Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias de la Computación

MINERÍA DE DATOS

MC Beatriz Beltrán Martínez

INDICE

INTRODUCCIÓN

Hoy día es cada vez más frecuente la generación masiva de datos económicos, comerciales, científicos, etc., que son almacenados de forma sistemática en grandes bases de datos (data warehouse de entidades financieras y empresas de marketing, bases de datos de experimentos científicos, etc.). Esta cantidad excepcional de información (en algunos casos del orden de Terabytes) dificulta enormemente la realización de cualquier que trate de poner de manifiesto algún aspecto relevante de la misma.

Tradicionalmente la exploración o búsqueda en datos se había realizado a través del análisis estadístico, que consistía en las clásicas rutinas estadísticas como correlación, regresión, etc. A finales de la década de 1980, el análisis estadístico se amplió con técnicas como lógica difusa, razonamiento heurístico y redes neuronales. Pero es necesario mayor poder y eficiencia en la exploración e identificación de datos útiles y en cómo obtenerlos. Actualmente, todas esas técnicas se han aprovechado en generar conocimiento a partir de un conjunto de datos.

Debido a los avances tecnológicos en cuanto a automatización de procesos, almacenamiento, etc., hoy nos encontramos "inundados" de datos. Hay empresas que guardan información como bitácoras de registros de años atrás, datos de clientes, etc. Esa gran cantidad de datos es una gran fuente de información y de conocimiento. El conocimiento es poder, el poder es la habilidad de tener control o influencia en los eventos. El conocimiento es tener una colección de reglas de cómo una cosa afecta a la otra, de acuerdo a la información disponible. La información se obtiene al detectar la utilidad de los datos. Los datos son simplemente registros en forma no mental de observaciones o sucesos.

La disciplina denominada **Minería de Datos** estudia métodos y algoritmos que permiten la extracción automática de información sintetizada que permite caracterizar las relaciones escondidas en la gran cantidad de datos; también se pretende que la información obtenida posea capacidad predictiva, facilitando el análisis de los datos de forma eficiente. Bajo la denominación de "minería de datos" se han agrupado recientemente diversas técnicas estadísticas y del aprendizaje automático (Inteligencia Artificial) enfocadas, principalmente, a la visualización, análisis, y modelización de información de bases de datos masivas.

La idea de obtener información a partir de un conjunto de datos no es nueva. Desde hace siglos, la gente ha tratado de entender cómo extraer datos valiosos o útiles a partir de una colección de datos y luego cómo hacer uso de la información obtenida para lograr un beneficio o un objetivo útil. Ese es también el propósito de la minería de datos.

Muchas de las aplicaciones de minería de datos se hacen sobre datos previamente recolectados, es decir, datos estáticos, que representan un estado en

un tiempo pasado. Los datos no están cambiando mientras están siendo analizados. Los resultados generados serán confiables y consistentes para ese conjunto de datos. El tiempo tomado no es relevante. Por lo anterior no se debe confundir la minería de datos con el monitoreo de datos. Este último opera sobre datos cambiantes y en tiempo real, y trata más que nada de reconocimiento de patrones y no de descubrimiento interactivo

El gran volumen de datos, generado actualmente, ha despertado el interés de los investigadores que buscan aprovechar esta circunstancia para la obtención de información útil. Parte de los esfuerzos se han dirigido al desarrollo de técnicas de extracción de datos y patrones que sirven de apoyo a las tareas de minería de datos y, en general, al proceso de descubrimiento de conocimiento en bases de datos u otras fuentes.

Como parte del proceso de descubrimiento de conocimiento, se requiere la selección de un método y de una técnica de minería de datos apropiada. Además, se debe determinar la estrategia para implementar la técnica de tal forma que sea eficiente teórica y empíricamente

CAPÍTULO I
INTRODUCCIÓN AL PROCESO DE DESCUBRIMIENTO DE CONOCIMIENTO EN BASES DE DATOS (KDD)

1.1 Proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD)

El término **Descubrimiento de Conocimiento en Bases de Datos** (Knowledge Discovery in Databases, o KDD para abreviar) empezó a utilizarse en 1989 para referirse al amplio proceso de búsqueda de conocimiento en bases de datos, y para enfatizar la aplicación a "alto nivel" de métodos específicos de minería de datos. En general, el descubrimiento es un tipo de inducción de conocimiento, no supervisado, que implica dos procesos:

- Búsqueda de regularidades interesantes entre los datos de partida,
- Formulación de leyes que las describan.

Entre la literatura dedicada al tema, se pueden encontrar varias definiciones para descubrimiento:

- ✓ **Descubrimiento** implica observar, recoger datos, formar hipótesis para explicar las observaciones, diseñar experimentos, comprobar la corrección de las hipótesis, comparar nuestros hallazgos con los de otros investigadores y repetir el ciclo. Las computadoras son capaces de observar y recoger datos, a veces mejor que los observadores humanos; los programas estadísticos pueden generar agrupaciones de forma automática entre los datos recogidos; también hay programas con cierta capacidad para diseñar experimentos; y algunos sistemas robóticos realizan las manipulaciones necesarias en ciertos experimentos. Pero ninguna computadora reúne todas estas habilidades ni es capaz de adaptarse para aplicarlas a nuevos problemas; en este sentido, las computadoras no serían capaces de descubrir. Sin embargo, el descubrimiento no requiere realizar simultáneamente todas estas tareas. De igual modo que un investigador puede descubrir nuevo conocimiento a través del análisis de sus datos, una computadora puede examinar los datos disponibles o recogidos por otras computadoras y encontrar relaciones y explicaciones previamente desconocidas, realizando así descubrimiento en un sentido más restringido. La capacidad de las computadoras para realizar búsquedas exhaustivas de forma incansable entre grandes cantidades de datos ofrece buenas expectativas para obtener descubrimiento de forma automática.

- ✓ **Descubrimiento de conocimiento** es la extracción no trivial de información implícita, previamente desconocida y potencialmente útil, a partir de un conjunto de datos. Dado un conjunto de hechos (datos) *H*, un lenguaje *L*, y alguna medida de la certidumbre *C*, definimos una regularidad

(pattern) como una sentencia *S* en *L* que describe relaciones dentro de un subconjunto H_S de *H* con una certidumbre *C*, de forma que *S* es más sencillo que la enumeración de todos los hechos de H_S. Una regularidad que sea interesante y bastante cierta (según criterios definidos por el usuario) se denomina conocimiento. Un sistema de descubrimiento será un programa que toma como entrada el conjunto de hechos y extrae las regularidades existentes. Cuando el conocimiento se extrae partiendo de los datos de una base de datos, se tiene KDD.

Los conceptos de lenguaje, certeza, simplicidad e interés, con los que se define el descubrimiento de conocimiento, son lo suficientemente vagos como para que esta definición cubra una amplia variedad de tendencias. Sin embargo, son ideas fundamentales que diferencian el KDD de otros sistemas de aprendizaje:

- ✓ *Lenguaje de alto nivel:* El conocimiento descubierto se representa en un lenguaje de alto nivel, inteligible desde el punto de vista humano. Por tanto, quedan descartadas, dentro del KDD, representaciones de bajo nivel como las generadas por redes neuronales (a pesar de que éstas son un método válido de minería de datos).

- ✓ *Precisión:* Los descubrimientos representan el contenido de la base de datos que, como reflejo de la realidad, puede contener imperfecciones y ruido. Por tanto, será raro que algún conocimiento se cumpla con todos los datos. El grado de certidumbre medirá el crédito o confianza que el sistema o usuario puede asignar a cierto descubrimiento; si la certeza no es lo suficientemente alta, los patrones descubiertos no llegarán a ser conocimiento.

- ✓ *Interés:* Aunque es posible extraer numerosos patrones de cualquier base de datos, sólo se consideran como conocimiento aquéllos que resulten interesantes según ciertos criterios del usuario. En particular, un patrón interesante debe ser nuevo, potencialmente útil y no trivial.

- ✓ *Eficiencia:* Son deseables procesos de descubrimiento que puedan ser eficientemente implementados en una computadora. Se considera que un algoritmo es eficiente cuando su tiempo de ejecución y el espacio de memoria requerido crecen de forma polinomial con el tamaño de los datos de entrada. No es posible aprender de forma eficiente cualquier concepto booleano (problema NP-completo), sin embargo, sí existen algoritmos eficientes para clases restringidas de conceptos, como los representables en forma conjuntiva, etc. Otra posibilidad es el uso de heurísticas y algoritmos aproximados para la inducción de conocimiento.

Cualquier algoritmo usado en un proceso de KDD debe considerar que conocimiento es una sentencia *S* expresada en un lenguaje *L*, cuyo interés (según criterios del usuario) supera cierto umbral *i* (definido también por el usuario). A su

vez, el interés depende de criterios de certeza, simplicidad y utilidad, establecidos por el usuario. Según se definan los umbrales para estos criterios, se puede enfatizar la búsqueda de información precisa (gran certeza), o útil, etc.

A pesar de que el Descubrimiento de Conocimiento tiene sus inicios en el Aprendizaje Automático o la estadística, hay ciertas componentes que lo hacen muy diferente. En particular, el objetivo fundamental es encontrar conocimiento útil, válido, relevante y nuevo sobre un fenómeno o actividad mediante algoritmos eficientes, dadas las crecientes órdenes de magnitud en los datos.

Al mismo tiempo hay un profundo interés por presentar los resultados de manera visual o al menos de manera que su interpretación sea muy clara. Otro aspecto es que la interacción humano-máquina deberá ser flexible, dinámica y colaborativa. El resultado de la exploración deberá ser interesante y su calidad no debe ser afectada por mayores volúmenes de datos o por ruido en los datos. En este sentido, los algoritmos de descubrimiento de información deben ser altamente robustos.

No cabe duda de que el valor táctico o estratégico de los grandes almacenes de datos está en proporción directa con la capacidad de analizarlos. Dada la gran gama de hipótesis plausibles que se ajustan a los datos, el problema computacional representa un reto hasta ahora poco enfrentado. Sin embargo, estas nuevas condiciones abren un nuevo mundo de oportunidades a la investigación y al desarrollo de nueva tecnología.

El campo del Descubrimiento de Conocimiento en Bases de Datos, denominado Knowledge Discovery in Data Bases en inglés y usualmente abreviado KDD, es la convergencia del Aprendizaje Automático, la Estadística, el Reconocimiento de Patrones, la Inteligencia Artificial, las Bases de Datos, la Visualización de Datos, los Sistemas para el Apoyo a la Toma de Decisiones, la Recuperación de Información, y otros muchos campos

1.2 Proceso KDD

Los principales pasos dentro del proceso interactivo e iterativo del KDD son los siguientes:

1. Desarrollo y entendimiento del dominio de la aplicación, el **conocimiento relevante** y los objetivos del usuario final. Este paso requiere cierta dependencia usuario/analista, pues intervienen factores como: conocer los cuellos de botella del dominio, saber qué partes son susceptibles de un procesado automático y cuáles no, cuáles son los objetivos, los criterios de rendimiento exigibles, para qué se usarán los resultados que se obtengan, compromisos entre simplicidad y precisión del conocimiento extraído, etc.

2. Creación del conjunto de **datos objetivo**, seleccionando el subconjunto de variables o ejemplos sobre los que se realizará el descubrimiento. Esto implica consideraciones sobre la homogeneidad de los datos, su variación a lo largo del tiempo, estrategia de muestreo, grados de libertad, etc.

3. **Preprocesado** de los datos: eliminación de ruido, estrategias para manejar valores ausentes, normalización de los datos, etc.

4. **Transformación y reducción** de los datos. Incluye la búsqueda de características útiles de los datos según sea el objetivo final, la reducción del número de variables y la proyección de los datos sobre espacios de búsqueda en los que sea más fácil encontrar una solución. Este es un paso crítico dentro del proceso global, que requiere un buen conocimiento del problema y una buena intuición, y que, con frecuencia, marca la diferencia entre el éxito o fracaso de la minería de datos.

5. Elección del tipo de sistema para **minería de datos**. Esto depende de sí el objetivo del proceso de KDD es la clasificación, regresión, agrupamiento de conceptos (clustering), detección de desviaciones, etc.

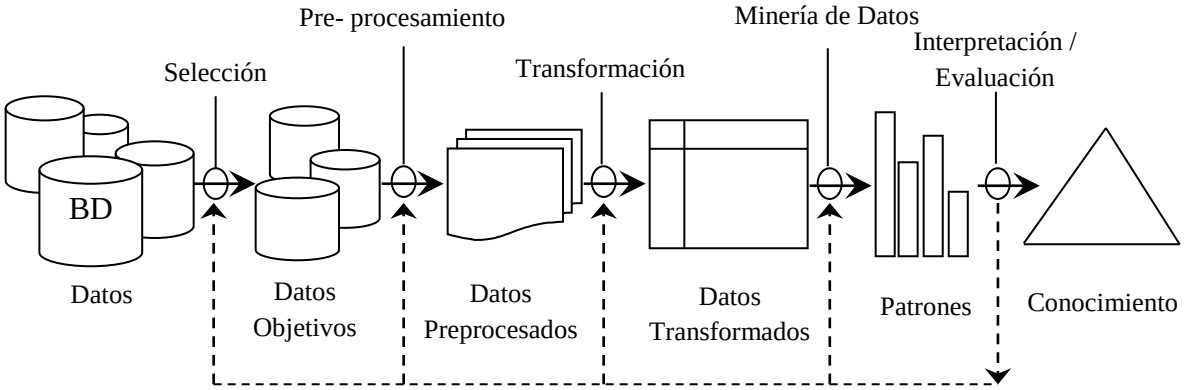


Figura 1. Proceso de KDD

6. Elección del algoritmo de minería de datos.

7. **Minería de datos**. En este paso se realiza la búsqueda de conocimiento con una determinada representación del mismo. El éxito de la minería de datos depende en gran parte de la correcta realización de los pasos previos: por parte del usuario.

8. **Interpretación del conocimiento** extraído, con posibilidad de iterar de nuevo desde el primer paso. La obtención de resultados aceptables dependerá de factores como: definición de medidas del interés del conocimiento (de tipo estadístico, en función de su sencillez, etc.) que permitan filtrarlo de forma automática, existencia de técnicas de visualización para facilitar la valoración de los resultados o búsqueda manual de conocimiento útil entre los resultados obtenidos.

9. Consolidación del **conocimiento descubierto**, incorporándolo al sistema, o simplemente documentándolo y enviándolo a la parte interesada. Este paso incluye la revisión y resolución de posibles inconsistencias con otro conocimiento extraído previamente.

Muchas veces los pasos que constituyen el proceso de KDD no están tan claramente diferenciados. Las interacciones entre las decisiones tomadas en diferentes pasos, así como los parámetros de los métodos utilizados y la forma de representar el problema suelen ser extremadamente complejos. Pequeños cambios en una parte pueden afectar fuertemente al resto del proceso.

Sin quitar importancia a ninguno de estos pasos del proceso KDD, se puede decir que la minería de los datos es la parte fundamental, en la que más esfuerzos se han realizado.

Históricamente, el desarrollo de la estadística nos ha proporcionado métodos para analizar datos y encontrar correlaciones y dependencias entre ellos. Sin embargo, el análisis de datos ha cambiado recientemente y ha adquirido una mayor importancia, debido principalmente a tres factores:

- ✓ *Incremento de la potencia de las computadoras*. Aunque la mayoría de los métodos matemáticos fueron desarrollados durante los años 60 y 70, la potencia de cálculo de las grandes computadoras de aquella época restringía su aplicación a pequeños ejemplos, fuera de los cuales los resultados resultaban demasiado pobres. Algo similar ha ocurrido con la capacidad de almacenamiento de los datos y su coste asociado.

- ✓ *Incremento del ritmo de adquisición de datos*. El crecimiento de la cantidad de datos almacenados se ve favorecido no sólo por el abaratamiento de los discos y sistemas de almacenamiento masivo, sino también por la automatización de muchos experimentos y técnicas de recolección de datos. Se estima que la cantidad de información almacenada en todo el mundo se duplica cada 20 meses; el número y tamaño de las bases de datos probablemente crece más rápidamente. Por ejemplo, se espera que los satélites de observación de la Tierra generen, aproximadamente un petabyte (10^{15} bytes) de datos diariamente, por lo que una persona trabajando 24 horas al día, todos los días del año, a un ritmo de procesamiento de una imagen por segundo, necesitaría varios años para mirar las imágenes generadas en sólo un día.

- ✓ Por último, han surgido *nuevos métodos*, principalmente de aprendizaje y representación de conocimiento, desarrollados por la comunidad de inteligencia artificial, estadística y física de dinámicas no lineales. Estos métodos complementan a las tradicionales técnicas estadísticas en el sentido de que son capaces de inducir relaciones cualitativas generales, o leyes, previamente desconocidas.

Estos nuevos métodos matemáticos y técnicas de software, para el análisis inteligente de datos y búsqueda de regularidades en los mismos, se denominan actualmente técnicas de **minería de datos** o **data mining**. A su vez, la minería de datos ha permitido el rápido desarrollo de lo que se conoce como descubrimiento de conocimiento en bases de datos.

