

MINERIA DE DATOS

|  |  |  |
| --- | --- | --- |
|  | **Elaborado Por:** | |
| **Profesor:  Coronel Castillo** | • | ***Arias Nicolas, Zamir*** |
| • | ***Medina Anchita, Jaime*** |
|  |
|  |
|
|  |
|  |
|  |
|  |
|  |
|  |
|  |

*Lima-Perú, diciembre del 2017.*

**ÍNDICE**

[CONCEPTOS MINERIA](#page7) DE DATOS [6](#page7)

[IMPORTANCIA DE LA DATA MINING](#page13) [12](#page13)

[DESCUBRIMIENTO DE CONOCIMIENTO Y MINERÍA DE DATOS](#page14) [13](#page14)

[Historia del KDD](#page15) [14](#page15)

[Metas](#page16) [15](#page16)

[Relación con otras disciplinas](#page16) [15](#page16)

[El proceso de KDD](#page16) [15](#page16)

[DATA MINING](#page17) [16](#page17)

[Grupos de técnicas principales](#page20) [19](#page20)

[Visualización.](#page20) [19](#page20)

[Verificación.](#page21) [20](#page21)

[Descubrimiento.](#page22) [21](#page22)

[Prediciendo el tamaño de las audiencias televisivas.](#page33) [32](#page33)

[En la universidad](#page33) [32](#page33)

[En investigaciones espaciales](#page34) [33](#page34)

[En los clubes deportivos](#page34) [33](#page34)

[Negocios](#page35) [34](#page35)

[Patrones de fuga](#page36) [35](#page36)

[Fraudes](#page36) [35](#page36)

[Recursos humanos](#page37) [36](#page37)

[Comportamiento en Internet](#page37) [36](#page37)

[Terrorismo](#page37) [36](#page37)

[Ciencia e Ingeniería](#page38) [37](#page38)

[Minería de datos y otras disciplinas análogas](#page39) [38](#page39)

[De la estadística](#page39) [38](#page39)

[De la informática](#page40) [39](#page40)

[HERRAMIENTAS DE SOFTWARE](#page41) [40](#page41)

[Software DTM - Minería de Datos y Textos (Lebart)](#page42) [41](#page42)

[CONCLUSIONES](#page44) [43](#page44)

[FUENTES CONSULTADAS](#page45) [44](#page45)

**INTRODUCCIÓN**

Las empresas suelen generar grandes cantidades de información sobre sus procesos productivos, desempeño operacional, mercados y clientes. Pero el éxito de los negocios depende por lo general de la habilidad para ver nuevas tendencias o cambios en las tendencias. Las aplicaciones de data mining pueden identificar tendencias y comportamientos, no sólo para extraer información, sino también para descubrir las relaciones en bases de datos que pueden identificar comportamientos que no son muy evidentes.

Cada día generamos una gran cantidad de información, algunas veces conscientes de que lo hacemos y otras veces inconscientes de ello porque lo desconocemos. Nos damos cuenta de que generamos información cuando registramos nuestra entrada en el trabajo, cuando entramos en un servidor para ver nuestro correo, cuando pagamos con una tarjeta de crédito o cuando reservamos un billete de avión. Otras veces no nos damos cuenta de que generamos información, como cuando conducimos por una vía donde están contabilizando el número de automóviles que pasan por minuto, cuando se sigue nuestra navegación por Internet o cuando nos sacan una fotografía del rostro al haber pasado cerca de una oficina gubernamental.

¿Con qué finalidad queremos generar información? Son muchos los motivos que nos llevan a generar información, ya que nos pueden ayudar a controlar, optimizar, administrar, examinar, investigar, planificar, predecir, someter, negociar o tomar decisiones de cualquier ámbito según el dominio en que nos desarrollemos. La información por sí misma está considerada un bien patrimonial. De esta forma, si una empresa tiene una pérdida total o parcial de información provoca bastantes perjuicios. Es evidente que la información debe ser protegida, pero también explotada.

¿Qué nos ha permitido poder generar tanta información? En los últimos años, debido al desarrollo tecnológico a niveles exponenciales tanto en el área de cómputo como en la de transmisión de datos, ha sido posible que se gestionen de una mejor manera el manejo y almacenamiento de la información. Sin duda existen cuatro factores importantes que nos han llevado a este suceso:



1. El abaratamiento de los sistemas de almacenamiento tanto temporal como permanente.
2. El incremento de las velocidades de cómputo en los procesadores.
3. Las mejoras en la confiabilidad y aumento de la velocidad en la transmisión de datos.

4. El desarrollo de sistemas administradores de bases de datos más poderosos.

Actualmente todas estas ventajas nos han llevado a abusar del almacenamiento de la información en las bases de datos. Podemos decir que algunas empresas almacenan un cierto tipo de datos al que hemos denominado *dato-escritura*, ya que sólo se guarda (o *escribe*) en el disco duro, pero nunca se hace uso de él. Generalmente, todas las empresas usan un dato llamado *dato-escritura-lectura*, que utilizan para hacer consultas dirigidas. Un nuevo tipo de dato al cual hemos denominado *dato-escritura-lectura-análisis* es el que proporciona en conjunto un verdadero conocimiento y nosapoya en las tomas de decisiones. Es necesario contar con tecnologías que nos ayuden a explotar el potencial de este tipo de datos.

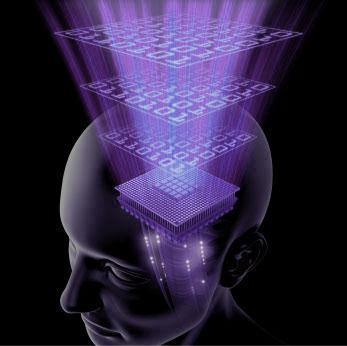
La cantidad de información que nos llega cada día es tan inmensa que nos resulta difícil asimilarla. Basta con ir al buscador *Google* y solicitar la palabra *información* para ver que existen 171.769.416 sitios donde nos pueden decir algo al respecto. Suponiendo que nos tomemos un minuto para ver el contenido de cada página, tardaríamos entonces 326 años en visitarlas todas. Esto es imposible, y, por lo tanto, existe una clara necesidad de disponer de tecnologías que nos ayuden en nuestros procesos de búsqueda y, aún más, de tecnologías que nos ayuden a comprender su contenido.

