TEMA 1: Introducción a la minería de datos

1.1. Introducción. Descubriendo conocimiento

La **minería de datos** es una disciplina que estudia el análisis de grandes cantidades de datos con el objetivo de obtener **conocimiento** a partir de ellos. Se trata, pues, una nueva tecnología de manejo y análisis de información que aprovecha la capacidad existente hoy día de procesamiento, almacenamiento y transmisión de datos a gran velocidad y bajo costo.

El origen del término está en otros de la década de 1960, cuando en Estadística se hacía referencia a términos como *data fishing* (buscar o "pescar" en los datos) o *data dredging* (dragado de datos) para referirse al análisis de grandes volúmenes de datos. En la década de 1990, aparece el término **data mining** (minería de datos[1]) con el que se conoce actualmente al proceso de obtención de conocimiento a partir de los datos por medio de su análisis. La aparición de la minería de datos como disciplina ha tenido mucho que ver con el cambio de concepción de los datos, unido a la gran cantidad de estos que se generan y almacenan continuamente en cualquier ámbito. Otro factor que también ha favorecido la consolidación de la minería de datos como disciplina es el gran avance en los últimos tiempos en las prestaciones y capacidad computacional.

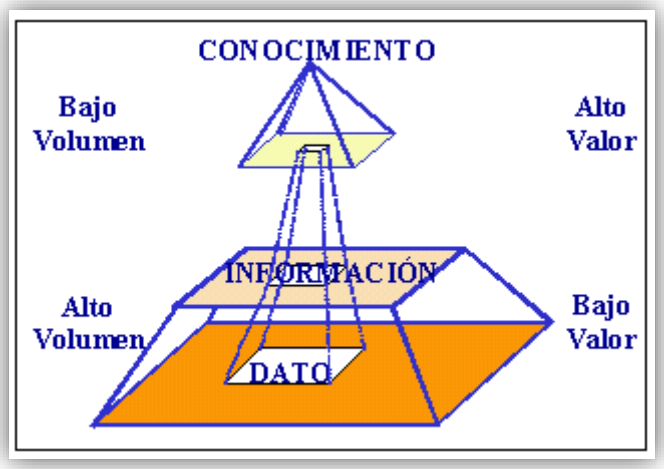
Tradicionalmente, las técnicas de minería de datos se aplicaban sobre información contenida en almacenes de datos. No obstante, actualmente está cobrando una importancia cada vez mayor la minería de datos desestructurados como es la información contenida en ficheros de texto (text mining), en Internet (web mining), etc. Además, hoy en día han surgido otras necesidades de tipo operativo, como la integración de los resultados obtenidos en los sistemas de información en línea, con la exigencia, por tanto, de que los procesos funcionen prácticamente en tiempo real; por ejemplo, la alerta temprana frente a alarmas en una cadena de montaje, la detección instantánea del fraude en operaciones bancarias, un sistema de recomendación de productos en una tienda en línea, etc.

Cuatro son los factores, pues, que nos han llevado a la aparición y desarrollo de esta disciplina en auge; a saber:

1. El abaratamiento de los sistemas de almacenamiento tanto temporal como permanente.
2. El incremento de las velocidades de cómputo en los procesadores.

[1] Quizá la expresión correcta atendiendo a su principal objetivo de obtención de conocimiento a partir de los datos sea la de *Minería de conocimiento*, ya que es este el que se desea extraer, del mismo modo que denominamos *Minería del carbón* a la obtención de este en una mina.

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.



Todas estas ventajas están haciendo que en la actualidad se abuse del almacenamiento de gran volumen de información en cualquier ámbito, que bien analizada, puede proporcionar en conjunto un verdadero conocimiento de gran ayuda en la toma de decisiones.
La figura de la izquierda ilustra la jerarquía que existe en una base de datos entre dato, 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. El data mining trabaja en el nivel superior buscando patrones, comportamientos, agrupaciones, secuencias, tendencias o asociaciones que puedan generar algún modelo que nos permita comprender mejor el dominio para ayudar en una posible toma de decisión.

