.25	0.59
Queso + Pepperoni --> Hongos			100	0.05	0.33	0.74

	Pueden descartarse por bajo soporte
	Reglas significativas

Census

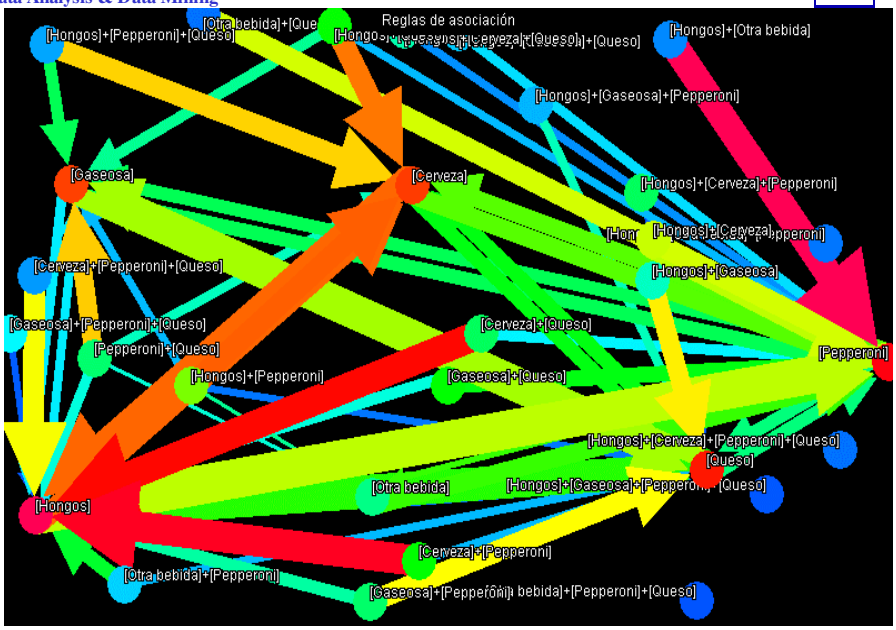
Data Analysis & Data Mining

Otro ejemplo de MBA

- La asociación se plantea entre los toppings de las pizzas y las bebidas.
- Los gráficos de reglas permiten visualmente identificar reglas con buen soporte, confianza y lift

Census

Data Analysis & Data Mining



Census

Data Analysis & Data Mining

Reglas

Soporte (%)	Confianza(%)		Tipo	Elevación	Cuerpo de regla		Cabecera de regla
3.1746	80.0000	+	1.7800	[Hongos]+[Otra bebida]	==>	[Pepperoni]
16.6667	81.8200	+	1.7200	[Cerveza]+[Pepperoni]	==>	[Hongos]
13.0688	78.4100	+	1.6500	[Cerveza]+[Queso]	==>	[Hongos]
16.6667	63.0000	.	1.5400	[Hongos]+[Pepperoni]	==>	[Cerveza]
29.8413	72.8700	+	1.5300	[Cerveza]	==>	[Hongos]
29.8413	62.6700	+	1.5300	[Hongos]	==>	[Cerveza]
13.0688	61.7500	.	1.5100	[Hongos]+[Queso]	==>	[Cerveza]
9.0476	57.0000	+	1.4000	[Pepperoni]+[Queso]	==>	[Gaseosa]
3.0159	57.0000	.	1.3900	[Hongos]+[Pepperoni]+[Queso]	==>	[Cerveza]
6.9312	56.9600	.	1.3500	[Hongos]+[Gaseosa]	==>	[Queso]
9.0476	56.4400	.	1.3300	[Gaseosa]+[Pepperoni]	==>	[Queso]

Census

Data Analysis & Data Mining

Web Mining

- El problema : se trata de analizar las transacciones y el perfil de los usuarios de un Web site de un comercio de venta por internet.

Census

Data Analysis & Data Mining

Modelos aplicados

- Asociación de páginas visitadas : canasta de productos
- Perfil de usuarios : clustering demográfico
- Potenciales compradores : árbol de decisión