El data mining surge como una tecnología que intenta ayudar a comprender el contenido de una base de datos. De forma general, los datos son la materia prima bruta. En el momento que el usuario les atribuye algún significado especial pasan a convertirse en información. Cuando los especialistas elaboran o encuentran un modelo, haciendo que la interpretación del confronto entre la información y ese modelo represente un valor agregado, entonces nos referimos al conocimiento [Molina, 2002].

La técnica usada para realizar estas operaciones en data mining se denomina modelado y es, simplemente, el acto de construir un modelo en una situación donde usted conoce la respuesta y luego la aplica en otra situación de la cual desconoce la respuesta. Principal diferencia: los algoritmos están adaptados para poder trabajar sobre grandes bases de datos El data mining es un conjunto de actividades utilizadas para encontrar en los datos contextos nuevos, ocultos o inesperados. Utilizando información contenida en un data Warehouse (o “depósito de datos”), el data mining puede responder a preguntas que un decisor no hubiera formulado de no contar con estas herramientas. Usando una combinación de técnicas que incluyen el análisis estadístico, la lógica neuronal, la lógica difusa, el análisis multidimensional, la visualización de datos y los agentes inteligentes, puede descubrir patrones útiles para

desarrollar modelos predictivos de conductas o de consecuencias, en una amplia variedad de dominios del conocimiento.

Nuestra capacidad para almacenar datos ha crecido en los últimos años a velocidades exponenciales. En contrapartida, nuestra capacidad para procesarlos y utilizarlos no ha ido a la par. Por este motivo, la data mining se presenta como una tecnología de apoyo para explorar, analizar, comprender y aplicar el conocimiento obtenido usando grandes volúmenes de datos. Descubrir nuevos caminos que nos ayuden en la identificación de interesantes estructuras en los datos es una de las tareas fundamentales en el data mining.



**CONCEPTO MINERIA DE DATOS**

La minería de datos es el proceso de detectar la información procesable de los conjuntos grandes de datos. Utiliza el análisis matemático para deducir los patrones y tendencias que existen en los datos. Normalmente, estos patrones no se pueden detectar mediante la exploración tradicional de los datos porque las relaciones son demasiado complejas o porque hay demasiado datos.

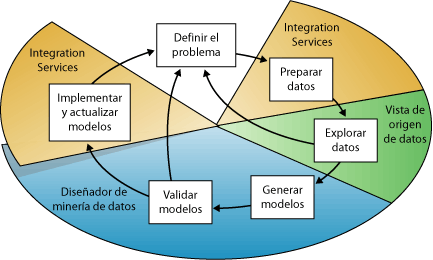
Estos patrones y tendencias se pueden recopilar y definir como un modelo de minería de datos. Los modelos de minería de datos se pueden aplicar en escenarios como los siguientes:

* **Pronóstico**: cálculo de las ventas y predicción de las cargas del servidor o del tiempo de inactividad del servidor.
* **Riesgo y probabilidad**: elección de los mejores clientes para la distribución de correo directo, determinación del punto de equilibrio probable para los escenarios de riesgo, y asignación de probabilidades a diagnósticos y otros resultados.
* **Recomendaciones**: determinación de los productos que se pueden vender juntos y generación de recomendaciones.
* **Búsqueda de secuencias**: análisis de los artículos que los clientes han introducido en el carrito de la compra y predicción de posibles eventos.
* **Agrupación**: distribución de clientes o eventos en grupos de elementos relacionados, y análisis y predicción de afinidades.

La generación de un modelo de minería de datos forma parte de un proceso mayor que incluye desde la formulación de preguntas acerca de los datos y la creación de un modelo para responderlas, hasta la implementación del modelo en un entorno de trabajo. Este proceso se puede definir mediante los seis pasos básicos siguientes:

1. [Definir el problema](https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-concepts#DefiningTheProblem)
2. [Preparar los datos](https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-concepts#PreparingData)
3. [Explorar los datos](https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-concepts#ExploringData)
4. [Generar modelos](https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-concepts#BuildingModels)
5. [Explorar y validar los modelos](https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-concepts#ValidatingModels)
6. [Implementar y actualizar los modelos](https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-concepts#DeployingandUpdatingModels)

El siguiente diagrama describe las relaciones existentes entre cada paso del proceso y las tecnologías de Microsoft SQL Server que se pueden usar para completar cada paso.

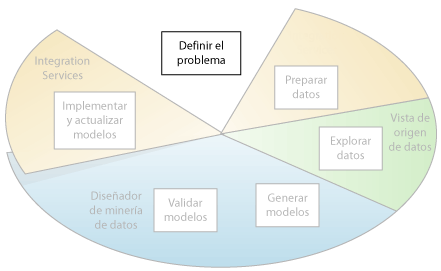


El proceso que se ilustra en el diagrama es cíclico, lo que significa que la creación de un modelo de minería de datos es un proceso dinámico e iterativo. Una vez que ha explorado los datos, puede que descubra que resultan insuficientes para crear los modelos de minería de datos adecuados y que, por tanto, debe buscar más datos. O bien, puede generar varios modelos y descubrir entonces que no responden adecuadamente al problema planteado cuando los definió y que, por tanto, debe volver a definir el problema. Es posible que deba actualizar los modelos una vez implementados debido a que haya más datos disponibles. Puede que haya que repetir cada paso del proceso muchas veces para crear un modelo adecuado.

La minería de datos de Microsoft SQL Server ofrece un entorno integrado para crear y trabajar con modelos de minería de datos. Este entorno incluye SQL Server Development Studio, que contiene algoritmos de minería de datos y herramientas de consulta que facilitan la generación de una solución completa para una serie de proyectos, y SQL Server Management Studio, que contiene herramientas que permiten examinar modelos y administrar objetos de minería de datos.

Definir el problema

El primer paso del proceso de minería de datos, tal como se resalta en el siguiente diagrama, consiste en definir claramente el problema y considerar formas de usar los datos para proporcionar una respuesta para el mismo.



