Implantación de un proyecto de Knowledge Center con una herramienta comercial (Synera)

Alumno : Valentina Luzón Calderón
ETIG
Consultor : Ramón Carihuelas

Barcelona, 18 de Junio del 2004

Indice:

1. - Plan de Trabajo

Proyecto :

- En una empresa los datos se pueden transformar en conocimiento, basándome en esta premisa una empresa comercial de venta de productos tangibles para realizar expositores, tiendas, etc., me ha encargado que le demuestre, como puedo mediante una herramienta comercial, convertir sus datos en conocimiento que les sirva para tomar decisiones respecto a su política comercial, marketing y de distribución del producto. Así como demostrar que la implantación de este proyecto en la empresa repercutirá positivamente en un futuro.

Herramienta Comercial:

- Usare el programa Synera Intelligent Exploration Suite

Objetivo General:

- Conseguir a través del estudio y del análisis de los datos dados por la empresa comercial, utilizando las técnicas de minería de datos, patrones validos, útiles y comprensibles para llevar a cabo el proyecto. Es decir la extracción de conocimiento útil en el ámbito comercial y de marketing de los datos, así como demostrar a la empresa comercial que si posee en el futuro este tipo de herramientas sus decisiones podrán ser tomadas más rápidamente y basándose en la realidad del mercado.

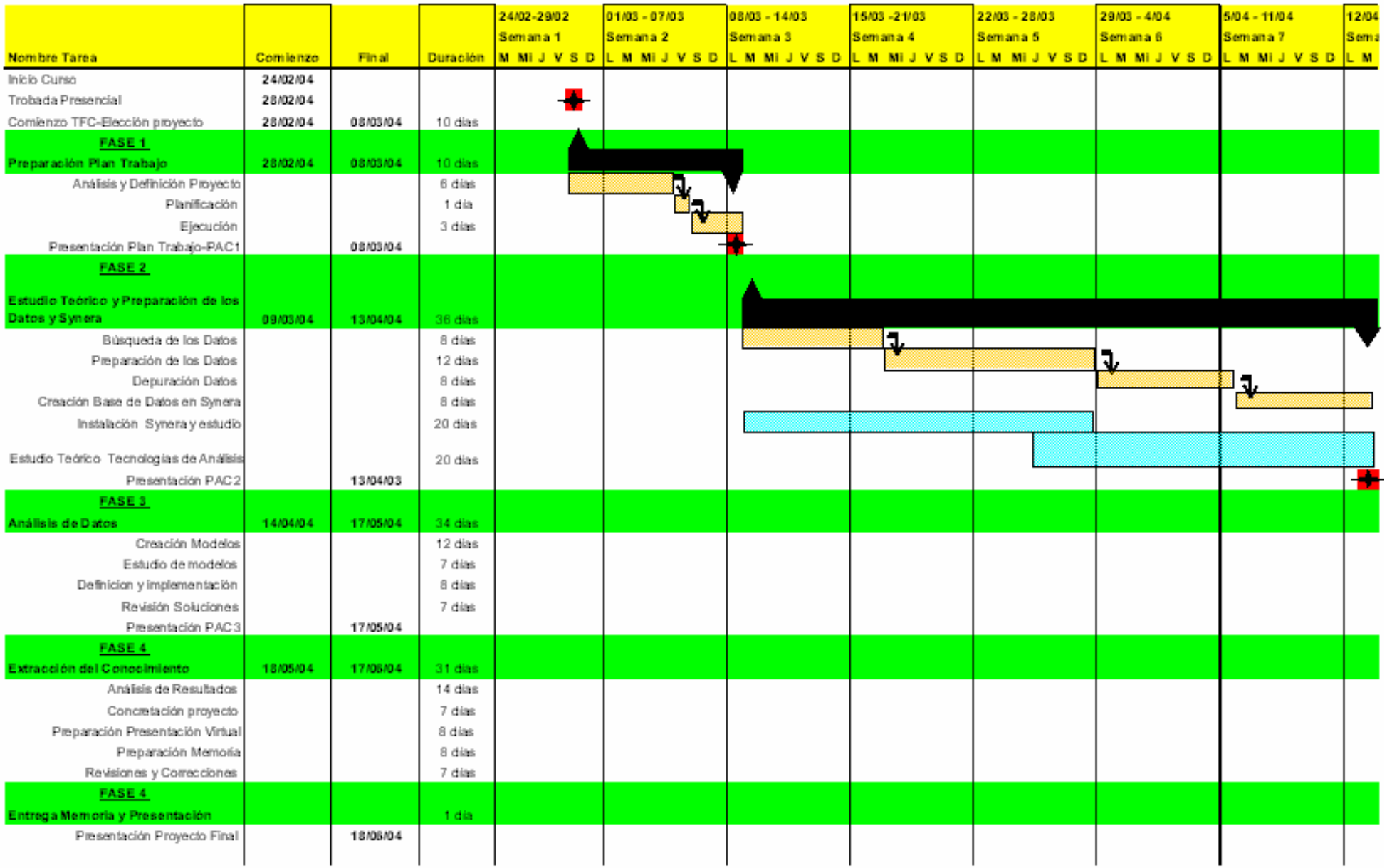
Objetivos y Tareas especificas:

- Conseguir los datos necesarios, depurarlos a fin de conseguir una base de datos la cual poder analizar mediante el Synera.
- Exportar los Datos al Synera y realizar los diferentes análisis utilizando los diferentes modelos (de agregación-clustering, arboles de decisión, redes neuronales, redes bayesianas, reglas de asociación
- Estudiar los diferentes modelos de análisis de datos, su teoría así como ver en la practica las diferencias entre ellos pudiendo llegar a analizar el porque en nuestro caso del uso de uno u otro.
- Estudiar teoría de la Gestión del conocimiento así como del Capital Intelectual.
- Estudiar y profundizar sobre las nuevas tecnologías como KDD, Data Warehouse, Data Mining, OLAP, etc.. que afectan directamente sobre nuestro proyecto e incluso pueden hacer que sea mejor.
- Instalar y estudiar el funcionamiento de la herramienta comercial Synera. Así como ver las aplicaciones más adecuadas para nuestro proyecto.

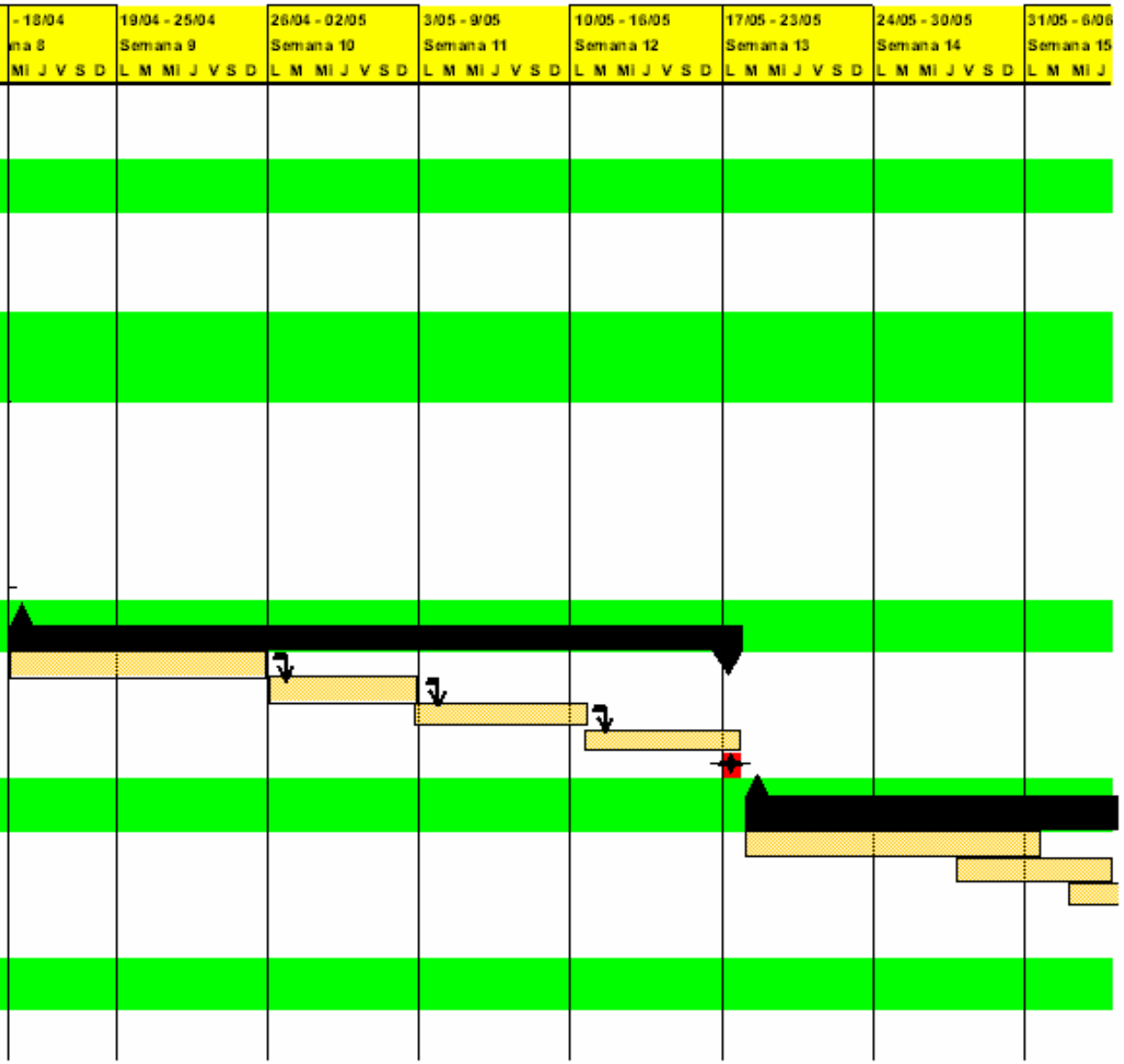
Temporalización:

Nombre Tarea	Comienzo	Final	Duración
Inicio Curso	**24/02/04**		
Trobada Presencial	**28/02/04**		
Comienzo TFC-Elección proyecto	**28/02/04**	**08/03/04**	10 días
FASE 1			
Preparación Plan Trabajo	**28/02/04**	**08/03/04**	10 días
Análisis y Definición Proyecto			6 días
Planificación			1 día
Ejecución			3 días
Presentación Plan Trabajo-PAC1		**08/03/04**	
FASE 2			
Estudio Teórico y Preparación de los Datos y Synera	**09/03/04**	**13/04/04**	36 días
Búsqueda de los Datos			8 días
Preparación de los Datos			12 días
Depuración Datos			8 días
Creación Base de Datos en Synera			8 días
Instalación Synera y estudio			20 días
Estudio Teórico Tecnologías de Análisis			20 días
Presentación PAC2		**13/04/03**	
FASE 3			
Análisis de Datos	**14/04/04**	**17/05/04**	34 días
Creación Modelos			12 días
Estudio de modelos			7 días
Definición y implementación			8 días
Revisión Soluciones			7 días
Presentación PAC3		**17/05/04**	
FASE 4			
Extracción del Conocimiento	**18/05/04**	**17/06/04**	31 días
Análisis de Resultados			14 días
Concretación proyecto			7 días
Preparación Presentación Virtual			8 días
Preparación Memoria			8 días
Revisiones y Correcciones			7 días
FASE 4			
Entrega Memoria y Presentación			1 día
Presentación Proyecto Final		**18/06/04**	

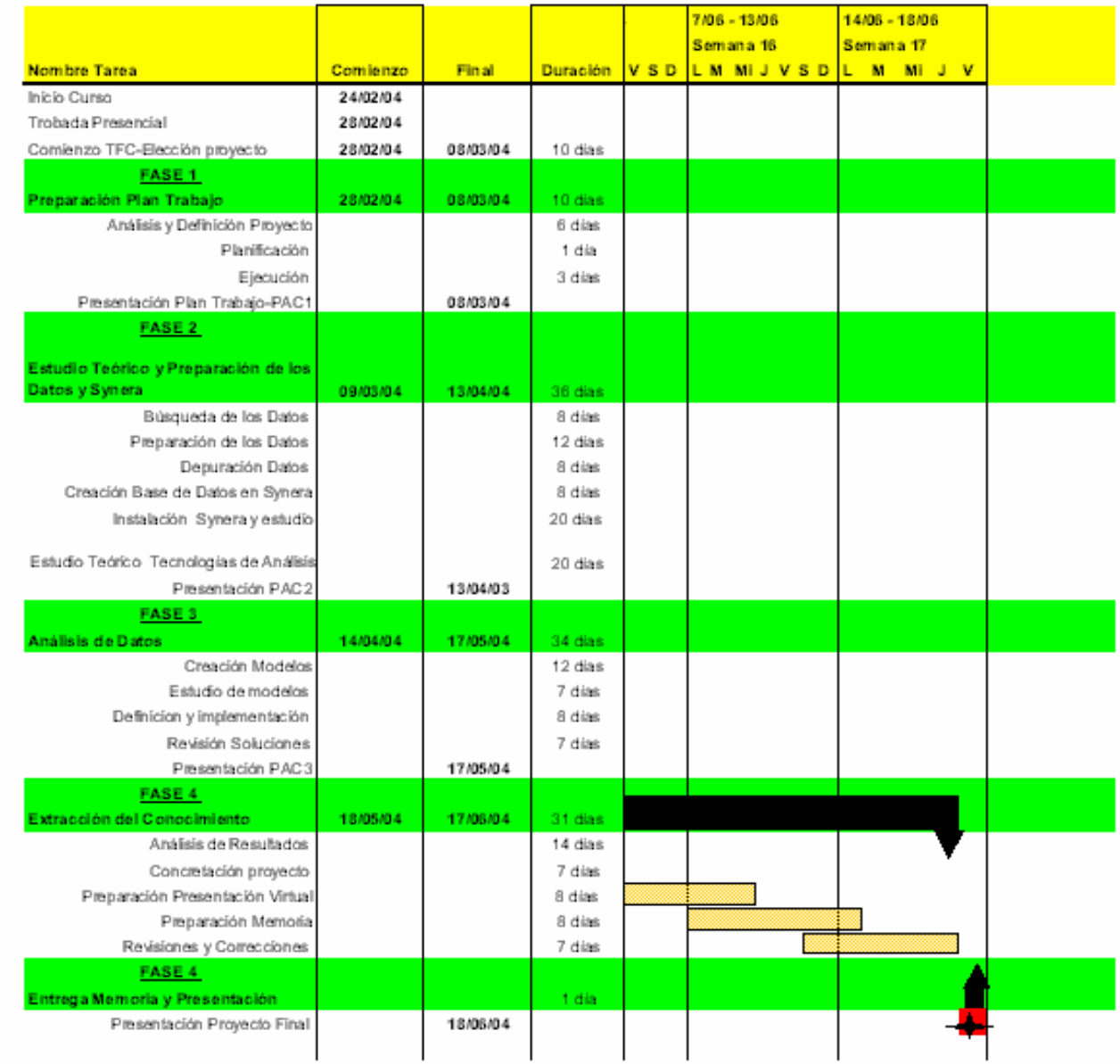
Nombre Tarea	Comienzo	Final	Duración
Inicio Curso	24/02/04		
Trobada Presencial	28/02/04		
Comienzo TFC-Elección proyecto	28/02/04	08/03/04	10 días
FASE 1			
Preparación Plan Trabajo	28/02/04	08/03/04	10 días
Análisis y Definición Proyecto			6 días
Planificación			1 día
Ejecución			3 días
Presentación Plan Trabajo-PAC1		08/03/04	
FASE 2			
Estudio Teórico y Preparación de los Datos y Synera	09/03/04	13/04/04	36 días
Búsqueda de los Datos			8 días
Preparación de los Datos			12 días
Depuración Datos			8 días
Creación Base de Datos en Synera			8 días
Instalación Synera y estudio			20 días
Estudio Teórico Tecnologías de Análisis			20 días
Presentación PAC2		13/04/03	
FASE 3			
Análisis de Datos	14/04/04	17/05/04	34 días
Creación Modelos			12 días
Estudio de modelos			7 días
Definición y implementación			8 días
Revisión Soluciones			7 días
Presentación PAC3		17/05/04	
FASE 4			
Extracción del Conocimiento	18/05/04	17/06/04	31 días
Análisis de Resultados			14 días
Concretación proyecto			7 días
Preparación Presentación Virtual			8 días
Preparación Memoria			8 días
Revisiones y Correcciones			7 días
FASE 4			
Entrega Memoria y Presentación			1 día
Presentación Proyecto Final		18/06/04	



Nombre Tarea	Comienzo	Final	Duración
Inicio Curso	24/02/04		
Trobada Presencial	28/02/04		
Comienzo TFC-Elección proyecto	28/02/04	08/03/04	10 días
FASE 1			
Preparación Plan Trabajo	28/02/04	08/03/04	10 días
Análisis y Definición Proyecto			6 días
Planificación			1 día
Ejecución			3 días
Presentación Plan Trabajo-PAC1		08/03/04	
FASE 2			
Estudio Teórico y Preparación de los Datos y Synera	09/03/04	13/04/04	36 días
Búsqueda de los Datos			8 días
Preparación de los Datos			12 días
Depuración Datos			8 días
Creación Base de Datos en Synera			8 días
Instalación Synera y estudio			20 días
Estudio Teórico Tecnologías de Análisis			20 días
Presentación PAC2		13/04/03	
FASE 3			
Análisis de Datos	14/04/04	17/05/04	34 días
Creación Modelos			12 días
Estudio de modelos			7 días
Definicion y implementación			8 días
Revisión Soluciones			7 días
Presentación PAC3		17/05/04	
FASE 4			
Extracción del Conocimiento	18/05/04	17/06/04	31 días
Análisis de Resultados			14 días
Concretación proyecto			7 días
Preparación Presentación Virtual			8 días
Preparación Memoria			8 días
Revisiones y Correcciones			7 días
FASE 4			
Entrega Memoria y Presentación			1 día
Presentación Proyecto Final		18/06/04	



Nombre Tarea	Comienzo	Final	Duración	V S D	7/06 - 13/06 Semana 16 L M Mi J V S D	14/06 - 18/06 Semana 17 L M Mi J V
Inicio Curso	24/02/04					
Trobada Presencial	28/02/04					
Comienzo TFC-Elección proyecto	28/02/04	08/03/04	10 días			
FASE 1						
Preparación Plan Trabajo	28/02/04	08/03/04	10 días			
Análisis y Definición Proyecto			6 días			
Planificación			1 día			
Ejecución			3 días			
Presentación Plan Trabajo-PAC1		08/03/04				
FASE 2						
Estudio Teórico y Preparación de los Datos y Synera	09/03/04	13/04/04	36 días			
Búsqueda de los Datos			8 días			
Preparación de los Datos			12 días			
Depuración Datos			8 días			
Creación Base de Datos en Synera			8 días			
Instalación Synera y estudio			20 días			
Estudio Teórico Tecnologías de Análisis			20 días			
Presentación PAC2		13/04/03				
FASE 3						
Análisis de Datos	14/04/04	17/05/04	34 días			
Creación Modelos			12 días			
Estudio de modelos			7 días			
Definicion y implementación			8 días			
Revisión Soluciones			7 días			
Presentación PAC3		17/05/04				
FASE 4						
Extracción del Conocimiento	18/05/04	17/06/04	31 días			
Análisis de Resultados			14 días			
Concretación proyecto			7 días			
Preparación Presentación Virtual			8 días			
Preparación Memoria			8 días			
Revisiones y Correcciones			7 días			
FASE 4						
Entrega Memoria y Presentación			1 día			
Presentación Proyecto Final		18/06/04				



Indice

2. - La Gestión del Conocimiento

Conceptos generales:

Tenemos inicialmente a través del diagrama que asocia el nivel del contexto con el nivel de entendimiento los elementos de la cadena informacional.
Luego la pirámide informacional explica el proceso de transformación asociado a la generación del conocimiento.

Figura 2 - Relaciones entre los componentes de la cadena informacional

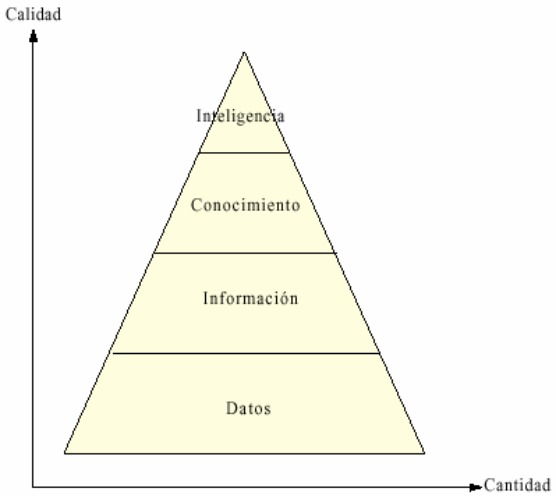


Figura 3 – Pirámide Informacional

Los datos no tienen un significado por sí mismos, ya que deben ser ordenados, agrupados, analizados e interpretados para entender potencialmente lo que nos quieren indicar. Cuando los datos son procesados, se convierten en información. Cuando la información es utilizada y puesta en el contexto o marco de referencia de una persona junto con su percepción personal se transforma en conocimiento. El conocimiento es la combinación de información, contexto y experiencia. El conocimiento resumido, una vez validado y orientado hacia un objetivo genera inteligencia (sabiduría), la cual pretende ser una representación de la realidad.
Estos factores están gobernados por dos criterios: Cantidad y Calidad.

Figura 4 - Del dato a la información

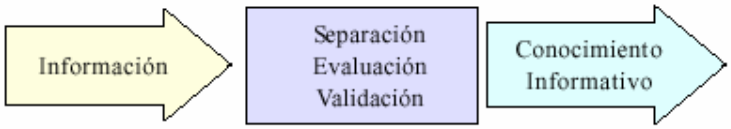


Figura 5 - De la información al conocimiento informativo

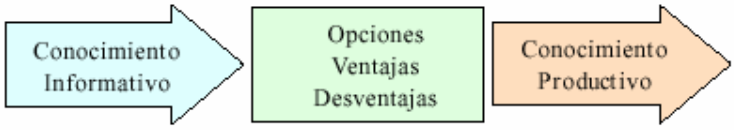


Figura 6 – Del conocimiento informativo al conocimiento productivo

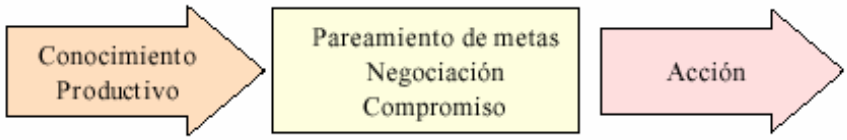


Figura 7 – Del conocimiento productivo a la acción

Por tanto tenemos también tipos de conocimiento:

Conocimiento Tácito (Subjetivo)	Conocimiento Explícito (Objetivo)
Conocimiento de las experiencias (Cuerpo)	Conocimiento del raciocinio (Mente)
Conocimiento simultaneo (Aquí y ahora)	Conocimiento secuencial (Allí y entonces)
Conocimiento Análogo (Práctica)	Conocimiento digital (Teoría)

Tabla 2 – Comparación entre el conocimiento tácito y explícito[27]

Diagrama 4 – Los cuatro modos de conversión del conocimiento[28]

Así llegamos al conocimiento organizacional que se define como lo que los integrantes de ella saben en su conjunto. Esta visión establece que son las personas que integran la organización las que son las poseedoras del conocimiento, el cual articula el funcionamiento de la organización y establece las bases para la 'Memoria Organizacional'.
Nonaka y Takeuchi establecen cuatro factores clave en torno a la creación del conocimiento organizacional:

- **Intención:** La organización debe tener la intención explícita de generar las condiciones optimas que permitan el crecimiento de la espiral de conocimiento organizacional. También se deben considerar los criterios necesarios para evaluar el valor y utilidad de los activos de conocimiento.
- **Autonomía:** La organización debe permitir algún nivel de autonomía en sus individuos, lo cual fomente la generación de nuevas ideas y visualización de nuevas oportunidades, motivando así a los participantes de la organización a generar nuevo conocimiento.

- **Fluctuación y caos creativo:** La organización debe estimular la interacción entre sus integrantes y el ambiente externo con el objeto de estimular nuevas perspectivas de cómo hacer las cosas. El caos se genera naturalmente cuando la organización sufre una crisis o cuando los administradores deciden establecer nuevas metas.
- **Redundancia:** La organización debe permitir niveles de redundancia dentro de su operativa. Esto genera que los diferentes puntos de vistas establecidos por las personas que conforman los equipos genere ideas robustas y nuevas posibilidades.

Esto nos lleva a crear una **"Organización capaz de aprender "**

Figura 9 - Tipos de aprendizaje

Así llegamos a la **Gestión del conocimiento**.
En primer lugar, el término 'Gestión' se define como "el proceso mediante el cual se obtiene, despliega o utiliza una variedad de recursos básicos para apoyar los objetivos de la organización."
Pero debido a lo novedoso del término "Gestión del Conocimiento", existen un sin número de definiciones:

- "Es el proceso sistemático de buscar, organizar, filtrar y presentar la información con el objetivo de mejorar la comprensión de las personas en una especifica área de interés", Thomas H. Davenport
- "Encarna el proceso organizacional que busca la combinación sinérgica del tratamiento de datos e información a través de las capacidades de las Tecnologías de Información, y las capacidades de creatividad e innovación de los seres humanos", Dr. Yogesh Malhotra
- "Es la habilidad de desarrollar, mantener, influenciar y renovar los activos intangibles llamados Capital de Conocimiento o Capital Intelectual", Hubert Saint-Onge.
- "Es el arte de crear valor con los activos intangibles de una organización", Phd. Karl E. Sveiby

Pero resumiendo tomaremos por valida la siguiente definición :
Gestión del Conocimiento: *Es el proceso sistemático de detectar, seleccionar, organizar, filtrar, presentar y usar la información por parte de los participantes de la organización, con el objeto de explotar cooperativamente los recursos de conocimiento basados en el capital intelectual propio de las organizaciones, orientados a potenciar las competencias organizacionales y la generación de valor.*

Los objetivos de la Gestión del conocimiento

Algunos objetivos de la Gestión del conocimiento son los siguientes:

- Formular una estrategia de alcance organizacional para el desarrollo, adquisición y aplicación del conocimiento.
- Implantar estrategias orientadas al conocimiento.

- Promover la mejora continua de los procesos de negocio, enfatizando en la generación y utilización del conocimiento.
- Monitorear y evaluar los logros obtenidos mediante la aplicación del conocimiento.
- Reducir los tiempos de ciclos en el desarrollo de nuevos productos, mejoras de los ya existentes y la reducción del desarrollo de soluciones a los problemas.
- Reducir los costos asociados a la repetición de errores.

Estos objetivos se ven complementados a través de actividades de apoyo, tales como el desarrollo de una gama de proyectos organizacionales, los cuales deben obedecer los objetivos generales en términos de los intereses y capacidades.

El Proceso de Gestión del Conocimiento

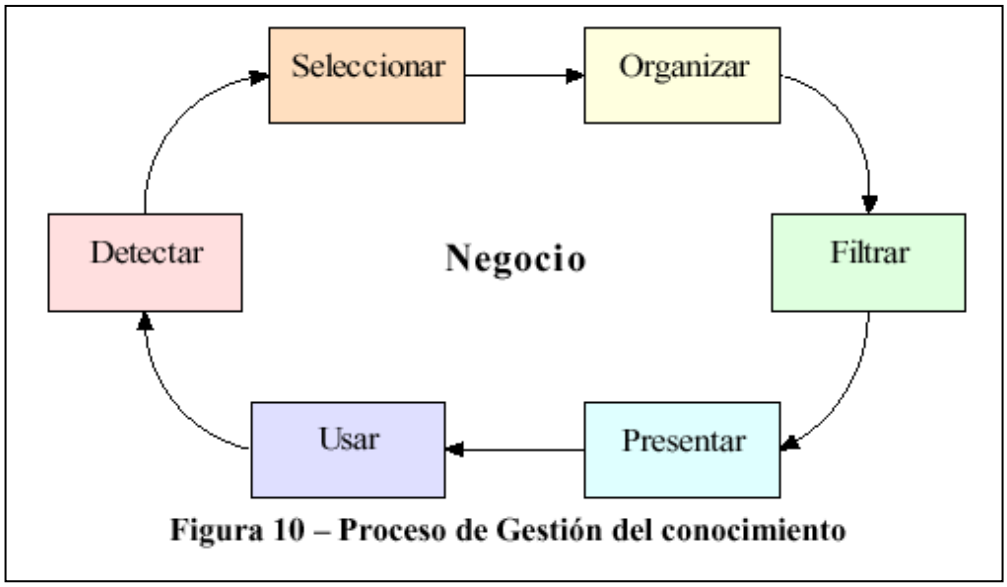


Figura 10 – Proceso de Gestión del conocimiento

donde:

- **Detectar:** Es el proceso de localizar modelos cognitivos y activos (pensamiento y acción) de valor para la organización, el cual radica en las personas. Las fuentes de conocimiento pueden ser generadas tanto de forma interna (I&D, proyectos, descubrimientos, etc.) como externa (fuentes de información periódica, Internet, cursos de capacitación, libros, etc.).
- **Seleccionar:** Es el proceso de evaluación y elección del modelo en torno a un criterio de interés. Los criterios pueden estar basados en criterios organizacionales, comunales o individuales, los cuales estarán divididos en tres grandes grupos: Interés, Práctica y Acción.
- **Organizar:** Es el proceso de almacenar de forma estructurada la representación explícita del modelo. Este proceso se divide en las siguientes etapas
  - **Generación:** Es la creación de nuevas ideas, el reconocimiento de nuevos patrones, la síntesis de disciplinas separadas, y el desarrollo de nuevos procesos.
  - **Codificación**: Es la representación del conocimiento para que pueda ser accedido y transferido por cualquier miembro de la organización a través de algún lenguaje de representación (palabras, diagramas, estructuras, etc.). Cabe destacar que la representación de codificación puede diferir de la representación de almacenamiento, dado que enfrentan objetivos diferentes: personas y máquinas.
  - **Trasferencia:** Es establecer el almacenamiento y la apertura que tendrá el conocimiento, ayudado por interfaces de acceso masivo (por ejemplo, la Internet o una Intranet), junto a establecer los criterios de seguridad y acceso. Además debe considerar aspectos tales como las barreras de tipo Temporales (Vencimiento), de Distancias y Sociales.
  - **Filtrar:** Una vez organizada la fuente, puede ser accedida a través de consultas automatizadas en torno a motores de búsquedas. Las búsquedas se basarán en

estructuras de acceso simples y complejas, tales como mapas de conocimientos, portales de conocimiento o agentes inteligentes.

- **Presentar:** Los resultados obtenidos del proceso de filtrado deben ser presentados a personas o máquinas. En caso que sean personas, las interfaces deben estar diseñadas para abarcar el amplio rango de comprensión humana. En el caso que la comunicación se desarrolle entre máquinas, las interfaces deben cumplir todas las condiciones propias de un protocolo o interfaz de comunicación.
- **Usar:** El uso del conocimiento reside en el acto de aplicarlo al problema objeto de resolver. De acuerdo con esta acción es posible evaluar la utilidad de la fuente de conocimiento a través de una actividad de retroalimentación.

Tipos de proyectos de Gestión del conocimiento

Existe una variedad de proyectos que contribuyen a implementar la gestión del conocimiento dentro de las organizaciones, donde cada uno de ellos contempla las características de las necesidades organizacionales.

- Diferencias entre la Gestión de información y la Gestión del conocimiento

La gestión del conocimiento está basada en parte en la gestión de información. En este contexto es necesario diferenciar la gestión de información y la gestión del conocimiento.

"mientras la información es definida como un flujo de mensajes, el conocimiento es la combinación de información y contexto en la medida que produce acciones."

Proyecto de gestión del conocimiento	Proyecto de gestión de la información
Las metas acentúan el valor agregado para los usuarios	Las metas acentúan la liberación y accesibilidad de la información
Apoya las mejoras operacionales y la innovación	Apoya las operaciones existentes
Agrega valor al contenido a través de filtros, sintetizado, interpretación, recorte de contenido.	Libera contenidos disponibles con pequeño valor agregado
Usualmente requiere contribuciones y feedback continuo	Enfatiza en trasferencias de información en un sentido
Enfoque balanceado entre los aspectos tecnológicos y culturales	Fuerte enfoque tecnológico
Variaciones en los sistemas de entrada imposibilitan automatizar el proceso de captura	Asume que la captura de información puede ser automatizada

Tabla 3 – Diferencias entre la gestión del conocimiento y la gestión de información[36]

¿Qué es un proyecto de Gestión del conocimiento?

Se define un proyecto de gestión del conocimiento como "la unidad básica de actividades que la empresa utiliza para generar valor desde los activos de conocimiento"

Algunos tipos son:

- **Capturar y rehusar conocimiento estructurado:** Este tipo de proyectos reconoce que el conocimiento se encuentra embebido en los componentes de salida de una organización, tales como diseño de productos, propuestas, reportes, procedimientos de implementación, código de software, entre otros.
- **Capturar y compartir lecciones aprendidas desde la práctica:** Este tipo de proyectos captura el conocimiento generado por la experiencia, el cual puede ser adaptado por un usuario para su uso en un nuevo contexto.
- **Identificar fuentes y redes de experiencia:** Este tipo de proyectos intenta capturar y desarrollar el conocimiento, permitiendo visualizar y acceder de la mejor manera a la experiencia, facilitando la conexión entre las personas que poseen el conocimiento y quienes lo necesitan.
- **Estructurar y mapear las necesidades de conocimiento para mejorar el rendimiento:** Este tipo de proyecto pretende apoyar los esfuerzos en el desarrollo de nuevos productos o el

rediseño de procesos haciendo explícito el conocimiento necesario para una etapa particular de una iniciativa .

- **Medir y manejar el valor económico del conocimiento:** Este tipo de proyecto reconoce que los activos tales como patentes, derechos de autor, licencias de software y bases de datos de clientes, crean tanto ingresos como costos para la organización, por lo que se orientan a administrarlos más juiciosamente.
- **Sintetizar y compartir conocimiento desde fuentes externas:** Este tipo de proyectos intentan aprovechar las fuentes de información y conocimiento externas, proveyendo un contexto para el gran volumen de datos disponible(Universidades).

Es importante destacar que los distintos proyectos descritos anteriormente concuerdan en una visión objetiva de negocios: la agregación de valor en torno a las necesidades de la organización.

Las ventajas competitivas

"la ventaja competitiva nace fundamentalmente del valor que una empresa es capaz de crear para sus compradores".

Las tres estrategias genéricas son:

- **Liderazgo en costos**

Esta estrategia fue muy popular en los años '70. Mantener el costo más bajo frente a los competidores y lograr un volumen alto de ventas. Se busca minimizar los costos en las áreas de I&D, red de ventas, publicidad, personal, entre otras. La competencia relacionada con la reducción de costos erosiona los márgenes de la competencia, estableciendo una barrera de entrada. Para lograr un posicionamiento basado en reducción de costos es frecuentemente necesario contar con un alto grado de participación del mercado con relación al competidor más cercano u otro tipo de ventaja tal como la cercanía con las materias primas. La desventaja de esta estrategia implica altos niveles de inversión inicial en tecnología, precios agresivos y reducción de márgenes.

- **Diferenciación**

Esta estrategia está basada en crearle al producto o servicio algo que sea percibido en todo el mercado como único. La diferenciación genera lealtad de marca, lo cual elimina las sensibilidades basadas en precio. Diferenciarse significa sacrificar participación de mercado, implementar actividades de investigación, diseño de productos, alta calidad, servicio al cliente, entre otras.
En esta estrategia es posible competir con bajos costos y diferenciarse, sólo que estará condicionado a las reacciones de los competidores. La desventaja de esta estrategia implica menor participación de mercado, altos niveles de inversión en I&D y Diseño de productos.

- **Focalización**

Esta estrategia está basada en concentrarse en un grupo especifico de clientes, en un segmento de mercado. La estrategia se basa en la premisa de que la organización está en condiciones de servir a un objetivo estratégico más reducido de forma más eficiente que los competidores de mayor cobertura. Como resultado, la empresa se diferenciara al atender mejor las necesidades de un mercado especifico. La desventaja de este estrategia es que implica menor participación de mercado, altos niveles de inversión en especialización y debilidades de diversificación.

Las Tecnologías de la Información (TI)

En la actualidad, entender cuál es el rol de las TI en torno a la gestión del conocimiento es la pieza clave para no cometer un error de concepto. Este error radica en entender la implantación de la Gestión del conocimiento como un tarea de la TI.

"Las TI proveen el marco, pero no el contenido. El contenido es una cuestión exclusiva de los individuos. La TI facilita el proceso, pero por si misma es incapaz de extraer algo de la cabeza de una persona"

El apoyo que pueden entregar las TI radica en instancias tecnológicas y culturales para ayudar a la dinámica del proceso de la Gestión del conocimiento. Estas pueden ser:

- **Generación de conocimiento:** Son las herramientas y técnicas que se enfocan a la exploración y análisis de datos para descubrir patrones interesantes dentro de ellos. Algunas herramientas/técnicas son Data Mining (DM), Knowledge Discovery in Databases (KDD) ,Text

Mining (TM), Web Mining (WM), Sistemas Inteligentes de Apoyo a las Decisiones (SAID), Sistemas Expertos (SE), Agentes Inteligentes (AI), entre muchos otras.

- **Facilitador de la generación de conocimiento:** Son las herramientas y técnicas que facilitan el libre flujo de conocimiento dentro de la organización. Algunas herramientas/técnicas son Lotus Notes, NetMeeting, Email, Intranets/Extranets & Portales, IdeaFisher, IdeaProcesor, Grupos de discusión, Servicio de mensajes, entre otras. Este tipo de tecnología se cataloga dentro del área de la Administración de la Información, comunicación, representación y Groupware.
- **Mediciones de conocimiento:** Son herramientas y técnicas que facilitan la 'visualización' de los conocimientos. Se pueden catalogar en tres categorías: actividades de conocimiento, resultados basados en conocimientos, e inversiones en conocimiento.

Para evaluar si la tecnología disponible, tanto en la organización como en el mercado, apoya a la Gestión de Información, la Gestión del Conocimiento y el Aprendizaje Organizacional, se debe tener en cuenta:

- Si apoyan a la estructuración de las fuentes de información en que se basan las decisiones.
- Si apoyan la generación de informes que resumen los datos útiles.
- Si los medios de comunicación entregan la información necesaria a las personas indicadas en el momento en que se necesita.
- Si apoyan las redes formales e informales de la organización.
- Si se integran fácilmente con el entorno y en los procesos de trabajo.
- Si posee interfaces factibles de usar y explotar.
- Si la apertura de la herramienta es suficiente como para interactuar con otras herramientas.
- Si apoyan la creación y transferencia de conocimiento tácito y explícito dentro de la organización.

El estado actual de la Gestión del conocimiento

Internacionalmente la Gestión del conocimiento está tomando cada vez mayor relevancia en el desarrollo de las empresas.

Estadísticas actuales

En los estudios realizados por KMPG del año 1998 y del año 2000, en que encuestó a 100 y 423 organizaciones respectivamente, se presentan una serie de estadísticas interesantes.

Algunos puntos interesantes son:

- El 61% de las empresas sufre de sobrecarga de información, lo cual provoca que sus integrantes no tengan el tiempo necesario para compartir conocimiento.
- El 81% de las empresas tiene, actualmente o consideran planificar, programas de Gestión del Conocimiento. El 38% tiene actualmente un programa de Gestión del Conocimiento, lo cual muestra que las empresas han empezado a considerar la necesidad de este tipo de proyectos.
- En las empresas que han implantado programas de Gestión del Conocimiento comentan que juega un rol 'extremadamente importante' o 'importante' en la mejora de las Ventajas competitivas (79%), en el Marketing (75%), en Mejorar el enfoque al cliente (72%), en el Desarrollo de los empleados (57%), en la Innovación de productos (64%) y en el incremento del crecimiento y las ganancias (ambas 63%).
- Las empresas con programas de Gestión del Conocimiento están mejor localizadas que las que no tienen.
- Las implementaciones de programas de Gestión del Conocimiento han generado una gran variedad de acciones. El 76% ha generado una Estrategia de conocimiento, el 64% ha adoptado por el entrenamiento, el 58% ha establecido compartir mejores prácticas, el 57% ha instaurado políticas de conocimiento y el 50% ha establecido redes formales de Gestión del Conocimiento.

Sin embargo, no todo han sido buenas noticias:

- Lamentablemente, los estudios revelan que las organizaciones aún siguen ciegas a las consideraciones de los empleados. De hecho, sólo el 33% de los programas de Gestión del

Conocimiento ha implementado políticas en torno al conocimiento - estipulando cuales elementos de conocimiento almacenar, actualizar y seleccionar - y menor aún (31%) gratificar a los trabajadores del conocimiento.

- Las empresas aún ven a la Gestión del conocimiento como una solución puramente tecnológica. Por ejemplo, la participación de la tecnología en las soluciones está marcada por el uso de Internet (93%), Intranet (78%), Data warehousing y Data Mining (63%), administración de documentos (61%), apoyo a decisiones (49%), Groupware (43%) y Extranets (38%), frente a un 44% de desarrollo de una estrategia de conocimiento, 33% de desarrollo de políticas y creación de redes formales en torno al conocimiento. Una investigación realizada por la consultora Arthur Andersen en torno a los factores críticos para la implantación de la Gestión del conocimiento indicó que "solo uno de los seis factores críticos para implementar eficazmente la Gestión del conocimiento está relacionado con la tecnología. La apertura y la confiabilidad de la alta gerencia encabezan la lista".
- Algunos beneficios esperados no se han cumplido. El 20% opina que la falta de comunicación entre los usuarios es uno de los motivos, el 19% opina que es debido a que el uso diario no se integra con el proceso normal de trabajo, el 18% opina que es debido a que los sistemas son muy complicados, el 15% piensa que es debido a la falta de entrenamiento, mientras que el 13% opina que es por que no se visualizan beneficios personales.

