MINERÍA DE DATOS

LA MINERÍA DE DATOS COMO HERRAMIENTA PARA LA TOMA DE DECISIONES ESTRATÉGICAS

Gustavo Adolfo Valencia Zapata

(info@gustavovalencia.com)

LA MINERÍA DE DATOS COMO HERRAMIENTA PARA LA TOMA DE DECISIONES ESTRATÉGICAS

Gustavo Adolfo Valencia Zapata

(info@gustavovalencia.com)

Resumen

El presente artículo pretende resaltar la importancia de la minería de datos como herramienta tecnológica empresaria y componente fundamental de la Inteligencia de Negocio (BI) para la toma de decisiones, destacando la gran importancia de la información como activo estratégico de las organizaciones y lo pertinente de un adecuado análisis para alcanzar la tan anhelada ventaja en los negocios. La minería de datos enmarca un completo conjunto de técnicas que buscan la extracción de conocimiento de diversas fuentes de información; ese conocimiento oculto entre millones de registros, es lo que análogamente se llamarían minerales preciosos que proporcionan a la compañía grandes ventajas competitivas y elementos contundentes para la toma de decisiones estratégicas.

Palabras Clave:

- Minería de Datos (Data Mining).
- CRISP-DM (Cross Industry Standard Process of data mininig).
- SEMMA (Sample, Explore, Modify, Model, Assess)
- Sistema de Información.
- Inteligencia de Negocios (Business Intelligence)

Introducción

En la actualidad las compañías manejan en sus sistemas de información altos volúmenes de datos, que en su mayoría no son aprovechados en profundidad para la toma de decisiones que ayudan al incremento de la rentabilidad. Sin darse cuenta, las empresas ignoran que bajos sus pies, se alojan grandes yacimientos de oportunidades con facilidades de explotación. Cada día se evidencia como la información es el motor vital de la empresa y su dependencia a los niveles de integridad, disponibilidad y confidencialidad de la misma, la convierten en un activo estratégico. ¿Cuál es la utilidad de la información cuando esta no es integral, cuando aloja registros errados o de poca validez?, ¿Qué sentido tiene almacenar información cuando se imposibilita el acceso a la misma? y ¿Qué validez o ventaja estratégica proporciona la información cuando carece de confidencialidad y mis competidores la conocen?

Las compañías deben procurar transformar los costos de la diferenciación en ventajas. Muchas actividades pueden hacerse más exclusivas añadiendo un pequeño costo extra. Una empresa puede ser capaz de diferenciase a si misma simplemente coordinándose mejor internamente (Porter, 1998). Un pensamiento estratégico en miras a la diferenciación y a la rentabilidad de la empresa, implica realizar una inversión inicial en tecnología y análisis de información; para el aprovechamiento de oportunidades y el mejoramiento continuo de los procesos.

La inteligencia de Negocios (BI) propone la integración de modelos matemáticos y metodologías de análisis que sistemáticamente explotan los datos disponibles en la infraestructura de TI (Tecnología de Información) para obtener información, conocimiento del negocio y oportunidades en la toma estratégica de decisiones.

Con una adecuada administración de la información, con herramientas como la minería de datos es posible encontrar en aquellos "mares o montañas" de datos, los "minerales" preciosos difícilmente perceptibles; identificando relaciones no aparentes o difíciles de detectar de forma tradicional. Surge entonces la pregunta: ¿Cómo utilizar la minería de datos al interior de la compañía?

Inicialmente es importante crear una conciencia, donde los sistemas de información han sido creados e implementados para brindar ventajas competitivas y servir como insumo estratégico; olvidando el estigma impuesto bajo excusas regulatorias, contabilidad, actividades secundarias, entre otros procesos considerados como apoyo y de carácter peyorativo. Asimilada la importancia de la información como activo estratégico, prosigue 'mapear' lo ofrecido por la minería de datos versus la razón de ser de la compañía, es decir, ¿cómo el proceso de descubrimiento de conocimiento con el enfoque del negocio y el despliegue predictivo se interceptan con los propósitos de la empresa? Encontradas estas innegables coincidencias, se procederá al desenvolvimiento de las actividades propias de la minería de datos.

