San Marcos

MIEMBRO DE LA RED ILUMNO

ANÁLISIS EXPLORATORIO DE DATOS Y REPROCESAMIENTO

ANÁLISIS EXPLORATORIO DE DATOS Y REPROCESAMIENTO (MINERÍA DE DATOS)

INTRODUCCIÓN A DATA MINING

DEFINICIONES Y CONCEPTOS

En estos tiempos, las sociedades son consideradas como la sociedad de la información, donde las tecnologías de hoy nos ayudan a la creación, distribución y manipulación de la Información, facilitan las actividades sociales, culturales y económicas de nuestros trabajos.

Según Senge (1990) la Teoría de la gestión del conocimiento,

> La gestión y la aplicación del conocimiento es un concepto actualmente aplicado en las organizaciones y las empresas, que procura compartir el conocimiento y la experiencia de sus recursos humanos, de tal formal que quede disponible y pueda ser utilizado por otros miembros de dicha organización. (p. 40)

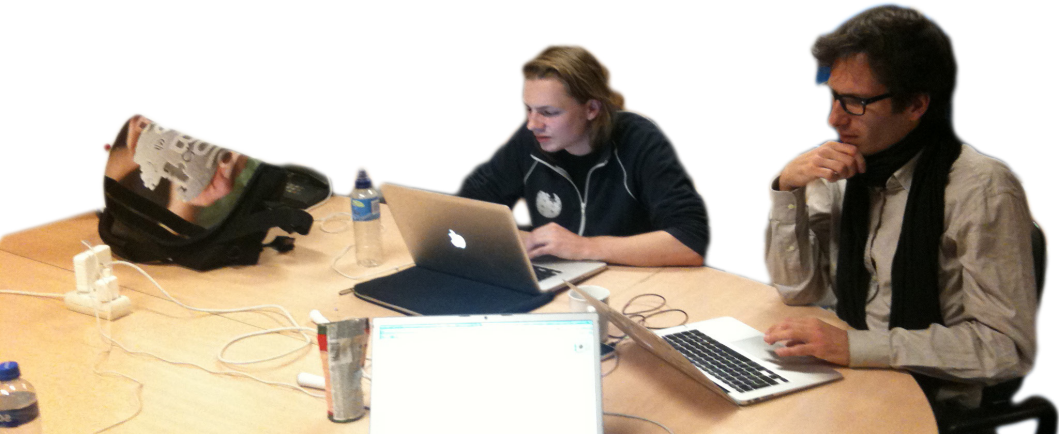
Dentro de los servicios de minería de datos deben contemplar en sus procesos cuatro etapas muy importantes para su desarrollo:

- **La preparación de la información**, acondicionamiento de los datos y análisis previo de los datos de partida sobre los que se generarán los modelos de información.
- **La modelización en sí misma**, entendiendo en esta etapa la construcción de los modelos mediante el procesado de la información de partida.
- **La validación de los modelos generados**, tanto desde un punto de vista técnico como de negocio.
- **La puesta en producción y aplicación de los modelos en el entorno final**, ya sea informacional u operacional.

Antes de continuar debemos manejar algún glosario sobre las terminologías de la *Data Mining:*

- **Algoritmos genéticos:** "técnicas de optimización que usan procesos tales como combinación genética, mutación y selección natural en un diseño basado en los conceptos de evolución natural." (Benzecri, 2015, p. 22)
- **Algoritmo de fuerza bruta (en inglés *Brute Force Algorithm*):** "técnica que utiliza la repetición exhaustiva de pasos simples con el fin de encontrar una solución óptima. Está en contraste con técnicas más complejas más caras y difíciles de construir pero mucho más eficientes." (Benzecri, 2015, p. 22)
- **Análisis de series de tiempo *(time-series)*:** "análisis de una secuencia de medidas hechas a intervalos específicos. El tiempo es usualmente la dimensión dominanate de los datos. " (Benzecri, 2015, p. 22)
- **Análisis prospectivo de datos:** "análisis de datos que predice futuras tendencias, comportamientos o eventos basado en datos históricos." (Benzecri, 2015, p. 22)
- **Análisis exploratorio de datos:** "uso de técnicas estadísticas tanto gráficas como descriptivas para aprender acerca de la estructura de un conjunto de datos." (Benzecri, 2015, p. 22)
- **Análisis retrospectivo de datos:** "análisis de datos que provee una visión de las tendencias, comportamientos o eventos basado en datos históricos." (Benzecri, 2015, p. 22)