Este paso incluye analizar los requisitos empresariales, definir el ámbito del problema, definir las métricas por las que se evaluará el modelo y definir los objetivos concretos del proyecto de minería de datos. Estas tareas se traducen en preguntas como las siguientes:

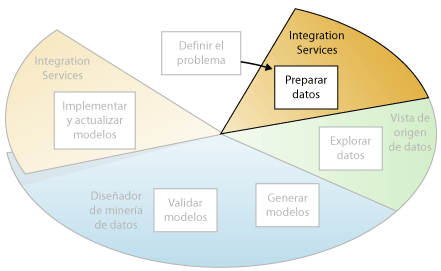
* ¿Qué está buscando? ¿Qué tipos de relaciones intenta buscar?
* ¿Refleja el problema que está intentando resolver las directivas o procesos de la empresa?
* ¿Desea realizar predicciones a partir del modelo de minería de datos o solamente buscar asociaciones y patrones interesantes?
* ¿Qué resultado o atributo desea predecir?
* ¿Qué tipo de datos tiene y qué tipo de información hay en cada columna? En caso de que haya varias tablas, ¿cómo se relacionan? ¿Necesita limpiar, agregar o procesar los datos antes de poder usarlos?
* ¿Cómo se distribuyen los datos? ¿Los datos son estacionales? ¿Los datos representan con precisión los procesos de la empresa?

Para responder a estas preguntas, puede que deba dirigir un estudio de disponibilidad de datos para investigar las necesidades de los usuarios de la empresa con respecto a los datos disponibles. Si los datos no abarcan las necesidades de los usuarios, podría tener que volver a definir el proyecto.

También debe considerar las maneras en las que los resultados del modelo se pueden incorporar en los indicadores de rendimiento clave (KPI) que se utilizan para medir el progreso comercial.

## Preparar los datos

El segundo paso del proceso de minería de datos, como se indica en el siguiente diagrama, consiste en consolidar y limpiar los datos identificados en el paso [Definir el problema](https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-concepts#DefiningTheProblem) .



Los datos pueden estar dispersos en la empresa y almacenados en formatos distintos; también pueden contener incoherencias como entradas que faltan o incorrectas. Por ejemplo, los datos pueden mostrar que un cliente adquirió un producto incluso antes que se ofreciera en el mercado o que el cliente compra regularmente en una tienda situada a 2.000 kilómetros de su casa.

La limpieza de datos no solamente implica quitar los datos no válidos o interpolar valores que faltan, sino también buscar las correlaciones ocultas en los datos, identificar los orígenes de datos que son más precisos y determinar qué columnas son las más adecuadas para el análisis. Por ejemplo, ¿debería utilizar la fecha de envío o la fecha de pedido? ¿Qué influye más en las ventas: la cantidad, el precio total o un precio con descuento? Los datos incompletos, los datos incorrectos y las entradas que parecen independientes, pero que de hecho están estrechamente correlacionadas, pueden influir en los resultados del modelo de maneras que no espera.

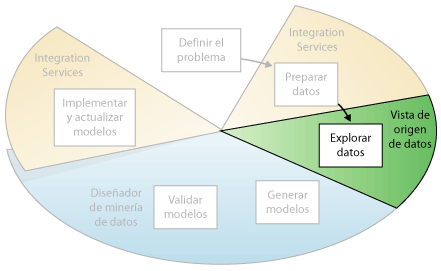
Por consiguiente, antes de empezar a generar los modelos de minería de datos, debería identificar estos problemas y determinar cómo los corregirá. En la minería de datos, por lo general se trabaja con un conjunto de datos de gran tamaño y no se puede examinar la calidad de los datos de cada transacción; por tanto, es posible que necesite usar herramientas de generación de perfiles de datos, y de limpieza y filtrado automático de datos, como las que se proporcionan en Integration Services, Microsoft SQL Server 2012 Master Data Services o SQL Server Data Quality Services para explorar los datos y buscar incoherencias. Para obtener más información, vea estos recursos:

* [Integration Services en Business Intelligence Development Studio](https://technet.microsoft.com/library/ms174181(v=sql.110).aspx)
* [Introducción a Master Data Services (MDS)](https://docs.microsoft.com/es-es/sql/master-data-services/master-data-services-overview-mds)
* [Data Quality Services](https://docs.microsoft.com/es-es/sql/data-quality-services/data-quality-services)

Es importante tener en cuenta que los datos que se usan para la minería de datos no necesitan almacenarse en un cubo de procesamiento analítico en línea (OLAP), ni siquiera en una base de datos relacional, aunque puede usar ambos como orígenes de datos. Puede realizar minería de datos mediante cualquier origen de datos definido como origen de datos de Analysis Services. Por ejemplo, archivos de texto, libros de Excel o datos de otros proveedores externos.

## Explorar los datos

El tercer paso del proceso de minería de datos, como se resalta en el siguiente diagrama, consiste en explorar los datos preparados.



Debe conocer los datos para tomar las decisiones adecuadas al crear los modelos de minería de datos. Entre las técnicas de exploración se incluyen calcular los valores mínimos y máximos, calcular la media y las desviaciones estándar, y examinar la distribución de los datos. Por ejemplo, al revisar el máximo, el mínimo y los valores de la media se podrían determinar que los datos no son representativos de los clientes o procesos de negocio, y que por consiguiente debe obtener más datos equilibrados o revisar las suposiciones que son la base de sus expectativas. Las desviaciones estándar y otros valores de distribución pueden proporcionar información útil sobre la estabilidad y exactitud de los resultados. Una desviación estándar grande puede indicar que agregar más datos podría ayudarle a mejorar el modelo. Los datos que se desvían mucho de una distribución estándar se podrían sesgar o podrían representar una imagen precisa de un problema de la vida real, pero dificultar el ajustar un modelo a los datos.

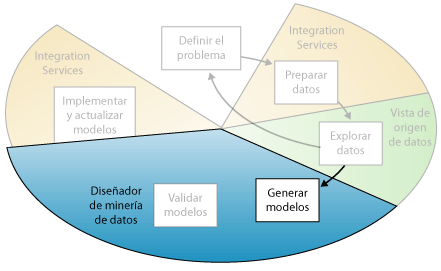
Al explorar los datos para conocer el problema empresarial, puede decidir si el conjunto de datos contiene datos defectuosos y, a continuación, puede inventar una estrategia para corregir los problemas u obtener una descripción más profunda de los comportamientos que son típicos de su negocio.

Puede usar herramientas como Master Data Services para sondear los orígenes de datos disponibles y determinar su disponibilidad para la minería de datos. Puede usar herramientas como SQL Server Data Quality Services, o el generador de perfiles de datos de Integration Services, para analizar la distribución de los datos y solucionar problemas, como la existencia de datos incorrectos o la falta de datos.