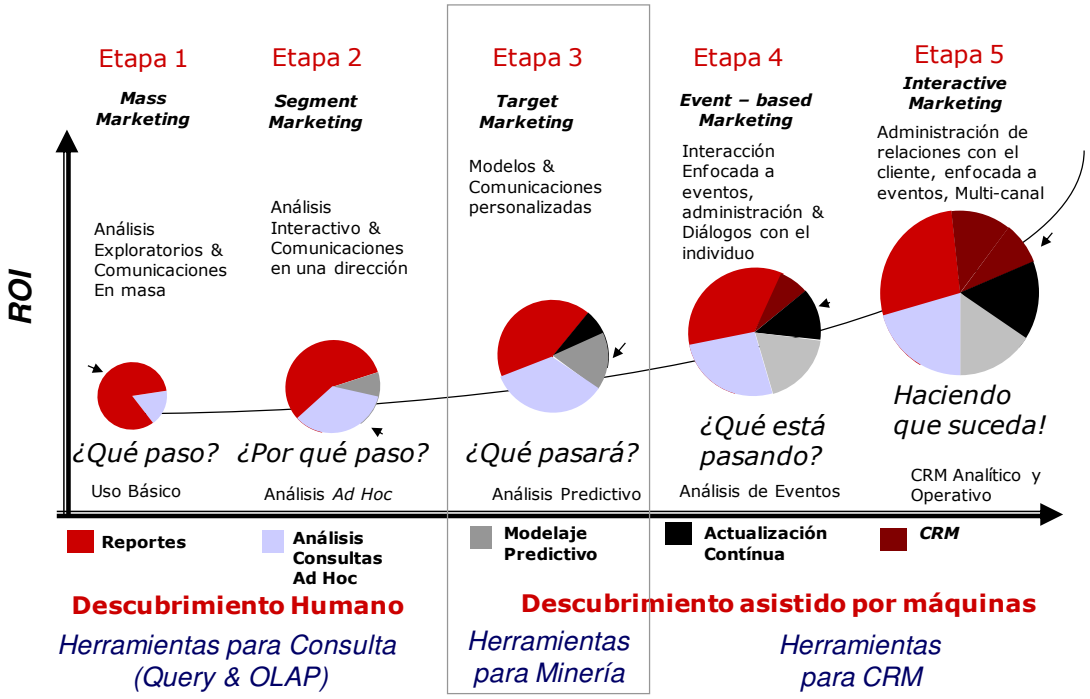
Concepto de minería de datos

Según Muñoz (2003) en su artículo sobre sistemas de información de las empresas, durante las últimas décadas se ha realizado múltiples esfuerzos y estudios fundamentados en analizar la información como factor clave para la toma de decisiones en la empresa.

Se podría visualizar una línea de tiempo donde se evidencia lo citado en el párrafo anterior, inicialmente la finalidad de los sistemas de información estaba orientada al soporte a los procesos básicos de las compañías (regulaciones, ventas, personal, etc.); surgiendo posteriormente los sistema de información para la toma de decisiones (EIS, OLAP, Consultas, reportes, informes, entre otros) brindando a los líderes, una visión inmediata (diariamente) sobre el estado y las actividades de gestión por medio de indicadores.

La minería de datos puede definirse inicialmente, como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos (Pérez, 2007).

Figura 1. Evolución de los análisis de datos



Nota. SPSS, Introducción a Clementine (2008). Extraído de Junio, 23, 2008.