- **Arbol de decisión:** "estructura en forma de árbol que representa un conjunto de decisiones. Estas decisiones generan reglas para la clasificación de un conjunto de datos. Ver *CART y CHAID*." (Benzecri, 2015, p. 22)
- ***Assembled to order* (ATO):** "estrategia que permite a un producto o servicio ser hecho bajo órdenes específicas, así un gran número de productos puede ser hecho a partir de un número limitado de componentes comunes. Esto exige una planeación sofisticada de los procesos para anticiparse a la demanda cambiante para componentes internos o accesorios mientras se enfoca en el ensamblaje final del producto para proveer un producto hecho a la medida para los usuarios." (Benzecri, 2015, p. 22)
- **Arriendo de vehículo *full service*:** "es un sistema que le proporciona al cliente un vehículo y una variedad de servicios de apoyo con un solo pago del arriendo mensual. Los arriendos de servicio full service pueden incluir características como el mantenimiento preventivo, atención de emergencia y reparaciones en el camino, evaluaciones de equipo y especificaciones, combustible, apoyo administrativo, apoyo al conductor y programas de seguridad." (Benzecri, 2015, p. 22)
- **Arriendo del camión:** "una transacción a corto plazo, generalmente de doce meses que le permiten el uso de un camión por un período especificado de tiempo a un cliente, generalmente medido en "días de arriendo". El arriendo puede usarse para complementar una flota privada o arrendada durante períodos cortos de alta necesidad, para ejecutar órdenes rápidas o volumen en exceso, o para probar nuevas rutas y cauces de la distribución." (Benzecri, 2015, p. 23)
- **Arriendo financiado:** "a menudo, un acuerdo de pleno-pago en el cual el cliente, al final del término del arriendo, asume propiedad del vehículo o se proporciona con una opción de compra. El arrendatario es normalmente responsable por gastos de mantenimiento, impuestos y seguros." (Benzecri, 2015, p. 23)
- ***Application Programming Interface* (API):** "interficie de lenguaje de programación (que relaciona o permite extender el programa)." (Benzecri, 2015, p. 23)
- **Base de datos multidimensional:** "base de datos diseñada para procesamiento analítico *on-line (OLAP)*; estructurada como un hipercubo con un eje por dimensión." (Benzecri, 2015, p. 23)
- ***Backhaul:*** "el movimiento del retorno de un vehículo de su destino hacia atrás a su punto de origen con una carga útil." (tecnologías-información.com).



- ***Benchmarking:*** "el proceso de comparar el desempeño contra las prácticas de otras compañías, con el propósito de mejorar la actuación. Las compañías también pueden hacer una referencia interna. Rastreando y comparando la actuación actual con actuaciones del pasado." (tecnologías-información.com)
- ***Bill of Lading (BOL o B/L):*** "un contrato de envío entre un cargador (el consignador) para depositar una carga a un portador o entregar en otra parte (el consignatario)." (Benzecri, 2015, p. 24)
- ***CART* Árboles de clasificación y regresión:** "una técnica de árbol de decisión usada para la clasificación de un conjunto da datos. Provee un conjunto de reglas que se pueden aplicar a un nuevo (sin clasificar) conjunto de datos para predecir cuáles registros darán un cierto resultado. Segmenta un conjunto de datos creando 2 divisiones. Requiere menos preparación de datos que *CHAID*." (Benzecri, 2015, p. 24)
- ***CHAID* Detección de interacción automática de Chi cuadrado:** "una técnica de árbol de decisión usada para la clasificación de un conjunto da datos. Provee un conjunto de reglas que se pueden aplicar a un nuevo (sin clasificar) conjunto de datos para predecir cuáles registros darán un cierto resultado. Segmenta un conjunto de datos utilizando tests de chi cuadrado para crear múltiples divisiones. Antecede y requiere más preparación de datos que *CART*." (Benzecri, 2015, p. 24)
- **Clasificación:** "proceso de dividir un conjunto de datos en grupos mutuamente excluyentes de tal manera que cada miembro de un grupo esté lo "más cercano" posible a otro, y grupos diferentes estén lo "más lejos" posible unos de otros, donde la distancia está medida con respecto a variable(s) específica(s), las cuales se están tratando de predecir. Por ejemplo, un problema típico de clasificación es el de dividir una base de datos de compañías en grupos que son lo más homogéneos posibles con respecto a variables como "posibilidades de crédito" con valores tales como "Bueno" y "Malo"." (Benzecri, 2015, p. 24)
- ***Clustering* (agrupamiento):** "proceso de dividir un conjunto de datos en grupos mutuamente excluyentes de tal manera que cada miembro de un grupo esté lo "más cercano" posible a otro, y grupos diferentes estén lo "más lejos" posible unos de los otros, donde la distancia está medida con respecto a todas las variables disponibles." (Benzecri, 2015, p. 24)
- **Computadoras con multiprocesadores:** "una computadora que incluye múltiples procesadores conectados por una red. Ver procesamiento paralelo." (Benzecri, 2015, p. 24)