Con todo lo anterior podemos decir que Data Mining es el proceso de descubrir patrones de información interesante y potencialmente útiles, inmersos en una gran base de datos en la que se interactúa constantemente. Data Mining es una combinación de procesos como:

- ✓ Extracción de datos
- ✓ Limpieza de datos.
- ✓ Selección de características.
- ✓ Algoritmos.
- ✓ Análisis de resultados.

Una definición tradicional es la siguiente:

"Un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos".

Desde otro punto de vista se define como

"*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 decisiones*".

La minería de datos no surge como un área completamente nueva, sino más bien como la mezcla de conceptos procedentes de otras muchas y diferentes disciplinas; a saber:

- **Estadística:** muchas de las técnicas que se aplican en la minería de datos son o tienen su raíz en la Estadística. Se podría decir que la Estadística es la *madre* de la minería de datos.
- **Bases de datos:** el proceso de KDD parte de datos que, habitualmente, se encuentran almacenados en bases de datos.
- **Visualización:** el objetivo final de la minería de datos es obtener conocimiento que sea útil. Para lograrlo, es un requisito fundamental que ese conocimiento pueda ser visualizado por los expertos de cada dominio. De ahí la importancia de las técnicas de visualización (diagramas, gráficos, resúmenes, etc.)
- **Aprendizaje automático:** se encuentra profundamente ligado con la minería de datos, ya que ambos, de alguna manera, persiguen la obtención de modelos por medio de mecanismos automáticos.
- **Otras:** sistemas de apoyo a la decisión, recuperación de información, procesamiento de señales, etc.

La utilidad de Data Mining se puede dar dentro de los siguientes aspectos:

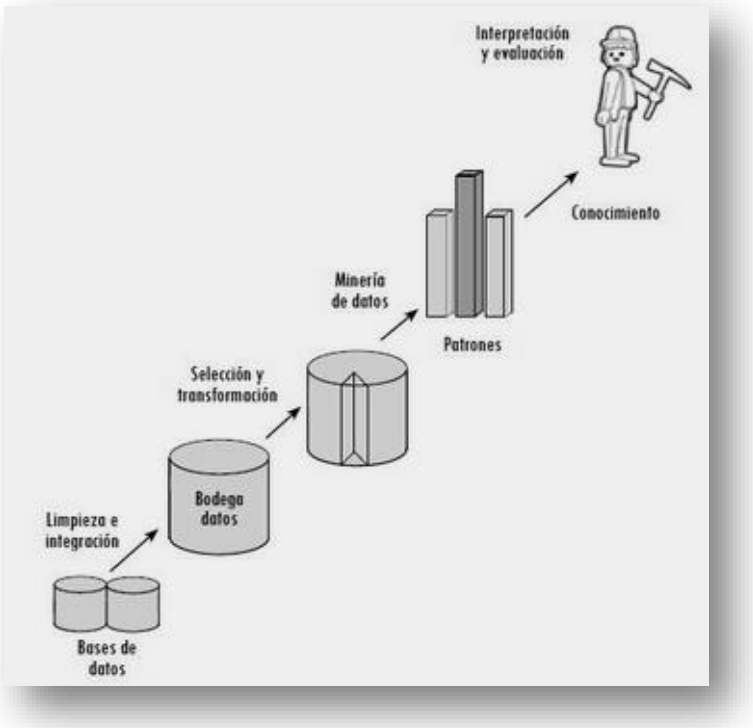
- ✓ *Sistemas parcialmente desconocidos:* En estos casos habrá una parte del sistema que es conocida y habrá una parte aparentemente de naturaleza aleatoria. Bajo ciertas circunstancias, a partir de una gran cantidad de datos asociada con el sistema, existe la posibilidad de encontrar nuevos aspectos previamente desconocidos del modelo.
- ✓ *Enorme cantidad de datos:* al contar con mucha información, es importante encontrar la forma de analizarla y que ello produzca algún tipo de beneficio.
- ✓ *Potente hardware y software:* la capacidad y disponibilidad computacional ha aumentado considerablemente el desempeño del proceso de buscar y analizar información (el cual a veces debe vérselas con producciones de datos del orden de los Gbytes/hora)

1.2. El proceso de KDD

La idea de data mining no es nueva. Ya desde los años sesenta los estadísticos manejaban términos como data fishing, data mining o data archaeology con la idea de encontrar correlaciones sin una hipótesis previa en bases de datos con ruido. A principios de los años ochenta, RakeshAgrawal, GioWiederhold, Robert Blum y Gregory Piatetsky-Shapiro, entre otros, empezaron a consolidar los términos de data mining y **KDD**o descubrimiento de conocimiento en bases de datos(en inglés, ***KnowledgeDiscovery*** *in* ***Databases***, más conocido como **proceso de KDD**). La minería de datos es una etapa de dicho proceso. Realmente, es la más importante.

