Minería de Datos para Principiantes

Webimar
Software Guru
23 de octubre de 2013
Por Rafael Gregorio Gamboa Hirales
ITAM

Agenda

- La Minería de Datos. Objetivo.
- Modelos, fundamentos y técnicas.
- Herramientas Comerciales para MD.
  - Software libre
  - Software propietario
- Caso de ejemplo de un modelo de clasificación con aplicación comercial.
- Conclusiones

La Minería de Datos. Objetivo.

El objetivo de la MD es el desarrollo y aplicación de procesos de estimación de valores objetivo. Se tienen dos situaciones:

1. Obtener la estimación del "grado de pertenencia" de un elemento a una clase.
2. Obtener la estimación del valor de una variable que representa una cantidad directa y objetivamente medible.

Muy importante: La MD se basa en casos (datos) de experiencias pasadas en los que ya sabemos el valor de la "variable objetivo".

Modelos

Con estas dos situaciones en mente podemos desarrollar modelos para:

i. ***Clasificar*** clientes como los que están en el grupo que responde a una promoción con cierta "probabilidad" (o que la densidad de casos "exitosos" está arriba de cierta densidad).

ii. ***Pronosticar*** o ***estimar*** el valor esperado de una variable del "negocio" estando esta variable en un rango continuo de valores.

iii. Obtener ***grupos*** de clientes de acuerdo a sus características y/o comportamiento.

Modelos, fundamentos y técnicas.

El fundamento teórico ad-hoc es la Estadística.

Ello nos permite elaborar pruebas de hipótesis y validar nuestros modelos.

Sin embargo en ocasiones los negocios nos demandan desarrollo mas rápidos de los modelos.

Por ello, en MD procedemos “partiendo” nuestro conjunto de datos en al menos dos subconjuntos y los utilizamos para desarrollar y “verificar”, “probar” o “validar” nuestro modelo.

Pre - requisitos

Los pasos que seguiremos suponen que:

1. Conocemos a la perfección las variables que definen nuestro conjunto de datos.
2. Los datos son de “buena calidad”.
3. Se han eliminado variables “redundantes”.
4. Los dos conjuntos elegidos tienen características similares al del conjunto original y estas características se conservan en el conjunto al cual se ha de aplicar el modelo.

Proceso de elaboración del modelo

Paso 0: Partir los datos en los dos subconjuntos mencionados. Diremos que los subconjuntos son conjunto de entrenamiento y conjunto de prueba.

Paso 1: "Entrenar", - i.e. obtener los parámetros del modelo que hacen que la V.O. se "calcule" en términos de las variables de soporte,- uno o más modelos con el conjunto de entrenamiento.

Paso 2) “Validar” o probar el modelo aplicándolo al conjunto de (datos de) prueba.

Paso 3) Si el resultado es “aceptable” ya acabamos, en caso contrario debemos iterar eligiendo otros “Modeladores” y/o transformando las variables de soporte o modificando el enfoque mismo del problema.

¿Cómo saber si el modelo es “aceptable”?

La bondad (de ajuste) del modelo tiene que ver con el objetivo de “negocio” a obtener. Por ejemplo, maximizar utilidad, minimizar costo, o bien se define una función de utilidad ad-hoc (no necesariamente monetaria).

Otra situación se plantea en el desarrollo de un modelo para recomendar el tratamiento (de entre cinco posibles tratamientos) para un padecimiento. Es posible que para ciertos casos no sea muy relevante el “entrecruzamiento” de la decisión, i.e. si a un paciente en lugar de recomendarle el mejor tratamiento según sus características se le recomienda un sub-óptimo. Pero si el paciente es diabético si que puede ser muy importante. Por ello conviene penalizar estos casos para que el modelo se “equivoque” lo menos posible en ellos.

Campaña de promoción de la venta de un producto o servicio.

Se desea realizar una campaña promoviendo un producto/servicio. Consideraremos el caso más simple:

El costo de promoción es c unidades monetarias, $c > 0$.

El ingreso es f unidades monetarias. Supondremos $f > c$.

Esto implica que en caso de “hacer hit” se tendrá una utilidad $u = f – c$ unidades monetarias.

En caso de “no hacer hit” se pierden c unidades monetarias.