Dada la anterior definición, es posible establecer con mayor claridad el objetivo de la minería de datos, el cual dirige sus esfuerzos a descubrir patrones, perfiles, anomalías y tendencias por medio del análisis de datos. No bastaría entonces con el resultado arrojado por un aplicativo de minería de datos para garantizar el éxito y la rentabilidad, sería necesario entonces, la toma de decisión basada en los resultados, la medición de los efectos generados (incremento de ventas, ahorros, efectividad en procesos, etc.) y la realimentación y calibración de los modelos construidos bajo las técnicas estadísticas empleadas en el modelo de minería de datos.

La minería de datos o descubrimiento de conocimiento en bases de datos (KDD, "knowledge discovery in databases"), es una poderosa herramienta informática de gran alcance con un incalculable potencial para la extracción de información previamente

desconocida y potencialmente útil a partir de grandes bases de datos. La minería de datos automatiza el proceso de búsqueda de relaciones y patrones en los datos y proporciona resultados que pueden ser utilizados en un sistema de apoyo a las decisiones estratégicas del negocio.

Fernández (2010) en su compendio "Statistical Data Mining Using SAS Aplication" menciona que un sinnúmero de empresas exitosas emplean técnicas de minería de datos para la toma de decisiones. La minería de datos permite la extracción de "minerales preciosos" de conocimiento a partir de los datos de negocio, que pueden ayudar a mejorar la gestión de relaciones con clientes (CRM) y ayudar a estimar el retorno de la inversión (ROI).

Según SPSS Inc (2008) compañía reconocida a nivel mundial en inteligencia de negocios, las situaciones de negocio a resolver bajo el análisis de los datos, son las siguientes:

- Predecir: valores de categorías o valores numéricos. Un ejemplo sería el riesgo crediticio, es decir, se puede predecir qué tan riesgoso es un cliente o negocio financieramente para la compañía.

- Segmentar: agrupar elementos con base a sus características. Algunos ejemplos serían identificar diferentes tipos de televidentes de un canal de televisión, diferentes clientes de una entidad financiera, etc.

- Asociar: eventos que ocurren juntos o en secuencia. Un ejemplo sería cuando se identifica la relación en ventas de dos productos en un almacén de cadena, como por ejemplo pañales y cerveza, es decir, se detecta que ciertos clientes cuando compran pañales tiene una tendencia a comprar cerveza.

- Anomalías: identificar eventos que no tienen un comportamiento esperado. Un ejemplo sería un fraude realizado a un cliente en el sector financiero.

Decisiones gerenciales basadas en los resultados de minería de datos

Para la toma de decisión basada en resultados de minería de datos, es indispensable contemplar algunos aspectos claves del proyecto de minería de datos al interior de la compañía. Según SPSS Inc (2008), Es conveniente tener claros los objetivos que persigue el negocio y también tener claro la naturaleza de los datos.

Algunos aspectos claves son:

- Disponibilidad de los datos. Responder preguntas como ¿Dónde están los datos?, ¿Cuál es su naturaleza?, ¿Existen motores de bases de datos?

- ¿Los datos cubren los factores relevantes para el análisis?

- ¿Los datos poseen mucho ruido? Una etapa de reparación de los datos aportará mejores resultados para la toma de decisiones.

