UNIVERSIDAD TECNOLÓGICA NACIONAL
FACULTAD REGIONAL MENDOZA

DEPARTAMENTO DE INGENIERÍA EN SISTEMAS DE INFORMACIÓN

CÁTEDRA DE GESTIÓN DE DATOS
3º AÑO

TRABAJO ESPECIAL

"ANÁLISIS DE AMBIENTES OPERATIVOS EN DATA MINING"

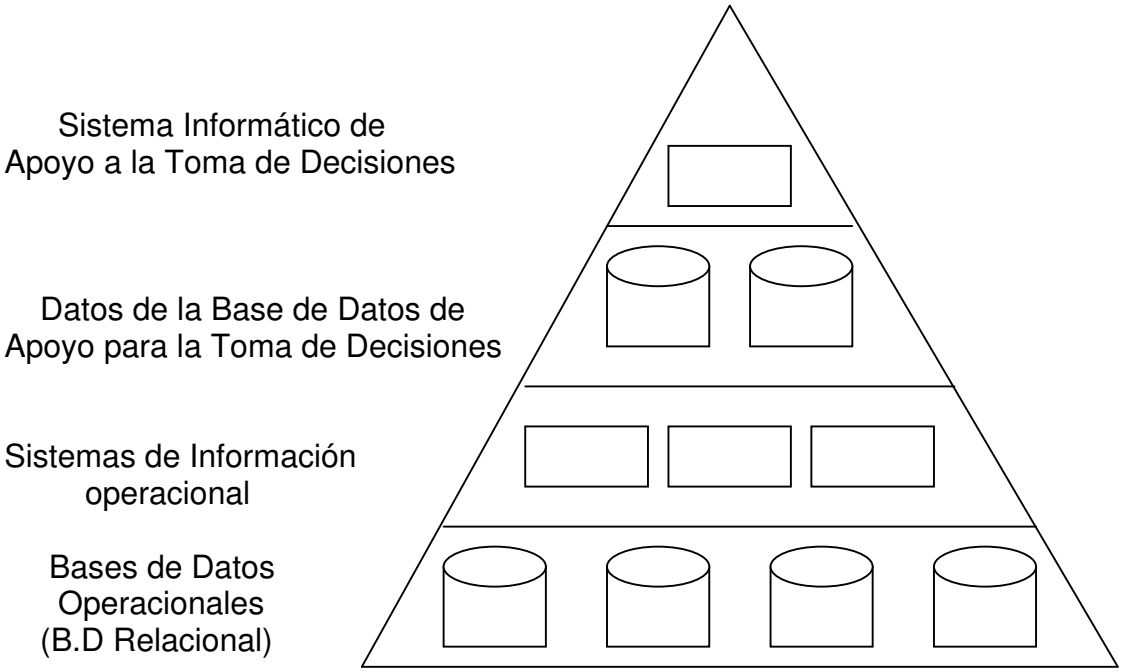
Ing. Santiago C. PÉREZ
Laura Noussan Lettry
Carlos Campos
Natalia Adaro
Carolina Pennisi
Andres Pozzi

INDICE

INTRODUCCIÓN A LOS SISTEMAS PARA APOYO A LA TOMA DE DECISIONES

La tecnología para el apoyo a la toma de decisiones no es en realidad parte de la tecnología de base de datos por sí misma. Sí es un uso, muy útil de esta tecnología que tiene sus principios en áreas de investigación de las Ciencias Administrativas y que con el tiempo se ha beneficiado con el desarrollo tecnológico de la computación y el de las bases de datos.

El objetivo consiste en recolectar datos operacionales del negocio y reducirlos a una forma útil para poder analizar el comportamiento del mismo .



PREPARACIÓN DE LOS DATOS

Como los datos deben ser *extraídos* de diversas fuentes, una vez obtenidos deben ser *limpiados, transformados y consolidados* para finalmente ser *cargados* en la base de datos de apoyo a la toma de decisiones. Posteriormente deben ser *actualizados* en forma periódica.

1) Extracción: es un proceso que consiste en obtener los datos tanto de las bases de datos operacionales como de otras fuentes, como podrían ser listas de precios de la competencia, correspondencia vía mail de los clientes, etc.

2) Limpieza**:** tiene que ver con la calidad de los datos. Al provenir de diversas fuentes este es un requisito inevitable y que puede consistir en el llenado de valores faltantes (evitar nulos), la corrección de errores tipográficos, establecimiento de abreviaturas, formatos estándares, etc.

3) Transformación y consolidación: la transformación hace referencia a llevar o transformar los datos al formato requerido para el sistema de apoyo a la toma de

decisiones. Además, cuando los datos suelen provenir de diversas fuentes, al proceso en sí se lo llama consolidación. Todavía mayor complejidad presenta la *Sincronización en el Tiempo* que significa que aquellos datos que son fechas y horas deben mantenerse con el significado que los mismos tienen para los negocios, y que además deben correlacionarse entre las distintas fuentes.

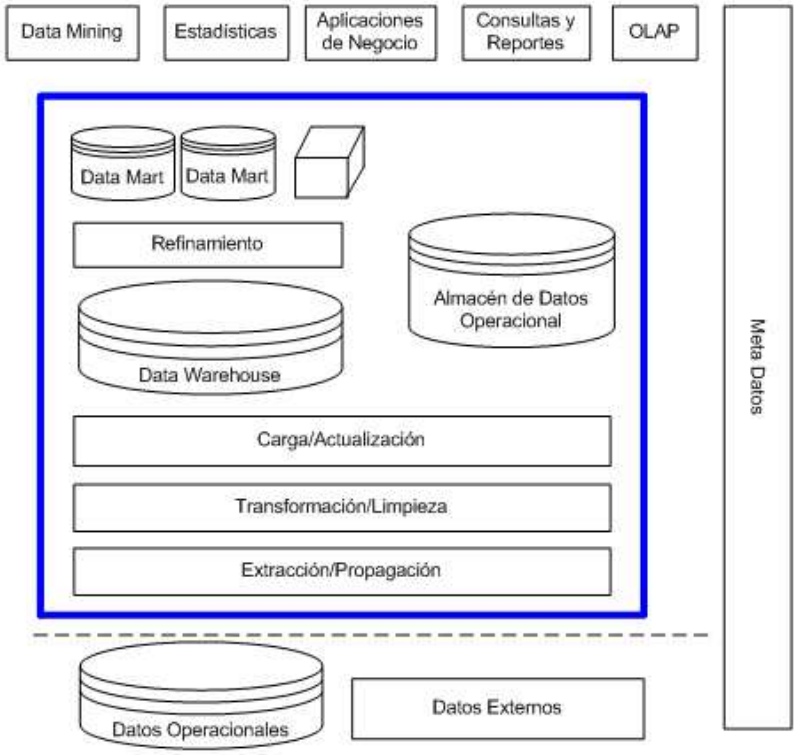
4) Carga: el proceso de carga está formado por tres pasos fundamentales:

a) el movimiento de los datos hacia la base de datos de apoyo para la toma de decisiones,
b) la verificación de su consistencia y
c) la construcción de cualquier índice necesario.

La consistencia de los datos suele ser importante para mantener la unicidad, y la construcción de índices es básicamente una cuestión de criterio que depende de la proporción del tamaño de la carga en relación a los datos que ya existen, y al tiempo de proceso que insume la carga. Así pues, cuando esta proporción es grande, suelen borrarse previamente todos los índices, y después se vuelven a crear ya que de lo contrario el proceso de carga se volvería sumamente lento, y cuando la proporción es mínima, no sería tan grave que los índices se fueran actualizando a medida que se van insertando los registros, método que suelen aplicar la mayoría de los productos comerciales.