Digamos VO e {0,1}, 1 = “hit”, si el cliente compra..

Densidad de umbral

Ahora consideremos el concepto de “densidad de umbral”, que es la densidad de casos exitosos que se requiere para que la campaña salga “tablas”. i.e. si d* es la densidad de umbral:

$$d^* \times u - (1 - d^*) \times c = 0;\ d^* \in [0,1].$$

Despejando d*:

$$d^* = c/(u+c)$$

Campaña intrínsecamente ganadora. Modelo de Clasificación.

Bajo estas características y conceptos una campaña será intrínsecamente ganadora si la densidad original de casos exitosos en la muestra es mayor que d*. (Aún en esos casos la MD puede hacer que la utilidad sea mejor…)

Si la densidad original de casos exitosos es menor que d* utilizamos la MD para tratar de obtener las características de subconjuntos de casos que tienen densidades mayores a la de umbral y poder calificar nuevos casos con este modelo.

Modeladores
Técnicas más populares

- CART (Árboles de decisión)
- Regresión Logística
- Análisis Discriminante
- Redes Neuronales
- Vecinos Cercanos
- Bayes Naive
- Redes Bayesianas

Resultado del modelado

El resultado del modelado es:

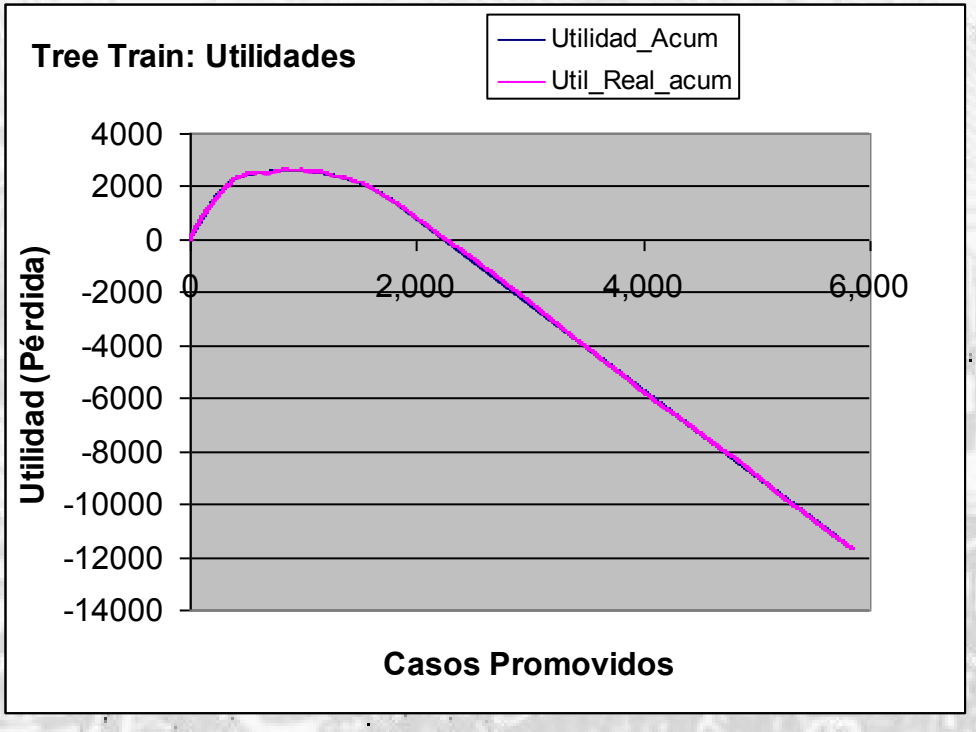
1. El modelo en sí, en ocasiones el código o parámetros que lo implementan para ser ejecutado por procesos autónomos o dependientes de la herramienta.
2. Las estadísticas del modelo y métricas de la "bondad de ajuste", como ROC, elevaciones o mejoras y matriz de confusión.
3. Los conjuntos de entrenamiento y prueba con sus "scores" o "Probabilidades de etiqueta", "P(VO=1)". A cada caso se le asigna su valor de P(VO=1) y el complemento es P(VO=0).