Census

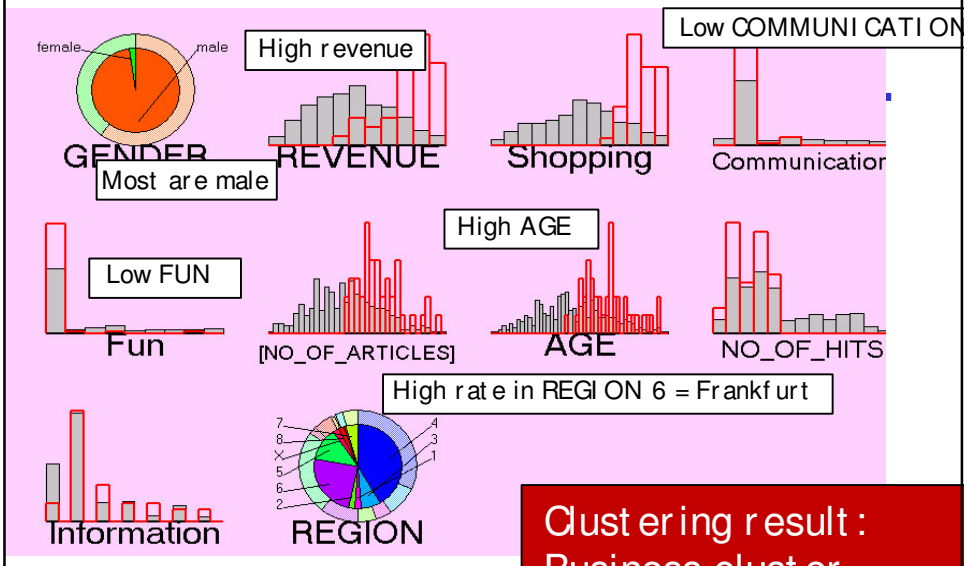
Data Analysis & Data Mining

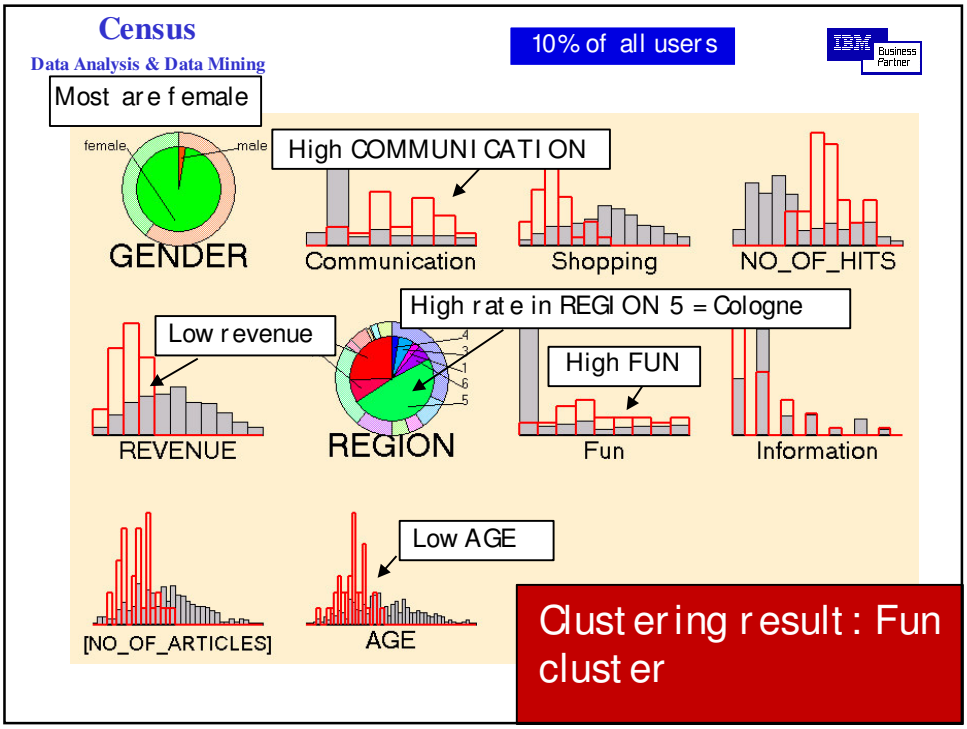
Asociación de páginas

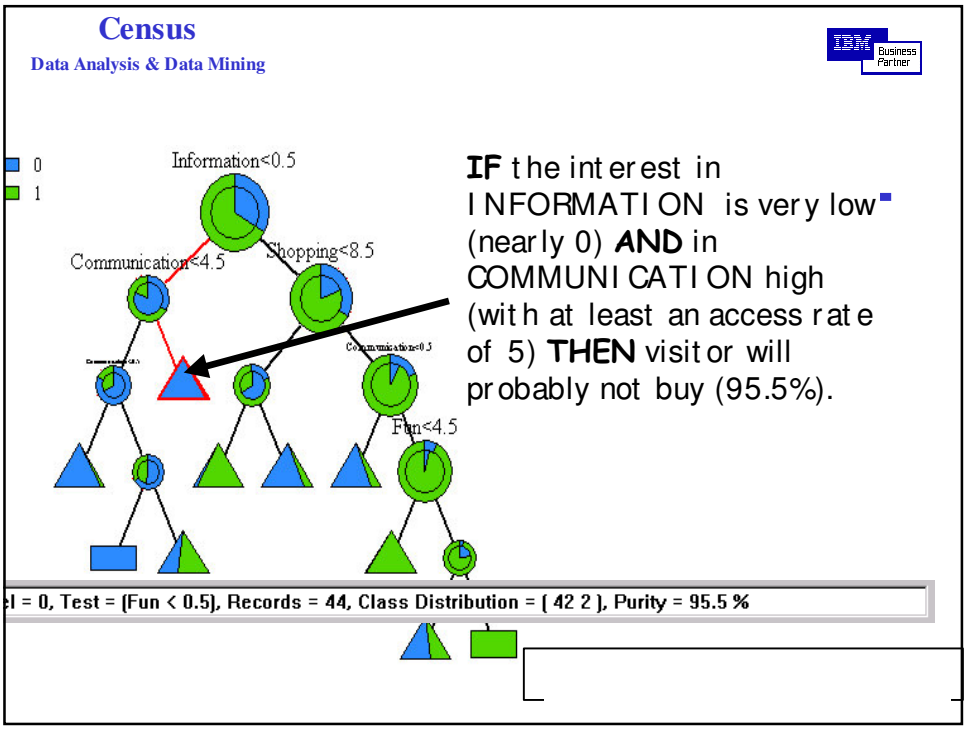
Soporte	Conf.	Elevación	Regla
11.2360	66.6700	3.1200	[food.html] ==> [chat.html]
11.2360	52.6300	3.1200	[chat.html] ==> [food.html]
10.1124	29.0300	1.2300	[womens-fashion.html]==> [music.html]
9.5506	44.7400	1.2800	[chat.html] ==>[womens-fashion.html]
9.5506	27.4200	1.2800	[womens-fashion.html]==> [chat.html]

Census

Data Analysis & Data Mining




Census
Data Analysis & Data Mining
10% of all users
Most are female
female
male
GENDER
High COMMUNICATION
Communication
Shopping
NO_OF_HITS
Low revenue
REVENUE
High rate in REGION 5 = Cologne
REGION
High FUN
Fun
Information
[NO_OF_ARTICLES]
Low AGE
AGE
Clustering result : Fun cluster


Census
Data Analysis & Data Mining
0
1
Information<0.5
Communication<4.5
Shopping<8.5
Fun<4.5
IF the interest in INFORMATION is very low (nearly 0) AND in COMMUNICATION high (with at least an access rate of 5) THEN visitor will probably not buy (95.5%).
el = 0, Test = (Fun < 0.5), Records = 44, Class Distribution = (42 2), Purity = 95.5 %

Census

Data Analysis & Data Mining

Secuencia de clicks

17.241	[gourmet.html] [email.html] [email.html]

Interpretation: in 17.2% (of all transactions) the user goes to GOURMET.html ; he then sends two emails out.