Además, otras características importantes reveladas en estos estudios son: no existe un consenso en torno a la definición de Gestión del conocimiento, las expectativas y resultados esperados, y la relación existente entre los activos intangibles y el valor de mercado.

Conclusiones extraídas del análisis de casos reales de empresas que han implementado programas de Gestión del Conocimiento

- Una alineación de las diferentes iniciativas en torno a la estrategia corporativa es primordial. Las necesidades de las variadas áreas de una organización pueden generar un sin número de iniciativas de Gestión del conocimiento, lo cual puede generar objetivos locales distintos. Estos objetivos deben ser congruentes con el objetivo general o corporativo, con el fin de "empujar todos para el mismo lado desde diferentes puntos".
- La tecnología cumple un rol estratégico como facilitador de la comunicación entre las personas. En la mayoría los casos la tecnología puede ser mal utilizada o sobredimensionada, por lo que es indispensable que ella se adapte a la operativa normal de la organización.
- Claramente una instancia de Gestión del conocimiento puede orientarse a reforzar los aspectos competitivos de una organización.
- No es necesario realizar una implantación brusca de la Gestión del conocimiento en la organización. Sólo será necesario establecer cual es la mejor oportunidad para iniciar una instancia de proyecto de Gestión del conocimiento para verificar la efectividad de los criterios utilizados, y que ayude a visualizar los resultados obtenidos y contrastarlos con los resultados esperados.
- Una de las alegres paradojas que presenta la Gestión del conocimiento es el hecho de generar ganancias/ventajas con recursos que siempre se han tenido a mano.

El Director de Conocimiento: Un nuevo rol estratégico

¿Qué es un Director de Conocimiento?

Es el encargado de "iniciar, impulsar y coordinar los programas de Gestión del conocimiento". Sin embargo, una definición tan sencilla puede llevar a
confusiones tales como entender que los proyectos de Gestión de conocimiento deben estar a cargo del Director Informático (Visión tecnológica) o del Director de Recursos Humanos (Visión organizacional).

Las responsabilidades del Director Informático - Estrategia de TI, Operaciones de TI, y

manejar los programas de las TI –han provocado la confusión debido a que inicialmente los proyectos de Gestión del conocimiento han sido asignados al área de TI.
Realmente el Director Informático tiene como objetivo supervisar el despliegue de las TI y el Director de Conocimiento se centra en maximizar las creación, el descubrimiento y la diseminación de conocimientos en la organización.

¿Por qué es necesario un Director de Conocimiento?
Sin duda, será necesario determinar si este nuevo puesto ejecutivo tiene fundamentos sostenibles para su implementación.
Algunas de sus funciones serán:

- Maximizar el retorno de las inversiones en conocimiento, tales como nuevas contrataciones, procesos y capital intelectual.
- Explotar los activos intangibles, tales como el know-how, patentes y relación de clientes.
- Repetir los éxitos pasados y compartir mejores prácticas.
- Mejorar la innovación (Comercialización de ideas).
- Evitar la perdida de conocimiento y las fugas producidas por las reestructuraciones organizacionales.

Sin embargo, destacamos una serie de situaciones en donde el Director de Conocimiento no será necesario.
Algunas de ellas son:

- El conocimiento no es importante en el negocio.
- Se está contento con las iniciativas locales (proyectos de Gestión del Conocimiento informales) y se espera que todo vaya bien.
- Existe una cultura de compartir conocimiento y un proceso sistémico de difusión.
- El liderazgo en conocimiento viene de la cima y es perseguido apasionadamente.
- Cada uno posee planes de desarrollo de conocimiento en sus planes de trabajo.
- Los sistemas de monitoreo de rendimiento poseen una dimensión explícita en

El contexto tecnológico de la Gestión del conocimiento

Tecnología/Herramienta	Nivel
Internet	93%
Intranet	78%
Data warehousing/mining	63%
Administración de documentos	61%
Sistemas de apoyo a la toma de decisiones	49%
Groupware	43%
Extranet	38%
Inteligencia Artificial	22%

Tabla 4 - KM y el rol de la Tecnología[33]

***KM =**

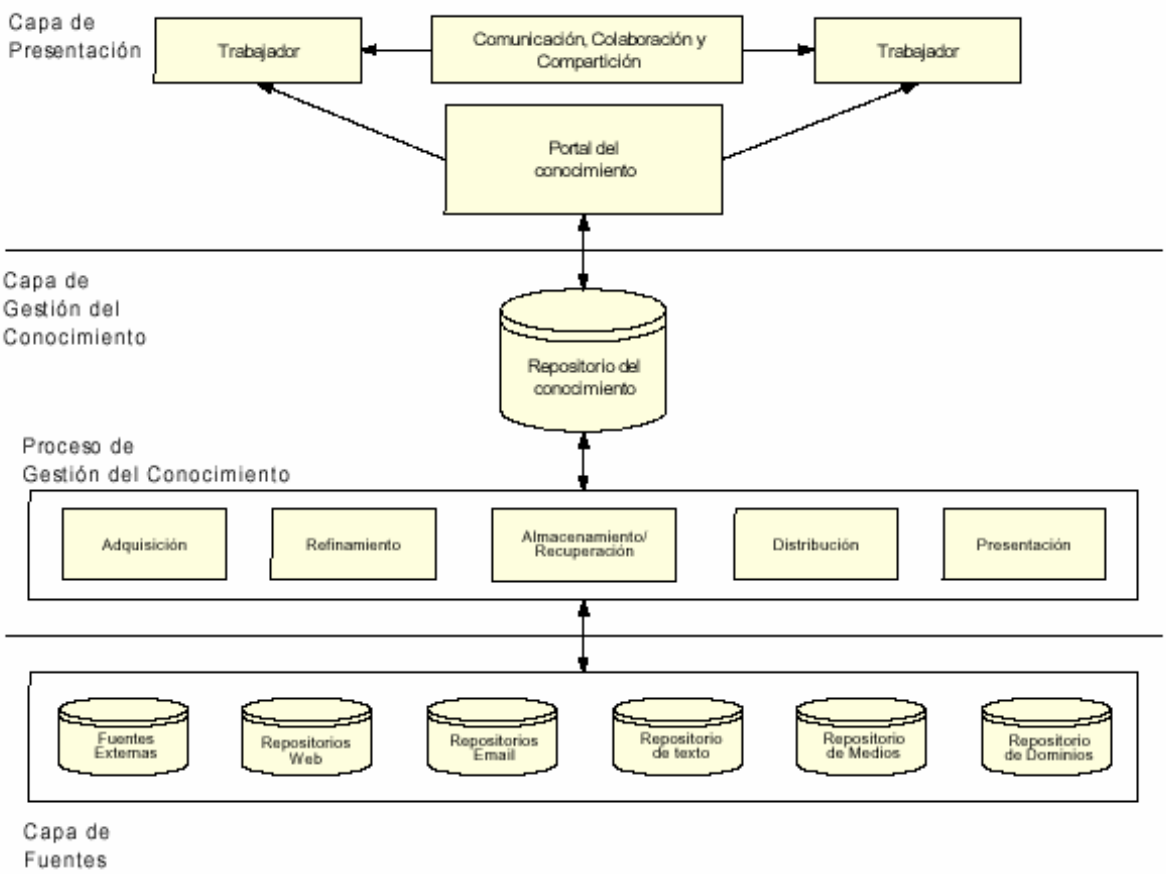


Figura 14 - Arquitectura de Gestión del conocimiento[36]

Análisis de debilidades

El esquema presentado anteriormente representa en gran medida la arquitectura sobre la cual se basan los diferentes proyectos de Gestión del Conocimiento.

Pero :

"las bases de datos sólo complementan las redes personales de aquellos que buscan las respuestas a los problemas. No importa cuán robusta sean las búsquedas o cuán personalizadas estén las bases de datos, la red de relaciones humanas de una persona a menudo determina cuál es el conocimiento que ella accede. La gente toma ventaja de las bases de datos sólo cuando los colegas lo dirigen a un punto especifico de ella".

Así descubrimos la necesidad de incorporar un nuevo factor dentro de la arquitectura, el cual considera los intereses de cada persona, el concepto de relación entre ellas a través de 'comunidades' y redes de conversación, y el comportamiento basado en compartir intereses comunes.

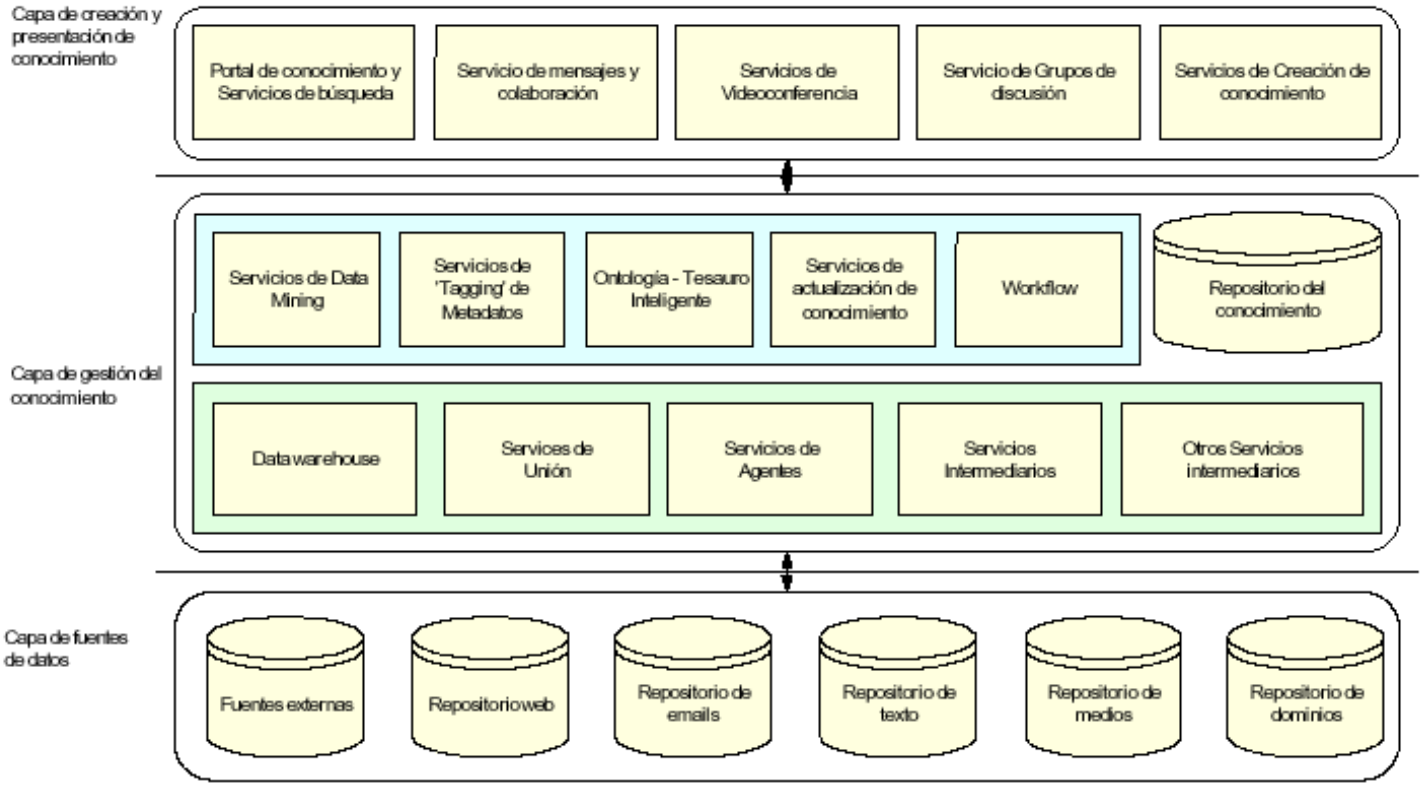


Figura 15 - Sistema de Gestión del conocimiento[37]

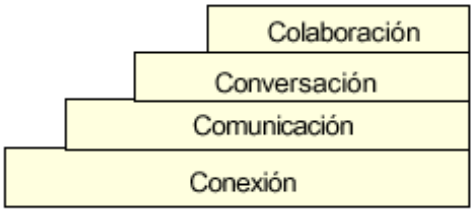


Figura 16 - Niveles de infraestructura de TI para el conocimiento[38]

Generación del modelo:

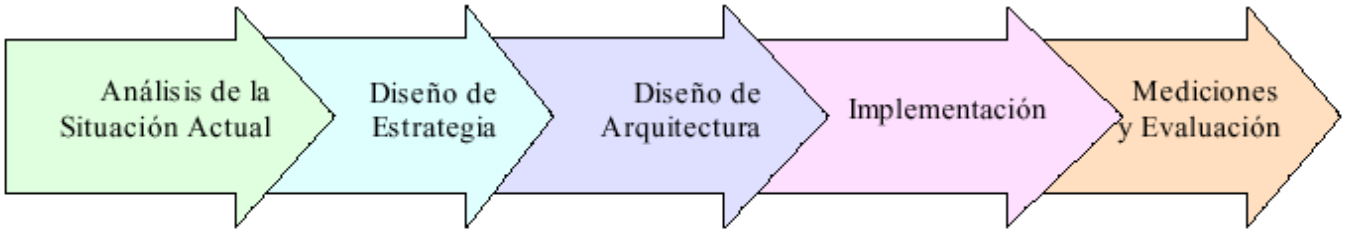


Figura 17 - Modelo propuesto

- **1º Etapa:** La etapa inicial nació de la necesidad de analizar la situación actual y la proyección futura de los recursos y capacidades de la organización.
- **2º Etapa:** Se basa en la importancia del conocimiento a nivel estratégico dentro de la organización.
- **3º Etapa:** Es la que incluye las necesidades y proyecciones establecidas en la estrategia de conocimiento, reconociendo el grado de adaptabilidad tecnológica necesaria para una evolución de los proyectos involucrados y un criterio de diseño e integración de largo plazo.
- **4º Etapa:** La etapa de implantación nace de la necesidad de coordinar todos los esfuerzos necesarios para el desarrollo de todo proyecto.
- **5º Etapa:** La etapa final, mediciones y evaluación, es necesaria debido a que es de vital importancia el visualizar los resultados obtenidos, ya sea desde el punto de vista valorativo (factores de rendimiento) como del punto de vista ambientalista (percepción de los resultados).

El modelo en su conjunto tiene como objetivo fomentar el desarrollo del aprendizaje de la organización, basado en el conocimiento y en la cultura que esta posee, donde el proyecto de Gestión del Conocimiento sea implantado con un criterio evolutivo

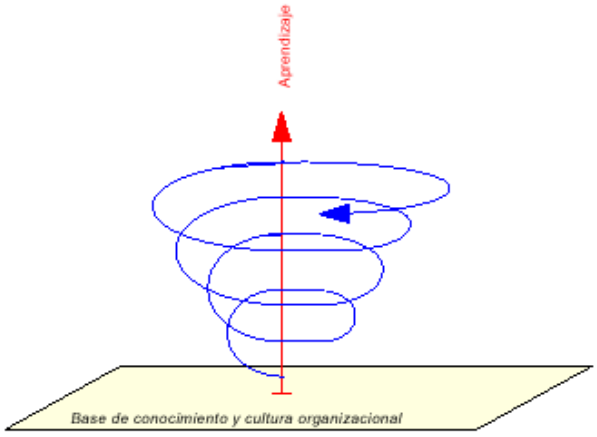


Figura 18 - Evolución de la implantación

Indice

3. - KNOWLEDGE DISCOVERY

Introducción:

Hoy en día, la cantidad de datos que ha sido almacenada en las bases de datos excede nuestra habilidad para reducir y analizar los datos sin el uso de técnicas de análisis automatizadas. Muchas bases de datos comerciales transaccionales y científicas crecen a una proporción gigantesca. El KDD [Knowledge Discovery in Databases] es el proceso completo de extracción de información, que se encarga además de la preparación de los
datos y de la interpretación de los resultados obtenidos. KDD se ha definido como
"el proceso no trivial de identificación en los datos de patrones válidos, nuevos, potencialmente útiles, y finalmente comprensibles"
Se trata de interpretar grandes cantidades de datos y encontrar relaciones o patrones.
Para conseguirlo harán falta técnicas de estadística, bases de datos, técnicas de representación del conocimiento, razonamiento basado en casos, razonamiento aproximado, adquisición de conocimiento, redes neurales y visualización de datos. Tareas comunes en KDD son la inducción de reglas, los problemas de clasificación y clustering, el reconocimiento de patrones, el modelado predictivo, la detección de dependencias, etc.
Los datos recogen un conjunto de hechos (una base de datos) y los patrones son expresiones que describen un subconjunto de los datos (un modelo aplicable a ese subconjunto). KDD involucra un proceso iterativo e interactivo de búsqueda de modelos, patrones o parámetros. Los patrones descubiertos han de ser válidos y potencialmente útiles.
Ha llegado un momento en el que disponemos de tanta información que nos vemos incapaces de sacarle provecho. Los datos tal cual se almacenan no suelen proporcionar beneficios directos. Su valor real reside en la información que podamos extraer de ellos: información que nos ayude a tomar decisiones o a mejorar nuestra comprensión de los fenómenos que nos rodean.
Una de las premisas mayores de KDD es que el conocimiento es descubierto usando técnicas de aprendizaje inteligente que van examinando los datos a través de procesos automatizados. Para que una técnica sea considerada útil para el descubrimiento del conocimiento, éste debe ser interesante; es decir, debe tener un valor potencial para el usuario.
KDD proporciona la capacidad para descubrir información nueva y significativa usando los datos existentes.

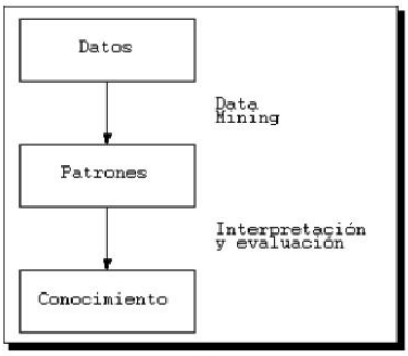


Figura 1.1: *Esquema del proceso de KDD*

El proceso de KDD

El proceso de KDD se inicia con la identificación de los datos. Para ello hay que imaginar qué datos se necesitan, dónde se pueden encontrar y cómo conseguirlos. Una vez que se dispone de datos, se deben seleccionar aquellos que sean útiles para los objetivos propuestos. Se preparan, poniéndolos en un formato adecuado.
Una vez se tienen los datos adecuados se procede a la minería de datos, proceso en el que se seleccionarán las herramientas y técnicas adecuadas para lograr los objetivos pretendidos. Y tras este proceso llega el análisis de resultados, con lo que se obtiene el conocimiento pretendido.

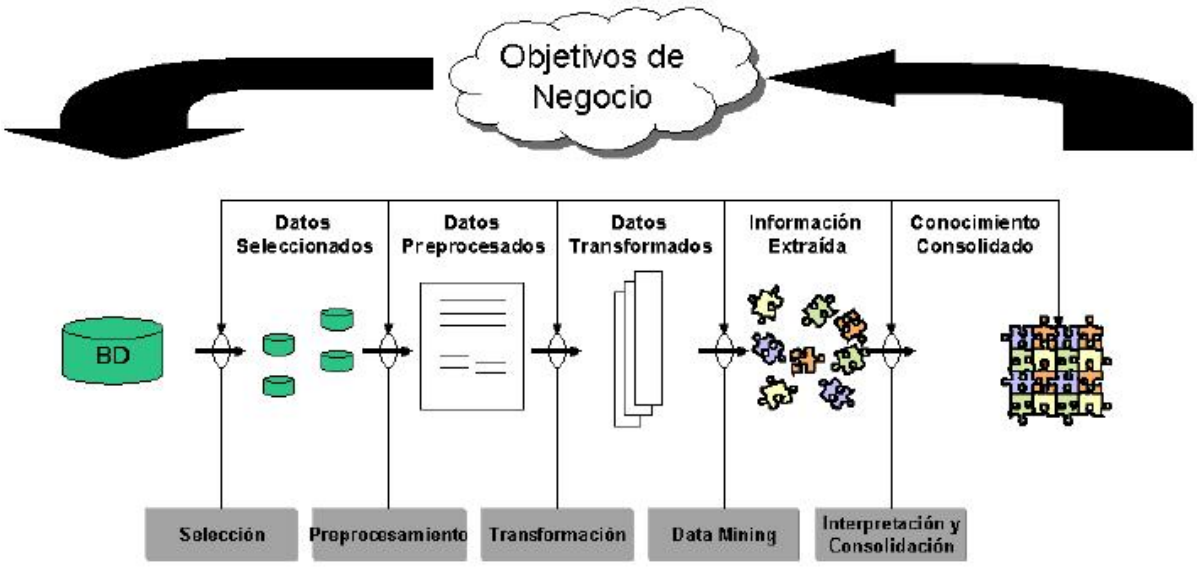


Figura 1.2: *Metodología para el descubrimiento de conocimiento en bases de datos.*

KDD es un proceso interactivo e iterativo, que involucra numerosos pasos e incluye muchas decisiones que deben ser tomadas por el usuario, y se estructura en las siguientes etapas:

- Comprensión del dominio de la aplicación, del conocimiento relevante y de los objetivos del usuario final.
- Creación del conjunto de datos: consiste en la selección del conjunto de datos, o del subconjunto de variables o muestra de datos, sobre los cuales se va a realizar el descubrimiento.
- Limpieza y preprocesamiento de los datos: Se compone de las operaciones, tales como: recolección de la información necesaria sobre la cual se va a realizar el proceso, decidir las estrategias sobre la forma en que se van a manejar los campos de los datos no disponibles, estimación del tiempo de la información y sus posibles cambios.
- Reducción de los datos y proyección: Encontrar las características más significativas para representar los datos, dependiendo del objetivo del proceso. En este paso se pueden utilizar métodos de transformación para reducir el número efectivo de variables a ser consideradas o para encontrar otras representaciones de los datos.
- Elegir la tarea de Minería de Datos: Decidir si el objetivo del proceso de KDD es: Regresión, Clasificación, Agrupamiento, etc.
- Elección del algoritmo(s) de Minería de Datos: Selección del método(s) a ser utilizado para buscar los patrones en los datos. Incluye además la decisión sobre que modelos y parámetros pueden ser los más apropiados.
- Minería de Datos: Consiste en la búsqueda de los patrones de interés en una determinada forma de representación o sobre un representaciones, utilizando para ello métodos de clasificación, reglas o árboles, regresión, agrupación, etc.
- Interpretación de los patrones encontrados. Dependiendo de los resultados, a veces se hace necesario regresar a uno de los pasos anteriores.
- Consolidación del conocimiento descubierto: consiste en la incorporación de este conocimiento al funcionamiento del sistema, o simplemente documentación e información a las partes interesadas.

El proceso de KDD puede involucrar varias iteraciones y puede contener ciclos entre dos de cualquiera de los pasos. La mayoría de los trabajos que se han realizado sobre KDD se centran en la etapa de minería. Sin embargo, los otros pasos se consideran importantes para el éxito del KDD. Por eso aunque la Minería de Datos es una parte del proceso completo de KDD, en

buena parte de la literatura los términos Minería de Datos y KDD se identifican como si fueran lo mismo.

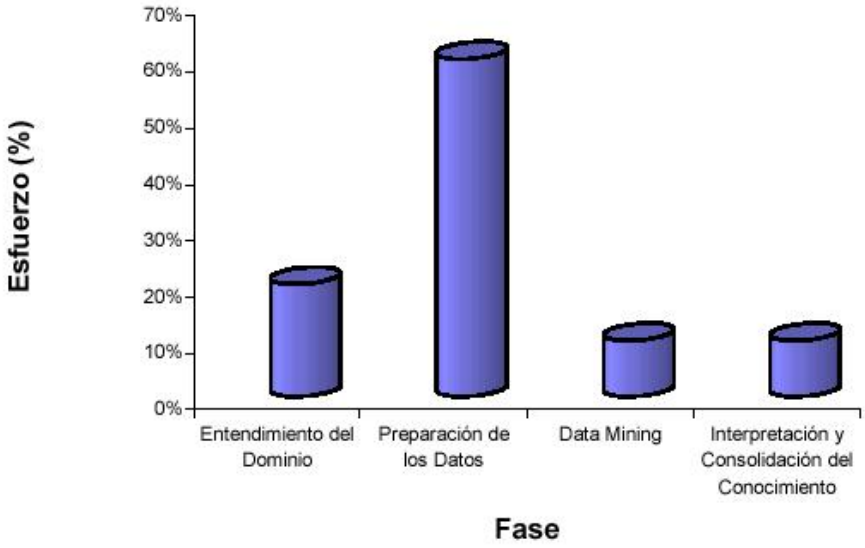


Figura 1.3: *Esfuerzo requerido por cada fase del proceso de KDD.*

Como se observa en la figura 1.3, gran parte del esfuerzo del proceso de KDD recae sobre la fase de preparación de los datos, fase crucial para tener éxito como ya se comentó anteriormente.

Indice

4. – Minería de Datos

Introducción:

Minería de Datos es un término genérico que engloba resultados de investigación, técnicas y herramientas usadas para extraer información útil de grandes bases de datos. La Minería de Datos es una parte del proceso completo de KDD.
Concretamente, el término Minería de Datos es usado comúnmente por los estadísticos, analistas de datos, y por la comunidad de administradores de sistemas informáticos como
todo el proceso del descubrimiento.
El análisis de la información recopilada (por ejemplo, en un experimento científico) es habitual que sea un proceso completamente manual (basado por lo general en técnicas estadísticas). Sin embargo, cuando la cantidad de datos de los que disponemos aumenta la resolución manual del problema se hace intratable. Aquí es donde entra en juego el conjunto de técnicas de análisis automático al que nos referimos al hablar de Minería de Datos o Data Mining.
Hasta ahora, los mayores éxitos en Minería de Datos se pueden atribuir directa o indirectamente a avances en bases de datos (un campo en el que los ordenadores superan a los humanos). No obstante, muchos problemas de representación del conocimiento y de reducción de la complejidad de la búsqueda necesaria (usando conocimiento a priori) están aún por resolver. Ahí reside el interés que ha despertado el tema entre investigadores de todo el mundo.
A continuación se presentan varias definiciones de Minería de Datos (MD):

- "MD es la extracción no trivial de información implícita, desconocida previamente, y potencialmente útil desde los datos"
- "MD es el proceso de extracción y refinamiento de conocimiento útil desde grandes bases de datos"
- "MD es el proceso de extracción de información previamente desconocida, válida y procesable desde grandes bases de datos para luego ser utilizada en la toma de decisiones"
- "MD es la exploración y análisis, a través de medios automáticos y semiautomáticos, de grandes cantidades de datos con el fin de descubrir patrones y reglas significativos"
- "MD es el proceso de planteamiento de distintas consultas y extracción de información útil, patrones y tendencias previamente desconocidas desde grandes cantidades de datos posiblemente almacenados en bases de datos".
- "MD es el proceso de descubrir modelos en los datos"

Para el estudio de la Minería de Datos se ha tomado la perspectiva orientada a datos, por dos razones. Primero porque la mayoría de los trabajos en Minería de Datos están enfocados hacia el *data warehouse* que proporciona el apoyo a la Minería de Datos organizando y estructurando los datos. Además, otras tecnologías de apoyo a la minería datos han sido utilizadas desde hace tiempo y la integración de estas tecnologías con la administración de datos ha contribuido mucho a mejorar la Minería de Datos.
Las más importantes entre estas tecnologías son los métodos estadísticos y el aprendizaje automático. Los métodos estadísticos han producido varios paquetes estadísticos para computar sumas, promedios, y distribuciones, que han ido integrándose con las bases de datos a explorar. El aprendizaje automático consiste en la obtención de reglas de aprendizaje y modelos de los datos, para lo cual a menudo se necesita la ayuda de la estadística. Por esta razón, los métodos estadísticos y el aprendizaje automático son los dos componentes más importantes de la Minería de Datos. Además existen otras tecnologías, entre las que se incluyen visualización, procesamiento paralelo, y apoyo a la toma de decisiones. Las técnicas de visualización ayudan a presentar los datos para facilitar la Minería de Datos. Las técnicas procesamiento paralelo ayudan a mejorar el rendimiento de la Minería de Datos. Los sistemas de apoyo a la toma de decisiones ayudan a discriminar los resultados y proporcionan los resultados esenciales para llevar a cabo las funciones de dirección.

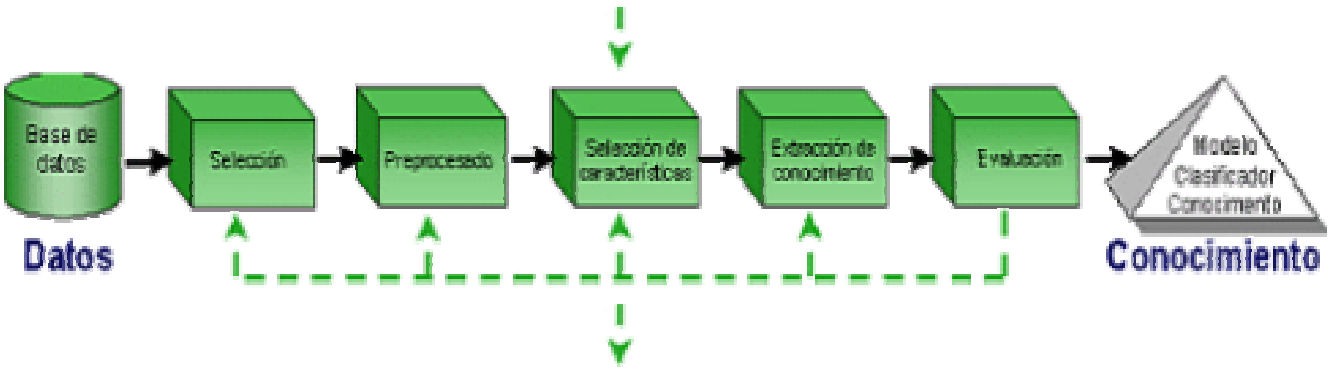
Tipología de Patrones de Minería de Datos

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. Cuidado ya que existen muchas dependencias nada interesantes (causalidades inversas). Ejemplo: que un paciente haya sido ingresado en maternidad determina su sexo.
- **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, grado 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.
- **Tendencias/Regresión:** El objetivo es predecir los valores de una variable continua a partir de la evolución sobre otra variable continua, generalmente el tiempo, o sobre un conjunto de variables. 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.
- **Información del Esquema:** (descubrir claves primarias alternativas.).
- **Reglas Generales:** patrones no se ajustan a los tipos anteriores. Recientemente los sistemas incorporan capacidad para establecer otros patrones más generales.

Fases de un Proyecto de Minería de Datos

Los pasos a seguir para la realización de un proyecto de minería de datos son siempre los mismos, independientemente de la técnica específica de extracción de conocimiento usada.



FILTRADO DE DATOS

El formato de los datos contenidos en la fuente de datos (base de datos, Data Warehouse...) nunca es el idóneo, y la mayoría de las veces no es posible ni siquiera utilizar ningún algoritmo de minería sobre los datos "en bruto".

Así que se filtran los datos (de forma que 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), o se reducen el número de valores posibles (mediante redondeo, clustering,...).

SELECCIÓN DE VARIABLES

Aún después de haber sido filtrados y limpiados los datos, en la mayoría de los casos se tiene una cantidad ingente de datos. La selección de características reduce el tamaño de los datos eligiendo las variables más influyentes en el problema, sin apenas sacrificar la calidad del modelo de conocimiento obtenido del proceso de minería.

Los métodos para la selección de características son básicamente dos:

- Los basados en la elección de los mejores atributos del problema.
- Los que buscan variables independientes mediante tests y/o algoritmos.

EXTRACCION 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. También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a una manipulación previa de los datos diferente.

INTERPRETACION Y EVALUACION

Una vez obtenido el modelo, se debe proceder a su validación, comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema. Si ninguno de los modelos alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos modelos.

TECNICAS DE MINERIA DE DATOS

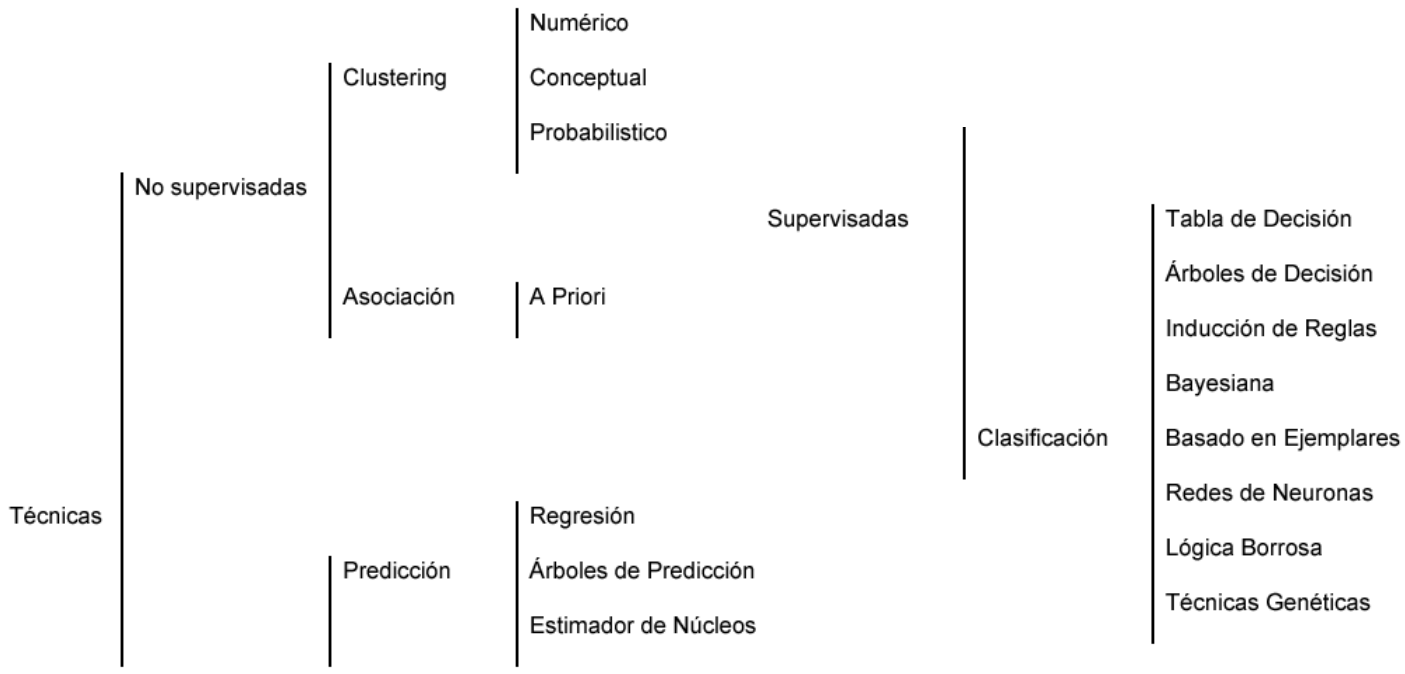


Figura 3.1: *Técnicas de la Minería de Datos*

Las técnicas de Minería de Datos se clasifican en dos grandes categorías :

- Supervisadas o Predictivas y
- No supervisadas o Descriptivas.

Una técnica constituye el enfoque conceptual para extraer la información de los datos, y, en general es implementada por varios algoritmos.
Las predicciones se utilizan para prever el comportamiento futuro de algún tipo de entidad mientras que una descripción puede ayudar a su comprensión.
De hecho, los modelos predictivos pueden ser descriptivos (hasta donde sean comprensibles por personas) y los modelos descriptivos pueden emplearse para realizar predicciones.

De esta forma, hay algoritmos o técnicas que pueden servir para distintos propósitos, por lo que la figura anterior únicamente representa para qué propósito son más utilizadas las técnicas. Por ejemplo, las redes de neuronas pueden servir para predicción, clasificación e incluso para aprendizaje no supervisado.
A continuación se presentan las principales técnicas (supervisadas y no supervisadas) de minería de datos

Clustering. ("Segmentación")

También llamada agrupamiento, permite la identificación de tipologías o grupos donde los elementos guardan gran similitud entre sí y muchas diferencias con los de otros grupos. Así se puede segmentar el colectivo de clientes, el conjunto de valores e índices financieros, el espectro de observaciones astronómicas, el conjunto de zonas forestales, el conjunto de empleados y de sucursales u oficinas, etc. La segmentación está teniendo mucho interés desde hace ya tiempo dadas las importantes ventajas que aporta al permitir el tratamiento de grandes colectivos de forma pseudoparticularizada, en el más idóneo punto de equilibrio entre el tratamiento individualizado y aquel totalmente masificado.
Las herramientas de segmentación se basan en técnicas de carácter estadístico, de empleo de algoritmos matemáticos, de generación de reglas y de redes neuronales para el tratamiento de registros. Para otro tipo de elementos a agrupar o segmentar, como texto y documentos, se usan técnicas de reconocimiento de conceptos. Esta técnica suele servir de punto de partida para después hacer un análisis de clasificación sobre los *cluster*s.
La principal característica de esta técnica es la utilización de una medida de similaridad que, en general, está basada en los atributos que describen a los objetos, y se define usualmente por proximidad en un espacio multidimensional. Para datos numéricos, suele ser preciso preparar los datos antes de realizar data mining sobre ellos, de manera que en primer lugar se someten a un proceso de estandarización.
Una de las técnicas empleadas para conseguir la normalización de los datos es utilizar la medida z *(z-score*) que elimina las unidades de los datos. Esta medida, z, es la que se muestra en la ecuación 2.1, donde *μf* es la media de la variable *f* y *ó f* la desviación típica de la misma.
Entre las medidas de similaridad destaca la distancia euclídea, ecuación 2.2.

$$z_{if} = \frac{x_{if} - \mu_f}{\sigma_f}$$

ecuación 2.1

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{n} (x_{il} - x_{jl})^2}$$

ecuación 2.2

Hay varios algoritmos de *clustering*. A continuación se exponen los más conocidos.

a) Clustering Numérico (k-medias)

Uno de los algoritmos más utilizados para hacer clustering es el k-medias (k-means), que se caracteriza por su sencillez. En primer lugar se debe especificar por adelantado cuantos clusters se van a crear, éste es el parámetro k, para lo cual se seleccionan *k* elementos aleatoriamente, que representaran el centro o media de cada cluster. A continuación cada

una de las instancias, ejemplos, es asignada al centro del cluster más cercano de acuerdo con la distancia Euclidea que le separa de él. Para cada uno de los clusters así construidos se calcula el centroide de todas sus instancias. Estos centroides son tomados como los nuevos centros de sus respectivos clusters. Finalmente se repite el proceso completo con los nuevos centros de los clusters. La iteración continúa hasta que se repite la asignación de los mismos ejemplos a los mismos clusters, ya que los puntos centrales de los clusters se han estabilizado y permanecerán invariables después de cada iteración.
El Algoritmo de K-meas es :
1. Dividir aleatoriamente los ejemplos en *k* conjuntos y calcular la media (el punto medio) de cada conjunto.
2. Reasignar cada ejemplo al conjunto con el punto medio más cercano.
3. Calcular los puntos medios de los *k* conjuntos.
4. Repetir los pasos 2 y 3 hasta que los conjuntos no varíen.