5) Actualización: suele ser periódica para mantenerlos con relativa y razonable vigencia. Presenta todos los problemas asociados con la carga ya que generalmente consiste en una carga parcial.

ARQUITECTURA de BUSINESS INTELLIGENCE



(Arquitectura de un Data Warehouse)

La arquitectura de un businnes intelligence considera varias componentes entre las que se destaca el data warehouse. Las **fuentes de datos** pueden ser bases de datos operacionales, datos históricos y datos externos. Asimismo, pueden ser bases de datos relacionales y también pueden residir en distintas plataformas teniendo información estructurada, como bases de datos relacionales, bases de datos no relacionales, como archivos de texto, imágenes, etc.

En este sentido, sería apropiado establecer algunas precisiones conceptuales:

a) Data Warehouse:

Para IBM el Data Warehouse estaría englobando no sólo al Data Warehouse propiamente dicho, sino también a otras estructuras derivadas o paralelas al mismo, como son los Almacenes de Datos Operacionales, los Data Marts y los Cubos.

Según C. J. Date un Data Warehouse es una base de datos especial "orientada a un tema, integrado y no volátil, variante en el tiempo que soporta decisiones de administración". El término no volátil lo emplea en el sentido de que una vez que los datos han sido insertados no pueden ser cambiados, aunque sí borrados. Este almacén está separado de la base de datos operacional justamente porque los requerimientos de uno y otro sistema no son compatibles (por ejemplo los operacionales tienen procesos de carga predecibles, rendimientos estrictos mientras que los de apoyo para la toma de decisiones presentan procesos de carga no tan predecibles y los rendimiento son variables).

b) Data Marts:

Un Data Marts contiene datos provenientes del data warehouse que cumplen con requerimientos específicos para determinado sector del negocio, una determinada función o una determinada aplicación. Su objetivo es dar una mayor eficiencia ya que dependiendo de las áreas interesadas en los datos de apoyo a la toma de decisiones, suelen repetirse consultas, motivo por el cual estos almacenes de datos están limitados a un propósito específico.

La definición aportada por Date hace hincapié en que se trata de un "Almacén de datos especializado, orientado a un tema, integrado, volátil y variante en el tiempo para apoyar un subconjunto específico de decisiones de administración".

En un Data Mart la información se encuentra previamente agregada y representa una vista de los datos para el usuario final, siendo también la interfase con el data warehouse.

La principal diferencia entre un data warehouse y un data mart estriba en que este último es un almacén de datos especializado y volátil; además puede ser creado en forma independiente del data warehouse.

c) Meta Datos:

Los Meta Datos estructuran la información en el data warehouse en categorías, tópicos, grupos, herencias y mucho más. Son utilizados para proveer información sobre los datos dentro de un data warehouse.

Desde la perspectiva de un DBA los meta datos son un gran repositorio o almacén, y la documentación de todo el contenido y los procesos que contiene el data warehouse; mientras que para el usuario final son el medio de acceso hacia la información contenida en el data warehouse. Por eso mismo, pueden distinguirse dos tipos de mata datos: los técnicos y los de negocios; siendo los primeros aquellos que tienen que ver con cuestiones críticas de mantenimiento y administración, y los segundos serían el vínculo entre el data warehouse y los usuarios de negocios, que proveen a éstos el acceso a los datos en él contenidos.

d) Almacén de Datos Operacional (ODS):

Los ODS puede definirse como un conjunto actualizable e integrado de datos utilizados para la toma de decisiones tácticas. Contiene datos vivos, no instantáneas, y sólo contiene datos históricos en forma mínima, a diferencia del data warehouse.

Por lo tanto, es un almacén de tipo volátil y además el tipo de información que contienen no está agregada sino detallada; siendo el grado de detalle variable según el problema que se intente resolver mediante el ODS.

e) On-Line Analitical Processing (OLAP):

Un data warehouse almacena información táctica que responde a las preguntas de quién y qué sobre eventos del pasado (recordar que almacena información de tipo histórica y por eso mismo no es volátil). Un sistema OLAP, al contrario que un data warehouse, utiliza vistas multidimensionales (generalmente los data warehouses se basan en la tecnología relacional) de datos agregados para proveer un rápido acceso a información de tipo estratégica para su posterior análisis.

En realidad los data warehouses y los sistemas OLAP son comoplementarios en el sentido de que el primero almacena y administra datos, y el segundo transforma los datos del data warehouse en información de tipo estratégica. La diferencia fundamental radica en que los sistemas OLAP tienen que permitir vistas multidimensionales de los datos, capacidad para establecer relaciones complejas entre los datos y una rápida respuesta a los cálculos que suelen ser muy intensivos, por lo que el tiempo de respuesta es crucial.

En las organizaciones de negocios es común que se utilice el sistema OLAP dependiendo de cada sector o función. Así el departamento de finanzas suele utilizarlo para sus aplicaciones de presupuestos, costeo y análisis financiero; el departamento de ventas para análisis de ventas y promociones; el departamento de marketing para el análisis de mercadeo, promociones de ventas, análisis de clientes y segmentación de mercado/clientes. También el departamento de producción suele utilizar esta herramienta en áreas como la planificación de la producción y el control de calidad.

f) Estadísticas:

Las estadísticas generalmente son utilizadas para sumarizar información contenida en la base de datos. Se utilizan habitualmente los siguientes métodos estadísticos: análisis de correlación, de factores y de regresión que intentan responder preguntas como ¿Cuáles son las aparentes relaciones de dependencia entre las variables y los datos? ¿cuál es la probabilidad de que un evento en particular vuelva a ocurrir?, o bien ¿cuáles son los patrones de datos significativos?

g) La minería de datos:

Analiza, a diferencia de las estadísticas, todos los datos relevantes de la base de datos y extrae patrones que hasta ese momento eran desconocidos. En cierto modo la minería de datos se basa en las estadísticas pero también extiende las técnicas y disciplinas utilizadas en los análisis estadísticos habituales.

DATA MINING (MINERIA DE DATOS)

¿Qué es Data Mining?

Data Mining, es la extracción de información oculta y predecible de grandes bases de datos. Es una poderosa tecnología nueva con gran potencial que ayuda a las compañías a concentrarse en la información más importante de sus Bases de Información (Data Warehouse).

Un Sistema Data mining es una tecnología de soporte para usuario final cuyo objetivo es extraer conocimiento útil y utilizable a partir de la información contenida en las bases de datos de las empresas. Es decir que su objetivo es descubrir cosas sobre el negocio o empresa desde los datos que almacena la empresa.

¿Cómo se desarrollan los sistemas Data Mining?

Los sistemas Data Mining se desarrollan bajo lenguajes de última generación basados en la inteligencia artificial y utilizan modelos matemáticos tales como:

- Redes neuronales artificiales: modelos predecible no-lineales que aprenden a través del entrenamiento y semejan la estructura de una red neuronal biológica.
- Árboles de decisión: estructuras de forma de árbol que representan conjuntos de decisiones. Estas decisiones generan reglas para la clasificación de un conjunto de datos. Métodos específicos de árboles de decisión incluyen Árboles de Clasificación y Regresión (CART: Classification And Regression Tree) y Detección de Interacción Automática de Chi Cuadrado (CHAI: Chi Square Automatic Interaction Detection)
- Algoritmos genéticos: técnicas de optimización que usan procesos tales como combinaciones genéticas, mutaciones y selección natural en un diseño basado en los conceptos de evolución.
- Método del vecino más cercano: una técnica que clasifica cada registro en un conjunto de datos basado en una combinación de las clases del de los k registro (s) más similares a él en un conjunto de datos históricos (donde k ? 1). Algunas veces se llama la técnica del vecino k-más cercano.
- Regla de inducción: la extracción de reglas if-then de datos basados en significado estadístico.

