Introducción a la Minería de Datos

La minería de datos es un conjunto de actividades orientado a descubrir conocimiento (patrones, relaciones, hechos, etc.) en bases de datos típicamente de gran volumen. La idea de la "minería" es una metáfora para resaltar el hecho de que se "encuentran" cosas "valiosas" mediante un proceso de prospección, como en las minas de minerales.

La analogía tiene implicaciones interesantes, entre otras:

- La minería comienza con algún tipo de objetivo o hipótesis sobre la existencia de minerales valiosos en una zona. En la minería de datos, también hay algunas ideas de negocio o hipótesis iniciales.
- La minería la hacen los mineros, pero se sirven de diferentes herramientas, cada vez más sofisticadas. En la minería de datos, hay muchas de estas herramientas, incluyendo diferentes modelos analíticos diseñados para trabajar con data de acuerdo al caso de uso en cuestión.
- La minería a veces tiene que excavar túneles alternativos. En la minería de datos, también hay un proceso de ensayo y error.
- En la minería es necesario extraer muestras de mineral y analizar su calidad. En la minería de datos, también es fundamental evaluar la calidad del conocimiento extraído.

La minería de datos incluye por tanto muchas actividades previas al trabajo de extracción de patrones, incluyendo la conformación de un Datawarehouse, procesamiento de datos, obtención de datos de fuentes externas, etc. Aquí nos centramos en los modelos analíticos como las herramientas del minero cuando está ya tratando con un conjunto de datos preparado y trata de extraer conocimiento valioso.

No obstante, es importante resaltar que en la minería de datos se necesitan dos tipos de competencias:

- Conocimiento del dominio. Hay muchos hechos y teorías que son conocidas de manera general o por los expertos que sirven como guía para el proceso de minería. Por ejemplo, el que las ventas en un sector dependan de algunas variables macroeconómicas.
- Conocimientos técnicos. Son los conocimientos en sí de las tareas técnicas de la minería de datos, que incluyen el uso de alguna herramienta tecnológica para aplicar la minera y el manejo de modelos analíticos.

Es importante también resaltar que la minería de datos en la mayoría de los casos es un proceso iterativo, que requiere una evaluación rigurosa antes de que se utilicen los modelos resultantes en la operación diaria. Por ejemplo, no se debe utilizar un modelo de predicción de fraude sin evaluar cuidadosamente su precisión, dado que si genera muchos falsos positivos, podría estar discriminando muchos clientes valiosos para el negocio.

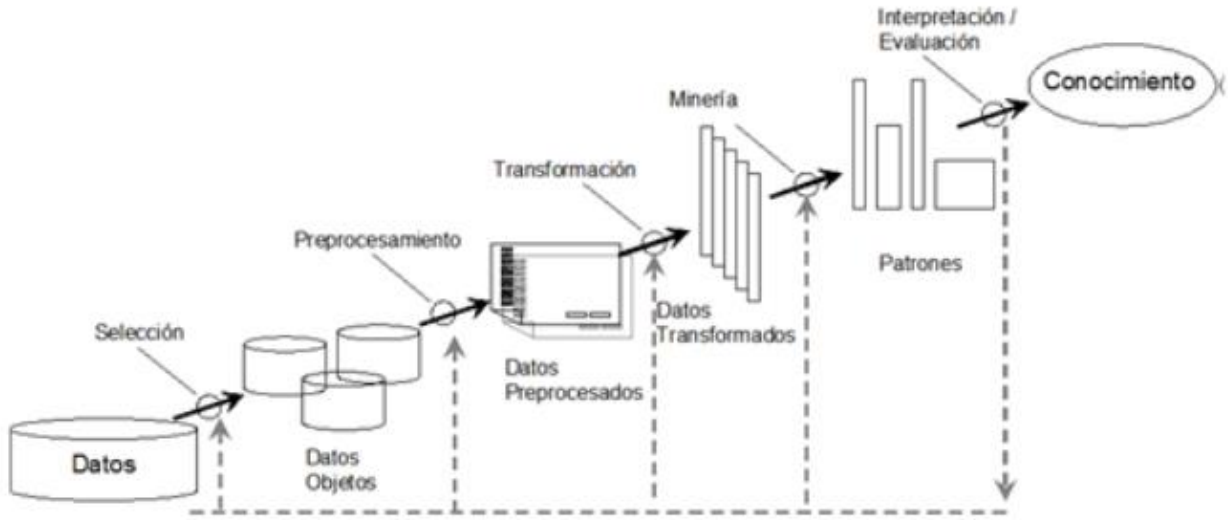
La disciplina del "***Descubrimiento de Conocimiento en Bases de Datos***" ("Knowledge Discovery in Databases" process, KDD) es el marco de actividades donde se encuadra la minería de datos y el aprendizaje automático (Machine Learning).