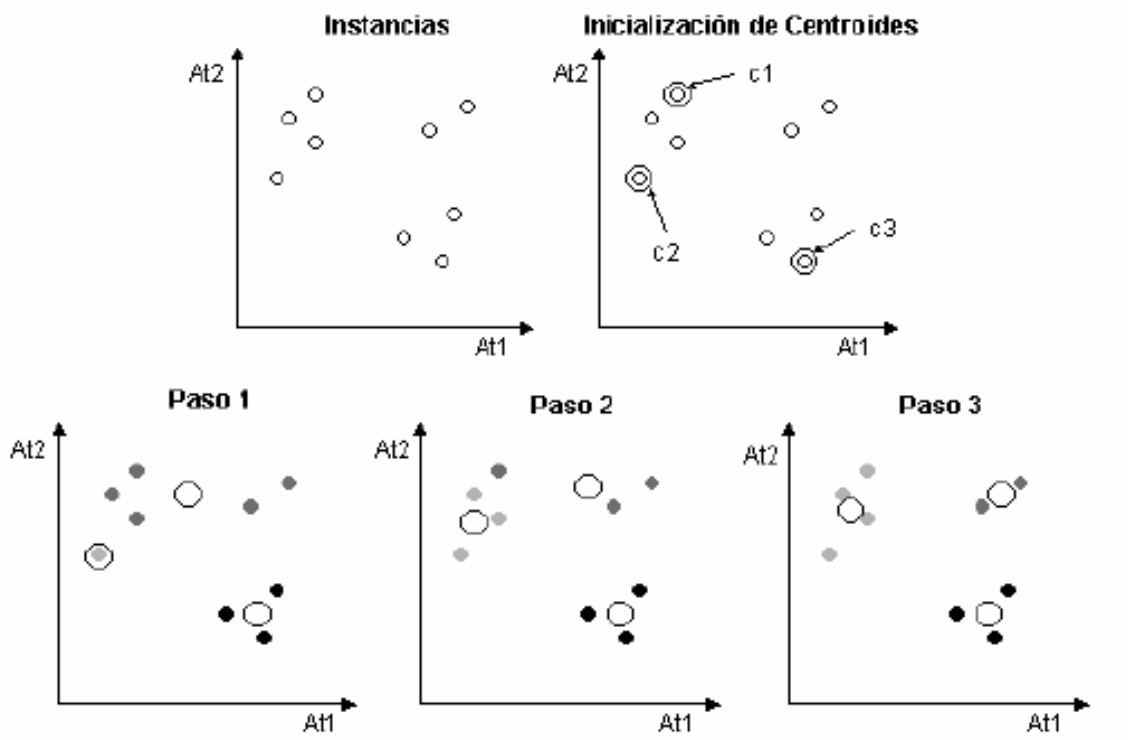


Figura 3.3: *Ejemplo de clustering con k-medias.*

b) Clustering Conceptual

El algoritmo de k-medias se encuentra con un problema cuando los atributos no son numéricos, ya que en ese caso la distancia entre ejemplares no está tan clara.
Para resolver este problema Michalski presenta la noción de clustering conceptual, que utiliza para justificar la necesidad de un clustering cualitativo frente al clustering cuantitativo, basado en la vecindad entre los elementos de la población. En buena interpretación conceptual (modelo cognitivo de jerarquías). Una de las principales motivaciones de la categorización de un conjunto de ejemplos, que básicamente supone la formación de conceptos, es la predicción de características de las categorías que heredarán sus subcategorías.

c) Clustering Probabilístico

Los algoritmos de clustering anteriores presentan ciertos defectos entre los que destacan la dependencia que tiene el resultado del orden de los ejemplos y la tendencia de estos

algoritmos al sobreajuste [overfitting]. Una aproximación estadística al problema del clustering resuelve estos problemas. La base de este tipo de clustering se encuentra en un modelo estadístico llamado mezcla de distribuciones [finite mixtures]. Cada distribución representa la probabilidad de que un objeto tenga un conjunto particular de pares atributo-valor, si se *supiera* que es miembro de ese cluster. Se tienen *k* distribuciones de probabilidad que representan los *k* clusters. La mezcla más sencilla se tiene cuando los atributos son numéricos con distribuciones gaussianas. Cada distribución (normal) se caracteriza por dos parámetros: la media (μ) y la varianza (σ^2). Además, cada distribución tendrá cierta probabilidad de aparición p, que vendrá determinada por la proporción de ejemplos que pertenecen a dicho cluster respecto del número total de ejemplos. En ese caso, si hay *k* clusters, habrá que calcular un total de *3k-1* parámetros: las *k* medias, *k* varianzas y *k-1* probabilidades de la distribución dado que la suma de probabilidades debe ser 1, con lo que conocidas *k-1* se puede determinar la *k-ésim*a.
Una vez obtenidos estos parámetros, si se deseara calcular la probabilidad de pertenencia de un determinado ejemplo de test a cada cluster, simplemente se aplicaría el teorema de Bayes.

Reglas de Asociación

Este tipo de técnicas se emplea para establecer las posibles relaciones o correlaciones entre distintas acciones o sucesos aparentemente independientes; pudiendo reconocer como la ocurrencia de un suceso o acción puede inducir o generar la aparición de otros. Son utilizadas cuando el objetivo es realizar *análisis exploratorio*s, buscando relaciones dentro del conjunto de datos. Las asociaciones identificadas pueden usarse para predecir comportamientos, y permiten descubrir correlaciones y co-ocurrencias de eventos. Debido a sus características, estas técnicas tienen una gran aplicación práctica en muchos campos como, por ejemplo, el comercial ya que son especialmente interesantes a la hora de comprender los hábitos de compra de los clientes y constituyen un pilar básico en la concepción de las ofertas y ventas cruzada, así como del "merchandising".
En otros entornos como el sanitario, estas herramientas se emplean para identificar factores de riesgo en la aparición o complicación de enfermedades. Para su utilización es necesario disponer de información de cada uno de los sucesos llevados a cabo por un mismo individuo o cliente en un determinado período temporal. Por lo general esta forma de extracción de conocimiento se fundamenta en técnicas estadísticas, como los análisis de correlación y de variación. Uno de los algoritmos mas utilizado es el algoritmo *A priori*.

a) Algoritmo A Priori

La generación de reglas de asociación se logra basándose en un procedimiento de *covering*. Las reglas de asociación son parecidas, en su forma, a las reglas de clasificación, si bien en su lado derecho puede aparecer cualquier par o pares *atributo-valor*. De manera que para encontrar ese tipo de reglas es preciso considerar cada posible combinación de pares *atributo-valor* del lado derecho. Para evaluar las reglas se emplean la medida del soporte, que indica el número de casos, ejemplos, que cubre la regla y la confianza, que indica el número de casos que predice la regla correctamente, y que viene expresado como el cociente entre el número de casos en que se cumple la regla y el número de casos en que se aplica, ya que se cumplen las premisas.

$$soporte(A \Rightarrow B) = P(A \cap B)$$

$$confianza(A \Rightarrow B) = P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

Las reglas que interesan son únicamente aquellas que tienen su valor de soporte muy alto, por lo que se buscan, independientemente de en qué lado aparezcan, pares *atributo-valor* que cubran

una gran cantidad de ejemplos. Un ejemplo típico de reglas de asociación es el análisis de la cesta de la compra. Básicamente consiste en encontrar asociaciones entre los productos que habitualmente compran los clientes.

La predicción

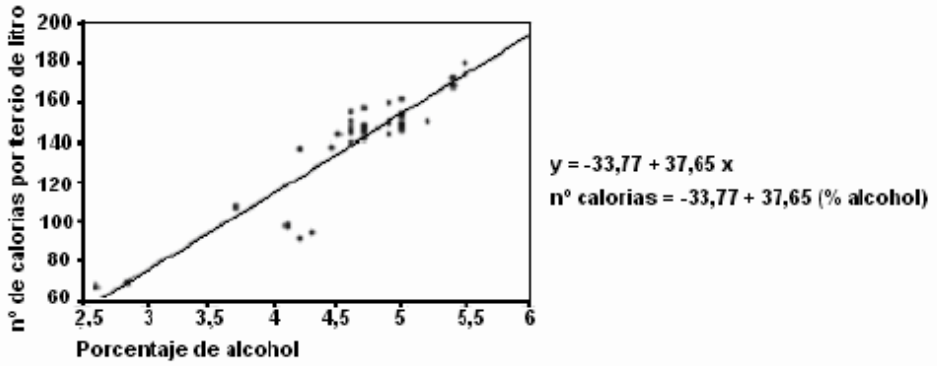
Es el proceso que intenta determinar los valores de una o varias variables, a partir de un conjunto de datos. La predicción de valores continuos puede planificarse por las técnicas estadísticas de regresión. Por ejemplo, para predecir las ventas potenciales de un nuevo producto dado su precio. Se pueden resolver muchos problemas por medio de la regresión lineal, y puede conseguirse todavía más aplicando las transformaciones a las variables para que un problema no lineal pueda convertirse a uno lineal. A continuación se presenta la regresión lineal, múltiple, y no lineal, así como la generalización a los modelos lineales.

Después, dentro de la clasificación, se ven varias técnicas de minería de datos que pueden servir para la predicción numérica. De entre todas ellas las más importantes se presentaran en la clasificación bayesiana, la basada en ejemplares y las redes de neuronas.

a) Regresión Lineal Simple

La regresión lineal es la forma más simple de regresión, ya que en ella se modelan los datos usando una línea recta. Se caracteriza, por tanto, por la utilización de dos variables, una aleatoria, *y* (llamada variable respuesta), que es función lineal de otra variable aleatoria, *x* (llamada variable predictora), formándose la ecuación $y=a+bx$

En esta ecuación la variación de *y* se asume que es constante, y *a* y *b* son los coeficientes de regresión que especifican la intersección con el eje de ordenadas, y la pendiente de la recta, respectivamente. Estos coeficientes se calculan utilizando el método de los mínimos cuadrados que minimizan el error entre los datos reales y la estimación de la línea.



Ejemplo Regresión lineal simple

b) Regresión Lineal Múltiple

La regresión Lineal Múltiple es una extensión de regresión lineal que involucra más de una variable predictora, y permite que la variable respuesta *y* sea planteada como una función lineal de un vector multidimensional.

Atributo 1 (X1)	Atributo 2 (X2)	Atributo 3 (X3)	Clase (Y)
1	3	-2	1
2	1	3	-2,3
4	6	5	-10
3	3	2	-1
2	4	-1	0,5
3	2	1	3,1

$$\Rightarrow Z=\begin{pmatrix}1&3&-2\\2&1&3\\4&6&5\\3&3&2\\1&4&-1\\3&2&1\end{pmatrix};\ Y=\begin{pmatrix}1\\-2,3\\-10\\-1\\0,5\\3,1\end{pmatrix};\ B=\begin{pmatrix}b1\\b2\\b3\end{pmatrix}$$

Paso 2. Obtención de los Coeficientes y Recta de Regresión

$$B=(Z^TZ)^{-1}Z^TY=\left[\begin{pmatrix}1&2&4&3&1&3\\3&1&6&3&4&2\\-2&3&5&2&-1&1\end{pmatrix}\begin{pmatrix}1&3&-2\\2&1&3\\4&6&5\\3&3&2\\1&4&-1\\3&2&1\end{pmatrix}\right]^{-1}\begin{pmatrix}1&2&4&3&1&3\\3&1&6&3&4&2\\-2&3&5&2&-1&1\end{pmatrix}\begin{pmatrix}1\\-2,3\\-10\\-1\\0,5\\3,1\end{pmatrix}=$$

$$\Rightarrow y=2,608x1-1,494x2-2,169x3$$

$$\begin{pmatrix}40&48&32\\48&75&31\\32&31&44\end{pmatrix}^{-1}\begin{pmatrix}-36,8\\-54,1\\-58,31\end{pmatrix}=\begin{pmatrix}0,22033&-0,1055&-0,0859\\-0,1055&0,06993&0,02788\\-0,0859&0,02788&0,06506\end{pmatrix}\begin{pmatrix}-36,8\\-54,1\\-58,31\end{pmatrix}=\begin{pmatrix}2,608\\-1,494\\-2,169\end{pmatrix}$$

Paso 3. Comprobación del ajuste de la recta de regresión

$$R^2=1-\frac{(Y-ZB)^T(Y-ZB)}{\sum_i^n(y_i-\bar{y})^2}=1-\frac{5,841}{104,595}=0,944$$

Figura 3.11: Ejemplo de obtención de una Regresión Lineal Múltiple.

c) Regresión no lineal.

En muchas ocasiones los datos no muestran una dependencia lineal. Esto es lo que sucede si, por ejemplo, la variable respuesta depende de las variables independientes según una función polinómica, dando lugar a una regresión polinómica que puede planearse agregando las condiciones polinómicas al modelo lineal básico. De está forma y aplicando ciertas transformaciones a las variables, se puede convertir el modelo no lineal en uno lineal que puede resolverse entonces por el método de mínimos cuadrados. Por ejemplo considérese una relación polinómica cúbica dada por: $y = a + b1x + b2\ x\ 2 + b3\ x\ 3$.
Para convertir esta ecuación a la forma lineal, se definen las nuevas variables:
$x1 = x\ x\ 2 = x\ 2\ x\ 3 = x\ 3$ Con lo que la ecuación anterior puede convertirse entonces a la forma lineal aplicando los cambios de variables, y resultando, que es resoluble por el método de mínimos cuadrados $y = a + b1\ x1 + b2\ x\ 2 + b3\ x\ 3.$
No obstante, algunos modelos son especialmente no lineales como, por ejemplo, la suma de términos exponenciales y no pueden convertirse a un modelo lineal. Para estos casos, puede ser posible obtener las estimaciones del mínimo cuadrado a través de cálculos extensos en formulas más complejas.

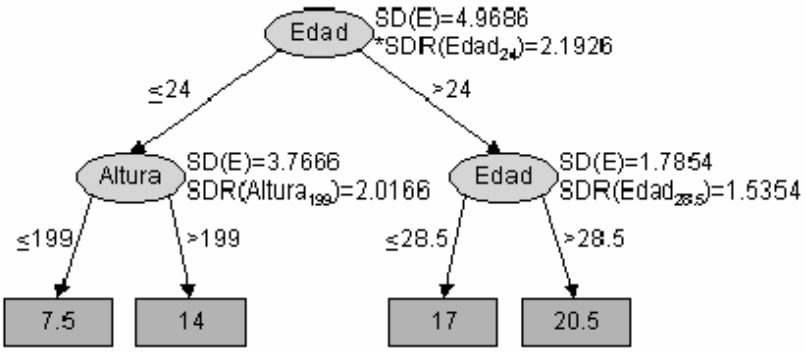
d) Árboles de Predicción

Los árboles de predicción numérica son similares a los árboles de decisión, que se verán más adelante, excepto en que la clase a predecir es continua. En este caso, cada nodo hoja almacena un valor de clase consistente en la media de las instancias que se clasifican con esa hoja, en cuyo caso estamos hablando de un *árbol de regresió*n, o bien un modelo lineal que predice el valor de la clase, en cuyo caso se habla de *árbol de modelo*s.

Ejemplos de Entrenamiento

Altura	Minutos Jugados	Edad	Clase (Puntos)
201	36	25	17
198	39	30	21
191	38	20	10
196	40	20	5
201	38	17	15
193	10	38	20
185	14	27	17
203	21	23	13

Árbol de Predicción Numérica



$$*SDR(Edad_{24}) = SD(E) - \sum_{I} \frac{|Edad_{24}|}{E} SD(Edad_{24}) = 4.9686 - \left(\frac{4}{8}3.7666 + \frac{4}{8}1.7854\right) = 2.1926$$

Figura 3.12: Ejemplo de generación del árbol de predicción con M5.

La clasificación

La clasificación es el proceso de dividir un conjunto de datos en grupos mutuamente excluyentes, de tal forma que cada miembro de un grupo esté lo mas cerca posible de otros y grupos diferentes estén lo más lejos posible de otros, donde la distancia se mide con respecto a las variables especificadas, que se quieren predecir.

Las principales técnicas de clasificación son:

a) Tabla de Decisión

La tabla de decisión constituye la forma más simple y rudimentaria de representar la salida de un algoritmo de aprendizaje, que es justamente representarlo como la entrada.

Estos algoritmos consisten en seleccionar subconjuntos de atributos y calcular su precisión (accuracy) para predecir o clasificar los ejemplos. Una vez seleccionado el mejor de los subconjuntos, la tabla de decisión estará formada por los atributos seleccionados (más la clase), en la que se insertarán todos los datos únicamente con el subconjunto de atributos elegido. Si hay dos ejemplos con exactamente los mismos pares *atributo-valor* para todos los atributos del subconjunto, la clase que se elija será la media de los ejemplos (en el caso de una clase numérica) o la que mayor probabilidad de aparición tenga (en el caso de una clase simbólica).

La precisión de un subconjunto *S* de atributos para todos los ejemplos de entrenamientos se calculará mediante la ecuación

$$precisión(S) = \frac{ejemplos\ bien\ clasificados}{ejemplos\ totales}$$

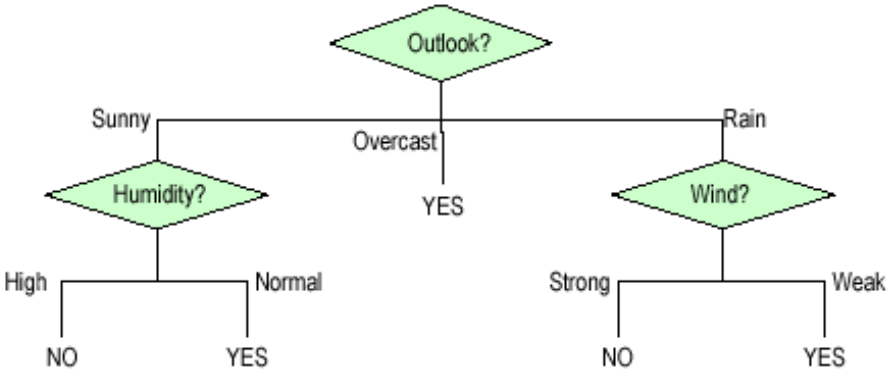
para el caso de que la clase sea simbólica o mediante la ecuación
en el caso de que la clase sea numérica:

$$precisión(S) = -RMSE = -\sqrt{\frac{\sum_{i \in I} (y_i - \hat{y}_i)^2}{n}}$$

b) Árboles de Decisión

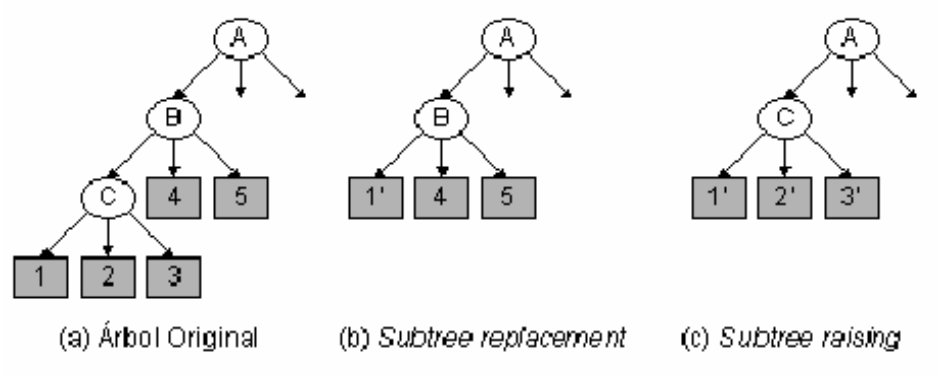
Un árbol de decisión puede interpretarse esencialmente como una serie de *reglas* compactadas para su representación en forma de árbol. Dado un conjunto de ejemplos, estructurados como vectores de pares ordenados atributo-valor, de acuerdo con el formato general en el aprendizaje inductivo a partir de ejemplos, el concepto que estos sistemas adquieren durante el proceso de aprendizaje consiste en un árbol. Cada eje está etiquetado con un par atributo-valor y las hojas con una clase, de forma que la trayectoria que determinan desde la raíz los pares de un caso de estudio alcanzan una hoja etiquetada -normalmente- con la clase del ejemplo. La clasificación de un ejemplo nuevo del que se desconoce su clase se hace con la misma técnica, solamente que en ese caso al atributo clase, cuyo valor se desconoce, se le asigna de acuerdo con la etiqueta de la hoja a la que se accede con ese caso.

- Ejemplo C4.5 con datos discretos:



.ej., la instancia:
(Outlook = sunny, Temperature = cool, Humidity = high, Wind = strong)
es NO.

El árbol de decisión ha sido construido a partir de un conjunto de ejemplos, por tanto, reflejará correctamente todo el grupo de casos. Sin embargo, como esos ejemplos pueden ser muy diferentes entre sí, el árbol resultante puede llegar a ser bastante complejo, con trayectorias largas y muy desiguales. Para facilitar la comprensión del árbol puede realizarse una *poda* del mismo. Se puede efectuar la poda después de haber desarrollado el árbol completo *(post-poda*), o realizar la construcción del árbol y la poda a la vez (pre-poda).



(a) Árbol Original (b) *Subtree replacement* (c) *Subtree raising*

c)Reglas de Clasificación

Las técnicas de Inducción de Reglas surgieron hace más de dos décadas y permiten la generación y contraste de árboles de decisión, o reglas y patrones a partir de los datos de entrada. La información de entrada será un conjunto de casos donde se ha asociado una clasificación o evaluación a un conjunto de variables o atributos. Con esa información estas técnicas obtienen el árbol de decisión o conjunto de reglas que soportan la evaluación o clasificación. En los casos en que la información de entrada posee algún tipo de “ruido" o defecto (insuficientes atributos o datos, atributos irrelevantes o errores u omisiones en los datos) estas técnicas pueden habilitar métodos estadísticos de tipo probabilístico para generar árboles de decisión recortados o podados. También en estos casos pueden identificar los atributos irrelevantes, la falta de atributos discriminantes o detectar "gaps" o huecos de conocimiento.
La inducción de reglas se puede lograr fundamentalmente mediante dos caminos: Generando un árbol de decisión y extrayendo de él las reglas, o bien mediante una estrategia de *covering*, consistente en tener en cuenta cada vez una clase y buscar las reglas necesarias para cubrir (cover) todos los ejemplos de esa clase; cuando se obtiene una regla se eliminan todos los ejemplos que cubre y se continúa buscando más reglas hasta que no haya más ejemplos de la clase.

d) Clasificación Bayesiana

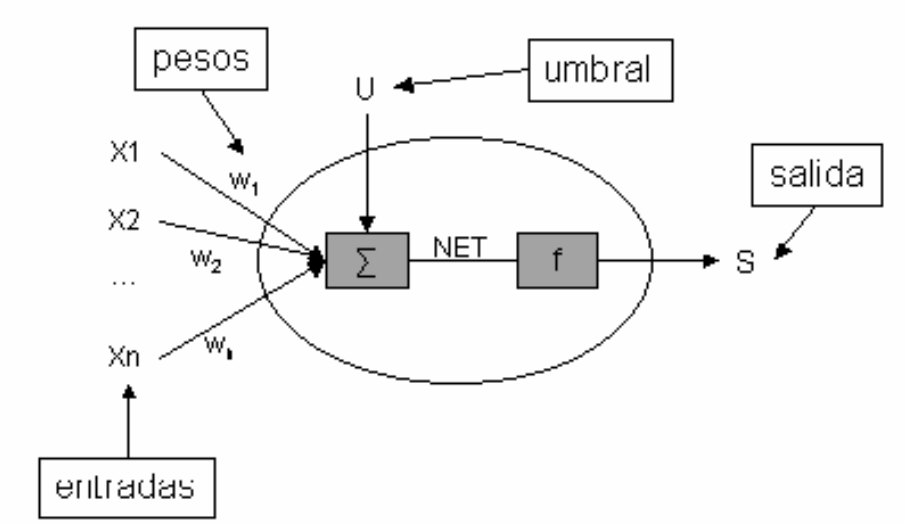
Los clasificadores Bayesianos son clasificadores estadísticos, que pueden predecir tanto las probabilidades del número de miembros de clase, como la probabilidad de que una muestra dada pertenezca a una clase particular. La clasificación Bayesiana se basa en el teorema de Bayes, y los clasificadores Bayesianos han demostrado una alta exactitud y velocidad cuando se han aplicado a grandes bases de datos.

e) Redes Neuronales

Las redes neuronales constituyen una nueva forma de analizar la información con una diferencia fundamental con respecto a las técnicas tradicionales: son capaces de detectar y aprender complejos patrones y características dentro de los datos. Se comportan de forma parecida a nuestro cerebro aprendiendo de la experiencia y del pasado, y aplicando tal conocimiento a la resolución de problemas nuevos. Presentan además, una eficiencia y fiabilidad similar a los métodos estadísticos y sistemas expertos, si no mejor, en la mayoría de los casos. En aquellos casos de muy alta complejidad las redes neuronales se muestran como especialmente útiles dada la dificultad de modelado que supone para otras técnicas. Sin embargo las redes neuronales tienen el inconveniente de la dificultad de acceder y comprender los modelos que generan y presentan dificultades para extraer reglas de tales modelos. Otra característica es que son capaces de trabajar con datos incompletos e, incluso, contradictorios lo que, dependiendo del problema, puede resultar una ventaja o un inconveniente. Las redes neuronales poseen las dos formas de aprendizaje: supervisado y no supervisado.

Estructura de las Redes de Neuronas

Las redes neuronales se construyen estructurando en una serie de niveles o capas (al menos tres: entrada, procesamiento u oculta y salida) compuestas por nodos o "neuronas", que tienen la siguiente estructura



Tanto el umbral como los pesos son constantes que se inicializarán aleatoriamente y durante el proceso de aprendizaje serán modificados. La salida de la neurona se define tal y como se muestra en las siguientes ecuaciones

$$NET = \sum_{i=1}^{N} X_i w_i + U$$

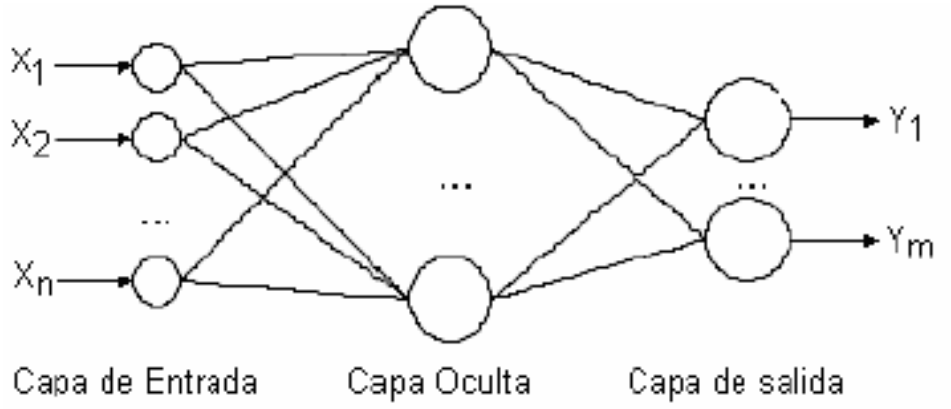
Como función f se suele emplear una función sigmoidal, bien definida entre *0* y *1* o entre *–1* y *1.*

$$S = f(NET)$$

$$f(x) = \frac{1}{1+e^{-x}}$$

$$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

Cada neurona está conectada a todas las neuronas de las capas anterior y posterior a través de los pesos o "dendritas"



Cuando un nodo recibe las entradas o "estímulos" de otras los procesa para producir una salida que transmite a la siguiente capa de neuronas. La señal de salida tendrá una intensidad fruto de la combinación de la intensidad de las señales de entrada y de los pesos que las transmiten. Los pesos o dendritas tienen un valor distinto para cada par de neuronas que conectan pudiendo así fortalecer o debilitar la conexión o comunicación

entre neuronas particulares. Los pesos son modificados durante el proceso de adiestramiento.
El diseño de la red de neuronas consistirá, entre otras cosas, en la definición del número de neuronas de las tres capas de la red. Las neuronas de la capa de entrada y las de la capa de salida vienen dadas por el problema a resolver, dependiendo de la codificación de la información. En cuanto al número de neuronas ocultas (y/o número de capas ocultas) se determinará por prueba y error. Por último, debe tenerse en cuenta que la estructura de las neuronas de la capa de entrada se simplifica, dado que su salida es igual a su entrada: no hay umbral ni función de salida.

f) Lógica Borrosa

La lógica borrosa surge de la necesidad de modelar la realidad de una forma más exacta evitando precisamente el determinismo o la exactitud. En otras palabras permite el tratamiento probabilístico de la categorización de un colectivo.
Así, para establecer una serie de grupos, segmentos o clases en los cuales se puedan clasificar a las personas por la edad, lo inmediato sería proponer unas edades límite para establecer tal clasificación de forma disjunta. Así los niños serían aquellos cuya edad fuera menor a los 12 años, los adolescentes aquellos entre 12 y 17 años, los jóvenes aquellos entre 18 y 35, las personas maduras entre 36 y 45 años y así sucesivamente. Se habrían creado unos grupos disjuntos cuyo tratamiento, a efectos de clasificación y procesamiento, es muy sencillo: basta comparar la edad de cada persona con los límites establecidos. Sin embargo enseguida se observa que esto supone una simplificación enorme dado que una persona de 16 años 11 meses y veinte días pertenecería al grupo de los adolescentes y, seguramente, es más parecido a una persona de 18 (miembro de otro grupo) que a uno de 12 (miembro de su grupo). Lógicamente no se puede establecer un grupo para cada año, dado que sí se reconocen grupos, y no muchos, con comportamientos y actitudes similares en función de la edad. Lo que implícitamente se esta descubriendo es que las clases existen pero que la frontera entre ellas no es clara ni disjunta sino “difusa” y que una persona puede tener aspectos de su mentalidad asociados a un grupo y otros asociados a otro grupo, es decir que implícitamente se está distribuyendo la pertenencia entre varios grupos. Cuando esto se lleva a una formalización matemática surge el concepto de distribución de posibilidad, de forma que lo que entendería como función de pertenencia a un grupo de edad serían unas curvas de posibilidad. Por tanto, la lógica borrosa es aquella técnica que permite y trata la existencia de barreras difusas o suaves entre los distintos grupos en los que se categoriza un colectivo o entre los distintos elementos, factores o proporciones que concurren en una situación o solución.

g) Algoritmos Genéticos

Estos algoritmos representan el modelado matemático de como los cromosomas en un alcanzan la estructura y composición más óptima en aras de la supervivencia. Los algoritmos Genéticos hacen uso de las técnicas biológicas de reproducción (mutación y cruce) para ser utilizadas en todo tipo de problemas de búsqueda y optimización. Se da la mutación cuando alguno o algunos de los genes cambian bien de forma aleatoria o de forma controlada vía funciones y se obtiene el cruce cuando se construye una nueva solución a partir de dos contribuciones procedentes de otras soluciones "padre". En cualquier caso, tales transformaciones se realizan sobre aquellos especímenes o soluciones más aptas o mejor adaptadas.
Los Algoritmos Genéticos transforman los problemas de búsqueda y optimización de soluciones en un proceso de evolución de unas soluciones de partida.
Las soluciones se convierten en cromosomas, transformación que se realiza pasando los datos a formato binario, y a los mejores se les van aplicando las reglas de evolución (funciones probabilísticas de transición) hasta encontrar la solución óptima.
El uso de estos algoritmos no está tan extendido como otras técnicas, pero van siendo cada vez más utilizados.

SECTORES QUE UTILIZAN LA MINERIA DE DATOS

La minería de datos se utilizan en diversos sectores como:

Marketing

Actualmente con la generación de los puntos de ventas informatizados y conectados a un ordenador central, y el constante uso de las tarjetas de créditos se genera gran cantidad de información que hay que analizar. Con ello se puede emplear la minería de datos para:

- Identificar patrones de compra de los clientes: Determinar cómo compran, a partir de sus principales características, conocer el grado de interés sobre tipos de productos, si compran determinados productos en determinados momentos,...
- Segmentación de clientes: Consiste en la agrupación de los clientes con características similares, por ejemplo demográficas. Es una importante herramienta en la estrategia de marketing que permite realizar ofertas acordes a diferentes tipos de comportamiento de los consumidores.
- Predecir respuestas a campañas de *mailing*: Estas campañas son caras y pueden llegar a ser molestas para los clientes a los que no le interesan el tipo de producto promocionado por lo que es importante limitarlas a los individuos con una alta probabilidad de interesarse por el producto. Está por ello muy relacionada con la segmentación de clientes.
- Análisis de cestas de la compra [market-basket analysis]: Consiste en descubrir relaciones entre productos, esto es, determinar qué productos suelen comprarse junto con otros, con el fin de distribuirlos adecuadamente.

Compañías de Seguros

En el sector de las compañías de seguros y la salud privada, se pueden emplear las técnicas de minería de datos, por ejemplo para:

- Análisis de procedimientos médicos solicitados conjuntamente.
- Predecir qué clientes compran nuevas pólizas.
- Identificar patrones de comportamiento para clientes con riesgo.
- Identificar comportamiento fraudulento.

Banca

En el sector bancario la información que puede almacenarse es, además de las cuentas de los clientes, la relativa a la utilización de las tarjetas de crédito, que puede permitir conocer hábitos y patrones de comportamiento de los usuarios. Esta información puede aplicarse para:

- Detectar patrones de uso fraudulento de tarjetas de crédito.
- Identificar clientes leales: Es importante para las compañías de cualquier sector mantener los clientes. Y es que hay estudios que demuestran que es cuatro veces más caros obtener nuevos clientes que mantener los existentes.
- Predecir clientes con probabilidad de cambiar su afiliación.
- Determinar gasto en tarjeta de crédito por grupos.
- Encontrar correlaciones entre indicadores financieros.
- Identificar reglas de mercado de valores a partir de históricos.

Telecomunicaciones

En el sector de las telecomunicaciones se puede almacenar información interesante sobre las llamadas realizadas, tal como el destino, la duración, la fecha,... en que se realiza la llamada, por ejemplo para:

- Detección de fraude telefónico: Mediante por ejemplo el agrupamiento o *clustering* se pueden detectar patrones en los datos que permitan detectar fraudes.

Medicina

También en el campo médico se almacena gran cantidad de información, sobre los pacientes, tal como enfermedades pasadas, tratamientos impuestos, pruebas realizadas, evolución,...

Se pueden emplear técnicas de minería de datos con esta información, por ejemplo, para:

- Identificación de terapias médicas satisfactorias para diferentes enfermedades.
- Asociación de síntomas y clasificación diferencial de patologías.

- Estudio de factores (genéticos, precedentes, hábitos, alimenticios,...) de riesgo para la salud en distintas patologías.
- Segmentación de pacientes para una atención más inteligente según su grupo.
- Estudios epidemiológicos, análisis de rendimientos de campañas de información, prevención, sustitución de fármacos,...
- Identificación de terapias médicas y tratamientos erróneos para determinadas enfermedades.

Industria farmacéutica

En el sector químico y farmacéutico se almacenan gran cantidad de información:

- Bases de datos de dominio público conteniendo información sobre estructuras y propiedades de componentes químicos.
- Resultados de universidades y laboratorios publicadas en revistas técnicas.
- Datos generados en la realización de los experimentos.
- Datos propios de la empresa.

Los datos son almacenados en diferentes categorías y a cada categoría se le aplica un diferente trato. Se podrían realizar, entre otras, las siguientes operaciones con la información obtenida:

- *Clustering* de moléculas: Consiste en el agrupamiento de moléculas que presentan un cierto nivel de similitud, con lo que se pueden descubrir importantes propiedades químicas.
- Búsqueda de todas las moléculas que contienen un patrón específico: Se podría introducir una subestructura (un patrón), devolviendo el sistema todas las moléculas que son similares a dicha estructura.
- Búsqueda de todas las moléculas que vincula un camino específico hacia una molécula objetivo: Realizar una búsqueda exhaustiva puede ser impracticable, por lo que se pueden usar restricciones en el espacio de búsqueda.
- Predicción de resultado de experimentos de una nueva molécula a partir de los datos almacenados: A través de determinadas técnicas de inteligencia artificial es posible predecir los resultados a nuevos experimentos a partir de los datos, con el consiguiente ahorro de tiempo y dinero.

Biología

Con la finalización en los próximos años del Proyecto Genoma Humano y el almacenamiento de toda la información que está generando en bases de datos accesibles por Internet, el siguiente reto consiste en descubrir cómo funcionan nuestros genes y su influencia en la salud. Existen nuevas tecnologías (chips de ADN, proteómica, genómica funcional, variablidad genética individual) que están posibilitando el desarrollo de una “nueva biología” que permite extraer conocimiento biomédicos a partir de bases de datos experimentales en el entorno de un ordenador básicamente mediante técnicas de minería de datos y visualización. Estos trabajos forman parte de los desarrollos de la **Bioinformática**.

Tendencias de la Minería de Datos

El interés que despierta la Minería de Datos para el análisis de la información especialmente en el área comercial hace que se busquen nuevas aplicaciones basadas en esta tecnología. Algunas de las principales nuevas aplicaciones basadas en la Minería de Datos son:

- Minería de Textos (Text Mining) surge ante el problema cada vez más apremiante de extraer información automáticamente a partir de masas de textos. Se trata así de extraer información de datos no estructurados: texto plano. Un ejemplo de aplicación basada en Minería de Textos es la generación automática de índices en documentos. Otras más complicadas consistirían en escanear completamente un texto y mostrar un mapa en el que las partes más relacionadas, o los documentos más relacionados se coloquen cerca unos de otros. En este caso se trataría de analizar las palabras en el contexto en que se encuentren.
- Minería de datos Web (Web Mining) es una tecnología usada para descubrir conocimiento interesante en todos los aspectos relacionados a la Web. Es uno de los mayores retos. El enorme volumen de datos en la Web generado por la explosión de usuarios y el desarrollo de librerías digitales hace que la extracción de la información útil sea un gran problema. Cuando el usuario navega por la web se encuentra frecuentemente saturado por los datos. La

integración de herramientas de minería de datos puede ayudar a la extracción de la información útil. La Minería de datos Web se puede clasificar en tres grupos distintos no disjuntos, dependiendo del tipo de información que se quiera extraer, o de los objetivos :

- Minería del *Contenido* de la Web (Web Content Mining): Extraer información del contenido de los documentos en la web. Se puede clasificar a su vez en:
  - Text Mining: Si los documentos son textuales (planos).
  - Hypertext Mining: Si los documentos contienen enlaces a sí mismos o a otros documentos
  - Markup Mining: Si los documentos son semiestructurados (con marcas).
  - Multimedia Mining: Para imágenes, audio, vídeo,...
- Minería de la Estructura de la Web (Web Structure Mining): Se intenta descubrir un modelo a partir de la tipología de enlaces de la red. Este modelo puede ser útil para clasificar o agrupar documentos.
- Minería del Uso de la Web (Web Usage Mining): Se intenta extraer información (hábitos, preferencias, etc. de los usuarios o contenidos y relevancia de documentos) a partir de las sesiones y comportamiento de los usuarios navegantes

Evaluación de una Herramienta para Minería de Datos

No tiene sentido preocuparse acerca de la precisión del sistema para aumentar un poco las ganancias cuando la base de datos misma está corrompida por culpa de copias y transferencias o cuando el modelo de negocio está mal definido y lleva a la empresa en la dirección equivocada.
Aunque la precisión predictiva sea la meta final de Minería de Datos, se pueden diferenciar tres medidas claves necesarias para una evaluación completa de la herramienta. Estas tres medidas son:

- **Precisión**
  La herramienta de Minería de Datos debe generar un modelo lo más preciso posible, pero reconociendo que las pequeñas diferencias en las distintas técnicas pueden deberse a fluctuaciones en muestreo aleatorio (incluso si se usa la base de datos completa para el modelo) o pueden ser despreciables en la dinámica del mercado en el que se despliegan los modelos.
- **Explicación**
  La herramienta de Minería de Datos tiene que ser capaz de **explicar** al usuario final de un modo claro cómo funciona el modelo para que pueda desarrollar la intuición. De este modo, las intuiciones y el sentido común serán fácilmente controlados y confirmados. Asimismo, la explicación del beneficio o el cálculo del rendimiento de la inversión tienen que ser fáciles y claros.
- **Integración**
  La herramienta de Minería de Datos debe integrarse en el proceso real de negocio, flujos de datos e información de la empresa. La solicitud de copias de datos y reprocesamiento masivo de datos aumenta la posibilidad de error mientras que una integración rigurosa reduce significativamente esta posibilidad.

Indice

5. - DATA WAREHOUSE

¿Qué es un Data Warehouse?

Tras las dificultades de los sistemas tradicionales en satisfacer las necesidades informacionales, surge el concepto de Data Warehouse, como solución a las necesidades informacionales globales de la empresa. Este término acuñado por Bill Inmon, se traduce literalmente como Almacén de Datos. No obstante si el Data Warehouse fuese exclusivamente un almacén de datos, los problemas seguirían siendo los mismos que en los Centros de Información.
La ventaja principal de este tipo de sistemas se basa en su concepto fundamental, la estructura de la información. Este concepto significa el almacenamiento de información homogénea y fiable, en una estructura basada en la consulta y el tratamiento jerarquizado de la misma, y en un entorno diferenciado de los sistemas operacionales. Según definió Bill Inmon, el Data Warehouse se caracteriza por ser:

- ***Integrado*:** Los datos almacenados en el Data Warehouse deben integrarse en una estructura consistente, por lo que las inconsistencias existentes entre los diversos sistemas operacionales deben ser eliminadas. La información suele estructurarse también en distintos niveles de detalle para adecuarse a las distintas necesidades de los usuarios.
- ***Temático*:** Sólo los datos necesarios para el proceso de generación del conocimiento del negocio se integran desde el entorno operacional. Los datos se organizan por temas para facilitar su acceso y entendimiento por parte de los usuarios finales. Por ejemplo, todos los datos sobre clientes pueden ser consolidados en una única tabla del Data Warehouse. De esta forma, las peticiones de información sobre clientes serán más fáciles de responder dado que toda la información reside en el mismo lugar.
- ***Histórico*:** El tiempo es parte implícita de la información contenida en un Data Warehouse. En los sistemas operacionales, los datos siempre reflejan el estado de la actividad del negocio en el momento presente. Por el contrario, la información almacenada en el Data Warehouse sirve, entre otras cosas, para realizar análisis de tendencias. Por lo tanto, el Data Warehouse se carga con los distintos valores que toma una variable en el tiempo para permitir comparaciones.