Las técnicas de minería de datos han surgido a partir de sistemas de aprendizaje inductivo en computadoras, siendo la principal diferencia entre ellos los datos sobre los que se realiza la búsqueda de nuevo conocimiento. En el caso tradicional de aprendizaje en computadoras (machine learning), se usa un conjunto de datos pequeño y cuidadosamente seleccionado para entrenar al sistema. Por el contrario, en la minería de datos se parte de una base de datos, generalmente grande, en la que los datos han sido generados y almacenados para propósitos diferentes del aprendizaje con los mismos.

1.3 Fases del KDD

Hay distintas técnicas de las diferentes disciplinas que se utilizan en las diversas fases. Si bien los términos Minería de datos y descubrimiento de conocimiento en bases de datos son usados como sinónimos, el término KDD describe el proceso completo de extracción de conocimiento a partir de los datos. Mientras que Data Mining, se refiere exclusivamente al estadio de descubrimiento de un proceso general KDD.

En este contexto, descubrimiento de conocimiento significa la identificación de relaciones y patrones existenciales en los datos.

Un proceso KDD consiste en la extracción no trivial de conocimiento previamente desconocido y potencialmente útil a partir de un conjunto de datos.

En el proceso KDD es posible definir al menos 6 estados: Recolección de datos, Selección, Limpieza y Transformación de datos, Minería de datos, Evaluación y Validación, Interpretación y Difusión, Actualización y Monitorización.

1.3.1 Recolección de Datos

Las primeras fases del KDD determinan que las fases sucesivas sean capaces de extraer conocimiento válido y útil a partir de la información original. Generalmente, la información que se requiere investigar sobre un cierto dominio de la organización se encuentra:

- ✓ En bases de datos y otras fuentes muy diversas.
- ✓ Tanto internas como externas.
- ✓ Muchas de estas fuentes son las que se utilizan para el trabajo transaccional.

El análisis posterior será mucho más sencillo si la fuente es unificada, accesible (interna) y desconectada del trabajo transaccional. De tal manera que el proceso subsiguiente de minería de datos:

- ✓ Depende mucho de la fuente:
  - OLAP u OLTP.
  - Datawarehouse o copia con el esquema original.
  - ROLAP o MOLAP.

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

- ✓ Recolección de Información Externa. Aparte de información interna de la organización, los almacenes de datos pueden recoger informaciones externas:
  - Demografía (censo), páginas amarillas, psicogeografías, gráficos web, información de otras organizaciones.
  - Datos compartidos en una industria o área de negocio, organizaciones y colegios profesionales, catálogos, etc.
  - Datos resumidos de áreas geográficas, distribución de la competencia, evolución de la economía, información de calendarios y climatológicas, programaciones televisivas, deportivas, catástrofes.
  - Bases de datos externas compradas a otras compañías.

1.3.2 Selección, Limpieza y Transformación de Datos.

Limpieza (data cleansing) y *criba* (selección) de datos:

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

- ✓ Histogramas (detección de datos anómalos).
- ✓ Selección de datos (muestreo, ya sea verticalmente, eliminando atributos u horizontalmente, eliminando tuplas).
- ✓ Redefinición de atributos (agrupación o separación).

Acciones ante datos anómalos (outlíers):

- ✓ *Ignorar*: algunos algoritmos son robustos a datos anómalos (por ejemplo: árboles)

- ✓ *Filtrar* (eliminar o reemplazar) la columna: Solución extrema, pero a veces existe otra columna dependiente con datos de mayor calidad. Preferible a eliminar la columna es reemplazarla por una columna discreta diciendo si el valor era normal o outlier (por encima o por debajo).
- ✓ *Filtrar la fila*: Claramente sesga los datos, porque muchas veces las causas de un dato erróneo están relacionadas con casos o tipos especiales.
- ✓ *Reemplazar el valor*: Por el valor 'nulo' si el algoritmo lo trata bien o por máximos o mínimos, dependiendo por donde es el outlier, o por medias. A veces se puede predecir a partir de otros datos, utilizando cualquier técnica de ML.
- ✓ *Discretizar*: transformar un valor continuo en uno discreto (por ejemplo: muy alto, alto, medio, bajo, muy bajo) hace que los outliers caigan en 'muy alto' o 'muy bajo' sin mayores problemas.

Acciones ante datos faltantes (missing values):

- ✓ *Ignorar*: algunos algoritmos son robustos a datos faltantes (por ejemplo: árboles).
- ✓ *Filtrar* (eliminar o reemplazar) la columna: solución extrema, pero a veces existe otra columna dependiente con datos de mayor calidad. Preferible a eliminar la columna, es reemplazarla por una columna booleana diciendo si el valor existía o no.
- ✓ *Filtrar la fila*: claramente sesga los datos, porque muchas veces las causas de un dato faltante están relacionadas con casos o tipos especiales.
- ✓ *Reemplazar* el valor por medias. A veces se puede predecir a partir de otros datos, utilizando cualquier técnica de ML.
- ✓ *Segmentar*: se segmentan las tuplas por los valores que tienen disponibles. Se obtienen modelos diferentes para cada segmento y luego se combinan.
- ✓ *Modificar* la política de calidad de datos y esperar hasta que los datos faltantes estén disponibles.

Razones sobre datos faltantes (missing values):

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

- ✓ Algunos *valores faltantes* expresan características relevantes: por ejemplo: La falta de teléfono puede representar en muchos casos un deseo de que no se moleste a la persona en cuestión, o un cambio de domicilio reciente.
- ✓ *Valores no existentes:* muchos valores faltantes existen en la realidad, pero otros no. Por ejemplo: el cliente que se acaba de dar de alta no tiene consumo medio de los últimos 12 meses.

- ✓ *Datos incompletos*: si los datos vienen de fuentes diferentes, al combinarlos se suele hacer la unión y no la intersección de campos, con lo que muchos datos faltantes representan que esas tuplas vienen de una(s) fuente(s) diferente(s) al resto.

1.3.3 Minería de Datos.

Características Especiales de los Datos:

Aparte del gran volumen, ¿por qué las técnicas de aprendizaje automático y estadísticas no son directamente aplicables?

- ✓ Los datos residen en el disco. No se pueden escanear múltiples veces.
- ✓ Algunas técnicas de muestreo no son compatibles con algoritmos no incrementales.
- ✓ Muy alta dimensionalidad (muchos campos).
- ✓ Evidencia Positiva.
- ✓ Datos Imperfectos

Aunque algunos se aplican casi directamente, el interés en la investigación en minería de datos está en su adaptación.

Patrones a descubrir:

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

En el primer caso, los propios sistemas de minería de datos se encargan generalmente de elegir el algoritmo más idóneo entre los disponibles para un determinado tipo de patrón a buscar.

1.3.4 Evaluación y Validación.

La fase anterior produce una o más hipótesis de modelos. Para seleccionar y validar estos modelos es necesario el uso de criterios de evaluación de hipótesis. Por ejemplo:

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

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

1.3.5 Interpretación y Difusión.

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

- ✓ El modelo puede requerir implementación (por ejemplo: Tiempo real detección de tarjetas fraudulentas).
- ✓ El modelo es descriptivo y requiere interpretación (por ejemplo: Una caracterización de zonas geográficas según la distribución de los productos vendidos).
- ✓ El modelo puede tener muchos usuarios y necesita difusión: el modelo puede requerir ser expresado de una manera comprensible para ser distribuido en la organización (por ejemplo: Las cervezas y los productos congelados se compran frecuentemente en conjunto → ponerlos en estantes distantes).

1.3.6 Actualización y Monitorización.

Los procesos derivan en un mantenimiento:

- ✓ *Actualización*: Un modelo válido puede dejar de serio: cambio de contexto (económicos, competencia, fuentes de datos, etc.).
- ✓ *Monitorización*: Consiste en ir revalidando el modelo con cierta frecuencia sobre nuevos datos, con el objetivo de detectar si el modelo requiere una actualización.

Producen realimentaciones en el proceso KDD.

CAPÍTULO II
INTRODUCCIÓN A LA MINERÍA DE DATOS

2.1 Evolución Histórica

Aunque los componentes clave del Data Mining ó Minería de datos (DM) existen desde hace décadas en la investigación en áreas como la inteligencia artificial, la estadística o el aprendizaje automático, se puede afirmar que ahora estamos asistiendo al reconocimiento de la madurez de estas técnicas, lo que, junto al espectacular desarrollo de los motores de bases de datos y las herramientas para integración de información justifican su introducción en la esfera empresarial.

Las raíces de la DM se remontan a los años 50. Los departamentos de informática preparaban resúmenes de la información, principalmente de tipo comercial, que se encontraba en los ficheros del ordenador central, con el propósito de facilitar la labor directiva. Así nacieron los sistemas de información para la dirección, que sin embargo, eran voluminosos, poco flexibles, y difíciles de leer para los no informáticos. En los 60 nacen los sistemas gestores de base de datos que aun se mostraban rígidos y carecían de flexibilidad para realizar consultas. Luego aparecieron los motores relacionales resolviendo estos problemas, aunque los informes resultaban muy laboriosos de preparar y depurar, perdiéndose relevancia por su bajo nivel de actualización. Otro grave problema era la diversidad de bases de datos no integradas establecidas por los diferentes departamentos de una organización. Nadie reparaba en la posible utilidad futura de un sistema interdependiente.

El Data Warehouse (DW) viene a solucionar este problema en los finales de los 80. La existencia de DW ha estimulado el desarrollo de los enfoques de DM, en los que las tareas de análisis se automatizan y dan un paso más al posibilitar la extracción de conocimiento inductivo.

Etapa	Cuestión planteada	Tecnologías	Características
Recolección de datos {Años 60}	'Dime mis beneficios totales en los últimos 4 años.'	Ordenadores, cintas, discos.	Retrospectivo, datos estáticos.
Acceso a los datos. {Años 80}	'Ventas en Cataluña durante las últimas Navidades'	Bases de Datos Relacionales (SQL) ODBC	Retrospectivo, datos dinámicos a nivel de registro.
Data Warehouse y soporte a la toma de decisiones. {Años 90}	'Ventas en Andalucia detalle por delegación y descender a nivel tienda.	{OLAP}, bases de datos multi-dimensionales, data warehouse	Retrospectivo, obtención dinámica de datos a múltiples niveles.

Data Mining	Justifica la tendencia de venta en Castilla para el próximo año	Algoritmos avanzados, ordenadores, multiprocesadores, bases de datos masivas.	Prospectivo, obtención proactiva de información.

Tabla 1. Evolución de las tecnologías relacionadas con Data Mining

2.2 ¿Qué es Minería de Datos?

Hace tan solo unos años los datos de las empresas estaban orientados principalmente, a alimentar sus sistemas contables y financieros así como inventarios, los procesos de producción, recursos humanos y ventas.

En la medida que los negocios mundiales se han hecho más competitivos, los datos cada vez cobran más vida y se han convertido en información vital y estratégica para la toma de decisiones. En tal sentido, las empresas han venido evolucionando y han querido agregarle valor a la gran cantidad de información que tienen almacenada en sus bases de datos. Para ello, se han interesado en automatizar los procesos y poder así descubrir información valiosa, que de otra manera seguiría siendo subutilizada o simplemente desperdiciada.

Con la gran velocidad a la cual ha venido evolucionando la tecnología, las empresas de hoy disponen de herramientas de software y hardware cada vez más sofisticadas que posibilitan el almacenamiento de grandes cantidades de información y el análisis de la misma. El avance tecnológico, sumado a la aparición de mercados cada vez más competidos, sugiere a las empresas el mejoramiento continuo de sus esquemas de administración y toma de decisiones, explotando una de las más grandes fuentes de competitividad como lo es la información.

Existen diferentes técnicas que posibilitan la explotación de los datos, extrayendo información que no es detectada a simple vista. Una de estas técnicas es la denominada Minería de Datos, la cual combina técnicas semiautomáticas de inteligencia artificial, análisis estadístico, bases de datos y visualización gráfica, para la obtención de información que no esté representada explícitamente en los datos.

La Minería de Datos descubre relaciones, tendencias, desviaciones, comportamientos atípicos, patrones y trayectorias ocultas, con el propósito de soportar los procesos de toma de decisiones con mayor conocimiento. La Minería de Datos se puede ubicar en el nivel más alto de la evolución de los procesos tecnológicos de análisis de datos.

La Minería de Datos (Data Mining) debe su nombre a. la analogía entre una montaña y la gran cantidad dé datos almacenados en cualquier empresa. Dentro de la montaña, ocultos entre piedras y tierra, se encuentran diamantes de gran valor que mediante actividades de minería son encontrados y aprovechados.

2.2.1 Algunas definiciones de Minería de Datos

La bibliografía es muy amplia a la hora de dar definiciones sobre Data Mining, resaltamos las siguientes:

1. Conjunto de técnicas; que automatizan la detección de patrones relevantes.

2. Proceso que permite transformar información en conocimiento útil para el negocio, a través del descubrimiento y cuantificación de relaciones en una gran base de datos.

3. La técnica denominada como Data Mining puede ser definida como el proceso de extracción de información y patrones de comportamiento que permanecen ocultos entre grandes cantidades de información: Es un proceso iterativo en el que a los avances que se van produciendo en cada paso se les denomina descubrimientos (KDD – Knowledge Discovery in Database)

4. El Data Mining es un proceso que, a través del. descubrimiento y cuantificación de relaciones predictivas en los datos, permite transforma la información disponible en conocimiento útil, Constituye una de .las vías principales de explotación de Data Warehouse.

5. El Data Mining surgió como una integración de múltiples tecnologías tales como la estadística, el soporte a la toma de decisiones, el aprendizaje automático, la gestión y almacenamiento de bases datos y procesamiento en paralelo. Para la realización de estos procesos se aplican técnicas procedentes de muy diversas áreas, como pueden ser los algoritmos genético, las redes neuronales, los árboles de decisión, etc.

6. Se ha llamado minería de datos, al análisis de archivo y bitácoras de transacciones con el fin de descubrir patrones, relaciones, reglas, asociaciones o incluso excepciones que sean útiles para la toma de decisiones.

7. DM es el proceso de descubrimiento significativo de nuevas correlaciones, patrones y tendencias de grandes cantidades datos almacenados en repositorios, utilizando tecnologías de reconocimiento de patrones, así como también técnicas estadísticas y matemáticas.

8. DM es la exploración y análisis, a través de medios automáticos y semiautomáticos, de grandes cantidades de datos con el fin de descubrir patrones y reglas significativos.

9. DM es el uso de herramientas estadísticas avanzadas para investigar en las bases de datos existentes en una compañía, por medio del descubrimiento

de patrones y relaciones que pueden ser explotadas en el contexto de los negocios.

10. DM es la combinación de métodos poderosos que ayudan a la reducción de costos y riesgos, así como también al incremento de las rentas por la extracción de la información estratégica desde los datos disponibles

2.3 ¿Qué no es minería de datos?

Data mining no es estadística

Comúnmente ambos términos se confunden, de hecho, según algunos, Data mining es el sucesor de la estadística tal como se usa actualmente. La estadística y el Data Mining tienen el mismo objetivo, que es construir "modelos" compactos y comprensibles que rindan cuenta de las relaciones establecidas entre la descripción de una situación y un resultado relacionado con dicha descripción. Sin embargo, existen claras diferencias entre ellos:

Fundamentalmente, la diferencia entre ambas consiste 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 del 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 contratar ni ponerse de acuerdo con un estadístico, proporciona gran libertad a los usuarios profesionales.

La Minería de Datos aventaja a la estadística en los siguientes supuestos:

- ✓ Minería de datos utiliza técnicas estadísticas, pero también utiliza técnicas de Inteligencia Artificial.
- ✓ Las técnicas estadísticas se centran generalmente en técnicas confirmatorias, mientras que las técnicas de minería de datos son generalmente exploratorias. Así, cuando el problema al que pretendemos dar respuesta es refutar o confirmar una hipótesis, podemos utilizar ambas ciencias, sin embargo, cuando el objetivo es meramente exploratorio surge la necesidad de delegar parte del conocimiento analítico de la empresa en técnicas de aprendizaje utilizando minería de datos.
- ✓ A mayor dimensionalidad del problema, minería de datos ofrece mejores soluciones. Cuantas más variables entren en el problema, mas difícil resulta encontrar hipótesis de partida interesantes.
- ✓ Las técnicas de minería de datos son menos restrictivas que las estadísticas. Una vez encontrado un punto de partida interesante y dispuestos a utilizar análisis estadístico en particular, puede suceder que los datos no satisfagan los requerimientos del análisis estadístico. Entonces las variables deberán ser examinadas para determinar que tratamiento

permite adecuarlas al análisis, no siendo posible o conveniente en todos los casos. La minería de datos permite ser utilizada con los mínimos de supuestos posibles.

Data mining* no es *OLAP

Las herramientas OLAP permiten navegar rápidamente por los datos, pero no se genera información en el proceso. Se llaman sistemas OLAP (On Line Analytical Processing) a aquellos sistemas que deben:

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

La funcionalidad de los sistemas OLAP se caracteriza por ser un análisis multidimensional de datos mediante navegación del usuario por los mismos de modo asistido.

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

La arquitectura MD-OLAP usa bases de datos multidimensionales, la arquitectura ROLAP implanta OLAP sobre bases de datos relacionales. La arquitectura MDOLAP requiere unos cálculos intensivos de compilación. Lee de datos precompilados, y tiene capacidades limitadas de crear agregaciones *dinámicamente* o de hallar ratios que no se hayan precalculados y almacenados previamente.

