FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

EXAMEN DE SUFICIENCIA PROFESIONAL

“MINERÍA DE DATOS PARA LA INTELIGENCIA DE NEGOCIOS”

PARA OPTAR EL TÍTULO PROFESIONAL DE

INGENIERO DE SISTEMAS E INFORMÁTICA

PRESENTADO POR:

MILAGROS CELESTINA ZURIMA REÁTEGUI CURTO

IQUITOS, PERU

2014

UNIVERSIDAD NACIONAL DE LA AMAZONIA PERUANA
FACULTAD DE INGENIERIA DE SISTEMAS E INFORMATICA

ACTA DE EXAMEN ORAL DE SUFICIENCIA PROFESIONAL

Siendo las 21:00 horas del día 23 de Agosto del 2014, en las instalaciones del Auditorio de la Facultad de Ingeniería de sistemas e Informática de la Universidad Nacional de la Amazonia Peruana, sito en la calle Moore Nº 280 - Iquitos, el Jurado Examinador, compuesto por los siguientes miembros:

Presidente : Ing. Saúl Flores Nunta
Primer Miembro : Eco. Wilson del Águila Panaifo
Segundo Miembro : Ing. Rafael Vilca Barbaran

Se procedió, al Acto Académico del Examen Oral de Suficiencia Profesional del Bachiller: **MILAGROS CELESTINA ZURIMA REATEGUI CURTO,** quien sustento el tema **"Mineria de Datos para la Inteligencia de Negocios"**, para optar el Título Profesional De Ingeniero de Sistema e Informática, de acuerdo a lo establecido en el Reglamento de Grados y Título de la Facultad.

Posteriormente, al Acto de Sustentación del Informe Final del bachiller se procedió al cálculo de Calificación y Condición Final, obteniéndose el siguiente resultado:

	Calificaciones	
	En número	En letras
Promedio de la Calificación Final de las Asignaturas.	13.25	Trece Y 25/100
Calificación de la Sustentación del Informe Final.	14.00	catorce Y 00/100
Calificación Final	13.63	Trece Y 63/100

Se desprende que la Condición Final del Bachiller es (marcar el que corresponde):

( ) Aprobado con excelencia (18 a 20 puntos).
( ) Aprobado por unanimidad (15 a 17.9 puntos).
(X) Aprobado por mayoría (12 a 14.9 puntos).
( ) Desaprobado (Menos de 12 puntos).

Siendo las 21:40 Horas del mismo día, se da por concluido el acto, firmando en conformidad los miembros del Jurado Examinador.

Ing. Saúl Flores Nunta
Presidente

Eco. Wilson del Águila Panaifo
Primer Miembro

Ing. Rafael Vilca Barbaran
Segundo Miembro

RESUMEN

En los últimos años, ha existido un gran crecimiento en nuestras capacidades de generar y colectar datos, debido básicamente al gran poder de procesamiento de las máquinas como a su bajo costo de almacenamiento. Dentro de estas enormes masas de datos existe una gran cantidad de información oculta, de gran importancia estratégica, a la que no se puede acceder por las técnicas clásicas de recuperación de la información.

El descubrimiento de esta información oculta es posible gracias a la Minería de Datos (Data Mining), que entre otras sofisticadas técnicas aplica la inteligencia artificial para encontrar patrones y relaciones dentro de los datos permitiendo la creación de modelos, es decir, representaciones abstractas de la realidad, pero es el descubrimiento del conocimiento (KDD, por sus siglas en inglés) que se encarga de la preparación de los datos y la interpretación de los resultados obtenidos, los cuales dan un significado a estos patrones encontrados.

El presente trabajo brinda un panorama de la minería de datos, sus procesos de descubrimiento de conocimientos, las herramientas y aplicaciones disponibles y su evolución futura.

Palabras Claves: Minería de Datos, Descubrimiento de Conocimiento, Aplicaciones y evolución.

Contenido

1. OBJETIVOS.

1.1 General.

El objetivo del presente trabajo es presentar la Minería de Datos como una tecnología poderosa habilitadora de la Inteligencia de Negocios, explicado su campo de acción, sus métodos y tecnologías empleadas, indicar las herramientas disponibles y ahondar en sus potencialidades y advertir sobre sus limitaciones y mitos asociados a ella.

1.2 Específicos.

- Explicar la naturaleza de la minería de datos
- Presentar los modelos de procesos de minería de datos.
- Indicar las herramientas disponibles para la minería de datos.
- Futura evolución de la minería de datos.