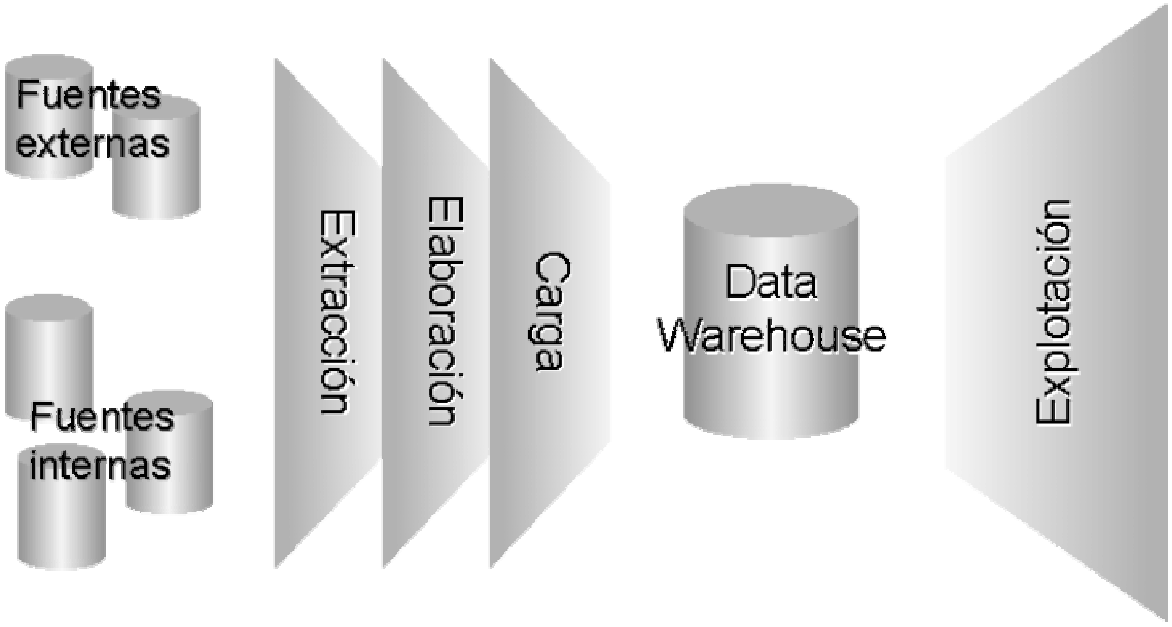
- ***No volátil*:** El almacén de información de un Data Warehouse existe para ser leído, y no modificado. La información es por tanto permanente, significando la actualización del Data Warehouse la incorporación de los últimos valores que tomaron las distintas variables contenidas en él sin ningún tipo de acción sobre lo que ya existía.

Otra característica del Data Warehouse es que contiene datos relativos a los datos, concepto que se ha venido asociando al término de metadatos. Los metadatos permiten mantener información de la procedencia de la información, la periodicidad de refresco, su fiabilidad, forma de cálculo, etc., relativa a los datos de nuestro almacén.
Estos metadatos serán los que permitan simplificar y automatizar la obtención de la información desde los sistemas operacionales a los sistemas informacionales.
Los objetivos que deben cumplir los metadatos, según el colectivo al que va dirigido, serían:

- Soportar al usuario final, ayudándole a acceder al Data Warehouse con su propio lenguaje de negocio, indicando qué información hay y qué significado tiene. Ayudar a construir consultas, informes y análisis, mediante herramientas de navegación.
- Soportar a los responsables técnicos del Data Warehouse en aspectos de auditoría, gestión de la información histórica, administración del Data Warehouse, elaboración de programas de extracción de la información, especificación de las interfaces para la realimentación a los sistemas operacionales de los resultados obtenidos, etc.

PROCESOS QUE CONFORMAN UN DW

Para comprender el concepto de Data Warehouse, es importante considerar los procesos que lo conforman.



- ***Extracción*:** obtención de información de las distintas fuentes tanto internas como externas.
- ***Elaboración*:** filtrado, limpieza, depuración, homogeneización y agrupación de la información.
- ***Carga*:** organización y actualización de los datos y los metadatos en la base de datos.
- ***Explotación*:** extracción y análisis de la información en los distintos niveles de agrupación.

Las diferencias de un Data Warehouse con un sistema tradicional

SISTEMA TRADICIONAL	DATA WAREHOUSE
• Predomina la actualización	• Predomina la consulta
• La actividad más importante es de tipo operativo (día a día)	• La actividad más importante es el análisis y la decisión estratégica
• Predomina el proceso puntual	• Predomina el proceso masivo
• Mayor importancia a la estabilidad	• Mayor importancia al dinamismo
• Datos en general desagregados	• Datos en distintos niveles de detalle y agregación
• Importancia del dato actual	• Importancia del dato histórico
• Importante del tiempo de respuesta de la transacción instantánea	• Importancia de la respuesta masiva
• Estructura relacional	• Visión multidimensional
• Usuarios de perfiles medios o bajos	• Usuarios de perfiles altos
• Explotación de la información relacionada con la operativa de cada aplicación	• Explotación de toda la información interna y externa relacionada con el negocio

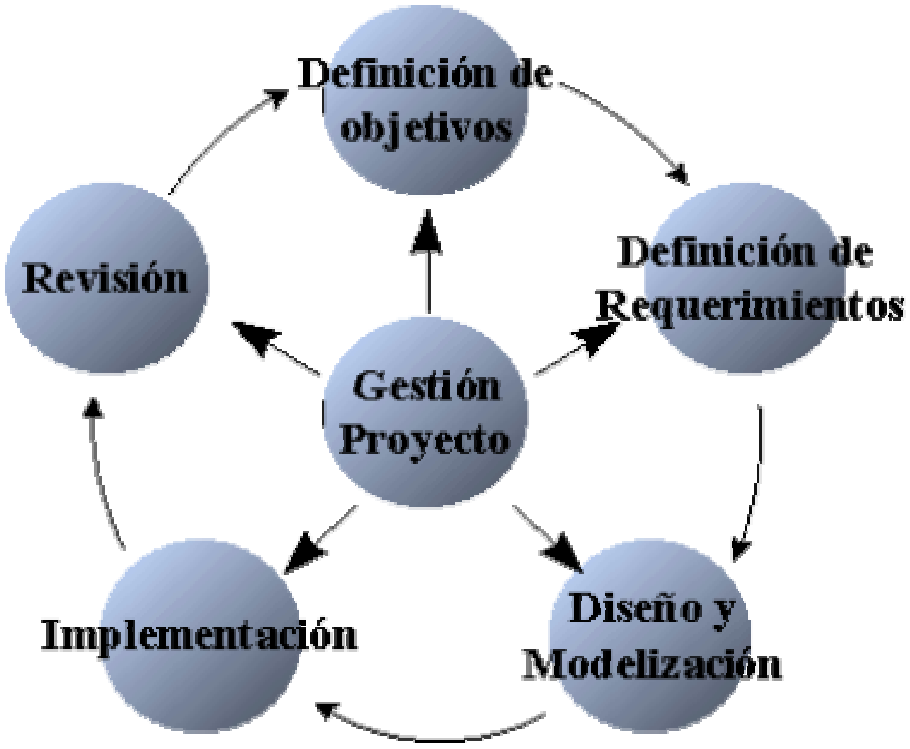
BENEFICIOS DE UN DW

- Proporciona una herramienta para la toma de decisiones en cualquier área funcional, basándose en información integrada y global del negocio.
- Facilita la aplicación de técnicas estadísticas de análisis y modelización para encontrar relaciones ocultas entre los datos del almacén; obteniendo un valor añadido para el negocio de dicha información.

- Proporciona la capacidad de aprender de los datos del pasado y de predecir situaciones futuras en diversos escenarios.
- Simplifica dentro de la empresa la implantación de sistemas de gestión integral de la relación con el cliente.
- Supone una optimización tecnológica y económica en entornos de Centro de Información, estadística o de generación de informes con retornos de la inversión espectaculares.

FASES DE IMPLANTACIÓN DE UN DATA WAREHOUSE

Tal y como aparecía en un artículo en ComputerWorld: "*Un Data Warehouse no se puede comprar, se tiene que construir*".
Planteamos aquí la metodología propuesta por SAS Institute: la "Rapid Warehousing Methodology". Dicha metodología es iterativa, y está basada en el desarrollo incremental del proyecto de Data Warehouse dividido en cinco fases:



- **Definición de los objetivos**

  En esta fase se definirá el equipo de proyecto que debe estar compuesto por representantes del departamento informático y de los departamentos usuarios del Data Warehouse además de la figura de jefe de proyecto.
  Se definirá el alcance del sistema y cuales son las funciones que el Data Warehouse realizará como suministrador de información de negocio estratégica para la empresa. Se definirán así mismo, los parámetros que permitan evaluar el éxito del proyecto.
  Definición de los requerimientos de información
  Durante esta fase se mantendrán sucesivas entrevistas con los representantes del departamento usuario final y los representantes del departamento de informática. Se realizará el estudio de los sistemas de información existentes, que ayudaran a comprender las carencias actuales y futuras que deben ser resueltas en el diseño del Data Warehouse
  Asimismo, en esta fase el equipo de proyecto debe ser capaz de validar el proceso de entrevistas y reforzar la orientación de negocio del proyecto. Al finalizar esta fase se obtendrá el documento de definición de requerimientos en el que se reflejarán no solo las

necesidades de información de los usuarios, sino cual será la estrategia y arquitectura de implantación del Data Warehouse.

- **Diseño y modelización**

  Los requerimientos de información identificados durante la anterior fase proporcionarán las bases para realizar el diseño y la modelización del Data Warehouse.

  En esta fase se identificarán las fuentes de los datos (sistema operacional, fuentes externas,..) y las transformaciones necesarias para, a partir de dichas fuentes, obtener el modelo lógico de datos del Data Warehouse. Este modelo estará formado por entidades y relaciones que permitirán resolver las necesidades de negocio de la organización.

  El modelo lógico se traducirá posteriormente en el modelo físico de datos que se almacenará en el Data Warehouse y que definirá la arquitectura de almacenamiento del Data Warehouse adaptándose al tipo de explotación que se realice del mismo.

  La mayor parte estas definiciones de los datos del Data Warehouse estarán almacenadas en los metadatos y formarán parte del mismo.

- **Implementación**

  La implantación de un Data Warehouse lleva implícitos los siguientes pasos:

  - Extracción de los datos del sistema operacional y transformación de los mismos.
  - Carga de los datos validados en el Data Warehouse. Esta carga deberá ser planificada con una periodicidad que se adaptará a las necesidades de refresco detectadas durante las fases de diseño del nuevo sistema.
  - Explotación del Data Warehouse mediante diversas técnicas dependiendo del tipo de aplicación que se de a los datos:
    - Query & Reporting
    - On-line analytical processing (OLAP)
    - Executive Information System (EIS) ó Información de gestión
    - Decision Support Systems (DSS)
    - Visualización de la información
    - Data Mining ó Minería de Datos, etc.

  La información necesaria para mantener el control sobre los datos se almacena en los metadatos técnicos (cuando describen las características físicas de los datos) y de negocio (cuando describen cómo se usan esos datos). Dichos metadatos deberán ser accesibles por los usuarios finales que permitirán en todo momento tanto al usuario, como al administrador que deberá además tener la facultad de modificarlos según varíen las necesidades de información.

  Con la finalización de esta fase se obtendrá un Data Warehouse disponible para su uso por parte de los usuarios finales y el departamento de informática.

- **Revisión**

  La construcción del Data Warehouse no finaliza con la implantación del mismo, sino que es una tarea iterativa en la que se trata de incrementar su alcance aprendiendo de las experiencias anteriores.

- **Diseño de la estructura de cursos de formación**

  Con la información obtenida de reuniones con los distintos usuarios se diseñarán una serie de cursos a medida, que tendrán como objetivo el proporcionar la formación estadística necesaria para el mejor aprovechamiento de la funcionalidad incluida en la aplicación. Se realizarán prácticas sobre el desarrollo realizado, las cuales permitirán fijar los conceptos adquiridos y servirán como formación a los usuarios.

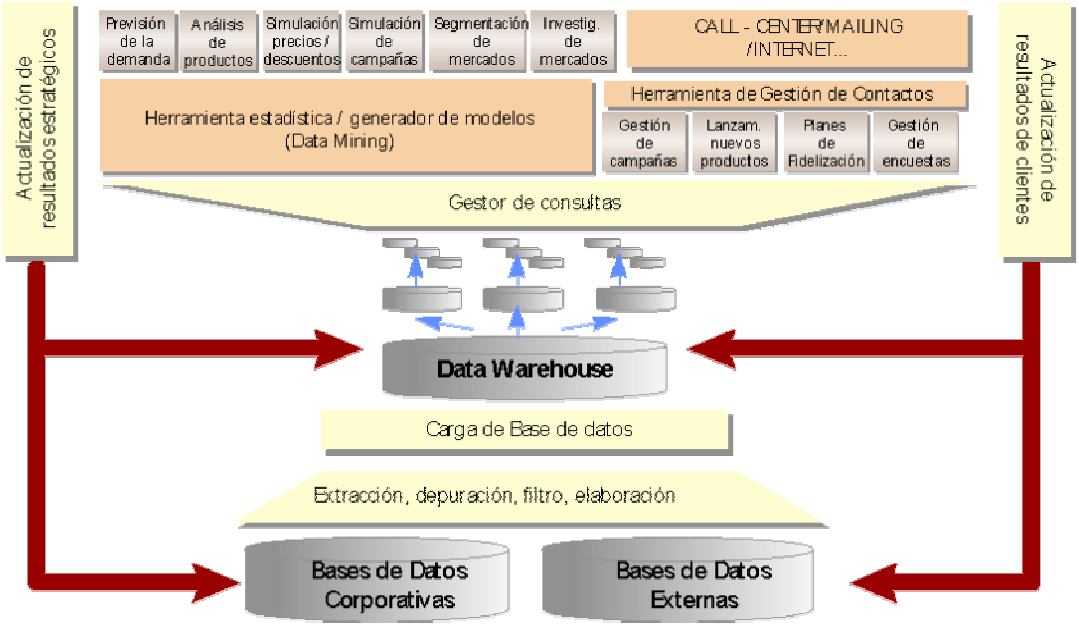
DATA MART

En un contexto de Data Warehouse, el término duplicación se refiere a la creación de Data Marts locales o departamentales basados en subconjuntos de la información contenida en el Data Warehouse central o maestro.
Los Data Marts, tienen las mismas características de integración, no-volatilidad, orientación temática y no-volatilidad que el Data Warehouse. Representan una estrategia de "divide y vencerás" para ámbitos muy genéricos de un Data Warehouse.
Esta estrategia es particularmente apropiada cuando el Data Warehouse central crece muy rápidamente y los distintos departamentos requieren sólo una pequeña porción de los datos contenidos en él. La creación de estos Data Marts requiere algo más que una simple réplica de los datos: se necesitarán tanto la segmentación como algunos métodos adicionales de consolidación.

TIPOS DE APLICACIONES EN LAS QUE UTILIZAR LAS TÉCNICAS DISPONIBLES SOBRE EL DW

- **Marketing**

La aplicación de tecnologías de Data Warehouse supone un nuevo enfoque de Marketing, haciendo uso del Marketing de Base de Datos. En efecto, un sistema de Marketing Warehouse implica un marketing científico, analítico y experto, basado en el conocimiento exhaustivo de clientes, productos, canales y mercado.
Este conocimiento se deriva de la disposición de toda la información necesaria, tanto interna como externa, en un entorno de Data Warehouse, persiguiendo con toda esta información, la optimización de las variables controladas del Marketing Mix y el soporte a la predicción de las variables no controlables (mediante técnicas de Data Mining). Basándose en el conocimiento exhaustivo de los clientes se consigue un tratamiento personalizado de los mismos tanto en el día a día (atención comercial) como en acciones de promoción específicas.



Las áreas en las que se puede aplicar las tecnologías de Data Warehouse a Marketing son, entre otras:

- Investigación Comercial
- Segmentación de mercados

- Identificación de necesidades no cubiertas y generación de nuevos productos, o modificación de productos existentes
- Fijación de precios y descuentos
- Definición de la estrategia de canales de comercialización y distribución
- Definición de la estrategia de promoción y atención al cliente
- Relación con el cliente:
- Programación, realización y seguimiento de acciones comerciales
- Lanzamiento de nuevos productos
- Campañas de venta cruzada, vinculación, fidelización, etc.
- Apoyo al canal de venta con información cualificada

- **Análisis de Riesgo Financiero**

El Data Warehouse aplicado al análisis de riesgos financieros ofrece capacidades avanzadas de desarrollo de aplicaciones para dar soporte a las diversas actividades de gestión de riesgos. Es posible desarrollar cualquier herramienta utilizando las funciones que incorpora la plataforma, gracias a la potencionalidad estadística aplicada al riesgo de crédito.
Así se puede usar para llevar a cabo las siguientes funcionalidades:

- **Para la gestión de la posición:** Determinación de la posición, Cálculo de sensibilidades, Análisis what/if, Simulaciones, Monitorización riesgos contra límites, etc.
- **Para la medición del riesgo:** Soporte metodología RiskMetrics (Metodología registrada de J.P. Morgan / Reuters), Simulación de escenarios históricos, Modelos de covarianzas, Simulación de Montecarlo, Modelos de valoración, Calibración modelos valoración, Análisis de rentabilidad, Establecimiento y seguimiento. de límites, Desarrollo/modificación modelos, Stress testing, etc.

- **Análisis de Riesgo de Crédito**

La información relativa a clientes y su entorno se ha convertido en fuente de prevención de Riesgos de Crédito. En efecto, existe una tendencia general en todos los sectores a recoger, almacenar y analizar información crediticia como soporte a la toma de decisiones de Análisis de Riesgos de Crédito.
Los avances en la tecnología de Data Warehouse hacen posible la optimización de los sistemas de Análisis de Riesgo de Crédito:
Para la gestión del riesgo de crédito los sistemas operacionales han ofrecido:

- Sistemas de Información para Gerencia (MIS) e informes de Soporte a la Decisión de Problemas (DSS) estáticos y no abiertos a nuevas relaciones y orígenes de datos, situación en la que la incorporación de nuevas fuentes de información ha sido un problema en lugar de una ventaja.
- Exploraciones de datos e informes cerrados y estáticos.
- Análisis sin inclusión de consideraciones temporales lo que imposibilita el análisis del pasado y la previsión del futuro.
- Herramientas de credit-scoring no flexibles, construidas sobre algoritmos difícilmente modificables, no adaptados al entorno de la empresa, o exclusivamente basados en la experiencia personal no contrastada, con lo que los sistemas han ayudado a repetir los errores en vez de a corregirlos.

- **Otras áreas de aplicación**

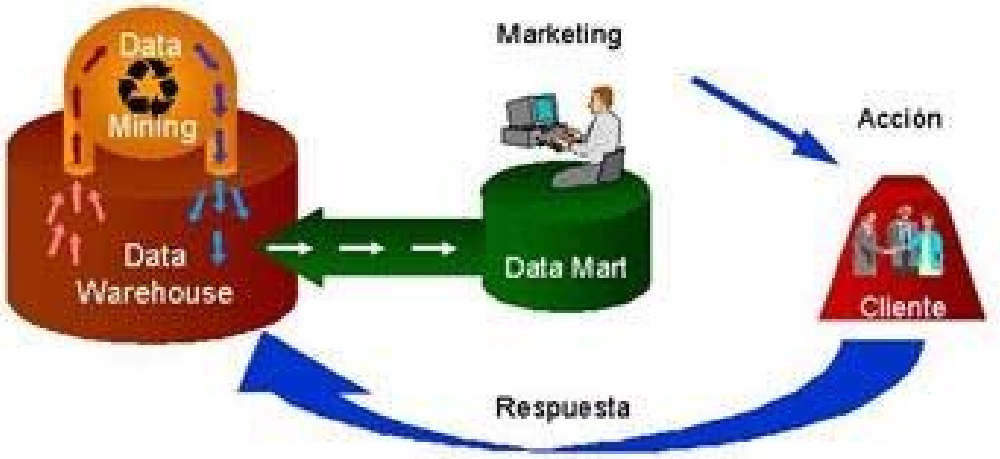
Otras áreas de la empresa han aplicado las soluciones que proporciona la tecnología Data Warehouse para mejorar gran parte de sus procesos actuales. Entre ellas destacamos:

- **Control de Gestión:**
  Sistemas de Presupuestación, Análisis de Desviaciones, Reporting (EIS, MIS, etc.)
- **Logística:**

Mejora de la relación con proveedores, Racionalización de los procesos de control de inventarios, Optimización de los niveles de producción, Previsión de la demanda en infraestructura.

- **Recursos Humanos**
  Planificación de incorporaciones, Gestión de carreras profesionales, Asignación de recursos a proyectos alternativos, etc.

Situación **IDEAL** de los flujos de datos dentro de una empresa



OLAP

Los sistemas de soporte a la decisión usando tecnologías de Data Warehouse, se llaman sistemas OLAP (siglas de On Line Analytical Processing (OLAP). En general, estos sistemas OLAP deben:

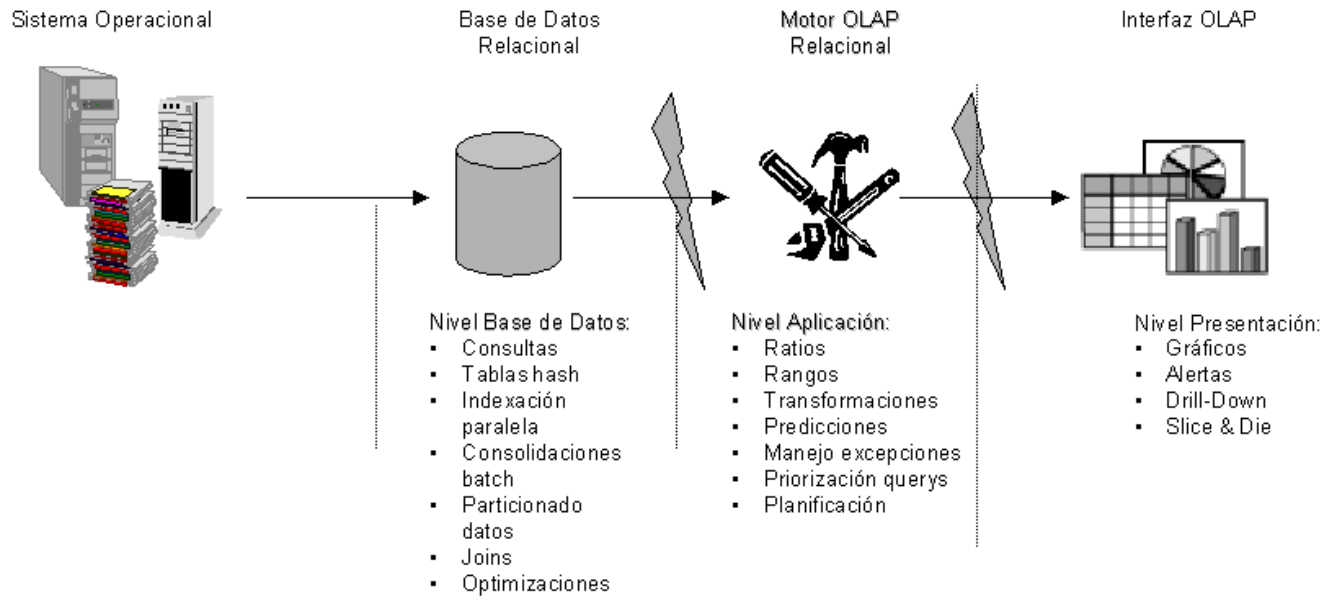
- Soportar requerimientos complejos de análisis
- Analizar datos desde diferentes perspectivas
- Soportar análisis complejos contra un volumen ingente de datos

La funcionalidad de los sistemas OLAP se caracteriza por ser un análisis multidimensional de datos corporativos, que soportan los análisis del usuario y unas posibilidades de navegación, seleccionando la información a obtener.

Normalmente este tipo de selecciones se ve reflejada en la visualización de la estructura multidimensional, en unos campos de selección que nos permitan elegir el nivel de agregación (jerarquía) de la dimensión, y/o la elección de un dato en concreto, la visualización de los atributos del sujeto, frente a una(s) dimensiones en modo tabla, pudiendo con ello realizar, entre otras las siguientes acciones:

- **Rotar (*Swap*):** alterar las filas por columnas (permutar dos dimensiones de análisis)
- **Bajar (*Down*):**bajar el nivel de visualización en las filas a una jerarquía inferior
- **Detallar (*Drilldown*):**informar para una fila en concreto, de datos a un nivel inferior
- **Expandir (*Expand*):** id. anterior sin perder la información a nivel superior para éste y el resto de los valores
- **Colapsar (*Collapse*):**operación inversa de la anterior.

Existen dos arquitecturas diferentes para los sistemas OLAP: OLAP multidimensional (MOLAP) y OLAP relacionales (ROLAP).



Indice

6. - Análisis de Datos

Origen de Datos de nuestro proyecto:

Los datos facilitados por la compañía han sido:

- Listado Clientes general
- Ventas de la Familia de productos a estudiar de los años 2002 y 2003
- Compras de la Familia de productos a estudiar de los años 2002 y 2003
- Tarifa de precios de la familia de productos a estudiar de los años 2002 y 2003
- Listado Familia de productos con la división en subfamilias.
- Listado de Ventas con el código del cliente ordenado por cliente.

De estos listados entregados en Excel, tanto los de ventas como los de compras que estaban por separado para los diferentes años, he procedido a unirlos en un único listado, también he realizado el cambio de formato de fecha y he repasado y limpiado los mismos a fin de borrar aquellos que no tenían todos los campos significativos.
Con el Listado de Clientes la tarea a sido más ardua primeramente porque se trataba de 15.000 líneas, y porque el estado en que se encontraba no era muy bueno. He tenido que eliminar aquellos clientes que no eran de España, ya que nuestro cliente quiere que concentremos el estudio en nuestro País. Se ha tenido que incluir en el campo Provincia muchas de ellas que por ejemplo al tratarse de capitales de provincia no se habían incluido. También faltaba gran cantidad de códigos postales.
Asimismo he creado un nuevo listado basándome en las ventas donde se relacionen los vendedores existentes en la empresa con la sección. Ya que esto no se me había facilitado.
Por ultimo he unido el listado de productos dividido por subfamilias con el de precio a fin de no tener tantos datos desperdigados.
Ya con los datos en unas condiciones bastantes aceptables he procedido a realizar la importación de cada listado por separado y he creado una base de datos.
También se ha tenido que solicitar bastante más información de la inicial a fin de poder realizar hipótesis y suposiciones validas para poder extraer reglas eficientes.
Se ha procedido en un segundo paso a convertir todos los ficheros de Excel a una base de datos previa de Acces a fin de volver ha realizar nuevos filtros mediante consultas ya que aunque se

había realizado una limpieza inicial, en los primeros análisis se vio que existía mucho "ruido" y datos deficientes o que no existían.
Así mismo se solicito más información por ejemplo en cuanto a la cuestión de los vendedores ya que existían duplicados y faltaban datos así como de las secciones, creando con la nueva información una nueva tabla más clara.
También se concreto focalizar los puntos de búsqueda en cuanto a todo lo relacionado con las Ventas pero concentrando esfuerzos en extraer conocimiento de las zonas de acción de la compañía como Barcelona, Madrid y Bilbao, así mimo se interesaron por zonas cercanas como Cataluña en general y Levante. De las ventas también la periodicidad o la diferencia anual o mensual será de interés.
Respecto al tema de producto poder extraer información respecto a consumos de subfamilias, y ver si existe relación de esta con las zonas (o vendedores) y con la periodicidad, y como no si se pudiera con el precio.
Evidentemente toda información extra que podamos aportar será bien recibida aunque se entiende que la falta de datos y la mala calidad de estos esta siendo determinante a la hora de no conseguir los objetivos inicialmente propuestos.

Que ofrece el programa Synera:

El programa Synera trabaja con :

- Todos los datos de la base introducida y no con una muestra de los mismos como pasa con otros programas.
- En muchos de los análisis nos provee de parámetros para el mismo, sin necesidad de tener que realizar nosotros el posible calculo, aún así existe casi en todos la opción de poder ser el usuario también el que indique los parámetros.
- El poder crear relaciones entre tablas de diferentes datos, a fin de poder hacer extensivo el análisis a estas a través de estas relaciones.

Los diferentes métodos de análisis que usa el Synera son:

- El Análisis Asociativo, donde aplica el análisis combinatorio.
- La Segmentación, usa estadísticas (el algoritmo K-Means).

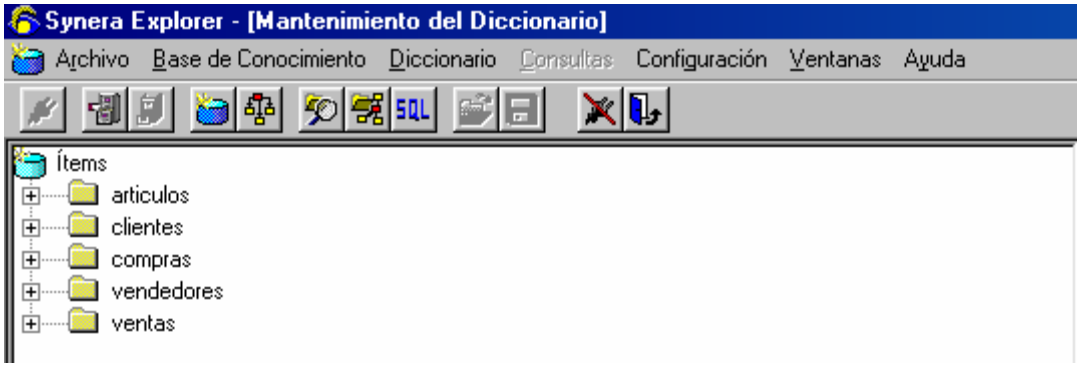
Por otra parte tenemos la aplicación del Synera Discovery, que es la que nos permite realizar los procesos de Data Mining(Minería de Datos). En la que usa los siguientes métodos explicados teóricamente en el apartado 3.

- Técnica de Cluster
- Análisis Asociativo (MBA)
  Se debe binarizar los links antes de proceder a ejecutar dichos análisis ya que los resultados de esta forma están garantizados. En este análisis tenemos la posibilidad de cambiar:
    - Soporte y Confianza
    - Links a incluir en el análisis
    - Incorporación o no de los valores nulos.

Solo es posible realizarlos sobre links numéricos.

Creación de la Base de Conocimiento en el Synera:

La importación la he realizado inicialmente del Excel creando los diferentes Links y atributos. Aunque luego a fin de que la limpieza de datos fuera más efectiva he creado una base de datos previa en el Acces y con la misma he pasado ha realizar unos nuevos filtros mediante el uso de las consultas que me han resultado más efectivos.



Mi premisa ha sido después de varias cargas infructuosas y numerosos problemas, la sencillez. Como se ve tengo los links : Ventas, Compras, Clientes, Artículos y Vendedores.

Resumen de nuestra Base de datos indicando las características de cada link y atributos:

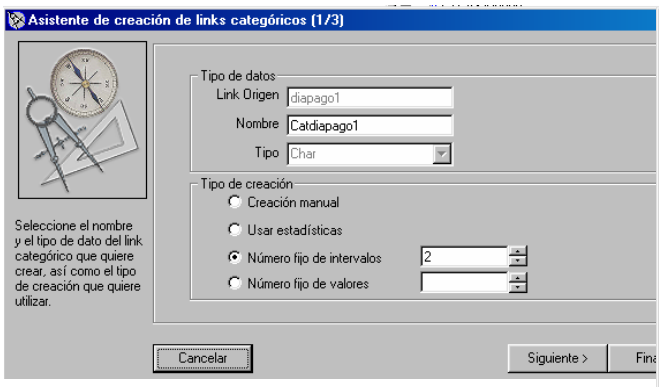
ARTICULOS		**DATOS GENERALES DE LOS ARTICULOS**
CODIGO	Char	Código que identifica al articulo
NOMBRE	Char	Descripción del articulo
FAMILIA	Integer	Código numérico de la familia
SUBFAMILIA	Integer	Código numérico que especifica las subfamilias
PVP2002	Integer	Precio Venta del articulo en el año 2002
PVP2003	Integer	Precio Venta del articulo en el año 2002
CLIENTES		**DATOS GENERALES DE LOS CLIENTES**
REFCLIENTE	Char	Código que identifica al cliente
CP	Char	Código Postal del cliente
POBLACION	Char	Población del Cliente
PROVINCIA	Char	Provincia del Cliente
VENDEDOR	Char	Código del vendedor que abrió la ficha
CREDITO	Integer	Importe del crédito que tiene el cliente
DTO	Integer	Descuento concedido al cliente
DIAPAGO1	Integer	Día de pago que posee el cliente
DIAPAGO2	Integer	Segundo día de pago que puede poseer el cliente
FECHALTA	DateTime	Fecha de apertura del cliente
PORTES	Char	Tipo de portes que posee (Debidos o Pagados)
TIPOFAC	Integer	Numero que identifica el tipo de facturación que posee el cliente puede ser 01, 02, 03, contado, recibo, pagare
TIPIVA	Integer	Numero que identifica el tipo de iva, 00 (sin iva extranjero), 01(7% IVA), 03(16% IVA) y 04(Reducido 4% , Canarias)
COMPRAS		**DATOS DE LAS COMPRAS AÑOS 2002 y 2003**
ARTICULO	Char	Código que identifica al articulo comprado
DESCRIPCION	Char	Descripción del articulo comprado
PROVEEDOR	Char	Código que identifica al proveedor
FECHACOMPRA	DateTime	Fecha de la compra
ALBARANCOMPRA	Char	Código del albarán de compra

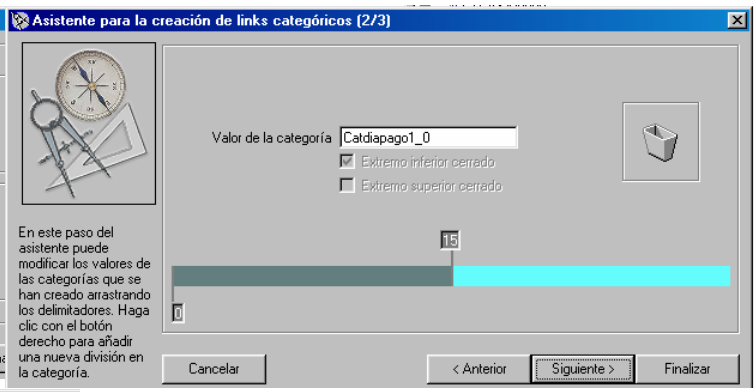
UNIDS	Integer	Unidades del articulo compradas
PUCOSTE	Integer	Precio de Coste Unitario del articulo
IMPORTE	Integer	Importe de la compra (Unidades x Precio Coste Unitario)
FACTURACOMPRA	Char	Código de la factura de compra
VENDEDORES		**DATOS DE LOS VENDEDORES**
VENDEDOR	Char	Código que identifica al vendedor
NOMBRE	Char	Nombre del Vendedor
SEXO	Char	Sexo del Vendedor
SECCION	Char	Código de la sección a que pertenece
NOMBRESECCION	Char	Nombre de la Sección a la que pertenece
VENTAS		**DATOS DE LAS VENTAS AÑOS 2002 y 2003**
VDOR	Char	Código que identifica al vendedor que ha realizo la venta
CODCLIENTE	Char	Código que identifica al cliente que ha realizo la compra
CODARTICULO	Char	Código que identifica al articulo vendido
DEFINICION	Char	Descripción del articulo vendido
FECHAVENTA	DateTime	Fecha de la venta
ALBARANVENTA	Char	Código del albarán de venta
SECCION	Char	Código de la sección a que pertenece la venta
UNIDADES	Integer	Unidades del articulo vendidas
PUVENTA	Integer	Precio de Venta Unitario del articulo
IMPORTEVENTA	Integer	Importe de la venta(Unidades x Precio Venta Unitario)
FACTURAVENTA	Char	Código de la factura de venta

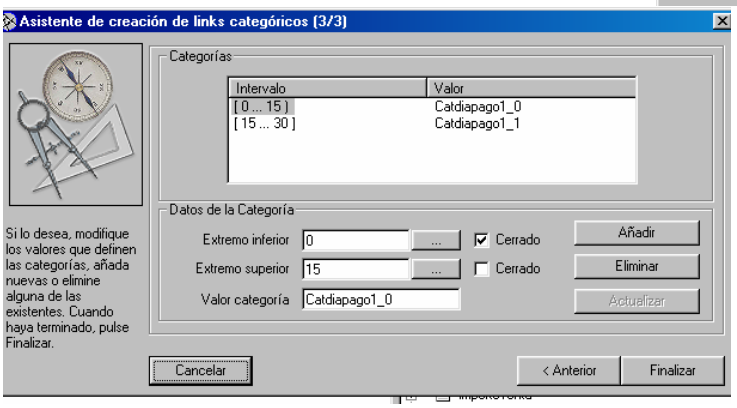
Antes de proceder a relacionar los links, he realizado una categorización de todos los ítems numéricos, ya que para los posteriores análisis en el Synera Discovery me serán necesarios.

Categorización de ítems:

Pongo a continuación un ejemplo, ya que así es como procedo en la mayoría de ítems numéricos, aunque en algunos casos realizo un cambio de intervalos manual, o elijo los extremos, aunque siempre realizo una categorización con dos valores fijos (0, y 1) a fin de crear la binarización necesaria para el proceso de análisis del Synera Discovery.

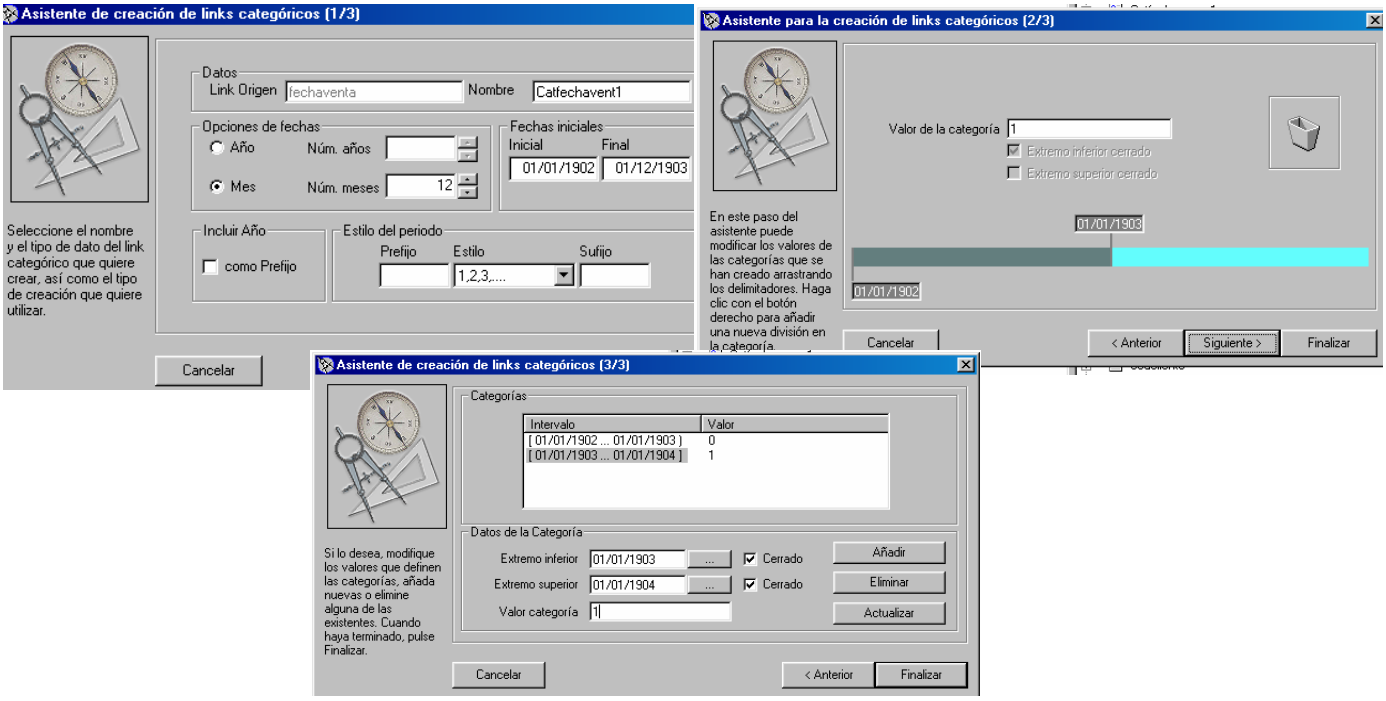




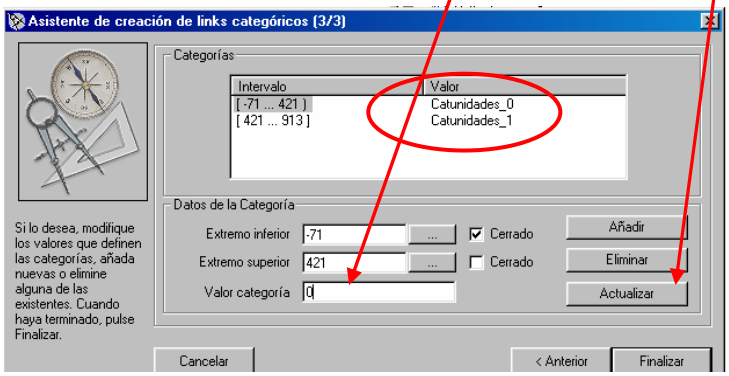


Con los ítems que no son numéricos, pero que son fechas, también intento una categorización, (los años aunque he cambiado en varias ocasiones la configuración a fin de que coja el 2000, sigue poniendo el 1900, por lo cual en realidad cuando sale 1901 este es el 2001, 1902 este es el 2002, y el 1903 este es el 2003).

Por ejemplo en el caso de ventas y compras creo una categorización dividiendo en dos intervalos, uno correspondiente al año 2002 (1902) que vale 0 y el otro al año 2003 (1903) que vale 1, también realizo la misma categorización poniendo como valor el año.