La arquitectura ROLAP, accede a los datos almacenados en un Data Warehouse para proporcionar los análisis OLAP. La premisa de los sistemas ROLAP es que las capacidades OLAP se soportan mejor contra las bases de datos relacionales.

Los usuarios finales ejecutan sus análisis multidimensionales a través del motor ROLAP, que transforma dinámicamente sus consultas a consultas SQL. Se ejecutan estas consultas SQL en las bases de datos relacionales, y sus resultados se relacionan mediante tablas cruzadas y conjuntos multidimensionales para devolver los resultados a los usuarios. ROLAP es una arquitectura flexible y general que crece para dar soporte a amplios requerimientos OLAP. El MOLAP es una solución particular, adecuada para soluciones departamentales con unos volúmenes de información y número de dimensiones más modestos

2.4 ¿Qué promete la Minería de Datos?

La tecnología informática es infraestructura fundamental de las grandes organizaciones y permite, hoy en día, registrar con lujo de detalle, los elementos

de todas las actividades con asombrosa facilidad. La tecnología de bases de datos permite almacenar cada transacción, y muchos otros elementos que reflejan la interacción de la organización con todos sus interlocutores, ya sean otras organizaciones, sus clientes, o internamente, sus divisiones, sus empleados, etcétera. Tenemos pues, un registro bastante completo del comportamiento de la organización. Pero, ¿cómo traducir ese voluminoso acervo en experiencia, conocimiento y sabiduría corporativa que apoye efectivamente la toma de decisiones, especialmente al nivel gerencial que dirige el destino de las grandes organizaciones? ¿Cómo comprender el fenómeno, tomando en cuenta grandes volúmenes de datos?

La Minería de Datos ha surgido del potencial del análisis de grandes volúmenes de información, con el fin de obtener resúmenes y conocimiento que apoye la toma de decisiones y que pueda construir una experiencia a partir de los millones de transacciones detalladas que registra una corporación en sus sistemas informáticos.

Por razones que detallaremos más adelante, la Minería de Datos parece ser más efectiva cuando los datos tienen elementos que pueden permitir una interpretación y explicación en concordancia con la experiencia humana. Lo anterior se facilita mucho si estos elementos son el espacio y el tiempo. Afortunadamente, se estima que el 80% de los datos registrados en una base de datos tiene la posibilidad de geo-referenciarse y, el1 00%, de puntualizarse temporalmente. ¿Qué quiere decir esto? En primer lugar, que en la mayoría de los casos es posible asociar un punto en el espacio, un domicilio, unas coordenadas geográficas con la entidad que representa el dato, una fecha o punto en el tiempo. En segundo lugar, que los patrones o inferencias sobre los datos son usualmente interesantes, en la medida en que son patrones en el tiempo o en el espacio. Por ejemplo, qué productos se comercializan mejor en la temporada navideña, en qué regiones es productivo sembrar café, etc.

La tecnología promete analizar con facilidad grandes volúmenes de datos y reconocer patrones en tiempo y espacio que soportarán la toma de decisiones y construirán un conocimiento corporativo de alto nivel. La tecnología de Minería de Datos parece robusta y lista para su aplicación, dado el gran crecimiento de empresas que comercializan software con diferentes técnicas. Más aún, gran parte de estas técnicas son una combinación directa de madurez en tecnología de bases de datos y "data warehousing", con técnicas de aprendizaje automático y de estadística. Sin embargo, la tecnología enfrenta aún varios retos.

2.4.1 Primer Reto de la Minería de Datos

El primero de estos retos, es la facilidad con que se puede caer en una falsa interpretación; para explicarlo, basta reconocer que las primeras y más maduras técnicas para el análisis de datos, con el fin de modelar un fenómeno, provienen de la estadística. Todos sabemos que existe la posibilidad de ser engañados por la estadística; no todos tenemos un sólido entendimiento de las

matemáticas, los supuestos y el modelado para entender a la perfección el riesgo o margen de error en un ejercicio de inferencia estadística, pero todos operamos y funcionamos con resúmenes e indicadores estadísticos generalmente muy simples. Cuando decimos que una gran decisión se basó en la información disponible, típicamente es una serie de promedios y estimadores estadísticos que presentan una generalización de un gran volumen de datos, donde se hace una inferencia.

La estadística es una herramienta poderosa, y es elemento crucial en el análisis de datos. Sin embargo, a veces enfrentamos problemas muy serios en la interpretación de sus resultados. El ejemplo típico es que, usualmente, no recordamos que estos resultados se aplican a grupos (poblaciones) y no a individuos. Estos peligros se ven amplificados en el uso de software de Minería de Datos. Dichas herramientas informáticas pueden poner a disposición de un "analista" (o minero de datos), la posibilidad de crear fácilmente indicadores, resúmenes, gráficas, y aparentes tendencias, sin un verdadero entendimiento de lo que se está reflejando. Es decir, resulta más fácil hacer creíble una falsedad, posiblemente porque la produjo una computadora, con muchas gráficas y con base en muchos datos, eso sí, en un instante.

Así que el reto es doble. ¿Cómo hacer las herramientas de minería de datos accesibles a cualquiera, hasta aquel que no sabe lo más mínimo de estadística, pero que sus resultados e interpretaciones sean válidos? Nótese que es importante que la herramienta tenga un gran elemento de accesibilidad para que su producción sea rentable. Un ejemplo de esto son las bases de datos relacionales, pues su diseño, modelado, y las herramientas alrededor de los manejadores, han hecho posible que no se requiera de una gran especialización para tener una gran cantidad de usuarios y que, por lo tanto, el mercado sea extenso para mantener a los que producen manejadores de datos.

Naturalmente, es importante que las inferencias sean válidas. Esto nos trae a un segundo punto crítico, o segundo reto. ¿Si con la estadística enfrentamos el problema de que es relativamente fácil equivocarse, lo vamos a lograr con la Minería de Datos?

2.4.2 Segundo Reto: ¿Por qué es más fácil equivocarse con la Minería de Datos?

La primera razón es porque, aun con la estadística, el hallar una correlación (estadísticamente significativa) no significa haber encontrado una relación causa – efecto. El contraejemplo clásico lo constituyen los datos anuales de edad, de las personas fallecidas en los Estados Unidos, por estados. Los análisis estadísticos más abundantes encuentran que el estado de Florida tiene, año tras año, la edad promedio más avanzada en que la gente fallece, y con todo el rigor (y significancia estadística) que se desee. ¿Es acaso esto un indicador de que nacer en Florida garantiza longevidad? ¿Se vive más si se muda uno a la península? De ninguna manera; la verdadera explicación es que Florida alberga a una gran cantidad de pensionistas, retirados, etcétera. La gente se va a morir a Florida, pero para

entonces, ya es muy mayor. Si se muere antes de ser pensionista, se muere en su lugar de origen, forzando el promedio de su estado a bajar; si vive mucho, le alcanza para mudarse a Florida y subir allí el promedio.

El software de Minería de Datos está diseñado para hallar correlaciones, para olfatearlas. Su tarea consiste en encontrar aquella proyección de los datos, aquella perspectiva donde aparece una correlación y, lamentablemente, en muchos casos, presentarla como una relación .causa-efecto. Esto es especialmente cierto en los sistemas que generan reglas de asociación, de tal forma:

"SI ESTADO = FLORIDA, ENTONCES, EDAD AL FALLECIMIENTO = ANCIANA"

Esto se deriva de que la Minería de Datos sigue una filosofía muy diferente a como se hace la ciencia. La ciencia, generadora del conocimiento y fundamento de nuestra sorprendente tecnología, opera con base en el método científico. Este método postula que la hipótesis se genera con antelación a la colección de los datos. La Minería de Datos genera hipótesis a partir de los datos. No es catastrófico que se generen hipótesis a partir de los datos. En realidad, el formular creencias a partir de una experiencia finita y limitada es un elemento fundamental del aprendizaje, pero el otro elemento crucial consiste en la revisión de las hipótesis a la luz de nuevos datos y nuevas experiencias.

La Minería de Datos es una herramienta explorativa y no explicativa. Es decir, explora los datos para sugerir hipótesis. Es incorrecto aceptar dichas hipótesis como explicaciones o relaciones causa-efecto. Es necesario coleccionar nuevos datos y validar las hipótesis generadas ante los nuevos datos, y después descartar aquellas que no son confirmadas por los nuevos datos.

Pero la Minería de Datos no puede ser experimental. En muchas circunstancias, no es posible reproducir las condiciones que generaron los datos (especialmente sí son datos del pasado, y una variable es el tiempo).

2.4.3 Tercer Reto: Tiempo, Espacio y Privacidad

La modelación en computadora del tiempo y el espacio son problemas complejos, especialmente para hacer inferencias. Esto hace que las técnicas de, Aprendizaje Automático enfrenten mayores dificultades cuando abordan los temas que parecen más interesantes, de descubrimiento de patrones.

A esto se añaden varios tipos de problemas. El primero, es el de Minería de Datos con relaciones en el tiempo. Es muy posible que se deseen hacer inferencias y análisis de datos sobre un periodo determinado, pero que durante dicho periodo no se haya registrado el mismo número de variables, o que éstas no tengan la misma precisión, o carezcan de la misma interpretación. En ciertos casos puede que se haya hecho un ejercicio de Minería de Datos en el pasado y que los datos se hayan descartado o destruido, pero que se desee hacer una

comparación con datos más recientes. Nótese que un ejercicio de Minería de Datos puede traer a la luz relevancia de variables y factores, pero que sea imposible recopilar estas variables y completar adecuadamente conjuntos de datos del pasado. Otros problemas de análisis de datos con relación al tiempo, son asociados a la granularidad de: los datos con respecto al tiempo. En este sentido, no se conocen todos los datos en el continuo del tiempo. Por ejemplo, si se hacen recopilaciones mensuales, es imposible hacer una predicción semanal.

Finalmente, quisiera tocar el punto de lo que es 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, sin embargo, no existe un marco jurídico que haya mantenido el paso con el avance tecnológico. Esto es, hoy en día, las corporaciones comercializan con millones de perfiles personales, sin que aquellos a que se refieren los datos intercambiados, estén .en posibilidad de intervenir. Cada llamada telefónica, cada transacción bancaria, cada compra en un supermercado, es registrada en una computadora, y si la compañía de teléfonos, el banco y el club de supermercados combinan sus bases de datos, están en condiciones de elaborar un perfil muy completo. Este perfil definiría a más de una persona. Si a esto añadimos qué sitios de WEB visita, qué y dónde se compró con la tarjeta de crédito, etcétera, no existe ninguna privacidad. El problema va desde las definiciones de qué constituye privacidad y quién es el propietario de los datos, hasta qué tanto de un individuo, está en posibilidad real de compilarse.

2.5 Tareas 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 error dentro del modelo.

Los algoritmos de minería de datos realizan en general tareas de predicción (de datos desconocidos) y de descripción (de patrones).

Las tareas principales son:

- ✓ *Agrupamiento* o *Identificación de Clases:* Es una tarea en donde se busca identificar un conjunto de categorías o conjuntos para describir los datos. Las categorías pueden ser mutuamente exclusivas y exhaustivas, o consistir de una representación jerárquica, o permitir solapamientos. Entre los ejemplos que utilizan agrupamiento en aplicaciones de KDD, se incluyen las subpoblaciones homogéneas de consumidores en una base de datos de mercados y la identificación de subcategorías en el espectro de alguna medida. Se utilizan algoritmos de clustering.

- ✓ *Clasificación:* Es la habilidad para adquirir una función que mapee (clasifica) un elemento de dato a una de varias clases predefinidas. Ejemplos de

métodos de Clasificación usados como parte de las aplicaciones de KDD, se encuentran la clasificación de" tendencias en los mercados financieros y la identificación automática de objetos de interés en grandes bases de datos de imágenes.

- ✓ *Condensación o Descripción de Conceptos:* Consiste en encontrar un método que permita encontrar una descripción compacta de un subconjunto de datos. Los métodos más sofisticados involucran reglas de compactación técnicas de visualización multivariada y descubrimiento de relaciones funcionales entre las variables. Estas técnicas son a menudo aplicadas en el análisis datos en forma interactiva y en la generación de reportes automáticos.

- ✓ *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.

- ✓ *Modelado de Dependencias:* Consiste en encontrar un modelo el cual describa las dependencias significantes entre las variables. Los modelos de dependencia existen a dos niveles: A nivel estructural del modelo específica, a menudo en forma gráfica, cuáles variables son localmente dependientes sobre cada una de las otras, mientras que a nivel cuantitativo especifica la robustez de las dependencias, usando una escala numérica. Por ejemplo, las redes de dependencia probabilística usan una independencia condicional para especificar el aspecto estructural del modelo y probabilidades y correlaciones para especificar la solidez de las dependencias. Las redes de dependencia probabilística se encuentran en aplicaciones, tales como: los sistemas expertos probabilístico para bases de datos médicas, recuperación de información y en el modelado de los fenómenos humanos.

- ✓ *Regresión:* Consiste en adquirir una función que mapee un elemento de dato a una variable de predicción de valor real. Existen muchas aplicaciones, como por ejemplo, predecir la cantidad de biomasa presente en un bosque dado, sensado remotamente por vía microondas, estimar la probabilidad de que un paciente no muera, dados los resultados de un conjunto de pruebas de diagnósticos, predecir la demanda de un consumidor por un nuevo producto como una función del gasto publicitario.

2.6 Proceso de Minería de Datos

El proceso de Data Mining se inicia con la identificación de los datos. Para ello hay que imaginar qué datos se necesitan, dónde se pueden encontrar (en una o varias bases de datos; en papel, etc.) y cómo conseguirlos. Una vez que se

dispone de los datos, se deben preparar, poniéndolos en bases de datos en un formato adecuado o construir una warehouse. Esta es una de las tareas más difíciles del Data Mining. Una vez que se tiene los datos en el formato adecuado hay que realizar una selección de los datos esenciales y eliminación de los innecesarios.

Antes de proceder al análisis de los datos por Data Mining, conviene tener una idea de qué es lo que interesa averiguar, qué herramientas se necesitan y cómo proceder. Tras aplicar la herramienta elegida o construida por nosotros mismos hay que saber interpretar los resultados o patrones obtenidos para saber los que son significativos y cómo podarlos para extraer únicamente los resultados útiles. Tras examinar los resultados útiles hay que identificar las acciones que deben de ser tomadas, discutirlas y pensar en los procedimientos para llevarlas a cabo e implementarlas.

Una vez implementadas hay que evaluarlas para ello hay que observar los resultados, los beneficios y el coste para poder reevaluar el procedimiento completo. Para entonces los datos pueden haber cambiado, nuevas herramientas pueden estar disponibles y probablemente habrá que planificar el siguiente ciclo de minería.

La actividad de minería de datos es en realidad un proceso que involucra ajustar modelos o determinar patrones a partir de datos. El ajuste en modelos normalmente es de tipo estadístico, en el sentido que se permite un cierto ruido o error dentro del modelo, en tanto que el establecimiento de patrones es generalmente de tipo determinístico o probabilístico. Los pasos a seguir para la realización de un proyecto de minería de datos son siempre los mismos, independientemente de la técnica específica de extracción de conocimiento usada.

A continuación se proporciona una visión general del proceso de minería de datos y una breve descripción de cada estado:

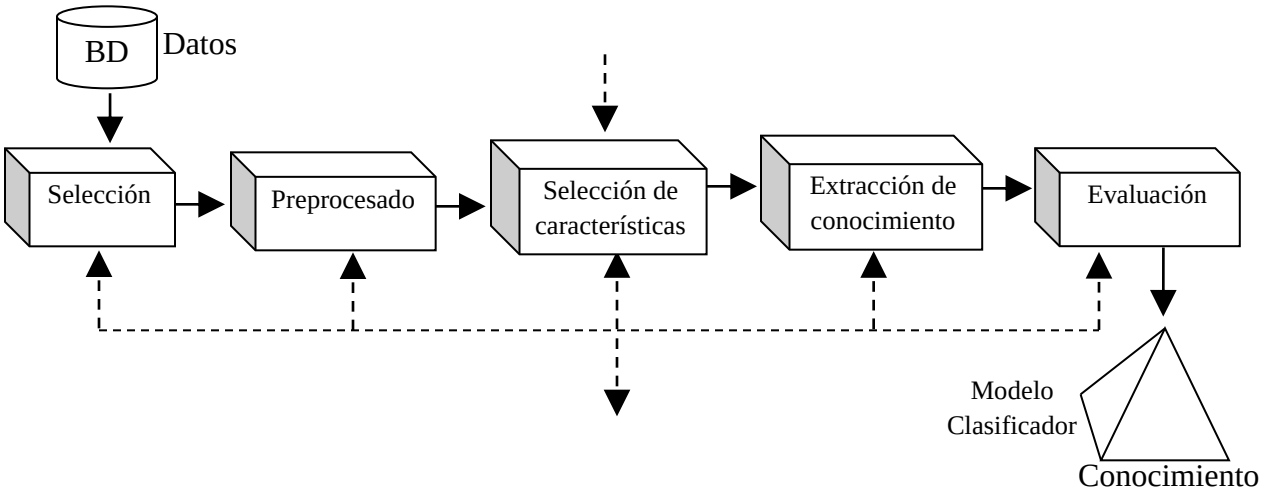


Figura 2. Proceso de la Minería de Datos

El proceso de minería de datos pasa por los siguientes estados:

- ✓ Procesado de los datos.
- ✓ Selección de características.
- ✓ Uso de un algoritmo de extracción de conocimiento
- ✓ Interpretación y evaluación.

