Data mining
Minería de datos
Autor: Alfredo Daza Vergaray

Coordinación de edición:
Magaly Ramon Quiroz

Diseño de portada:
Rudy Herrera Torres

Corrección de estilo:
Martín Vargas Canchanya

Diagramación:
Julissa Ventocilla Fernández

Edición a cargo de:
© Empresa Editora Macro EIRL
Av. Paseo de la República N.° 5613, Miraflores, Lima, Perú

Teléfono: (511) 748 0560
E-mail: proyectoeditorial@editorialmacro.com
Página web: www.editorialmacro.com

Primera edición: julio 2016
Tiraje: 1200 ejemplares

Impresión
Talleres gráficos de la Empresa Editora Macro EIRL
Jr. San Agustín N.° 612-624, Surquillo, Lima, Perú

ISBN N.° 978-612-304-417-6
Hecho el depósito legal en la Biblioteca Nacional del Perú N.° 2016-08276

Índice

1.1 Minería de datos

La minería de datos se ha definido de diferentes maneras. A continuación, se mencionarán algunos de estos conceptos para un mejor entendimiento, en especial, por aquellas personas que recién se estén iniciando en el maravilloso mundo de la extracción del conocimiento:

A. Primera definición

La minería de datos se define como aquel proceso que consiste en extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. En ese sentido, la tarea fundamental de la minería de datos es encontrar modelos inteligibles a partir de los datos recogidos (Hernández *et al.*, 2004).

B. Segunda definición

Según Hernández *et al.* (2004), la minería de datos implica un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias que son observadas al examinar grandes cantidades de información.

C. Tercera definición

Según Han y Kamber (2006) la minería de datos se refiere a la extracción de conocimiento o «minería» de grandes cantidades de datos. Sin embargo, de acuerdo con su perspectiva el nombre sería, en realidad, inapropiado, ya que, por ejemplo, la actividad minera que extrae oro de las rocas o de la arena se conoce como «minería de oro» en vez de «minería de roca» o «de extracción de arena». Por ende, partiendo de una lógica similar, la minería de datos debería haber recibido el nombre más apropiado de «minería de datos del conocimiento», el cual, por desgracia, es un poco largo. Ahora bien, sucede que «minería» es una palabra que porta la idea de un proceso por el cual se extrae un pequeño conjunto de elementos (pepitas) poseedores de una cierta cantidad de materia prima (metales preciosos). Así, a pesar de ser un nombre poco apropiado, al vincular las ideas de datos y extracción, «minería de datos» se ha convertido en una opción más popular. Frente a esto, lo único que cabe advertir es que existen muchas otras expresiones similares a esta, las cuales, empero, tienen un diferente matiz de sentido, tales como la minería de datos de conocimiento, la extracción de conocimientos, análisis de datos, análisis de patrones, arqueología de datos y filtración de información.

D. Cuarta definición

Según González (2005), la minería de datos es el proceso por el cual se genera un modelo útil para la predicción. Dicho modelo se construye teniendo como fundamento los datos que se encuentran en una base de datos, a los cuales se le ha aplicado algún algoritmo justamente con el fin de plantear un modelo.

En conclusión, se podría decir que la minería de datos es un proceso que integra los datos de diferentes fuentes (SQL Server, Oracle, Excel, etc.) para, posteriormente, extraer un importante conocimiento, es decir, identificar información trascendente, valiosa y útil, a partir de lo cual las instituciones van a poder tomar alguna significativa decisión.

1.2 Procesos de minería de datos (KDD)

Las etapas para la realización de la minería de datos siempre son las mismas independientemente de la técnica específica a usar. El conjunto de las partes de este proceso, también conocido por sus siglas en inglés «KDD» (*knowledge discovery in databases* o «descubrimiento de conocimiento en bases de datos»), es descrito de la siguiente manera por Fayyad *et al.* (1996).

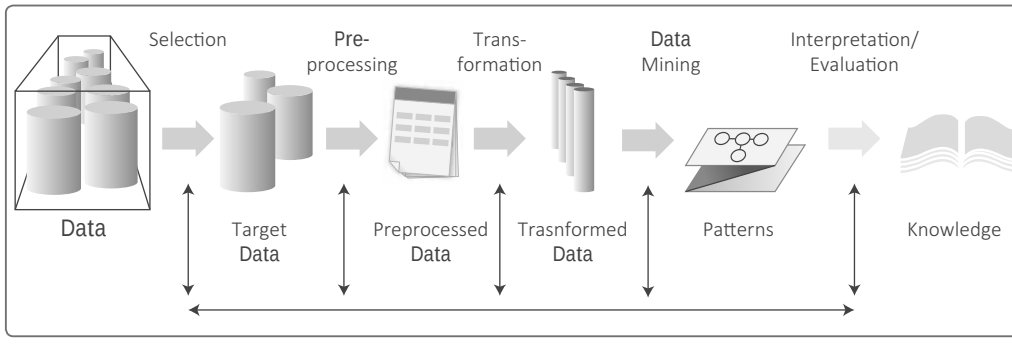

Figura 1.1 Descripción general de los pasos que constituyen el proceso KDD
Fuente: Fayyad *et al*. (1996).

El proceso de KDD es interactivo e iterativo (con muchas decisiones tomadas por el usuario) e implica numerosos pasos que se resumen así:

1. Aprendizaje del dominio de la aplicación: este paso incluye la adquisición del conocimiento previo relevante y el planteo de los objetivos de la aplicación.
2. Creación de un conjunto de datos de destino: por medio de esto se escoge el conjunto de datos o se elige el subconjunto de variables o muestras de datos en los cuales el descubrimiento se va a realizar.
3. Limpieza de datos y preprocesamiento: aquí se dan las operaciones básicas como la eliminación de ruido, la recogida de la información necesaria para modelar, la determinación de estrategias para el manejo de los campos de datos que faltan, la contabilidad de la información en tiempo y secuencia de los cambios conocidos, la decisión en torno al uso de DBMS (tales como tipos de datos y esquemas) y la asignación de valores desaparecidos y desconocidos.
4. Reducción de datos y de proyección: en este punto se lleva a cabo la búsqueda de las características útiles para representar los datos, dependiendo del objetivo y de la tarea. Para ello, se ejecuta la reducción de dimensionalidad o se aplican métodos de transformación que permitan acortar el número efectivo de las variables en estudio o encontrar las representaciones invariables para los datos.