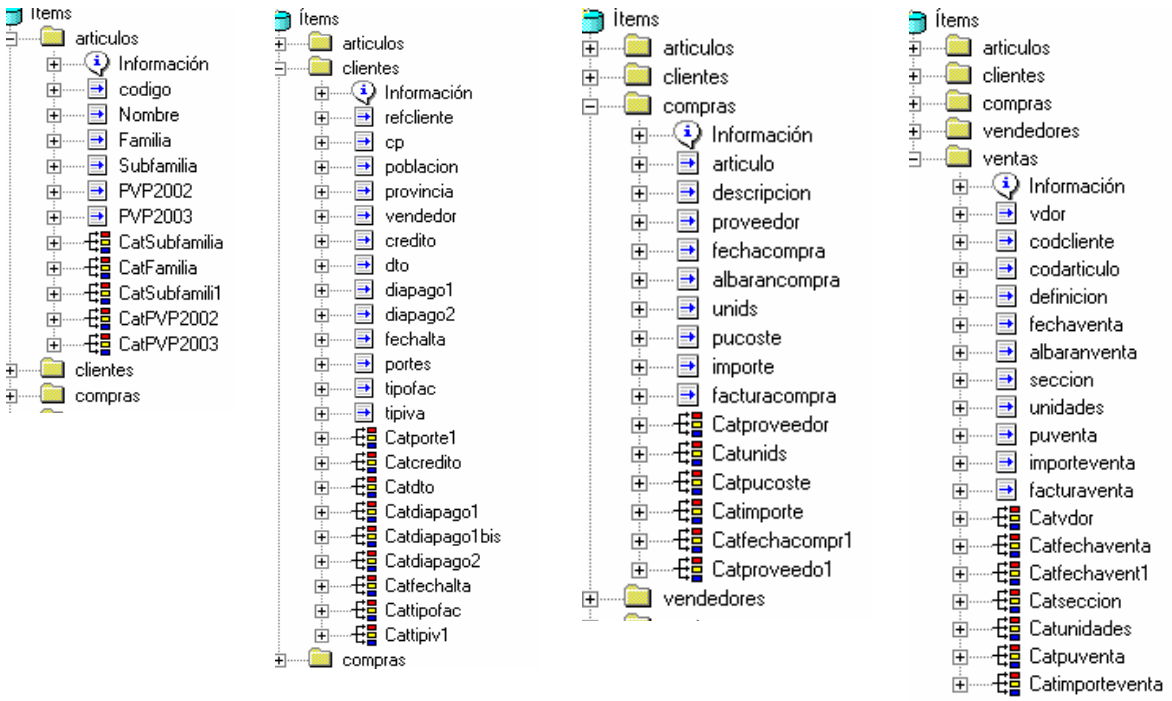


Un detalle importante es que en el ultimo paso de las categorizaciones, la mayoría de veces te pone en valor el nombre de la categorización (ejemplo CatImporte_0), evidentemente lo idóneo es que el valor sea un numero = 0 ó 1 para que los análisis sean validos, por tanto procedo al cambio marcando el intervalo y cambiándolo en la parte inferior, pero si antes de pasar a cambiar el otro intervalo, no se le da a actualizar, no procede a realizar el cambio



También en algunos casos de ítems que son char, procedo a realizar una categorización siempre y cuando puedo crear dos intervalos o tienen pocos valores.

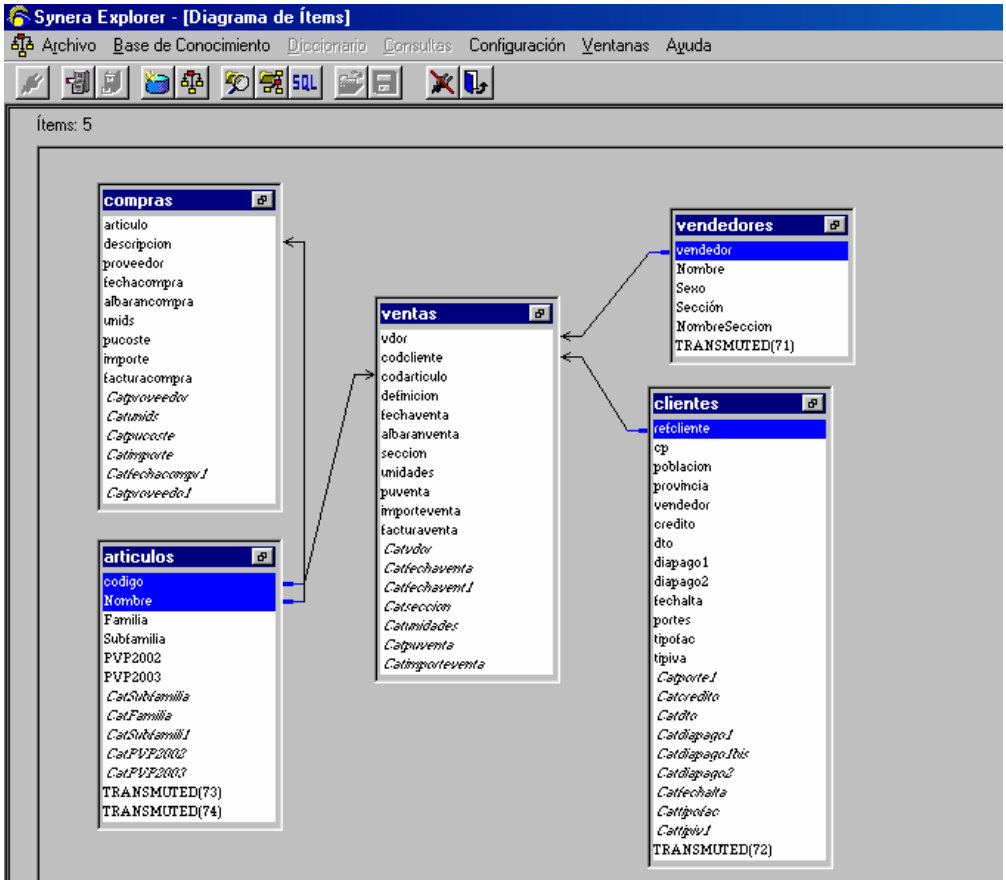
De cada link las categorizaciones realizadas son :



*Nota: En el link de vendedores al ser todos Char no he podido categorizar ninguno.

Relaciones entre ítems en el Synera Explorer:

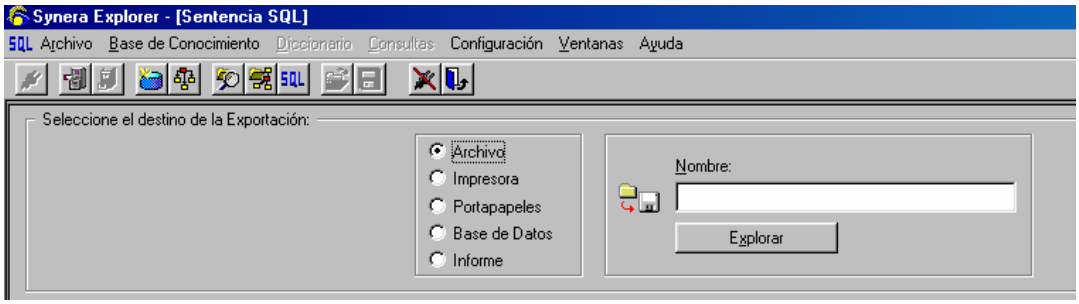
A continuación se observa las relaciones creadas entre ítems



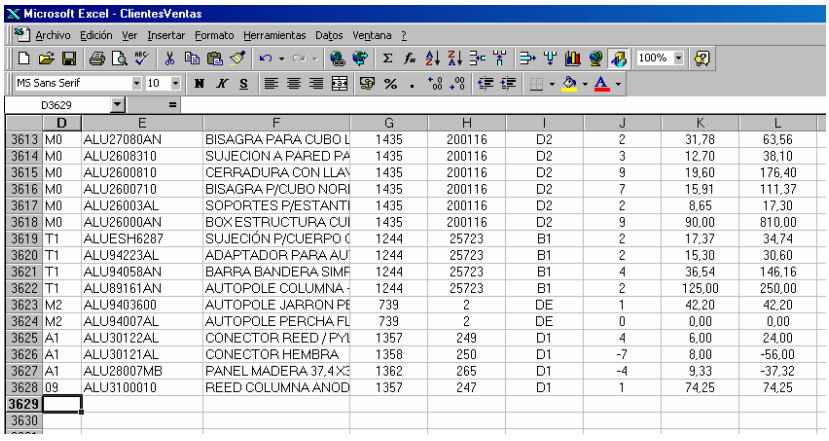
En todas ellas se ha realizado sin marcar la opción de transmutación. Pero aún así surge el atributo “Transmuted”

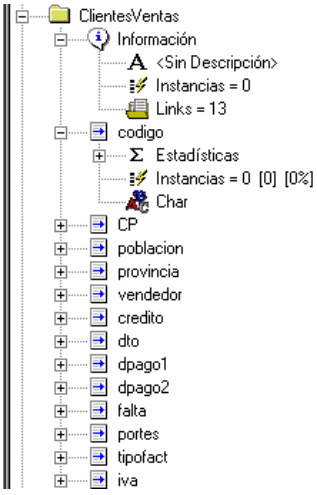
Uso de SQL en el Synera Explorer

Debido a que me interesaba mezclar datos entre los diferentes ítems que poseo en la base de datos y probar en profundidad todas las opciones del programa, inicialmente mi idea era realizar nuevos ítems, que fueran una selección (SELECT) de atributos de mi interés con unas condiciones predeterminadas (WHERE).
Parecía fácil o eso me suponía yo, pero aunque el manual de Synera explica este apartado, no existe ningún apartado de gramática o donde especifique las aceptaciones del SQL del Synera, por lo cual he tenido que ir probando, (a ciegas literalmente).
Por fin consigo realizar una selección aunque no totalmente como yo quería (es imposible poner los ORDER BY o GROUP BY)
Y la exporto pero como no puedo hacerlo en un ítem, realizo una hoja Excel, que intento volver a importar en el Synera como ítem,

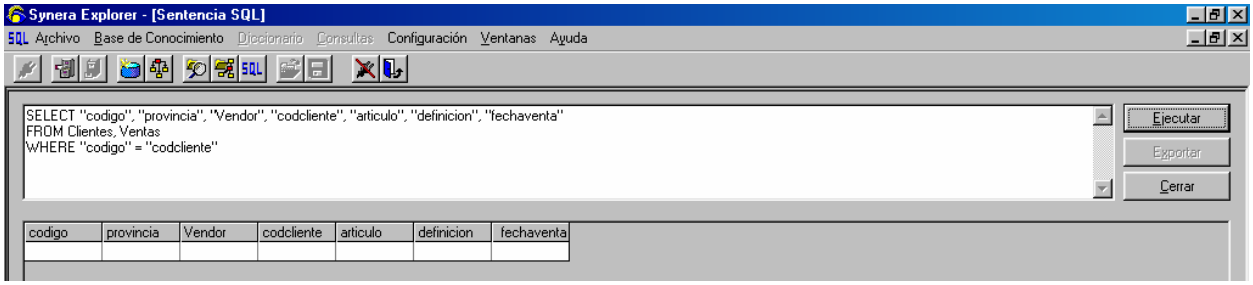


pero me da constantemente errores y no me carga ningún dato, realizo una limpieza y completo algunas filas vacías, que me hacen percibir que los datos que poseo facilitados por el cliente están todavía peor que en un inicio percibí.
Aún así tengo 3629 filas que importa, pero como da errores, el ítem que me realiza queda vacío (Instancias 0) como se ve.





Otras de las veces, que no da errores, en cambio se queda pensando más de 5 minutos y no aparece ningún dato.



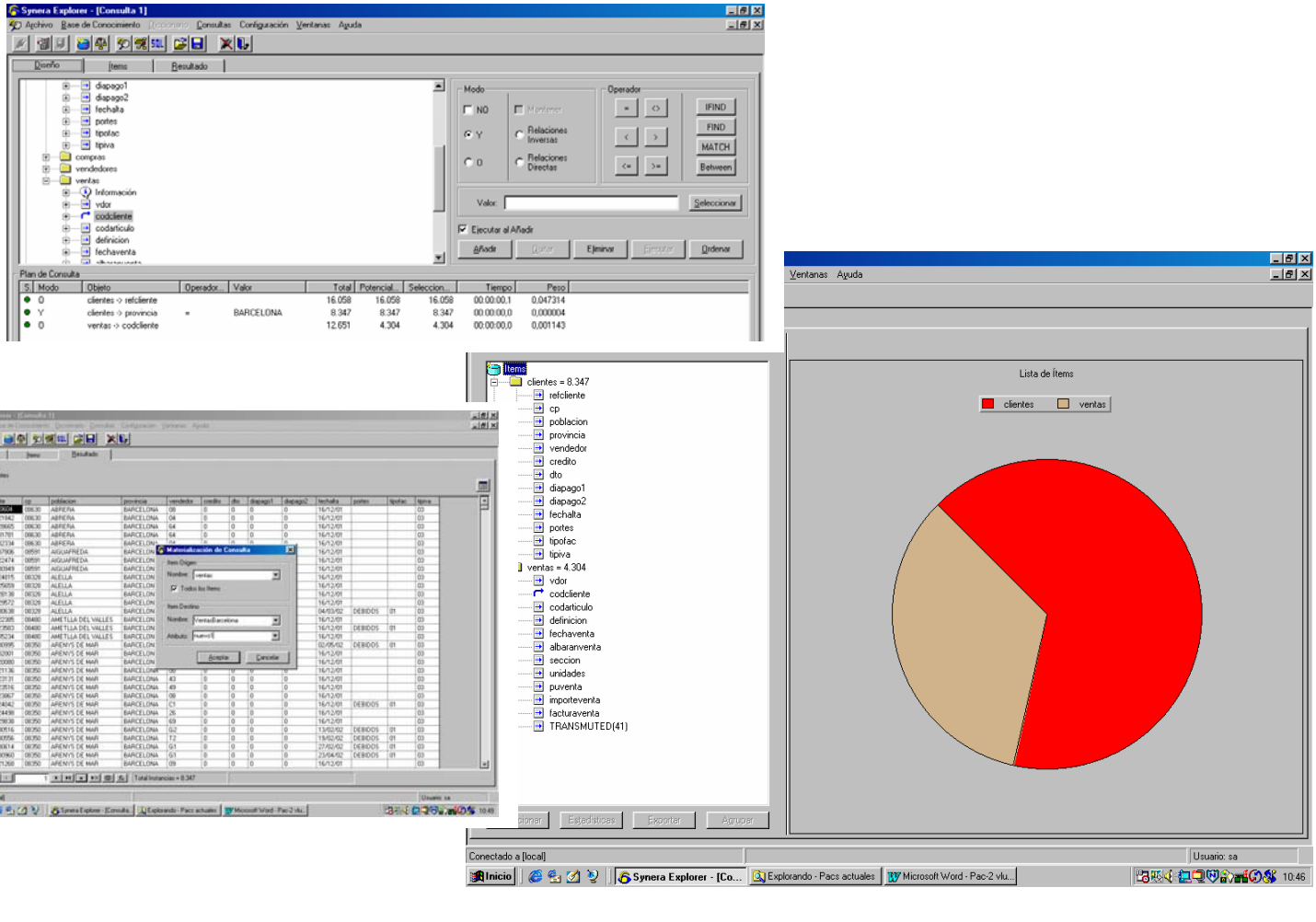
Uso de Consultas

He realizado numerosas consultas sin conseguir que contuvieran alguna instancia, aun así he conseguido bastantes con resultados.
O también al intentar hacer un join, unión, intersección con varias consultas como no puedo cerrarlas ya que si lo realizo se eliminan, debo dejarlas abiertas y abrir de nuevas, primero que es confuso cuando ya llevas varias, y además ha provocado a veces que el Synera se me quedara colgado.

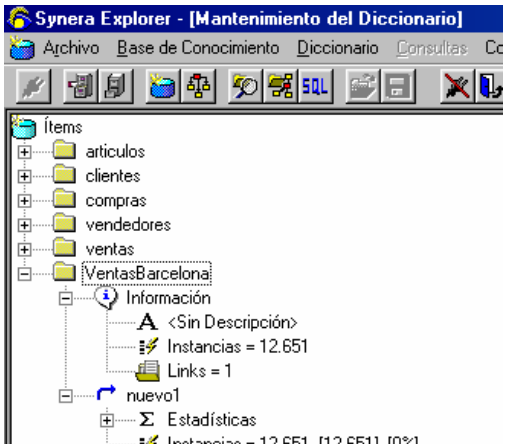
A continuación detallo algunas de las consultas realizadas:

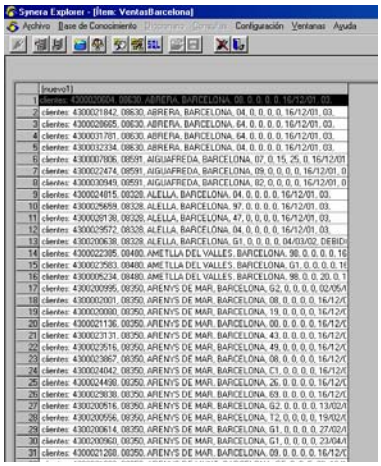
Las búsquedas sobretodo que me interesan para el trabajo es ver si existe una relación de las ventas por provincia, así como el porcentaje, eso también respecto a las fechas, es decir las ventas son homogéneas en los meses o existe una variación. También enlazar el tema de vendedores.

- Consulta de Clientes = BARCELONA y que además tengan ventas

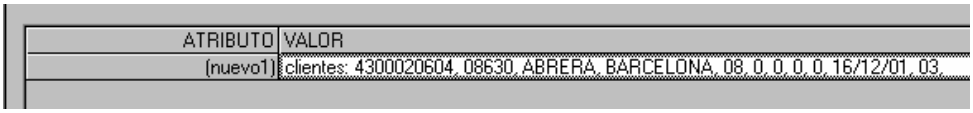


Como se ve intento materializar esta consulta en un nuevo ítem, pero no puedo mezclar datos de cliente y ventas que a mí me gustaría ya que quería saber si existe alguna regla o relación. Así que extraigo los datos de ventas y le llamo al nuevo ítem “ClientesBarcelona” ya que quiero realizar la misma consulta para las diferentes provincias más importantes o de interés para el trabajo.

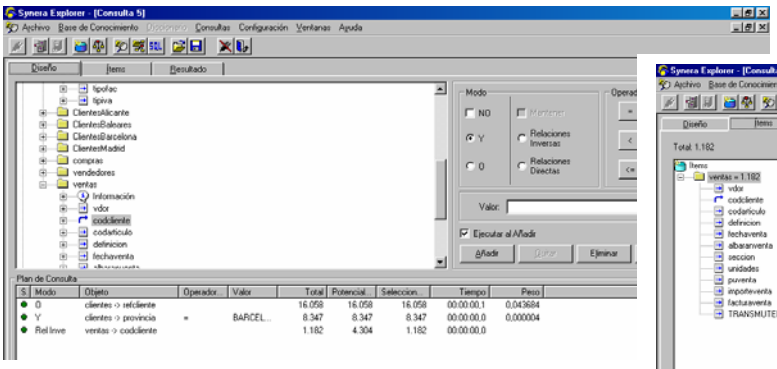


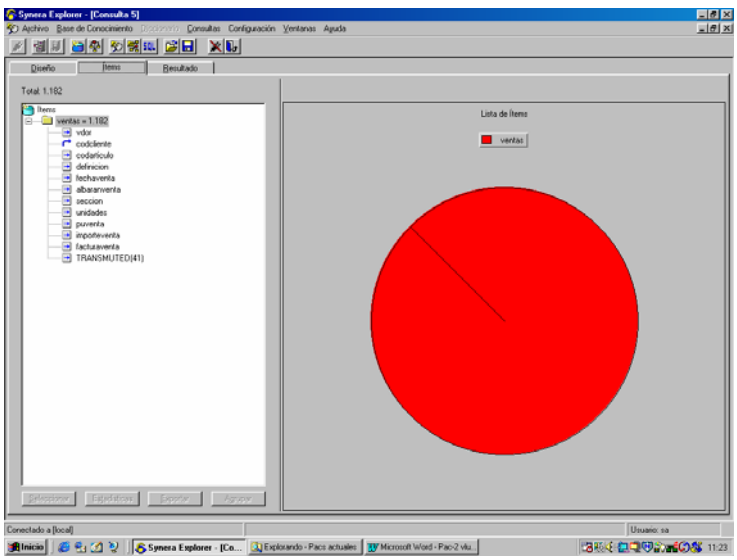


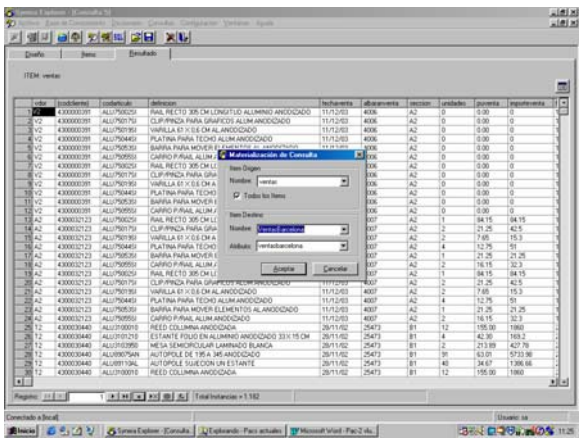
Como se ve realiza la materialización esta vez, y crea un único atributo que contiene toda la información pero de clientes, ahora realizare lo mismo pero realizando la materialización al revés, y me da lo mismo, así que quizás lo bueno sea luego realizar una intersección entre consultas para completar los datos que faltan de ventas que creo que son los que pueden aportar reglas de comportamientos.



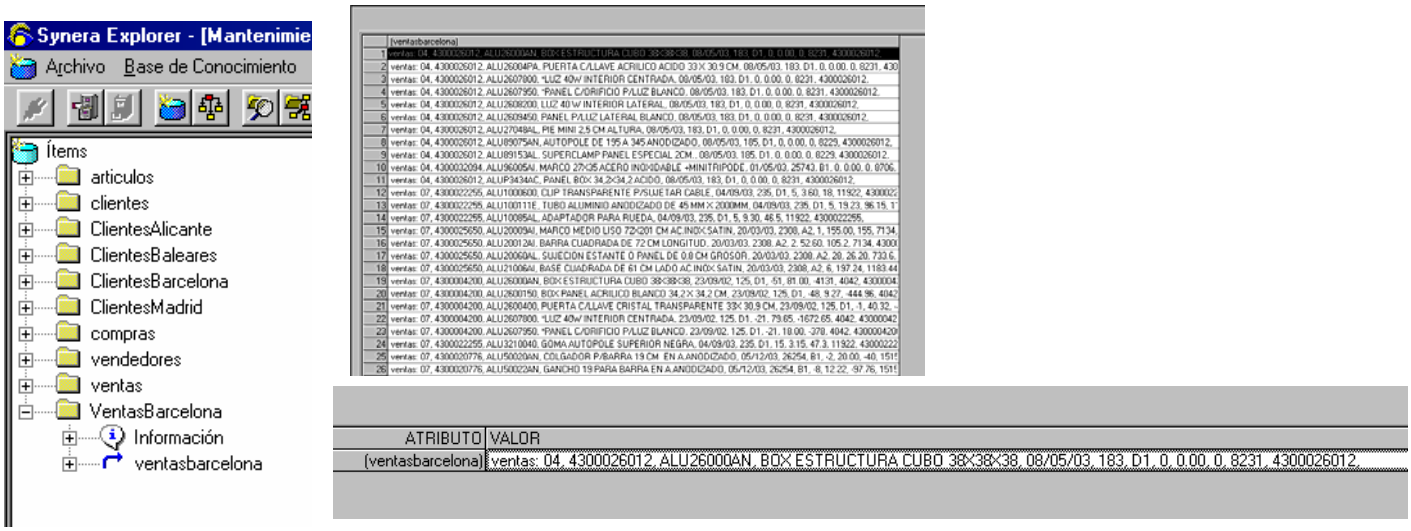
Para este tipo de consultas he probado con los operadores Y, y relación directa y no me produce ninguna instancia. En cambio realizo la consulta con Relación inversa y aquí veo el resultado de ventas de clientes de solo de Barcelona.



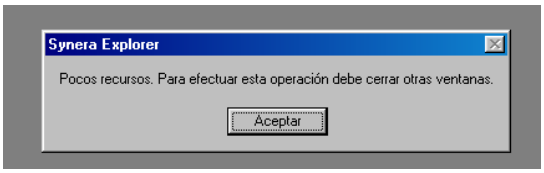


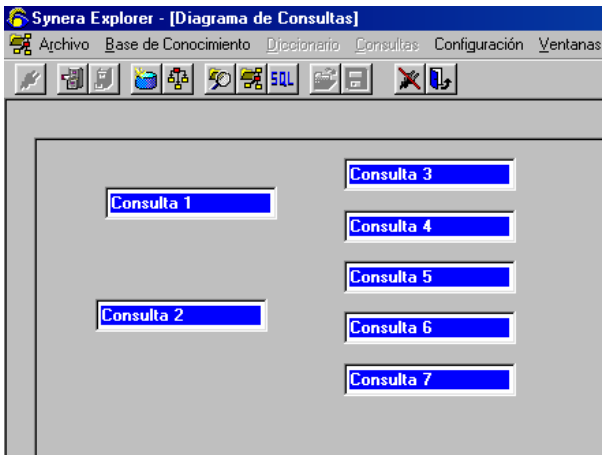


Por tanto estas consultas las materializo en ítems “VentasBarcelona”

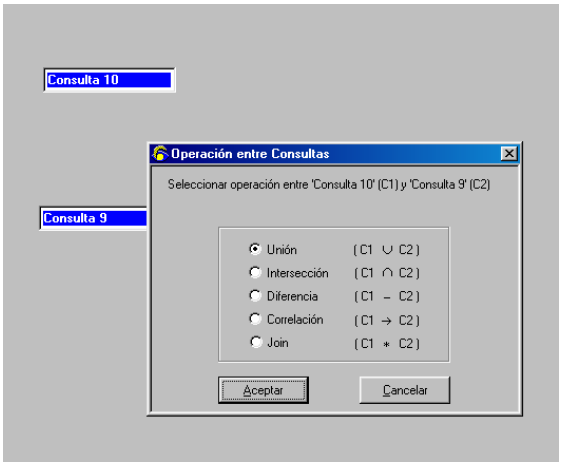


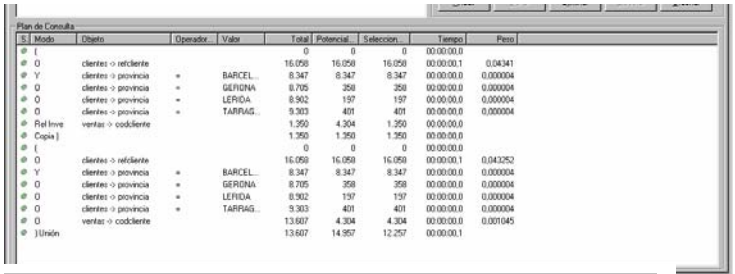
Debido a las numerosas consultas que muchas veces he realizado, me ha pasado que el Synera se queda sin recursos, no deja realizar muchas más de 7 consultas.

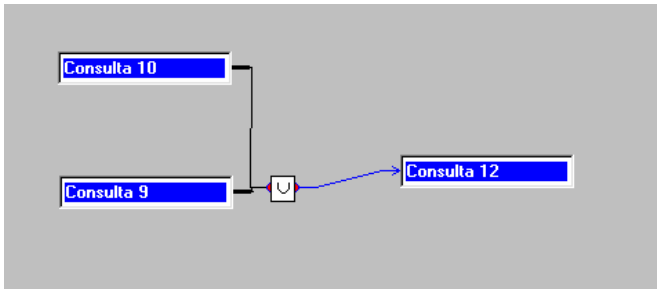




También realizo por ejemplo, operaciones entre consultas que realizo, aunque en algunos casos debo probar que operación será la mejor

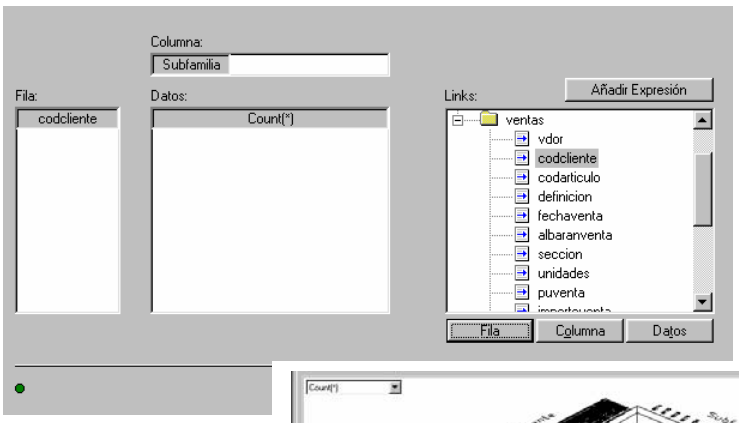


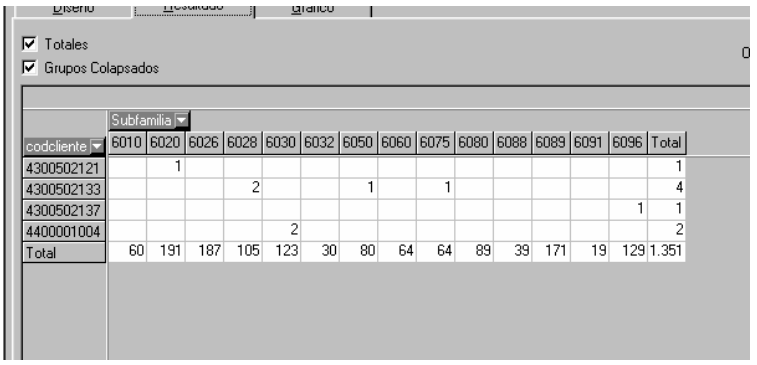


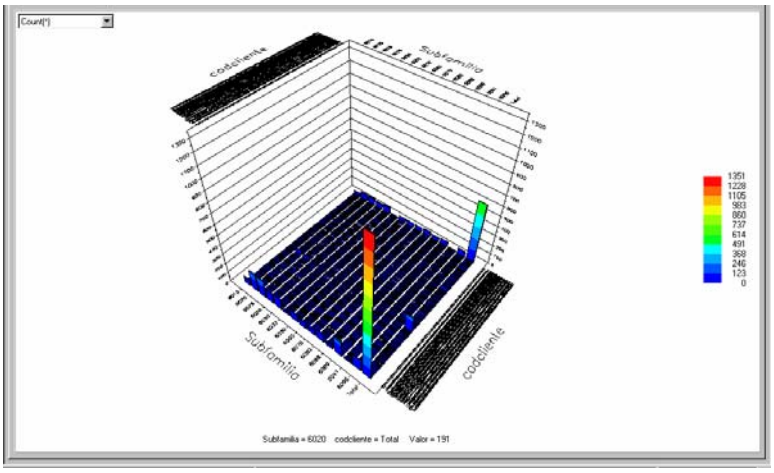


Uso de los cubos de datos.

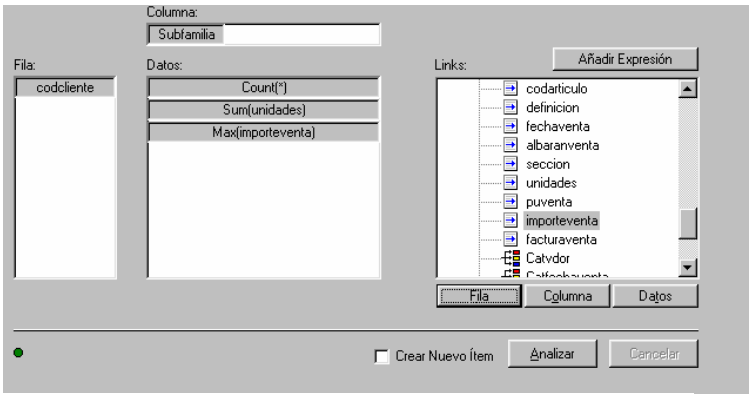
Inicialmente realizo la relación o el análisis de artículos (subfamilia) con ventas, para ver las subfamilias de productos que han tenido más ventas.

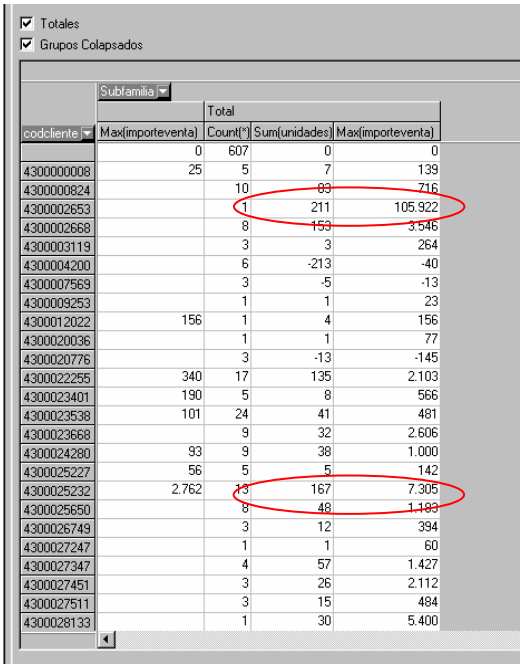


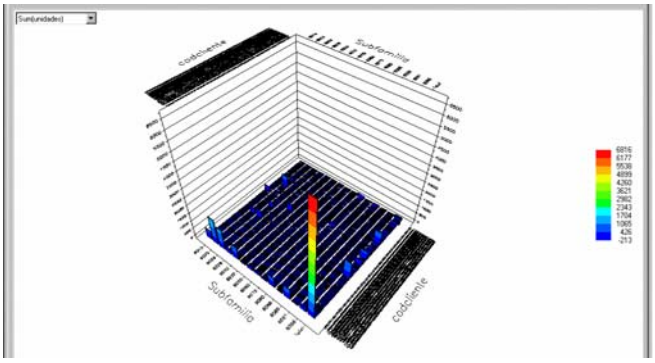


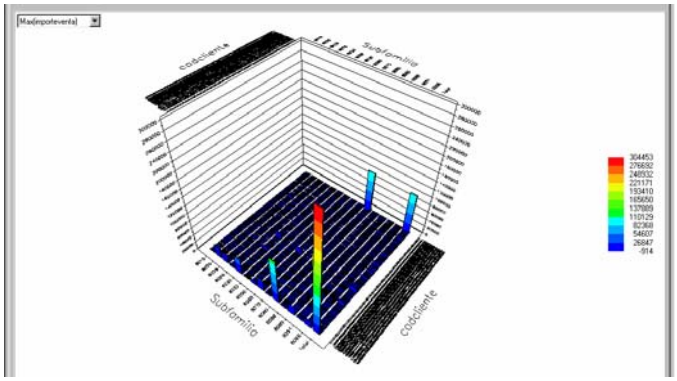


Vemos claramente que las subfamilias por orden de mayor a menor han sido la 6020, 6026 y 6089. Ahora incorporare más complejidad al cubo extraído de datos poniendo la suma de unidades vendidas y el máximo importe, para detectar tanto el hecho de ver que subfamilia proporciona pedidos mas grandes, así como ver si estos se corresponden a unas ventas de grandes cantidades de material o a pedidos de material de gran importe.









Por una parte vemos que los importes máximos
Y las unidades no son significativo, eso quiere dec
Que existen materiales de precio elevado.
Y el máximo importe se encuentra localizado en
La subfamilia 6080. Por tanto seguro que nos
Encontramos ante un pedido especial y puntual
De un material con elevado precios.

Aparte podríamos añadir más columnas y filas , realizando una combinación de 4 o más atributos como por ejemplo : añadiendo familia y sección

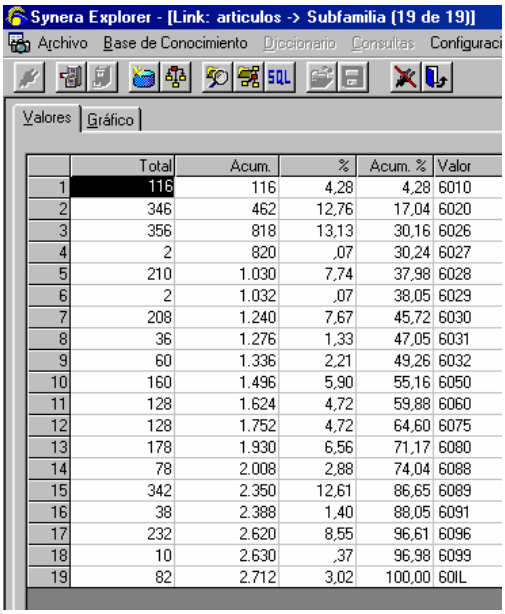
Totales
Grupos Colapsados
Orden: Materializar Opciones Exportar

		Subfamilia		Familia																	
		6010		6020		6026			6028		6030	6032	6050	6060	6075	6080	6088	6089	6091	6096	Total
codcliente	seccion	60	Total	60	Total	60	61	Total	60	Total											
		31	31	63	63	64	1	65	55	55	67	11	38	40	43	49	38	52	4	51	607
	Total	31	31	63	63	64	1	65	55	55	67	11	38	40	43	49	38	52	4	51	607
4300000008	D2																	1		1	2
	D3															1		2			3
	Total															1		3		1	5
4300000824	E1	3	3															7			10
	Total	3	3															7			10
4300002653	E1															1					1
	Total															1					1
4300002668	D1																	3			3
	F1												1					4			5
	Total												1					7			8
4300003119	A2										3										3
	Total										3										3
4300004200	D1					6		6													6

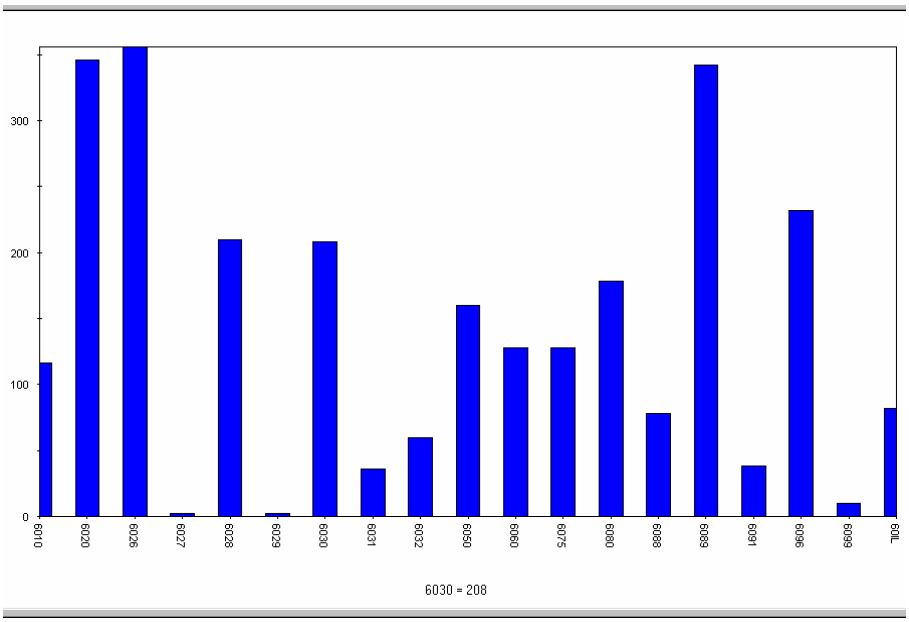
Análisis de ítems a través del Synera Explorer

Item : Artículos

Aquí tenemos los valores del atributo de subfamilia, que nos indica, que subfamilia dentro de los productos tiene más peso.