Según la definición comúnmente aceptada, el **proceso de KDD** persigue la extracción automatizada de conocimiento no trivial, implícito, previamente desconocido y potencialmente útil a partir de grandes volúmenes de datos.

Tras la preparación, los datos pasan a la fase de minería de datos, en la que se aplican una serie de técnicas para obtener modelos (representaciones simbólicas de la realidad que representan los datos de entrada) para, finalmente, ser validados e interpretados y así poder obtener de ellos el ansiado conocimiento. Es éste un proceso iterativo, no lineal, que se retroalimenta.



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 facilitan el desarrollo de un proyecto. Sin embargo, casi siempre acaban complementándose con otra herramienta

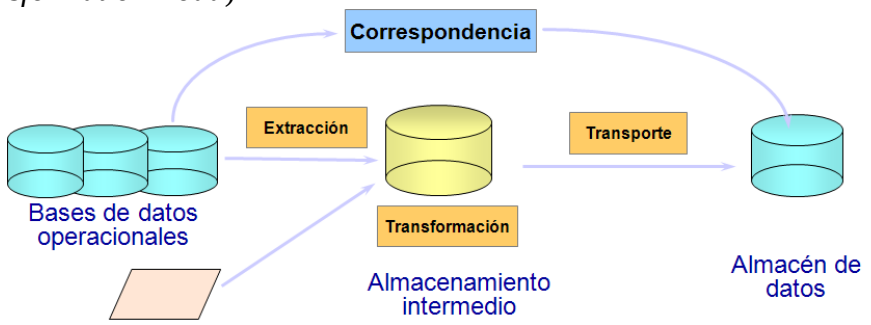
Las cuatro características que ha de poseer el conocimiento extraído por el proceso de KDD son:

- ✓ **No trivial.** De nada sirve extraer conocimiento conocido por todos o que carezca de importancia.
- ✓ **Implícito.** Se encuentra oculto en los datos.
- ✓ **Previamente desconocido.** Nada nuevo aporta si el conocimiento extraído ya había sido descubierto anteriormente.
- ✓ **Útil.** El conocimiento extraído debe servir para algo, de lo contrario no tiene ningún sentido invertir esfuerzos en extraerlo.

El proceso de KDD se compone de las siguientes fases:

i. **Recopilación de datos.** En esta fase, los datos, procedentes de diferentes fuentes, se integran en un mismo y único repositorio de datos, denominado *almacén de datos*, más conocido como ***data warehouse***. El resultado final de esta fase es, precisamente, ese data warehouse.

   Los datos generalmente se encuentran en múltiples fuentes diseñadas según esquemas desnormalizados y guardan información en torno a *hechos*, cada uno de los cuales está caracterizado por una serie de *dimensiones*, esquema que se denomina **modelo multidimensional**.
   Para integrarlas en un mismo almacén, es necesario disponer de un proceso que lea los datos de las diferentes fuentes, los limpie y los adecúe a la estructura que tiene el data warehouse para su almacenamiento. Este tipo de proceso se lleva a cabo mediante un sistema conocido como **sistema ETL** (*Extraction-Transformation-Load*)



   Una vez almacenados los datos, es posible tener que volver a repetir el proceso ETL para integrar nuevas fuentes de datos.

ii. **Filtrado de los datos y selección de variables.**El formato de los datos contenidos en la fuente de datos nunca es el idóneo, y la mayoría de las veces no es posible utilizar ningún algoritmo de minería. En esta fase se realiza una selección de los datos integrados en el data warehouse, pues dichos datos pueden no estar limpios, contener atributos irrelevantes, etc. Y se limpian y transforman de cara a fases posteriores.