Muchas de estas tecnologías han estado en uso por más de una década en herramientas de análisis especializadas que trabajan con volúmenes de datos relativamente pequeños. Estas capacidades están ahora evolucionando para integrarse directamente con herramientas OLAP y de Data Warehousing.

¿Qué son capaces de hacer las herramientas del Data Mining?

Las herramientas de Data Mining predicen futuras tendencias y comportamientos, permitiendo en los negocios tomar decisiones proactivas y conducidas por un conocimiento acabado de la información (knowledge-driven). Los análisis prospectivos automatizados ofrecidos por un producto así van más allá de los eventos pasados provistos por herramientas retrospectivas típicas de sistemas de soporte de decisión.

Las herramientas de Data Mining pueden responder a preguntas de negocios que tradicionalmente consumen demasiado tiempo para poder ser resueltas y a los cuales los usuarios de esta información casi no están dispuestos a aceptar. Estas herramientas exploran las bases de datos en busca de patrones ocultos, encontrando información predecible que un experto no puede llegar a encontrar porque se encuentra fuera de sus expectativas.

Una vez que las herramientas de Data Mining fueron implementadas en computadoras cliente servidor de alto performance o de procesamiento paralelo, pueden analizar bases de datos masivas para brindar respuesta a preguntas tales como, "¿Cuáles clientes tienen más probabilidad de responder al próximo mailing promocional, y por qué? y presentar los resultados en formas de tablas, con gráficos, reportes, texto, hipertexto, etc.

Las técnicas de Data Mining pueden ser implementadas rápidamente en plataformas ya existentes de software y hardware, para acrecentar el valor de las fuentes de información existentes, y pueden ser integradas con nuevos productos y sistemas pues son traídas en línea (on-line).

El Alcance del Data Mining

Dadas bases de datos de suficiente tamaño y calidad, la tecnología de Data Mining puede generar nuevas oportunidades de negocios al proveer estas capacidades:

- Predicción automatizada de tendencias y comportamientos.
- Automatización del proceso para encontrar información predecible en grandes bases de datos.
- Preguntas que tradicionalmente requerían un intenso análisis manual, ahora pueden ser contestadas directa y rápidamente desde los datos. Un típico ejemplo de problema predecible es el marketing apuntado a objetivos (targeted marketing).
- Data Mining usa datos en mailing promocionales anteriores para identificar posibles objetivos para maximizar los resultados de la inversión en futuros mailing.
- Otros problemas predecibles incluyen pronósticos de problemas financieros futuros y otras formas de incumplimiento, e identificar segmentos de población que probablemente respondan similarmente a eventos dados.
- Descubrimiento automatizado de modelos previamente desconocidos.
- Las herramientas de Data Mining barren las bases de datos e identifican modelos previamente escondidos en un sólo paso.

- Otros problemas de descubrimiento de modelos incluye detectar transacciones fraudulentas de tarjetas de créditos e identificar datos anormales que pueden representar errores de tipeado en la carga de datos.
- las técnicas de Data Mining pueden redituar los beneficios de automatización en las plataformas de hardware y software existentes y puede ser implementadas en sistemas nuevos a medida que las plataformas existentes se actualicen y nuevos productos sean desarrollados.
- Cuando las herramientas de Data Mining son implementadas en sistemas de procesamiento paralelo de alto performance, pueden analizar bases de datos masivas en minutos. Procesamiento más rápido significa que los usuarios pueden automáticamente experimentar con más modelos para entender datos complejos. La alta velocidad hace que sea práctico para los usuarios analizar inmensas cantidades de datos. Grandes bases de datos, a su vez, producen mejores predicciones.
- Las bases de datos pueden ser grandes tanto en profundidad como en ancho:

  Más columnas. Los analistas muchas veces deben limitar el número de variables a examinar cuando realizan análisis manuales debido a limitaciones de tiempo. Sin embargo, variables que son descartadas porque parecen sin importancia pueden proveer información acerca de modelos desconocidos. Un Data Mining de alto rendimiento permite a los usuarios explorar toda la base de datos, sin preseleccionar un subconjunto de variables.

  Más filas. Muestras mayores producen menos errores de estimación y desvíos, y permite a los usuarios hacer inferencias acerca de pequeños pero importantes segmentos de población.

¿Cómo Trabaja el Data Mining?

¿Cuán exactamente es capaz Data Mining de decirle cosas importantes que usted desconoce o que van a pasar? La técnica usada para realizar estas hazañas en Data Mining se llama Modelado. Modelado es simplemente el acto de construir un modelo en una situación donde usted conoce la respuesta y luego la aplica en otra situación de la cual desconoce la respuesta. Por ejemplo, si busca un galeón español hundido en los mares lo primero que podría hacer es investigar otros tesoros españoles que ya fueron encontrados en el pasado. Notaría que esos barcos frecuentemente fueron encontrados fuera de las costas de Bermuda y que hay ciertas características respecto de las corrientes oceánicas y ciertas rutas que probablemente tomara el capitán del barco en esa época. Usted nota esas similitudes y arma un modelo que incluye las características comunes a todos los sitios de estos tesoros hundidos. Con estos modelos en mano sale a buscar el tesoro donde el modelo indica que en el pasado hubo más probabilidad de darse una situación similar. Con un poco de esperanza, si tiene un buen modelo, probablemente encontrará el tesoro.

Este acto de construcción de un modelo es algo que la gente ha estado haciendo desde hace mucho tiempo, seguramente desde antes del auge de las computadoras y de la tecnología de Data Mining. Lo que ocurre en las computadoras, no es muy diferente de la manera en que la gente construye modelos. Las computadoras son cargadas con mucha información acerca de una variedad de situaciones donde una respuesta es conocida y luego el software de Data Mining en la computadora debe

correr a través de los datos y distinguir las características de los datos que llevarán al modelo. Una vez que el modelo se construyó, puede ser usado en situaciones similares donde usted no conoce la respuesta.

Si alguien le dice que tiene un modelo que puede predecir el uso de los clientes, ¿Cómo puede saber si es realmente un buen modelo? La primera cosa que puede probar es pedirle que aplique el modelo a su base de clientes - donde usted ya conoce la respuesta. Con Data Mining, la mejor manera para realizar esto es dejando de lado ciertos datos para aislarlos del proceso de Data Mining. Una vez que el proceso está completo, los resultados pueden ser testeados contra los datos excluidos para confirmar la validez del modelo. Si el modelo funciona, las observaciones deben mantenerse para los datos excluidos.

Arquitectura para Data Mining

Para aplicar mejor estas técnicas avanzadas, éstas deben estar totalmente integradas con el data warehouse así como con herramientas flexibles e interactivas para el análisis de negocios. Varias herramientas de Data Mining actualmente operan fuera del warehouse, requiriendo pasos extra para extraer, importar y analizar los datos.

Además, cuando nuevos conceptos requieren implementación operacional, la integración con el warehouse simplifica la aplicación de los resultados desde Data Mining. El Data warehouse analítico resultante puede ser aplicado para mejorar procesos de negocios en toda la organización, en áreas tales como manejo de campañas promocionales, detección de fraudes, lanzamiento de nuevos productos, etc.