Cuando tenga definido los orígenes, combínelos en una vista del origen de datos con el Diseñador de vistas del origen de datos de SQL Server Data Tools.  Este diseñador también contiene algunas herramientas que podrá usar para explorar los datos y comprobar que funcionarán a la hora de crear un modelo.

## Generar modelos

El cuarto paso del proceso de minería de datos, como se resalta en el siguiente diagrama, consiste en generar el modelo o modelos de minería de datos. Usará los conocimientos adquiridos en el paso [Explorar los datos](https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-concepts#ExploringData) para definir y crear los modelos.



Deberá definir qué columnas de datos desea que se usen; para ello, creará una estructura de minería de datos. La estructura de minería de datos se vincula al origen de datos, pero en realidad no contiene ningún dato hasta que se procesa. Al procesar la estructura de minería de datos, Analysis Services genera agregados y otra información estadística que se puede usar para el análisis. Cualquier modelo de minería de datos que esté basado en la estructura puede utilizar esta información.

Antes de procesar la estructura y el modelo, un modelo de minería de datos simplemente es un contenedor que especifica las columnas que se usan para la entrada, el atributo que está prediciendo y parámetros que indican al algoritmo cómo procesar los datos. El procesamiento de un modelo a menudo se denomina *entrenamiento*. El entrenamiento hace referencia al proceso de aplicar un algoritmo matemático concreto a los datos de la estructura para extraer patrones. Los patrones que encuentre en el proceso de entrenamiento dependerán de la selección de los datos de entrenamiento, el algoritmo que elija y cómo se haya configurado el algoritmo. SQL Server 2017 contiene muchos algoritmos diferentes. Cada uno está preparado para un tipo diferente de tarea y crea un tipo distinto de modelo. .

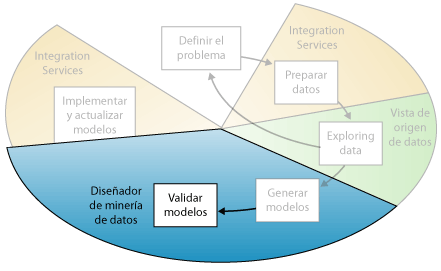
También puede utilizar los parámetros para ajustar cada algoritmo y puede aplicar filtros a los datos de entrenamiento para utilizar un subconjunto de los datos, creando resultados diferentes. Después de pasar los datos a través del modelo, el objeto de modelo de minería de datos contiene los resúmenes y modelos que se pueden consultar o utilizar para la predicción.

Puede definir un modelo nuevo mediante el Asistente para minería de datos de SQL Server Data Tools o con el lenguaje DMX (Extensiones de minería de datos).

Es importante recordar que siempre que los datos cambian, debe actualizar la estructura y el modelo de minería de datos. Al actualizar una estructura de minería de datos volviéndola a procesar, Analysis Services recupera los datos del origen, incluido cualquier dato nuevo si el origen se actualiza dinámicamente, y vuelve a rellenar la estructura de minería de datos.

## Explorar y validar los modelos

El quinto paso del proceso de minería de datos, como se resalta en el siguiente diagrama, consiste en explorar los modelos de minería de datos que ha generado y comprobar su eficacia.



Antes de implementar un modelo en un entorno de producción, es aconsejable probar si funciona correctamente. Además, al generar un modelo, normalmente se crean varios con configuraciones diferentes y se prueban todos para ver cuál ofrece los resultados mejores para su problema y sus datos.

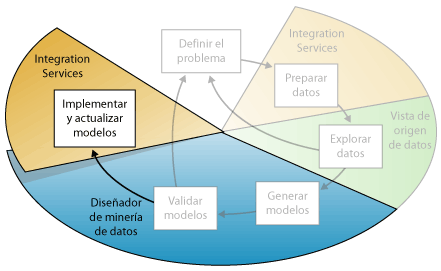
Analysis Services proporciona herramientas que ayudan a separar los datos en conjuntos de datos de entrenamiento y pruebas, para que pueda evaluar con precisión el rendimiento de todos los modelos en los mismos datos. El conjunto de datos de entrenamiento se utiliza para generar el modelo y el conjunto de datos de prueba para comprobar la precisión del modelo mediante la creación de consultas de predicción. Estas particiones pueden hacer automáticamente mientras se genera el modelo de minería de datos.

Puede explorar las tendencias y patrones que los algoritmos detectan mediante los visores del diseñador de minería de datos de SQL Server Data Tools. Para obtener más información, vea [Visores de modelos de minería de datos](https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-model-viewers). También puede comprobar si los modelos crean predicciones correctamente mediante herramientas del diseñador como el gráfico de mejora respecto al modelo predictivo y la matriz de clasificación. Para comprobar si el modelo es específico de sus datos o se puede usar para realizar inferencias en la población general, puede usar la técnica estadística denominada *validación cruzada* para crear automáticamente subconjuntos de los datos y probar el modelo con cada uno.

Si ninguno de los modelos que ha creado en el paso [Generar modelos](https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-concepts#BuildingModels) funciona correctamente, puede que deba volver a un paso anterior del proceso y volver a definir el problema o volver a investigar los datos del conjunto de datos original.

## Implementar y actualizar los modelos

El último paso del proceso de minería de datos, como se resalta en el siguiente diagrama, consiste en implementar los modelos que funcionan mejor en un entorno de producción.



Una vez que los modelos de minería de datos se encuentran en el entorno de producción, puede llevar acabo diferentes tareas, dependiendo de sus necesidades. Las siguientes son algunas de las tareas que puede realizar:

* Use los modelos para crear predicciones que luego podrá usar para tomar decisiones comerciales. SQL Server pone a su disposición el lenguaje DMX, que podrá usar para crear consultas de predicción, y el Generador de consultas de predicción, que le ayudará a generar las consultas.
* Crear consultas de contenido para recuperar estadísticas, reglas o fórmulas del modelo.
* Incrustar la funcionalidad de minería de datos directamente en una aplicación. Puede incluir Objetos de administración de análisis (AMO), que contiene un conjunto de objetos que la aplicación pueda utilizar para crear, cambiar, procesar y eliminar estructuras y modelos de minería de datos. También puede enviar mensajes XML for Analysis (XMLA) directamente a una instancia de Analysis Services.
* Utilizar Integration Services para crear un paquete en el que se utilice un modelo de minería de datos para dividir de forma inteligente los datos entrantes en varias tablas. Por ejemplo, si una base de datos se actualiza continuamente con clientes potenciales, puede utilizar un modelo de minería de datos junto con Integration Services para dividir los datos entrantes en clientes que probablemente compren un producto y clientes que probablemente no compren un producto.
* Crear un informe que permita a los usuarios realizar consultas directamente en un modelo de minería de datos existente.
* Actualizar los modelos después de la revisión y análisis. Cualquier actualización requiere que vuelve a procesar los modelos.
* Actualizar dinámicamente los modelos, cuando entren más datos en la organización, y realizar modificaciones constantes para mejorar la efectividad de la solución debería ser parte de la estrategia de implementación.