   Las técnicas de selección tienen por objeto filtrar aquellos datos que no son relevantes para el análisis posterior, filtrado que se puede realizar a varios niveles:

   - **Filtrado de atributos.** Es posible que algunos de los atributos de los datos a analizar no sean de interés.
   - **Filtrado de registros.** En ocasiones, el objetivo puede ser eliminar algunos de los registros almacenados y quedarse sólo con los relevantes, pues con un subconjunto menos de registros (**muestra**) se podría hacer un análisis igual de efectivo, pero mucho más eficiente desde el punto de vista computacional. En este caso, se suelen aplicar técnicas de muestreo, como muestreo aleatorio simple, aleatorio estratificado o muestreo de grupos.

Por su parte, Las tareas de limpieza de datos van, normalmente, encaminadas a resolver dos problemas bastantes habituales:

- **La ausencia de valores.** Es muy habitual que, para muchos de los registros analizados, falte cierta información (datos faltantes o *missing values*) que, en ocasiones, aportan información interesante. Ante esta situación, se pueden adoptar diferentes alternativas:
    - Pasar por alto el valor faltante y continuar con el análisis.
    - Filtrar toda la columna asociada a dicho atributo.
    - Filtrar el registro que contiene el valor faltante.
    - Asignar un valor al atributo en cuestión, mediante uno de los posibles procedimientos de imputación automática.
- **La existencia de valores erróneos.** También suele ser común encontrar valores que, claramente, son erróneos. Aunque existen diferentes técnicas para detectar este tipo de valores (especificación de *edits*), la realidad es que, la mayoría de las veces, se realiza mediante procedimientos artesanales "ad-hoc". Una vez localizados los valores erróneos, las opciones más comunes para su tratamiento son las siguientes:
    - Pasar por alto el valor erróneo y continuar con el análisis.
    - Filtrar toda la columna asociada al valor erróneo.
    - Filtrar el registro que contiene el valor erróneo.
    - Reemplazar el valor erróneo por un valor correcto, seguramente mediante el uso de alguna técnica específica de predicción.

Algunos autores consideran la identificación de objetos atípicos como un problema de detección de valores erróneos.

El resultado de esta fase un subconjunto limpio y transformado de los datos sobre el que ya se puede aplicar las técnicas de data mining en la siguiente fase.

Finalmente, las técnicas de transformación de datos ofrecen soluciones a problemas que se pueden presentar como que los datos se encuentren en un determinado formato no adecuado para el uso de los diferentes algoritmos del data mining, formateando los datos según se necesite.

Existen múltiples técnicas de transformación de datos. Algunas de las más aplicadas son:

- **Numerización.** Consiste en transformar un atributo de tipo cualitativo en uno equivalente cuantitativo.
- **Discretización.** Consiste en transformar un atributo cuantitativo en uno cualitativo ordinal.
- **Creación de características.** Consiste en la creación de un nuevo atributo en los datos, normalmente calculado como función de otros atributos ya existentes.

- **Normalización.** Consiste en la transformación del rango de valores que tomaun determinado atributo. El caso más común de normalización es la *normalización lineal uniforme*, que transforma los valores de un atributo a una escala uniforme en el intervalo [0,1], utilizando $Valor\ normalizado = \frac{Valor\ inicial - Valor\ mínimo}{Valor\ máximo - valor\ mínimo}$.
- **Reducción de la dimensionalidad**. Las técnicas de reducción de la dimensionalidad buscan reducir el número de atributos sobre los que realizar el análisis posterior. Para ello, existen múltiples técnicas, aunque quizá la más conocida es la técnica de **análisis de componentes principales (PCA)**, que proyecta los atributos iniciales en un espacio de dimensionalidad mucho menor, de forma que en los nuevos atributos recogen la mayor parte de la información relevante de los originales, pero con la ventaja adicional de que se eliminan las posibles redundancias y dependencias que había. Esto permite un análisis más eficiente de los datos en términos de coste computacional.

Mediante el preprocesado, se filtran los datos (se eliminan valores incorrectos, no válidos, desconocidos, etc.), se obtienen muestras de los mismos (mayor velocidad de respuesta del proceso), o se reducen el número de valores posibles (mediante redondeo, agrupamiento, etc.)