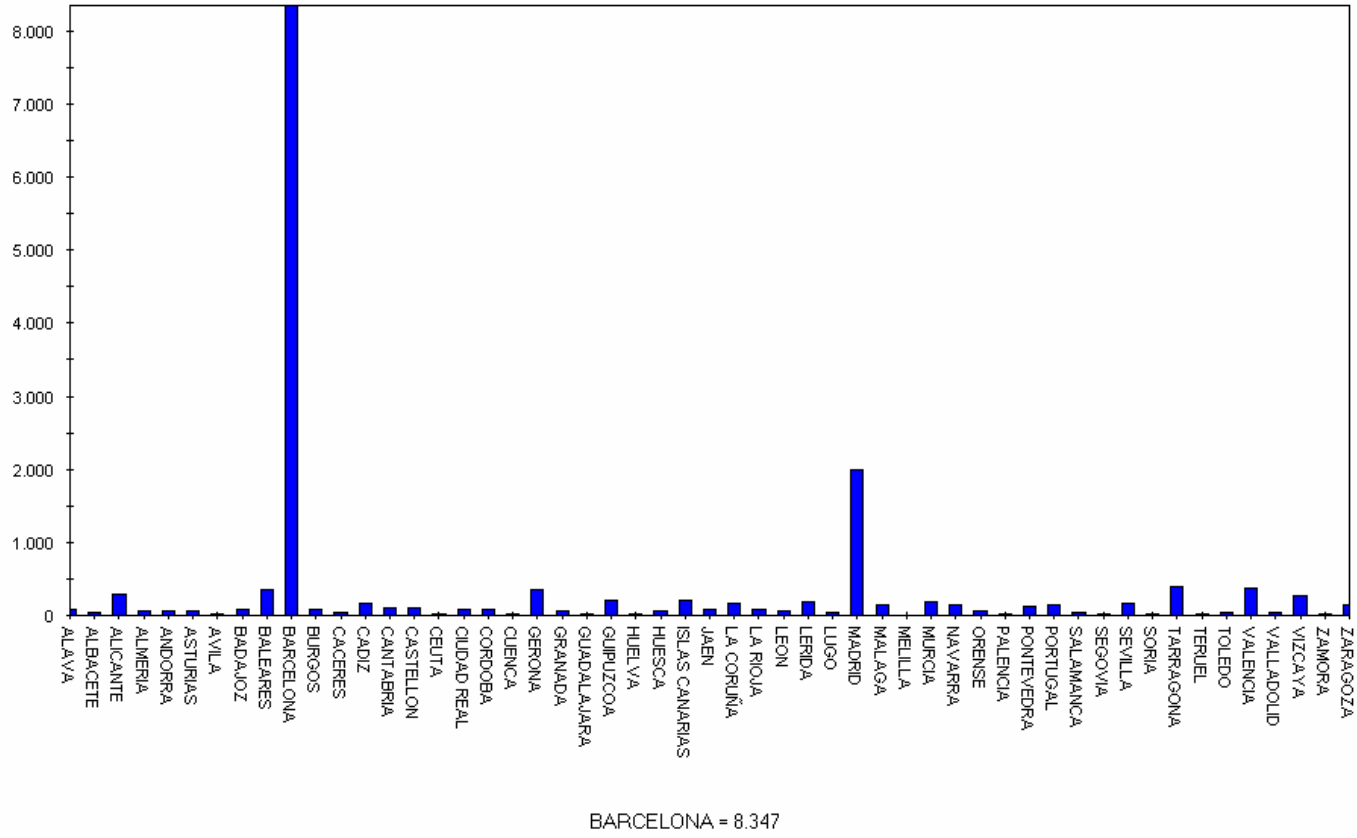


	Total	Acum.	%	Acum. %	Valor
1	116	116	4,28	4,28	6010
2	346	462	12,76	17,04	6020
3	356	818	13,13	30,16	6026
4	2	820	,07	30,24	6027
5	210	1.030	7,74	37,98	6028
6	2	1.032	,07	38,05	6029
7	208	1.240	7,67	45,72	6030
8	36	1.276	1,33	47,05	6031
9	60	1.336	2,21	49,26	6032
10	160	1.496	5,90	55,16	6050
11	128	1.624	4,72	59,88	6060
12	128	1.752	4,72	64,60	6075
13	178	1.930	6,56	71,17	6080
14	78	2.008	2,88	74,04	6088
15	342	2.350	12,61	86,65	6089
16	38	2.388	1,40	88,05	6091
17	232	2.620	8,55	96,61	6096
18	10	2.630	,37	96,98	6099
19	82	2.712	3,02	100,00	60IL



Vemos claramente que las subfamilias, 6026, 6020 y 6089 son las que poseemos más productos. Podemos realizar un informe, también he imprimirlo.

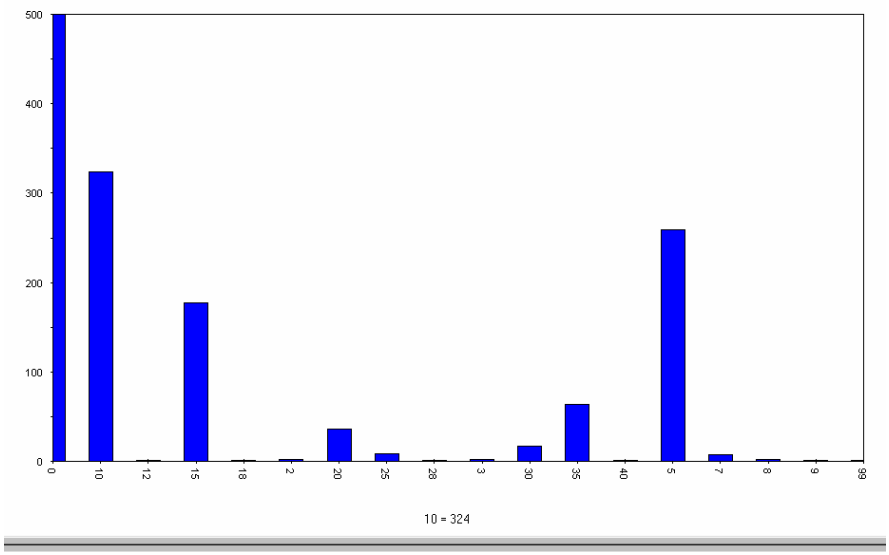
En el **ítem Clientes**, en el atributo provincia, vemos donde se tienen concentrados los mismos, por tanto nos indica claramente donde tendríamos que actuar más a nivel de campañas de publicidad o marketing para incrementar en aquellas provincias con menor presencia



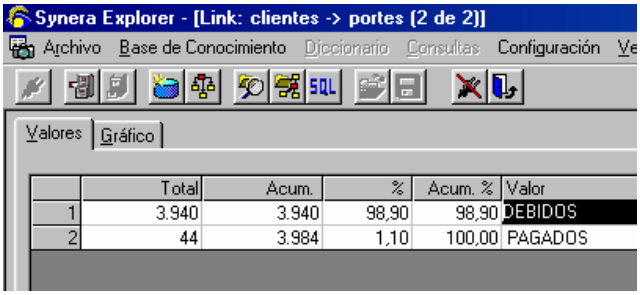
Aunque el gráfico nos aportara una información a simple vista más clara, ya vemos claramente que la empresa tiene concentrados los clientes en Cataluña (Barcelona la provincia con mayor peso), Madrid, y luego le sigue Valencia

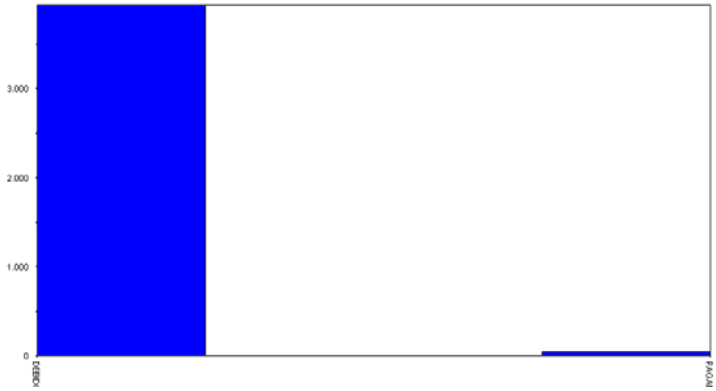
Por el atributo de Descuento también del **ítem de Clientes,** vemos que la política de la empresa mayoritariamente es no dar descuentos, como se ve existe claramente una mayoría que no lo tienen, el resto o se da un 10% o un 5%. (en el gráfico en máximo ha sido modificado para poder apreciar mejor los otros valores que no eran 0)

	Total	Acum.	%	Acum. %	Valor
1	15.153	15.153	94,37	94,37	0
2	324	15.477	2,02	96,39	10
3	1	15.478	,01	96,39	12
4	177	15.655	1,10	97,50	15
5	1	15.656	,01	97,50	18
6	2	15.658	,01	97,52	2
7	36	15.694	,22	97,74	20
8	8	15.702	,05	97,79	25
9	1	15.703	,01	97,80	28
10	2	15.705	,01	97,81	3
11	17	15.722	,11	97,91	30
12	64	15.786	,40	98,31	35
13	1	15.787	,01	98,32	40
14	259	16.046	1,61	99,93	5
15	7	16.053	,04	99,98	7
16	2	16.055	,01	99,99	8
17	1	16.056	,01	99,99	9
18	1	16.057	,01	100,00	99

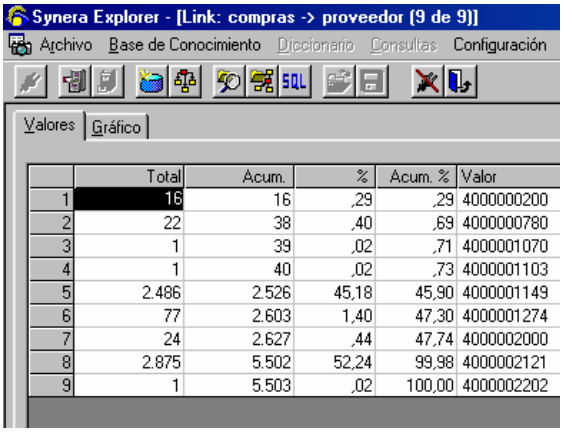


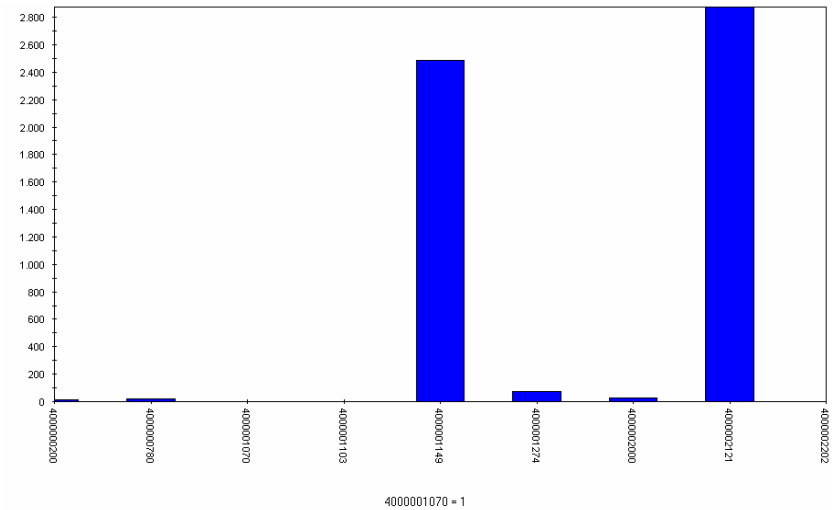
También con el atributo Portes, vemos que la empresa mayoritariamente, es bastante apiastante cobra los portes es decir son debidos ya que solo un 1,10 % los tiene pagados, aunque considero que este dato seguro que a la practica no es real del todo.



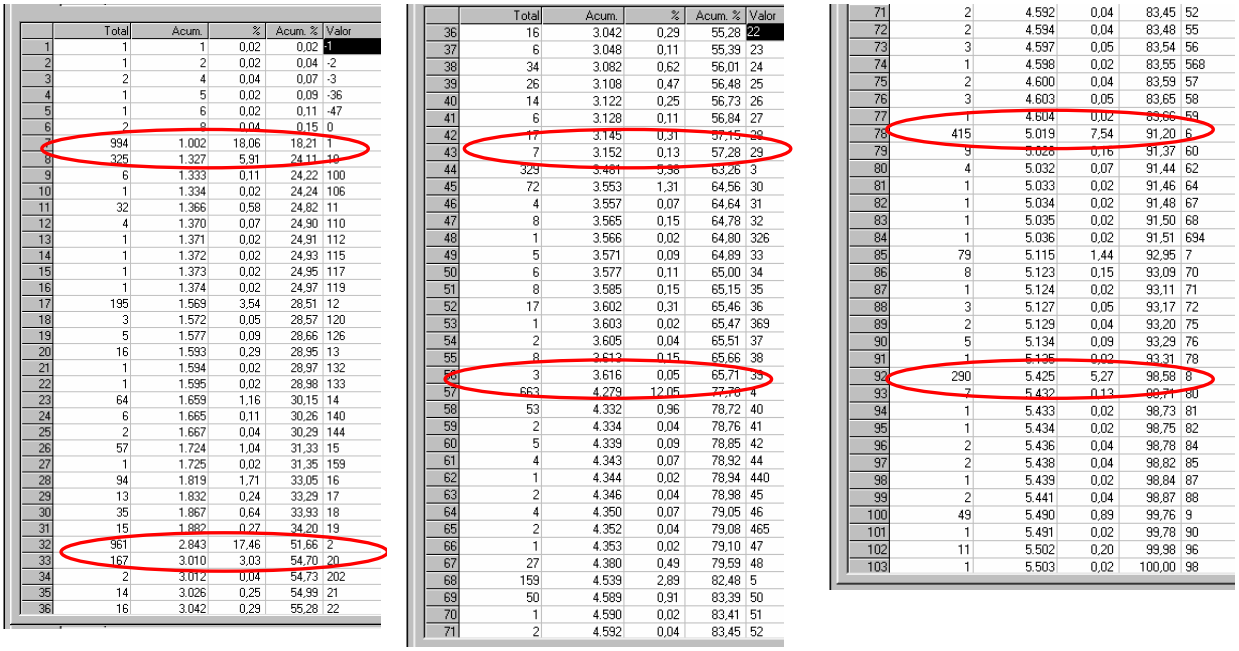


En el **ítem de Compras** vemos con el atributo proveedor que para esta gama de artículos o familia, la concentración de las compras esta en dos proveedores. El 1149 y el 2121.





En el **ítem de Compras** vemos con el atributo unidades que normalmente las cantidades que se piden de cada artículo son 1, 2, 4,6,8,3 (en orden de mayor a menor)
El gráfico debido a que es muy extenso no lo copio.

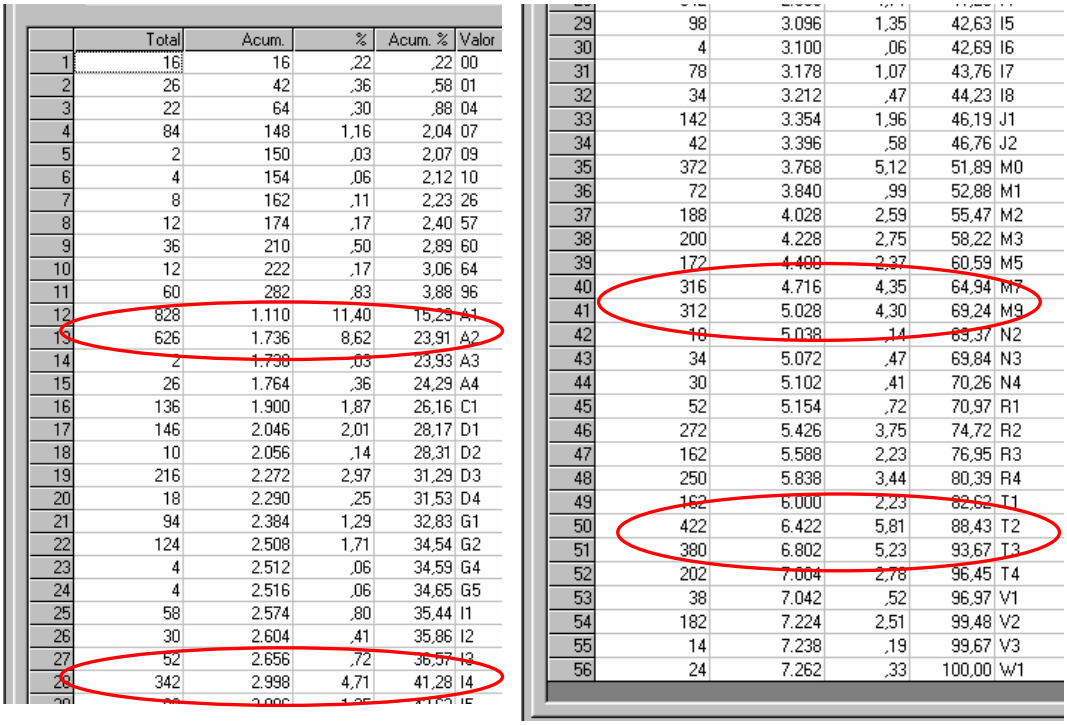


	Total	Acum.	%	Acum. %	Valor
1	1	1	0,02	0,02	1
2	1	2	0,02	0,04	-2
3	2	4	0,04	0,07	-3
4	1	5	0,02	0,09	-36
5	1	6	0,02	0,11	-47
6	2	8	0,04	0,15	0
7	994	1.002	18,06	18,21	1
8	325	1.327	5,91	24,11	19
9	6	1.333	0,11	24,22	100
10	1	1.334	0,02	24,24	106
11	32	1.366	0,58	24,82	11
12	4	1.370	0,07	24,90	110
13	1	1.371	0,02	24,91	112
14	1	1.372	0,02	24,93	115
15	1	1.373	0,02	24,95	117
16	1	1.374	0,02	24,97	119
17	195	1.569	3,54	28,51	12
18	3	1.572	0,05	28,57	120
19	5	1.577	0,09	28,66	126
20	16	1.593	0,29	28,95	13
21	1	1.594	0,02	28,97	132
22	1	1.595	0,02	28,98	133
23	64	1.659	1,16	30,15	14
24	6	1.665	0,11	30,26	140
25	2	1.667	0,04	30,29	144
26	57	1.724	1,04	31,33	15
27	1	1.725	0,02	31,35	159
28	94	1.819	1,71	33,05	16
29	13	1.832	0,24	33,29	17
30	35	1.867	0,64	33,93	18
31	15	1.882	0,27	34,20	19
32	961	2.843	17,46	51,66	2
33	167	3.010	3,03	54,70	20
34	2	3.012	0,04	54,73	202
35	14	3.026	0,25	54,99	21
36	16	3.042	0,29	55,28	22

	Total	Acum.	%	Acum. %	Valor
36	16	3.042	0,29	55,28	22
37	6	3.048	0,11	55,39	23
38	34	3.082	0,62	56,01	24
39	26	3.108	0,47	56,48	25
40	14	3.122	0,25	56,73	26
41	6	3.128	0,11	56,84	27
42	17	3.145	0,31	57,15	28
43	7	3.152	0,13	57,28	29
44	329	3.481	5,98	63,26	3
45	72	3.553	1,31	64,56	30
46	4	3.557	0,07	64,64	31
47	8	3.565	0,15	64,78	32
48	1	3.566	0,02	64,80	326
49	5	3.571	0,09	64,89	33
50	6	3.577	0,11	65,00	34
51	8	3.585	0,15	65,15	35
52	17	3.602	0,31	65,46	36
53	1	3.603	0,02	65,47	369
54	2	3.605	0,04	65,51	37
55	8	3.613	0,15	65,66	38
56	3	3.616	0,05	65,71	39
57	663	4.279	12,05	77,76	4
58	53	4.332	0,96	78,72	40
59	2	4.334	0,04	78,76	41
60	5	4.339	0,09	78,85	42
61	4	4.343	0,07	78,92	44
62	1	4.344	0,02	78,94	440
63	2	4.346	0,04	78,98	45
64	4	4.350	0,07	79,05	46
65	2	4.352	0,04	79,08	465
66	1	4.353	0,02	79,10	47
67	27	4.380	0,49	79,59	48
68	159	4.539	2,89	82,48	5
69	50	4.589	0,91	83,39	50
70	1	4.590	0,02	83,41	51
71	2	4.592	0,04	83,45	52

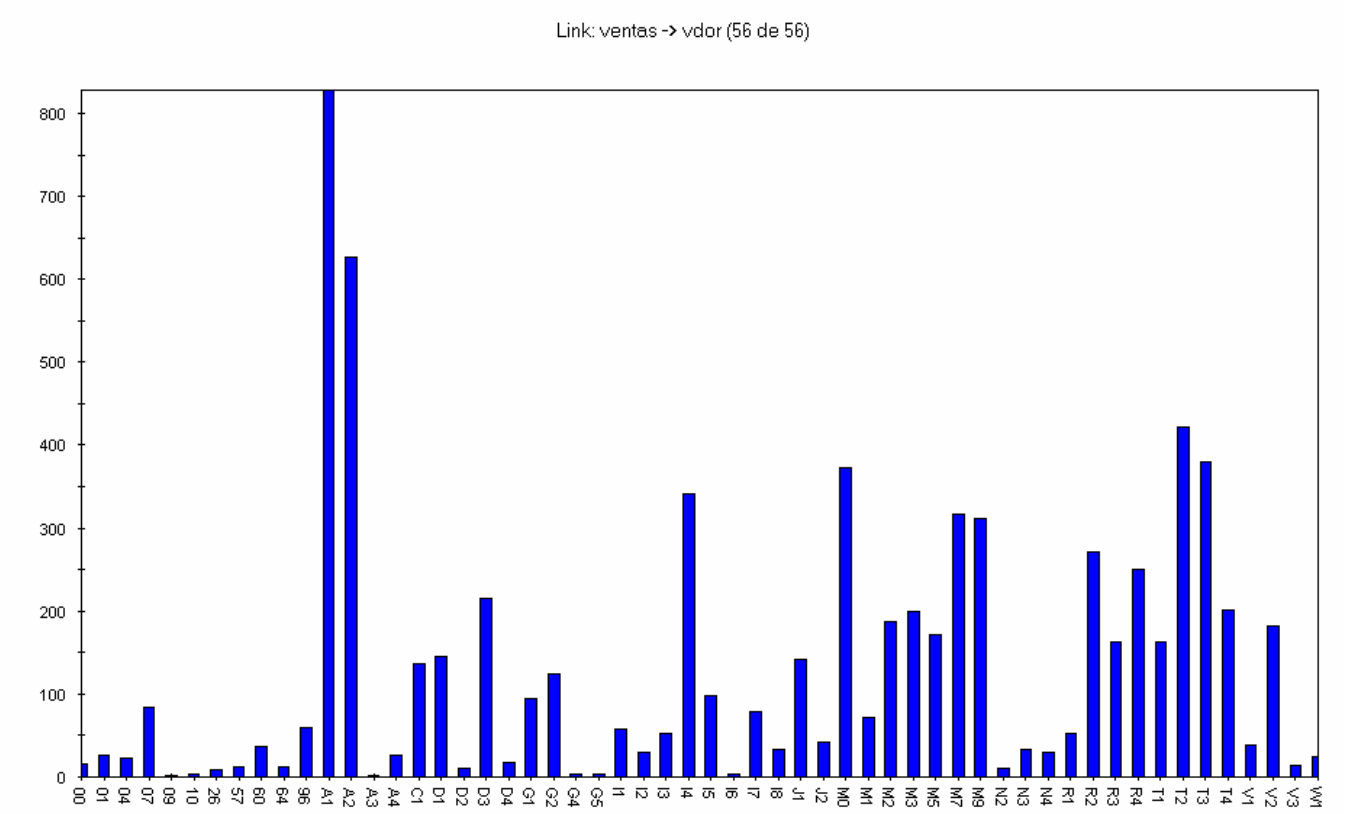
71	2	4.592	0,04	83,45	52
72	2	4.594	0,04	83,48	55
73	3	4.597	0,05	83,54	56
74	1	4.598	0,02	83,55	568
75	2	4.600	0,04	83,59	57
76	3	4.603	0,05	83,65	58
77	1	4.604	0,02	83,66	59
78	415	5.019	7,54	91,20	6
79	9	5.028	0,16	91,37	60
80	4	5.032	0,07	91,44	62
81	1	5.033	0,02	91,46	64
82	1	5.034	0,02	91,48	67
83	1	5.035	0,02	91,50	68
84	1	5.036	0,02	91,51	694
85	79	5.115	1,44	92,95	7
86	8	5.123	0,15	93,09	70
87	1	5.124	0,02	93,11	71
88	3	5.127	0,05	93,17	72
89	2	5.129	0,04	93,20	75
90	5	5.134	0,09	93,29	76
91	1	5.135	0,02	93,31	78
92	290	5.425	5,27	98,58	8
93	7	5.432	0,13	98,71	80
94	1	5.433	0,02	98,73	81
95	1	5.434	0,02	98,75	82
96	2	5.436	0,04	98,78	84
97	2	5.438	0,04	98,82	85
98	1	5.439	0,02	98,84	87
99	2	5.441	0,04	98,87	88
100	49	5.490	0,89	99,76	9
101	1	5.491	0,02	99,78	90
102	11	5.502	0,20	99,98	96
103	1	5.503	0,02	100,00	98

En el **ítem de Ventas** vemos con el atributo vendedores los que realmente han conseguido realizar mayores ventas, A1, A2, T2, T3, I4, M7, M9,



	Total	Acum.	%	Acum. %	Valor
1	16	16	,22	,22	00
2	26	42	,36	,58	01
3	22	64	,30	,88	04
4	84	148	1,16	2,04	07
5	2	150	,03	2,07	09
6	4	154	,06	2,12	10
7	8	162	,11	2,23	26
8	12	174	,17	2,40	57
9	36	210	,50	2,89	60
10	12	222	,17	3,06	64
11	60	282	,83	3,88	96
12	828	1.110	11,40	15,29	A1
13	626	1.736	8,62	23,91	A2
14	2	1.738	,03	23,93	A3
15	26	1.764	,36	24,29	A4
16	136	1.900	1,87	26,16	C1
17	146	2.046	2,01	28,17	D1
18	10	2.056	,14	28,31	D2
19	216	2.272	2,97	31,29	D3
20	18	2.290	,25	31,53	D4
21	94	2.384	1,29	32,83	G1
22	124	2.508	1,71	34,54	G2
23	4	2.512	,06	34,59	G4
24	4	2.516	,06	34,65	G5
25	58	2.574	,80	35,44	I1
26	30	2.604	,41	35,86	I2
27	52	2.656	,72	36,57	I3
28	342	2.998	4,71	41,28	I4

29	98	3.096	1,35	42,63	I5
30	4	3.100	,06	42,69	I6
31	78	3.178	1,07	43,76	I7
32	34	3.212	,47	44,23	I8
33	142	3.354	1,96	46,19	J1
34	42	3.396	,58	46,76	J2
35	372	3.768	5,12	51,89	M0
36	72	3.840	,99	52,88	M1
37	188	4.028	2,59	55,47	M2
38	200	4.228	2,75	58,22	M3
39	172	4.400	2,37	60,59	M5
40	316	4.716	4,35	64,94	M7
41	312	5.028	4,30	69,24	M9
42	10	5.038	,14	69,37	N2
43	34	5.072	,47	69,84	N3
44	30	5.102	,41	70,26	N4
45	52	5.154	,72	70,97	R1
46	272	5.426	3,75	74,72	R2
47	162	5.588	2,23	76,95	R3
48	250	5.838	3,44	80,39	R4
49	162	6.000	2,23	82,62	T1
50	422	6.422	5,81	88,43	T2
51	380	6.802	5,23	93,67	T3
52	202	7.004	2,78	96,45	T4
53	38	7.042	,52	96,97	V1
54	182	7.224	2,51	99,48	V2
55	14	7.238	,19	99,67	V3
56	24	7.262	,33	100,00	W1



En el **ítem de Ventas** vemos con el atributo clientes que las ventas están concentradas realmente en 264 clientes, esto realmente supone 1,64 % de los clientes totales de la compañía, es decir que evidentemente casi seguro que podríamos dar a conocer mejor esta familia a los clientes de la empresa en general ya que seguro que muchos no lo conocen. Otro dato es que de 4304 líneas de ventas solo existan esos 264 clientes diferentes, entonces tenemos que existe muchos clientes repetitivos o pedidos con gran cantidad de productos diferentes. Ya que como podemos observar hay clientes que han comprado 936,148, 132, 70,68,56, 54 productos diferentes.

Valores | Gráfico

	Total	Acum.	%	Acum. %	Valor
1	54	54	1,25	1,25	430002366
2	10	64	0,23	1,49	430020235
3	148	212	3,44	4,93	430002365
4	2	214	0,05	4,97	430003208
5	10	224	0,23	5,20	430050142
6	132	356	3,07	8,27	430002523
7	6	362	0,14	8,41	430050147
8	26	388	0,60	9,01	430003073
9	16	404	0,37	9,39	430003347
10	8	412	0,19	9,57	430050008
11	8	420	0,19	9,76	430017024
12	6	426	0,14	9,90	430003066
13	4	430	0,09	9,99	430002861
14	68	498	1,58	11,57	430000266
15	4	502	0,09	11,66	430002389
16	2	504	0,05	11,71	430020081
17	4	508	0,09	11,80	430002545
18	8	516	0,19	11,99	430002651
19	14	530	0,33	12,31	430003344
20	10	540	0,23	12,55	430003177
21	66	606	1,53	14,08	430002428
22	12	618	0,28	14,36	430000035
23	12	630	0,28	14,64	430003212
24	10	640	0,23	14,87	430003044
25	36	676	0,84	15,71	430000311
26	20	696	0,46	16,17	430002601
27	4	700	0,09	16,26	430002921
28	4	704	0,09	16,36	430003218
29	4	708	0,09	16,45	430020225
30	6	714	0,14	16,59	430003238
31	2	716	0,05	16,64	430003230
32	2	718	0,05	16,68	430002983
33	10	728	0,23	16,91	430003344
34	6	734	0,14	17,05	430020075
35	8	742	0,19	17,24	430003308
36	2	744	0,05	17,29	430020084

	Total	Acum.	%	Acum. %	Valor
37	14	758	0,33	17,61	4300026642
38	4	762	0,09	17,70	4300031092
39	2	764	0,05	17,75	4300032064
40	22	786	0,51	18,26	4300027511
41	8	794	0,19	18,45	4300031269
42	4	798	0,09	18,54	4300200341
43	2	800	0,05	18,59	4300000004
44	4	804	0,09	18,68	4300029396
45	10	814	0,23	18,91	4300030692
46	4	818	0,09	19,01	4300030755
47	4	822	0,09	19,10	4300200022
48	42	864	0,98	20,07	4300033437
49	4	868	0,09	20,17	4300202057
50	6	874	0,14	20,31	4300202265
51	14	888	0,33	20,63	4300031554
52	2	890	0,05	20,68	4300028904
53	8	898	0,19	20,86	4300030313
54	36	934	0,84	21,70	4300030756
55	6	940	0,14	21,84	4300030757
56	44	984	1,02	22,86	4300031291
57	2	986	0,05	22,91	4300033442
58	2	988	0,05	22,96	4300202081
59	4	992	0,09	23,05	4300033476
60	4	996	0,09	23,14	4300033421
61	28	1.024	0,65	23,79	4300004200
62	2	1.026	0,05	23,84	4300023704
63	6	1.032	0,14	23,98	4300200972
64	2	1.034	0,05	24,02	4300028326
65	16	1.050	0,37	24,40	4300031268
66	2	1.052	0,05	24,44	4300170027
67	56	1.108	1,30	25,74	4300031890
68	4	1.112	0,09	25,84	4300033486
69	16	1.128	0,37	26,21	4300007569
70	2	1.130	0,05	26,25	4300028133
71	2	1.132	0,05	26,30	4300033414
72	14	1.146	0,33	26,63	4300200002

Conectado a [local]

	Total	Acum.	%	Acum. %	Valor
73	8	1.154	0,19	26,81	4300025309
74	2	1.156	0,05	26,86	4300027776
75	4	1.160	0,09	26,95	4300200841
76	2	1.162	0,05	27,00	4300202050
77	2	1.164	0,05	27,04	4300027917
78	70	1.234	1,63	28,67	4300022556
79	12	1.246	0,28	28,95	4300025651
80	4	1.250	0,09	29,04	4300012022
81	2	1.252	0,05	29,09	4300027073
82	2	1.254	0,05	29,14	4300030438
83	18	1.272	0,42	29,55	4300202066
84	2	1.274	0,05	29,60	4300202290
85	12	1.286	0,28	29,88	4300020776
86	10	1.296	0,23	30,11	4300200722
87	38	1.334	0,88	30,99	4300028389
88	16	1.350	0,37	31,37	4300032327
89	2	1.352	0,05	31,41	4300022671
90	2	1.354	0,05	31,46	4300022584
91	6	1.360	0,14	31,60	4300025277
92	4	1.364			
93	6	1.370			
94	2	1.372			
95	18	1.390			
96	6	1.396			
97	2	1.398			
98	8	1.406			
99	2	1.408			
100	4	1.412			
101	12	1.424			
102	6	1.430			
103	16	1.446			
104	12	1.458			
105	34	1.492			
106	10	1.502			
107	2	1.504			
108	8	1.512			

	Total	Acum.	%	Acum. %	Valor
229	8	2.944	0,19	68,40	4300501427
230	2	2.946	0,05	68,45	4300501372
231	22	2.968	0,51	68,96	4300031946
232	6	2.974	0,14	69,10	4300029867
233	18	2.992	0,42	69,52	4300202363
234	6	2.998	0,14	69,66	4300030338
235	2	3.000	0,05	69,70	4300030631
236	18	3.018	0,42	70,12	4300028857
237	16	3.034	0,37	70,49	4300031937
238	4	3.038	0,09	70,59	4300024646
239	2	3.040	0,05	70,63	4300033443
240	4	3.044	0,09	70,72	4300030246
241	10	3.054	0,23	70,96	4300029881
242	4	3.058	0,09	71,05	4300501112
243	2	3.060	0,05	71,10	4300029017
244	40	3.100	0,93	72,03	4300033478
245	6	3.106	0,14	72,17	4300031938
246	2	3.108	0,05	72,21	4300200607
247	6	3.114	0,14	72,35	4300200879
248	12	3.126	0,28	72,63	4300030451
249	2	3.128	0,05	72,68	4300200539
250	4	3.132	0,09	72,77	4300202031
251	2	3.134	0,05	72,82	4300023678
252	8	3.142	0,19	73,00	4300029441
253	22	3.164	0,51	73,51	4300027347
254	6	3.170	0,14	73,65	4300002423
255	4	3.174	0,09	73,75	4300031523
256	66	3.240	1,53	75,28	4300024333
257	22	3.262	0,51	75,79	4300027247
258	2	3.264	0,05	75,84	4300030481
259	6	3.270	0,14	75,98	4300030410
260	936	4.206	21,75	97,72	4300200204
261	8	4.214	0,19	97,91	4300033016
262	20	4.234	0,46	98,37	4300031917
263	60	4.294	1,39	99,77	4300022357
264	10	4.304	0,23	100,00	4300032313

Valores | Gráfico

	Total	Acum.	%	Acum. %	Valor
109	2	1.514	0,05	35,18	430002
110	54	1.568	1,25	36,43	430003
111	18	1.586	0,42	36,85	430003
112	34	1.620	0,79	37,64	430003
113	4	1.624	0,09	37,73	430003
114	4	1.628	0,09	37,83	430020
115	4	1.632	0,09	37,92	430003
116	54	1.686	1,25	39,17	430003
117	2	1.688	0,05	39,22	430002
118	2	1.690	0,05	39,27	430020
119	22	1.712	0,51	39,78	430017
120	2	1.714	0,05	39,82	430003
121	2	1.716	0,05	39,87	430003
122	8	1.724	0,19	40,06	430002
123	4	1.728	0,09	40,15	430003
124	16	1.744	0,37	40,52	430003
125	8	1.752	0,19	40,71	430002
126	10	1.762	0,23	40,94	430003
127	4	1.766	0,09	41,03	430020
128	22	1.788	0,51	41,54	430017
129	8	1.796	0,19	41,73	430050
130	10	1.806	0,23	41,96	430000
131	8	1.814	0,19	42,15	430003
132	10	1.824	0,23	42,38	430003
133	12	1.836	0,28	42,66	430002
134	16	1.852	0,37	43,03	430003
135	12	1.864	0,28	43,31	430002
136	12	1.876	0,28	43,59	430003
137	2	1.878	0,05	43,63	430003
138	12	1.890	0,28	43,91	430003
139	4	1.894	0,09	44,01	430002
140	2	1.896	0,05	44,05	430003
141	2	1.898	0,05	44,10	430003
142	2	1.900	0,05	44,14	430020
143	18	1.918	0,42	44,56	430000
144	2	1.920	0,05	44,61	430002

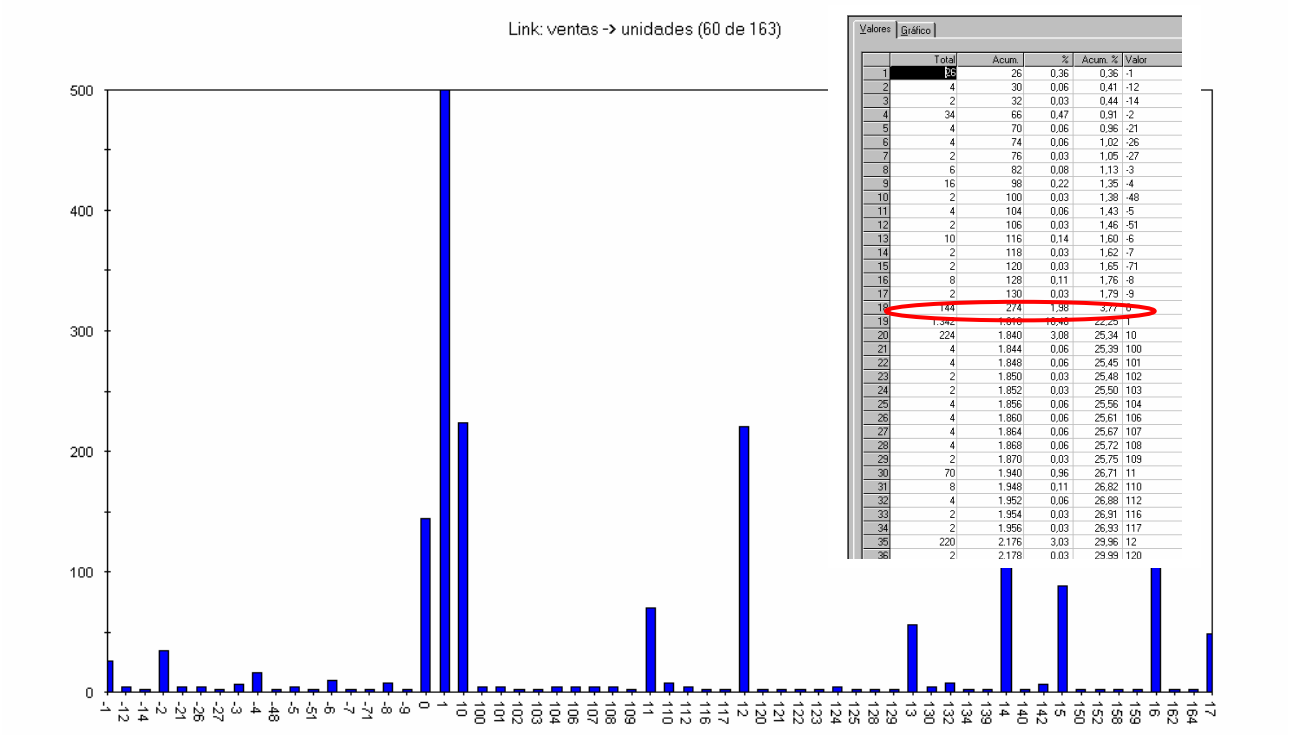
	Total	Acum.	%	Acum. %	Valor
145	2	1.922	0,05	44,66	4300021599
146	4	1.926	0,09	44,75	4300200098
147	2	1.928	0,05	44,80	4300027615
148	8	1.936	0,19	44,98	4300030765
149	4	1.940	0,09	45,07	4300032161
150	4	1.944	0,09	45,17	4300033048
151	18	1.954	0,23	45,48	4300030470
152	56	2.010	1,30	46,70	4300500842
153	13	2.023	0,30	46,96	4300032151
154	68	2.090	1,58	48,56	4300023401
155	6	2.096	0,14	48,70	4300029853
156	20	2.116	0,46	49,16	4300005899
157	2	2.118	0,05	49,21	4300032141
158	4	2.122	0,09	49,30	4300030247
159	8	2.130	0,19	49,49	4300500051
160	18	2.148	0,42	49,91	4300032004
161	4	2.152	0,09	50,00	4300029492
162	2	2.154	0,05	50,05	4300030697
163	4	2.158	0,09	50,14	4300031221
164	8	2.166	0,19	50,33	4300500098
165	14	2.180	0,33	50,65	4300020036
166	62	2.242	1,44	52,09	4300500133
167	2	2.244	0,05	52,14	4300029919
168	18	2.262	0,42	52,56	4300033400
169	2	2.264	0,05	52,60	4300501471
170	12	2.276	0,28	52,88	4300031087
171	4	2.280	0,09	52,97	4300033015
172	10	2.290	0,23	53,21	4300500242
173	2	2.292	0,05	53,25	4300033436
174	10	2.302	0,23	53,49	4300027567
175	4	2.306	0,09	53,58	4300030848
176	10	2.316	0,23	53,81	4300500035
177	4	2.320	0,09	53,90	4300500125
178	36	2.356	0,84	54,74	4300500175
179	10	2.366	0,23	54,97	4300501469
180	4	2.370	0,09	55,07	4300031608

Conectado a [local]

Valores | Gráfico

	Total	Acum.
181	4	2.374
182	14	2.388
183	64	2.452
184	2	2.454
185	8	2.462
186	8	2.470
187	4	2.474
188	12	2.486
189	4	2.490
190	2	2.492
191	16	2.508
192	2	2.510
193	4	2.514
194	24	2.538
195	4	2.542
196	4	2.546
197	4	2.550
198	8	2.558
199	28	2.586
200	8	2.594
201	8	2.602
202	12	2.614
203	24	2.638
204	2	2.640
205	2	2.642
206	6	2.648
207	8	2.656
208	4	2.660
209	28	2.688
210	4	2.692
211	4	2.696
212	6	2.702
213	72	2.774
214	6	2.780
215	12	2.792
216	28	2.820

En el **ítem de Ventas** vemos con el atributo unidades que las ventas están concentradas realmente en la venta unitaria de productos, aunque como se ve en el gráfico (he bajado el rango para que se aprecie mejor los pequeños)

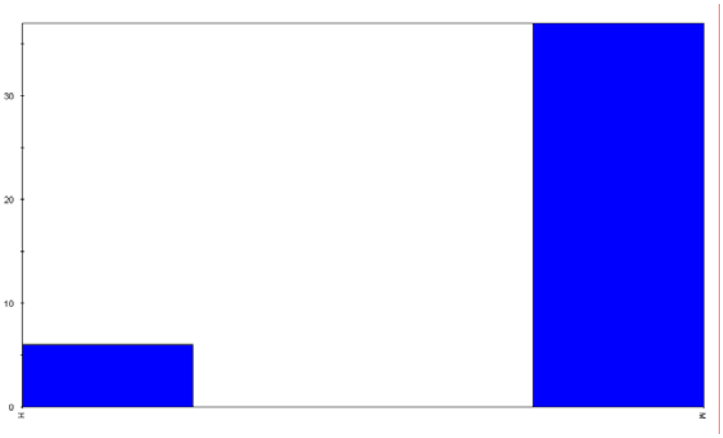


Podemos comprobar que aparte del valor de 1 unidad (aparece 1.342 veces), tenemos también valores importantes en 12, 10, 16,14,15,13.