El punto de inicio ideal es un data warehouse que contenga una combinación de datos de seguimiento interno de todos los clientes junto con datos externos de mercado acerca de la actividad de los competidores. Información histórica sobre potenciales clientes también provee una excelente base para prospecting. Este warehouse puede ser implementado en una variedad de sistemas de bases relacionales y debe ser optimizado para un acceso a los datos flexible y rápido.

Un server multidimensional OLAP permite que un modelo de negocios más sofisticado pueda ser aplicado cuando se navega por el data warehouse. Las estructuras multidimensionales permiten que el usuario analice los datos de acuerdo a como quiera mirar el negocio - resumido por línea de producto, u otras perspectivas claves para su negocio. El server de Data Mining debe estar integrado con el data warehouse y el server OLAP para insertar el análisis de negocios directamente en esta infraestructura.

Un avanzado, metadata centrado en procesos define los objetivos del Data Mining para resultados específicos tales como manejos de campaña, prospecting, y optimización de promociones. La integración con el data warehouse permite que decisiones operacionales sean implementadas directamente y monitoreadas. A medida que el data warehouse crece con nuevas decisiones y resultados, la organización puede "minar" las mejores prácticas y aplicarlas en futuras decisiones.

Este diseño representa una transferencia fundamental desde los sistemas de soporte de decisión convencionales. Más que simplemente proveer datos a los usuarios

finales a través de software de consultas y reportes, el server de Análisis Avanzado aplica los modelos de negocios del usuario directamente al warehouse y devuelve un análisis proactivo de la información más relevante. Estos resultados mejoran los metadatos en el server OLAP proveyendo una estrato de metadatos que representa una vista fraccionada de los datos. Generadores de reportes, visualizadores y otras herramientas de análisis pueden ser aplicadas para planificar futuras acciones y confirmar el impacto de esos planes.

¿Por qué usar Data Mining?

Sin duda alguna que el uso de Data Mining:

- Contribuye a la toma de decisiones tácticas y estratégicas proporcionando un sentido automatizado para identificar información clave desde volúmenes de datos generados por procesos tradicionales y de e-Business.
- Permite a los usuarios dar prioridad a decisiones y acciones mostrando factores que tienen un mayor en un objetivo, qué segmentos de clientes son desechables y qué unidades de negocio son sobrepasados y por qué.
- Proporciona poderes de decisión a los usuarios del negocio que mejor entienden el problema y el entorno y es capaz de medir la acciones y los resultados de la mejor forma.
- Genera Modelos descriptivos : En un contexto de objetivos definidos en los negocios permite a empresas, sin tener en cuenta la industria o el tamaño, explorar automáticamente, visualizar y comprender los datos e identificar patrones, relaciones y dependencias que impactan en los resultados finales de la cuenta de resultados (tales como el aumento de los ingresos, incremento de los beneficios, contención de costes y gestión de riesgos)
- Genera Modelos predictivos: permite que relaciones no descubiertas e identificadas a través del proceso del Data Mining sean expresadas como reglas de negocio o modelos predictivos. Estos outputs pueden comunicarse en formatos tradicionales (presentaciones, informes, información electrónica compartida, embebidos en aplicaciones,...) para guiar la estrategia y planificación de la empresa.

Técnicas de Minería de Datos

Básicamente existen dos grandes grupos de técnicas:

- Descubrimiento
- Predicción

1) Minería de datos de descubrimiento:

Estas técnicas intentan encontrar patrones en los datos sin un previo conocimiento de los patrones existentes. Seguidamente, describimos brevemente cada una de ellas:

a) Segmentación: Esta técnica agrupa a una familia de técnicas que busca agrupar registros de datos según su similitud. Por ejemplo, considerando una tabla de clientes, lo que haría la segmentación sería agrupar a los clientes según sus características, apuntando a que cada segmento sea lo más disímil de los otros.

b) Asociación: Bajo esta categoría encontramos a una familia de técnicas que establecen asociaciones entre los registros de datos. La técnica más conocida es la del

Análisis de la Canasta de Mercado, según la cual los registros de datos son todos los productos comprados por un cliente durante la misma transacción y por lo tanto la técnica en sí permite descubrir combinaciones de ítems que han sido adquiridos por distintos clientes y por asociación se puede establecer qué productos son habitualmente adquiridos con qué otros.

c) Secuenciación: Estas técnicas se aplican a registros de datos ordenados en el tiempo o bien cuando cada registro puede considerase ordenado. Intentan descubrir secuencias o subsecuencias similares en los datos ordenados.

2) Minería de Datos Predictiva:

Estas técnicas intentan encontrar relaciones entre una variable específica y las otras variables existentes en los datos. Seguidamente describimos brevemente cada una de las técnicas involucradas:

a) Clasificación: Asigna los registros de datos dentro de categorías previamente definidas; por ejemplo categorizando clientes en segmentos de mercado. Entre las técnicas de clasificación más utilizadas pueden mencionarse, a los árboles de decisión, funciones neuronales, etc.

b) Predicción de valor: Se utiliza para predecir el valor de una variable continua a partir de otras variables presentes en los registros de datos.

Conclusiones

Un Sistema Data mining nos permite analizar factores de influencia en determinados procesos, predecir o estimar variables o comportamientos futuros, segmentar o agrupar ítems similares, además de obtener secuencias de eventos que provocan comportamientos específicos.

La llegada del Data Mining se considera como la última etapa de la introducción de métodos cuantitativos, científicos en el mundo del comercio, industria y negocios. Desde ahora, todos los no-estadísticos -es decir el 99,5% de nosotros - pueden construir modelos exactos de algunas de sus actividades, para estudiarlas mejor, comprenderlas y mejorarlas.

IBM DB2 Intelligent Miner

Estos productos dan soporte a una rápida habilitación de las posibilidades de análisis de Intelligent Miner que incorporan Business Intelligence (BI), Commerce o programas de aplicación de OLTP tradicionales.

La realización del modelado y su aplicación incluye:

- IBM DB2 Intelligent Miner Modeling V8.1 se denomina IM Modeling.
- IBM DB2 Intelligent Miner Scoring V8.1 se denomina IM Scoring.
- IBM DB2 Intelligent Miner Visualization V8.1 se denomina IM Visualization.

Con IM Modeling se pueden crear modelos de minería de datos.

Mediante IM Scoring pueden desplegarse modelos de PMML (Lenguaje de Código de Modelo Predictivo Bazado en XML) creados por uno de los productos de Intelligent Miner o por otras aplicaciones o herramientas que dan soporte al funcionamiento conjunto por medio de modelos de PMML.

Puede utilizar IM Visualization para examinar modelos de PMML creados por uno de los productos de Intelligent Miner o por otras aplicaciones o herramientas que dan soporte al funcionamiento conjunto por medio de modelos de PMML.

PMML es un formato estándar para los modelos de minería de datos. Basado en XML, PMML proporciona un estándar que permite que las aplicaciones de diferentes proveedores compartan los modelos de minería de datos. La intención es la de proporcionar un método de definir modelos que sea independiente del proveedor. De este modo, las tecnologías exclusivas de un producto o fabricante y las incompatibilidades dejan de ser una barrera para el intercambio de modelos entre aplicaciones.

IM Modeling

IBM DB2 Intelligent Miner Modeling V8.1 es una interfaz de programación de aplicaciones de SQL para DB2. Está formada por un conjunto de objetos de base de datos que le permite crear modelos de minería de datos a partir de información contenida en bases de datos DB2 de IBM. La interfaz de SQL para DB2 da soporte a la rápida creación de aplicaciones de modelado de minería de datos.

