

**UNIVERSIDAD**  
**NACIONAL DEL**  
**NORDESTE**

**TRABAJO DE INVESTIGACIÓN BIBLIOGRÁFICA:**  
***MINERÍA DE DATOS***

**DEPARTAMENTO DE INFORMÁTICA**

**PROFESOR: MASTER DAVID LUIS LA RED MARTÍNEZ**

**ALUMNO: RAMÓN DAVID E. LEZCANO**

## **OBJETIVOS**

- ◆ Analizar y entender qué es la Minería de Datos.
- ◆ Cómo la Minería de Datos se relaciona con el KDD o descubrimiento de conocimientos.
- ◆ Reconocer la problemática del análisis de grandes volúmenes de datos y de los beneficios de su uso sistemático para la obtención de modelos y patrones predictivos o descriptivos.
- ◆ Diferenciar entre Estadística y Minería de Datos.
- ◆ Conocer las aplicaciones habituales de la Minería de Datos.
- ◆ Conocer por qué su importancia hoy en día.
- ◆ Conocer la relación de la Minería de Datos con otras disciplinas.

## **INTRODUCCIÓN**

La medición del software está adquiriendo una gran importancia debido a que cada vez es mayor la necesidad de obtener datos objetivos que permitan evaluar, predecir y mejorar la calidad del software, así como el tiempo y coste de desarrollo del mismo.

Asimismo, en los últimos años se ha visto un gran crecimiento en la capacidad de generación y almacenamiento de información, debido a la creciente automatización de procesos y los avances en las capacidades de almacenamiento de información. Gran parte de esa información es histórica, es decir, representa transacciones o situaciones que se han producido. Aparte de su función de “Memoria de la Organización”, la información ésta histórica, es útil para predecir información futura, ya que la mayoría de las decisiones de empresas, organizaciones e instituciones se basan en información de experiencias pasadas, extraídas de fuentes muy diversas.

Desgraciadamente, no se ha visto un desarrollo equivalente en las técnicas de análisis de información, por lo que existe la necesidad de una nueva generación de técnicas y herramientas computacionales con la capacidad de asistir a usuarios en el análisis automático e inteligente de datos. El procesar automáticamente grandes cantidades de datos para encontrar conocimiento útil para un usuario y satisfacerle sus metas, es el objetivo principal del área de Descubrimiento de Conocimiento en Bases de Datos o KDD (Knowledge Discovery from Data base). Este es el campo que está evolucionando para proporcionar soluciones al análisis automatizado, al que también podemos definirlo como: Un proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia comprensible a partir de datos o como la extracción no trivial de información implícita, desconocida, y potencialmente útil de los datos.

Es muy importante comprender al KDD, ya que el tema que vamos a tocar (Minería de Datos), no es más que una fase del mismo, fase que integra los métodos de aprendizaje y estadísticas para obtener hipótesis de patrones y modelos, además por que las técnicas de minería de datos surgen como las mejores herramientas para realizar exploraciones más profundas y extraer información nueva, útil y no trivial que se encuentra oculta en grandes volúmenes de datos. Es importante también aclarar que vulgarmente se asimila a KDD con Minería de Datos.

## **MINERÍA DE DATOS (DM). DESCUBRIMIENTO DE CONOCIMIENTOS (KD)**

Se puede decir que un sistema Data Mining es una tecnología soporte para usuario final cuyo objetivo es extraer conocimiento útil y utilizable a partir de la información contenida en bases de datos; también se llama minería de datos (data mining) al análisis de archivos y bitácoras de transacciones que sean útiles para la toma de decisiones. La especie humana posee habilidades extremadamente sofisticadas para detectar patrones y descubrir tendencias. Por tal motivo una imagen nos dice más que mil palabras y una gráfica nos permite, de una mirada, identificar tendencias en el tiempo o relaciones entre dos mediciones de un fenómeno. Por otro lado, no es claro que nuestras habilidades puedan realizar, con la misma eficiencia, la tarea de analizar los trillones de datos almacenados electrónicamente al monitorear las transacciones comerciales de una base de datos.

Dada de la tecnología actual, resulta más o menos sencillo coleccionar grandes volúmenes de información. Con el uso de lectura óptica y código de barras, las cadenas de supermercados pueden fácilmente coleccionar la información de cada canasta de compra, es decir, cual es el conjunto de artículos que el cliente compra. Un concepto similar es el estado de cuenta mensual de una tarjeta de crédito en el que se describe un conjunto de artículos que el cliente adquirió ese mes. De igual manera, gobiernos, instituciones públicas y privadas, están en la posibilidad de juntar millones y millones de datos de actividades individuales que contienen información altamente detallada sobre montos, fechas, horas, lugares, productos y servicios.

Esta información cruda es tan voluminosa que resulta inútil, pues no aporta conocimiento o fundamento para la toma de decisiones. El resumir datos para la toma de decisiones ha sido el campo tradicional de la estadística pero hoy en día existen nuevas técnicas, una de ella es la Minería de Datos, la que revela patrones o asociaciones que usualmente nos eran desconocidas y se le ha llamado también descubrimiento de conocimiento (KD Knowledge Discovery).

El descubrir patrones o relaciones útiles en una colección de datos ha recibido tradicionalmente muchos nombres. El término data mining llegó incluso a ser muy desprestigiado en la estadística, pues representaba “masajear” suficientemente los datos hasta que los mismos confirmasen lo que uno quería postular. En ese sentido, la minería de datos es un proceso que invierte la dinámica del método científico en el siguiente sentido.

En el método científico, primero se formulan las hipótesis y luego se diseña el experimento para coleccionar los datos que confirmen o refuten la hipótesis. Si esto se hace con la formalidad adecuada (cuidando cuáles son las variables controladas y cuáles experimentales), se obtiene un nuevo conocimiento.

En la minería de datos, se coleccionan los datos y esperamos que de ellos emerjan hipótesis.

Al hablar de descubrimiento de conocimientos en base de datos decimos que es un *proceso de extracción no trivial para identificar patrones que sean válidos, novedosos, potencialmente útiles y entendibles, a partir de datos.*

- Proceso: KDD involucra varios pasos y es interactivo, al encontrar información útil en los datos, se realizan mejores preguntas.
- Válido: se utilizan principalmente los datos y se espera que los patrones puedan aplicarse en el futuro.
- Novedoso: desconocido con anterioridad.
- Útil: aplicable y cumpliendo las metas del usuario.
- Entendible: que nos lleve a la comprensión, muchas veces medido por el tamaño.

El proceso de KDD consiste en usar métodos de minería de datos (algoritmos) para extraer (identificar) lo que se considera como conocimiento de acuerdo a la especificación de ciertos parámetros usando una base de datos junto con pre-procesamientos y post-procesamientos.

Se estima que la extracción de patrones (minería) de los datos ocupa solo el 15% - 20% del esfuerzo total del proceso de KDD.

## **MINERÍA DE DATOS VERSUS ESTADÍSTICA**

El *Data Mining* es el descendiente y -según algunos- el sucesor de la estadística tal y como ésta se utiliza actualmente.

Pero lo que se pretende en este punto es explicar las diferencias entre data mining y estadística, desde una perspectiva constructiva en el uso de ambas herramientas analíticas y bajo un contexto empresarial.

Estadística y *Data Mining* conducen al mismo objetivo, el de efectuar “modelos” compactos y comprensibles que rindan cuenta de las relaciones establecidas entre la descripción de una situación y un resultado (o un juicio) relacionado con dicha descripción. También apunta a mejorar la toma de decisiones mediante un conocimiento del entorno.

Este entorno lo facilitan los datos almacenados en la compañía, cuantitativos o cualitativos y mediante información de terceras empresas.

Fundamentalmente, la diferencia entre ambas reside en que las técnicas del *Data Mining* construyen el modelo de manera automática mientras que las técnicas estadísticas “clásicas” necesitan ser manejadas - y orientadas - por un estadístico profesional.

Las técnicas de *Data Mining* permiten ganar tanto en performance como en manejabilidad e incluso en tiempo de trabajo. La posibilidad de realizar uno mismo sus propios modelos sin necesidad de sub-contratar ni ponerse de acuerdo con un estadístico proporciona una gran libertad a los usuarios profesionales.