En el **ítem de Vendedores** vemos con el atributo sexo que la mayoría de los vendedores que componen la compañía son mujeres

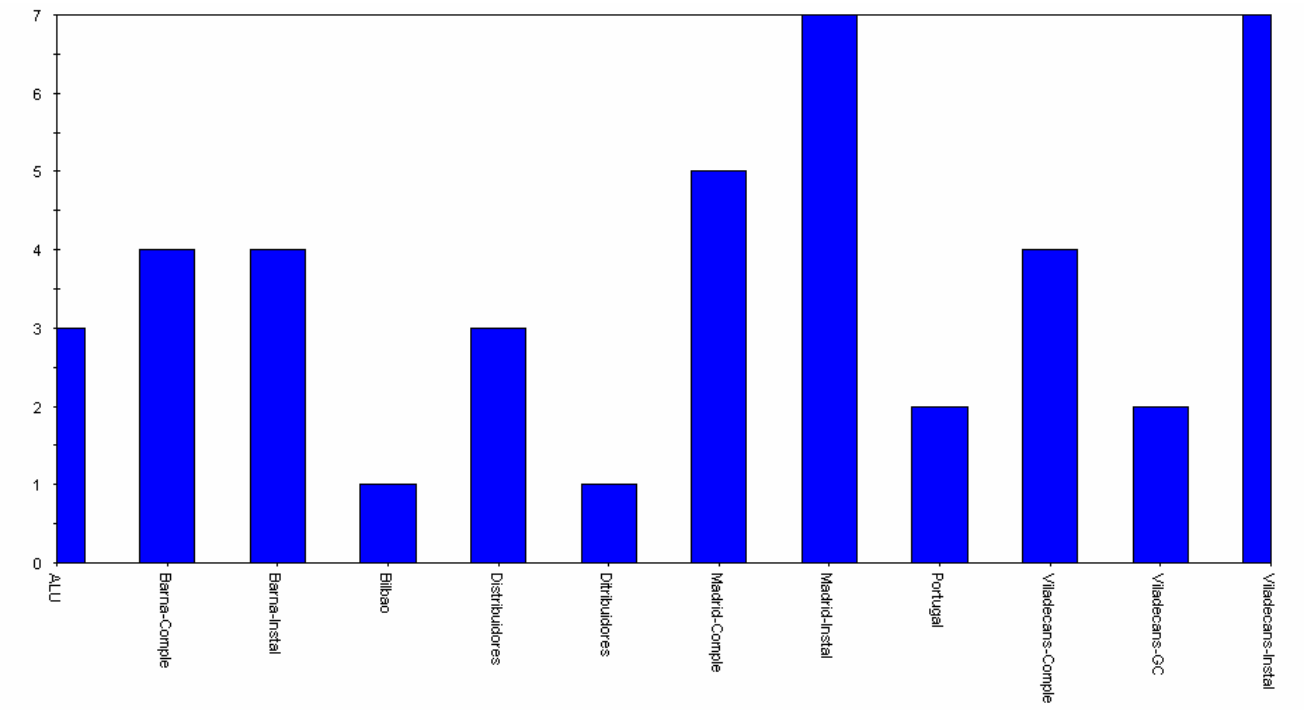
Valores | Gráfico

	Total	Acum.	%	Acum. %	Valor
1	6	6	13,95	13,95	H
2	37	43	86,05	100,00	M



En este mismo ítem , por las secciones (nombres) , vemos las que poseen mayor numero de vendedores

	Total	Acum.	%	Acum. %	Valor
1	3	3	6,98	6,98	ALU
2	4	7	9,30	16,28	Barna-Comple
3	4	11	9,30	25,58	Barna-Instal
4	1	12	2,33	27,91	Bilbao
5	3	15	6,98	34,88	Distribuidores
6	1	16	2,33	37,21	Ditribuidores
7	5	21	11,63	48,84	Madrid-Comple
8	7	28	16,28	65,12	Madrid-Instal
9	2	30	4,65	69,77	Portugal
10	4	34	9,30	79,07	Viladecans-Comple
11	2	36	4,65	83,72	Viladecans-GC
12	7	43	16,28	100,00	Viladecans-Instal

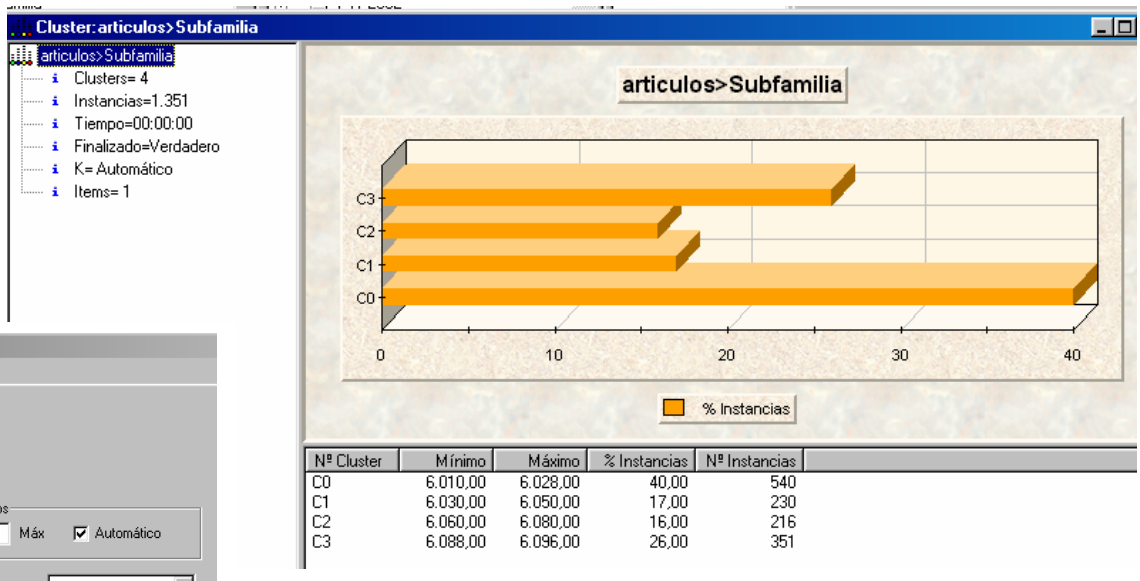


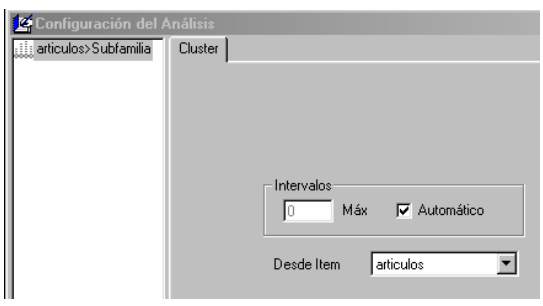
Que vemos son Madrid Instal y Viladecans Instal, aunque habría que ver si en los dos años de ventas estos vendedores estaban activos, es decir si realmente continúan o no, ya que la base de datos supongo mantiene todos los vendedores y no tiene en cuenta las bajas.

Análisis de ítems a través del Synera Discovery- Clusters

Primero realizo todos los análisis clusters sobre los atributos numéricos de los diferentes ítems, para ver las reglas que surgen y poder materializar los que nos interesen para binarizar o realizar otros análisis posteriores.

Item: Artículos , atributo :Subfamilia

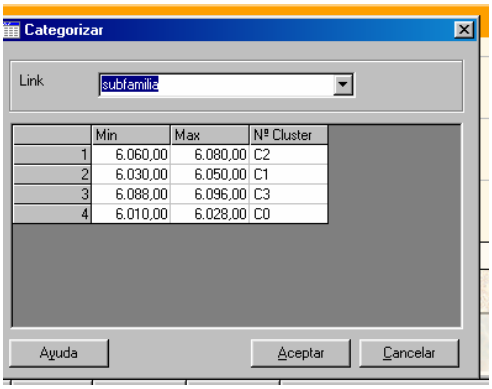




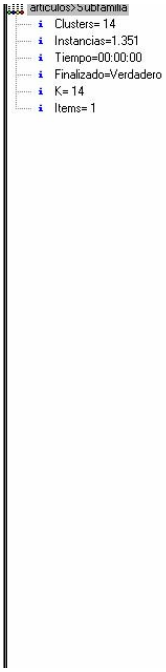
De las opciones posibles una es Categorizar , y así lo realizo

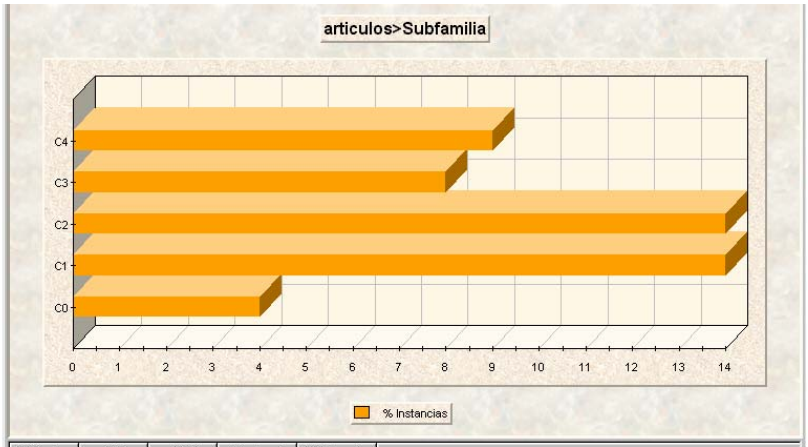
Nº Cluster	Mínimo	Máximo	% Instancias	Nº Instancias
C2	6.060,00	6.080,00	16,00	216
C1	6.030,00	6.050,00	17,00	230
C3	6.088,00	6.096,00	26,00	351
[illegible]		28,00	40,00	540

Categorizar...
Mostrar Instancias...
✔ Ordenar por Porcentaje
Ordenar por Id Cluster



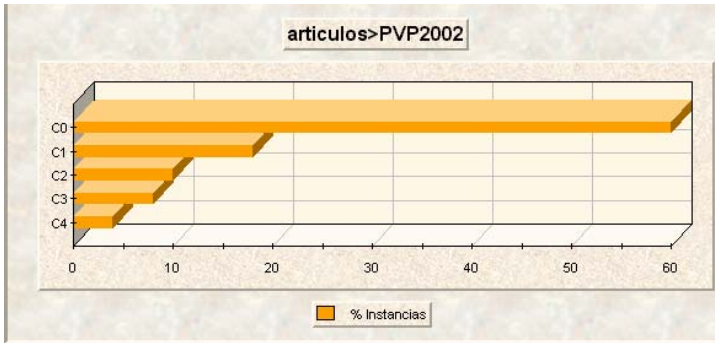
Como se ve tenemos otras opciones de ordenar por dos conceptos diferentes, mostrar las instancias o la categorización.
Realizo el mismo análisis cluster pero poniendo en vez de automático 14 , que son el numero de subfamilias que tenemos , y vuelvo a categorizarlo



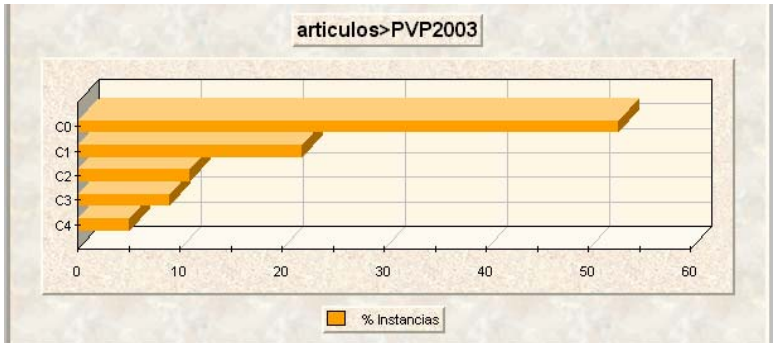


Nº Cluster	Mínimo	Máximo	% Instancias	Nº Instancias
C0	6.010,00	6.010,00	4,00	54
C1	6.020,00	6.020,00	14,00	189
C2	6.026,00	6.026,00	14,00	189
C3	6.028,00	6.028,00	8,00	108
C4	6.030,00	6.030,00	9,00	122
C5	6.032,00	6.032,00	2,00	27
C6	6.050,00	6.050,00	6,00	81
C7	6.060,00	6.060,00	5,00	68
C8	6.075,00	6.075,00	5,00	68
C9	6.080,00	6.080,00	7,00	95
C10	6.088,00	6.088,00	3,00	41
C11	6.089,00	6.089,00	13,00	176
C12	6.091,00	6.091,00	1,00	14
C13	6.096,00	6.096,00	10,00	135

Realizo lo mismo con los **atributos PVP2002 y PVP2003** y también los categorizo



Nº Cluster	Mínimo	Máximo	% Instancias	Nº Instancias
C4	526,00	2.479,00	4,00	54
C3	244,00	511,00	8,00	108
C2	135,00	240,00	10,00	135
C1	53,00	133,00	18,00	243
C0	0,00	52,00	60,00	811

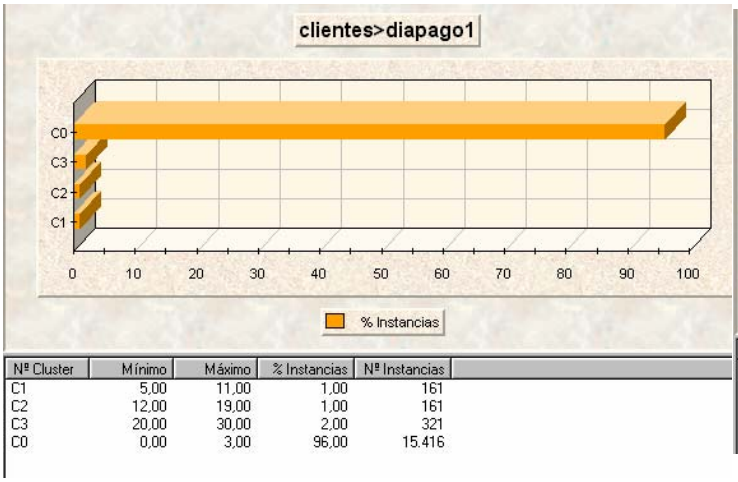


Nº Cluster	Mínimo	Máximo	% Instancias	Nº Instancias
C4	568,00	2.700,00	5,00	68
C3	253,00	550,00	9,00	122
C2	144,00	250,00	11,00	149
C1	62,00	140,00	22,00	297
C0	0,00	60,00	53,00	716

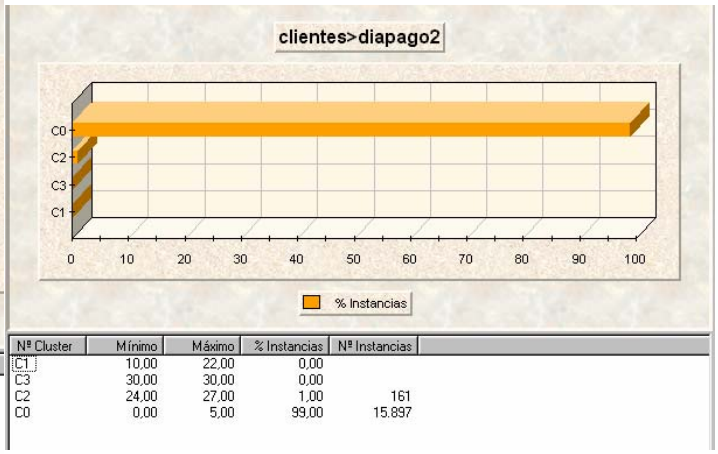
En ambos he ordenado por porcentaje, ya que me es más fácil detectar donde se concentran los precios, que en ambos casos es entre 0 y 60,00 Euros (PVP2003) y 0 y 52,00 Euros(PVP2002), aunque se ve claramente que hay más instancias en el año 2002 que en el 2003. También vemos que el importe máximo ha aumentado, por lo que es fácil deducir que ha habido un aumento de precios entre un año y otro, ya que todos los intervalos el máximo ha variado.

Item: Clientes, atributo : diapago1 y diapago2

Aquí realizo el análisis cluster automático y como se ve claramente, la mayoría de nuestros clientes no tienen indicado el día de pago, ya que la mayoría esta entre 0 y 3, después de los que si que lo poseen se concentran entre los días 20 y 31 de cada mes. En el caso del diapago2 todavía es más evidente la falta de datos, ya que solo existen 161 casos entre las fechas 24 y 27 de cada mes y el resto es entre 0 y 5. (en este caso no categorizo)



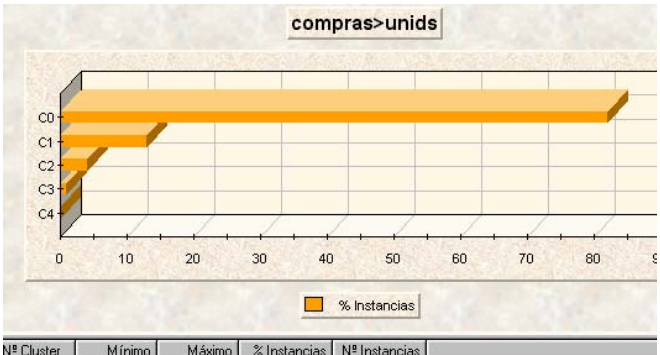
Nº Cluster	Mínimo	Máximo	% Instancias	Nº Instancias
C1	5,00	11,00	1,00	161
C2	12,00	19,00	1,00	161
C3	20,00	30,00	2,00	321
C0	0,00	3,00	96,00	15.416



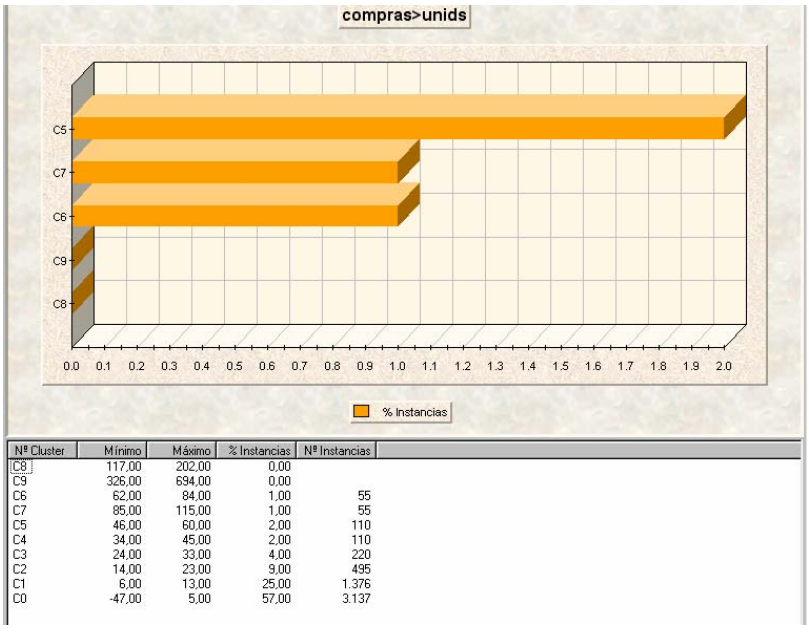
Nº Cluster	Mínimo	Máximo	% Instancias	Nº Instancias
C1	10,00	22,00	0,00	
C3	30,00	30,00	0,00	
C2	24,00	27,00	1,00	161
C0	0,00	5,00	99,00	15.897

Item: Compras, atributo : Unids (unidades)

Aquí podemos apreciar que las unidades de compra más usuales están entre los márgenes de –47 unidades y 12 (el –47 así como el resto de negativos nos informan de que se han producido devoluciones de material), aún así se confirma que se compran bastantes artículos pero en cantidades pequeñas. Como automáticamente me ha realizado solo 5 intervalos pruebo en forma manual con más intervalos para concretar más la información, así con el doble de intervalos, vemos cosas curiosas como que existen dos y dos (4 intervalos) con los mismos porcentajes



Nº Cluster	Mínimo	Máximo	% Instancias	Nº Instancias
C4	326,00	694,00	0,00	
C3	76,00	202,00	1,00	55
C2	34,00	75,00	4,00	220
C1	13,00	33,00	13,00	715
C0	-47,00	12,00	82,00	4.512

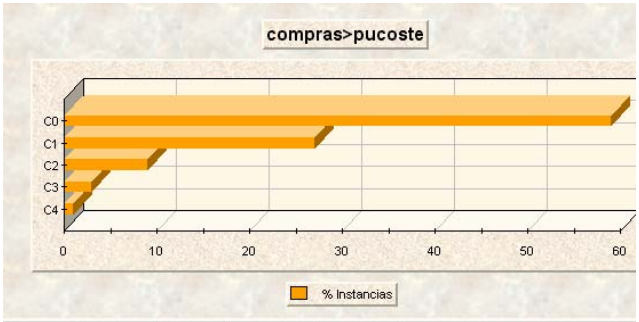


Nº Cluster	Mínimo	Máximo	% Instancias	Nº Instancias
C8	117,00	202,00	0,00	
C9	326,00	694,00	0,00	
C6	62,00	84,00	1,00	55
C7	85,00	115,00	1,00	55
C5	46,00	60,00	2,00	110
C4	34,00	45,00	2,00	110
C3	24,00	33,00	4,00	220
C2	14,00	23,00	9,00	495
C1	6,00	13,00	25,00	1.376
C0	-47,00	5,00	57,00	3.137

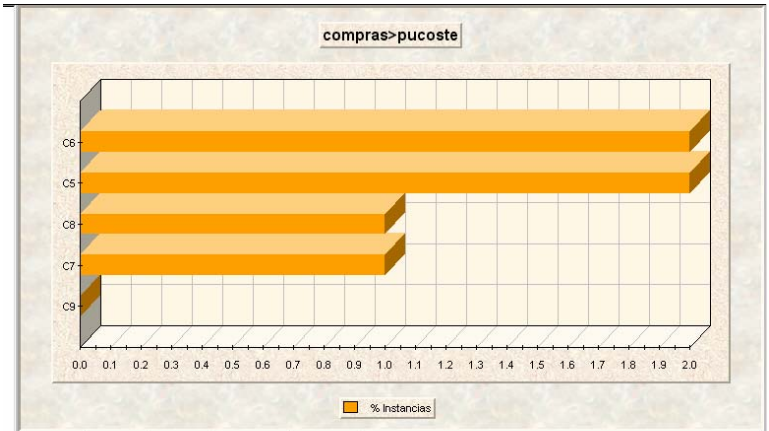
Vemos que el C6 y C7, tiene 55 instancias iguales y el C5 y C4, tienen 110 instancias, y confirmamos la concentración en pedidos de 0 a 5 unidades (el –47 no lo tengo en cuenta por tratarse seguro de devoluciones y casos puntuales es decir datos que se tendrían que haber quitado de la base ya que las devoluciones no son compras y nos incluyen ruido en la base)

Item: Compras, atributo : pucoste(preciounitariocoste)

Aquí también lo realizamos de forma automática y la concentración se encuentra entre 0 y 27 Euros, es decir vemos que la mayoría de productos comprados tienen costes pequeños. Comprobamos realizando el análisis manual doblando el intervalo como hemos realizado con las unidades si pasa lo mismo que antes ya que eso nos puede dar información de que ha existido una negociación especial al concentrarse las unidades y el precio en las mismas franjas. Efectivamente se confirma, que existen 4 intervalos (dos a dos con el mismo numero de instancias que antes)



Nº Cluster	Mínimo	Máximo	% Instancias	Nº Instancias
C4	256,00	578,00	1,00	55
C3	120,00	250,00	3,00	165
C2	61,00	119,00	9,00	495
C1	28,00	60,00	27,00	1.486
C0	0,00	27,00	59,00	3.247



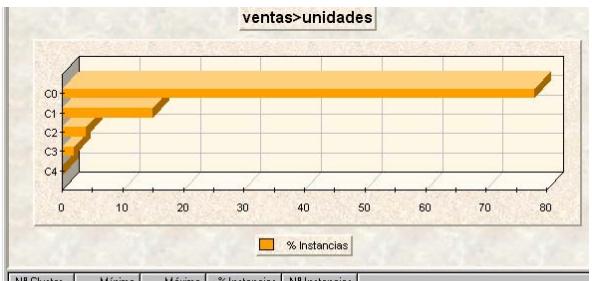
Nº Cluster	Mínimo	Máximo	% Instancias	Nº Instancias
C9	345,00	578,00	0,00	
C7	162,00	219,00	1,00	55
C8	225,00	330,00	1,00	55
C5	94,00	123,00	2,00	110
C6	124,00	160,00	2,00	110
C4	71,00	92,00	4,00	220
C3	52,00	70,00	8,00	440
C2	35,00	51,00	14,00	770
C1	19,00	34,00	24,00	1.321
C0	0,00	18,00	45,00	2.476

Item: Compras, atributo : importe(importe)

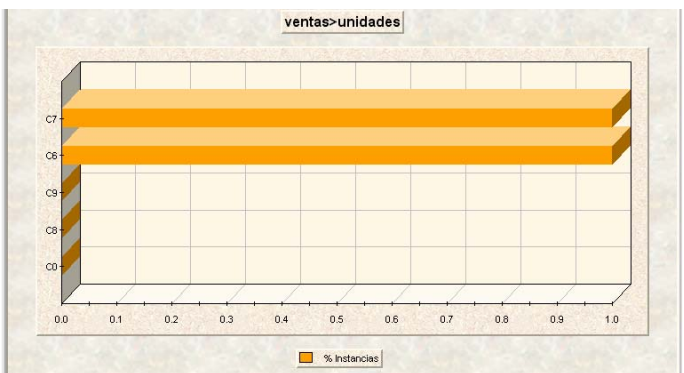
Aquí también lo realizamos de forma automática y me confirma lo visto anteriormente ya que si multiplicamos el valor del máximo del intervalo que tiene más instancias en unidades X preciounitario de coste, es el importe.

Item: Ventas, atributo : unidades

Aquí podemos deberíamos tener una concordancia con las unidades anteriormente analizadas de compras , ya que se supone que se compran los productos para venderlos así que comprobamos que las unidades de compra más usuales estaban entre los márgenes de –47 unidades y 12 , y vemos que las de ventas están entre –71 y 14 . si nos olvidamos de los negativos estaríamos



Nº Cluster	Mínimo	Máximo	% Instancias	Nº Instancias
C4	506,00	913,00	0,00	
C3	101,00	283,00	2,00	71
C2	43,00	100,00	4,00	142
C1	15,00	42,00	15,00	533
C0	-71,00	14,00	78,00	2.773

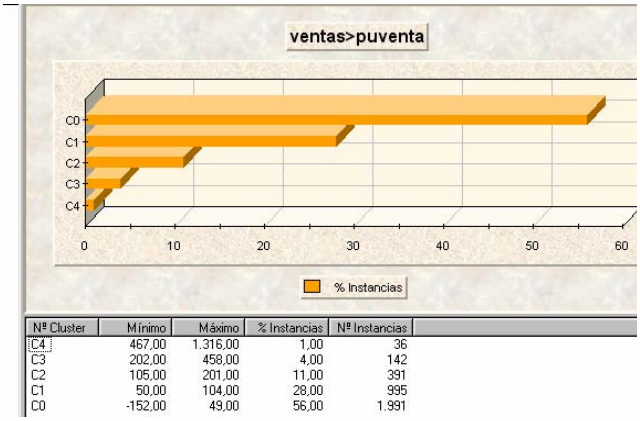


Nº Cluster	Mínimo	Máximo	% Instancias	Nº Instancias
C0	-71,00	-21,00	0,00	
C8	178,00	283,00	0,00	
C9	506,00	913,00	0,00	
C6	88,00	120,00	1,00	36
C7	121,00	171,00	1,00	36
C5	60,00	87,00	2,00	71
C4	37,00	57,00	3,00	107
C3	21,00	36,00	7,00	249
C2	9,00	20,00	18,00	640
C1	-14,00	8,00	67,00	2.382

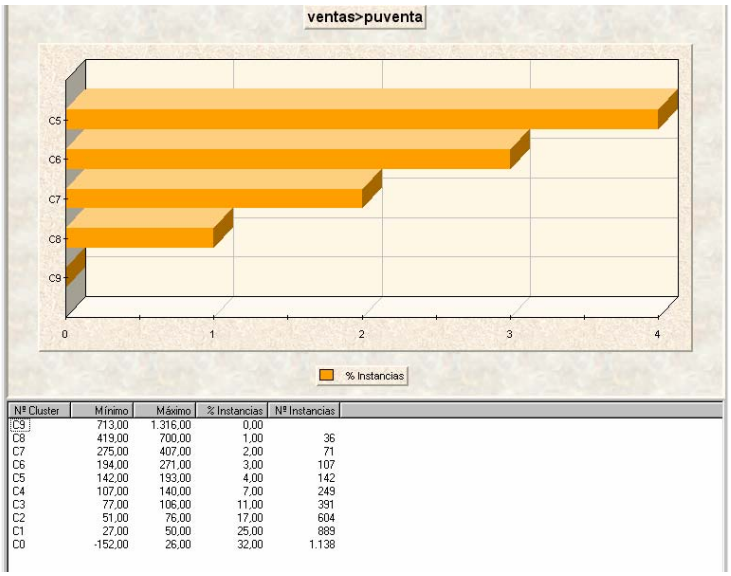
Más o menos equilibrados, lo veremos con una análisis de intervalos más amplio , como hicimos también anteriormente. Y nos encontramos con elementos similares como por ejemplo que existan dos intervalos C6 y C7 con el mismo numero de instancias , aunque en compras fueran 4 (dos a dos).

Item: Ventas, atributo : puventa (precio unitario de venta)

Aquí deberíamos tener una concordancia con las unidades de ventas igual que con el precio de coste unitarios aunque como uno es coste y el otro es venta debemos de tener en cuenta el margen de beneficio y evidentemente los intervalos estarán desplazados con respecto al coste este margen.



Nº Cluster	Mínimo	Máximo	% Instancias	Nº Instancias
C4	467,00	1.316,00	1,00	36
C3	202,00	458,00	4,00	142
C2	105,00	201,00	11,00	391
C1	50,00	104,00	28,00	995
C0	-152,00	49,00	56,00	1.991



Nº Cluster	Mínimo	Máximo	% Instancias	Nº Instancias
C9	713,00	1.316,00	0,00	
C8	419,00	700,00	1,00	36
C7	275,00	407,00	2,00	71
C6	194,00	271,00	3,00	107
C5	142,00	193,00	4,00	142
C4	107,00	140,00	7,00	249
C3	77,00	106,00	11,00	391
C2	51,00	76,00	17,00	604
C1	27,00	50,00	25,00	889
C0	-152,00	26,00	32,00	1.138

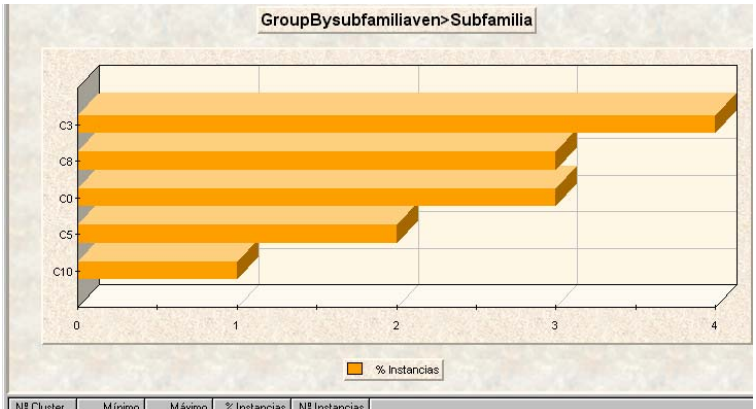
Realmente vemos que existe una concordancia , así como en la cuestión de las unidades.

Item: Ventas, atributo : importe

Igual que en compras este atributo es la composición de los anteriores y aunque compruebo los datos no incluyo las capturas por ser igual que multiplicar los dos anteriores.

Item: GroupBysubfamiliaven, atributo : subfamilia

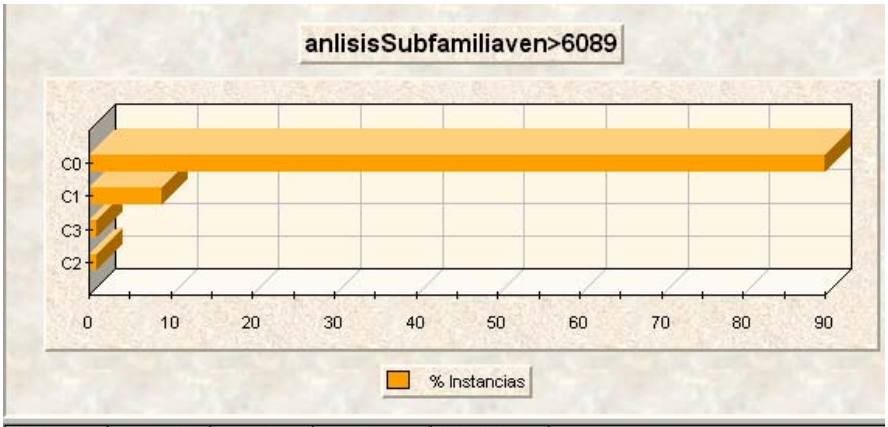
Este ítem analizado es uno de los originados a través del análisis de cubos , donde relacionaba la familia de productos y ventas a fin de confirmar que familia de productos es la que más se vende, y al hacer el análisis de cluster vemos claramente que es la subfamilia 6089. Viendo que le sigue de cerca la subfamilia 6096.Luego tenemos las subfamilias 6020 y 6026.



Nº Cluster	Mínimo	Máximo	% Instancias	Nº Instancias
C10	6.088,00	6.088,00	1,00	3
C5	6.032,00	6.032,00	2,00	6
C0	6.010,00	6.010,00	3,00	9
C8	6.075,00	6.075,00	3,00	9
C3	6.028,00	6.028,00	4,00	12
C7	6.060,00	6.060,00	4,00	12
C12	6.091,00	6.091,00	4,00	12
C6	6.050,00	6.050,00	5,00	15
C4	6.030,00	6.030,00	8,00	24
C9	6.080,00	6.080,00	9,00	26
C2	6.026,00	6.026,00	11,00	32
C1	6.020,00	6.020,00	12,00	35
C13	6.096,00	6.096,00	16,00	47
C11	6.089,00	6.089,00	20,00	59

Item: analisisSubfamiliaven , atributo : 6089

Este ítem analizado es uno de los originados a través del análisis de cubos, donde relacionaba La familia de productos y ventas a fin de confirmar que familia de productos es la que más se vende. Pero aquí en vez de estar agrupado como en el anterior estaba por separado, así que lo que pretendo al hacer el análisis de cluster es ver si el consumo de esta familia ha sido por pedidos grandes o un goteo de pedidos pequeños. Claramente se ve que se han tenido 4 pedidos grandes ya que tenemos 2 instancias de valor 52 y 2 de valor 171, el resto se mueve mayoritariamente entre 0 y 2.



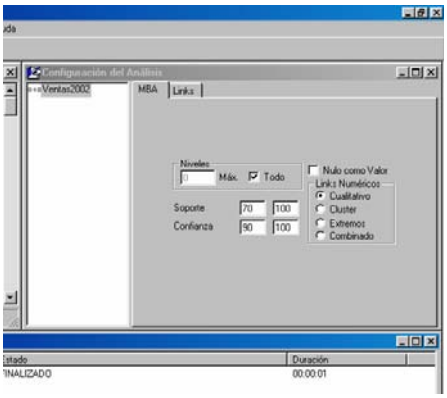
Nº Cluster	Mínimo	Máximo	% Instancias	Nº Instancias
C2	52,00	52,00	1,00	2
C3	171,00	171,00	1,00	2
C1	3,00	11,00	9,00	16
C0	0,00	2,00	90,00	157

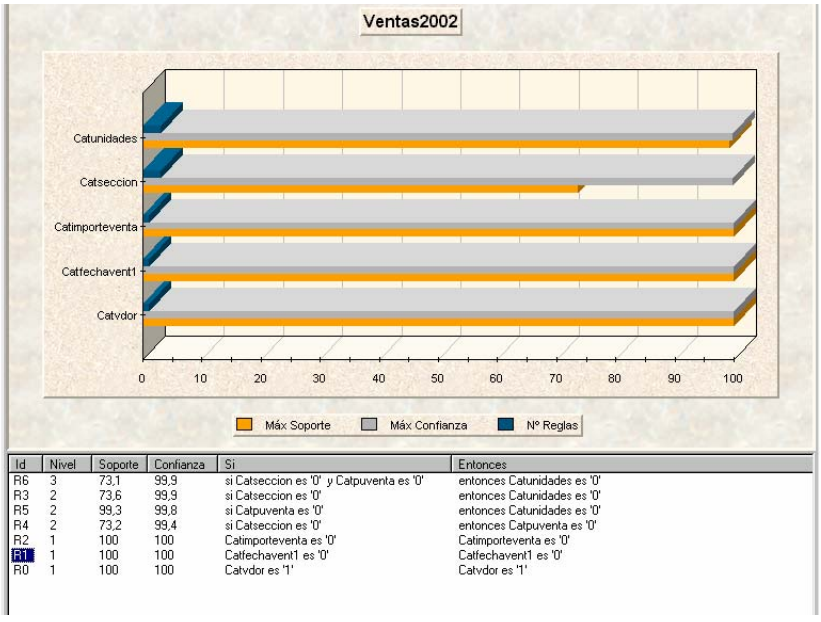
Análisis de ítems a través del Synera Discovery- MBA

Inicialmente analizo

Item: Ventas2002

Extraído de ventas, filtrando solo las ventas que se hicieron en el 2002 , para ver si podemos extraer alguna regla interesante. (atributos categorizados / binarizados) El primer análisis que realizo es con 30 de soporte y 50 de confianza y me salen numerosas reglas así que realizo otro con 70 de soporte y 90 de confianza



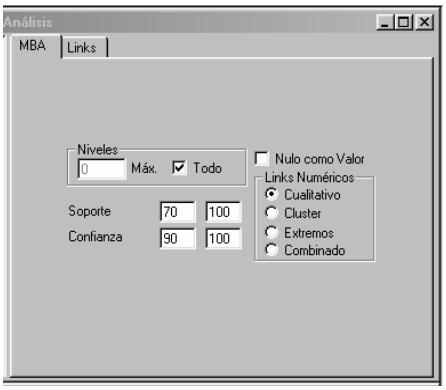


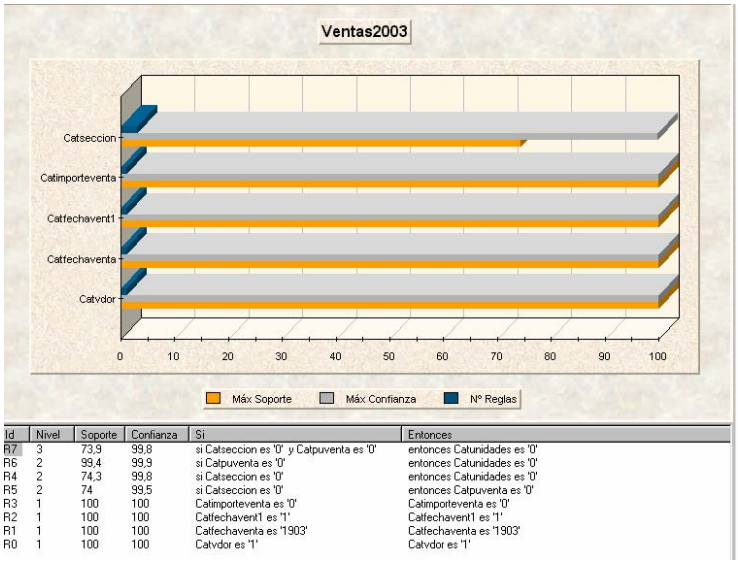
Id	Nivel	Soporte	Confianza	Si	Entonces
R6	3	73,1	99,9	si Catseccion es '0' y Catpuventa es '0'	entonces Catunidades es '0'
R3	2	73,6	99,9	si Catseccion es '0'	entonces Catunidades es '0'
R5	2	99,3	99,8	si Catpuventa es '0'	entonces Catunidades es '0'
R4	2	73,2	99,4	si Catseccion es '0'	entonces Catpuventa es '0'
R2	1	100	100	Catimporteventa es '0'	Catimporteventa es '0'
R1	1	100	100	Catfechavent1 es '0'	Catfechavent1 es '0'
R0	1	100	100	Catvdor es '1'	Catvdor es '1'

De las reglas extraídas no veo ninguna significativa que aporte más información interesante al estudio.