Funciones de minería de datos

IM Modeling le permite crear modelos de minería para tres funciones de minería de datos. Estas funciones de minería son:

- Asociación
- Clasificación
- Agrupación

Naturalmente, la decisión respecto a la función de minería que ha de utilizarse para un determinado problema empresarial es una decisión muy importante.

Asociación

El objeto de la función de minería de Asociación es encontrar elementos que estén asociados entre sí de una forma significativa. Por ejemplo, puede analizar las transacciones de compra para descubrir combinaciones de productos que a menudo se compran juntos. La función de minería de Asociación intenta determinar qué producto o productos es probable que estén presentes en una transacción si en ésta se dan determinados productos.

Las relaciones descubiertas por esta función de minería de Asociación se expresan en forma de normas de asociación. En una aplicación comercial típica, la función de minería descubre asociaciones y también asigna probabilidades.

Por ejemplo, puede descubrir que si un cliente compra pintura, existe una probabilidad del 20% de que también compre una brocha de pintar. También descubre asociaciones múltiples; por ejemplo, si un cliente compra pintura y brochas de pintar, existe una probabilidad del 40% de que también compre disolvente de pintura.

Cuando el usuario analiza las normas de asociación, debe interpretarlas y decidir si son:

- Relaciones fortuitas. Por ejemplo, dos productos estaban a la venta a la mitad de precio el mismo día, y por tanto se creó una correlación aleatoria.
- Relaciones conocidas. Por ejemplo, la correlación de la pintura y las brochas es una asociación ya conocida.
- Relaciones no conocidas, pero de escasa importancia. Por ejemplo, una correlación entre pintura roja brillante y pintura roja mate puede ser desconocida, pero sin importancia.
- Relaciones desconocidas e importantes. Por ejemplo, una correlación entre pintura y pelotas de baloncesto puede ser desconocida hasta ese momento.

Puede también ser muy útil en la organización de la publicidad y en la disposición de los productos dentro de la tienda.

Aplicaciones: El descubrimiento de normas de Asociación se utiliza en el análisis de la cesta de la compra, la planificación de la ubicación de los productos y en la planificación de ventas promocionales, entre otras muchas aplicaciones.

Clasificación

La clasificación es el proceso de crear automáticamente un modelo de clases a partir de un conjunto de registros que contienen etiquetas de clases. La función de minería de Clasificación analiza registros que ya se sabe que pertenecen a clases determinadas. Luego crea perfiles para miembros de cada clase a partir de las características comunes de los registros. Puede utilizar una herramienta de minería de datos para aplicar el modelo a nuevos registros, es decir, registros que no se han clasificado todavía. Esto le permite predecir la clase a la que pertenecen los nuevos registros.

La función de minería de Clasificación se utiliza a menudo en CRM (Gestión de Relación con el Cliente). Algunas de las aplicaciones comerciales de esta función de minería son la calificación del riesgo crediticio, la clasificación de clientes para campañas de correo dirigidas, la predicción del riesgo crediticio en nuevos clientes y la predicción del desgaste.

IM Modeling utiliza el algoritmo de Inducción en árbol de la función de minería de IM Modeling le permite crear un modelo de minería y también probar su exactitud.
Ejemplo: En los datos de que dispone sobre sus clientes, una compañía de seguros tiene datos sobre los clientes que han dejado que su seguro caduque por falta de pago

y sobre los que no lo han hecho. ¿Cómo puede la compañía hacer el mejor uso de esta información para identificar en el futuro posibles asegurados que dejan de pagar?

Los asegurados que dejan de pagar ya pertenecen a una clase determinada: se les clasificados como clientes que dejan que su seguro caduque. La compañía puede utilizar la función de minería de Clasificación para crear un perfil de grupo de riesgo en forma de modelo de minería de datos. Este perfil o modelo contiene los atributos comunes de los clientes que han dejado caducar sus seguros, en comparación con los demás clientes. La compañía puede luego aplicar este perfil a los nuevos clientes (todavía 'sin clasificar') para determinar si pertenecen a este grupo de riesgo.

Puede utilizar la función de minería de Clasificación para hacer lo siguiente, por ejemplo:

- Aprobar o denegar peticiones de seguro
- Detectar fraudes de tarjeta de crédito
- Identificar defectos en imágenes de componentes manufacturados
- Diagnosticar condiciones de error

Aplicaciones: Son la venta dirigida, la diagnosis médica, la efectividad del tratamiento médico, la reposición de inventarios y la planificación de la ubicación de las tiendas.

Agrupación

La función de minería de Agrupación consta de un conjunto de algoritmos que agrupan los registros de datos de acuerdo con su grado de similitud. Por ejemplo, un registro de datos puede incluir información sobre un cliente. En este caso, la Agrupación agruparía los clientes similares, con lo que al mismo tiempo se resaltan las diferencias entre los diferentes grupos de clientes formados de esta manera.

Los grupos formados de esta manera se denominan agrupaciones. Cada agrupación proporciona información específica referente a la identidad o comportamiento de los clientes, de su trasfondo demográfico o de sus productos o combinaciones de productos preferidos. De esta forma, los clientes que son similares se incluyen en grupos homogéneos para fines comerciales u otros procesos empresariales.

Cada algoritmo de la función de minería de Agrupación utiliza un método propio para formar agrupaciones con los datos. El algoritmo utilizado en IM Modeling es la Agrupación demográfica. Este sería un posible ejemplo de su utilización:

Ejemplo: Un banco ofrece a sus clientes un determinado tipo de cuenta (cuenta X). Sin embargo, la aceptación de la cuenta es inferior a la esperada por el banco. ¿Cómo puede el banco aumentar el número de sus clientes que eligen la cuenta X sin llevar a cabo una campaña de marketing a gran escala?

El banco puede utilizar la Agrupación demográfica para crear primero un modelo de minería de datos de todos sus clientes. Este modelo contendrá una o más agrupaciones de clientes que utilizan la cuenta X, en la que cada agrupación tendrá un perfil demográfico específico. A continuación el banco puede apuntar de modo selectivo

a los clientes de otras agrupaciones que tengan perfiles demográficos similares a los que tienen la cuenta X pero que en cambio no tienen este tipo de cuenta.

Aplicaciones: La información proporcionada por esta función de minería permite a las empresas ofrecer servicios y productos específicos, personalizados, a sus clientes. En el entorno empresarial, la Agrupación se utiliza en los sectores siguientes:

- Comercialización cruzada
- Venta cruzada
- Planes de comercialización personalizados para diferentes tipos de clientes
- Toma de decisiones sobre el tipo de medio a utilizar
- Análisis de los objetivos de compra

IM Scoring

Proporciona técnicas de evaluación en forma de extensores de base de datos, extensores DB2 y cartuchos Oracle. Permite a los programas de aplicación aplicar modelos PMML a bases de datos grandes, subconjuntos de bases de datos o a filas individuales (casos). Los programas de aplicación utilizan la API de SQL, que consta de funciones definidas por el usuario (las UDF) y métodos definidos por el usuario (los UDM), para realizar la operación de evaluación (scoring).