Pero es importante aclarar que la estadística se utiliza para validar o para matizar un modelo sugerido y preexistente, no para generarlo.

La data mining aventaja a la estadística en los siguientes supuestos:

- Las técnicas estadísticas se centran generalmente en técnicas confirmatorias, mientras que las técnicas de data mining son generalmente exploratorias. Así, cuando el problema al que pretendemos dar respuesta es refutar o confirmar una hipótesis, podremos utilizar ambas ciencias (diferentes conclusiones y más robusta la estadística). Sin embargo, cuando el objetivo es meramente exploratorio (para concretar un problema o definir cuáles son las variables más interesantes en un sistema de información) surge la necesidad de delegar parte del conocimiento analítico de la empresa en técnicas de aprendizaje (inteligencia artificial), utilizando data mining. Aquí hemos detectado una primera diferencia de aplicación de ambas herramientas: data mining se utilizará cuando no partamos de supuestos de partida y pretendamos buscar algún conocimiento nuevo y susceptible de proporcionar información novedosa en la toma de decisiones.
- A mayor dimensionalidad del problema la data mining ofrece mejores soluciones. Cuantas más variables entran en el problema, más difícil resulta encontrar hipótesis de partida interesantes. O, aún cuando pudiera, el tiempo necesario no justificará la inversión. En ese caso, utilizar técnicas de data mining como árboles de decisión nos permitirá encontrar relaciones inéditas para luego concretar la investigación sobre las variables más interesantes.
- Las técnicas de data mining son menos restrictivas que las estadísticas. Una vez encontrado un punto de partida interesante y dispuestos a utilizar algún análisis estadístico en particular (por ejemplo, discriminante para diferenciar segmentos de mercado), puede suceder que los datos no satisfagan los requerimientos del análisis estadístico. Entonces, las variables deberán ser examinadas para determinar qué tratamiento permite adecuarlas al análisis, no siendo posible o conveniente en todos los casos. Aquí también destaca la data mining, puesto que es menos restrictivo que la estadística y permite ser utilizado con los mínimos supuesto posibles (permite ‘escuchar’ a los datos).

Cuando los datos de la empresa son muy ‘dinámicos’ las técnicas de data mining inciden sobre la inversión y la actualización del conocimiento de nuestro negocio. Un almacén de datos poco ‘dinámico’ permite que una inversión en un análisis estadístico quede justificada -personal cualificado en estadística, metodología rígida y respuestas a preguntas muy concretas- dado que las conclusiones van a tener un ciclo de vida largo. Sin embargo, en un almacén ‘muy dinámico’ las técnicas de data mining permiten explorar cambios y determinar cuándo una regla de negocio ha cambiado. Permitiendo abordar diferentes cuestiones a corto / medio plazo.

Exponemos ahora aquellos contextos en los que es más adecuado el análisis estadístico que el de data mining:

- El objetivo de la investigación es encontrar causalidad. Si se pretende determinar cuáles son las causas de ciertos efectos (por ejemplo, si invertir más en la publicidad de cierto producto tiene como consecuencia un incremento de ventas o si es más determinante el ofrecer un descuento a los clientes), deberemos utilizar técnicas de estadística (por ejemplo, ecuaciones estructurales). Las relaciones complejas que subyacen a técnicas de data mining impiden una interpretación certera de diagramas causa-efecto.
- Se pretende generalizar sobre poblaciones desconocidas en su globalidad. Si las conclusiones han de ser extensibles a otros elementos de poblaciones similares habrán de utilizarse técnicas de inferencia estadística. Esto viene relacionado con situaciones en las que se dispone exclusivamente de muestras (con el consiguiente problema de aportar validez a las muestras). En data mining, se generarán modelos y luego habrán de validarse con otros casos conocidos de la población, utilizando como significación el ajuste de la predicción sobre una población conocida (es lo habitual cuando queremos predecir perfiles de clientes, que ya disponemos de antecedentes para poder validarlos, aunque no siempre es posible acceder a dicha información o no siempre es correcto aplicar ciertas muestras).

Se han detallado algunos argumentos acerca de cuándo es conveniente utilizar data mining o estadística. Llegado a este punto deseamos destacar que ambas perspectivas constituyen una sinergia y que no son excluyentes una de la otra. En este sentido, la metodología de un proyecto de data mining ha de contener referencias a la estadística en dos partes destacables del proceso:

- Preparación de los datos (tratamiento de valores erróneos, valores omitidos,...) y aproximación a las variables de estudio.
- Despliegue del proyecto y posible generación de hipótesis a refutar con una metodología y técnica estadística.

Así pues, data mining y estadística son técnicas complementarias que permiten obtener conocimiento inédito en nuestros almacenes de datos o dar respuestas a cuestiones concretas de negocio.

## **TIPOLOGÍA DE LAS TÉCNICAS DE MINERÍA DE DATOS**

Las técnicas de minerías de datos crean dos modelos:

Tenemos los Modelos Predictivos o basados en la Memoria y los Modelos Descriptivos.

### **Modelos Predictivos o Basados en la Memoria**

Técnicas: Clasificación, Predicción de valores.

Ejemplos: ¿Cuál es el riesgo de este cliente?. ¿Se quedará el cliente?.

Los modelos predictivos requieren de un set de pruebas y de interacciones de entrenamiento:

1. Selección de pruebas.
2. Minado inicial.
3. Resultado.
4. Aplicación de una segunda muestra representativa.
5. Análisis de los resultados.
6. Interacciones hasta lograr un modelo consistente.
7. Aplicar al negocio.

### **Modelos Descriptivos**

Técnicas: Asociación, Segmentación o 'Clustering'.

Ejemplos: Un cliente que compra productos dietéticos es tres veces más probable que compre caramelos.

## **Componentes básicas de los métodos de Minería de Datos**

Sus componentes básicos son:

1. **Lenguaje de representación del modelo:** Es muy importante que se sepan las suposiciones y restricciones en la representación empleada para construir modelos.
2. **Evaluación del modelo:** En cuanto a predictividad se basa en técnicas de validación cruzada (*cross validation*); en cuanto a calidad descriptiva del modelo se basan en principios como el de máxima verosimilitud (*maximum likelihood*) o en el principio de longitud de descripción mínima o MDL (*minimum description length*).
3. **Método de búsqueda:** Se puede dividir en búsqueda de parámetros y búsqueda del modelo y determinan los criterios que se siguen para encontrar los modelos (hipótesis).

## TÉCNICAS DE MINERÍA DE DATOS

La minería de datos ha dado lugar a una paulatina sustitución del análisis de datos dirigido a la verificación por un enfoque de análisis de datos dirigidos al descubrimiento del conocimiento. La principal diferencia entre ambos se encuentra en que en el último se descubre información sin necesidad de formular previamente una hipótesis.

La aplicación automatizada de algoritmos de minería de datos permite detectar fácilmente patrones en los datos, razón por la cual esta técnica es mucho más eficiente que el análisis dirigido a la verificación cuando se intenta explorar datos procedentes de repositorios de gran tamaño y complejidad elevada. Dichas técnicas emergentes se encuentran en continua evolución como resultado de la colaboración entre campos de investigación tales como bases de datos, reconocimiento de patrones, inteligencia artificial, sistemas expertos, estadísticas, visualización, recuperación de la información y computación de altas prestaciones.

Los algoritmos de minería de datos se clasifican en dos grandes categorías: supervisados o predictivos y no supervisados o de descubrimiento del conocimiento.

Los algoritmos **supervisados o predictivos** predicen el valor de un atributo (etiqueta) de un conjunto de datos, conocidos otros atributos (atributos descriptivos). A partir de datos cuya etiqueta se conoce, se induce una relación entre dicha etiqueta y otras series de atributos. Esas relaciones sirven para realizar la predicción en datos cuya etiqueta es desconocida. Esta forma de trabajar se conoce como aprendizaje supervisado y se desarrolla en dos fases: entrenamiento (construcción de un modelo usando un subconjunto de datos con etiqueta conocida) y prueba (prueba del modelo sobre el resto de los datos).