- ***Data cleansing:*** "proceso de asegurar que todos los valores en un conjunto de datos sean consistentes y correctamente registrados." (Benzecri, 2015, p. 24)
- ***Data Mining:*** "la extracción de información predecible escondida en grandes bases de datos." (Benzecri, 2015, p. 24)
- ***Data Warehouse:*** "sistema para el almacenamiento y distribución de cantdades masivas de datos". (Benzecri, 2015, p. 24)
- **Datos anormales:** "datos que resultan de errores (por ej.: errores en el tipeado durante la carga) o que representan eventos inusuales." (Benzecri, 2015, p. 24)
- **Dimensión:** "en una base de datos relacional o plana, cada campo en un registro representa una dimensión. En una base de datos multidimensional, una dimensión es un conjunto de entidades similares; por ejemplo: una base de datos multidimensional de ventas podría incluir las dimensiones producto, tiempo y ciudad." (Benzecri, 2015, p. 24)
- **Modelo analítico:** "una estructura y proceso para analizar un conjunto de datos. Por ejemplo, un árbol de decisión es un modelo para la clasificación de un conjunto de datos." (Benzecri, 2015, p. 24)
- **Modelo lineal:** "un modelo analítico que asume relaciones lineales entre una variable seleccionada (dependiente) y sus predictores (variables independientes)." (Benzecri, 2015, p. 24)
- **Modelo no lineal:** "un modelo analítico que no asume una relación lineal en los coeficientes de las variables que son estudiadas." (Benzecri, 2015, p. 24)
- **Modelo predictivo:** "estructura y proceso para predecir valores de variables especificadas en un conjunto de datos." (Benzecri, 2015, p. 24)
- **Navegación de datos:** "proceso de visualizar diferentes dimensiones, "fetas" y niveles de una base de datos multidimensional. Ver OLAP. " (Benzecri, 2015, p. 24)
- **OLAP Procesamiento analítico *on-line (On Line Analitic prossesing)*:** "Se refiere a aplicaciones de bases de datos orientadas a array que permite a los usuarios ver, navegar, manipular y analizar bases de datos multidimensionales." (Benzecri, 2015, p. 24)
- ***Outlier:*** "un ítem de datos cuyo valor cae fuera de los límites que encierran a la mayoría del resto de los valores correspondientes de la muestra. Puede indicar datos anormales. Deberían ser examinados detenidamente; pueden dar importante información." (Benzecri, 2015, p. 24)