**IMPORTANCIA DE LA DATA MINING**

En el ámbito comercial, resulta interesante encontrar patrones ocultos de consumo de los clientes para poder explorar nuevos horizontes. Saber por ejemplo, que un vehículo deportivo corre un riesgo de accidente casi igual al de un vehículo normal cuando su dueño tiene un segundo vehículo en casa ayuda a crear nuevas estrategias comerciales para ese grupo de clientes. Asimismo, predecir el comportamiento de un futuro cliente, basándose en los datos históricos de clientes que presentaron el mismo perfil, ayuda a poder retenerlo durante el mayor tiempo posible.

Vivimos permanentemente bombardeados por informes sobre la importancia del conocimiento y la información para obtener ventajas competitivas en un mundo globalizado. Las actividades de vigilancia e inteligencia, en el contexto de la "sociedad de la Información", se considera como "la transformación de la información en conocimiento, y del conocimiento en acción". Es por esto la importancia del "análisis de la información" en el seno de todas las actividades en donde se trata de transformar los datos brutos con el fin de extraer los conocimientos que puedan ser explotados y útiles en un determinado campo de acción.

**DESCUBRIMIENTO DE CONOCIMIENTO Y MINERÍA DE DATOS**

El término descubrimiento de conocimiento en bases de datos (knowledge discovery in databases, KDD para abreviar) se refiere 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.

Proceso KDD

1. Pre-procesamiento de Datos: Limpieza, integración y transformación.
2. Minería de Datos: Uso de métodos inteligentes para extraer conocimiento (búsqueda de oro) .
3. Evaluación de patrones encontrados y presentación

Los principales pasos dentro del proceso interactivo e iterativo del KDD pueden verse en la figura 3

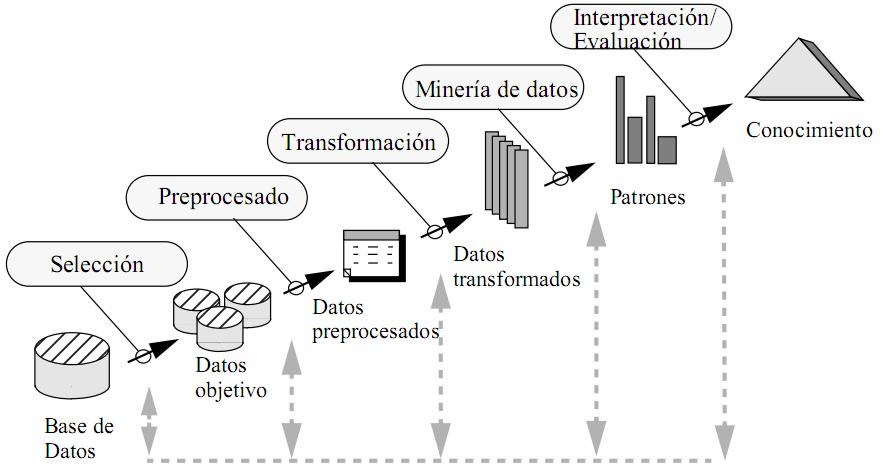


Figura 3.

**Historia del KDD**

En los últimos años, ha existido un gran crecimiento en nuestras capacidades de generar y colectar datos, debido básicamente al gran poder de procesamiento de las máquinas como a su bajo costo de almacenamiento. Sin embargo, dentro de estas enormes masas de datos existe una gran cantidad de información oculta, de gran importancia estratégica, a la que no se puede acceder por las técnicas clásicas de recuperación de la información. El descubrimiento de esta información oculta es posible gracias a la Minería de Datos (DataMining), que entre otras sofisticadas técnicas aplica la inteligencia artificial para encontrar patrones y relaciones dentro de los datos permitiendo la creación de modelos, es decir, representaciones abstractas de la realidad, pero es el descubrimiento del conocimiento (KDD, por sus siglas en inglés) que se encarga de la preparación de los datos y la interpretación de los resultados obtenidos, los cuales dan un significado a estos patrones encontrados. Así el valor real de los datos reside en la información que se puede extraer de ellos, información que ayude a tomar decisiones o mejorar nuestra comprensión de los fenómenos que nos rodean. Hoy, más que nunca, los métodos analíticos avanzados son el arma secreta de muchos negocios exitosos. Empleando métodos analíticos avanzados para la explotación de datos, los negocios incrementan sus ganancias, maximizan la eficiencia operativa, reducen costos y mejoran la satisfacción del cliente.

En la figura 4 se ilustra la jerarquía que existe en una base de datos entre datos, información y conocimiento. Se observa igualmente el volumen que presenta en cada nivel y el valor que los responsables de las decisiones le dan en esa jerarquía. El área interna dentro del triángulo representa los objetivos que se han propuesto. La separación del triángulo representa la estrecha unión entre dato e información, no así entre la información y el conocimiento.

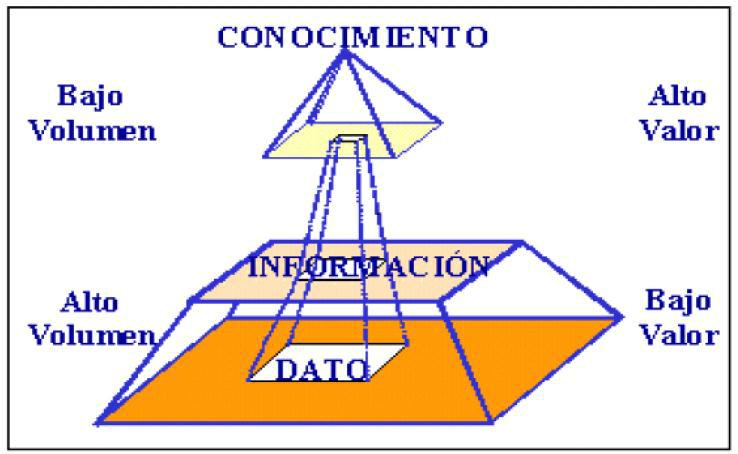


Figura 4: Jerarquía del Conocimiento. **Relación entre dato, información y** **conocimiento (Molina, 1998).**

**Metas**

Las metas del KDD son:

* Procesar automáticamente grandes cantidades de datos crudos.
* Identificar los patrones más significativos y relevantes.
* Presentarlos como conocimiento apropiado para satisfacer las metas del usuario.