2.6.1 Procesado de los Datos

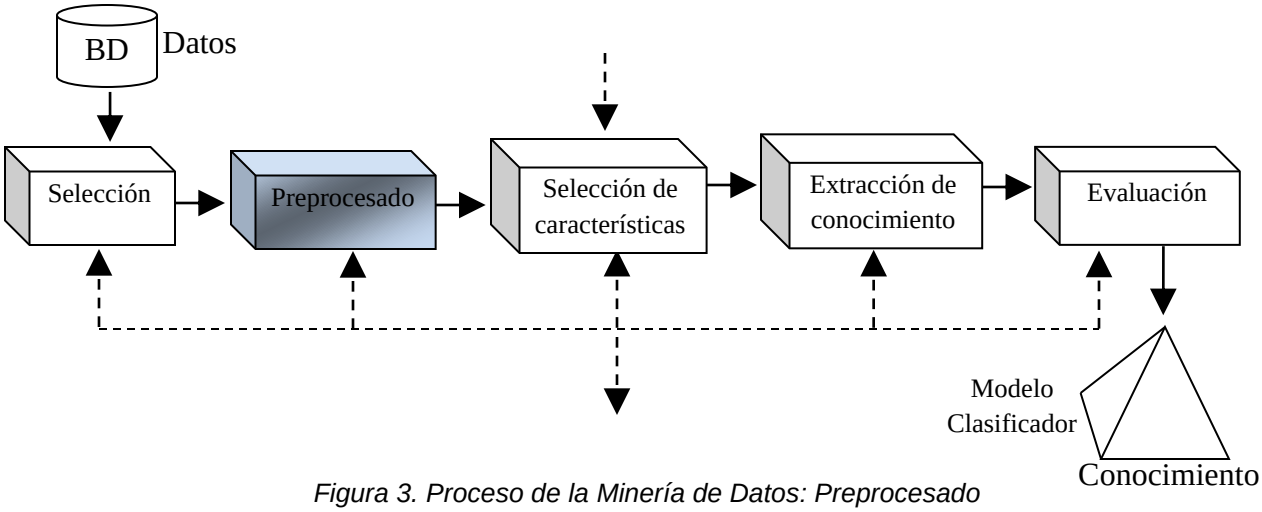


Figura 3. Proceso de la Minería de Datos: Preprocesado

El formato de los datos contenidos en la fuente de datos (base de datos, Data Warehouse, etc.) nunca es el idóneo, y la mayoría de las veces no es posible ni siquiera utilizar ningún algoritmo de minería sobre los datos "en bruto". Mediante el **preprocesado**, se filtran los datos (de forma que se eliminan valores incorrectos, no válidos, desconocidos; según las necesidades y el algoritmo a usar), se obtienen muestras de los mismos (en busca de una mayor velocidad de respuesta del proceso), o se reducen el número de valores posibles (mediante redondeo, clustering, etc.).

2.6.2 Selección de Características

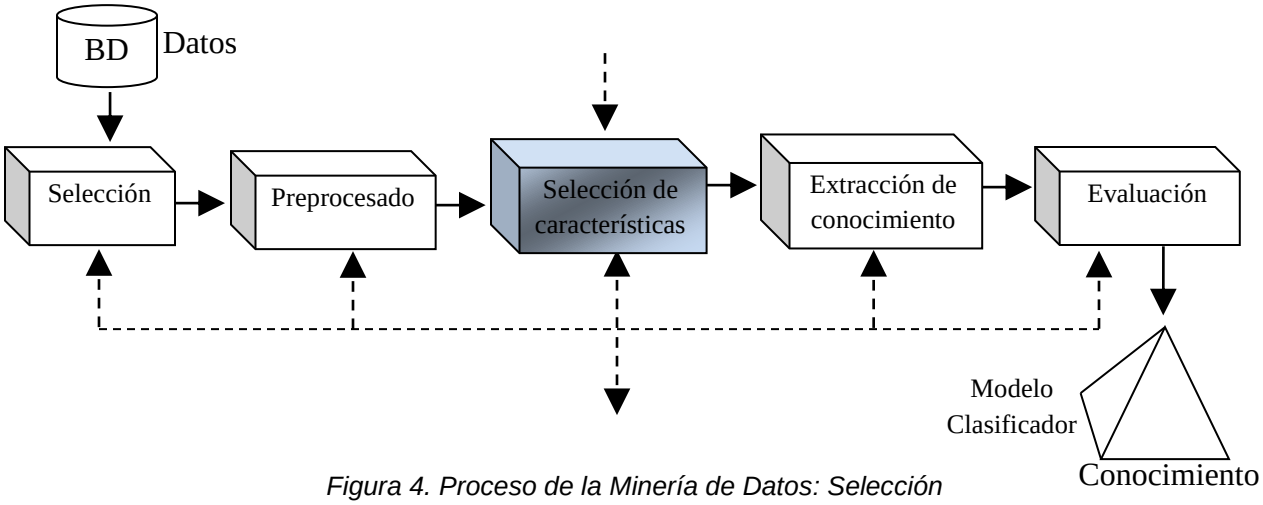


Figura 4. Proceso de la Minería de Datos: Selección

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

Los métodos para la selección de características son básicamente dos:

- ✓ Aquellos basados en la elección de los mejores atributos del problema,
- ✓ Y aquellos que buscan variables independientes mediante tests de sensibilidad, algoritmos de distancia o heurísticos.

2.6.3 Algoritmos de Aprendizaje

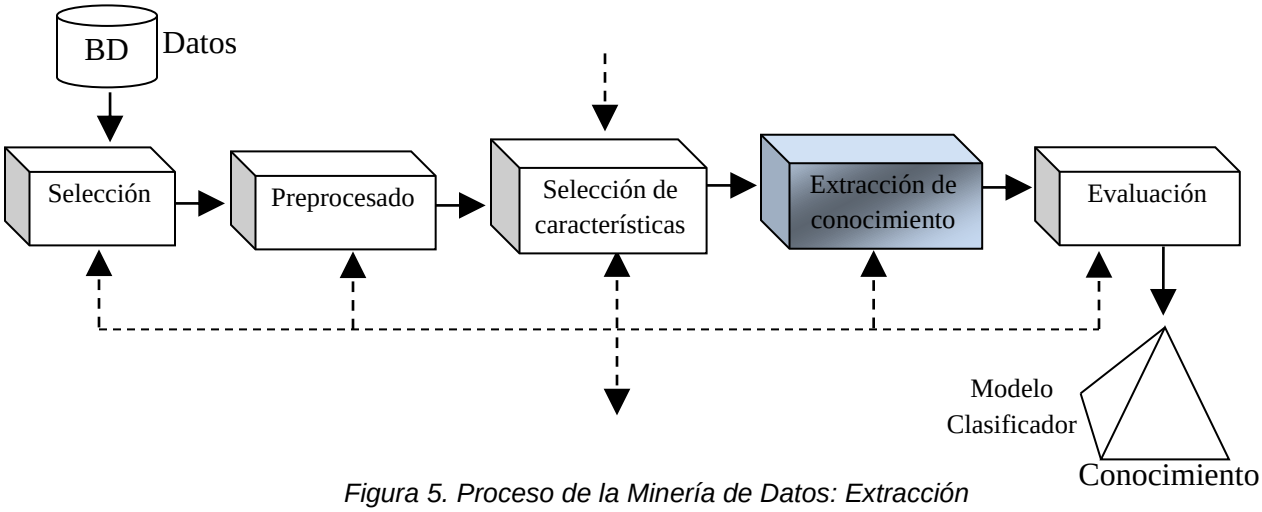


Figura 5. Proceso de la Minería de Datos: Extracción

Mediante una técnica de minería de datos, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un preprocesado diferente de los datos.

2.6.4 Evaluación y Validación

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

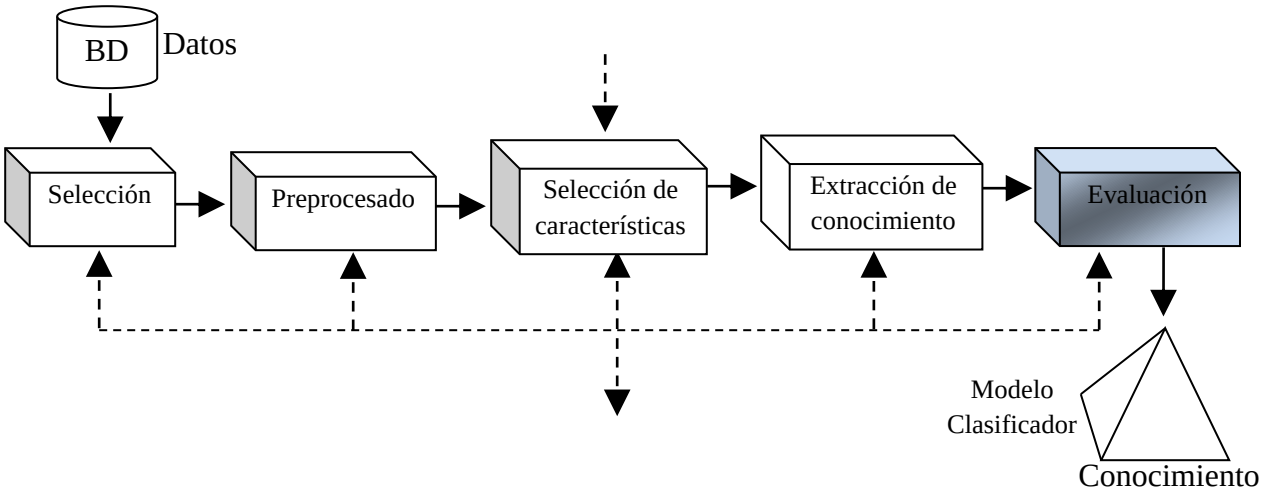


Figura 6. Proceso de la Minería de Datos: Evaluación

La minería de datos se puede definir como un proceso analítico diseñado para explorar grandes cantidades de datos (generalmente datos de negocio y mercado) con el objetivo de detectar patrones de comportamiento consistentes o relaciones entre las diferentes variables para aplicarlos a nuevos conjuntos de datos.

Como se puede apreciar en los estados anteriores, la esencia del problema consiste en "escarbar" en la información almacenada para "descubrir" los elementos de utilidad.

CAPÍTULO III
MÉTODOS Y TÉCNICAS DE MINERÍA DE DATOS

3.1 Tipología de Técnicas de Minería de Datos

Las técnicas de minería de datos crean modelos que son *predictivos* y/ o *descriptivos.* Un modelo predictivo responde preguntas sobre datos futuros. Un modelo descriptivo proporciona información sobre las relaciones entre los datos y sus características.

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 otra serie 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.

Ejemplo de Modelo Predictivo: Queremos saber si jugar o no jugar esta tarde al tenis.

Example	Sky	Temperature	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes

11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Tabla 2. Atributos para el ejemplo "Play Tennis"

- ✓ Pasamos estos ejemplos a un algoritmo de aprendizaje de árboles de decisión, señalando el atributo "Play Tennis" como la clase.

- ✓ El resultado del algoritmo es el siguiente modelo:

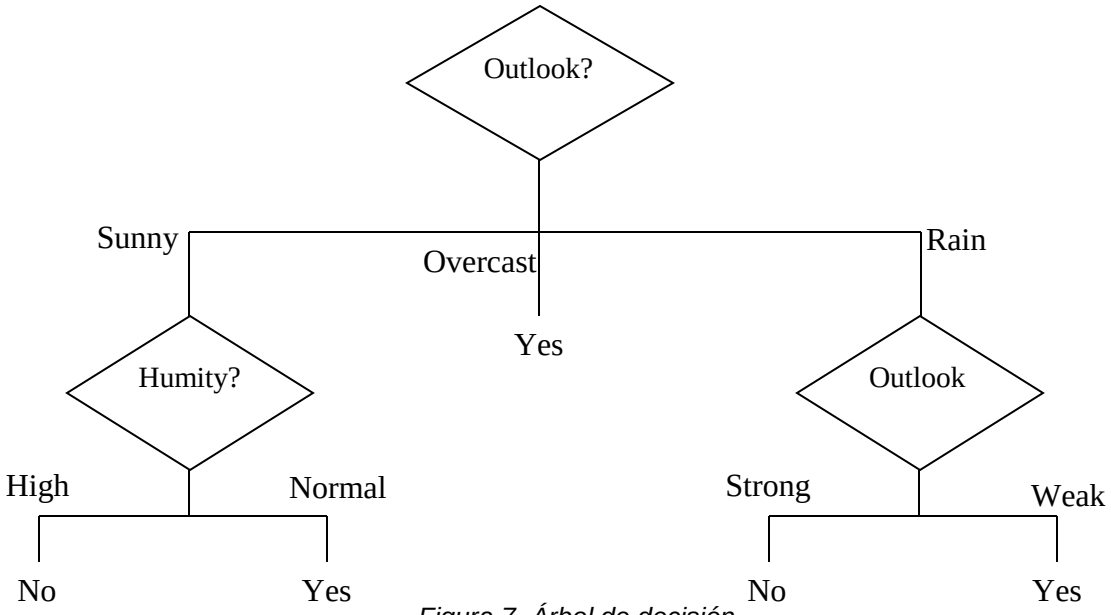


Figura 7. Árbol de decisión

Ahora podemos utilizar este modelo para predecir si esta tarde jugamos o no al tenis.

Por ejemplo, la instancia: (Outlook = sunny, Temperature = hot, Humidity = high, Wind = strong) es No.

#Ej	Sueldo	Casado	Coche	Hijos	Alq/Prop	Sindico	Bajas/Año	Antigüedad	Sexo
1	10000	SI	No	0	Alquiler	No	7	15	H
2	20000	No	SI	1	Alquiler	SI	3	3	M
3	15000	SI	SI	2	Prop	SI	5	10	H
4	30000.	SI	SI	1	Alquiler	No	15	7	M
5	10000	SI	SI	0	Prop	SI	1	6	H
6	40000	No	SI	0	Alquiler	SI	3	16	M
7	25000	No	No	0	Alquiler	SI	0	8	H
8	20000	No	SI	0	Prop	SI	2	6	M
9	20000	SI	SI	3	Prop	No	7	5	H

10	30000	SI	SI	2	Prop	No	1	20	H
11	50000	No	No	0	Alquiler	No	2	12	M
12	8000	SI	SI	2	Prop	No	3	1	H
13	20000	No	No	0	Alquiler	No	27	5	M
14	10000	No	SI	0	Alquiler	SI	0	7	H
15	8000	No	SI	0	Alquiler	No	3	2	H

Tabla 3. Categorías de Empleados

Ejemplo de Modelo Descriptivo: Queremos categorizar nuestros empleados. Tenemos estos datos de los empleados:

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

Cluster 1:	5 examples
Sueldo:	226000
Casado: No → 0.8	
	Si → 0.2
Coche:	No → 0.8
	Si → 0.2
Hijos:	0
Alq/Prop:	Alquiler → 1.0
Sindic:	No → 0.8
	Si → 0.2
Baja / Año:	8
Antigüedad:	8
Sexo:	H → 0.6
	M →0.4

Cluster 2:	4 examples
Sueldo:	225000
Casado: No → 1.0	
Coche:	Si → 1.0
Hijos:	0
Alq/Prop:	Alquiler → 0.75
	Prop → 0.25
Sindic:	Si → 1.0
Baja / Año:	2
Antigüedad:	8
Sexo:	H → 0.25
	M →0.75

Cluster 3:	6 examples
Sueldo:	188333
Casado:	Si → 1.0
Coche:	Si → 1.0
Hijos:	2
Alq/Prop:	Alquiler → 0.2
	Prop → 0.8
Sindic:	No → 0.67
	Si → 0.33
Baja / Año:	5
Antigüedad:	8
Sexo:	H → 0.83
	M →0.0.17

Tabla 4. Cluster generados

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

3.2 Taxonomía de las Técnicas de Minería de Datos

Clasificación de las técnicas de aprendizaje

Cualquier problema de aprendizaje inductivo se puede presentar (más o menos directamente) de cualquiera de estas cuatro formas:

- ✓ ***Interpolación*:** una función continua sobre varias dimensiones.

- ✓ ***Predicción secuencial:*** las observaciones están ordenadas secuencialmente. Se predice el siguiente valor de la secuencia. Caso particular de interpolación con 2 dimensiones, una discreta y regular.