Item: Ventas2003

Extraído ventas, filtrando solo las ventas que se hicieron en el 2003 , para ver si podemos extraer alguna regla interesante. (atributos categorizados / binarizados) El primer análisis que realizo es con 30 de soporte y 50 de confianza y me salen numerosas reglas así que realizo otro con 70 de soporte y 90 de confianza

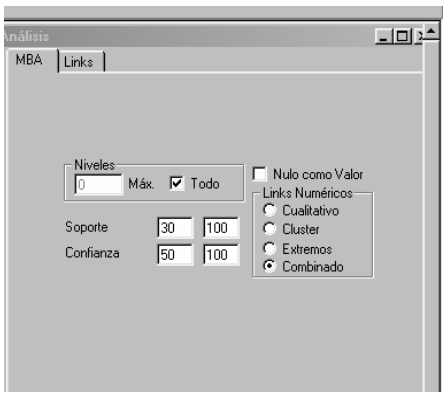


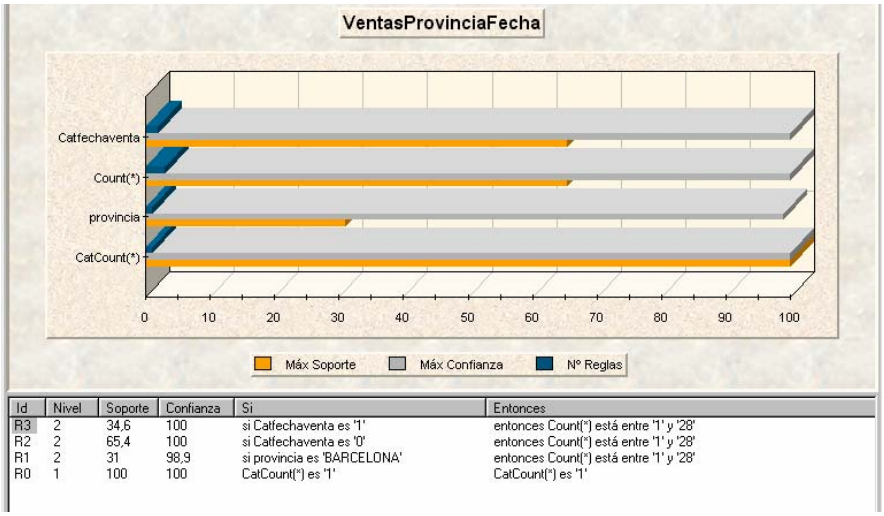


Id	Nivel	Soporte	Confianza	Si	Entonces
R7	3	73,9	99,8	si Catseccion es '0' y Catpuventa es '0'	entonces Catunidades es '0'
R6	2	99,4	99,9	si Catpuventa es '0'	entonces Catunidades es '0'
R4	2	74,3	99,8	si Catseccion es '0'	entonces Catunidades es '0'
R5	2	74	99,5	si Catseccion es '0'	entonces Catpuventa es '0'
R3	1	100	100	Catimporteventa es '0'	Catimporteventa es '0'
R2	1	100	100	Catfechavent1 es '1'	Catfechavent1 es '1'
R1	1	100	100	Catfechaventa es '1903'	Catfechaventa es '1903'
R0	1	100	100	Catvdor es '1'	Catvdor es '1'

Item: VentasProvinciaFecha

Extraído de la fusión de provincia de clientes y de fecha de venta de ventas,(atributos categorizados / binarizados) Realizo un análisis combinado.

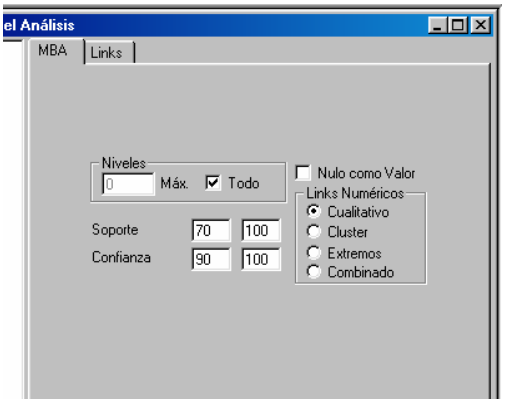


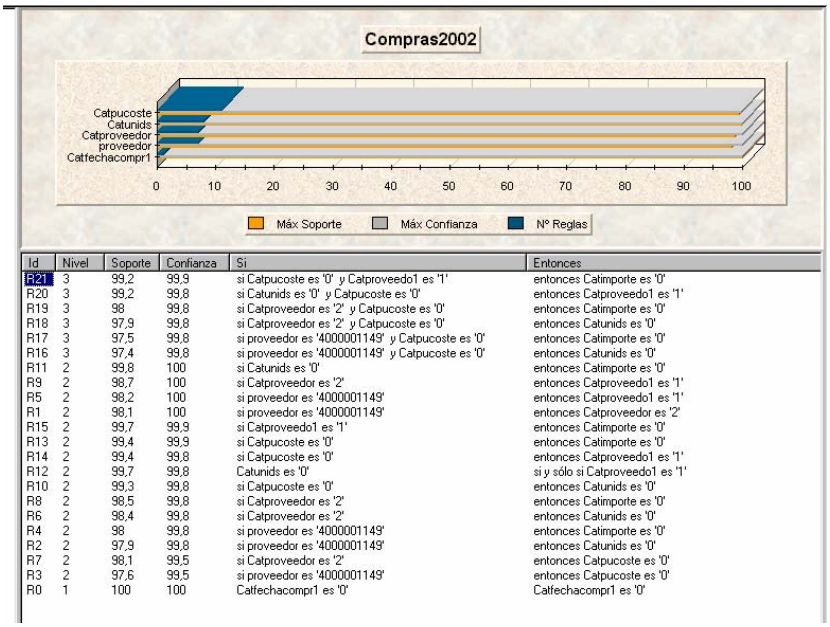


Id	Nivel	Soporte	Confianza	Si	Entonces
R3	2	34,6	100	si Catfechaventa es '1'	entonces Count(*) está entre '1' y '28'
R2	2	65,4	100	si Catfechaventa es '0'	entonces Count(*) está entre '1' y '28'
R1	2	31	98,9	si provincia es 'BARCELONA'	entonces Count(*) está entre '1' y '28'
R0	1	100	100	CatCount(*) es '1'	CatCount(*) es '1'

Item: Compras2002

Este link también es de una consulta filtrada donde solo salen las compras del año 2002, Aquí se ve las reglas que surgen, aunque ninguna es significativa.





Id	Nivel	Soporte	Confianza	Si	Entonces
R21	3	99,2	99,9	si Catpucoste es '0' y Catproveedo1 es '1'	entonces Catimporte es '0'
R20	3	99,2	99,8	si Catunids es '0' y Catpucoste es '0'	entonces Catproveedo1 es '1'
R19	3	98	99,8	si Catproveedor es '2' y Catpucoste es '0'	entonces Catimporte es '0'
R18	3	97,9	99,8	si Catproveedor es '2' y Catpucoste es '0'	entonces Catunids es '0'
R17	3	97,5	99,8	si proveedor es '4000001149' y Catpucoste es '0'	entonces Catimporte es '0'
R16	3	97,4	99,8	si proveedor es '4000001149' y Catpucoste es '0'	entonces Catunids es '0'
R11	2	99,8	100	si Catunids es '0'	entonces Catimporte es '0'
R9	2	98,7	100	si Catproveedor es '2'	entonces Catproveedo1 es '1'
R5	2	98,2	100	si proveedor es '4000001149'	entonces Catproveedo1 es '1'
R1	2	98,1	100	si proveedor es '4000001149'	entonces Catproveedor es '2'
R15	2	99,7	99,9	si Catproveedo1 es '1'	entonces Catimporte es '0'
R13	2	99,4	99,9	si Catpucoste es '0'	entonces Catimporte es '0'
R14	2	99,4	99,8	si Catpucoste es '0'	entonces Catproveedo1 es '1'
R12	2	99,7	99,8	Catunids es '0'	si y sólo si Catproveedo1 es '1'
R10	2	99,3	99,8	si Catpucoste es '0'	entonces Catunids es '0'
R8	2	98,5	99,8	si Catproveedor es '2'	entonces Catimporte es '0'
R6	2	98,4	99,8	si Catproveedor es '2'	entonces Catunids es '0'
R4	2	98	99,8	si proveedor es '4000001149'	entonces Catimporte es '0'
R2	2	97,9	99,8	si proveedor es '4000001149'	entonces Catunids es '0'
R7	2	98,1	99,5	si Catproveedor es '2'	entonces Catpucoste es '0'
R3	2	97,6	99,5	si proveedor es '4000001149'	entonces Catpucoste es '0'
R0	1	100	100	Catfechacompr1 es '0'	Catfechacompr1 es '0'

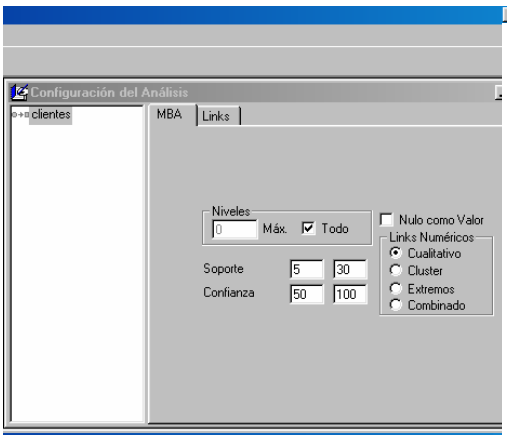
Item: Ventas

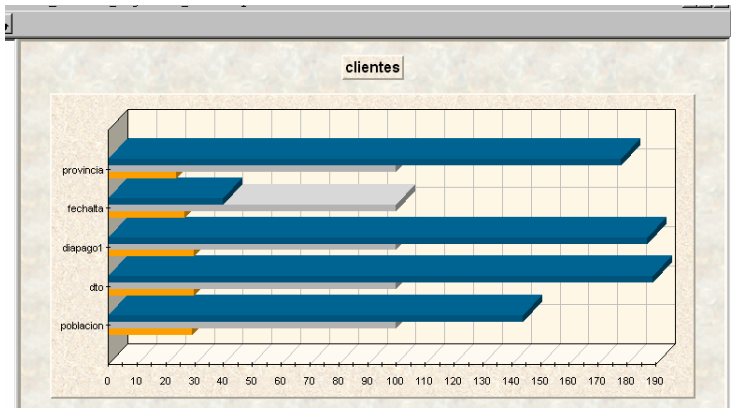
Es el link general que miro haber si surge algo diferentes, y la verdad es que confirma afirmaciones extraídas en los análisis clusters , pero no veo ninguna regla significativa.

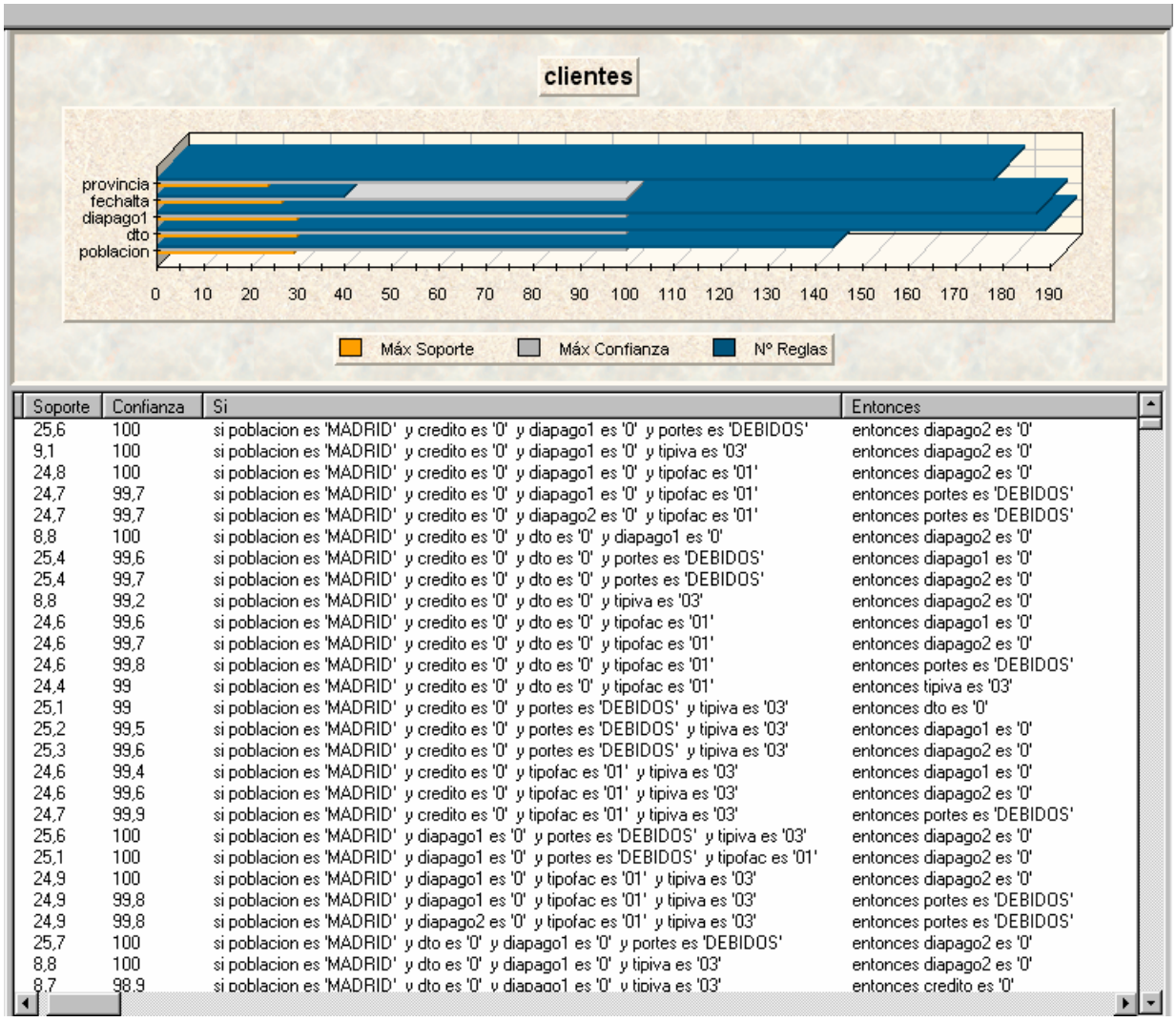
Id	Nivel	Soporte	Confianza	Si	Entonces
R54	3	41,8	99,8	si Catfechavent1 es '0' y Catpuventa es '0'	entonces Catunidades es '0'
R52	3	31	99,9	si Catfechavent1 es '0' y Catseccion es '0'	entonces Catunidades es '0'
R53	3	30,9	99,4	si Catfechavent1 es '0' y Catseccion es '0'	entonces Catpuventa es '0'
R57	3	57,5	99,9	si Catfechavent1 es '1' y Catpuventa es '0'	entonces Catunidades es '0'
R59	3	57,6	100	si Catfechavent1 es '1' y Catpuventa es '0'	entonces Catimporteventa es '0'
R49	3	33,6	78	si Catfechavent1 es '1' y Catseccion es '0'	entonces unidades está entre '-71' y '14'
R55	3	43	99,8	si Catfechavent1 es '1' y Catseccion es '0'	entonces Catunidades es '0'
R56	3	42,8	99,5	si Catfechavent1 es '1' y Catseccion es '0'	entonces Catpuventa es '0'
R58	3	57,7	100	si Catfechavent1 es '1' y Catunidades es '0'	entonces Catimporteventa es '0'
R60	3	37,4	99,4	si Catfechavent1 es '1' y clusterunidades es 'C1'	entonces Catpuventa es '0'
R61	3	73,6	99,8	si Catseccion es '0' y Catpuventa es '0'	entonces Catunidades es '0'
R62	3	50,1	99,2	si Catseccion es '0' y clusterunidades es 'C1'	entonces Catpuventa es '0'
R63	3	99,3	100	si Catunidades es '0' y Catpuventa es '0'	entonces Catimporteventa es '0'
R46	3	32,5	99,7	si puventa está entre '-152' y '49' y Catfechavent1 es '1'	entonces Catunidades es '0'
R45	3	30	75,5	si puventa está entre '-152' y '49' y Catseccion es '0'	entonces unidades está entre '-71' y '14'
R47	3	39,7	99,7	si puventa está entre '-152' y '49' y Catseccion es '0'	entonces Catunidades es '0'
R48	3	33,8	99,3	si unidades está entre '-71' y '14' y Catfechavent1 es '0'	entonces Catpuventa es '0'
R50	3	44,2	99,4	si unidades está entre '-71' y '14' y Catfechavent1 es '1'	entonces Catpuventa es '0'
R51	3	58,4	99,3	si unidades está entre '-71' y '14' y Catseccion es '0'	entonces Catpuventa es '0'
R16	2	34	80,6	si Catfechavent1 es '0'	entonces unidades está entre '-71' y '14'
R23	2	31,1	73,7	si Catfechavent1 es '0'	entonces Catseccion es '0'
R24	2	42,1	99,8	si Catfechavent1 es '0'	entonces Catunidades es '0'
R25	2	41,9	99,5	si Catfechavent1 es '0'	entonces Catpuventa es '0'
R26	2	42,1	100	si Catfechavent1 es '0'	entonces Catimporteventa es '0'
R17	2	44,5	76,9	si Catfechavent1 es '1'	entonces unidades está entre '-71' y '14'
R27	2	43	74,4	si Catfechavent1 es '1'	entonces Catseccion es '0'
R28	2	57,8	99,9	si Catfechavent1 es '1'	entonces Catunidades es '0'
R29	2	57,6	99,6	si Catfechavent1 es '1'	entonces Catpuventa es '0'
R30	2	57,8	100	si Catfechavent1 es '1'	entonces Catimporteventa es '0'
R31	2	37,7	65,1	si Catfechavent1 es '1'	entonces clusterunidades es 'C1'

Item: Clientes

Aquí he parado el análisis pero la diferencia con los otros es que me salen muchisimas reglas unas 457 .... más o menos







Soporte	Confianza	Si	Entonces
25,6	100	si poblacion es 'MADRID' y credito es '0' y diapago1 es '0' y portes es 'DEBIDOS'	entonces diapago2 es '0'
9,1	100	si poblacion es 'MADRID' y credito es '0' y diapago1 es '0' y tipiva es '03'	entonces diapago2 es '0'
24,8	100	si poblacion es 'MADRID' y credito es '0' y diapago1 es '0' y tipofac es '01'	entonces diapago2 es '0'
24,7	99,7	si poblacion es 'MADRID' y credito es '0' y diapago1 es '0' y tipofac es '01'	entonces portes es 'DEBIDOS'
24,7	99,7	si poblacion es 'MADRID' y credito es '0' y diapago2 es '0' y tipofac es '01'	entonces portes es 'DEBIDOS'
8,8	100	si poblacion es 'MADRID' y credito es '0' y dto es '0' y diapago1 es '0'	entonces diapago2 es '0'
25,4	99,6	si poblacion es 'MADRID' y credito es '0' y dto es '0' y portes es 'DEBIDOS'	entonces diapago1 es '0'
25,4	99,7	si poblacion es 'MADRID' y credito es '0' y dto es '0' y portes es 'DEBIDOS'	entonces diapago2 es '0'
8,8	99,2	si poblacion es 'MADRID' y credito es '0' y dto es '0' y tipiva es '03'	entonces diapago2 es '0'
24,6	99,6	si poblacion es 'MADRID' y credito es '0' y dto es '0' y tipofac es '01'	entonces diapago1 es '0'
24,6	99,7	si poblacion es 'MADRID' y credito es '0' y dto es '0' y tipofac es '01'	entonces diapago2 es '0'
24,6	99,8	si poblacion es 'MADRID' y credito es '0' y dto es '0' y tipofac es '01'	entonces portes es 'DEBIDOS'
24,4	99	si poblacion es 'MADRID' y credito es '0' y dto es '0' y tipofac es '01'	entonces tipiva es '03'
25,1	99	si poblacion es 'MADRID' y credito es '0' y portes es 'DEBIDOS' y tipiva es '03'	entonces dto es '0'
25,2	99,5	si poblacion es 'MADRID' y credito es '0' y portes es 'DEBIDOS' y tipiva es '03'	entonces diapago1 es '0'
25,3	99,6	si poblacion es 'MADRID' y credito es '0' y portes es 'DEBIDOS' y tipiva es '03'	entonces diapago2 es '0'
24,6	99,4	si poblacion es 'MADRID' y credito es '0' y tipofac es '01' y tipiva es '03'	entonces diapago1 es '0'
24,6	99,6	si poblacion es 'MADRID' y credito es '0' y tipofac es '01' y tipiva es '03'	entonces diapago2 es '0'
24,7	99,9	si poblacion es 'MADRID' y credito es '0' y tipofac es '01' y tipiva es '03'	entonces portes es 'DEBIDOS'
25,6	100	si poblacion es 'MADRID' y diapago1 es '0' y portes es 'DEBIDOS' y tipiva es '03'	entonces diapago2 es '0'
25,1	100	si poblacion es 'MADRID' y diapago1 es '0' y portes es 'DEBIDOS' y tipofac es '01'	entonces diapago2 es '0'
24,9	100	si poblacion es 'MADRID' y diapago1 es '0' y tipofac es '01' y tipiva es '03'	entonces diapago2 es '0'
24,9	99,8	si poblacion es 'MADRID' y diapago1 es '0' y tipofac es '01' y tipiva es '03'	entonces portes es 'DEBIDOS'
24,9	99,8	si poblacion es 'MADRID' y diapago2 es '0' y tipofac es '01' y tipiva es '03'	entonces portes es 'DEBIDOS'
25,7	100	si poblacion es 'MADRID' y dto es '0' y diapago1 es '0' y portes es 'DEBIDOS'	entonces diapago2 es '0'
8,8	100	si poblacion es 'MADRID' y dto es '0' y diapago1 es '0' y tipiva es '03'	entonces diapago2 es '0'
8,7	98,9	si poblacion es 'MADRID' y dto es '0' y diapago1 es '0' y tipiva es '03'	entonces credito es '0'

Vemos que existen algunas reglas muy curiosas que ya nos indican información del departamento o tienda que no rellena demasiado bien las fichas de clientes, seguimos mirando

Soporte	Confianza	Si	Entonces
5,4	100	si provincia es 'BARCELONA' y vendedor es '04' y diapago1 es '0'	entonces credito es '0' y diapago2 es '0'
5,4	100	si provincia es 'BARCELONA' y vendedor es '04' y diapago1 es '0'	entonces credito es '0' y tipiva es '03'
5,4	100	si provincia es 'BARCELONA' y vendedor es '04' y diapago1 es '0'	entonces diapago2 es '0' y tipiva es '03'
5,4	100	si provincia es 'BARCELONA' y vendedor es '04' y diapago1 es '0' y fechalta es '1...	entonces credito es '0'
5,4	100	si provincia es 'BARCELONA' y vendedor es '04' y diapago1 es '0' y fechalta es '1...	entonces diapago2 es '0'
5,4	100	si provincia es 'BARCELONA' y vendedor es '04' y diapago1 es '0' y fechalta es '1...	entonces tipiva es '03'
5,6	100	si provincia es 'BARCELONA' y vendedor es '04' y diapago2 es '0'	entonces credito es '0' y tipiva es '03'
5,6	100	si provincia es 'BARCELONA' y vendedor es '04' y diapago2 es '0' y fechalta es '1...	entonces credito es '0'
5,6	100	si provincia es 'BARCELONA' y vendedor es '04' y diapago2 es '0' y fechalta es '1...	entonces tipiva es '03'
5,6	100	si provincia es 'BARCELONA' y vendedor es '04' y dto es '0'	entonces credito es '0' y tipiva es '03'
5,4	100	si provincia es 'BARCELONA' y vendedor es '04' y dto es '0' y diapago1 es '0'	entonces credito es '0'
5,4	100	si provincia es 'BARCELONA' y vendedor es '04' y dto es '0' y diapago1 es '0'	entonces diapago2 es '0'
5,4	99,9	si provincia es 'BARCELONA' y vendedor es '04' y dto es '0' y diapago1 es '0'	entonces fechalta es '16/12/01'
5,4	100	si provincia es 'BARCELONA' y vendedor es '04' y dto es '0' y diapago1 es '0'	entonces tipiva es '03'
5,6	100	si provincia es 'BARCELONA' y vendedor es '04' y dto es '0' y diapago2 es '0'	entonces credito es '0'
5,5	99,9	si provincia es 'BARCELONA' y vendedor es '04' y dto es '0' y diapago2 es '0'	entonces fechalta es '16/12/01'
5,6	100	si provincia es 'BARCELONA' y vendedor es '04' y dto es '0' y diapago2 es '0'	entonces tipiva es '03'
5,6	100	si provincia es 'BARCELONA' y vendedor es '04' y dto es '0' y fechalta es '16/12/01'	entonces credito es '0'
5,6	100	si provincia es 'BARCELONA' y vendedor es '04' y dto es '0' y fechalta es '16/12/01'	entonces tipiva es '03'
5,6	100	si provincia es 'BARCELONA' y vendedor es '04' y fechalta es '16/12/01'	entonces credito es '0' y tipiva es '03'
6	100	si provincia es 'BARCELONA' y vendedor es '07' y credito es '0' y diapago1 es '0'	entonces diapago2 es '0'
6	100	si provincia es 'BARCELONA' y vendedor es '07' y credito es '0' y diapago1 es '0'	entonces tipiva es '03'
6,4	100	si provincia es 'BARCELONA' y vendedor es '07' y credito es '0' y diapago2 es '0'	entonces tipiva es '03'
6,1	100	si provincia es 'BARCELONA' y vendedor es '07' y credito es '0' y dto es '0'	entonces tipiva es '03'
6	100	si provincia es 'BARCELONA' y vendedor es '07' y diapago1 es '0'	entonces diapago2 es '0' y tipiva es '03'
5,6	99,9	si provincia es 'BARCELONA' y vendedor es '07' y dto es '0' y diapago1 es '0'	entonces credito es '0'
5,6	100	si provincia es 'BARCELONA' y vendedor es '07' y dto es '0' y diapago1 es '0'	entonces diapago2 es '0'

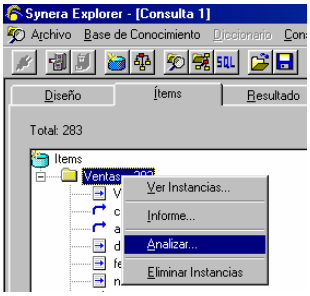
*Este primer análisis del link de clientes , fue realizado sin datos binarizados , ni categorizados.

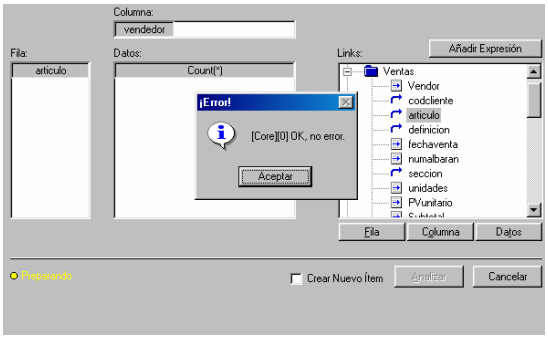
Filtro Clientes=Barcelona con Ventas=año2002

Realmente como para realizar los análisis MBA , primero hay que binarizar o categorizar , como para nuestro cliente lo importante es ver reglas comportamientos y/o conocimiento que podamos extraer de datos entre ventas (fechas) y provincias (localizaciones geográficas) , empiezo realizando una consulta con estos parámetros de filtro (los indicados arriba)

Plan de Consulta

S.	Modo	Objeto	Operador...	Valor	Total	Potencial...	Seleccion...	Tiempo	Peso
●	0	Clientes -> provincia	=	BARCEL...	8.347	8.347	8.347	00:00:00,0	0,000005
●	Rel Inve	Ventas -> codcliente			591	3.631	591	00:00:00,1	
●	Y	Ventas -> fechaventa	Between	03/01/1...	283	3.631	283	00:00:00,1	0,000979





Exporto la consulta en un ítem , y a parte la analizo, el problema es que cada vez que analizo , mejor dicho lo intento ponga lo que le ponga , me acaba saliendo , el error capturado. Por tanto no puedo realizar nada , ni exportar , para binarizar fuera, ni nada. Ya que tampoco me deja agrupar.

De Datos a Conocimiento:

Una vez realizado los diferentes análisis en el Synera de ellos podemos deducir muchas cosas que transformamos en conocimiento dado que nos podrá permitir en nuestro caso aconsejar a la empresa que nos facilito los datos estrategias o medidas a tomar para que sean más efectivas sus acciones así como que seria lo aconsejable dependiendo de la política de la misma.

A continuación expondré un resumen de puntos importantes encontrados:

- Se ha detectado una centralización de Clientes en una o dos zonas geográficas, eso evidentemente esta asociado a donde la empresa posee tienda directa, por lo cual habría que plantearse o bien la apertura de nuevas tiendas o la entrada de más comerciales que cubrieran las zonas con menos influencia, además de hacer campañas publicitarias muy concretas. (Al no tener el sector en los datos facilitados no podemos detectar que sector es el de mayor consumo, cual el de menos y los intermedios, ya que este dato a nivel de las campañas publicitarias es muy importante a fin de centralizar esfuerzos.)
- En cuanto a las ventas podemos ver que la mayoría de productos se venden unitariamente por lo cual, habría que ver de realizar pedidos mayores. Aunque han existido varios pedidos de bastantes unidades, aunque el consumo normal se centra entre un margen de 1 a 15 unidades. Para profundizar más en este aspecto del porque deberíamos tener más datos y saber más del producto.
- En cuanto a los vendedores vemos que también esta muy centralizado a pocas personas, en este punto habría quizás que conseguir más datos ya que a lo mejor estas son las responsables de las tiendas y no salen en los datos facilitados todas las vendedoras que están involucradas en una venta. También se ve que el porcentaje más elevado del equipo de ventas son mujeres y que existe una mayor cantidad de ventas en los departamentos de Madrid-Instal y Viladecans-Instal , aunque también puede suponer que existe mayor rotación ya que consideramos que en la base facilitada están todos los vendedores y que habría que quizás profundizar más en este aspecto viendo quien ha vendido en el 2002 que ya no figura en el 2002. Aunque para eso los datos que poseemos son insuficientes.
- También observamos que las ventas de esta familia de productos esta concentrada en pocos clientes, aunque repetitivos, para poder saber más sobre conductas y el porque de la venta seria muy interesante tener más datos de dichos clientes. Aunque se ve que la fidelización al producto existe y deberíamos seguir incidiendo en este punto.
- Respecto a los artículos vemos que las ventas se centran en la subfamilia 6089 y 6096, siguiéndoles la 6020 y 6026, por lo cual se tendría que mirar de promocionar las otras subfamilias de manera de tener unas mayores ventas en las mismas y analizar los motivos de que se vendan menos.

También debido a las dificultades encontradas con los datos le aconsejamos varias cuestiones a la empresa a fin de que en un futuro se pueda extraer mucho más conocimiento de sus datos que debido al formato, la falta de limpieza de los mismos, el no mantenimiento de los mismos, y el desorden así como la falta de otros datos que serian importantes incorporar no podemos ver más conductas ni de clientes, ni de vendedores.

- Actualización de Datos
- Ampliación de Datos de los Clientes, como por ejemplo actividad y/o sector.
- Planteamiento de cambio de la Base de Datos actuales o posibilidad de plantearse un programa informático más integral con una buena base de datos y evidentemente un responsable de la misma para asegurar que los datos que están en ella sean validos.
- Estudiar alguna herramienta que este más enfocada a nivel Marketing / Ventas.

Indice

7. – Otras Herramientas Comerciales para Data Mining :

Después de múltiples búsquedas en Internet he localizado información a cerca de las siguientes herramientas para realizar procesos de Minería de Datos

Darwin (Thinking Machines): www.oracle.com/ip/analyza/warehouse/datamining/index.html

Herramientas:

- StarTree: construye árboles de decisión usando el criterio de CART
- StarNet: entrena una red neuronal *feed-forward*. El usuario especifica el número de capas y las neuronas por capa. La regla de entrenamiento puede ser: *backpropagation*, *modified Newton*, *steepest descent* y *conjugate gradient*.
- StarMatch: encuentra los ejemplos prototípicos usando razonamiento basado en casos o instancias usando la medida de los vecinos más cercanos (*k-nearest neighbours*). La distancia es Euclideana y los pesos los puede asignar el usuario.
- StarGene: usa algoritmos genéticos para optimizar los parámetros asociados con otras técnicas (número de capas ocultas, pesos de los parámetros en StarMatch, etc).
- StarView: herramientas diversas de visualización de datos.
- StarDB: interface a bases de datos.
- StarData: herramientas para manipular bases de datos. Sirve de interface entre Darwin y manejadores comerciales de bases de datos y deja una base de datos en un formato entendible por todas las herramientas de Darwin.

Plataformas :

- Win NT y Unix

Interface:

- Oracle

MineSet (Silicon Graphics):www.sgi.com/software/miniset/

Herramientas: Algoritmos de minería de datos:

- árboles de decisión
- árboles de opción (árboles de decisión con varias opciones en cada nodo)
- naive Bayes: determina la probabilidad de un evento basado en un atributo dado
- generador de reglas

Plataformas :

- Unix

Interface:

- Oracle, Sybase, Informix

Clementine (Intelligenza, S.A.): www.spss.com

Tiene menús para selección de:

- datos: ASCII o tablas de bases de datos tomadas de ORACLE, Ingres, Sybase, Informix, etc).
- registros: selecciona, mezcla (*merge*), muestrea, balancea
- campos: filtro, deriva nuevos campos, selecciona por tipo, llena información faltante
- gráficas: gráfica, histograma, distribución, red
- salidas: tablas, análisis, matriz, estadísticas

Herramientas :

- red neuronal, C4.5, Kohonen y generación de reglas

Plataformas :

- Unix

Interface:

- ODBC

DBMiner (Simon Fraser University, Canadá)

Herramientas:

- Caracterizador: encuentra relaciones generales entre datos
- Discriminador: encuentra reglas que distinguen clases
- Clasificador: construye modelos de clases basados en reglas
- Reglas de asociación: del tipo, donde *X* y *Y* son conjuntos e implican que *Y* ocurre cuando ocurre *X*.
- Meta-reglas: usa formato de lógica de segundo orden para buscar relaciones en los datos
- Predictor: predice valores faltantes basándose en información relacionada
- Evaluador de evolución de datos: encuentra tendencias en los datos
- Evaluador de desviaciones: encuentra desviaciones de tendencias en los datos
  También utiliza *Data-Cube* (una generalización de queries en SQL).

DataMine (Rutgers University)

Herramientas:

Encuentra reglas de asociación con medidas de soporte (cuantos ejemplos la satisfacen) y confianza (relación entre cuantos ejemplos satisfacen la regla y cuantos satisfacen sólo la parte izquierda).
Usa extensiones a SQL, el operador MINE, que encuentra todas las reglas que satisfacen ciertas condiciones (e.g., intervalos de confianza y de soporte).

Quest o **IBM Intelligent Miner** (IBM) www.ibm.com/software/data/iminer

Herramientas:

- reglas de asociación del tipo.(donde *X* y *Y* son conjuntos)
- patrones secuenciales
- clustering de series de tiempo
- clasificación basada en árboles de decisión herramientas de partición de datos
- algunos de los algoritmos paralelizados en IBM-SP2

Plataformas :

- Unix

Interface:

- IBM y DB2

INLEN (Michalski et al.)

Consiste en una base de datos conectada a una base de conocimiento y un conjunto de operadores.
Tiene varios operadores para manejar datos y conocimiento: seleccionar, crear, proyectar, insertar, unir, cambiar, combinar, borrar, interceptar.
Operadores de generación de conocimiento:

- genrule: basado en AQ15c
- gentree: genera estructuras de decisión. Son como árboles, pero los nodos pueden tener conjuntos de pruebas de decisión y las hojas pueden tener varias decisiones
- geneq: genera ecuaciones algebraicas
- genhier: genera clusters y jerarquías basadas en Cluster/2
- transform: realiza varias transformaciones en los resultados, tales como generalizaciones y especializaciones.

Otros operadores relacionados:

- genatr: genera nuevos atributos combinando algunos o mediante abstracciones
- geneve: genera ejemplos
- analyze: realiza comparaciones entre ejemplos para evaluar similaridades, relaciones de implicación, etc.
- test: prueba los resultados en los ejemplos
- visualize: herramientas de visualización

KNOWLEDGE SEEKER (Angoss) www.angoss.com
Herramientas
- Arboles de Decisión y Estadísticas

Plataformas :
- Win NT

Interface:
- ODBC

CART (Salford Systems) www.salford-systems.com
Herramientas
- Arboles de Decisión

Plataformas :
- Win NT /Unix

DATA SURVEYOR (Data Distilleries) www.datadistilleries.com
Herramientas
- Un amplio abanico de ellas.

Plataformas :
- Unix

Interface:
- ODBC

GAINSMARTS (Urban Science) www.urbanscience.com
Especializado en gráficos de ganancias
Herramientas
- Arboles de Decisión, Estadísticas Lineales y Regresiones

Plataformas :
- Unix y Win NT

Nota : Existen más herramientas, solo he indicado algunas de las más conocidas.

Indice

8.-Bibliografía

Autores mencionados en los textos y/o consultados y artículos:

Bueno, E. (1998), "El Capital Intangible como clave estratégica en la competencia actual",
Bueno, E. (1999a), "Gestión del Conocimiento, Aprendizaje y Capital Intelectual",
Bueno, E. (1999b), "¿Por qué Gestión del Conocimiento?"
Bueno, E. (2000), "La Era de la Información, del Conocimiento y del Aprendizaje"
Carbone, P (1998) "Data Mining"
Davenport, Thomas O, "Capital Humano: Creando ventajas competitivas a través de las personas"
Davenport, T.H (1998) "Successful Knowledge management projects"
Hang, J (1998) "Data Mining"
Malhotra, Yogesh Doctor "Knowledge Management, Knowledge Organizations & Knowledge Workers: A View from the Front Lines " http://www.brint.com/interview/maeil.htm
Nonaka, I (1995) "The knowledge creating company"
Saint-Onge, Hubert (2000) "Capacidad Estrategica" y "Organización en evolución"
Senge, P (1990) "Aprendizaje organizacional"
Sueiby, Phd.Karl E. "What is Knowledge Management?"

KPMG Knowlegde Management Reseach Report (2000) http://www.kpmg.es/principal.asp
Canals, Agustí (2003) http://www.uoc.edu/dt/20251/index.html
Serradell Lopez, Enric y Juan Perez, Angel A (2003) http://www.uoc.edu/dt/20133/index.html
Ortoll, Eva (2003) http://www.uoc.edu/dt/20343/index.html

Gestión del conocimiento
Portal genérico de este tema con artículos, ejemplos prácticos etc.
http://www.gestiondelconocimiento.com/
Portal Sobre Gestión Documental
http://www.ecm-spain.com/home.asp
Gestión del Capital Intelectual por José María Viedma Marti
http://www.terra.es/personal7/jm_viedma/emenuinicio.htm
Revista Robotiker
http://revista.robotiker.com/revista_articulos/gc.jsp
El Faro, Servicio Bibliotecario
http://nutabe.udea.edu.co/~elfaro/areas/gest.html
Data Mining /KDD / Data Warehouse
Buscador de monografías de múltiples temas
http://www.monografias.com/
Otro buscador:
http://www.tectimes.com/ppal.asp
Data Mining Institute, S.L.
http://www.estadistico.com/about.html
Revista SQL Server
http://www.w2000mag.com/sqlmag/atrasados/04_mar01/articulos/portada_1.htm
Novatica Revista de ATI (Asociación de Técnicos de Informática)
http://www.ati.es/novatica/1999/138/nv138sum.html
Data Mining (Portal) en Ingles
http://www.dmreview.com/portals/portal.cfm?topicId=230005
Buscador en Ingles
http://www.kdnuggets.com/websites/data-mining.html

Temas empresariales, nuevas tendencias, etc..
Buscador sobre relaciones humanas, cursos , trabajo etc.
http://www.sht.com.ar/archivo/temas/conocimiento.htm#Autor
Buscador por temas empresariales generales
http://www.gestiopolis.com/educacion/
Maping Interactivo
http://www.mappinginteractivo.com/plantilla-ante.asp?id_articulo=30#

Universidades y Centros Oficiales:

Es de la Facultad de Ciencias Exactas y Naturales de Argentina
http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MineriaDatosBressan.htm#Data Warehouse
Universidad Buenos Aires - Departamento de Computación
http://www.dc.uba.ar/academic/gen_mat.php3
Universidad Carlos III de Madrid /Departamento de informática y base de datos
http://galahad.plg.inf.uc3m.es/~scalab/
http://basesdatos.uc3m.es/
Universidad de Oviedo (Asturias)

http://webuniovi.innova.uniovi.es/
Secretaria del Consejo Superior de Informática
http://www.csi.map.es/csi/silice/Datwar.html
Servicio de Estadística de la UAB
http://einstein.uab.es/_c_serv_estadistica/cat/index.html
Uned - Universidad a distancia
http://www.uned.es/VIICongresoMetodologia/comunicaciones/actosecojueves.htm#datosI
Departamento de sistemas de información y computación - Universidad de Valencia
http://www.dsic.upv.es/~jorallo/master/
Facultad de Ciencias Contables de Lima (Perú)– Biblioteca Digital UNMSM
http://sisbib.unmsm.edu.pe/bibvirtual/publicaciones/quipukamayoc/2002/Segundo/indice.htm
CiberConta y 5Campus dependen de la universidad de Zaragoza pero es un buscador para alumnos donde puedes encontrar desde artículos, trabajos, revistas etc...
http://ciberconta.unizar.es/ ó http://www.5campus.com/ ó http://www.5campus.org/
Tecnológico de Monterrey, campus Cuernavaca- Méjico
http://www.mor.itesm.mx/principal/inicio.htm
Universidad de Murcia
http://www.um.es/fccd/anales/ad05/ad0515.pdf

Webs de Programas Informáticos:
Microsoft Server System /SQL Server 2000
http://www.microsoft.com/spain/servidores/sql/productinfo/sql2000_metas.asp
SPSS, programa data mining
http://www.spss.com/
Salford Systems
http://www.salford-systems.com/
Synera
http://www.synerasystems.com/
***Las otras encontradas aparecen en el apartado 7) dedicado a las herramientas.**

Indice