2. JUSTIFICACIÓN.

En los últimos años, ha existido un gran crecimiento en nuestras capacidades de generar y colectar datos, debido básicamente al gran poder de procesamiento de las máquinas como a su bajo costo de almacenamiento.

Sin embargo, dentro de estas enormes masas de datos existe una gran cantidad de información oculta, de gran importancia estratégica, a la que no se puede acceder por las técnicas clásicas de recuperación de la información.

El descubrimiento de esta información oculta es posible gracias a la Minería de Datos (Data Mining), que entre otras sofisticadas técnicas aplica la inteligencia artificial para encontrar patrones y relaciones dentro de los datos permitiendo la creación de modelos, es decir, representaciones abstractas de la realidad, pero es el descubrimiento del conocimiento (KDD, por sus siglas en inglés) que se encarga de la preparación de los datos y la interpretación de los resultados obtenidos, los cuales dan un significado a estos patrones encontrados.

Figura 01: Valor del Conocimiento y del dato
Fuente:

Así el valor real de los datos reside en la información que se puede extraer de ellos, información que ayude a tomar decisiones o mejorar nuestra comprensión de los fenómenos que nos rodean (Figura 01). Hoy, más que nunca, los métodos analíticos avanzados son el arma secreta de muchos negocios exitosos. Empleando métodos analíticos avanzados para la explotación de datos, los negocios incrementan sus ganancias, maximizan la eficiencia operativa, reducen sus costos y la incertidumbre.

Algo peor que no tener información disponible es tener mucha información y no saber qué hacer con ella. La **minería de datos o data mining** es la solución a ese problema, ya que por medio de dicha información puede generar escenarios, pronósticos y reportes que apoyen a la toma de decisiones, lo que se traduce en una ventaja competitiva.

3. LA INTELIGENCIA DE NEGOCIOS

Se puede definir la Inteligencia de Negocios como el proceso de analizar los bienes o datos acumulados en la empresa y extraer una cierta inteligencia o conocimiento de ellos. Dentro de la categoría de bienes se incluyen las bases de datos de clientes, información de la cadena de suministro, ventas personales y cualquier actividad de marketing o fuente de información relevante para la empresa. BI apoya a los tomadores de decisiones con la información correcta, en el momento y lugar correcto, lo que les permite tomar mejores decisiones de negocios. La información adecuada en el lugar y momento adecuado incrementa efectividad de cualquier empresa.

De forma general la **Minería de Datos** es una de las maneras de desarrollar la Inteligencia de Negocios (BI) de la empresa de la data que una organización recolecta, organiza y almacena. Se cuenta con un amplio espectro de técnicas de minería de datos que son empleados por las organizaciones para comprender , por ejemplo, el comportamiento de los consumidores y clientes, o para administrar sus propias operaciones y resolver complejos problemas organizacionales.

4. ELEMENTOS DE LA INTELIGENCIA DE NEGOCIO

La Inteligencia de Negocios suele definirse como la transformación de los datos de la compañía en conocimiento para obtener una ventaja competitiva (Gartner). Desde un punto de vista más pragmático, y asociándolo directamente a las tecnologías de la información, podemos definir la Inteligencia de Negocios como el conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información desestructurada (interna y externa a la compañía) en información estructurada, para su explotación directa (reporting, análisis OLAP...) o para su análisis y conversión en conocimiento soporte a la toma de decisiones sobre el negocio. (Ibermática, 2012)

Esta definición pretende abarcar y describir el ámbito integral del entorno BI, reflejado resumidamente en el gráfico que aquí se muestra. Es importante considerar cualquier proyecto BI como un modelo objetivo integral. Algunas organizaciones han desarrollado proyectos

parciales BI, sin tener en cuenta esta visión global, comprometiendo la calidad y efectividad de los resultados obtenidos.

Según la fuente citada solución integral BI se compone de los siguientes elementos:

4.1 **Diseño conceptual de los sistemas.** Para resolver el diseño de un modelo BI, se deben contestar a tres preguntas básicas: cuál es la información requerida para gestionar y tomar decisiones; cuál debe ser el formato y composición de los datos a utilizar; y de dónde proceden esos datos y cuál es la disponibilidad y periodicidad requerida. En otras palabras, el diseño conceptual tiene diferentes momentos en el desarrollo de una plataforma BI: En la fase de construcción del datawarehouse y datamarts, primarán los aspectos de estructuración de la información según potenciales criterios de explotación. En la fase de implantación de herramientas de soporte a la alta dirección, se desarrolla el análisis de criterios directivos: misión, objetivos estratégicos, factores de seguimiento, indicadores clave de gestión o KPIs, modelos de gestión... en definitiva, información para el qué, cómo, cuándo, dónde y para qué de sus necesidades de información. Estos momentos no son, necesariamente, correlativos, sino que cada una de las etapas del diseño condiciona y es condicionada por el resto.