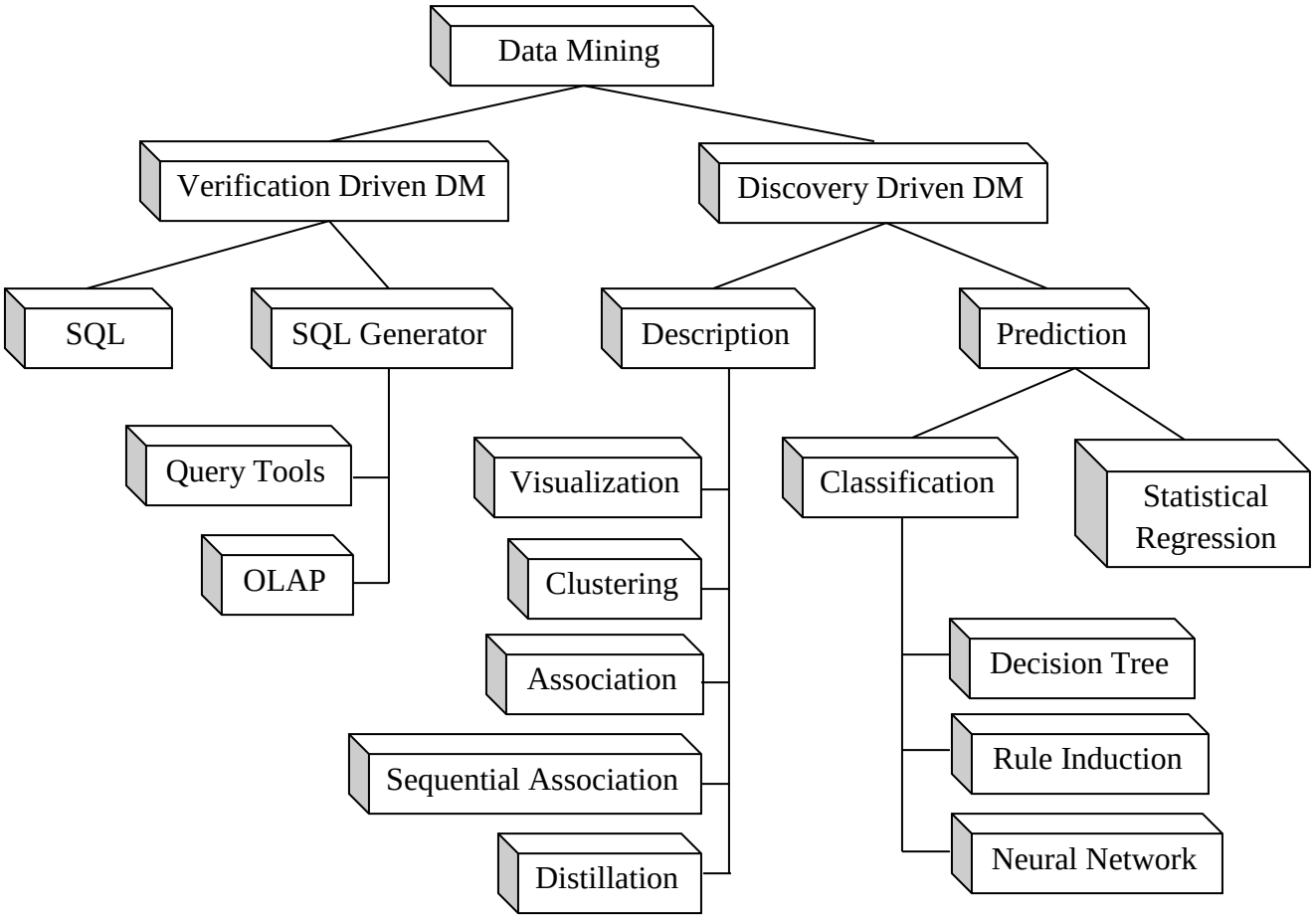


Figura 8. Taxonomía de las técnicas de DM

- ✓ ***Aprendizaje supervisado:*** cada observación incluye un valor de la clase a la que corresponde. Se aprende un clasificador. Caso particular de interpolación: la clase es discreta.

- ✓ ***Aprendizaje no supervisado:*** el conjunto de observaciones no tiene clases asociadas. El objetivo es detectar regularidades en los datos de cualquier tipo: agrupaciones, contornos, asociaciones, valores anómalos.

- ✓ ***Abducción* o *Aprendizaje Analítico:*** *B* es muy importante. El objetivo es explicar la evidencia respecto a *B.*

A continuación haremos una clasificación de los algoritmos Predictivos y Descriptivos:

Predictivo

- ✓ Interpolación y Predicción Secuencial.

  Generalmente las mismas técnicas:

  - Datos continuos (reales):
    Regresión Lineal:
    Regresión lineal global (clásica).
    Regresión lineal ponderada localmente.
    Regresión No Lineal: logarítmica, pick & mix, etc.

  - Datos discretos:
    No hay técnicas específicas: se suelen utilizar técnicas de algoritmos genéticos o algoritmos de enumeración refinados.

- ✓ Aprendizaje supervisado.

  Dependiendo de sí se estima una función o una correspondencia:

  - Clasificación: se estima una función (las clases son disjuntas).
  - Categorización: se estima una correspondencia (las clases pueden solapar).

  Dependiendo del número y tipo de clases:

  - Clase discreta: se conoce como "clasificación".
    Ejemplo: Determinar el grupo sanguíneo a partir de los grupos sanguíneos de los padres.

  - Si sólo tiene dos valores (V y F) se conoce como "concept learning".
    Ejemplo: Determinar si un compuesto químico es cancerígeno.

  - Clase continua o discreta ordenada: se conoce como "estimación".
    Ejemplo: Estimar el número de hijos de una familia a partir de otros ejemplos de familias.

- ✓ Aprendizaje supervisado (Clasificación).

  Técnicas:

  - K – NN (Nearest Neighbor).
  - K – means (competitive learning).
  - Perceptron Learning.
  - Multilayer ANN methods (e. g. backpropagation).
  - Radial Basis Functions.
  - Decision Tree Leaming (e. g. 103, C4.5, CART).
  - Bayes Classifiers.
  - Center Splitting Methods.

- Rules (CN2).
- Pseudo- relational: Supercharging, Pick- and- Mix.
- Relational: ILP, IFLP, SCIL.

Descriptivo

✓ Análisis Exploratorio

Técnicas:

- Estudios correlacionales.
- Dependencias.
- Detección datos anómalos.
- Análisis de dispersión.

✓ Segmentación (Aprendizaje no supervisado)

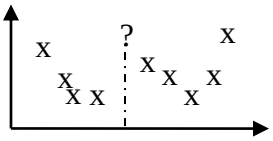
Técnicas de clustering:

- K – means (competitive learning).
- redes neuronales de Kohonen
- EM (Estimated Means) (Dempster 1977).
- Cobweb (Fisher 1987).
- AUTOCLASS

Ejemplos:

Predictivos

- ***Interpolación:***



f(2,2) = ?

- ***Predicción Secuencial:*** 1, 2, 3, 5, 7, 11, 13, 17, 19, … ?
- ***Aprendizaje Supervisado:***

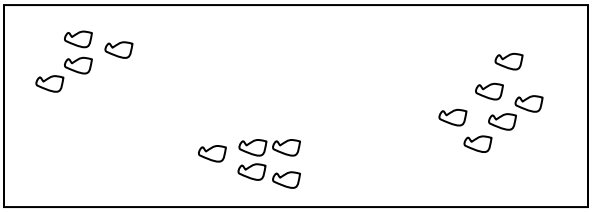
1 3 - > 4.

3 5 - > 8. 4 2 - > ?

7 2 - > 9.

Descriptivos

- ***Segmentación (Aprendizaje no supervisados):***



¿Cuántos grupos hay?
¿Qué grupos formo?

- ***Análisis Exploratorio:*** Correlaciones, Asociaciones y Dependencia

3.2.1 Técnicas No Supervisadas y Descriptivas.

Métodos Descriptivos

- ✓ Correlación y Asociaciones (análisis exploratorio o links analysis):

  Coeficiente de correlación:

$$Cor(\bar{x}, \bar{y}) = \frac{Cov(\bar{x}, \bar{y})}{\sigma_x \cdot \sigma_y}$$

  Donde

$$Cov(\bar{x}, \bar{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

  Asociaciones (cuando los atributos son discretos).

  Ejemplo: tabaquismo y alcoholismo están asociados.

  Dependencias funcionales: asociación unidireccional.

  Ejemplo: el nivel de riesgo de enfermedades cardiovasculares depende del tabaquismo y alcoholismo (entre otras cosas).

- ✓ Correlaciones y Estudios Factoriales

  Permiten establecer relevancia! irrelevancia de factores y si aquélla es positiva o negativa respecto a otro factor o variable a estudiar.

  Ejemplo (Kiel 2000): Estudio de visitas: 11 pacientes, 7 factores:

  - *Health:* salud del paciente (referida a la capacidad de ir a la consulta). (1- 10)
  - *Need:* convicción del paciente que la visita es importante. (1- 10)
  - *Transportation:* disponibilidad de transporte del paciente al centro. (1- 10)
  - *Child Care:* disponibilidad de dejar los niños a cuidado. (1- 10)
  - *Sick Time:* si el paciente está trabajando, puede darse de baja. (1- 10)
  - *Satisfaction:* satisfacción del cliente con su médico. (1- 10)
  - *Ease:* facilidad del centro para concertar cita y eficiencia de la misma. (1- 10)
  - *No – Show:* indica si el paciente no se ha pasado por el médico durante el último año (O se ha pasado, 1 no se ha pasado).

Matriz de correlaciones:

	Health	Need	Transportation	Child Care	Sick Time	Satisfaction	Ease	No – Show
Health	1							
Need	-0.7378	1						
Transportation	0.3116	-0.1041	1					
Child Care	0.3116	-0.104 1	1	1				
Sick Time	0.2771	0.0602	0.6228	0.6228	1			
Satisfaction	0.22008	-0.1337	0.6538	0.6538	0.6257	1		
Ease	0.3887	-0.0334	0.6504	0.6504	0.6588	0.8964	1	
No – Show	0.3955	-0.5416	-0.5031	-0.5031	-0.7249	-0.3988	-0.3278	1

Coeficientes de Regresión:

Independent Variable	Coefficient
Health	.6434
Need	.0445
Transportaltion	-.2391
Child Care	-.0599
Sick Time	-.7584
Satisfaction	.3537
Ease	-.0786

Indica que un incremento de 1 en el factor Health aumenta la probabilidad de que no aparezca el paciente en un 64.34%

✓ Reglas de Asociación y Dependencia de Valor:

La terminología no es muy coherente en este campo (Fawad, por Ejemplo, suele llamar asociaciones a todo y regla de asociación a las dependencias).

- Asociaciones: Se buscan asociaciones de la siguiente forma:

$$(X_1 = a) \sqsubset (X_4 = b)$$

De los n casos de la tabla, que las dos comparaciones sean verdaderas o *falsas será cierto en rc casos:* Un parámetro T_c (confidence): T_c = *certeza de la regla* = *rc/n.*

Si consideramos valores nulos, tenemos también un número de casos en los que se aplica satisfactoriamente (diferente de T_c) y denominado T_s.

- *Dependencias de Valor:* Se buscan dependencias de la siguiente forma (if *Ante* then *Cons):*

Por ejemplo: If $(X_1 = a,\ X_3 = c,\ X_5 = d)\text{then}(X_4 = b, X_2 = a)$

De los n casos de la tabla, el antecedente se puede hacer cierto en r a casos y *de estos en rc casos se hace también el consecuente, tenemos:*

Dos parámetros T_c (confidence/ accuracy) y T_s (support):

T_c = certeza de la regla = rc/ra, fuerza o *confianza P* (*Cons I Ante)*
T_s = mínimo número de casos o *porcentaje en los que se aplica satisfactoriamente (rc o rc/n respectivamente).*
Llamado también prevalencia: P (cons ⊏ *Ante).*

Ejemplo:

DNl	Renta Familiar	Ciudad	Profesión	Edad	Hijos	Obeso	Casado
11251545	5.000.000	Barcelona	Ejecutivo	45	3	S	S
30512526	1.000.000	Melilla	Abogado	25	0	S	N
22451616	3.000.000	León	Ejecutivo	35	2	S	S
25152516	2.000.000	Valencia	Camarero	30	0	S	S
23525251	1.500.000	Benidorm	Animador	30	0	N	N
			Parque				
			Temático				

Tabla 5. Ejemplo de personas y sus características

Asociaciones:

Casado e (Hijos > 0) están asociados (67%, 2 casos). Obeso y casado están asociados (75%, 3 casos)

Dependencias:

(Hijos > 0) Æ Casado (100%, 2 casos). Casado Æ Obeso (100%, 3 casos).

Complejidad de los algoritmos de asociaciones y dependencias:

Temporal: bajo ciertas condiciones de dispersión y para atributos discretos se pueden encontrar en casi tiempo lineal.

✓ Algoritmos de búsqueda de asociaciones y dependencias.

La mayoría se basa en descomponer el problema en dos fases:

- FASE A: BÚSQUEDA DE "LARGE ITEMSETS". Se buscan conjuntos de atributos con 'support' >= al support deseado, llamados 'large itemsets' (conjuntos de atributos grandes). De momento no se busca separarlos en parte izquierda y parte derecha.

- FASE B: ESCLARECIMIENTO DE DEPENDENCIAS (REGLAS). Se hacen particiones binarias y disjuntas de los itemsets y se calcula la confianza de cada uno. Se retienen aquellas reglas que tienen confianza >= a la confianza deseada.

- ✓ Otros tipos de asociaciones:

  Asociaciones entre jerarquías. Si existen jerarquías entre los ítems (por ejemplo: Las familias de productos de un comercio o de un supermercado) a veces sólo es interesante buscar asociaciones inter – jerarquía y no intra – jerarquía. Esto puede reducir mucho el espacio de búsqueda.

  Asociaciones negativas. A veces nos interesa conocer asociaciones negativas, por ejemplo: "80% de los clientes que compran pizzas congeladas no compran lentejas". El problema es mucho más difícil en general, porque, cuando hay muchos ítems, existen muchas más combinaciones que no se dan que las que se dan.

  Asociaciones con valores no binarios yl o continuos. Se deben binarizar. Por ejemplo: Si se tiene un atributo *a* con *k* posibles valores v_1, ..., v_k (k > 2) se sustituye por *k* atributos con la condición ($a = v_i$). Con los atributos continuos se discretizan en rangos (0 – 5, 6 – 10, 11 – 15, ...) y luego se hace el mismo procedimiento.

  Asociaciones relacionales.

- ✓ Métodos Representativos**:**

  - AprioriAII
  - AprioriSome
  - DynamicSome

  Problema: los usuarios quieren especificar restricciones sobre el tiempo máximo y mínimo entre eventos secuenciales.

  Extensiones:

  - Minería de patrones secuenciales con restricciones.

- ✓ Dependencias Funcionales:

$$A \wedge B \wedge C \odot D$$

  Significa: para los mismos valores de *A*, *B* y *C* tenemos un solo valor de *D*. Es decir *D* es función de *A*, *B* y *C*.

  Si representamos la parte izquierda como un conjunto de condiciones, podemos establecer una relación de orden entre las dependencias funcionales. Esto genera un semi- retículo.

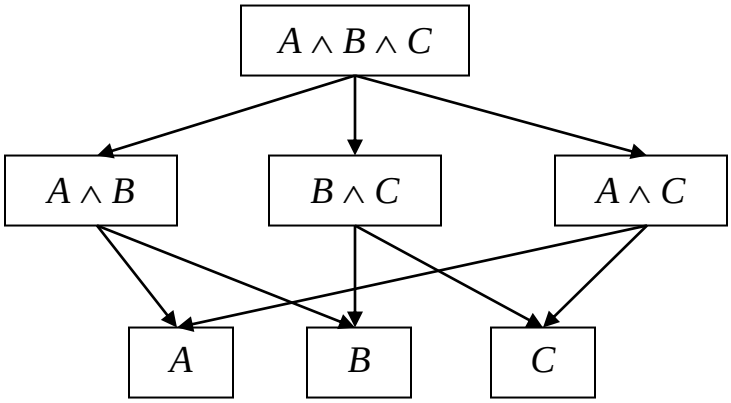


Figura 9. Semi – retículo

La búsqueda se realiza en este retículo. El *coste exponencial* FDEP incluye:

a) simple top- down algorithm,
b) bottom- up algorithm, and
c) bi- directional algorithm.

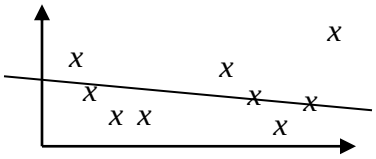
3.2.2 Técnicas supervisadas y predictivas.

Métodos Predictivos

- ✓ Interpolación y Predicción Secuencial.
  - Regresión Lineal Global.

Se buscan los coeficientes de una función lineal

$$\hat{f}(x) = w_0 + w_1 x_1 + \ ... + w_n x_n$$



Una manera fácil (sí es lineal simple, sólo dos dimensiones *x* e *y*):

$$w_1 = \frac{n(\sum xy)(\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \qquad w_0 = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

Obteniendo $y = w_0 + w_1 x$

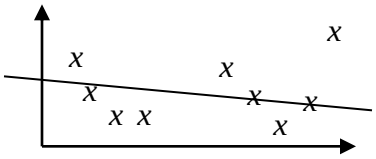
Error típico de una regresión lineal simple:

$$E_{típico} = \sqrt{\left[\frac{1}{n(n-2)}\right]\left[n\left(\sum y^2\right) - \left(\sum y\right)^2 - \frac{[(n\sum xy) - (\sum x)(\sum y)]^2}{n(\sum x^2) - (\sum x)^2}\right]}$$

- Regresión Lineal Global por Gradient Descent.