Aún después de haber sido preprocesados, se sigue teniendo 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 dos:

1. Los basados en la elección de los mejores atributos del problema.
2. Los que buscan variables independientes mediante tests de sensibilidad, algoritmos de distancia o heurísticos.

iii. **Extracción de conocimiento (Data mining).** El siguiente paso consiste en aplicar técnicas concretas de minería de datos para obtener modelos.

Una vez seleccionados, limpiados y transformados los datos, se obtiene el denominado conjunto de *datos minables*, y el siguiente paso del proceso de KDD consiste en aplicar técnicas de data mining para obtener modelos que representen a dichos datos.

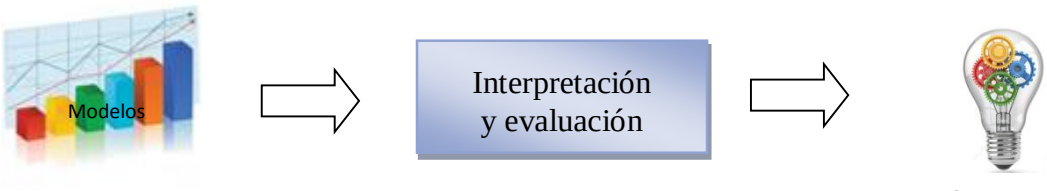
En la etapa del data mining se pueden aplicar diferentes técnicas para resolver diferentes tipo de problemas, a los que se les conoce aquí con el nombre de **tareas**, que se suelen clasificar dependiendo del tipo de modelo que son capaces de generar; a saber:

- **Tareas predictivas.** Son aquellas que se utilizan para predecir el valor desconocido de uno o varios atributos para uno o varios registros de losdatos minables. Entre ellas se encuentran, entre otras:
  - *Clasificación.* Consiste en encontrar un modelo que, aplicado a un nuevo ejemplo sin clasificar, lo clasifique dentro de un conjunto predefinido de clases. Normalmente, el atributo a predecir es de tipo cualitativo, y recibe el nombre de ***atributo de clase***.
  - *Regresión.* Es similar a la de clasificación, con la diferencia de que, en este caso, el atributo es de tipo cuantitativo.
- **Tareas descriptivas.** Son aquellas que generan modelos que, de alguna forma, describen los datos, sin llevar a cabo ningún tipo de predicción. Entre las más importantes se encuentran:
  - *Clustering.* Pretende dividir una población heterogénea de objetos en grupos homogéneos, denominados ***clústeres*** de forma que los objetos de cada grupo sean muy similares entre sí. También se denomina *segmentación o agrupamiento.*
  - *Asociación.* Pretende encontrar reglas que muestran la relación que existe entre los distintos atributos de los datos analizados, denominadas ***reglas de asociación***.
  - *Detección de atípicos.* Consiste en encontrar objetos que, dentro de un conjunto, manifiesten características significativamente diferentes a las del resto de los objetos del conjunto.

Para abordar cada una de las tareas anteriores, existen numerosas **técnicas** o **algoritmos**; y además de las anteriores, existen muchas tareas específicas para tipos de datos no convencionales.

Mediante una técnica 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.

iv. **Interpretación y evaluación.** Los modelos obtenidos en la fase anterior han de ser evaluados. Tras la obtención de los modelos de data mining, el último paso del proceso de KDD consiste en evaluar la calidad de dichos modelos y realizar una interpretación de los mismos para obtener el conocimiento buscado.



La evaluación de los modelos obtenidos en la fase anterior es una tarea crucial. No todos los modelos obtenidos cumplirán con las características esperadas en ellos para poder obtener conocimiento. En particular, los modelos han de ser precisos, comprensibles e interesantes.

Aunque dependerá de la tarea de data mining en cuestión, en general, para evaluar un modelo se suele utilizar un enfoque consistente en reservar un pequeño subconjunto de los datos (***conjunto de prueba***) que se utilizará para validar el modelo construido con el resto de los datos (***conjunto de entrenamiento***), denominado **validación simple**.