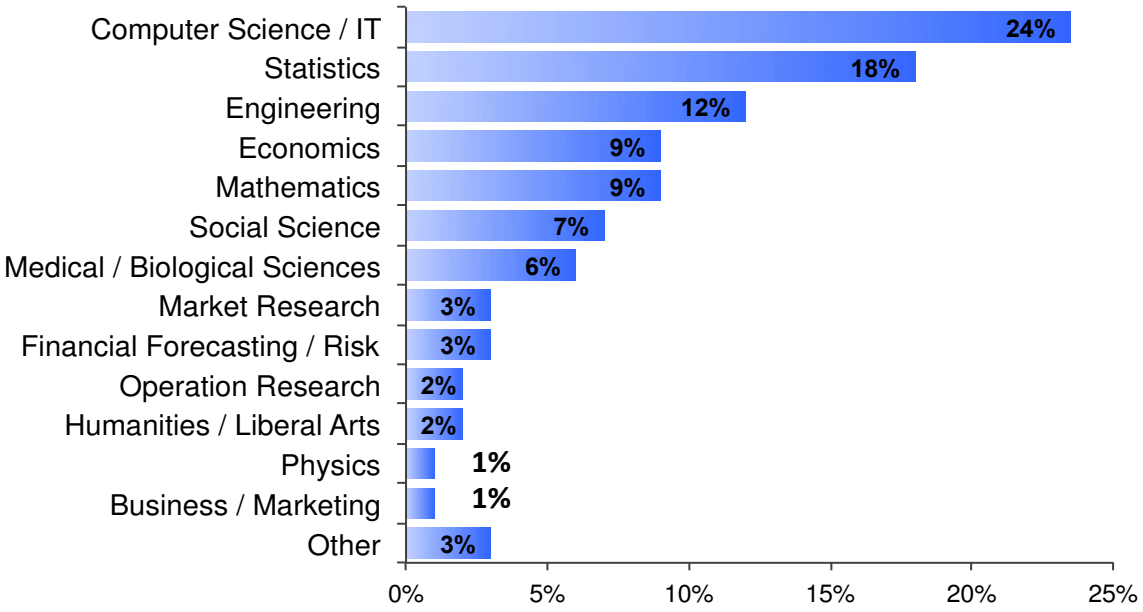
- ¿Hay suficientes datos? Un adecuado conocimiento del negocio da solución a este cuestionamiento, resaltando elementos como el sector empresarial, el tamaño de la compañía o el tipo de bien y/o servicio que se ofrece.

- Conocimiento de los datos disponibles. No necesariamente el experto en minería de datos es el experto en el tema a ser analizado por medio de este proceso, es decir, es importante que un proyecto de minería de datos garantice la presencia del conocimiento y experticia en el negocio a ser intervenido; lo anterior generalmente es manejado por un equipo de trabajo interdisciplinario.

Sobre este último aspecto, se podría aludir que en el sector de los negocios es cada vez más común encontrar profesionales adquiriendo competencias en análisis de información, estadística y técnicas de minería de datos; orientando este tipo de conocimiento a un estado transversal y de apoyo a cualquier disciplina. Es así, como ingenieros, médicos, economista, entre otros, se forman en este tipo de saberes. Rexer Analitics (2011) en su encuesta realizada a mineros de datos a nivel mundial se

identificó que el 24% de los mineros son profesionales en informática, seguidos por profesionales en estadística con un 18% y otras ingenierías con un 12%. La Figura 2 esboza la distribución de profesiones en dicha encuesta.

Figura 2. Profesión de los mineros de datos



Nota. Rexer Analytics., 4th Annual Data Miner Survey, 2011. www.rexeranalytics.com. Extraído de Agosto, 2, 2012.

Por otra parte, el escenario de la investigación (la academia) aunque no alude esta práctica, se determina por la conformación de equipos interdisciplinarios, cuyos miembros tienen competencias específicas altamente desarrolladas y bajo un objetivo común.

Una metodología como CRISP-DM (Cross Industry Standard Process of data mininig) sugerida por SPSS, no solo garantizaría una adecuada planeación sino una mayor efectividad en los resultados de un proyecto de minería de datos. Se debe tener presente que a pesar de disponer de la tecnología o las herramientas más sofisticadas; un factor determinante en el éxito recaerá sobre un individuo o grupo con un conocimiento profundo del negocio. Bajo la metodología CRISP-DM, se deben responder las siguientes preguntas (Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer & Wirth, 2007):

- ¿Cuál es el principal objetivo que se persigue resolver?
- ¿Qué datos se tendrán disponibles y cuales son relevantes en cuestión?
- ¿Qué clase de depuración de datos es requerida?
- ¿Qué técnica de minería de datos se empleará?
- ¿Cómo se evaluarán los resultados?