- **Procesamiento paralelo:** "uso coordinado de múltiples procesadores para realizar tareas computacionales. El procesamiento paralelo puede ocurrir en una computadora con múltiples procesadores o en una red de estaciones de trabajo o PCs." (Benzecri, 2015, p. 25)
- ***Raid:*** "formación redundante de discos baratos (*Redundant Array of inexpensive disks).* Tecnología para el almacenamiento paralelo eficiente de datos en sistemas de computadoras de alto rendimiento." (Benzecri, 2015, p. 25)
- **Regresión lineal:** "técnica estadística utilizada para encontrar la mejor relación lineal que encaja entre una variable seleccionada (dependiente) y sus predicados (variables independientes)." (Benzecri, 2015, p. 25)
- **Regresión logística:** "una regresión lineal que predice las proporciones de una variable seleccionada categórica, tal como "tipo de consumidor" en una población." (Benzecri, 2015, p. 25)
- **Vecino más cercano:** "técnica que clasifica cada registro en un conjunto de datos basado en una combinación de las clases del/de los k registro (s) más similar/es a él en un conjunto de datos históricos (donde k ? 1). Algunas veces se llama la técnica del vecino k-más cercano." (Benzecri, 2015, p. 25)
- **SMP Multiprocesador simétrico *(Symmetric multiprocessor)*:** "tipo de computadora con multiprocesadores en la cual la memoria es compartida entre los procesadores". (Benzecri, 2015, p. 25)
- ***Stock Keeping Unit (SKU)*:** "sistema de numeración que hace a un producto o artículo discernible de todos los otros." (Benzecri, 2015, p. 25)
- **Precisión (en inglés *Accuracy*):** "se define como la medida de un modelo predictivo que refleja la proporción número de veces que el modelo es correcto cuando se aplica a los datos." (Benzecri, 2015, p. 25)
- **Inteligencia artificial (en inglés *Artificial Intelligence*):** "campo de la ciencia que concierne a la creación de comportamiento inteligente en una máquina." (Benzecri, 2015, p. 25)
- **Red neuronal artificial (en inglés *Artificial Neural Network (ANN)*:** véase red neuronal.
- **Regla de asociación (en inglés *Association Rule*):** "regla en la forma "si esto entonces" que asocia acontecimientos en una base de datos. Por ejemplo, hábitos de compra." (Benzecri, 2015, p. 25)
- **Retropropagación (en inglés *Back Propagation*):** "uno de los algoritmos más comunes en la formación de redes neuronales consistente consiste en minimizar un error (comúnmente cuadrático) por medio de gradiente descendiente." (Benzecri, 2015, p. 25)
- **3PL *(Third Party Logistic)*:** "transportación, almacenaje y otros servicios relacionados con la logística, que son proporcionados por compañías empleadas para asumir tareas que previamente fueron realizadas por el cliente." (Benzecri, 2015, p. 25

LOS CONCEPTOS DE DATO, INFORMACIÓN Y CONOCIMIENTO DE LA INFORMACIÓN ESTADÍSTICA

De acuerdo con Ramírez y Perri (2004) algunos sistemas que son solo parcialmente conocidos, producen una cantidad inmensa de datos, datos que con frecuencia contienen información valiosa que puede resultar muy útil a ejecutivos de una empresa a la hora de la toma de decisiones y de resolver problemas de negocio como:

- Procesos y análisis de procesos.
- Detección de anomalías (fraudes).
- Gestión de riesgos.
- Segmentación de clientes.
- Personalización de la publicidad.
- Previsión.

Esto nos ha indicado que usualmente en estos procesos informáticos y de análisis de datos implica una variedad de técnicas para capturar información, organizarla y almacenar el conocimiento del personal de la organización para transformarlo en un activo intelectual que brinde beneficios y se pueda compartir. Las tecnologías de la información permiten contar con herramientas que apoyan en el proceso de toma de desición, la recolección, la transferencia y la administración sistemática de la información, esto con el proceso de la capacitación del profesional que interpreta la información obtenida.

EL OBJETIVO DE LA MINERÍA DE DATOS ES OFRECER INFORMACIÓN AL PROFESIONAL QUE ANALIZA A LAS EMPRESAS PARA MEJORAR SUS OPERATIVAS POR MEDIO DE UN MAYOR ENTENDIMIENTO DE SU ENTORNO EN EL MERCADO.

Todo este proceso implica que la información suministrada por este conjunto de herramientas sea clave de la organización para apoyar la toma de decisiones y reducir el riesgo vinculado a tomar decisiones equivocadas durante el proceso. (p. 137)

En el ámbito de las nuevas tecnologías de la información y de actividades de consultoría relacionadas con el tema de la inteligencia competitiva de las organizaciones, la gestión del conocimiento cobra una importancia vital. La administración del conocimiento se ha convertido en un asunto primordial en las empresas ya que se han percatado de que una gran parte de su valor como entidades que brindan un servicio de valor agregado al usuario, depende de la capacidad de las mismas para crear y administrar el conocimiento. Existen estudios que han determinado que una parte importante del valor de una organización, se relaciona con sus activos intangibles, de los cuales el conocimiento es un activo fundamental.