Cuando una aplicación no es lo suficientemente madura no tiene el potencial necesario para una solución predictiva, en ese caso hay que recurrir a los métodos **no supervisados o de descubrimiento del conocimiento** que descubren patrones y tendencias en los datos actuales (no utilizan datos históricos). El descubrimiento de esa información sirve para

llevar a cabo acciones y obtener un beneficio (científico o de negocio) de ellas. La tabla siguiente muestra algunas de las técnicas de minería de ambas categorías.

<b>SUPERVISADOS</b>	<b>NO SUPERVISADO</b>
Árboles de decisión	Detección de desviaciones
Introducción neuronal	Segmentación
Regresión	Agrupamiento (“clustering”)
Series temporales	Reglas de asociación
	Patrones secuenciales

La aplicación de los algoritmos de minería de datos requiere la realización de una serie de actividades previas encaminadas a preparar los datos de entrada debido a que, en muchas ocasiones, dichos datos proceden de fuentes heterogéneas, no tienen el formato adecuado o contienen ruido. Por otra parte, es importante interpretar y evaluar los resultados obtenidos.

El proceso completo consta de los pasos que se detallan seguidamente.

## **DATA MINING: PASOS A SEGUIR:**

Los pasos a seguir son:

### **Paso 1: Identificar el problema de negocios:**

- Analizar si un objetivo es inútil.
- Metas sin significado son solo sueños.

### **Paso 2: Preparación de los Datos:**

- ¿Dónde se encuentran los datos?.
- ¿Cómo están estructurados?.
- ¿Cuándo están disponibles?.
- ¿Qué significado tienen los datos?.
- ¿Los datos están relacionados a los objetivos de negocio?.
- ¿Podemos realizar muestras al azar para reducir el volumen?.
- ¿Qué variables y registros son apropiados como datos de entrada para el proceso de minería?.

### **Paso 3: Construir el Modelo de Minería de Datos:**

Las consideraciones son las siguientes:

- Técnicas necesarias.
- Secuencia de Técnicas.
- Mejores algoritmos.
- Modos - entrenamiento, pruebas, aplicación.
- Estimar la duración de la corrida de la minería.

#### **Paso 4: Análisis y Validación de los Resultados:**

La herramienta de minería entrega:

- Resultados de los datos.

Debemos decidir:

- Significado de los resultados.
- Importancia de los resultados.
- Suficiencia de los resultados.
- Debemos determinar cómo los resultados se relacionan con el problema original planteado.

#### **Paso 5: Implementar y Monitorear:**

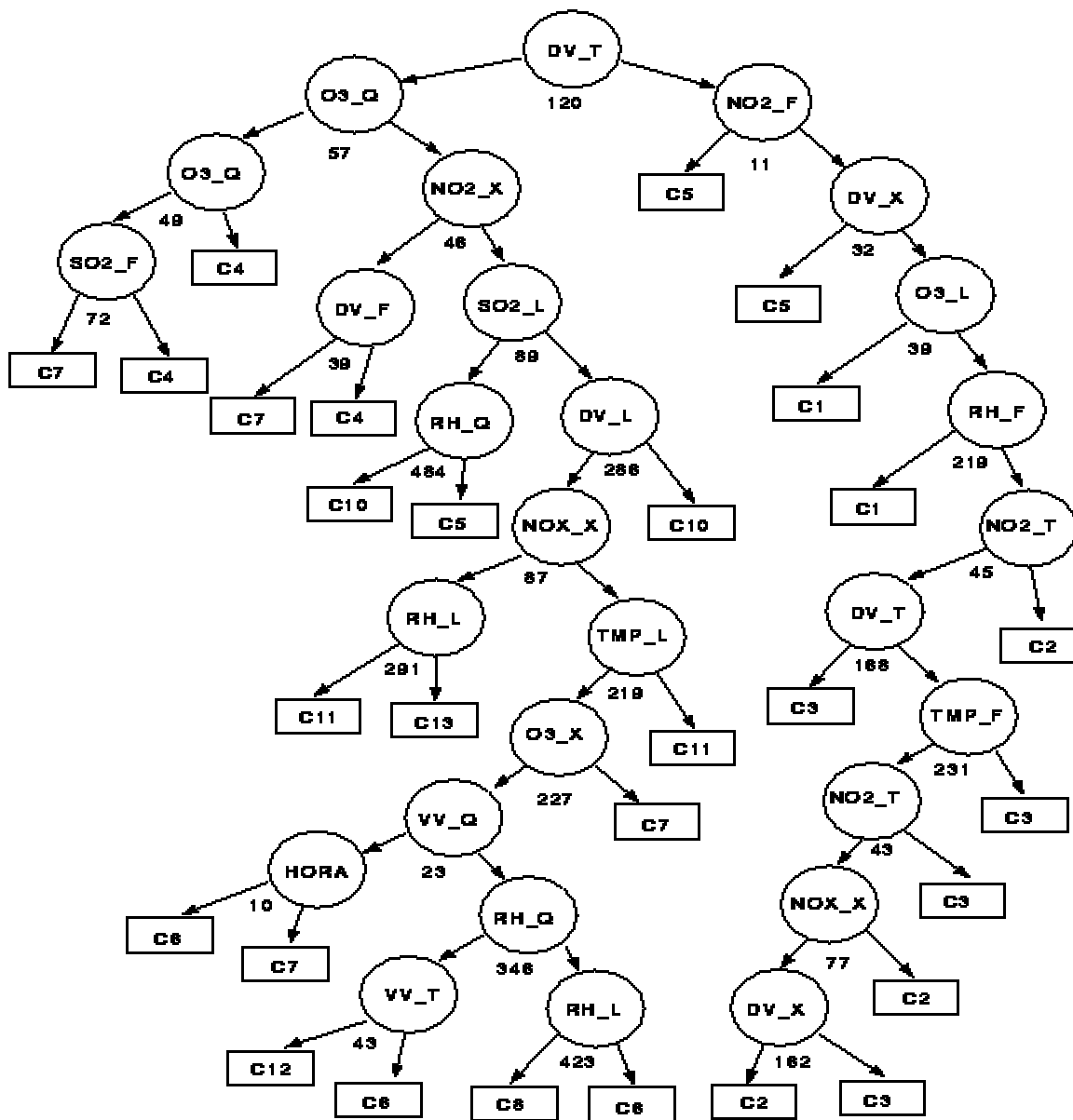
- Riesgo crediticio (predictivo).
- Pérdida y adquisición de clientes (predictivo).
- Morosidad de pagos (predictivo).
- Promociones de productos en conjunto (descriptivo).
- Campañas a grupos o segmentos específicos (descriptivo).
- Comunicaciones por correo (predictivo).
- Determinar el mercado para un nuevo servicio (descriptivo).

Aunque los pasos anteriores se realizan en el orden que aparecen, el proceso es altamente iterativo, estableciéndose retroalimentación entre los mismos. Además, no todos los pasos requieren el mismo esfuerzo; generalmente la etapa de procesamiento es la más costosa ya

que representa aproximadamente el 60% del esfuerzo total mientras que la etapa de minería de datos solo representa el 10%.