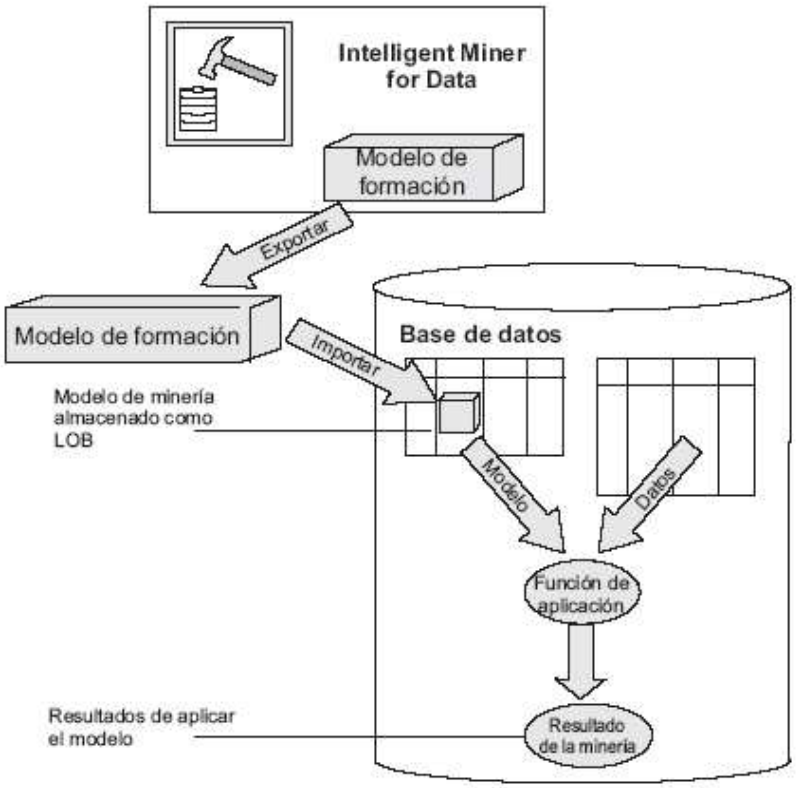
La tabla siguiente muestra los diversos modelos PMML que se pueden aplicar mediante diversas funciones de minería.

Tipo de modelo PMML	Algoritmo de minería
Técnica de agrupación basada en un centro	Algoritmo de agrupación neuronal
Técnica de agrupación basada en la distribución	Algoritmo de agrupación demográfica
Redes neuronales	Algoritmo de clasificación neuronal, algoritmo de predicción neuronal
Árbol de decisión	Algoritmo de clasificación en árbol
Regresión	Algoritmo de regresión logística, algoritmo de regresión polinómica, algoritmo de regresión lineal

Funciones de minería soportadas por IM Scoring

IM Scoring da soporte a la modalidad de aplicación para las siguientes funciones estadísticas y de minería de **IM for Data**:

- Agrupación demográfica y neuronal
- Clasificación en árbol y neuronal
- Predicción RBF y neuronal
- Regresión polinómica



IM Visualization

Proporciona los visualizadores Java siguientes para presentar los resultados de modelado de datos para su análisis:

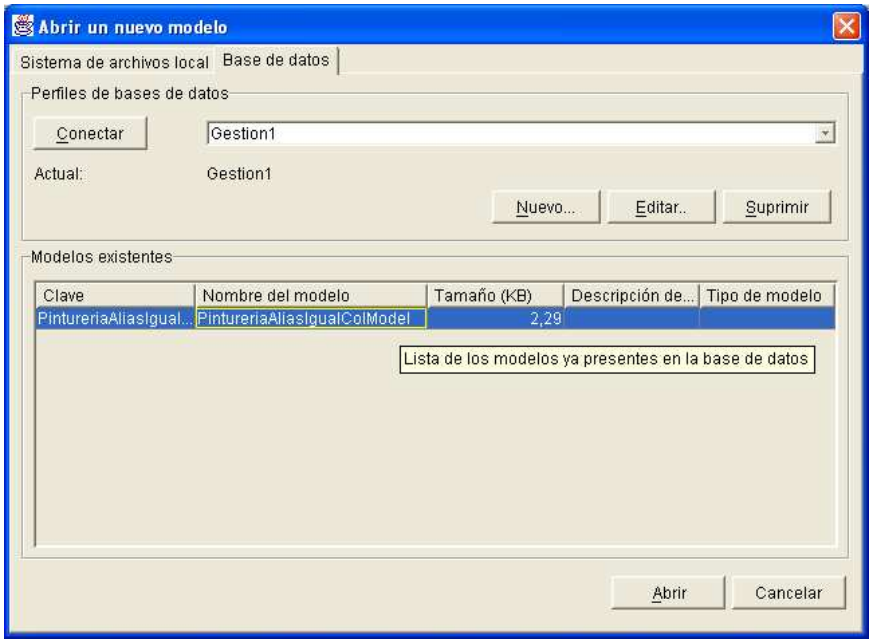
- Visualizador de asociaciones
- Visualizador de clasificación
- Visualizador de clústeres (agrupación)

Puede utilizar Intelligent Miner Visualization para visualizar modelos de minería que se ajustan a PMML. Las aplicaciones pueden llamar a estos visualizadores para presentar los resultados de modelos o los visualizadores pueden desplegarse como applets en un navegador Web para su diseminación inmediata.

Cada visualizador puede mostrar el mismo modelo en vistas diferentes. Cada una de estas vistas contiene información que no está disponible o que es difícil representar en otra vista. Las vistas están sincronizadas. Esto significa que cuando se ocultan ítems en una vista, dichos ítems también están ocultos en las demás vistas.

Ejemplo Práctico: Visualizador de Asociación

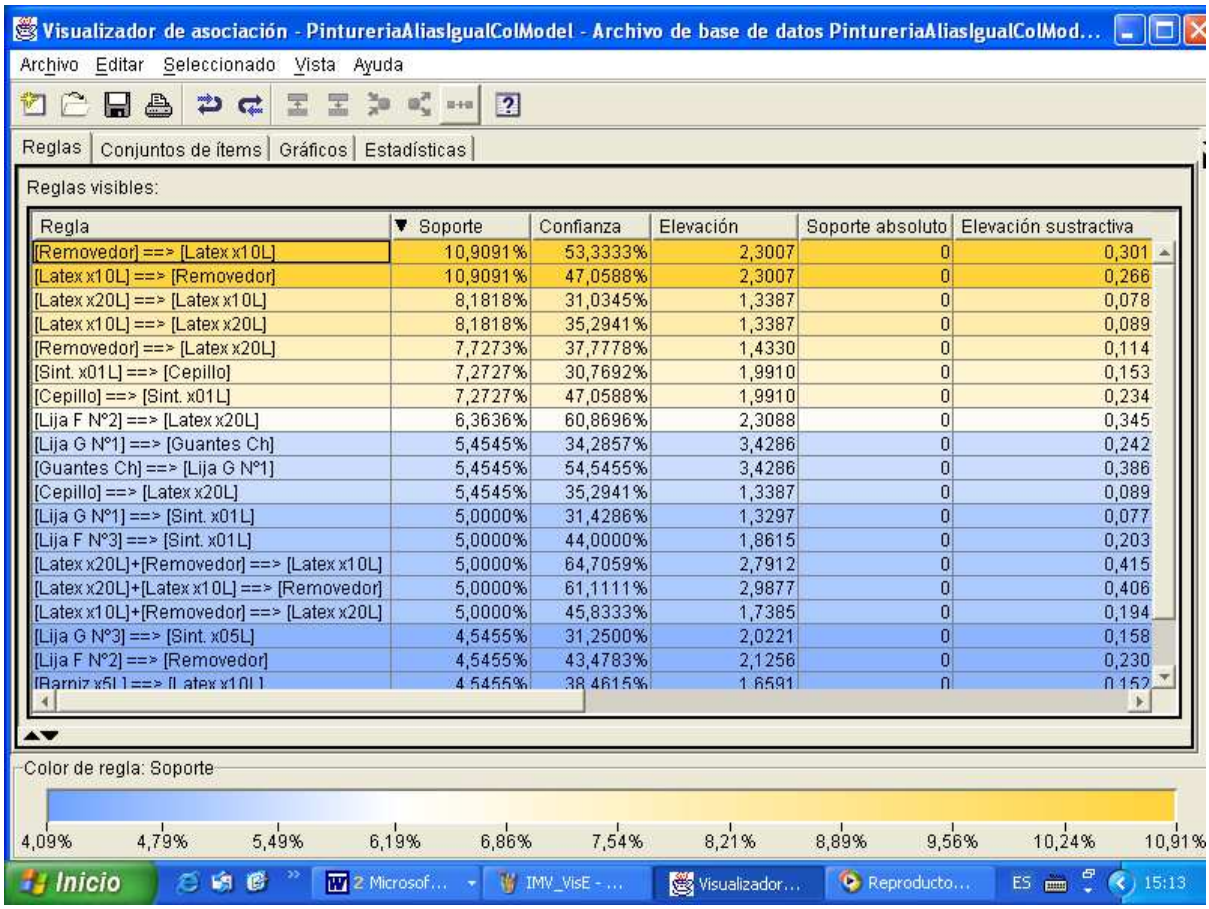
El modelo de asociación que seguidamente le presentamos se basa en un ejemplo práctico desarrollado por la Cátedra partiendo de la creación del modelo de minería de datos utilizando para ello I.M. Modeling, como puede apreciarse en la siguiente imagen.