Es el conocer la gestión del conocimiento de información, se refiere al conjunto de procesos de negocios desarrollados en una organización para crear, almacenar, transferir y aplicar el conocimiento; incluye una variedad de técnicas en sus distintas fases entre las que se encuentra la minería de datos o ***Data Mining***.

En este trabajo se explorará la relación entre la minería de datos y cómo esta última contribuye al proceso total de la dicha gestión de análisis y de procesos.

LA MINERÍA DE DATOS Y LAS ETAPAS DE LA GESTIÓN DEL CONOCIMIENTO

De acuerdo con Méndez (2011) *Data Mining*, también denominada extracción de datos, es la práctica por medios automáticos o semiautomáticos de la búsqueda y la exploración en grandes almacenes de datos de relaciones no visualizadas previamente, dando por resultado el descubrimiento de patrones significativos entre los mismos y reglas. Para lograr este propósito, la *Data Mining* emplea técnicas estadísticas, de automatización del conocimiento y de reconocimiento de patrones (observar datos de una sola fuente, recursos de información, otros.).

PROCESO KDD (*KNOWLEDGE DISCOVERY DATABASES*), CARACTERÍSTICAS Y FACES

La extracción de conocimiento está principalmente relacionada con el proceso de descubrimiento conocido como *Knowledge Discovery in Databases (KDD)*, que se refiere al proceso no-trivial de descubrir conocimiento e información potencialmente útil dentro de los datos contenidos en algún repositorio de información. No es un proceso automático, es un proceso iterativo que exhaustivamente explora volúmenes muy grandes de datos para determinar relaciones. Es un proceso que extrae información de calidad que puede usarse para dibujar conclusiones basadas en relaciones o modelos dentro de los datos.

EL OBJETIVO ES BRINDAR INFORMACIÓN AL NEGOCIO ASISTIENDO A LAS EMPRESAS PARA MEJORAR SUS OPERACIONES POR MEDIO DE UN MAYOR ENTENDIMIENTO DE SU ENTORNO.

La siguiente figura ilustra las etapas del proceso KDD. (Hernández, Ramírez y Perri, 2004, p. 235)

Figura 1. Fases del proceso KDD
Fuente. Elaboración Propia.

PROCESO KDD

Como muestra la figura anterior, las etapas del proceso KDD se dividen en 5 fases y son:

1. **Selección de datos.** En esta etapa se determinan las fuentes de datos y el tipo de información a utilizar. Es la etapa donde los datos relevantes para el análisis son extraídos desde la o las fuentes de datos.
2. **Preprocesamiento.** Esta etapa consiste en la preparación y limpieza de los datos extraídos desde las distintas fuentes de datos en una forma manejable, necesaria para las fases posteriores. En esta etapa se utilizan diversas estrategias para manejar datos faltantes o en blanco, datos inconsistentes o que están fuera de rango, obteniéndose al final una estructura de datos adecuada para su posterior transformación.
3. **Transformación.** Consiste en el tratamiento preliminar de los datos, transformación y generación de nuevas variables a partir de las ya existentes con una estructura de datos apropiada. Aquí se realizan operaciones de agregación o normalización, consolidando los datos de una forma necesaria para la fase siguiente.

4. ***Data Mining.*** Es la fase de modelamiento propiamente tal, en donde métodos inteligentes son aplicados con el objetivo de extraer patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles y que están contenidos u "ocultos" en los datos.
5. **Interpretación y evaluación.** Se identifican los patrones obtenidos, basándose en algunas medidas, y se realiza una evaluación de los resultados obtenidos.

Además de las fases descritas, frecuentemente se incluye una fase previa de análisis de las necesidades de la organización y definición del problema, en la que se establecen los objetivos de la minería de datos. De acuerdo con Laudon (2008), también es usual incluir una etapa final donde los resultados obtenidos se integran al negocio para la realización de acciones comerciales. (p. 123)

En un próximo artículo se revisarán cada una de estas fases, incluyendo la fase previa de identificación de objetivos y la fase final de integración al negocio, de tal forma que todas las etapas son:

1. **IDENTIFICACIÓN DE LOS OBJETIVOS**
2. **SELECCIÓN DE DATOS E INFORMACIÓN**
3. **PREPROCESAMIENTO**
4. ***DATA MINING***
5. **IDENTIFICACIÓN DE LOS OBJETIVOS**
6. **INTERPRETACIÓN Y EVALUACIÓN**
7. **INTEGRACIÓN AL NEGOCIO**

VALIDACIÓN DE DATOS

Estimar la precisión de un clasificador inducido por algoritmos de aprendizaje supervisado es importante tanto para evaluar su futura precisión de clasificación como para elegir un clasificador óptimo de un conjunto dado.