Post procesamos estos registros y obtenemos la "Curva de Utilidad" la cual indica la utilidad obtenida al ir procesando secuencialmente los casos habiendo ordenado los casos de mayor a menor según el score o "P(VO=1)".

Curva de utilidad

a) Ordenar descendentemente los casos del subconjunto en cuestión (entrenamiento o prueba) de acuerdo a “P(VO=1)”.
b) Para cada caso, si VO=1 ganamos la cantidad u, si VO=0 perdemos la cantidad c.
c) Vamos acumulando la utilidad y pérdida desde el “primer” caso hasta el caso del registro actual.
d) Graficamos el numero de caso en el eje horizontal y la utilidad o pérdida acumulada en el eje vertical. (Gráfica de dispersión para el caso general).

Curva de Utilidad
CART
post procesado propio



Curva de utilidad

Con esta gráfica podemos determinar hasta dónde debemos de llevar a cabo la promoción.

Debemos considerar que el corte debe hacerse en un lugar donde podamos diferenciar el valor del score. Esta consideración es muy importante en el caso de los árboles donde todos los casos que caen en el mismo nodo tienen el mismo valor de “score” o “P(VO=1)”.

Caso de ejemplo

Modelo de clasificación con aplicación comercial.

Consideremos el popular caso “Insurance”. Se desea realizar una campaña de venta de un seguro y se tiene una campaña realizada con anterioridad que se supone con las mismas características a la que se desea llevar a cabo actualmente. Los parámetros son:

Densidad de la muestra: 2%.

Costo individual de promoción: $4

Utilidad por caso exitoso: $96

La densidad de umbral es 4%

Si se ejecuta la campaña sin realizar la preselección de los clientes se obtiene un resultado de 2% $96 – 98% $4 = $1.92-$3.92 = -$2.

¡¡¡La casa pierde!!!!
¿Se puede hacer algo al respecto?

Demo con SAS Enterprise Miner

SAS define su metodología con las siglas

“SEMMA”: Sample, Exploration, Modificaction, Modeling, Assesment.

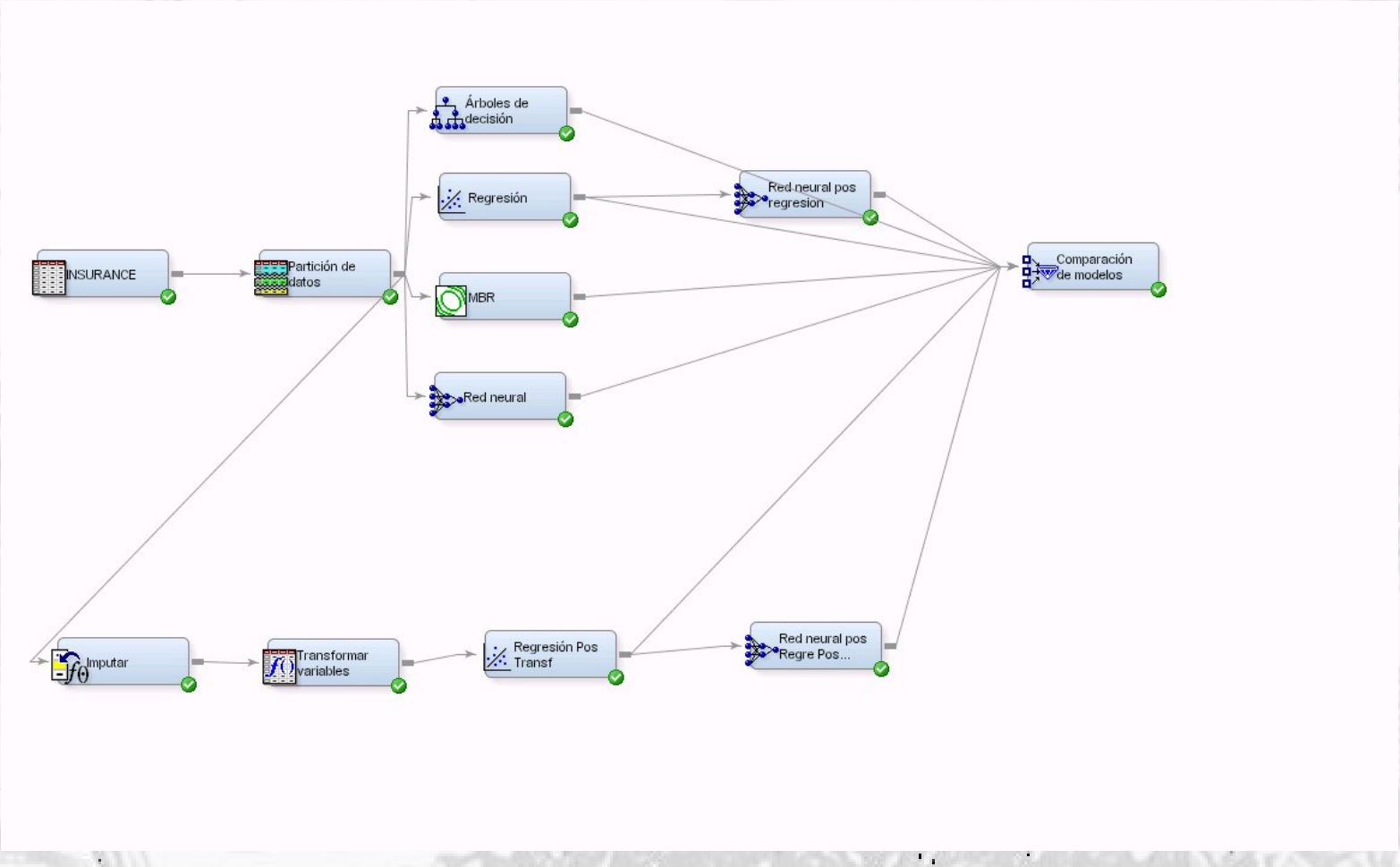
En castellano: Muestrear, Explorar, Modificar, Modelar y Evaluar. “MEMEE”.