56.897	[sports.html] [chat.html] [womens-fashion.html]

Interpretation: in 56.9% (of all transactions) the user goes first SPORTS.html ; he then uses the chat as a communication medium; finally, he focus his attention to Fashion.

25.862	[womens-fashion.html] [postcards.html] [womens-fashion.html]

Interpretation: In 25.9% (of all transactions) the user goes first to womens-fashion.html ; he then sends a postcard, and goes to womens-fashion.html back again.

Census

Data Analysis & Data Mining

Detección temprana de mora

- **El problema:Se trata de identificar anticipadamente los clientes con mayor posibilidad de entrar en mora para anticipar las acciones preventivas de cobranza y recupero.**

Census

Data Analysis & Data Mining

Las soluciones posibles

- Reglas para identificar a los segmentos de clientes con mayor propensión a mora
- Scoring de riesgo de morosidad

Census

Data Analysis & Data Mining

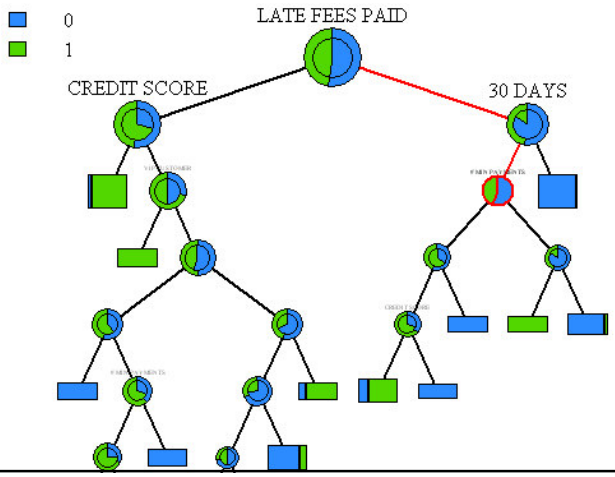
Modelos aplicables

- Para las reglas : árbol de clasificación.
- Para el scoring : modelo neuronal

Census

Data Analysis & Data Mining

Arbol muestra 50/50

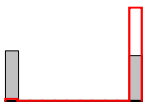


Census

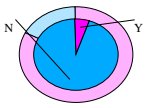
Data Analysis & Data Mining

Morosos

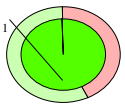
Mora 60 dias Región 90-98 9,39% de población