**Relación con otras disciplinas**

* KDD nace como interfaz y se nutre de diferentes disciplinas:
* Sistemas de información / bases de datos: tecnologías de bases de datos y bodegas de datos, maneras eficientes de almacenar, accesar y manipular datos.
* Estadística, aprendizaje automático / IA (redes neuronales, lógica difusa, algoritmos genéticos, razonamiento probabilístico): desarrollo de técnicas para extraer conocimiento a partir de datos.
* Reconocimiento de patrones: desarrollo de herramientas de clasificación.
* Visualización de datos: interfaz entre humanos y datos, y entre humanos y patrones.
* Computación paralela / distribuida: cómputo de alto desempeño, mejora de desempeño de algoritmos debido a su complejidad y a la cantidad de datos.
* Interfaces de lenguaje natural a bases de datos.

**El proceso de KDD**

El proceso de KDD consiste en usar métodos de minería de datos (algoritmos) para extraer (identificar) lo que se considera como conocimiento de acuerdo a la especificación de ciertos parámetros usando una base de datos junto con preprocesamiento y post-procesamientos. En la figura 3. Se ilustra el proceso de KDD.

Se estima que la extracción de patrones (minería) de los datos ocupa solo el 15% - 20% del esfuerzo total del proceso de KDD. El proceso de descubrimiento de conocimiento en bases de datos involucra varios pasos:

* Determinar las fuentes de información: que pueden ser útiles y dónde conseguirlas.
* Diseñar el esquema de un almacén de datos (Data Warehouse): que consiga unificar de manera operativa toda la información recogida.
* Implantación del almacén de datos: que permita la navegación y visualización previa de sus datos, para discernir qué aspectos puede interesar que sean estudiados. Esta es la etapa que puede llegar a consumir el mayor tiempo.
* Selección, limpieza y transformación de los datos que se van a analizar: la selección incluye tanto una criba o fusión horizontal (filas) como vertical (atributos).La limpieza y preprocesamiento de datos se logra diseñando una estrategia adecuada para maneja ruido, valores incompletos, secuencias de tiempo, casos extremos (si es necesario), etc.
* Seleccionar y aplicar el método de minería de datos apropiado: esto incluye la selección de la tarea de descubrimiento a realizar, por ejemplo, clasificación, agrupamiento o clustering, regresión, etc. La selección de él o de los algoritmos a utilizar. La transformación de los datos al formato requerido por el algoritmo específico de minería de datos. Y llevar a cabo el proceso de minería de datos, se buscan patrones que puedan expresarse como un modelo o simplemente que expresen dependencias de los datos, el modelo encontrado depende de su función (clasificación) y de su forma de representarlo (árboles de decisión, reglas, etc.), se tiene que especificar un criterio de preferencia para seleccionar un modelo dentro de un conjunto posible de modelos, se tiene que especificar la estrategia de búsqueda a utilizar (normalmente está predeterminada en el algoritmo de minería)
* Evaluación, interpretación, transformación y representación de los patrones extraídos:

Interpretar los resultados y posiblemente regresar a los pasos anteriores. Esto puede involucrar repetir el proceso, quizás con otros datos, otros algoritmos, otras metas y otras estrategias. Este es un paso crucial en donde se requiere tener conocimiento del dominio. La interpretación puede beneficiarse de procesos de visualización, y sirve también para borrar patrones redundantes irrelevantes.

* Difusión y uso del nuevo conocimiento. Incorporar el conocimiento descubierto al sistema (normalmente para mejorarlo) lo cual puede incluir resolver conflictos potenciales con el conocimiento existente. El conocimiento se obtiene para realizar acciones, ya sea incorporándolo dentro de un sistema de desempeño o simplemente para almacenarlo y reportarlo a las personas interesadas. En este sentido, KDD implica un proceso interactivo e iterativo involucrando la aplicación de varios algoritmos de minería de datos.

**DATA MINING**

Aunque desde un punto de vista académico el término data mining es una etapa dentro de un proceso mayor llamado extracción de conocimiento en bases de datos (Knowledge Discovery in Databases o KDD) en el entorno comercial, ambos términos se usan de manera indistinta. Lo que en verdad hace el data mining es reunir las ventajas de varias áreas como la Estadística, la Inteligencia Artificial, la Computación Gráfica, las Bases de Datos y el Procesamiento Masivo, principalmente usando como materia prima las bases de datos. Una definición tradicional es la siguiente: "la integración de un conjunto de áreas que tienen como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten un sesgo hacia la toma de decisión".

La idea de data mining no es nueva. Ya desde los años sesenta los estadísticos manejaban términos como data fishing, data mining o data archeology con la idea de encontrar correlaciones sin una hipótesis previa en bases de datos con ruido. A principios de los años ochenta, Rakesh Agrawal, Gio Wiederhold, Robert Blum y Gregory Piatetsky-Shapiro, entre otros, empezaron a consolidar los términos de data mining y KDD. A finales de los años ochenta sólo existían un par de empresas dedicadas a esta tecnología; en 2002 existen más de 100 empresas en el mundo que ofrecen alrededor de 300 soluciones. Las listas de discusión sobre este tema las forman investigadores de más de ochenta países. Esta tecnología ha sido un buen punto de encuentro entre personas pertenecientes al ámbito académico y al de los negocios.

El data mining es una tecnología compuesta por etapas que integra varias áreas y que no se debe confundir con un gran software. Durante el desarrollo de un proyecto de este tipo se usan diferentes aplicaciones software en cada etapa que pueden ser estadísticas, de visualización de datos o de inteligencia artificial, principalmente. Actualmente existen aplicaciones o herramientas comerciales de data mining muy poderosas que contienen un sinfín de utilerías que facilitan el desarrollo de un proyecto. Sin embargo, casi siempre acaban complementándose con otra herramienta. La data mining es la etapa de descubrimiento en el proceso de KDD: Paso consistente en el uso de algoritmos concretos que generan una enumeración de patrones a partir de los datos preprocesados (Fayyad et al., 1996) Aunque se suelen usar indistintamente los términos KDD y Minería de Datos.