4.2 **Construcción y alimentación del datawarehouse y/o de los datamarts.** Un datawarehouse es una base de datos corporativa que replica los datos transaccionales una vez seleccionados, depurados y especialmente estructurados para actividades de consulta y reporte. Un datamart (o mercado de datos) es una base de datos especializada, departamental, orientada a satisfacer las necesidades específicas de un grupo particular de usuarios (en otras palabras, un datawarehouse departamental, normalmente subconjunto del corporativo con transformaciones específicas para el área a la que va dirigido).

La vocación del data warehouse es aislar los sistemas operacionales de las necesidades de información para la gestión, de forma que cambios en aquéllos no afecten a éstas, y viceversa (únicamente cambiarán los mecanismos de alimentación, no la estructura, contenidos, etc.). No diseñar y estructurar convenientemente y desde un punto de vista corporativo el data warehouse y los datamarts generará problemas que pueden condenar al fracaso cualquier esfuerzo posterior: información para la gestión obtenida directamente a los sistemas operacionales, florecimiento de datamarts descoordinados en diferentes departamentos, etc.

4.3 **Herramientas de explotación de la información:** es el área donde más avances se han producido en los últimos años. Sin embargo, la proliferación de soluciones mágicas y su aplicación coyuntural para solucionar aspectos puntuales ha llevado, en ocasiones, a una situación de desánimo en la organización respecto a los beneficios de una solución BI. Sin entrar a detallar las múltiples soluciones que ofrece el mercado, a continuación se identifican los modelos de funcionalidad o herramientas básicas (cada

producto de mercado integra, combina, potencia, adapta y personaliza dichas funciones):

- **Query&reporting**: herramientas para la elaboración de informes y listados, tanto en detalle como sobre información agregada, a partir de la información de los datawarehouses y datamarts. Desarrollo a medida y/o herramientas para una explotación libre.
- **Cuadro de mando analítico (EIS tradicionales):** elaboración, a partir de datamarts, de informes resumen e indicadores clave para la gestión **(KPI)**, que permitan a los gestores de la empresa analizar los resultados de la misma de forma rápida y eficaz. En la práctica es una herramienta de query orientada a la obtención y presentación de indicadores para la dirección (frente a la obtención de informes y listados).
- **Cuadro de mando integral o estratégico (Balanced Scorecard):** este modelo parte de que la estrategia de la empresa es el punto de referencia para todo proceso de gestión interno. Con él los diferentes niveles de dirección y gestión de la organización disponen de una visión de la estrategia de la empresa traducida en un conjunto de objetivos, iniciativas de actuación e indicadores de evolución. Los objetivos estratégicos se asocian mediante relaciones causa- efecto y se organizan en cuatro áreas o perspectivas: financiera, cliente, procesos y formación o desarrollo. El cuadro de mando integral es una herramienta que permite alinear los objetivos de las diferentes áreas o unidades con la estrategia de la empresa y seguir su evolución.
- **OLAP (On-Line Analytical Processing):** herramientas que manejan interrogaciones complejas de bases de datos relacionales, proporcionando un acceso multidimensional a los datos, capacidades intensivas de cálculo y técnicas de indexación especializadas. Permiten a los usuarios trocear sus datos planteando consultas sobre diferentes atributos o ejes. Utilizan un servidor intermedio para almacenar los datos multidimensionales precalculados de forma que la explotación sea rápida.
- **Datamining (minería de datos):** Son auténticas herramientas de extracción de conocimiento útil, a partir de la información contenida en las bases de datos de cualquier empresa. El objetivo que se persigue es descubrir patrones ocultos, tendencias y correlaciones, y presentar esta información de forma sencilla y accesible a los usuarios finales, para solucionar, prever y simular problemas del negocio. La Minería de Datos incorpora la utilización de tecnologías basadas en redes neuronales, árboles de decisión, reglas de inducción, análisis de series temporales... y visualización de datos.