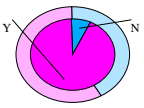
MORA 60



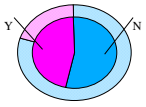
VIP CUSTOMER



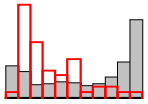
LATE FEES PAID



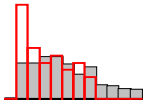
30 DAYS



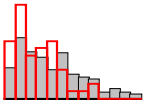
OVER CREDIT LIMIT



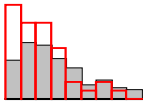
CREDIT SCORE



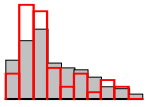
CUSTOMER AGE



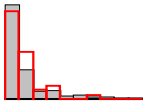
CREDIT LIMIT



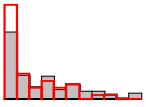
INCOME



MEMBER (MONTHS)



PURCHASES/WEEK



CASH LIMIT

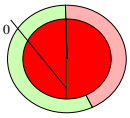
Census

Data Analysis & Data Mining

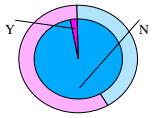
No Morosos

Mora 60 dias Región 0-2 6,81% de población

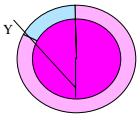
MORA 60



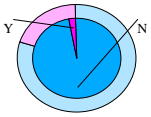
LATE FEES PAID



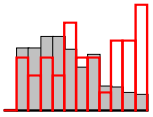
30 DAYS



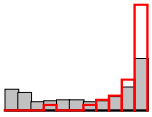
VIP CUSTOMER



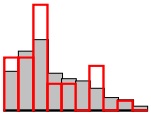
OVER CREDIT LIMIT



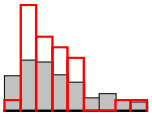
CUSTOMER AGE



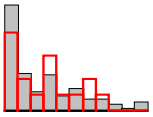
CREDIT SCORE



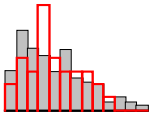
MEMBER (MONTHS)



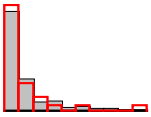
INCOME



CASH LIMIT



CREDIT LIMIT

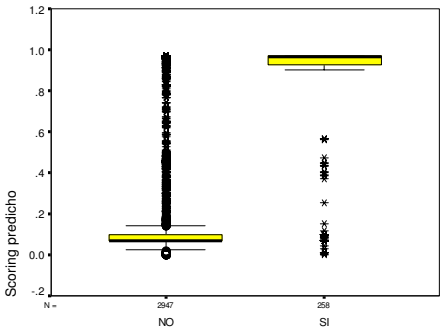


PURCHASES/WEEK

Census

Data Analysis & Data Mining

Verificación



Mora real

El scoring que predice la red está netamente diferenciado para morosos y pagadores

Referencias

- Data Mining Techniques for Marketing, Sales and Customer Support. Michael Berry, Gordon Linoff. Wiley, USA, 1997
- Data Mining with Neural Networks. Joseph Bigus. Mc Graw Hill, USA, 1996
- Data Mining: a hands-on approach for business professionals. Robert Groth. Prentice Hall, USA, 1998
- Mastering Data Mining. Michael Berry, Gordon Linoff. Wiley USA, 2000

Census
Data Analysis & Data Mining

Referencias

- Data preparation for Data Mining. Dorian Pyle. Morgan Kaufmann Publishers Inc. San Francisco, USA ,1999
- Análisis Multivariante.Hair, Anderson, Tatham, Black. Prentice Hall. Madrid, 1999
- Building Data Mining applications for CRM. A. Berson, S. Smith, K. Thearling. Mc Graw Hill, 2000

Census

Data Analysis & Data Mining

Referencias

- **IBM:**
  - www.ibm.com/software/data/iminer/fordata/
  - www.dmg.org
  - www.ibm.com/redbooks
- **The Data Mine (www.the-data-mine.com)**
- **KDD Mine (www.kdnuggets.com)**
- **chb@census.com.ar**