Para probar un modelo se parten los datos en dos conjuntos. Por un lado, se tiene el conjunto de entrenamiento o training set. Este grupo de instancias serviría para enseñar al modelo cuál es el comportamiento tipo del sistema, haciéndose una clasificación por el analista de dichas instancias. Por otro lado, se tiene el conjunto de prueba o test set que será el conjunto sobre el que se aplicarán los métodos una vez adquirido el conocimiento previo a través del *training set*.

De acuerdo con Breiman, Friedman, Olsen y Stone (1984), si no se usa esta metodología, la precisión del modelo será sobrestimada, es decir, tendremos estimaciones muy optimistas. (Laudon, 2008, p. 128.)

Se pueden establecer tres tipos fundamentales de métodos de validación:

1. VALIDACIÓN SIMPLE

Utiliza un conjunto de muestras para construir el modelo del clasificador, y otro diferente para estimar el error, con el fin de eliminar el efecto de la sobreestimación.

> Entre la variedad de porcentajes utilizados, uno de los más frecuentes es tomar 2/3 de las muestras para el proceso de aprendizaje y el 1/3 restante para comprobar el error del clasificador. El hecho de que solo se utiliza una parte de las muestras disponibles para llevar a cabo el aprendizaje es el inconveniente principal de esta técnica, al considerar que se pierde información útil en el proceso de inducción del clasificador.Esta situación se deteriora si el número de muestras para construir el modelo es muy reducido. (Laudon, 2008, p. 232)

2. VALIDACIÓN CRUZADA *(CROSS-VALIDATION)*

También conocida como validación cruzada de n particiones. Se plantea para evitar la ocultación de parte de las muestras al algoritmo de inducción y la consiguiente pérdida de información. En ella se dividen los datos disponibles en tantas particiones como indique el parámetro n y se entrena n veces promediando el error de cada prueba. Según Laudon (2008) el esquema del proceso seguido para una validación 10fold. En general, este es el número de particiones más utilizado.

De acuerdo con Méndez (2011) una posible mejora en la utilización de la validación cruzada es la estratificación que consiste en mantener en cada una de las particiones una distribución de las etiquetas similar a la existente en el conjunto de aprendizaje, para evitar una alta varianza en la estimación. Además, es una práctica común repetir la validación cruzada con k particiones un número determinado de veces para hacer más estable la estimación de la precisión.

3. ENTRENAMIENTO Y VALIDACIÓN

Data Mining como un conjunto de técnicas estadísticas. No existe una única definición del término *Data Mining* (DM). Se puede decir que DM se refiere a un conjunto de métodos estadísticos que proporcionan información (correlaciones o patrones) cuando se dispone de muchos datos (de aquí viene el nombre Minería de Datos). Esta idea de DM lleva a la siguiente estructura de conocimiento:

EL FUNDAMENTO DE LA MINERÍA DE DATOS ES LA EXPLORACIÓN Y EL ANÁLISIS DE LOS PROCESOS DE INFORMACIÓN POR MEDIOS INFORMÁTICOS O SEMI-AUTOMÁTICOS QUE INTERPLETAN LOS PROCESOS QUE SON SIGNIFICATIVOS PARA SOLUCIONAR PROBLEMAS DE LA EMPRESA.

Según algunos autores, el *Data Mining* es aquella parte de la estadística que se usa para los problemas que se presentan actualmente en análisis de datos de la información recolectada. Los problemas actuales se diferencian de los clásicos en que el número de datos a analizar es mucho mayor y, como consecuencia, las técnicas estadísticas clásicas no pueden ser aplicadas.

Generalmente, la minería de datos es el proceso de analizar datos desde diferentes perspectivas con el objetivo de resumir los datos en segmentos de información útiles. Esta información que puede ser usada para incrementar réditos o beneficios, reducir costos, entre otros. El *Data Mining* permite a los usuarios analizar datos desde diferentes dimensiones o ángulos, categorizándolos y resumiendo las relaciones identificadas. (JIISIC, 2002, p. 132)

Con estas técnicas es posible, a veces, hacer evidente las relaciones ocultas entre sucesos. Un ejemplo simple sería averiguar la relación entre la compra de pañales y de cerveza el sábado por la tarde en los supermercados.