Esta herramienta permite definirle entre otras, la matriz de costos - utilidades y decirle si la muestra está “sobrecargada” para que “compense” la salida del modelo, entregando la situación como será en la realidad.

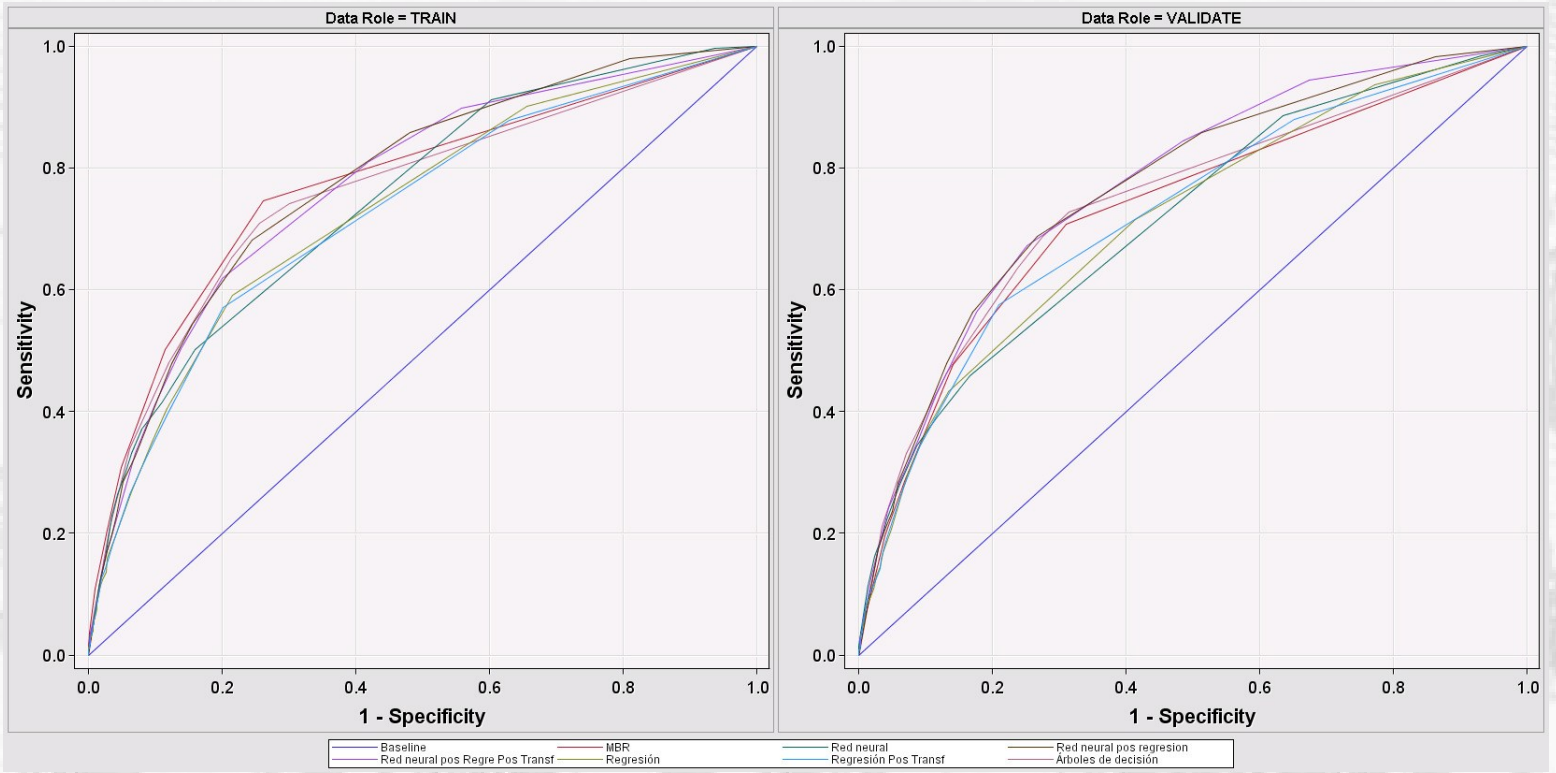
La herramienta se maneja gráficamente, de una manera muy ad-hoc para el trabajo con los modelos de MD.