Vista Reglas

Muestra las reglas de asociaciones y los conjuntos de ítems con varios valores de campos, tales como el soporte o la confianza.

Pueden mostrarse las reglas tanto en forma textual como tabular.



Asimismo el usuario puede establecer no sólo los colores sino también la ubicación de los valores de campo, etc.

Una **norma de asociación** consta de:

- Dos conjuntos afines de elementos: el cuerpo de la norma y la cabecera de la norma
- El **soporte** de la norma, que es un valor estadístico en forma de porcentaje
- La **fiabilidad** de la norma, que es asimismo un valor estadístico en forma de porcentaje

Por ejemplo, del modelo de la pinturería puede apreciarse que:

- [Látex x20L][Removedor] Þ [Látex x10L]
- Soporte = 5 %
- Fiabilidad = 64,7 %

En este caso:

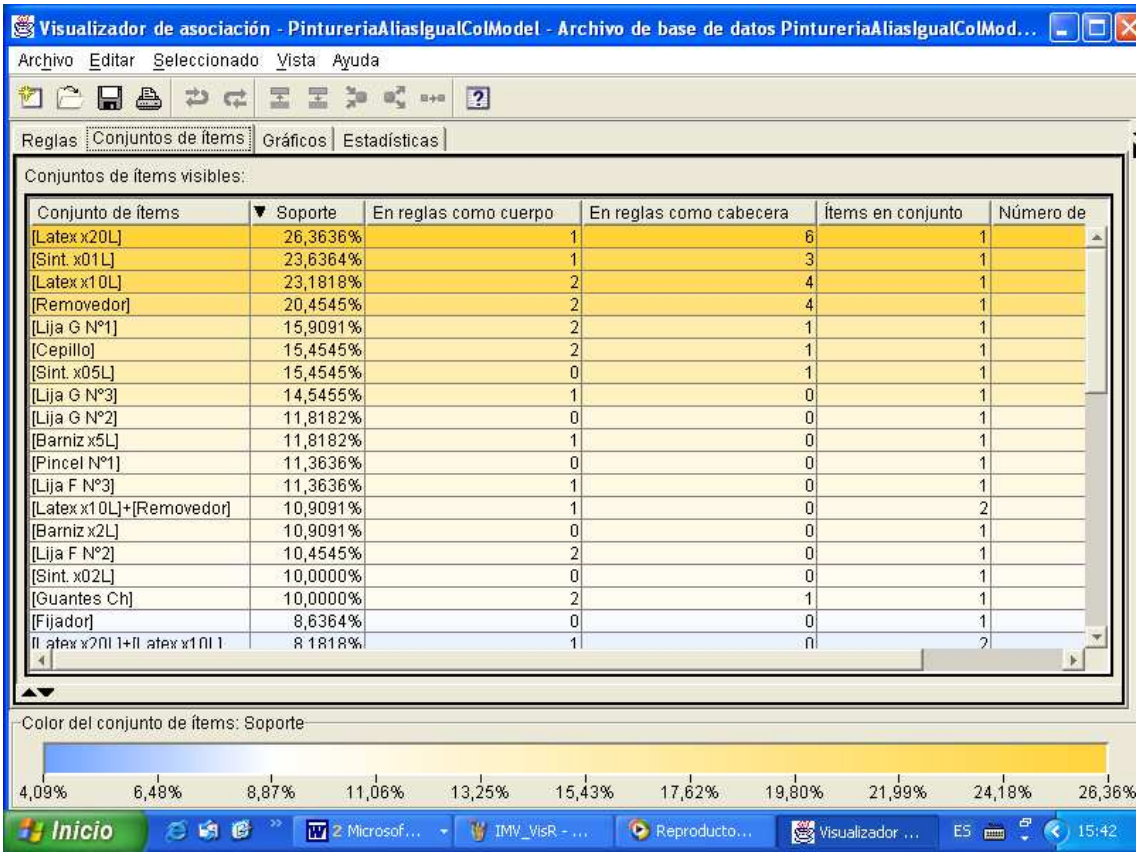
- [Látex x20L][Removedor] es el **cuerpo de la norma**
- [Látex x10L] es la **cabecera de la norma**

El conjunto de elementos [Látex x20L][Removedor][Látex x10L] estaba presente en un 5% de las transacciones de compra consideradas. Este es el **valor de soporte.**

En las transacciones donde aparecían juntos los elementos [Látex x20L][Removedor], también estaba presente el elemento [Látex x10L] en un 64,7% de los casos. Este es el **valor de fiabilidad**.

Vista Conjuntos de ítems

Muestra los conjuntos de ítems que se incluyen en una regla de asociación.