Una manera usual es utilizando "gradient descenf'. Se intenta minimizar la suma de cuadrados:

$$E = \frac{1}{2}\sum_{x \in D} (f(x) - \hat{f}(x))^2$$



Derivando:

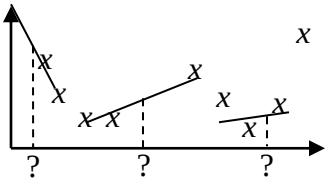
$$\Delta w_j = r \cdot \sum_{x \epsilon D} (f(x) - \hat{f}(x)) x_j$$

Iterativamente se van ajustando los coeficientes y reduciendo el error.

- Regresión Lineal Ponderada Localmente.

La función lineal se aproxima para cada punto x_q a interpolar:

$$\hat{f}(x) = w_0 + w_1 x_1 + \ \ldots + w_m x_m$$



Se intenta minimizar la suma de cuadrados de los *k* más cercanos

$$E = \frac{1}{2} \sum_{x \in [\text{los } k \text{ puntos más cercanos}]} \left(f(x) - \hat{f}(x)\right)^2 K(d(x_q, x))$$

donde $d(\cdot,\cdot)$ es una distancia y *K* es una función que disminuye con la distancia (una función Kernel), por ejemplo $1/ d^2$.

Gradient descent:

$$\Delta w_j = r \cdot \sum_{x \in [\text{los } k \text{ puntos más cercanos}]} (f(x) - \hat{f}(x)) \cdot K(d(x_q, x)) \cdot x_j$$

A mayor *K* más global, a menor *K* más local.

- Regresión Adaptativa.

Son casos particulares de regresión local, en el que se supone un orden y se utiliza preferentemente para predecir futuros valores de una serie:

Muy utilizada en compresión de sonido y de vídeo, en redes, etc. (se predicen las siguientes tramas).

Algoritmos mucho más sofisticados (cadenas de Markov), Algoritmo MARS (Multiple Adaptive Regression Splines).

- Regresión No Lineal.

Estimación Logarítmica (se sustituye la función a obtener por *y = ln(f)*):

$$y = w_0 + w_1 x_1 + \ ... + \ w_m x_m$$

Se hace regresión lineal para calcular los coeficientes y a la hora de predecir se calcula la $f = ey$.

- Pick and Mix – Supercharging

Se añaden dimensiones, combinando las dadas. Por ejemplo: Si tenemos cuatro dimensiones: x_1, x_2, x_3 (además de *y*) podemos definir $x_4=x_1 \cdot x_2$, $x_5=x_3^2$, $x_6=x_1^{x_2}$ y obtener una función lineal de x_1, x_2, x_3, x_4, x_5, x_6.

✓ Aprendizaje supervisado.

- k- NN (Nearest Neighbour):

1. Se miran los *k* casos más cercanos.
2. Si todos son de la misma clase, el nuevo caso se clasifica en esa clase.
3. Si no, se calcula la distancia media por clase o se asigna a la clase con más elementos.

El valor de *k* se suele determinar heurísticamente.

- (On-line) k- means clustering:

Aunque lo vimos como una técnica no supervisada, también se puede utilizar para aprendizaje supervisado, si se utiliza convenientemente

Elegir un *k* mayor que el número de clases pero no mucho mayor.

- Perceptron Leaming:

Computan una función lineal para cada y_i es:

$$y'_j = \sum_{i=1}^{n} w_{i,j} \cdot x_i$$

- Perceptron Learning (Gradient Descent).

El algoritmo Gradiente Descendiente ajusta así:

1. Se asigna a los $W_{i,j}$ un valor pequeño aleatorio entre 0 y 1.
2. Hasta que la condición de terminación se cumple, hacer:
3. Para todos los *p* ejemplos (x_k, y_k) *t* se calcula la matriz de error ($et_k = yt_k - y't_k$)
4. Se recalculan los pesos siguiendo Least – Mean Square (LMS), con un rango de aprendizaje (*r*):
$$w_{i\cdot j}^{t+1} = w_{i\cdot j}^{t} + \sum_{k:1..p} r(x_{i\cdot k}^{t} \cdot e_{j\cdot k}^{t})$$
5. *t* := *t* + *1*, ir a 2.

r es un valor generalmente pequeño (0.05) y se determina heurísticamente. A mayor *r* converge más rápido pero puede perderse en valles locales.

El algoritmo Perceptron (versión *incremental* o *aproximación estocástica al Gradiente Descendiente):*

1. Se asignan aleatoriamente los $w_{i,j}$ entre 0 y 1 (o se pone 0.5)
2. *t* = 1 (se toma el primer ejemplo).
3. Para el ejemplo (*x,* y) *t* se calcula el vector error (*et* = *yt – y't*)
4. Se recalculan los pesos siguiendo Least – Mean Squares (LMS), llamada regla delta, Adaline o *Widrow- Hoff:*
$$w_{i\cdot j}^{t+1} = w_{i\cdot j}^{t} + r(x_{i}^{t} \cdot e_{j}^{t})$$
5. *t* : = *t* + 1, ir a 2 hasta que no queden ejemplos o el error medio se ha reducido a un valor deseado.

En general, esta versión es más eficiente que la anterior y evita algunos mínimos locales.

- Multilayer Perceptron (redes neuronales artificiales multicapas, ANN)

El **perceptron** de una capa no es capaz de aprender las funciones más sencillas. Se añaden capas internas, se introducen diferentes funciones de activación e incluso recientemente se introducen bucles y retardos.

En el caso más sencillo, con la función de activación *sgn*, el número de unidades internas *k* define exactamente el número de límites que la función global puede calcular por cada salida.

El valor de *k* se suele determinar heurísticamente. Pero, ¿cómo entrenar este tipo de red?

Para poder extender el Gradiente Descendiente necesitamos una función

de activación continua. La más usual es la función sigmoidal:

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

Esto permite particiones no lineales.

- Radial- Basis Function (Clustering Method + LMS).

PRIMER PASO: Algoritmo Clustering:

1. Dividir aleatoriamente los ejemplos en *k* conjuntos y calcular la media (el punto medio) de cada conjunto.
2. Reasignar cada ejemplo al conjunto. Con punto medio más cercano.
3. Calcular los puntos medios de los *k* conjuntos.
4. Repetir los pasos 2 y 3 hasta que los conjuntos no varíen.

SEGUNDO PASO: Recodificar los ejemplos como distancias a los centros y normalizar (cada ejemplo pasa a ser un vector de *k* elementos).

TERCER PASO: Con un perceptron de *k* elementos de entrada y una salida, aplicar el algoritmo visto antes.

- Métodos pseudo – relacionales

Una manera de poder abordar problemas con poca separabilidad es transformar los datos, mediante recodificaciones o aumento de la dimensionalidad.

> *Super – charging:* aumentar la dimensionalidad separando todos los atributos en dígitos binarios.
>
> *Pick and Mix:* crear nuevos atributos como combinación de los ya existentes.
>
> *SCIL*: iterar múltiples veces un método clásico fence & fill seguido de recodificación.

Pero estos métodos pseudo – relacionales no son capaces de:

> Detectar relaciones entre varios ejemplos o entre partes complejas del mismo ejemplo.
>
> Aprender funciones recursivas.

- Aprendizaje Recursivo

Modelos Estructurales de Grammar Learning:

Es uno de los campos más antiguos de ML: (las frases gramaticales son de la clase true).

Aprendizaje de autómatas aceptadores de gramáticas.

Gramáticas regulares estocásticas.

Lenguajes probabilísticos: cadenas de Markov, algoritmo de Viterbi.

- Aprendizaje Relacional y Recursivo

IFP (Inductive Functional Programming): Se aprenden reglas de la forma: $g(f(a), X) \sqsubset b$.
Existen aproximaciones con LISP, el lenguaje ML y otros (70s).

ILP (Inductive Logic Programming). El lenguaje representacional es cláusulas de Horn: *p(x, Y, b) : - q(f(x, Y), c)*.
Inicio en los 80 (Shapiro) y gran desarrollo en la década de los 90.

IFLP (Inductive Functional Logic Programming): $g(f(a), X) \sqsubset b$ *: - p(x, b) = true*, *q(x, X) = a*.
Mayor naturalidad y expresividad. Ventaja con problemas de clasificación.

✓ Combinación de hipótesis.

- BOOSTING:

Se utiliza el MISMO algoritmo para aprender distintas hipótesis sobre distintas particiones de los datos.

Luego se combinan las distintas hipótesis.

- VOTING / ARBITER / COMBINER:

Se utiliza DISTINTOS algoritmos para aprender distintas hipótesis sobre todo el conjunto de los datos.

Luego se combinan las distintas hipótesis.

- Maneras de Combinar hipótesis:

WEIGHTING MAJORITV: el valor se obtiene haciendo la media (caso

continuo) o la mediana (caso discreto).

STACKING / CASCADE: se utiliza cada hipótesis como una variable y se utiliza otro algoritmo (por ejemplo: Una red neuronal para asignar diferentes pesos a las diferentes hipótesis).

3.3 Métodos de Minería de Datos

Los métodos de minería de datos tienen como metas primarias (en un alto nivel) la predicción de datos desconocidos y la descripción de patrones

Pueden emplearse diferentes criterios para clasificar los sistemas de minería de datos y, en general, los sistemas de aprendizaje inductivo en computadoras:

- ✓ Dependiendo del objetivo para el que se realiza el aprendizaje, pueden distinguirse sistemas para: **clasificación** (clasificar datos en clases predefinidas), **regresión** (función que convierte datos en valores de una función de predicción), **agrupamiento de conceptos** (búsqueda de conjuntos en los que agrupar los datos), **compactación** (búsqueda de descripciones más compactas de los datos), **modelado de dependencias** (dependencias entre las variables de los datos), **detección de desviaciones** (búsqueda de desviaciones importantes de los datos respecto de valores anteriores o medios), etc.

- ✓ Dependiendo de la tendencia con que se aborde el problema, se pueden distinguir tres grandes líneas de investigación o paradigmas: **sistemas conexionistas** (redes neuronales), **sistemas evolucionistas** (algoritmos genéticos) y **sistemas simbólicos**.

- ✓ Dependiendo del lenguaje utilizado para representar del conocimiento, se pueden distinguir: representaciones basadas en la **lógica de proposiciones**, representaciones basadas en **lógica de predicados** de primer orden, representaciones **estructuradas**, representaciones a través de **ejemplos** y representaciones **no simbólicas** como las redes neuronales.

A continuación describiremos con más detalle los diferentes métodos de representación del conocimiento que se emplean en la minería de datos, dado que el lenguaje de representación es uno de los aspectos importantes para el proceso de KDD.

3.3.1 Agrupamiento ("Clustering"):

También llamada **Segmentación**, esta herramienta permite la identificación de tipologías o grupos donde los elementos guardan similitud entre sí y diferencias

con aquellos de otros grupos. Para alcanzar las distintas tipologías o grupos existentes en una base de datos, estas herramientas requieren, como entrada, información sobre el colectivo a segmentar. Esta información corresponderá a los valores concretos, para cada elemento en un momento del tiempo, de una serie de variables ("Segmentación estática") o a través del comportamiento en el tiempo de cada uno de los elementos del colectivo ("Segmentación dinámica").

Como resultado del tratamiento de la información, estas herramientas presentan los distintos grupos detectados junto con los valores característicos de las variables. Este tipo de herramientas se basan en técnicas de carácter estadístico, de empleo de algoritmos matemáticos, de generación de reglas y de redes neuronales para el tratamiento de registros.

Para otro tipo de elementos a agrupar o segmentar, como texto y documentos, se usan técnicas de reconocimiento de conceptos.

3.3.2 Asociación (" Association Pattern Discovery"):

Este tipo de herramientas establece las posibles relaciones o correlaciones entre distintas acciones o sucesos aparentemente independientes, pudiendo reconocer como la ocurrencia de un suceso o acción puede inducir o generar la aparición de otros. .

Normalmente este tipo de herramientas se fundamenta en técnicas estadísticas como los análisis de correlación y de variación.

3.3.3 Secuenciamiento (“Sequential Pattern Discovery"):

Esta herramienta permite identificar como, en el tiempo, la ocurrencia de una acción desencadena otras posteriormente. Es muy similar a la anteriormente analizada si bien, en este caso, el tiempo es una variable crítica e imprescindible a introducir en la información a analizar.

3.3.4 Reconocimiento de Patrones ("Pattern Matching"):

Estas herramientas permiten la asociación de una señalo información de entrada con aquella o aquellas con las que guarda mayor similitud y que están catalogadas en el sistema.

Estas herramientas son usadas por elementos que son tan habituales como un procesador de texto o un despertador. Los patrones pueden ser cualquier elemento de información que deseemos.

En el ámbito particular del DM estas herramientas pueden ayudamos en la identificación de problemas e incidencias y de sus posibles soluciones toda vez que dispongamos de la base de información necesaria en la cual buscar.

Estas herramientas se sustentan en las técnicas de Redes Neuronales y Algoritmos Matemáticos.

3.3.5 Previsión ("Forecasting"):

La Previsión establece el comportamiento futuro más probable dependiendo de la evolución pasada y presente.

Esta herramienta tiene su uso fundamental en el tratamiento de Series Temporales y las técnicas asociadas disponen de una importante madurez.

Las herramientas de Previsión utilizan bien la propia información histórica, o bien, la información histórica relativa a otras variables de las cuales la primera depende.

3.3.6 Simulación:

Las herramientas de Simulación forman parte también del conjunto de herramientas veteranas de la investigación científica. Como ejemplo están las herramientas de diseño y producción asistidas por ordenador, "CAD" - "CAM", en las cuales se revisan los diseños sometiéndoles a una amplísima serie de condiciones reales normales y extremas.

Ello permite no sólo ajustar y adaptar el diseño sino posteriormente establecer márgenes y límites de funcionamiento.

La simulación se puede definir como la generación de múltiples escenarios o posibilidades sujetas, normalmente, a unas reglas o esquemas con el objeto de analizar la idoneidad y comportamiento de una decisión o prototipo en un marco de posibles condiciones futuras o para analizar todas las posibles variaciones o alternativas a una decisión o situación y también se usa para el cálculo numérico.

3.3.7 Optimización:

Al igual que la Previsión y la Simulación, las herramientas de Optimización tienen una amplia tradición de uso.

La optimización ha sido y es extensivamente usada en la resolución de los problemas asociados a la logística de distribución y a la gestión de "Stocks" en los negocios y en la determinación de parámetros teóricos a partir de los experimentos en la investigación científica.

La optimización resuelve el problema de la minimización o maximización de una función que depende de una serie de variables, encontrando los valores de éstas que satisfacen esa condición de máximo, típicamente beneficios, o mínimo, normalmente costes.

Habitualmente estos problemas conllevan, adicionalmente, una serie de "ligaduras" o estricciones de forma que no todas las posibles soluciones son aceptables, ello se traduce en que debemos reducir nuestro universo de búsqueda a aquellas soluciones que satisfagan tales restricciones.

3.3.8 Clasificación ("Clasification", "Prediction" o "Scoring"):

La clasificación agrupa todas aquellas herramientas que permiten asignar a un elemento la pertenencia a un grupo o clase. Ello se instrumenta a través de la dependencia de la pertenencia a las clases en los valores de una serie de atributos o variables.

A través del análisis de un colectivo de elementos, o casos de los cuales conocemos la clase a la que pertenecen, se establece un mecanismo que establece la pertenencia a tales clases en función de los valores de las distintas variables y nos permite establecer el grado de discriminación o influencia de éstas.

También se utiliza para estas herramientas la denominación de Predicción o Evaluación para aquellos casos donde se aplican técnicas, normalmente numéricas, que establecen para cada elemento un valor dependiente de los valores que tengan las variables en tal elemento.

Las herramientas de Clasificación hacen uso de técnicas como algoritmos matemáticos, análisis discriminante y de variaciones, sistemas expertos y sistemas de conocimiento e inducción de reglas.

Como se ha podido apreciar, normalmente es necesaria la conjunción e integración de varios tipos de herramientas a efectos de brindar una solución completa a nuestros problemas.

Métodos apropiados

- ✓ No estructurados:
  - Métodos bayesianos.
  - Otros métodos estadísticos
  - Métodos relacionales

- ✓ Semi estructurados

  - Gramaticales.
  - Métodos relacionales con constructores.

Métodos no apropiados

Sin una profunda transformación de los datos, muchas técnicas de aprendizaje automático son útiles para muchas aplicaciones

- ✓ Métodos de clasificación (árboles de decisión,...) están basados. en una clase dependiente de un número de atributos predeterminados.

- ✓ Métodos numéricos (regresión, redes neuronales,...), los datos son simbólicos, no numéricos.

- ✓ Métodos por casos (KNN, CBR,...) tiempos de respuesta serían muy altos.

3.4 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 dirigido 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, estadística, visualización, recuperación de información, y computación de altas prestaciones.

Como mencionamos al principio, los algoritmos de minería de datos se clasifican en dos grandes categorías: supervisados o predictivos y no supervisados o de descubrimiento del conocimiento.