Entrega los modelos en SAS, C, java y DB2.

Flujo de proceso en SAS EM ™



ROC


Data Role = TRAIN
Data Role = VALIDATE
Sensitivity
1 - Specificity
0.0
0.2
0.4
0.6
0.8
1.0
Baseline
MBR
Red neural
Red neural pos regresion
Red neural pos Regre Pos Transf
Regresión
Regresión Pos Transf
Árboles de decisión

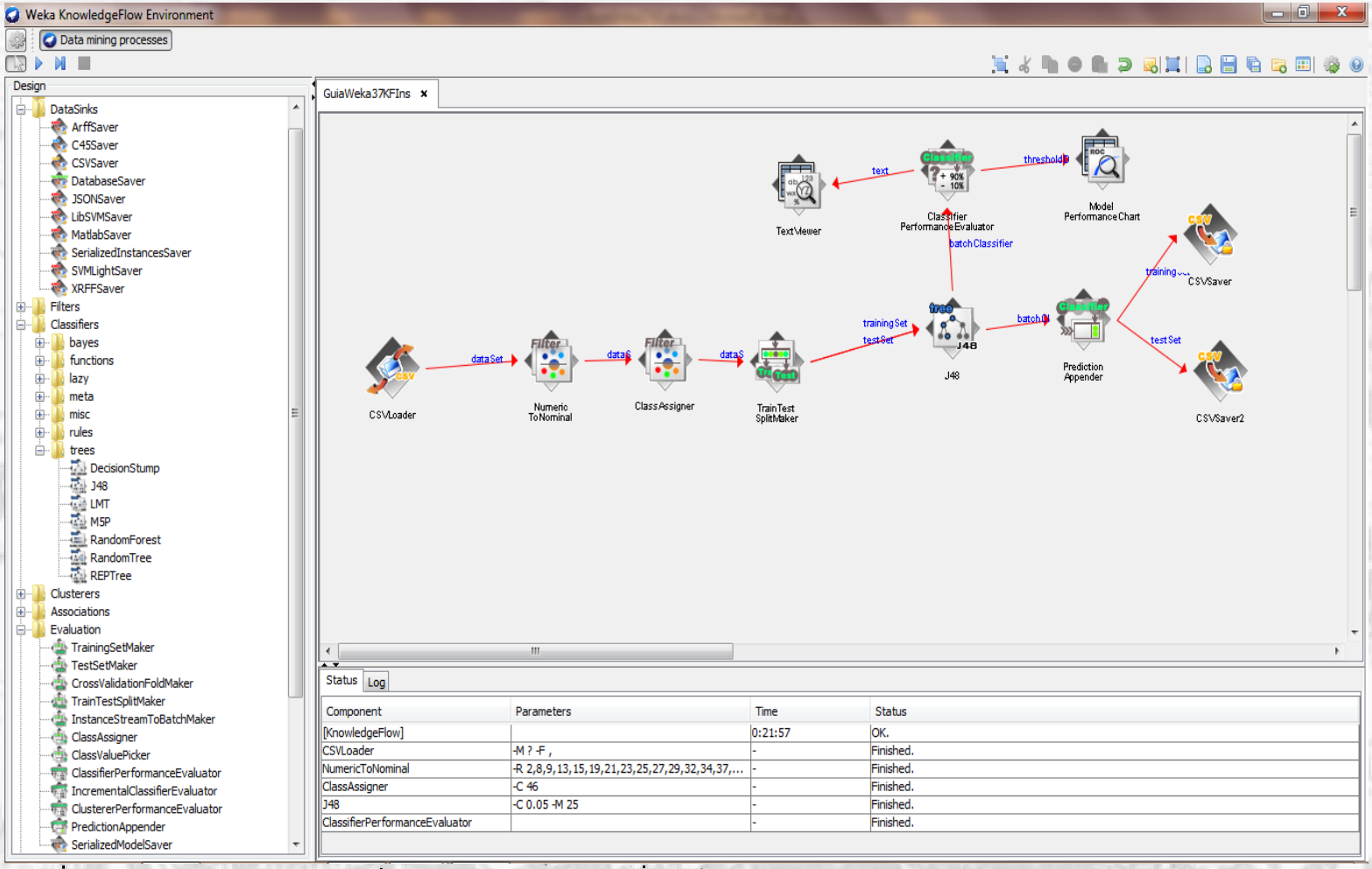
Demo con Weka

Weka es una de las herramientas libres para MD más populares. Está programada en java y ofrece procesos y procedimientos para poner en producción los modelos obtenidos.

Ofrece cuatro ambientes para trabajar con los modelos y los datos:

Explorer, Experimenter, Knowledge Flow y comandos “a pié” (texto).

Flujo en Weka



Herramientas Comerciales para MD.

- Software libre y/o gratuito:
  - Weka, R, Rapid Miner, Orange, etc…
- Software propietario:
  - SAS Enterprise Miner
  - Modeler de SPSS
  - Addendums a herramientas de BI.

Ahora se les llama “Analytics” e incorporan algunos elementos para llevar a cabo Minería de Textos y Herramientas para el análisis de características y relaciones sobre redes sociales.

Analytics

Se incorporan datos no estructurados en el sentido de un esquema de base de datos relacional.

Google es el pionero en la explotación de estos elementos, aunque los grandes jugadores ya se pusieron las pilas y ofrecen productos “llave en mano” para algunas de las necesidades donde hay más recursos económicos.

Conclusiones y Tendencias

- La MD es una técnica auxiliar en muchos campos de la investigación y los negocios.
- Requiere la conjunción de conocimientos de Computación, Estadística, Matemáticas y del área propia de aplicación.
- Actualmente debido a varios factores (reducción de precio del bit procesado, de los medios de almacenamiento y las telecomunicaciones) está en “ebullición” al tener datos explotables y manera de explotarlos.
- Los desarrollos actuales se encaminan a tener los resultados de manera temprana y oportuna, dándole ventajas a las organizaciones que se puedan apropiar de esta forma de trabajo dentro de su estrategia de negocio. Ej. High Performance Analytics de SAS.
- El recurso humano (capital intelectual) es de los más requeridos.
- Se destacan la formación en “Machine Learning” , “Data Science” en la parte de avanzada, trabajando con volúmenes grandes de datos, en ocasiones no estructurados, dispersos y con gran diversidad. A ello se le conoce como “Big Data”.

Muchas gracias

imagen de fondo tomada del Artículo sobre Minería de Wikipedia