pavel.hernandez@intec.edu.do **http://www.linkedin.com/pub/pavel-hernandez/70/79a/677/en**

Un proceso de KDD genérico normalmente incluye las siguientes actividades:

1. Determinar las fuentes de información que pueden ser útiles y dónde conseguirlas.
2. Diseñar el esquema de un almacén de datos (data warehouse) que consiga unificar de manera operativa toda la información recogida.
3. Implantación del almacén de datos que permita la "navegación" y visualización previa de sus datos.
4. Selección, limpieza y transformación de los datos.
5. Seleccionar y aplicar el método de minería de datos apropiado, que servirá para obtener patrones de los datos.
6. Evaluación, interpretación, transformación y representación de los patrones extraídos.
7. Comunicación y uso del nuevo conocimiento

La siguiente Figura esquematiza las actividades de KDD como un ciclo.

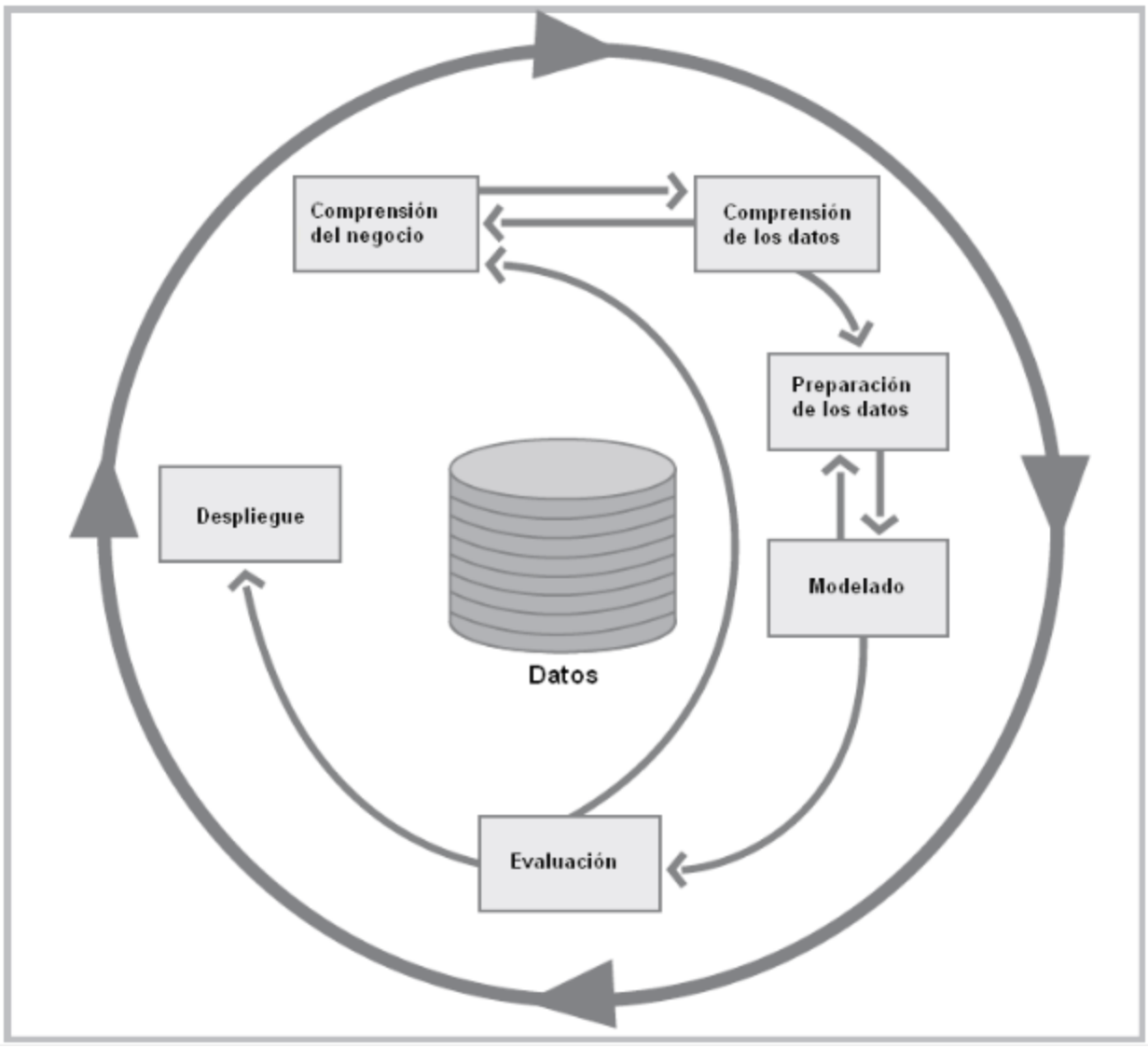


Actividades de la Minería de Datos

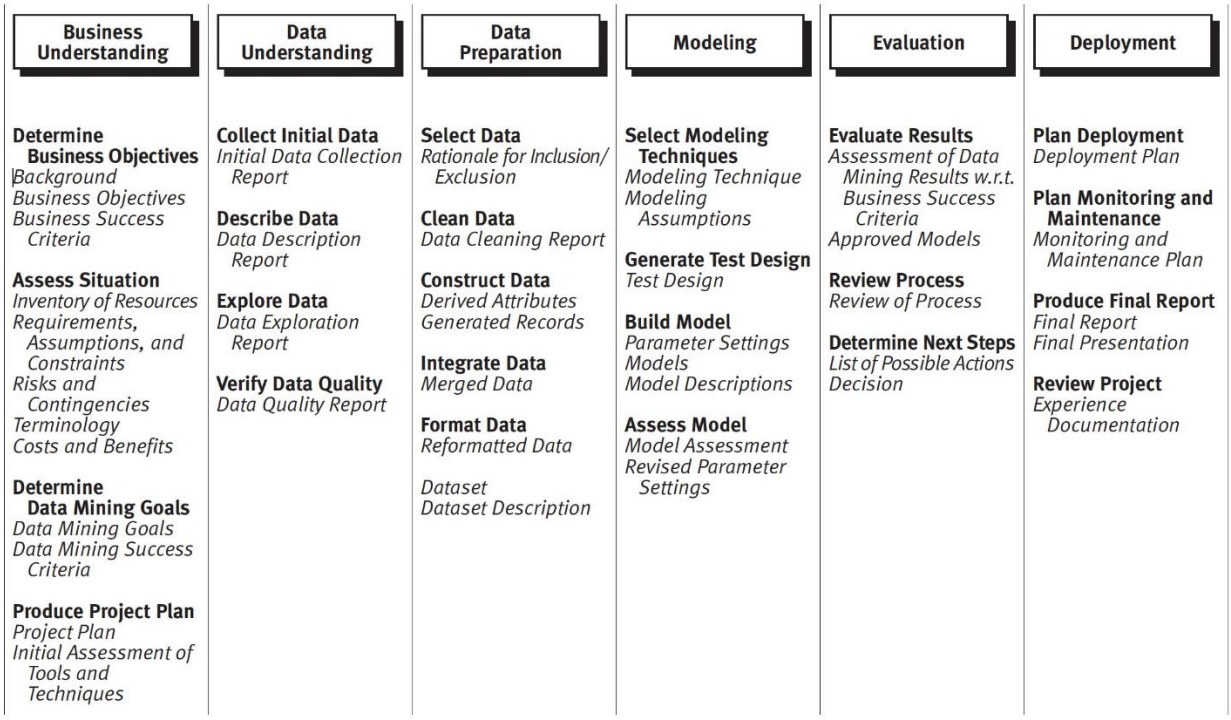
Si entendemos la minería de datos en sentido amplio, es decir, incluyendo el conjunto de actividades que normalmente se denominan KDD, tenemos un conjunto de actividades y tareas muy diversas, algunas relacionadas con comprender y hacer hipótesis sobre el negocio, y otras relacionadas con la selección, uso y evaluación técnica del modelo analítico utilizado. También incluyen tareas de limpieza, transformación y selección de datos.