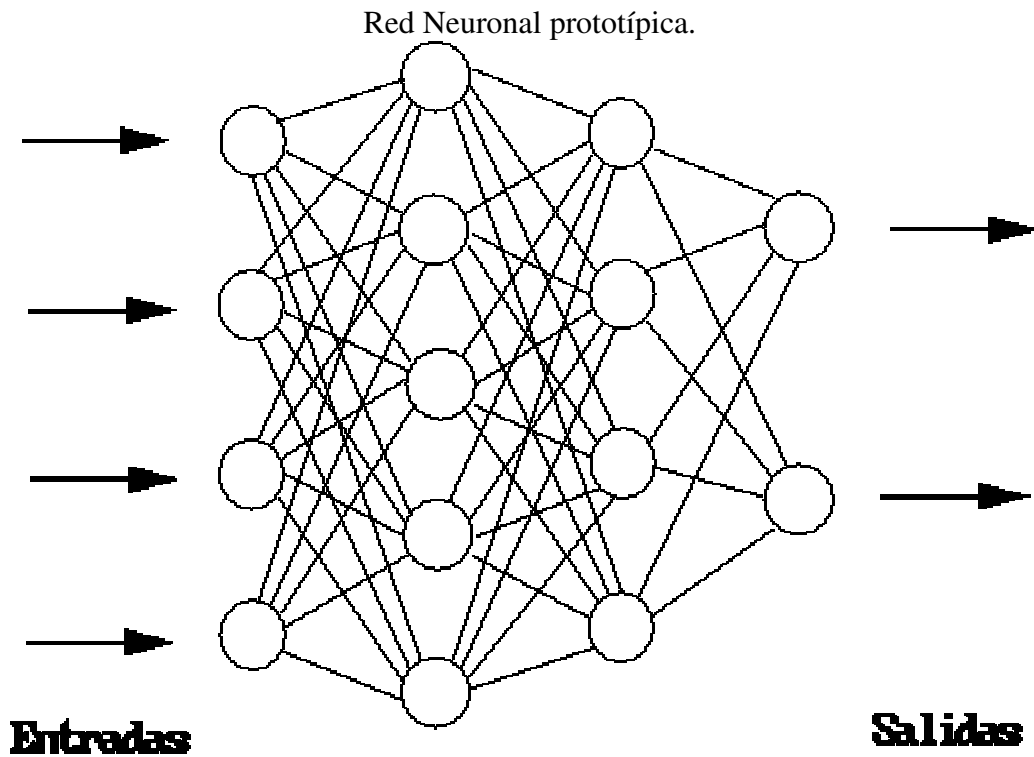
Para explicar las técnicas mencionadas anteriormente (supervisados y no supervisados) tenemos:

- Árboles de decisión y reglas de clasificación: realizan cortes sobre una variable (lo cual limita su expresividad, pero facilita su comprensión). Generalmente se usan técnicas heurísticas en su construcción.



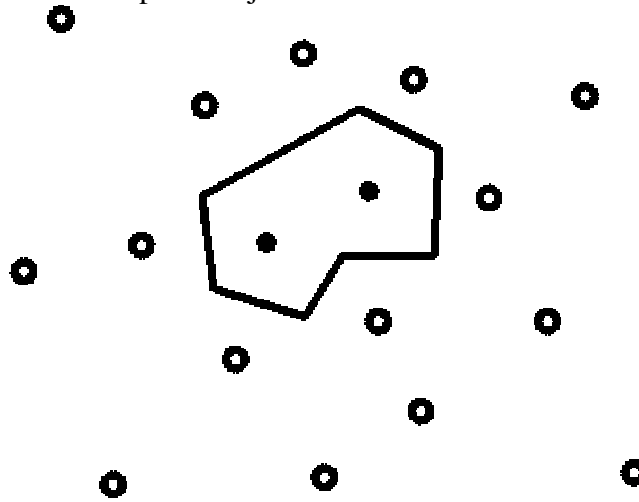
Predicción de Ozono en la Ciudad de México.

- Métodos de clasificación y regresiones no-lineales: tratan de ajustar combinaciones de funciones lineales y no-lineales, por ejemplo, redes neuronales (e.g., backpropagation), métodos de *splines* adaptativos, etc.



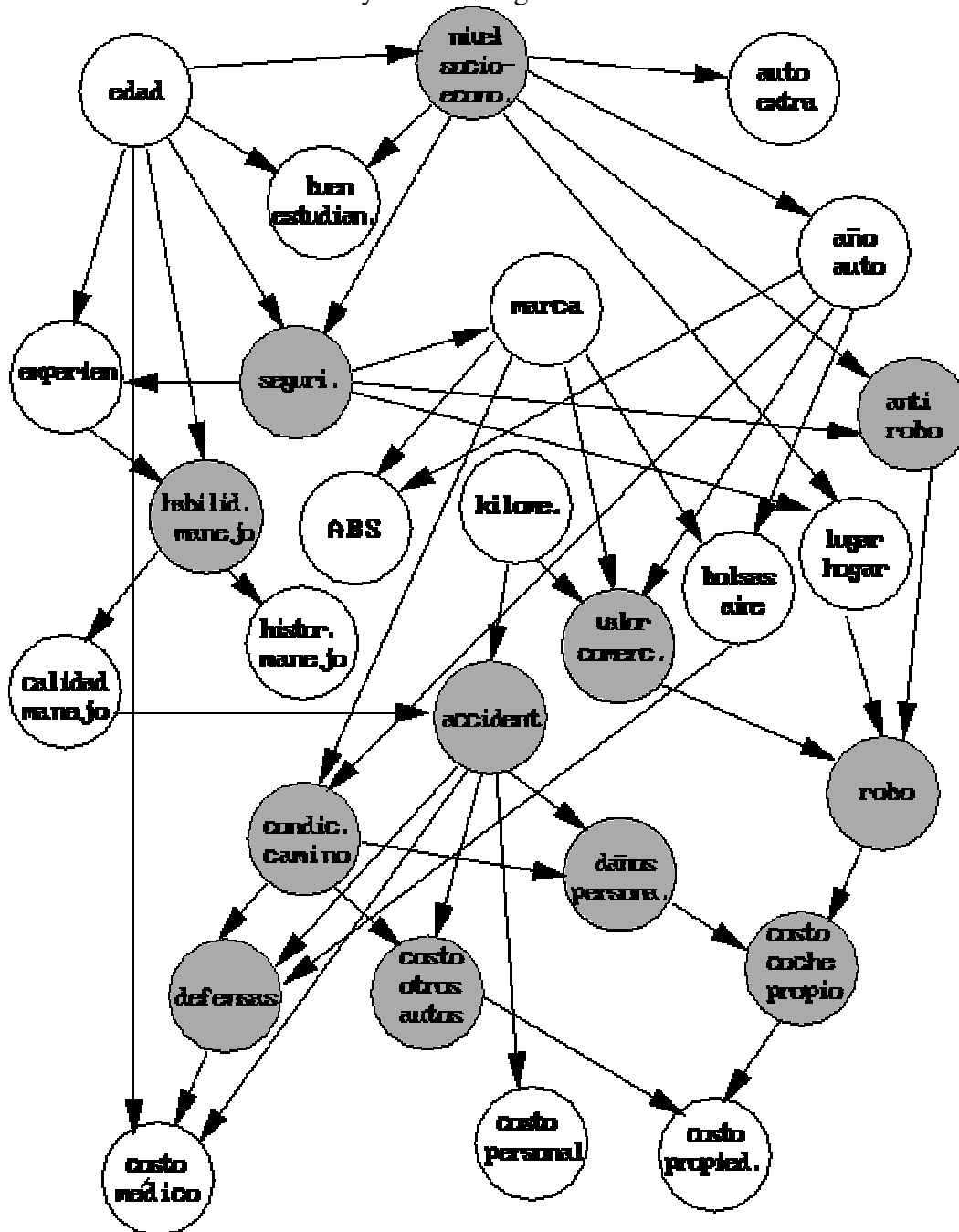
- Métodos basados en ejemplos prototípicos: se hacen aproximaciones sobre la base de los ejemplos o casos más conocidos (*exemplar-based learning* y *case-based reasoning*). El problema es cómo determinar una medida de similitud adecuada.

Aprendizaje basado en instancias.



- Modelos gráficos de dependencias probabilísticas: básicamente redes bayesianas, en donde la evaluación se basa en probabilidad y el encontrar el modelo en heurísticas.

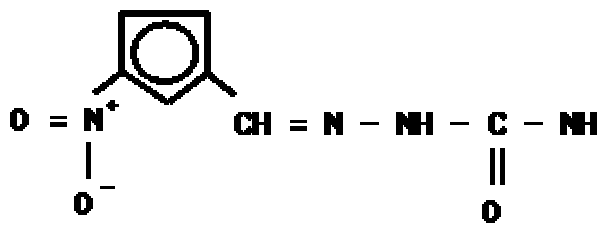
Red bayesiana de seguros de coches.