Una técnica algo más avanzada, a la vez que más utilizada, es la técnica de validación cruzada ***n-fold cross validation***. En este caso, para validar un modelo, se elige aleatoriamente el *n%* de los datos como conjunto de prueba y con el *(100-n)%* restante, como conjunto de entrenamiento, se construye el modelo; y el proceso se repite *n* veces variando cada vez los conjuntos de prueba y entrenamiento. Un valor muy habitual para *n* es 10.

En general, para cada tarea de data mining se utilizan unas métricas específicas que miden la calidad de los modelos obtenidos con ellas:

- **Clasificación.** Para evaluar un modelo de clasificación se mide su *precisiónpredictiva* o porcentaje de clasificaciones acertadas en el conjunto de prueba frente al total de las clasificaciones realizadas.
- **Regresión.** En este caso, la medida típica que se suele emplear es el error cuadrático medio.
- **Reglas de asociación.** En este caso se suele medir el porcentaje de las instancias que la regla predice correctamente.
- **Clustering.** La calidad de una segmentación de objetos en clústeres se suele medir por medio de la cohesión de los clústeres; es decir, mediante alguna métrica que, efectivamente, mida si los objetos de cada clúster son similares entre sí y diferentes a los objetos de otros clústeres.

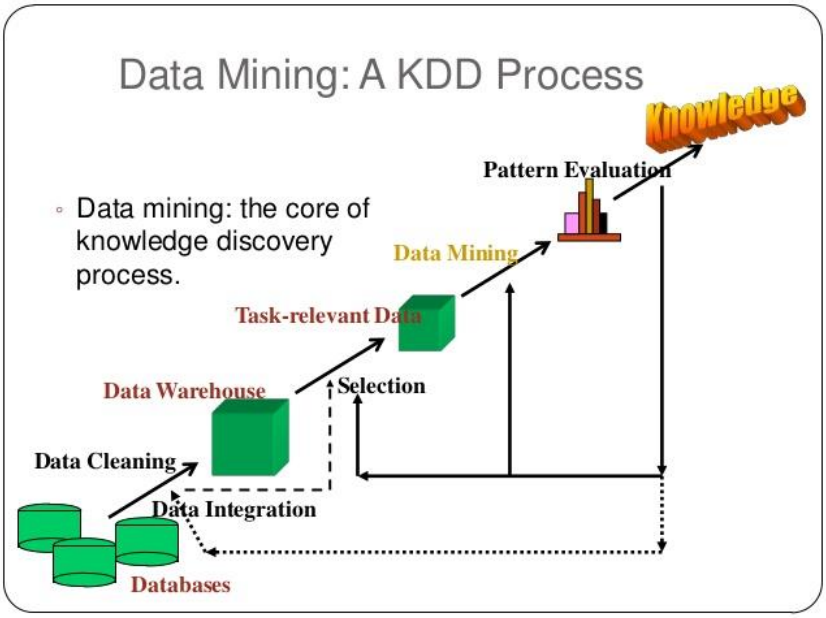
Una vez conocida la calidad de los modelos, es necesario expresarlos en términos del área de aplicación, para lo que es importante contar con técnicas de visualización de dichos modelos para que sean comprendidos e interpretados por los expertos de cada dominio. De esa manera, los expertos serán capaces de contrastar los modelos obtenidos con su propia visión de la realidad y transformarlos en conocimiento que será utilizado y difundido para el avance del área en cuestión.

Recogiendo todo lo expuesto sobre el proceso de KDD, tenemos la siguiente tabla resumen:

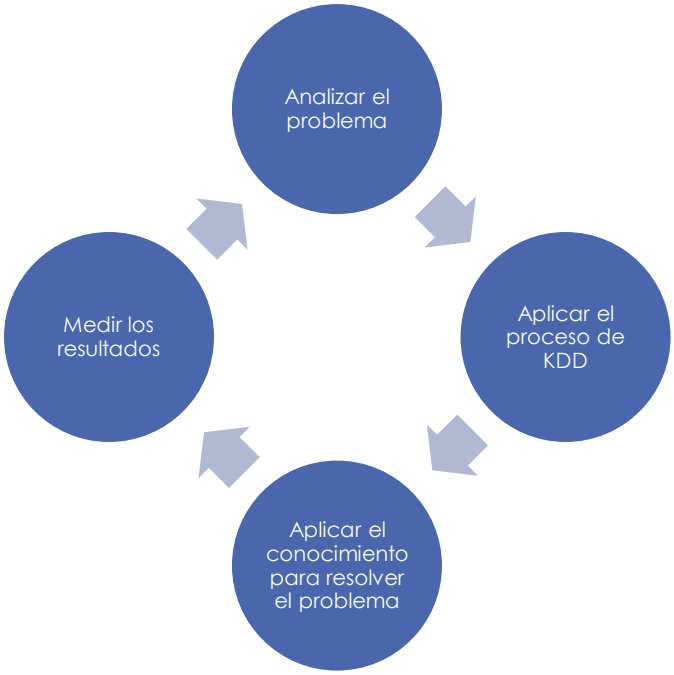
Etapa o fase	Entrada	Salida	Técnicas
Recopilación de los datos	Datos originales	Almacén de datos	Técnicas de diseño de data warehouse (modelo multidimensional)
Selección, limpieza y transformación de los datos	Almacén de datos	Datos seleccionados, limpios y transformados (minables)	Técnicas de selección (muestreo), de limpieza (tratamiento de datos faltantes) y de transformaciones (discretización, normalización, etc.)
Data mining	Datos minables	**Modelos** de data mining	**Algoritmos** para resolver las diferentes tareas de clasificación (C.45), clustering (K-medias),..
Interpretación y evaluación	Modelos de data mining	**Conocimiento**	Técnicas de interpretación y de evaluación (validación cruzada)

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, se alterará alguno de los procesos anteriores en busca de nuevos modelos.

Según algunos autores, la fase ii. se puede descomponer en varias fases, si bien lo importante es saber que el objetivo de la misma es preparar los datos para poder realizar data mining con ellos en la fase siguiente. El gráfico siguiente ilustra estas fases.



Una característica importante del proceso de KDD es su naturaleza iterativa. Esto significa que es posible tener que aplicar varias veces el proceso de KDD hasta obtener el conocimiento deseado. El gráfico ilustrativo de este ciclo podría ser del siguiente tipo:



1.3. Los datos

La información primaria de la minería de datos no es estrictamente la almacenada en las denominadas bases de datos relacionales (que también), sino la que se encuentra en un denominado sistema de almacenamiento desnormalizado. Y el proceso de KDD utiliza diferentes tipos de datos:

- **Cuantitativos:** cuyos valores representan magnitudes (discretos y continuos)
- **Cualitativos:** cuyos valores representan una categoría y no una cantidad (nominales y ordinales)
- **Otros:** necesitarán enfoques particulares
    - *Series temporales:* sucesiones de valores que representan la evolución de una determinada característica a lo largo de un periodo de tiempo, con mediciones, generalmente, a intervalos de tiempo regulares.
    - *Datos espaciales:* que representan la estructura espacial de algún objeto.
    - *Datos multimedia:* imágenes, videos, elementos de audio, etc.
    - *Documentos:* descripciones textuales de objetos como, por ejemplo, resúmenes, obras literarias completas, etc.
    - *Datos procedentes de la web:* información acerca de la estructura de los sitios web, de los patrones de navegación de los usuarios, etc.

1.4. Utilidad y aplicaciones de la Minería de Datos

La minería de datos puede resultar útil en una gran cantidad de situaciones; algunos problemas específicos que puede resolver son los siguientes:

- Encontrar grupos o tipologías de objetos, lo que recibe el nombre de **clustering**.
- Analizar y obtener modelos que se puedan utilizar de cara al futuro, lo que se conoce como técnicas de **clasificación** (por ejemplo, con minería de datos, al analizar una tabla se podría obtener un conjunto de reglas que recomiendan o no la concesión de un crédito, etc)
- Detectar grupos de variables o valores que están asociados (por ejemplo: los clientes de un supermercado que compran pan y azúcar, habitualmente también compran leche, etc.), lo que se conoce como **técnicas de asociación**.