Los proyectos de Data Mining tienen por objetivo extraer información útil a partir de grandes cantidades de datos y se aplican a todos los sectores y en todos los campos. Así existen proyectos de este tipo en sectores tan dispares como el comercio electrónico, la banca, las empresas industriales o la exploración petrolífera. La extracción de esta información útil es un proceso complejo, que requiere la aplicación de una metodología estructurada para la utilización ordenada y eficiente de las técnicas y herramientas disponibles.

La minería de datos (DM, Data Mining) consiste en la extracción no trivial de [información](http://es.wikipedia.org/wiki/Informaci%C3%B3n) que reside de manera implícita en los [datos.](http://es.wikipedia.org/wiki/Dato) Dicha información era previamente desconocida y podrá resultar útil para algún proceso. En otras palabras, la minería de datos prepara, sondea y explora los datos para sacar la información oculta en ellos.

Bajo el nombre de minería de datos se engloba todo un conjunto de técnicas encaminadas a la extracción de conocimiento procesable, implícito en las [bases de](http://es.wikipedia.org/wiki/Base_de_datos) [datos.](http://es.wikipedia.org/wiki/Base_de_datos) Está fuertemente ligado con la supervisión de procesos industriales ya que resulta muy útil para aprovechar los datos almacenados en las bases de datos.

Las bases de la minería de datos se encuentran en la [inteligencia artificial](http://es.wikipedia.org/wiki/Inteligencia_artificial) y en el análisis [estadístico.](http://es.wikipedia.org/wiki/Estad%C3%ADstica) Mediante los [modelos](http://es.wikipedia.org/wiki/Modelo) extraídos utilizando técnicas de minería de datos se aborda la solución a problemas de [predicción,](http://es.wikipedia.org/wiki/Predicci%C3%B3n) [clasificación](http://es.wikipedia.org/wiki/Clasificaci%C3%B3n) y [segmentación.](http://es.wikipedia.org/wiki/Segmentaci%C3%B3n)

Las **categorías básicas** de las técnicas de minería de datos actualmente en uso se pueden clasificar en: clasificación, asociación, secuencia y cluster.

* Clasificación: incluye los procesos de minería de datos que buscan reglas para definir si un ítem o un evento pertenecen a un *subset* particular o a una clase de datos. Esta técnica, probablemente la más utilizada, incluye dos subprocesos: la construcción de un modelo y la predicción.

En términos generales, los métodos de clasificación desarrollan un modelo compuesto por reglas IF-THEN y se aplican perfectamente, por ejemplo, para encontrar patrones de compra en las bases de datos de los clientes y construir mapas que vinculan los atributos de los clientes con los productos comprados. Con un conjunto apropiado de atributos predictivos, el modelo puede identificar los clientes con mayor propensión a realizar una determinada compra durante el próximo mes. Un caso típico de clasificación es el de dividir una base de datos de compañías en grupos homogéneos respecto a variables como "posibilidades de crédito" con valores tales como "bueno" y "malo".

* Asociación: incluye técnicas conocidas como linkage analysis, utilizadas para buscar patrones que tienen una probabilidad alta de repetición, como ocurre al analizar una canasta en la búsqueda de productos afines.

Se desarrolla un algoritmo asociativo que incluye las reglas que van a correlacionar un conjunto de eventos con otro. Por ejemplo, un supermercado podría necesitar información sobre los hábitos de compra de sus clientes, para reubicar los productos que se suelen comprar juntos, para localizar los productos nuevos en el mejor lugar, o para ofrecer promociones.

* Secuencia: los métodos de análisis de series de tiempo son usados para relacionar los eventos con el tiempo.

A título ilustrativo: como resultado de este tipo de modelo se puede aprender que las personas que alquilan una película de video tienden a adquirir los productos promocionales durante las siguientes dos semanas; o bien, que la adquisición de un horno de microondas se produce frecuentemente luego de determinadas compras previas.

* Cluster: Muchas veces resulta difícil o imposible definir los parámetros de una clase de datos. En ese caso, los métodos de clustering pueden usarse para crear particiones, de forma tal que los miembros de cada una de ellas resulten similares entre sí, según alguna métrica o conjunto de métricas.

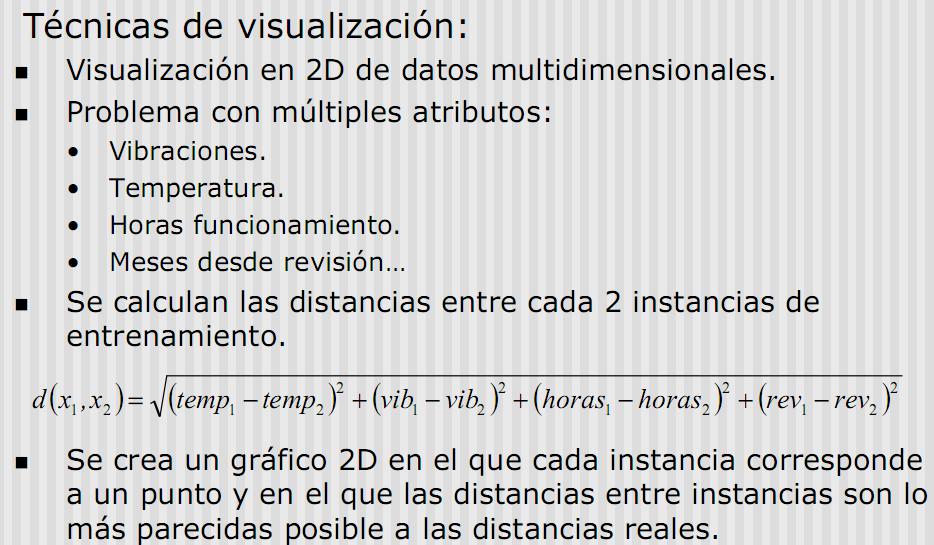
El análisis de clusters podría utilizarse, entre otras aplicaciones, al estudiar las compras con tarjetas de crédito, para descubrir digamos que los alimentos comprados con una tarjeta dorada de uso empresarial son adquiridos durante los días de semana y tienen un valor promedio de ticket de 152 pesos, mientras que el mismo tipo de compra, pero realizado con una tarjeta platino personal, ocurre predominantemente durante los fines de semana, por un valor menor, pero incluye una botella de vino más del 65 % de las veces.