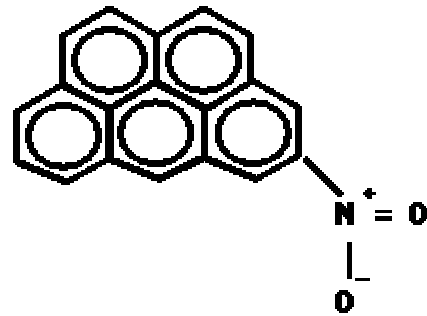
- Modelos relacionales: Programación lógica inductiva (o ILP), en donde la búsqueda del modelo se basa en lógica y heurísticas.

Predicción de muta génesis.

**ACTIVO**

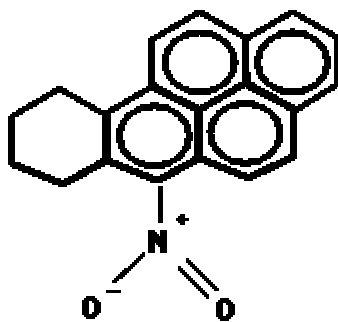


**nitrofurazone**

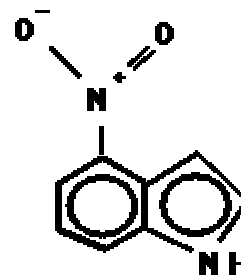


**4-nitropenta[cd]pyrene**

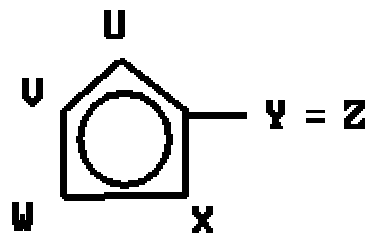
**INACTIVO**



**6-nitro-7,8,9,10-tetrahydrobenzo[a]pyrene**



**4-nitroindole**

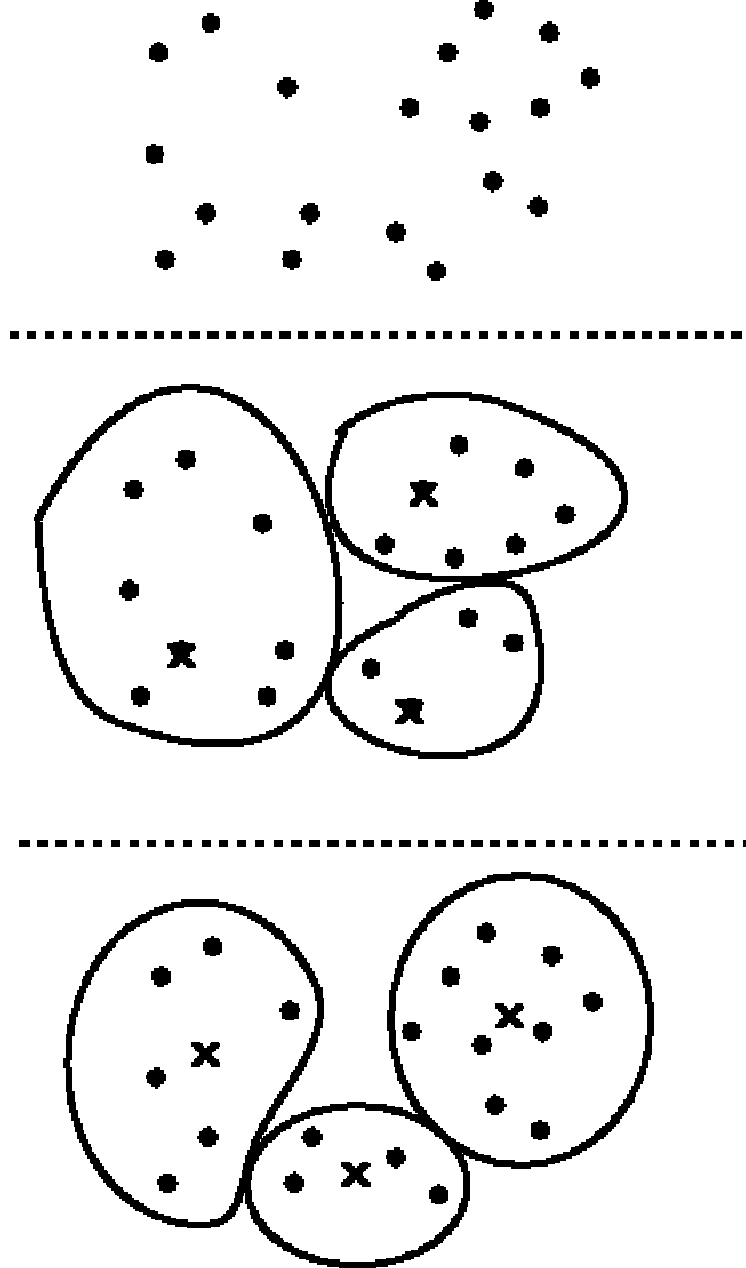


- Reglas de Asociación: reglas que relacionan un conjunto de pares atributo-valor con otros pares atributo-valor. Por ejemplo:

$edad(X, 20 \dots 29) \wedge ingresos(X, 20K..29K) \Rightarrow compra(X, CD)$   
 [soporte = 2%, confianza = 60%]

- Clustering: agrupan datos cuya distancia multidimensional intraclass es pequeña e interclass es grande.

Ejemplo de Clustering.



## **METAS DE LA MINERÍA DE DATOS**

El proceso de minería involucra ajustar modelos o determinar patrones a partir de datos. Este ajuste normalmente es de tipo estadístico en el sentido que se permite un cierto ruido o errores dentro del modelo.

Los algoritmos de minería de datos realizan en general tareas de descripción (de datos y patrones), de predicción (de datos desconocidos) y de segmentación (de datos). Otras como análisis de dependencias e identificación de anomalías se pueden usar tanto para descripción como para predicción:

- *Descripción*: normalmente es usada para análisis preliminar de los datos (resumen, características de los datos, casos extremos, etc.). Con esto, el usuario se sensibiliza con los datos y su estructura.
- *Predicción*: la podemos dividir en dos: Clasificación y Estimación:
  - *Clasificación*: los datos son objetos caracterizados por atributos que pertenecen a diferentes clases (etiquetas discretas).
  - *Estimación o Regresión*: las clases son continuas.
- *Segmentación*: separación de los datos en subgrupos o clases interesantes.
- *Detección de desviaciones, casos extremos o anomalías*: detectar los cambios más significativos en los datos con respecto a valores pasados o normales. Sirve para filtrar grandes volúmenes de datos que son menos probables de ser interesantes. El problema está en determinar cuándo una desviación es significativa para ser de interés.

## **AREAS DE APLICACIÓN**

En la actualidad existe una gran cantidad de aplicaciones en áreas como:

- Toma de decisiones (banca-finanzas-seguros, marketing, políticas sanitarias / demográficas). Estas más importantes industrialmente.
- Astronomía: clustering y clasificación de cuerpos celestes.
- Medicina.
- Biología molecular: predicción de sustancias cancerígenas, genoma humano, etc.
- Aspectos climatológicos: predicción de tormentas, etc.

- Industria y manufactura: diagnóstico de fallas.
- Mercadotecnia: identificar clientes susceptibles de responder a ofertas de productos y servicios por correo, selección de sitios de tiendas, etc.
- Inversión en casas de bolsa y banca: análisis de clientes, aprobación de préstamos, etc.
- Detección de fraudes y comportamientos inusuales: telefónicos, seguros, electricidad, etc.

Algunos de los ejemplos para los cuales se utiliza:

- En el Comercio / Marketing:
  - Identificar patrones de compras de los clientes.
  - Análisis de cestas de las compras, etc.
- En Banca:
  - Detectar patrones de uso fraudulentos de tarjetas de créditos.
  - Identificar clientes leales.
  - Predecir clientes con probabilidad de cambiar su afiliación.
  - Determinar gastos de tarjetas de créditos por grupo.
- Seguros y Salud Privada:
  - Análisis de Procedimientos médicos solicitados conjuntamente.
  - Predecir clientes que compran nuevas pólizas.
  - Identificar comportamiento fraudulento.
- Transporte:
  - Determinar la planificación de la distribución entre tiendas.
  - Analizar patrones de carga.
- Medicina:
  - Identificación de terapias médicas satisfactorias para diferentes enfermedades.
  - Asociación de síntomas y clasificación diferencial de patologías.
  - Estudio de factores (genéticos, precedentes, etc.) de riesgo / 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, etc.