Un riguroso proyecto en minería de datos, contempla las anteriores preguntas para aumentar las posibilidades de éxito en los resultados obtenidos. El objetivo del presente artículo no es abordar en detalle cada uno de estos cuestionamientos, sino dar un acercamiento a los elementos más importante y de fácil comprensión por parte del líder y estratega del área, equipo o compañía.

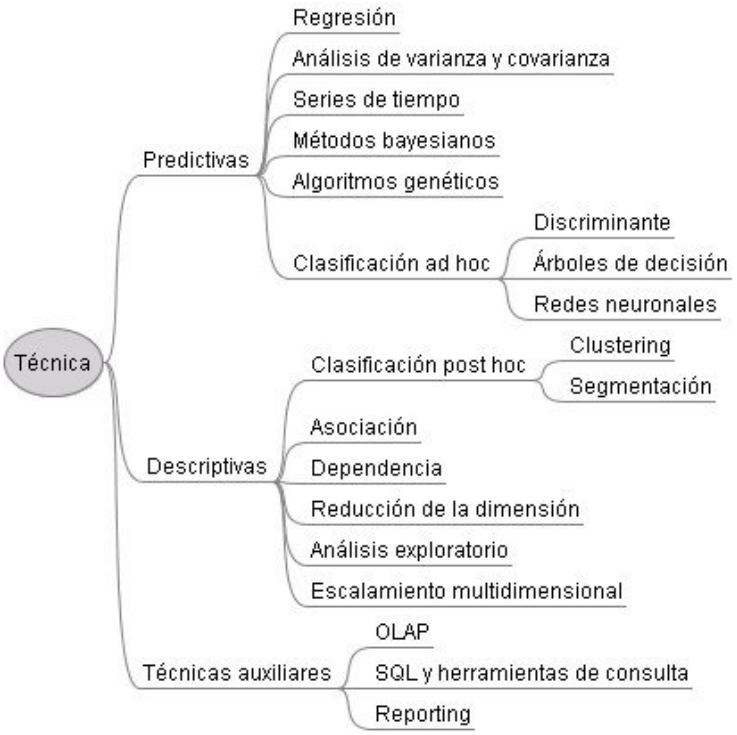
SAS es considerada actualmente por los expertos como la compañía líder en inteligencia de negocios a nivel mundial, sin embargo esta corporación considera su solución de minería de datos un proceso más que un conjunto de herramientas analíticas. La metodología SEMMA se refiere a una metodología que detalla este proceso; donde a partir de una muestra estadísticamente representativa de los datos, SEMMA hace fácil aplicar técnicas estadísticas de exploración y visualización, seleccionar y transformar las variables predictoras más importantes, el modelo de las variables para predecir los resultados y confirmar la exactitud de un modelo.

Según Pérez (2007), la clasificación inicial de las técnicas de minería de datos se distinguen entre técnicas predictivas (las variables pueden clasificarse inicialmente como dependientes o independientes), técnicas descriptivas (donde todas las variables al principio tienen el mismo estatus) y técnicas auxiliares.

Las técnicas predictivas especifican el modelo para los datos empleando un conocimiento teórico previo, las técnicas descriptivas no se asigna ningún papel predeterminado a las variables. Tanto las técnicas predictivas como las técnicas

descriptivas están enfocadas al descubrimiento del conocimiento embebido en los datos. Las técnicas auxiliares son herramientas de apoyo, más superficiales y limitadas. Se basan en técnicas estadísticas descriptivas, consultas e informes en general hacia la verificación.

Figura 3. Clasificación de las técnicas de minería de datos



Nota. Pérez López, César. (2007). Minería de datos: Técnicas y herramientas. Extraído de Julio, 4, 2008.