Como puede verse, en el mundo de la data mining las preguntas que se pueden hacer y responder son infinitas y las metodologías utilizadas para responderlas son diversas y cada vez más variadas. Así como existen muchas técnicas para dar soporte al proceso de decidir con la minería de datos, existen variadas tecnologías para construir los modelos

**Grupos de técnicas principales**

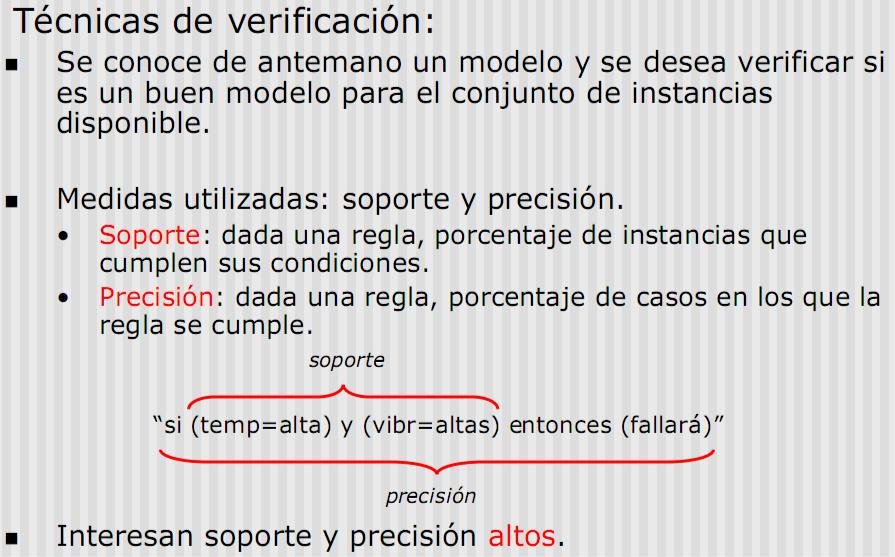
**Visualización.**

* Ayudas para el descubrimiento manual de información.
* Se muestran tendencias, agrupamientos de datos, etc.
* Funcionamiento semi-automático.



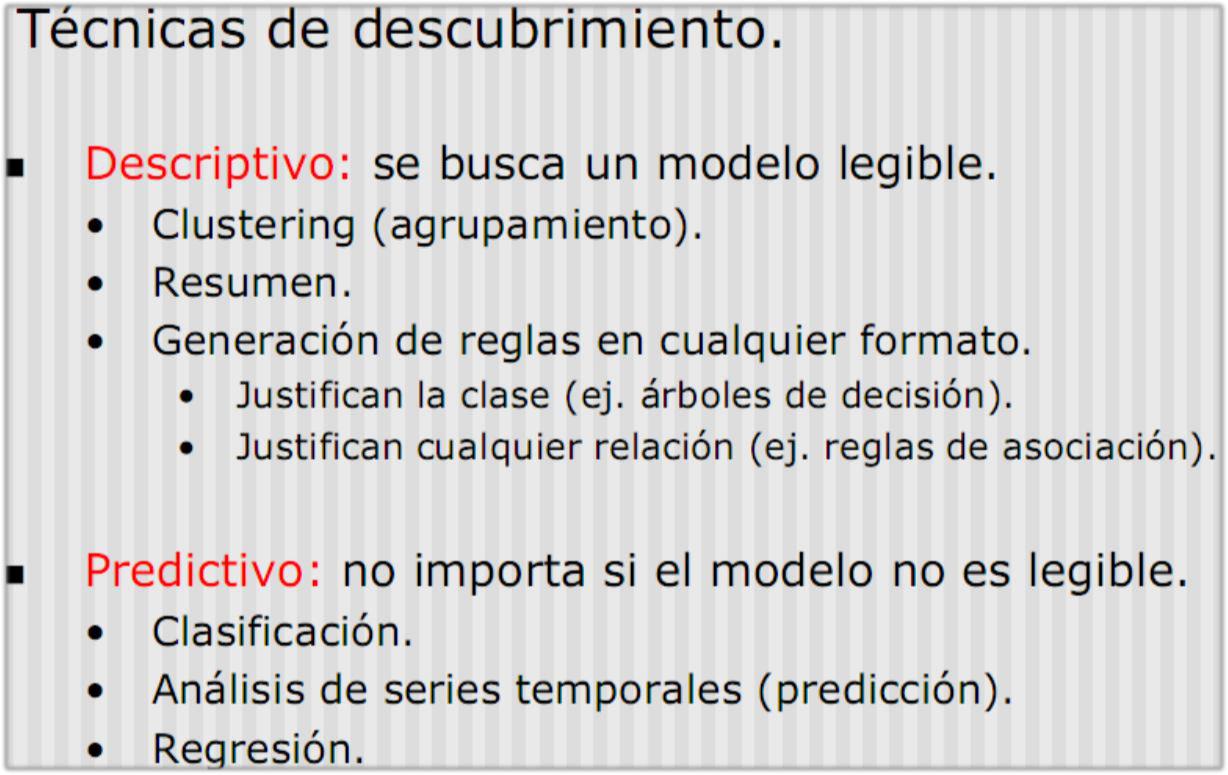
**Verificación.**

* Se conoce de antemano un modelo y se desea saber si los datos disponibles se ajustan a él.
* Se establecen medidas de ajuste al modelo.



**Descubrimiento.**

* Se busca un modelo desconocido de antemano.
* Descubrimiento descriptivo: se busca modelo legible.
* Descubrimiento predictivo: no importa que el modelo no sea legible.



Sea como sea la presentación del problema, una de las características presente en cualquier tipo de aprendizaje y en cualquier tipo de técnica de Minería de Datos es su carácter hipotético, es decir, lo aprendido puede, en cualquier momento, ser refutado por evidencia futura.

En muchos casos, los modelos no aspiran a ser modelos perfectos, sino modelos aproximados. En cualquier caso, al estar trabajando con hipótesis, es necesario realizar una evaluación de los patrones obtenidos, con el objetivo de estimar su validez y poder compararlos con otros. Por tanto, la Minería de Datos, más que verificar patrones hipotéticos, usa los datos para encontrar estos patrones. Por lo tanto, es un proceso inductivo.

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

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 gran- des 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.