## **PROBLEMAS DE APLICACIÓN.**

La data mining presenta los siguientes problemas de aplicación:

- Entrenamiento insuficiente.
- Herramientas de soporte inadecuadas.
- Abundancia de patrones.
- Cambios rápidos de los datos en el tiempo.
- Datos complejos (espaciales, imágenes, texto, audio, video).

## **ALCANCE DE LA MINERÍA DE DATOS**

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*

Data Mining automatiza el proceso de 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 redituvar los beneficios de automatización en las plataformas de hardware y software existentes y pueden 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. 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.

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

## **¿CÓMO SE DESARROLLA EL SISTEMA DE DATA MINING? ¿CÓMO TRABAJA Y QUÉ SON CAPACES DE HACER SUS HERRAMIENTAS?**

Los sistemas Data Mining se desarrollan bajo lenguajes de última generación basados en la inteligencia artificial y utilizan modelos matemáticos tales como las ya mencionadas redes neuronales, árboles de decisión, clasificación, etc.

¿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.

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 comparados contra los datos excluidos para confirmar la validez del modelo. Si el modelo funciona, las observaciones deben mantenerse para los datos excluidos.

## **¿Por que usar Data Mining?**

Proporciona poderes de decisión a los usuarios del negocio que mejor entienden el problema y el entorno y es capaz de medir las 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 de Data Mining sean expresadas como reglas de negocio o modelos predictivos. Estos output pueden comunicarse en formatos tradicionales (presentaciones, informes, información electrónica compartida, embebidos en aplicaciones, etc.) para guiar la estrategia y planificación de la empresa.

## **ALGUNO FACTORES IMPORTANTES SON:**

- El abaratamiento de los sistemas de almacenamiento tanto temporal como permanente.
- Mejora la calidad de datos.
- El incremento de las velocidades de cómputo en los procesadores.
- Mejora de consultas.
- Soporta el diseño de base de datos.

- Las mejoras en la confiabilidad y aumento de la velocidad en la transmisión de datos.
- El desarrollo de sistemas administradores de bases de datos más poderosos.

## **RETOS DE LA MINERÍA DE DATOS**

Los tres retos fundamentales son:

Primer reto: facilidad con que se puede caer en una falsa interpretación.

Tiempo y espacio: la modelación en computadora del tiempo y el espacio son problemas complejos, especialmente para hacer inferencias.

Privacidad: cuando la Minería de Datos era aún emergente, se llegó a pensar que no presentaba ningún peligro o riesgo para la privacidad de los clientes. Hoy en día, se piensa todo lo contrario.

Además podemos mencionar:

- Volumen de datos (mega, giga y hasta terabytes).
- Tratamiento de datos cambiantes: necesidad de revisión y extensión de patrones (incrementalidad).
- Alta dimensionalidad.
- Sobre ajuste (*over fitting*) de modelos en los datos.
- Evaluación de significancia estadística.
- Minería de datos con tipos no-estándar, multimedia u orientado a objeto.
- Datos y conocimiento dinámicos (datos en BD y los patrones encontrados cambian continuamente).
- Ruido, incertidumbre (tanto en datos como en conocimiento del dominio y en patrones descubiertos) y datos incompletos y / o esparcidos.
- Relaciones complejas entre campos, jerarquías, relaciones entre atributos, nuevos atributos, etc.
- Entendimiento de patrones.
- Incorporación de conocimiento del dominio.

- Interacción activa del usuario.
- Integración con otros sistemas.
- Información redundante (puede “descubrirse” erróneamente).

## **MINERÍA DE REPORTES Y DOCUMENTOS**

Los reportes de aplicaciones que se imprimen cada día contienen la mayoría de los datos necesarios para dar soporte a la decisión y a las iniciativas de inteligencia de negocio. Sin embargo, si los datos no se filtran, resumen o presentan como cada usuario necesita, son solo datos -no información-. La minería transforma los datos en informes creando y entregando vistas personalizadas de los datos del archivo de reporte de los reportes de aplicaciones existentes. Ahora se tiene una solución completamente automatizada, fácil de mantener para minería y entregar información valiosa dentro de los archivos de reportes de clases de producción.

Se puede mencionar el Modulo Ciprés Data Mining que filtra, ordena y resume datos del reporte a las necesidades de cada usuario, guarda la información como una hoja de trabajo Excel o archivos de texto conveniente, después automáticamente las rutea al buzón apropiado, a la impresora, al fax, al directorio, o a otro destino de la empresa basada en el contenido del reporte.

***Algunas de sus características son:***

***Aumentar la productividad del trabajador de conocimiento:*** Ciprés automatiza las tareas de minería de datos realizadas normalmente por los usuarios finales, permitiendo que los trabajadores de conocimiento pasen más tiempo en sus trabajos, y menos tiempo aprendiendo y operando software sofisticado de minería de datos basados en clientes.

***Transformar los datos heredados en ricos reportes relacionales:*** Ciprés da nueva vida a los datos de aplicaciones preexistentes heredadas. Los reportes de mainframe ricos en datos se pueden ahora filtrar y resumir para contestar a preguntas, a decisiones económicas y a necesidades específicas de la dirección.

***Salvaguardar el contenido sobre una base “necesitar conocer”:*** Porque la minería de datos de Ciprés se basa en el servidor Docu Vault, los usuarios ven solamente la información para la cual están autorizados, algo que las soluciones de minería de datos de usuario final generalmente no pueden proporcionar consistentemente.

Es importante mencionar que *la minería de datos es solo el principio.*

Una vez que se hayan procesado los archivos de reporte mediante el modulo Data Mining, el archivo resultante de Excel o texto tiene acceso a todas las ventajas de la arquitectura de Ciprés, incluyendo el archivo, indexación, enrutamiento basado en contenido, entrega en

Web y más. Los archivos de texto también tienen acceso a una variedad amplia de características de realce del documento tales como formatos, fuentes y más.

## **ALGO CURIOSO**

### **MINERÍA DE DATOS: EXPEDIENTES XBOX**

La gente de marketing de Xbox usa herramientas de minería de datos de digiMine para entender las necesidades de los visitantes. El objetivo es vender más y adaptar los sitios a los grupos individuales.

Casi todas las organizaciones tienen una base de datos, y la mayor parte tienen más de una; si las organizaciones pueden reunir la información de estas distintas bases de datos, pueden atrapar más criminales, vender más productos u operar con mayor eficiencia.

Compañías como digiMine, Autonomy, Clear Forest e diphase Technologies venden software que usa algoritmos complejos para buscar relaciones entre punto de datos dispersos en almacenes de datos o reunidos en uno solo.

Los organismos gubernamentales pueden beneficiarse con la minería de datos. De hecho, Autonomy y otras compañías proporcionan al Departamento de Seguridad Interna de Estados Unidos software para compartir y analizar información. Pero, al parecer, los detallistas utilizan mejor la tecnología. El sitio Web de Microsoft para su popular consola de juegos, Microsoft Xbox, usa digiMine para estudiar la actividad en la Web y compararlas con la información de mercadotecnia. El objetivo es vender más productos y adaptar el sitio a los gustos individuales.

“La información que reunimos es central para la mercadotecnia de Xbox”, dice Scott Picle, gerente del sitio Xbox en línea. “Para nosotros es importante segmentar a los usuarios para darles lo que quieren”.

Los blocs de los visitantes al sitio de Xbox se guardan en un almacén de datos digiMine, mientras que la información personal de los clientes se almacena en Microsoft. Los gerentes de sitio de 27 países, así como el personal de mercadotecnia de Xbox, pueden iniciar una consulta desde un visualizador Web. Después digiMine extrae los datos relevantes, como los jugadores de Halo que ya leyeron el artículo en el sitio sobre Star Wars: The Clone Wars y demostraron interés en comprar cualquier juego en los últimos diez días, haciendo clic por medio de uno de los socios detallistas de Xbox en línea. Una vez que digiMine regresa la información, el personal de mercadotecnia de Xbox compara la información con su base de datos interna para crear ofrecimientos dirigidos por correo electrónico.