Contemplando lo sugerido por la metodología y seleccionadas las variadas técnicas (Regresión, árboles, Quest, red neuronal, C5.0, red bayesiana, K-medias, etc.) es cuando el resultado por sí mismo ya contienen un valor estratégico para la compañía. La definición del objetivo del proyecto planteó el camino a seguir, empleando las diferentes técnicas de extracción, reparación y modelamiento de datos, y el resultado será tan contundente como las relaciones o correlaciones ocultas en las bases de datos; las cuales darán diferentes perspectivas sobre las estrategias planteadas con anterioridad por la compañía. Otro elemento diferenciador en los resultados, aborda el tema predictivo, donde el nivel de confianza entregado por el modelo, plantea las acciones que generarán un posible valor sobre los procesos de la compañía. Es el apetito de riesgo y la confiabilidad del modelo predictivo los que llevan a la compañía a

la toma de decisiones; adicional a lo anterior, se incluyen en las decisiones estratégica los posibles controles que mitiguen el riesgo de resultado impropios al momento de llevar el modelo de minería de datos a producción.

Un ejercicio saludable es el entrenamiento del modelo con diferentes periodos de tiempo según la realidad de cada compañía, lo anterior, para evidenciar que lo predicho por la herramienta sea coherente con la realidad de la empresa.

Según Pérez (2007) y Rexer Analitics (2012) en su encuesta anual realizada a mineros de datos a nivel mundial, las siguientes son algunas de las herramientas de minería de datos más reconocidas y utilizadas en el mercado:

- IBM SPSS Modeler
- STATISTICA
- R
- SAS
- Knowledge Seeker (Angoss)
- Matlab
- CART (Salford Systems)
- Enterprise Miner (SAS)
- Knime
- Data Surveyor
- Gain Smart
- Weka
- Intelligent Miner (IBM)
- Microstrategy
- Polyanalyst
- Darwin
- SGI MineSet
- Wizsoft/Wizwhy
- Rapid Miner
- IBM SPSS Statistics
- KXEN

¿Qué no es minería de datos?

Según SPSS Inc (2008), en su curso de introducción a la minería de datos, aborda conceptos erróneos atribuidos a la minería de datos. La minería de datos no es:

- Una búsqueda de patrones a ciegas
- El sustituto de personal.
- Un Data Warehousing
- SQL, Ad Hoc Queries, reportes
- OLAP
- Visualización de datos

Es común encontrar constantes confusiones entre minería de datos y análisis estadísticos, donde en realidad son ejercicios diferentes pero con cierta relación.

La estadística en el mundo de los negocios es considerada de gran importancia, ya que suministra los mejores instrumentos de investigación, no sólo para observar y recopilar toda gama informativa incubada dentro de la misma empresa o fuera de ella, sino también en el control de ciertas actividades de producción, ventas, proyecciones o estimaciones a corto plazo, mediano y largo plazo, en la formulación de hipótesis y en el análisis de procesos encaminados a facilitar la toma de decisiones por parte de los encargados de la buena marcha de la empresa (Martínez, 2008).

En pocas palabras el análisis estadístico es el proceso de validar o rechazar hipótesis y la minería de datos es el proceso de generar análisis a partir del comportamiento de los datos, es decir, la minería de datos hace uso de la estadística de forma implícita aplicando técnicas matemáticas y modelamiento que ayudará al entendimiento de los datos. Es verídico que los mineros más versados y reconocidos en el escenario de la investigación son profesionales formados en las ciencias estadísticas.

Algunas otras diferencias entre minería de datos y análisis estadístico son (SPSS Inc 2008):

Minería de datos	Análisis estadístico
• Poco interés en la formulación detallada los algoritmos. • Requiere entendimiento de los datos y el problema de negocio. • Encuentra patrones en grandes cantidades de datos. • No presta mucha atención a los supuestos estadísticos (esto puede ser altamente peligroso)	• Prueba de Hipótesis. • Requiere de fuertes habilidades estadísticas. • Trabaja bajo muestras; las técnicas no se encuentran optimizadas para trabajar con grandes cantidades de datos.