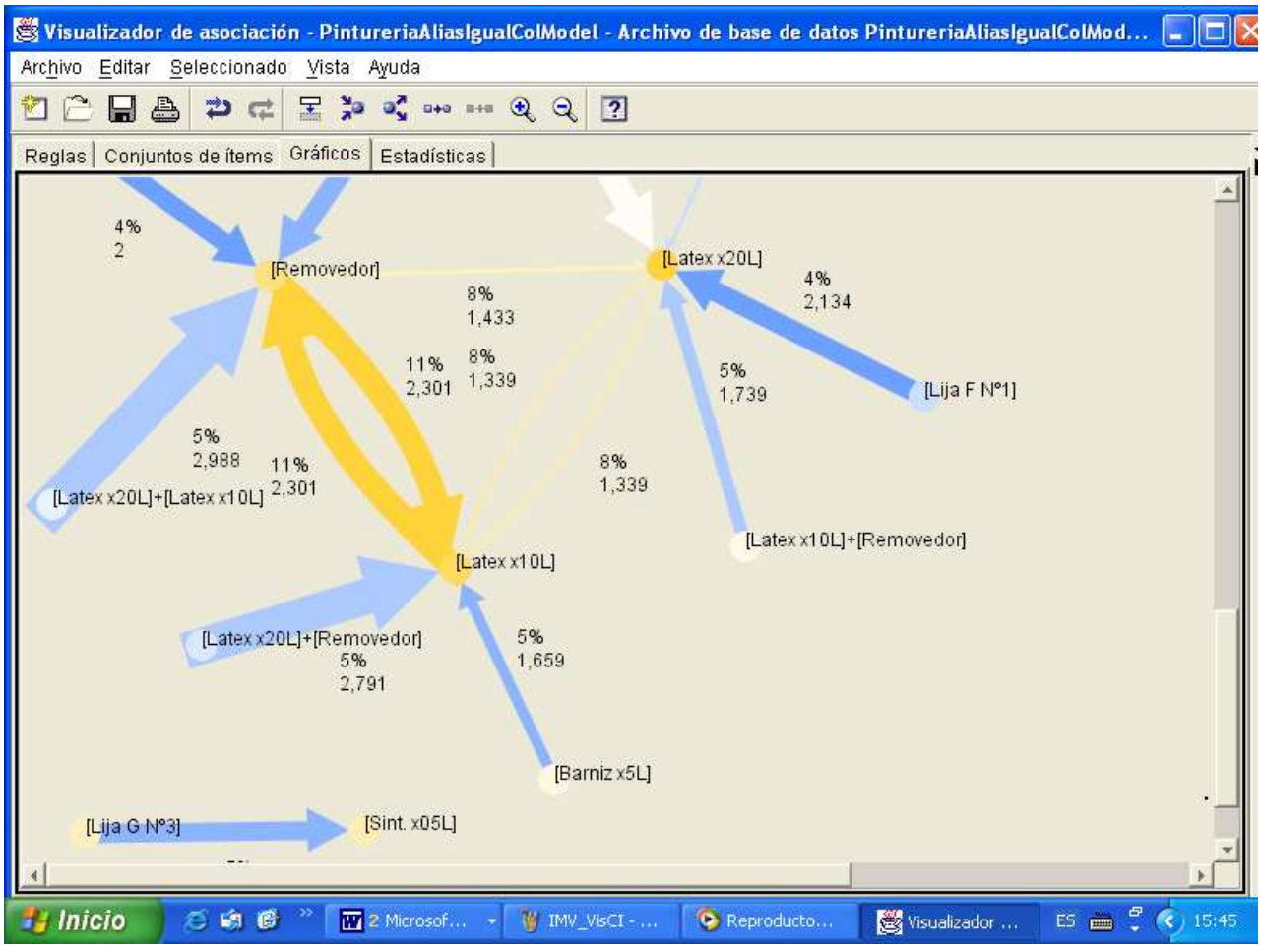
Puede apreciarse la siguiente información:

- Conjunto de ítems
- Soporte
- En reglas como cuerpo

Vista Gráficos

Los conjuntos de ítems se visualizan como nodos y las reglas de asociaciones como flechas. Las flechas conducen **desde** los conjuntos de ítems del **cuerpo** de la regla **a los conjuntos de la cabecera** de la regla.

El color de los nodos y el color de las flechas representa el valor de un parámetro en particular como, por ejemplo, “Soporte” o “En reglas como cuerpo”.

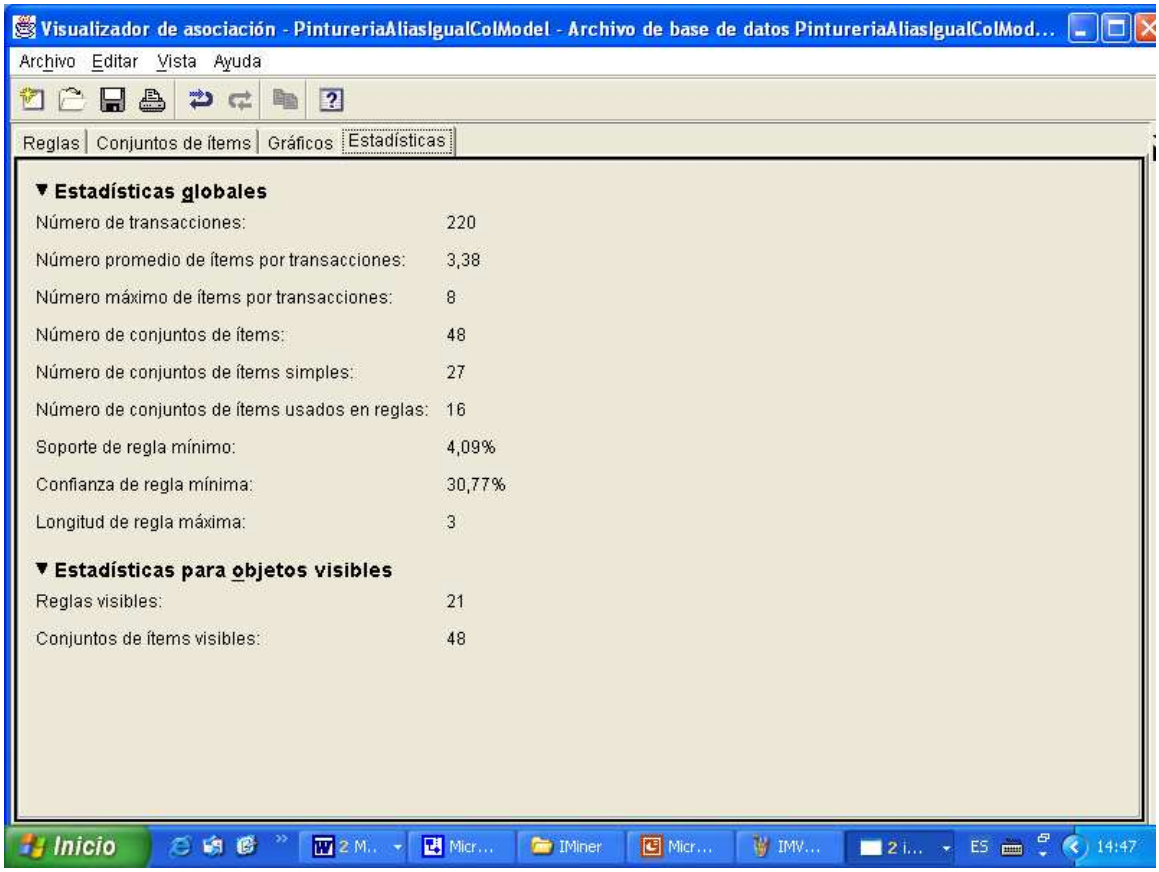


Vista Estadísticas

Incluye las secciones que pueden apreciarse en la siguiente imagen.

La Sección Estadísticas visibles le muestra la cantidad de reglas y conjuntos de reglas del modelo que son visibles en el Visualizador de asociación.

Si se han ocultado reglas o conjuntos de ítems, se visualizará la cantidad de reglas o conjuntos de ítems visibles. Si no se han ocultado reglas ni conjuntos de ítems, se mostrará la cantidad total de reglas y conjuntos de ítems que incluye el modelo.


Visualizador de asociación - PintureriaAliasIgualColModel - Archivo de base de datos PintureriaAliasIgualColMod...
Archivo Editar Vista Ayuda
Reglas Conjuntos de ítems Gráficos Estadísticas
▼ Estadísticas globales
Número de transacciones: 220
Número promedio de ítems por transacciones: 3,38
Número máximo de ítems por transacciones: 8
Número de conjuntos de ítems: 48
Número de conjuntos de ítems simples: 27
Número de conjuntos de ítems usados en reglas: 16
Soporte de regla mínimo: 4,09%
Confianza de regla mínima: 30,77%
Longitud de regla máxima: 3
▼ Estadísticas para objetos visibles
Reglas visibles: 21
Conjuntos de ítems visibles: 48
Inicio
2 M...
Micr...
IMiner
Micr...
IMV...
2 i...
ES
14:47