Algunas soluciones de minería de datos están dirigidas al cliente, pues ofrecen interfaces amigables para el autoservicio y la investigación.

Al final no importa si J. Crew en línea usa digiMine para comparar compradores potenciales con la línea de ropa más reciente, o si el FBI usa el software de Clear Forest para identificar terroristas, la minería de datos funciona igual. Buscando en montañas de información y analizando relaciones, las soluciones de minería de datos ayudan a dar sentido a un mundo basado en la información.

### **MINERÍA SOBRE TABLAS DINAMICAS**

Consiste en un informe de tabla dinámica que se puede utilizar para grandes volúmenes de datos. Podrá girar sus filas y columnas para ver diferentes datos de origen, filtrar los datos mostrando diferentes página, o mostrar las áreas de interés.

## **ANEXOS**

### **Data Warehouse**

*¿Qué es Data Warehousing?*

En la actualidad hay una gran confusión respecto a lo que es un Data Warehouse que, afortunadamente, está comenzando a despejarse. No obstante, parece que cada proveedor de un producto o servicio relacionado con tecnología informática tiene su definición y, lo que es peor, en su propia jerga no siempre comprensible.

Data Warehouse, Business Intelligence y Decision Support en realidad se consideran la solución integral y oportuna para desarrollar negocios.

El Data Warehouse se caracteriza por ser integrado, temático, histórico y no volátil.

*Definición:* es un proceso, no un producto. Es una técnica para consolidar y administrar datos de variadas fuentes con el propósito de responder preguntas de negocios y tomar decisiones, de una forma que no era posible hasta ahora.

Consolidar datos desde una variedad de fuentes. Dentro del marco conceptual de Data Warehousing, los agruparemos dentro del proceso de Transformación de Datos.

Manejar grandes volúmenes de datos de una forma que no era posible, o no era costo efectivo. A estos medios los agruparemos en Procesamiento y Administración de Datos.

Acceder a los datos de una forma más directa, en “el lenguaje del negocio”, y analizarlos para obtener relaciones complejas entre los mismos. Estos procesos se engloban en dos categorías: Acceso a los Datos y Descubrimiento o Data Mining.

Estos desarrollos tecnológicos, correctamente organizados e interrelacionados, constituyen lo que se ha dado en llamar un Data Warehouse o Bodega de Datos. Veamos un poco más en detalle los grupos mencionados.

Existen muchas definiciones para el DW, la más conocida fue propuesta por Inmon [Microst96] (considerado el padre de las Bases de Datos) en 1992: “Un DW es una colección de datos orientados a temas, integrados, no-volátiles y variante en el tiempo, organizados para soportar necesidades empresariales”.

En 1993, Susan Osterfeldt [Microst96] publica una definición que sin duda acierta en la clave del DW: “Yo considero al DW como algo que provee dos beneficios empresariales reales: integración y acceso de datos. DW elimina una gran cantidad de datos inútiles y no deseados, como también el procesamiento desde el ambiente operacional clásico”.

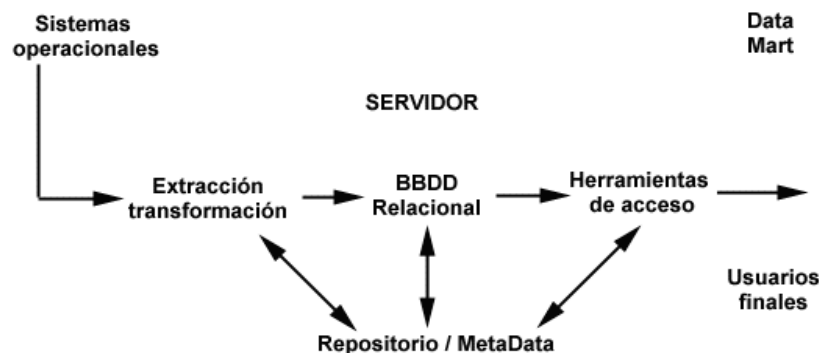
### Data Marts

Es un pequeños Data Warehouse, para un determinado número de usuarios, para un área funcional, específica de la compañía. También podemos definir que un Data Marts es un subconjunto de una bodega de datos para un propósito específico.

Su función es apoyar a otros sistemas para la toma de decisiones.

Los procesos que conforma el Datawarehouse son:

- Extracción.
- Elaboración.
- Carga.
- Explotación.



Componentes del Data Warehouse.

El éxito de DW no está en su construcción, sino en usarlo para mejorar procesos empresariales, operaciones y decisiones.

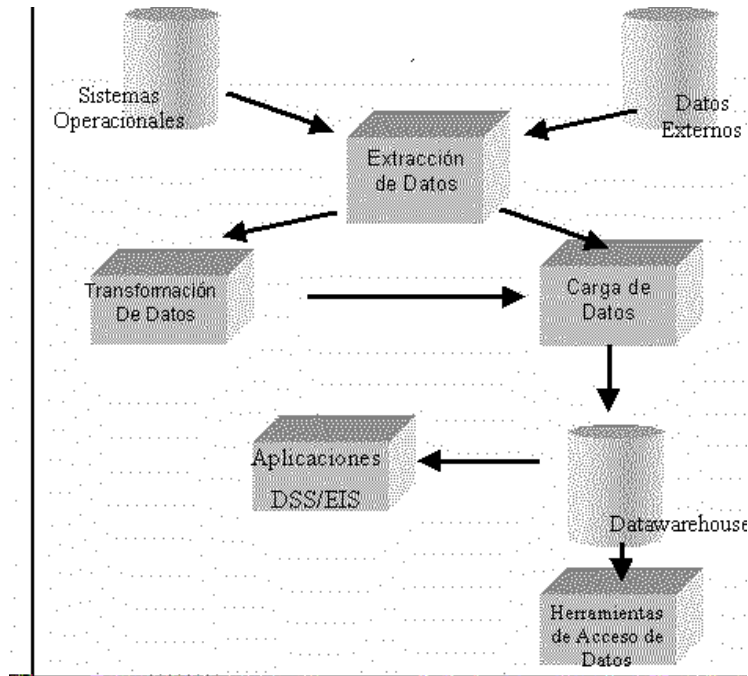


Diagrama de Funcionamiento.

*¿Cómo trabaja el Data Warehouse?*

- Extrae la información operacional.
- Transforma la información operacional a formatos consistentes.
- Automatiza las tareas de la información para prepararla a un análisis eficiente.

*¿En qué podemos usarlo?*

- Manejo de relaciones de marketing.
- Análisis de rentabilidad.
  - Reducción de costos.

## **Text Mining**

Las técnicas hasta ahora descritas sólo tratan datos numéricos o cualitativos. El text mining surge ante el problema cada vez más apremiante de extraer información automáticamente a partir de masas de textos. La enorme cantidad de referencias recogidas durante una búsqueda en Internet ilustra muy bien este problema.

La investigación literal simple se ha mostrado limitada desde hace ya mucho tiempo; hay muchos problemas como los errores de tipeado, la sinonimia, las acepciones múltiples, etc. En definitiva, es necesario inyectarle al ordenador un cierto sentido común o “conocimiento del mundo”. Aún en ese caso, la memoria y el poder de cálculo disponibles en nuestra época permiten ciertas soluciones que no siempre son las más elegantes pero sí potentes y rápidas.

Nuestras técnicas de “fuzzy string matching” y de búsqueda de contexto han dado excelentes resultados en la práctica.

### **Web Mining: (Minería de Web)**

Normalmente, el Web Mining puede clasificarse en tres dominios de extracción de conocimiento de acuerdo con la naturaleza de los datos:

**Web content mining** (minería de contenido web).

**Web structure mining** (minería de estructura web).

**Web usage mining** (minería de uso web).

## **TENDENCIAS**

❖ 80 y 90:

-OLAP: consultas predefinidas. El sistema OLAP como sistema para extraer gráficas y confirmar hipótesis. Técnicas fundamentalmente estadísticas.

- Se usa exclusivamente información interna a la organización.

❖ Finales de los 90:

-Data Mining: descubrimiento de patrones. Técnicas de aprendizaje automático para generar patrones novedosos.

-El Data Warehouse incluye información interna fundamentalmente.