En la tabla siguiente se muestran algunas de las técnicas de minería de datos en ambas categorías:

SUPERVISADOS	NO SUPERVISADOS
Árboles de decisión	Detección de Desviaciones
Inducción neuronal	Segmentación
Regresión	Agrupamiento (Clustering)
Series temporales	Reglas de Asociación
	Patrones Secuenciales

Tabla 6.Clasificación de las técnicas de minería de datos

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 necesario interpretar y evaluar los resultados obtenidos.

La siguiente tabla muestra algunas de las técnicas más comunes de Minería de Datos y a continuación describiremos cada una de ellas:

Métodos estadísticos	ANOVA
	Prueba Ji cuadrado
	Análisis de componentes principales
	Análisis de clusters
	Análisis discriminante
	Regresión lineal
	Regresión logística
Arboles de decisión	CHAID
	CART
Reglas do asociación	
Redes de neuronas artificiales	
Algoritmos genéticos	
Otros	Lógica difusa
	Series temporales

Tabla 7. Técnicas do Data Mining

3.4.1 Métodos Estadísticos:

La estadística es tradicionalmente la técnica que se ha usado para el tratamiento de grandes volúmenes de datos numéricos y nadie pone en duda su efectividad al poseer un amplísimo conjunto de modelos de análisis para cubrir el tratamiento de todo tipo de poblaciones y series de datos. Estos son algunos de los métodos estadísticos más utilizados:

- ✓ **ANOVA:** Análisis de la Varianza, contrasta si existen diferencias significativas entre las medidas de una o más variables continuas en grupos de población distintos.

- ✓ **Ji cuadrado:** Contrasta la hipótesis de independencia entre variables.

- ✓ **Componentes principales:** Permite reducir el número de variables observadas a un menor número de variables artificiales, conservando la mayor parte de la información sobre la varianza de las variables.

- ✓ **Análisis de clusters:** Permite clasificar una población en un número determinado de grupos, sobre la base de semejanzas y diferencias de perfiles existentes entre los diferentes componentes de dicha población.

- ✓ **Análisis discriminante:** Método de clasificación de individuos en grupos que previamente se han establecido, y que permite encontrar la regla de clasificación de los elementos de estos grupos, y por tanto identificar cuáles son las variables que mejor definan la pertenencia al grupo.

- ✓ **Regresión Lineal:** Técnica más básica del Data Mining. Un modelo de regresión lineal se implementa identificando una variable dependiente (y) y todas las variables independientes (X_1, X_2,...). Se asume que la relación entre estas y aquella es lineal. Todas las variables han de ser continuas. El resultado es la ecuación de la recta que mejor se ajusta al juego de datos y esta ecuación se interpreta o se usa para predicción.

- ✓ Regresión Logística: Puede trabajar con variables discretas. También requiere que todas las variables sean lineales.

3.4.2 Métodos Basados en Arboles de Decisión

Son herramientas analíticas empleadas para el descubrimiento de reglas y relaciones mediante la ruptura y subdivisión sistemática de la información contenida en el conjunto de datos. El árbol de decisión se construye partiendo el conjunto de datos en dos (CART) o más (CHAID) subconjuntos de observaciones a partir de los valores que toman las variables predictoras. Cada uno de estos subconjuntos vuelve después a ser particionado utilizando el mismo algoritmo.

Este proceso continúa hasta que no se encuentran diferencias significativas en la influencia de las variables de predicción de uno de estos grupos hacia el valor de la variable de respuesta.

La raíz del árbol es el conjunto de datos íntegro, los subconjuntos y los subsubconjuntos conforman las ramas del árbol. Un conjunto en el que se hace una partición se llama nodo.

El método CHAID (Chi Squared Automatic Interaction Detector) es útil en aquellas situaciones en las que el objetivo es dividir una población en distintos segmentos basándose en algún criterio de decisión.

3.4.3 Reglas de Asociación

Derivan de un tipo de análisis que extrae información por coincidencias. Este análisis a veces llamado "cesta de la compra" permite descubrir correlaciones o co-ocurrencias en los sucesos de la base de datos a analizar y se formaliza en la obtención de reglas de tipo; SI ... ENTONCES...

3.4.4 Redes Neuronales ("Neural Networks")

Las Redes Neuronales constituyen una técnica inspirada en los trabajos de investigación, iniciados en 1930, que pretendían modelar computacionalmente el aprendizaje humano llevado a cabo a través de las neuronas en el cerebro.

Las redes neuronales son una nueva forma de analizar la información con una diferencia fundamental con respecto a las técnicas tradicionales: son capaces de detectar y aprender patrones y características dentro de los datos.

Se comportan de forma parecida a nuestro cerebro aprendiendo de la experiencia y el pasado y aplicando tal conocimiento a la resolución de problemas nuevos.

Una vez adiestradas las redes neuronales pueden hacer previsiones, clasificaciones y segmentación.

Las redes neuronales se construyen estructurando en una serie de niveles o capas compuesta por nodos o "neuronas". Poseen dos formas de aprendizaje derivadas del tipo de paradigma que usan: el supervisado y el no supervisado.

Son métodos de proceso numérico en paralelo que tratan de modelizar el funcionamiento del cerebro. La red asigna pesos al azar a cada variable independiente y determina si existe algún patrón predictivo en los datos. Una vez que encuentra un patrón la red lo optimiza reforzando los pesos de las variables y comparando con los datos del grupo de validación. Luego prosigue el proceso y aprende de los resultados una y otra vez. Finalmente, se puede aplicar el modelo aprendido a cualquier nuevo conjunto de datos de entrada. Pueden manejar datos continuos y discretos, lineales y no-lineales simultáneamente. El único inconveniente que presentan es que no genera una ecuación o modelo que explique el comportamiento del sistema, siendo muy difícil determinar la influencia de cada variable en el comportamiento global del sistema.

3.4.5 Algoritmos Genéticos ("Genetic Algorithms")

Los Algoritmos Genéticos son otra técnica que debe su inspiración, de nuevo, a la Biología como las Redes Neuronales.

Estos algoritmos representan la modelización matemática de como los cromosomas en un marco evolucionista alcanzan la estructura y composición más óptima en aras de la supervivencia. Entendiendo la evolución como un proceso de búsqueda y optimización de la adaptación de las especies que se plasma en mutaciones y cambios en los genes o cromosomas.

Los Algoritmos Genéticos hacen uso de las técnicas biológicas de reproducción (mutación y cruce) para ser utilizadas en todo tipo de problemas de búsqueda y optimización.

Esta aproximación está enfocada a problemas de optimización. Se comienza con una población de partida y se va alterando y optimizando su composición para la solución de un problema particular mediante mecanismos tomados de la teoría de la evolución (introducir elementos aleatorios para la modificación de las variables o mutaciones). El material genético o información de los individuos puede ser transmitido a las siguientes generaciones, de diferentes formas que van optimizando el proceso. A través de la reproducción, los mejores segmentos perduran y su proporción crece de generación en generación. Al cabo

de cierto número de iteraciones, la población estará constituida por buenas soluciones al problema de optimización.

Esta herramienta se usa en las primeras fases del Data Mining, para seleccionar las variables que luego se emplearán con otra técnica, como las redes de neuronas o la regresión logística.

3.4.6 Lógica Difusa ("fuzzy logic")

La Lógica Difusa surge de la necesidad de modelizar la realidad de una forma más exacta evitando precisamente el determinismo o la exactitud.

La Lógica permite el tratamiento probabilístico de la categorización de un colectivo.

La Lógica Difusa es aquella técnica que permite y trata la existencia de barreras difusas o suaves entre los distintos grupos en los que categorizamos un colectivo o entre los distintos elementos, factores o proporciones que concurren en una situación o solución.

3.4.7 Series Temporales

Consisten en el estudio de una variable a través del tiempo para, a partir de ese conocimiento, y bajo el supuesto de que no van a producirse cambios estructurales, poder realizar predicciones. Suelen basarse en un estudio de la serie en ciclos, tendencias y estacionalidades, que se diferencian por el ámbito de tiempo abarcado, para, por composición, obtener la serie original. Se pueden aplicar enfoques híbridos con los métodos anteriores, en los que la serie se puede explicar no sólo en función del tiempo sino como combinación de otras variables de entorno más estables y, por lo tanto, más fácilmente predecibles.

3.4.8 Redes Bayesianas

Las redes bayesianas son una alternativa para minería de datos, la cual tiene varias ventajas:

- ✓ Permiten aprender sobre relaciones de dependencia y causalidad.
- ✓ Permiten combinar conocimiento con datos.
- ✓ Evitan el sobre-ajuste de los datos.
- ✓ Pueden manejar bases de datos incompletos.

El obtener una red bayesiana a partir de datos es un proceso de aprendizaje, el cual se divide, naturalmente, en dos aspectos:

1. **Aprendizaje paramétrico:** dada una estructura, obtener las probabilidades a priori y condicionales requeridas.

2. **Aprendizaje estructural:** obtener la estructura de la red Bayesiana, es decir, las relaciones de dependencia e independencia entre las variables involucradas.

Las técnicas de aprendizaje estructural dependen del tipo de estructura de red: árboles, poliárboles y redes multicomectadas. Otra alternativa es combinar cOI1ocimiento subjetivo del experto con aprendizaje. Para ello se parte de la estructura dada por el experto, la cual se valida y mejora utilizando datos estadísticos.

3.4.9 Inducción de Reglas

Las técnicas de Inducción de Reglas surgieron hace dos décadas y permiten la generación y contraste de árboles de decisión o reglas y patrones a partir de los datos de entrada.

Como información de entrada, tendremos un conjunto de casos donde se ha asociado una clasificación o evaluación a un conjunto de variables o atributos.

Con tal información estas técnicas obtienen el árbol de decisión o conjunto de reglas que soportan la evaluación o clasificación.

En los casos en que la información de entrada posee algún tipo de "ruido" o defecto estas técnicas pueden habilitar métodos estadísticos de tipo probabilístico para generar, en estos casos, árboles de decisión podados o recortados.

3.4.10 Sistemas basados en el Conocimiento y Sistemas Expertos ("Knowledge Based Systems" & "Expert Systems")

Estos sistemas son un clásico de la Inteligencia Artificial.

Estas técnicas permiten la formalización de árboles y reglas de decisión extraídas de la formalización del conocimiento de los expertos.

Poseen motores llamados "Motores de Inferencia" que se encargan de gestionar las distintas preguntas al ser realizadas de forma que el proceso de decisión sea lo más eficiente y rápido posible.

3.4.11 Algoritmos Matemáticos

Sin llegar a ser técnicas que den soporte a unas necesidades concretas como las anteriores, existe una amplia gama de algoritmos matemáticos que son especialmente útiles y eficaces en la resolución y tratamiento de problemas muy específicos y. puntuales y que, normalmente, son incorporados en alguna de aquellas técnicas con el objeto de mejorarlas.

CAPÍTULO IV
EVALUACIÓN DE LA MINERÍA DE DATOS

4.1 Cuándo Utilizar Minería de Datos

4.1.1 Selección de la Técnica Adecuada.

El mínimo aceptable para elegir una tecnología de DM y luego un producto efectivo depende de sí realmente el producto puede beneficiar el negocio, qué significa realmente el benefició, ingresos aumentados, costes disminuidos o rendimiento de las inversiones (ROl). Si la técnica y herramienta no proporcionasen ninguna de estas 4 ventajas de una forma perceptible, sería poco probable que alguien en el negocio tenga tiempo para extraer los datos. Para desarrollar con éxito en el negocio, el DM debe ser algo más que una simple búsqueda de los patrones deseados en el interior de grandes bases de datos.

4.1.1.1 Proceso de Data Mining

Para empezar, hay que ver todo el proceso de DM ofrecido por los investigadores. Éstos descubrieron que el DM (la generación moderna de modelos predictivos y patrones) es sólo un primer paso en un largo proceso que mueve los datos hacia el conocimiento. Los investigadores ofrecieron los siguientes pasos (hay que tener en cuenta que se distinguen del conjunto de pasos más simples ofrecidos aquí para fechar).

Este punto de vista sobre el DM muestra cómo llegar de los datos brutos hacia el descubrimiento de patrones útiles para el conocimiento. La mejor herramienta de DM más automatizada y menos dolorosa es ir por pasos. Este proceso de DM ha sido explicado en la sección 2.6.

4.1.1.2 Similitudes entre Distintas Técnicas de DM

Para hacer una selección inteligente de herramientas y tecnologías de DM hay que determinar dónde está la diferencia. Dado que hay demasiadas superposiciones una de las mejores vías para ver las significativas diferencias entre los algoritmos es ver que es lo qué tienen en común. Por ejemplo, cada algoritmo de DM tiene:

- ✓ Estructura de modelo

  La estructura que determina el modelo (un árbol, red neuronal o vecino más cercano) Es un modelo percibido porque la instantaneidad real del modelo puede ser SOL preguntas en caso de árboles de decisiones o sistemas basados en reglas; o ecuaciones matemáticas en caso de regresiones estadísticas.

- ✓ Un plan de investigación

  Como el algoritmo corrige o modifica el modelo todo el tiempo cuando más datos se hacen disponibles. Por ejemplo, redes neuronales investigan a través de espacio de peso de enlaces mediante algoritmos de retro propagación y el algoritmo genético investiga a través de ponderaciones aleatorias y recombinaciones genéticas.

- ✓ Un proceso de validación

  ¿Cuándo se acaba el algoritmo al haber creado un modelo valido? Por ejemplo, los árboles de decisiones CART utilizan validación cruzada para determinar el nivel óptimo de crecimiento del árbol. Por su parte, las redes neuronales no tienen una técnica especializada de validación para determinar la terminación, pero la validación cruzada a menudo se usa en el exterior de redes neuronales.

Esta descripción de algoritmos de DM ya podría ser útil para hacer una selección inteligente de herramientas y tecnologías de DM, porque hace más fácil ver cómo se distinguen los algoritmos gen éticos de otras técnicas mencionadas como los algoritmos de DM. Los algoritmos genéticos no tienen una inherente estructura de modelo. Son sólo la estrategia de optimización para cualquier estructura de modelo que está determinada para codificarse en el material genético. El cromosoma por si sólo no es sólo una estructura de modelo, sino además la interpretación de cromosoma. Esta diferencia existe incluso en Biología, donde el material gen ético se llama el genotipo y se convierte en el fenotipo. Dado que el algoritmo genético es realmente una estrategia de investigación sin estructura de modelo, puede utilizarse en combinación con cualquier otra técnica de DM

La Tabla 8 clasifica algunos de los algoritmos de DM según las características de árboles. Al organizar los algoritmos mediante estas tres características, está claro que una de las mayores diferencias consiste en el uso de validación. Unas técnicas no la tienen incorporada en absoluto; otras, evalúan directamente contra los datos y algunas usan test de significado estadístico para calcular el mejor modelo. Aunque útil, este punto de vista pasa por alto algunos rasgos importantes que pueden marcar toda la diferencia. En efecto, la estructura de modelo utilizado puede criticarse en función de sí el problema puede solucionarse por completo en un periodo de tiempo razonable.

Algoritmo	Estructura	Búsqueda	Validación
CART	Árboles binarios	Divisiones elegidas por entropía métrica de Gini.	Validación cruzada
CHAID	Árbol dividido en Multivías	Según test Chi-Cuadrado y Estadístico de Bonferroni.	Validación realizada en el momento de selección de partición.
Redes Neuronales	Red de propagación directa con umbral no lineal	Retro propagación de errores	No se aplica
Algoritmos Genéticos	No se aplica	Supervivencia de los más aptos en mutaciones y cruce genético	Por lo general validación cruzada
Generación de Reglas	Regla de si- entonces	Agrega nueva restricción a la regla y la retiene si esta empareja el criterio de "interesante" basado en precisión y cobertura	Test de chi.cuadrado con corte por significación estadística
El vecino más Cercano	Distancia de fototipo en n-dimensional espacio de rasgos	Por lo general no hay búsqueda	Validación cruzada para tasas de precisión comunicadas pero no para terminación de algoritmo

Tabla 8.Comparación de los algoritmos de DM

4.2 Evaluación de una Herramienta de Minería de Datos.

Como se ha visto, muchos de los puntos de vista actuales sobre el DM son erróneos. Los sistemas están altamente optimizados para mejorar un uno por mil en precisión cuando, en realidad, el mercado se mueve a tanta velocidad que es poco probable que algún modelo predictivo pueda seguir siendo tan preciso todo el tiempo. (Por esta razón, los recientes debates sobre grandes o pequeños datos y hardware de procesamiento masivo en paralelo pueden ser superfluos). No tiene sentido preocuparse acerca de la precisión del sistema para aumentar un poco las ganancias cuando la base de datos misma está corrompida por culpa de copias y transferencias o cuando el modelo de negocio está mal definido y lleva a la empresa en la dirección equivocada.

Aunque la precisión predictiva sea la meta final del DM, se pueden diferenciar tres medidas claves necesarias para una evaluación completa de la herramienta de DM. Estas tres medidas deberían ser las reglas de oro para el desarrollo de las herramientas de DM:

- ✓ **Precisión:** La herramienta de DM debe generar un modelo lo más preciso posible, pero reconociendo que las pequeñas diferencias en las distintas técnicas pueden deberse a fluctuaciones en muestreo aleatorio (incluso si se usa la base de datos completa para el modelo) o pueden ser despreciables en la dinámica del mercado en el que se despliegan los modelos.