Acercamiento a resultados de minería de datos

A continuación se esbozan ejemplos de objetivos y de decisiones estratégicas basadas en resultados de minería de datos:

- Objetivo: Desarrollo de modelos predictivos de fraude electrónico (tarjetas de crédito o virtual). Decisión estratégica: Realizar un bloqueo de productos o servicios a los clientes bancarios afectados
- Objetivo: Predicción en el aumento de eficiencia y mayor productividad al modificar el orden en procesos de manufactura. Decisión estratégica: Cambiar el orden del proceso y comparar la predicción con la realidad.
- Objetivo: Pronosticar la siguiente página que visitará un usuario de Internet. Decisión estratégica: invertir publicidad de productos y servicios en las páginas que serán visitadas posteriormente.
- Objetivo: Identificar el riesgo de crédito para clientes bancarios. Decisión estratégica: Evitar la aprobación del crédito a clientes con un alto nivel de riesgo.

Al tratar de referenciar casos de éxito basados en resultados o análisis vía minería de datos, se evidencia lo tangible y provecho que se convierte una inversión tecnológica en análisis de datos y el positivo retorno de inversión que entrega un proyecto de naturaleza. Para ser más descriptivos, se mencionan algunos casos contundentes de aplicación de minería de datos en la industria.

Según Da Cunha, C., Agard, B. y Kusiak, A (2006), por medio de un proyecto de minería de datos se pretende dar mejoras a la calidad de producción industrial. Inicialmente el proyecto tenía como objetivo determinar la secuencia de ensambles que minimice el riesgo de producir productos defectuosos; arrojando como resultados la identificación cinco reglas; las cuales determinaron los siguientes resultados:

- Se determina dos procesos que estaban invertidos, es decir, el que estaba primero debería ser el segundo.
- Hay una operación que sola necesita ser reelaborada.
- El modelo propuesto organiza la secuencia de los módulos de ensamble que precede o sucede inmediatamente a otra.

Al incorporar un proyecto de minería de datos con los diferentes agentes tecnológicos en un Shop Floor Control, se contribuye al dinamismo del proceso, al tratarse de un modelo de aprendizaje continuo. La modernización tecnológica en los procesos de manufactura, unidos a la facilidad de procesamiento y análisis de información que estos entregan continuamente, se convierte en recursos valiosos para la comprensión del comportamiento de los sistemas y el continuo mejoramiento del proceso (Harding, J. A., 2008).

En materia de seguridad empresarial y más específicamente en seguridad de la información, un proyecto de minería de datos como Clustering-Based Network Intrusion Detection, ayudó a la detectar comportamientos asociados a ataques en la red con relativa facilidad, superando el modelo tradicional supervisado; lo anterior debido al dinamismo de estos ataques, los modelos supervisados y parametrizables no aportarían el nivel de detección deseado. Al implementar modelos empleando técnicas de clustering no supervisado se logrará detectar nuevos tipos de intrusiones sobre la red, elemento constante y de naturaleza variable en los incidentes de seguridad reales a los que están expuestas las compañías de hoy (Zhong, S., Khoshgoftaar, Taghi M. y Seliya, N., 2008)

Adicional a los anteriores casos de éxito, se podrían mencionar miles de aplicaciones con diferentes focos de negocio en donde la minería de datos pueda entregar resultados lo suficientemente veraces para la toma de decisiones estratégicas.

CONCLUSIONES

- Actualmente se le ha reconocido a la **información** como uno de los activos más importantes para las organizaciones, dándole un tratamiento cuidadoso para temas relacionados a la seguridad de la información y estratégico para la toma de decisiones que apunten a los objetivos del negocio.

- La minería de datos es considerada en la actualidad como una herramienta tecnológica estratégica que pretender ver más allá de lo evidente al interior y exterior de las empresas, proporcionando resultados y argumentos a los líderes empresariales para la toma de decisiones.