❖ Principios del 2000:

-Técnicas de “scoring” y simulación: descubrimiento y uso de modelos globales. Estimación a partir de variables de entrada, de variables de salida (causa-efecto) utilizando simulación sobre el modelo aprendido.

-El Data Warehouse incluye información interna y externa (parámetros de la economía, poblaciones, geográficos, etc.).

## CONCLUSIONES

Existen métodos o técnicas no tradicionales con los que se pueden sacar u obtener información útil de grandes volúmenes de datos, además de la ya conocida estadística y que en muchos casos es mejor que esta; uno de ellos es la Minería de Datos o Data Mining.

El Data Mining brinda otra posibilidad para el análisis de datos y la obtención de modelos o información útil para diferentes acciones o finalidades.

Además, un sistema de Data Mining permite realizar diversidad de acciones, algunas de ellas son analizar factores de influencia en determinados procesos o estimar variables o comportamientos futuros, segmentar o agrupar ítems semejantes y lo que es más importante aún, no necesitar de un estadístico para lograr dichos objetivos o fines.

El Data Mining es sin duda una de las facilidades más útiles hoy en día disponibles para la extracción de conocimiento de grandes Almacenes de Datos, con claras aplicaciones en diversidad de organizaciones de distinto tipo.

## GLOSARIO

- **Algoritmos genéticos:** Técnicas de optimización que usan procesos tales como combinación genética, mutación y selección natural en un diseño basado en los conceptos de evolución natural.
- **Análisis de series de tiempo (time-series):** Análisis de una secuencia de medidas hechas a intervalos específicos. El tiempo es usualmente la dimensión dominante de los datos.
- **Análisis prospectivo de datos:** Análisis de datos que predice futuras tendencias, comportamientos o eventos basado en datos históricos.
- **Análisis exploratorio de datos:** Uso de técnicas estadísticas tanto gráficas como descriptivas para aprender acerca de la estructura de un conjunto de datos.
- **Análisis retrospectivo de datos:** Análisis de datos que provee una visión de las tendencias, comportamientos o eventos basado en datos históricos.
- **Árbol de decisión:** Estructura en forma de árbol que representa un conjunto de decisiones. Estas decisiones generan reglas para la *clasificación* de un conjunto de datos. Ver *CART* y *CHAID*.

- **Base de datos multidimensional:** Base de datos diseñado para procesamiento analítico on-line (*OLAP*). Estructurada como un hipercubo con un eje por dimensión.
- **CART Árboles de clasificación y regresión:** Una técnica de *árbol de decisión* usada para la *clasificación* de un conjunto de datos. Provee un conjunto de reglas que se pueden aplicar a un nuevo (sin clasificar) conjunto de datos para predecir cuáles registros darán un cierto resultado. Segmenta un conjunto de datos creando 2 divisiones. Requiere menos preparación de datos que *CHAID*.
- **CHAID Detección de interacción automática de Chi cuadrado:** Una técnica de *árbol de decisión* usada para la *clasificación* de un conjunto de datos. Provee un conjunto de reglas que se pueden aplicar a un nuevo (sin clasificar) conjunto de datos para predecir cuáles registros darán un cierto resultado. Segmenta un conjunto de datos utilizando tests de chi cuadrado para crear múltiples divisiones. Antecede, y requiere más preparación de datos, que *CART*.
- **Clasificación:** Proceso de dividir un conjunto de datos en grupos mutuamente excluyentes de tal manera que cada miembro de un grupo esté lo “más cercano” posible a otro, y grupos diferentes estén lo “más lejos” posible uno del otro, donde la distancia está medida con respecto a variable(s) específica(s) las cuales se están tratando de predecir. Por ejemplo, un problema típico de clasificación es el de dividir una base de datos de compañías en grupos que son lo más homogéneos posibles con respecto a variables como “posibilidades de crédito” con valores tales como “Bueno” y “Malo”.
- **Clustering (agrupamiento):** Proceso de dividir un conjunto de datos en grupos mutuamente excluyentes de tal manera que cada miembro de un grupo esté lo “más cercano” posible a otro, y grupos diferentes estén lo “más lejos” posible uno del otro, donde la distancia está medida con respecto a todas las variables disponibles.
- **Computadoras con multiprocesadores:** Una computadora que incluye múltiples procesadores conectados por una red. Ver *procesamiento paralelo*.
- **Data cleansing:** Proceso de asegurar que todos los valores en un conjunto de datos sean consistentes y correctamente registrados.
- **Data Mining:** La extracción de información predecible escondida en grandes bases de datos.
- **Data Warehouse:** Sistema para el almacenamiento y distribución de cantidades masivas de datos
- **Datos anormales:** Datos que resultan de errores (por ej. : errores en el tipeado durante la carga) o que representan eventos inusuales.
- **Dimensión:** En una base de datos relacional o plana, cada campo en un registro representa una dimensión. En una *base de datos multidimensional*, una dimensión es un conjunto de entidades similares; por ej. : una base de datos multidimensional de ventas podría incluir las dimensiones Producto, Tiempo y Ciudad.

- **Modelo analítico:** Una estructura y proceso para analizar un conjunto de datos. Por ejemplo, un *árbol de decisión* es un modelo para la *clasificación* de un conjunto de datos
- **Modelo lineal:** Un *modelo analítico* que asume relaciones lineales entre una variable seleccionada (dependiente) y sus preeditores (variables independientes).
- **Modelo no lineal:** Un *modelo analítico* que no asume una relación lineal en los coeficientes de las variables que son estudiadas.
- **Modelo predictivo:** Estructura y proceso para predecir valores de variables especificadas en un conjunto de datos.
- **Navegación de datos:** Proceso de visualizar diferentes dimensiones, “fetas” y niveles de una *base de datos multidimensional*. Ver *OLAP*.
- **OLAP Procesamiento analítico on-line (On Line Analytic Processing):** Se refiere a aplicaciones de bases de datos orientadas array que permite a los usuarios ver, navegar, manipular y analizar *bases de datos multidimensionales*.
- **Outlier:** Un ítem de datos cuyo valor cae fuera de los límites que encierran a la mayoría del resto de los valores correspondientes de la muestra. Puede indicar *datos anormales*. Deberían ser examinados detenidamente; pueden dar importante información.
- **Procesamiento paralelo:** Uso coordinado de múltiples procesadores para realizar tareas computacionales. El procesamiento paralelo puede ocurrir en una *computadora con múltiples procesadores* o en una red de estaciones de trabajo o PCs.
- **RAID:** Formación redundante de discos baratos (Redundant Array of Inexpensive Disk). Tecnología para el almacenamiento paralelo eficiente de datos en sistemas de computadoras de alto rendimiento.
- **Regresión lineal:** Técnica estadística utilizada para encontrar la mejor relación lineal que encaja entre una variable seleccionada (dependiente) y sus predicados (variables independientes).
- **Regresión logística:** Una regresión lineal que predice las proporciones de una variable seleccionada categórica, tal como Tipo de Consumidor, en una población.
- **Vecino más cercano:** 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. Algunas veces se llama la técnica del vecino  $k$ -más cercano.
- **SMP Multiprocesador Simétrico (Symmetric Multiprocessor):** Tipo de *computadora con multiprocesadores* en la cual la memoria es compartida entre los procesadores.
- **Terabyte:** Un trillón de bytes.

## **BIBLIOGRAFÍA**

- Trabajo de investigación de la Universidad Tecnológica de Queensland. Australia. Copyright 1997 Lania, Ac (Información de Internet).
- Minería de Datos para Reportes y Documentos. Copyright 2001. Cypress Corporation (información de Internet).
- Trabajo de investigación “Aplicación de Técnicas de Minería de Datos en la construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requisitos de software”; De María N. Moreno García, Luis A. Miguel Quintales, Francisco J. García Peñalvo y José Polo Martín. Universidad de Salamanca, Departamento de Informática y Automática.
- Descubrimientos de Conocimientos en Base de Datos (Información de Internet).
- Curso de Minería de Datos, por José Hernández Orallo, Master de cursos de postgrado del Dsic Universidad Politécnica de Valencia. Información de Internet.
- Revista PC Magazine en Español. “Minería de Datos: los Expedientes Xbox”.