Para tener una visión más general de las actividades de minería, lo mejor es mirar a los estándares. El modelo Cross Industry Standard Process for Data Mining (CRISP-DM) es probablemente el más completo y amplio, así como el más utilizado.

La siguiente Figura resume las fases de este modelo. Si hacemos un paralelo con las del proceso de KDD, será fácil encontrar analogías entre ellas.



Entrando en más detalle en las Fases, la siguiente Tabla nos muestra las actividades y los productos de cada actividad. Es importante resaltar que el ciclo comienza con una fase que es propia de las necesidades del negocio, y que determina cómo se evaluarán los resultados.



Puede apreciarse que la fase de Modelado es donde se utilizan las técnicas del modelo analítico seleccionado. No obstante, éstas tienen lugar en el centro de otras actividades. Por ejemplo, es muy importante entender que la evaluación de los modelos se realiza "con respecto a criterios de éxito de negocio", que fueron establecidos en la primera fase.

La minería de datos entendida de manera amplia como proceso de descubrimiento tiene similitudes con otros conceptos. Recientemente, con el énfasis en los datos que ha traído el Big Data, se ha popularizado el concepto de "científico de datos" (data scientist). En muchos aspectos, un data scientist dedicará parte de su tiempo a la minería de datos, aunque en otras ocasiones dedique su tiempo a tareas de análisis que no pueden clasificarse como minería de datos, sino que caen en la categoría de análisis estadístico más tradicional.

Es difícil trazar una línea entre la minería de datos y el trabajo del Data Scientist. No obstante, el elemento que diferencia cuándo se hace minería está en la búsqueda de nuevos patrones (modelos) en los datos. Un análisis estadístico descriptivo o un contraste de hipótesis tradicional, por ejemplo, no encajarían en esa definición.

Data Scientist es la práctica de la extracción de conocimiento generalizable a partir de los datos. Además de incorporar las técnicas y métodos del trabajo de la investigación científica, es intensiva en procesamiento estadístico, reconocimiento de patrones, visualización y modelización de la incertidumbre, entre otras técnicas.

pavel.hernandez@intec.edu.do http://www.linkedin.com/pub/pavel-hernandez/70/79a/677/en

El "científico de datos" normalmente trabaja sobre algún tipo de entorno computacional como el entorno R o SAS, por mencionar dos ejemplos. Estos entornos proporcionan lenguajes de programación adaptados o extendidos para el trabajo estadístico, y un amplio abanico de algoritmos y técnicas de visualización para trabajar de forma interactiva sobre los datos.

También en ocasiones el científico de datos trabaja sobre una infraestructura de procesamiento de "Big Data", o sobre un almacén de datos (datawarehouse), pero no en todos los casos.

En el siguiente video, Mike Gualtieri explica el rol profesional del data scientist y otros roles profesionales relacionados pero diferentes.

https://www.youtube.com/watch?feature=player_embedded&v=iQBat7e0MQs

pavel.hernandez@intec.edu.do **http://www.linkedin.com/pub/pavel-hernandez/70/79a/677/en**