De acuerdo con Méndez (2011) este ejemplo ilustra muy bien la necesidad de conocer el campo de trabajo para aplicar el *Data Mining*: solo un especialista que conozca a su cliente es capaz de interpretar una correlación bruta que permita realizar el retrato típico de una pareja haciendo sus compras. Encontrar las relaciones causales que llevan a correlaciones como la anterior puede ser más rápido y sencillo con el *Data Mining*.

Además, la minería de datos permite trabajar con grandes cantidades de observaciones sin ningún inconveniente. También permite tratar una gran cantidad de variables predictivas. Esto último es de gran utilidad para seleccionar variables.

CAMPOS DE APLICACIÓN DE LA MINERÍA DE DATOS

La minería de datos tiene muchos campos de aplicación pues puede ser útil en prácticamente todas las facetas de la actividad humana. Se van a indicar algunas cuestiones relevantes sobre la posible aplicación de la minería de datos:

- La minería de datos tiene utilidad empresarial: las empresas pueden optimizar procesos y mejorar sus productos y ventas utilizando minería de datos.
- Existen pocos especialistas o empresas especializadas en minería de datos. Teniendo en cuenta su importancia, es un campo de trabajo para emprendedores.
- La minería de datos es una disciplina que se está desarrollando con mayores capacidades gracias al avance en tecnología y a la vez con la más alta capacidad de computación en los ordenadores. Asimismo, constituye un campo amplio de investigación en el que cada vez trabajan más investigadores y equipos de investigación.

METODOLOGÍA DE LA MINERÍA DE DATOS

Hernández, Ramírez y Ferri (2004) mencionan que un trabajo de minería de datos consta de las siguientes partes:

- **Entendimiento del problema:** se trata de hablar con el cliente, conocer sus necesidades, conocer su negocio o actividad, conocer qué datos relevantes tiene disponibles y cuáles serían necesarios pero no están disponibles, etc.
- **Entendimiento de los datos:** hay que saber qué significan los datos, si son continuos o discretos, qué tipo de valores toman, qué utilidad futura pueden tener y saber si están bien capturados o no.
- **Preparación de datos:** se trata de reflexionar sobre cómo guardar los datos. Típicamente hablaremos de tablas con filas y columnas, pero hay que ver cómo se organizan las tablas, cómo se interrelacionan entre ellas, etc. En definitiva organizar los datos para poder sacarles partido.
- **Modelamiento:** una vez se tienen los datos organizados hay que definir los algoritmos que se van a utilizar para tratar los datos. Una vez tratados, los datos nos devolverán información útil.
- **Evaluación:** los resultados obtenidos deben de ser sometidos a comprobación, verificar que están libres de errores, ratificar que son útiles para los objetivos perseguidos, etc.
- **Despliegue funcional-comercial:** una vez se tiene automatizada la captura y tratamiento de datos para obtener unos resultados, se desarrollan herramientas, normalmente en forma de aplicaciones informáticas que permiten generar alertas, informes, estadísticas, etc. que tienen una utilidad directa para la toma de decisiones y sistema de información del cliente. (p. 345)

Algunas cosas que se pueden hacer con el DM; el usuario del DM usualmente busca los siguientes cuatro tipos de relaciones:

- **Clases:** a las observaciones de los datos se asignan a grupos predeterminados. El proceso de clasificación consiste en asignar un conjunto de datos a grupos fijados de manera que se minimice la probabilidad de una clasificación errónea. Por ejemplo, un problema típico de clasificación es el de dividir una base de datos de bancos en grupos que sean lo más homogéneos posibles con respecto a variables como posibilidades de crédito en términos de valores tales como bueno o malo.
- **Clusters:** en este proceso se construyen grupos de observaciones similares según un criterio prefijado. El proceso de clustering (agrupamiento) consiste en subdividir un conjunto de datos en grupos mutuamente excluyentes de tal manera que cada miembro de un grupo esté lo más cercano posible a otro elemento, y grupos diferentes estén lo más lejos posible entre sí, de modo que la distancia está medida respecto a todas las variables disponibles. Un típico ejemplo de aplicación de clustering es la clasificación de segmentos de mercado. Por ejemplo, una empresa quiere introducirse en el mercado de bebidas alcohólicas, pero antes hace una encuesta de mercado para averiguar si existen grupos de clientes con costumbres particulares en el consumo de bebidas. La empresa quiere introducirse en el grupo (si existe) que esté menos servido por la competencia. En este ejemplo no existen grupos de clientes predeterminados. (Ramírez y Ferri, 2004, p. 348)
- **Asociaciones:** en las observaciones deformaciones recolectadas son usadas para identificar asociaciones entre variables. La búsqueda de asociaciones es diferente a la búsqueda de relaciones causales.