- La estrategia tecnológica es uno de los caminos obligados en la búsqueda de la ventaja competitiva, siendo la innovación una de las formas más productivas para lograrlo; al tratar la información como activo fundamental de la compañía, proyectos tecnológicos como la minería de datos contribuirán al encuentro con la diferenciación y la rentabilidad.

BIBLIOGRAFÍA

Pérez López, César. (2007). Minería de datos: Técnicas y herramientas. España: Thomson Editores.

SPSS, (2008). Introducción a Clementine.

Chapman, Pete., Clinton, Julian., Kerber, Randy., Khabaza, Thomas., Reinartz, Thomas., Shearer, Colin. y Wirth, Rüdiger. (2007) CRISP-DM 1.0. USA: SPSS

Martínez Bencardino, Ciro. (2008). Estadística básica aplicada. Colombia: Ecoe editores

Porter E, Michael. (1998). Ventaja competitiva: Compañía editorial continental

Harding, J. A. (2008). A data mining integrated architecture for shop floor control. Proceedings of the Institute of Mechanical Engineers – Part B – Engineering Manufacture, 222 (5), 605-624. Recuperado el 18 de Mayo de 2008 desde las bases de datos EBSCO (Masterfile) en internet: http://web.ebscohost.com/ehost/detail?vid=6&hid=108&sid=b379cfad-bc09-47c1-bf38-6c97967de502%40sessionmgr104&bdata=JmFtcDtsYW5nPWVzJnNpdGU9ZWhvc3QtbGl2ZQ%3d%3d#db=buh&AN=32854795

Da Cunha, C., Agard, B. y Kusiak, A. (2006). Data mining for improvement of product quality. International Journal of Production Research, 44 (18- 9), 4027-4041. Recuperado el 10 de Octubre de 2008 desde las bases de datos EBSCO (Masterfile) en internet: En línea. http://web.ebscohost.com/ehost/detail?vid=1&hid=114&sid=2062378c-0a66-4b69-a66a-676144f468c8%40sessionmgr103&bdata=JmFtcDtsYW5nPWVzJnNpdGU9ZWhvc3QtbGl2ZQ%3d%3d#db=buh&AN=21782359

Muñoz, Antonio. (2007). Sistemas de información en las empresas. Extraído de noviembre, 08, 2008. del World Wide Web: http://www.hipertext.net/web/pag251.htm#2577.

Zhong, S., Khoshgoftaar, Taghi M. y Seliya, N. (2008). Clustering-Based Network Intrusion Detection. International Journal of Reliability, Quality & Safety Engineering, 14 (2), 169-187. Recuperado el 18 de Mayo de 2008 desde las bases de datos EBSCO (Masterfile) en internet: En línea. En línea. En línea. http://web.ebscohost.com/ehost/detail?vid=6&hid=108&sid=b379cfad-bc09-47c1-bf38-6c97967de502%40sessionmgr104&bdata=JmFtcDtsYW5nPWVzJnNpdGU9ZWhvc3QtbGl2ZQ%3d%3d#db=buh&AN=32854795

Parr Rud, Olivia. (2001). Data Mining Cookbook. Modeling Data for Marketing, Risk, and Customer Relationship Management. John Wilwy & Son.
Perner, Petra (2002). Data Mining on Multimedia Data. Springer.

Larose T, Daniel (2004). Discovering Knowledge in Data. An introduction to data mining. John Wilwy & Son.

Carlo Vecellis (2009). Business Intelligence: Data Mining and Optimization for Decision Making. John Wiley & Sons.

Gert H.N (2010). Laursen & J. Thorlund. Business Analytics for Managers: Taking Business Intelligence Beyond Reporting. John Wiley & Sons.

Rexer Analytics (2011). 4th Annual Data Miner Survey. www.rexeranalytics.com

Rexer Analytics (2012). 5th Annual Data Miner Survey. www.rexeranalytics.com