Por lo que respecta a las aplicaciones, en general, la minería de datos se puede aplicar y resulta útil en casi cualquier dominio, siempre y cuando haya una cantidad suficiente de datos de los que extraer conocimiento. Presentamos aquí algunos problemas típicos que abordaría la minería de datos en diferentes dominios:

Dominio	Problemas abordados con minería de datos
Negocios	✓ Fidelización de clientes. ✓ Publicidad personalizada. ✓ Captación de nuevos clientes. ✓ Aumento del volumen de ventas. ✓ Estudio de las tipologías de clientes.
Banca y finanzas	✓ Detección del uso fraudulento de tarjetas de crédito. ✓ Estudio de la concesión de créditos a clientes. ✓ Predicción de la evolución de un valor bursátil.
Compañías de seguros	✓ Detección de fraudes y simulaciones. ✓ Estudio de la concesión de coberturas a los clientes en función de sus características.
Supermercados	✓ Análisis de la cesta de la compra (identificar productos que se compran juntos) ✓ Ubicación de productos dentro del supermercado. ✓ Campañas de publicidad dirigidas.
Educación	✓ Predicción de la calificación de los estudiantes. ✓ Mejora del proceso de enseñanza-aprendizaje.
Medicina	✓ Ayuda al diagnóstico de enfermedades. ✓ Estudio de la evolución de pacientes. ✓ Estudio de la efectividad de un tratamiento.
Biología, genética y otras ciencias	✓ Estudio de las secuencias de genes en busca de patrones significativos. ✓ Predicción de catástrofes naturales. ✓ Predicción meteorológica.
Internet	✓ Análisis del comportamiento de los usuarios en la web. ✓ Estudio del contenido y estructura de los sitios web. ✓ Detección de correo basura (spam) ✓ Identificación de comportamientos fraudulentos en comercio electrónico. ✓ Radio personalizada en internet: Last.fm ✓ Flikr
Gobiernos	✓ "El FBI analizará las bases de datos comerciales para detectar terroristas"
Empresariales	✓ Detección de fraudes en las tarjetas de crédito ✓ Migración de clientes entre distintas compañías ✓ Predicción del tamaño de las audiencias televisivas ✓ Supermercados Wal-mart
Universidad	✓ ¿Llevan a cabo los recién titulados de una universidad actividades profesionales relacionadas con sus estudios?
Investigación espacial	✓ Proyecto SKYCAT
Textos	✓ Text Mining

1.5. Retos y tendencias de la Minería de Datos

La capacidad para almacenar datos ha crecido en los últimos años a velocidades exponenciales. En el otro extremo, nuestra capacidad para procesar esta enorme cantidad de datos para por utilizarlos eficazmente no ha ido a la par. Por este motivo, el 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.

Algunos de los retos y tendencias actuales de la minería de datos son:

- **El tratamiento de grandes volúmenes de datos.** La presencia de gran volumen de información dificulta cada vez más el tratamiento de los datos. Disponer de técnicas capaces de tratar esos grandes volúmenes de datos de alta dimensionalidad es uno de los desafíos más importantes de la minería de datos.
- **Logro de tiempos de respuesta bajos.** Estos sistemas, en ocasiones, se utilizan como sistemas de apoyo a la decisión en ámbitos donde el tiempo de respuesta es un factor esencial.
- **Funcionamiento en línea y en tiempo real.** Es el caso, por ejemplo, de un supuesto sistema de detección de fraude en las operaciones realizadas con tarjetas de crédito.
- **Tratamiento de datos con una estructura compleja.** Otro reto es el de la obtención de conocimiento de datos no convencionales (radiodiagnóstico, imágenes de satélites, etc.)
- **Automatización de tareas.** Retos como la automatización del procesado de datos puede tener consecuencias como reducción de tiempos y gastos.

La tendencia en minería de datos es, precisamente, la propuesta de soluciones encaminadas a dar respuesta a los retos descritos anteriormente, entre otros.

Existen iniciativas, como la propuesta por Kaggle, una plataforma *on-line* que propone competiciones de minería de datos. Proporciona un repositorio para que las compañías publiquen sus datos. A partir de ahí, comienza un concurso abierto para que los expertos en minería de datos de todo el mundo descarguen esos datos y propongan soluciones a los problemas de la compañía en cuestión. La mejor solución se hace con un premio cuantioso.