  Las relaciones causales son mucho más difíciles de encontrar que las asociaciones, debido a la presencia de variables no observadas. Las relaciones causales y asociaciones no son equivalentes: si hay asociaciones no tiene por qué haber causalidad.

- **Patrones secuenciales:** se trata de identificar patrones de comportamiento y tendencias. Un ejemplo sería intensidades de expresión en microarrays que permiten distinguir entre diferentes expresiones de genes para individuos con cancer o sin él.

  La presión competitiva es cada vez mayor, y los datos deben ser entendidos como un activo que le permitirá a las organizaciones proporcionar más y mejores servicios, predecir eventos futuros, anticiparse a ellos, entre otros.

REFERENCIAS BIBLIOGRÁFICAS

Alegret Rodríguez, M., Herrera, M., Grau Abalo, R. (2008). Las técnicas de estadística espacial en la investigación salubrista. Rev Cubana Sal Públ [revista en la Internet]. Disponible en: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-34662008000400003&lng=es

Barcellos, C y Buzai, G. (2006). La dimensión espacial de las desigualdades sociales en salud: aspectos de su evolución conceptual y metodológica. México: Universidad Nacional de Luján: Anuario de la División Geografía.

Benzecri, J. (2015). Gestión del conocimiento y minería de datos. PHC, París

Berry, M. y Linoff, G. (2014). Data Mining Techniques for Marketing Sales and Customer Support. USA: John Wile& Sons.

Chatfield, C. y Collins, A.J. (1999). Introduction to multivariate analysis. London: Chapman and Hall.

Kamber M. (2006). Data mining: concepts and techniques. Morgan Kaufmann;

Hernández C (2004). USA: Centricity Solution for Marketin Automation.

Hoaglin D.C., Mosteller F., Tukey J.W. (2005). N.Y: Exploring Data tables. Trends and Shapes, Wiley.

Ester, M., Kriegel, H., y Sander, J (1999). . Knowledge discovery in spatial databases. KI-99. Advanc Artif Intellig.

Jambu, M. (2000). Classification Automatique pour l'Analyse des données. París: Dunod.

Johnson Dallas, E. (2013). Métodos multivariados aplicados al análisis de datos. México: Thomson editores.

Johnson, R.A. y Wichern Dean, W. (2010). Applied Multivariare Statistical Analysis. 3rd De. USA: Prentice Hall Inc.

Lebart, L., Morineau, A. y Tabard, N. (2000). Téchniques de la description statistique. París: Dunod.

Lebart, L., Morineau, A. y Piron, M. (2005). Statistique exploratoire multidimensionelle. París: Dunod.

Macoe, J. (2016). 5 de los mejores software de minería de datos de Código Libre y Abierto. Recuperado de:

Http://blog.jmacoe.com/gestion_ti/base_de_datos/5-mejores-software-mineria-datos-codigo-libre-abierto/

Méndez, D . (2011). Estrategia para la Gestión Empresarial. McGraw Hill. México:

Pla, L. (2006). Análisis Multivariado: Método de Componentes Principales. Monografía Nº 27 Serie de matemática. USA: O.E.A.Washington.

Tecnologías-información.com. (2015). Minería de Datos. Recuperado de: http://www.tecnologias-informacion.com/mineria-de-datos.html

Vinnakota, S. y Lam, N. (2006). Socioeconomic inequality of cancer mortality in the United States: a spatial data mining approach. Internat J Heal Geogr.

Zhao, F., Zhu, R., Zhang, L., Zhang, Z, Li, Y. y He. M. (2011). Application of satscan in detection of schistosomiasis clusters in marshland and lake areas.