- ✓ **Explicación:** La herramienta de DM tiene que ser capaz de explicar al usuario final de un modo claro cómo funciona el modelo para que pueda desarrollar la intuición. De este modo, las intuiciones y el sentido común serán fácilmente controlados y confirmados. Asimismo, la explicación del beneficio o el cálculo del rendimiento de la inversión tienen que ser fáciles y claros.

- ✓ **Integración:** La herramienta de DM debe integrarse en el proceso real de negocio, flujos de datos e información de la empresa. La solicitud de copias de datos y reprocesamiento masivo de datos aumenta la posibilidad de error mientras que una integración rigurosa reduce significativamente esta posibilidad.

Si estos tres requisitos se cumpliesen debidamente, las herramientas de DM producirán modelos de alto rendimiento y larga duración.

Se puede decir que las herramientas de DM deberían evaluarse primero por su capacidad de alcanzar unos niveles de precisión aceptables, pero, a continuación, diseñarse para ser integradas sin transformaciones en el proceso de negocio existente mediante la explicación de los resultados e integración con la tecnología informática de proceso de datos de que se dispone.

4.2.1 Incrustación del DM dentro del Proceso de Negocio

Con estas nuevas reglas para DM eficaz dentro del proceso real de negocio habrá que introducir algunos cambios para que el DM consiga estos objetivos. Por ejemplo, a menudo se copia la base de datos, porque muchos algoritmos necesitan un pre-formateo manual de datos, antes de lanzar las herramientas, y se hacen los extractos de fichero plano para aumentar la velocidad de la construcción del modelo. Para evitarlo, el proceso de DM necesita ser integrado en el hardware, software y SGBD (Sistema de Gestión de Bases de Datos), donde se almacenan los datos. Esto significa reescribir el algoritmo de DM en SOL para el sistema de base de datos relacional o escribir procedimientos almacenados o funciones de acceso especial para nuevos tipos de datos.

Independientemente de la forma, lo principal es obtener un sistema de DM que no requiera extracciones de datos y reduzca cualquier reprocesamiento de la base de datos. Finalmente, el sistema de DM completamente integrado aprovechará al máximo la información de existentes diccionarios de datos y otros metadatos y preparará los metadatos que genera para que puedan guardarse en

el almacén de datos actual.

Para conseguir una buena explicación de DM se puede elegir entre:

- ✓ Crear potentes herramientas de visualización específicas.
- ✓ Mostrar los resultados de DM en objetos de visualización menos potentes, pero más familiares, aprovechando la interfaz de usuario existente que el usuario final suele utilizar para la resolución de problemas de negocio.

Ambas estrategias son válidas, pero la segunda aproximación, si tiene éxito, es con diferencia la más fácil para el usuario final, dado que no necesita aprender a manejar otras herramientas y puede contrastar los resultados de DM dentro de un entorno de trabajo en el que tiene las intuiciones desarrolladas.

Para conseguir tanto la integración como la explicación, el sistema de DM tiene que estar incrustado dentro de la estructura de almacén de datos existente y la herramienta de explicación tiene que estar incrustada en las herramientas existentes y en las aplicaciones de navegación por los datos, familiares para el usuario final.

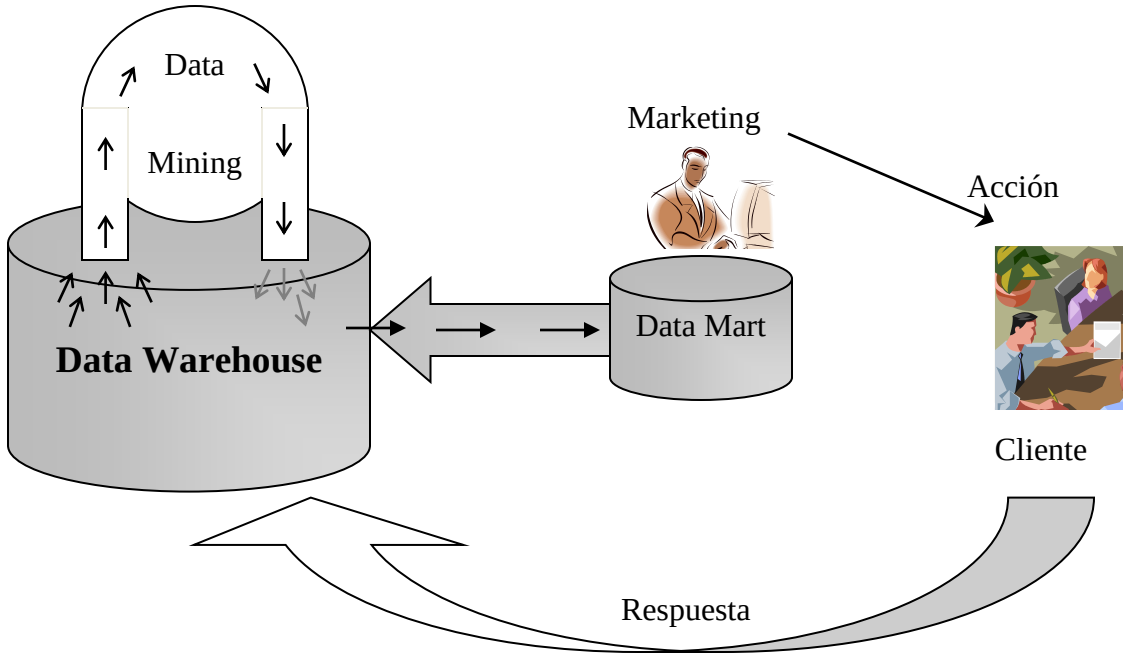


Figura 10. Situación IDEAL de los flujos de datos dentro de una empresa

Como ejemplo, puede servir el software Discovery Server de la empresa Pilot Software que permite incrustar el mecanismo de DM directamente en cualquier SGBDR (Sistema de Gestión de Bases de Datos Relacionales). Esto se consigue aplicando el algoritmo de árbol de decisión en SQL contra cualquier base de datos existente, sin importar si está completamente normalizada, no está en absoluto normalizada o lo está a medias. Sin hacer extractos de base de datos, el

modelo predictivo se almacenó en tablas dentro del SGBDR como metadata. De esta forma, el modelo puede utilizarse para realizar predicciones sobre nuevos datos dentro del SGBDR. Para conseguir un alto nivel de explicación del modelo predictivo, el árbol de decisión fue transformado en segmentación e incrustado como una dimensión jerárquica en la herramienta de visualización multidimensional de la base de datos.

Esto permitió a los usuarios finales utilizar el concepto familiar de segmentación para visualizar el modelo predictivo y les proporcionó la potencia del motor OLAP existente para navegar por el modelo predictivo al igual que en cualquier otra dimensión (Ver figura 10).

4.3 Evaluación de Minería de Datos.

Al realizar el DM desde dentro del almacén de datos, ya no hay necesidad de hacer extracciones de datos. De esta forma se evitan multitud de posibles errores y pasos de pre-procesamiento. Incrustando los resultados de DM en el mercado de datos de OLAP como nueva dimensión en una vista multidimensional, el usuario final entiende los resultados de DM y los usa.

4.3.1 Medición de Precisión

A la hora de evaluar la capacidad predictiva de una herramienta de DM, el parámetro más importante suele ser la precisión de las predicciones que realiza. Para los sistemas de aprendizaje no supervisado, como análisis de conglomerados o generación de reglas de asociación, la medición directa es más difícil debido a que el rendimiento del sistema depende mucho de las circunstancias particulares en las cuales se utiliza.

Los parámetros que habitualmente se utilizan para evaluar la predicción realizada por una herramienta de DM son los siguientes:

1. **Precisión:** La precisión es el porcentaje de las predicciones correctas con respecto a todas las realizadas. Si la variable predicha es discreta o binaria, es fácil calcular la precisión, porque el resultado o es correcto o es erróneo. Sin embargo, si la variable predicha es continua (como predicción de ingresos), el cálculo de la precisión puede ser complicado, pues hará falta definir los márgenes dentro de los que la predicción se considera como correcta (por ejemplo, la predicción de ingreso de 4.500.000 +/- 200.000 Pts).

2. **Tasa de error:** La tasa de error es el complemento de la precisión, mide el porcentaje de las predicciones que son erróneas. Se suele utilizar cuando los niveles de precisión son muy altos, pues resulta más fácil apreciar la mejora. Por ejemplo, la mejora de la precisión del 99,0% al 99,5% puede parecer menos importante que la mejora del 50% al 75%, sin embargo, en

ambos casos la tasa de error se reduce a la mitad (una mejora espectacular).

3. **Tasa de error en rechazo:** A menudo, al realizar la predicción, el algoritmo de DM proporciona tanto la predicción como la confianza de que dicha predicción sea correcta. Por ejemplo, el algoritmo del vecino más cercano puede proporcionar la misma predicción para todos los vecinos o para una mayoría. La predicción puede ser la misma en ambos casos, pero en el caso de unanimidad la confianza en la predicción es más alta. Las predicciones pueden ordenarse según su confianza y las que menos confianza "inspiran" pueden rechazarse. De esta forma, se puede duplicar la precisión rechazando un 80% de predicciones.

4. **Error cuadrático medio:** Para las variables continuas, el grado de mal emparejamiento entre la predicción y el valor real pueden calcularse restando los dos valores y elevando el resultado al cuadrado. Este "error cuadrado" puede promediarse sobre todas las predicciones para estimar la distancia entre los valores reales y las predicciones. La elevación al cuadrado tiene dos ventajas: por un lado, da un mayor peso a los errores graves; por otro lado, asegura que todos los errores son positivos y se suman a la hora de calcular la media.

5. **Elevación:** Mide el grado en el que el modelo de predicción aumenta la densidad de respuestas en el subconjunto dado de la base de datos en comparación con el que podría alcanzarse sin usar el modelo (selección al azar). Por ejemplo, si la densidad normal de respuesta en la población objeto de un mailing es de un 10%, pero enfocando la campaña en el primer 25% de la población que el modelo predice que va a responder, la respuesta aumenta hasta un 30%, entonces la elevación sería 3 (=30%/10%) para el primer cuartil de la base da datos. Debido a que este método para medir la elevación proporciona la mejora limitada a un de la respuesta para el modelo predictivo y mercado casual.

6. **Beneficio/ROI:** En realidad, el mejor método para medir el rendimiento de un sistema es calcular el beneficio máximo o el rendimiento de la inversión del modelo predictivo. Si el proceso de negocio dispone de un modelo de beneficio o rendimiento de la inversión, este parámetro simplifica espectacularmente la búsqueda del tamaño adecuado de la muestra.

4.3.2 Medición de Explicación

Una vez evaluada la precisión de un sistema de DM, es importante disponer de herramientas de explicación dentro del sistema. Las cuestiones a tener en cuenta son los siguientes:

1. **Generación automatizada de reglas:** Independientemente del sistema, a menudo es posible generar las reglas que expliquen el modelo predictivo.

Aunque no se puede garantizar que las reglas sean causales, ni que abarquen completamente el proceso de negocio, pueden ayudar a comprender el modelo.

2. **OLAP integrado:** Hay que comprobar si los resultados están a disposición del existente entorno de navegación OLAP. A menudo, OLAP puede dar nuevas ideas sobre cómo usar los "datos y confirmar intuiciones sobre el modelo.

3. **Evaluación de modelo:** El sistema de DM debería realizar la evaluación automática del modelo, mediante un conjunto de test o por validación cruzada, y permitir al usuario ver tanto modelos más complicados como más generales y sus rendimientos en datos evaluados.

4.3.3 Medición de Integración

Existen muchos factores involucrados en la integración de los procesos de negocio con los procesos informáticos. El propósito de todos ellos es simplificar la integración de DM en los procesos existentes minimizando las interrupciones de flujos de información y reduciendo las necesidades de ampliación de hardware, software o almacén de datos. Hay que tener en cuenta los siguientes momentos críticos:

1. **Extractos de datos propios:** Para maximizar el rendimiento de DM, es importante encontrar herramientas de DM que utilicen los datos tal y como están en la empresa. Las herramientas que necesitan copiar los datos, transformarlos en ficheros de texto u otros formatos introducen posibilidad de nuevos errores y complican mucho la ampliación y actualización del almacén de datos.

2. **Metadatos:** Seria muy ventajoso que el modelo predictivo se almacenara como metadatos en el almacén de datos, junto con las puntuaciones de los registros asignadas por el modelo predictivo. También sería interesante que los algoritmos de DM pudieran acceder y utilizar las estructuras existentes de metadatos.

3. **Pre-procesamento de predictores:** Muchos de los sistemas habituales requieren la modificación de predictores en la base de datos (por ejemplo, la conversión a valores entre 0,0 y 1,0 para redes neuronales). Hay que asegurarse de que el sistema es capaz de hacerlo automáticamente sin perder precisión o de que existan herramientas adicionales que lo hagan, fáciles de usar y de integrar en el sistema de DM.

4. **Tipos de predictor/predicción:** Hay que asegurarse de que la herramienta admite todos los tipos habituales de predictores de un modo comprensible y sin transformación: ordinales, categóricos, columnas de alta cardinalidad como código postal, continuos o binarios.

5. **Datos sucios:** Los datos suelen tener problemas que afectan significativamente a la precisión del modelo predictivo: pueden llevar a conclusiones erróneas, pueden no contener suficiente información, pueden no estar elegidos al azar o, simplemente, pueden ser erróneos. Las técnicas de DM suelen manejar los datos sucios con soltura y proporcionar resultados razonables casi con cualquier información disponible. La validación cruzada y la validación por conjuntos de test son unas herramientas muy potentes de localización y eliminación de errores en una base de datos. Sin embargo, existen casos en los que tanto los datos de aprendizaje como los de validación son erróneos, con lo que incluso la validación cruzada es incapaz de detectarlos. Esto suele deberse a errores de transformación y, por tanto, pueden evitarse en la mayoría de los casos.

6. **Valores perdidos:** Otra forma de ruido habitual en una base de datos son los valores perdidos. Todas las herramientas deberían ser capaces de manejar los valores perdidos, ya sea ignorándolos (omitiendo los registros que los contienen para predictores críticos), ya sea reconstruyéndolos (por ejemplo, usando el valor medio o prediciendo el valor perdido en base a los casos con predictores conocidos).

7. **Escalabilidad:** La herramienta de DM tendría que ser capaz de manejar tanto las bases de datos grandes como pequeñas, disponer de una robusta y automatizada rutina de muestreo y ser capaz de aprovechar el hardware de procesamiento paralelo e implementaciones de SGBDR.

4.3.4 Futuro del DM Incrustado

Una vez que el proceso de DM se hace fácil de. utilizar y se integra plenamente en los procesos de negocio y los flujos de información y datos de la empresa, surgirán nuevas aplicaciones y sinergias que convertirán el DM en una parte imprescindible de un almacén de datos moderno y funcional. Algunas de estas sinergias son las siguientes:

- ✓ **DM → BD multidimensional:** Uno de los problemas de OLAP y de BDMs (Base de Datos Multidimensional) es la dificultad de determinar qué columnas del almacén de datos tienen que ser dimensiones de la BD multidimensional (sobre todo, sí el usuario no tiene claro cuáles de ellas son importantes para el negocio). En este caso, el DM puede crear un modelo predictivo de dimensiones para BDM o calcular un ranking por importancia de las columnas de datos basado en algún objetivo de predicción de negocio.

- ✓ **DM → estructura de DW:** Uno de los problemas que surgen al pasar los datos de las transacciones al almacén de datos consiste en determinar cuáles son los datos relevantes para la mayoría de las cuestiones de negocio. En este caso, el DM permite hacer un primer filtrado,

determinando cuáles son los datos importantes para el negocio y, por tanto, tienen que añadirse al almacén de datos.

- ✓ **BD multidimensional + Datos resumidos → DM:** Cuantos más datos hay, mejor será el rendimiento de las técnicas de DM. De esta forma, hasta los más oscuros e interesantes patrones saldrán a la luz. Hay que tener en cuenta que la información extra que aportan el pre-procesamiento de datos en resúmenes y los metadatos en forma de jerarquías desplegables en dimensiones también mejoran el rendimiento de DM. Incluso un minado directo de una BDM resumida, sin datos detallados, será mucho más eficaz que la búsqueda manual de patrones en la BDM, como suele hacerse en sistemas OLAP.

BIBLIOGRAFÍA

- Data Mining: Concepts and Techniques. By Han, Jiawei / Kamber, Micheline Morgan Kaufmann Publishers; 08/2000

- Predictive Data Mining: A Practical Guide By Weiss. Sholom M. /Indurkhya, Nitin Morgan Kaufmann Pub; 08/1997

- Advances in Knowledge Discovery and Data Mining. By Fayyad, Usama M. (Edt) / Piatetsky-Shapiro, Gregory (Edt) / Smyth, Padhr MIT Pr; 03/1996

- Data Mining : Technologies, Techniques, Tools, and Trends By Thuraisingham, Bhavani CRC Pr; 12/1998

- Methodologies for Knowledge Discovery and Data Mining By Zhong, SPRINGER VERLAG INC; 07/1999

Ligas y Software

- http://www.kdnuggets.com

- http://www.cs.waikato.ac.nz/ml/weka/

- http://dms.irb.hr/

- http://www.oracle.com

- http://www.dbminer.com

- https://rapidminer.com/