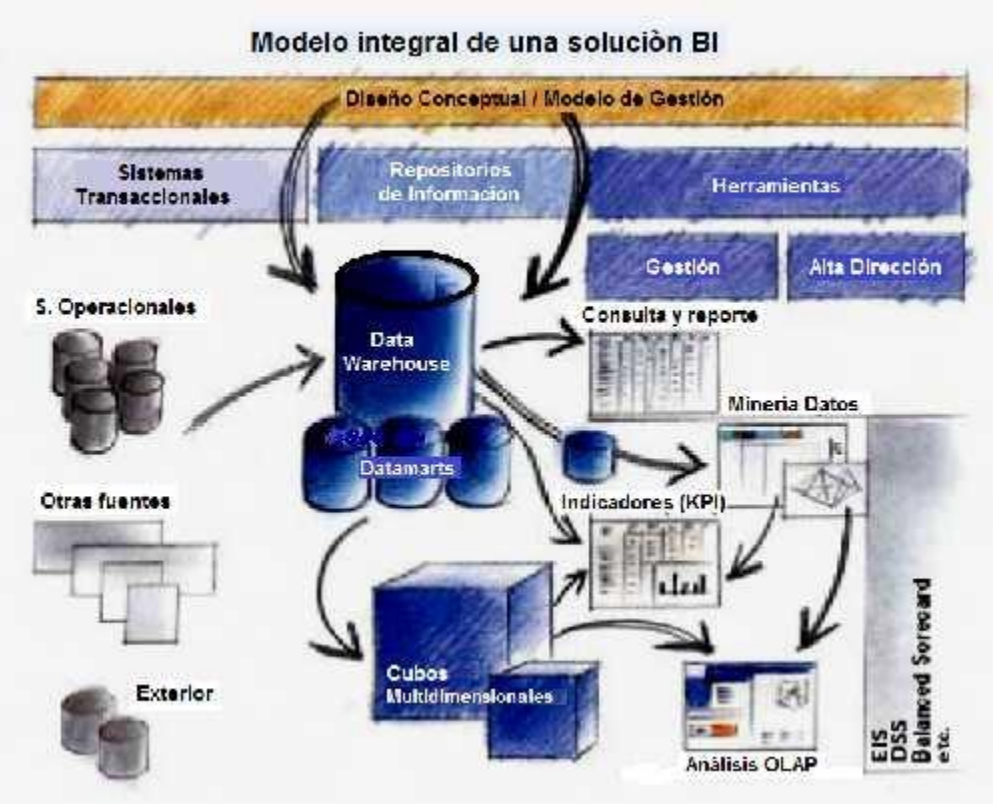

Figura 02: Arquitectura Inteligencia de Negocios
Fuente: Ibermática (2012)

5. EL PROCESO DE DESCUBRIMIENTO DE CONOCIMIENTO DE LOS DATOS (KDD).

Fayyad et al (1999), define el proceso KDD como: "*El proceso no trivial de identificación válida, novedosa, potencialmente útil, y en última instancia, los patrones comprensibles en los datos*".

Explica el autor que la expresión "*patrón*" va más allá de su sentido tradicional para incluir modelos o estructura de datos. En esta definición, datos comprende un conjunto de hechos (por ejemplo, casos en una base de datos), y el patrón es una expresión en algún lenguaje que describe un subconjunto de los datos (o un modelo aplicable a ese subconjunto). El proceso de término implica que hay muchos pasos que involucran la preparación de datos; búsqueda de patrones, evaluación de conocimientos y el refinamiento, todo repetido en múltiples iteraciones. El proceso se supone que es no trivial en que va más allá de forma cerrada de computación cantidades; es decir, debe implicar la búsqueda de estructura, modelos, patrones o parámetros. Los patrones descubiertos deben ser válidos para los nuevos datos con algún grado de certeza. También queremos patrones para ser novela (al menos para el sistema y preferiblemente para el usuario) y potencialmente útil para el usuario o tarea.

Por último, los patrones deben ser comprensibles, si no inmediatamente, entonces después de un post-procesamiento. Esta definición implica que podemos definir medidas cuantitativas para evaluar los patrones extraídos. En muchos casos, es posible definir las medidas de seguridad (por ejemplo, que se estima la precisión de clasificación) o la utilidad (por ejemplo, ganancia, tal vez en dólares ahorrados debido a mejores predicciones o aceleración en el tiempo de

respuesta de un sistema). Tales nociones como la novedad y comprensibilidad son mucho más subjetiva. En ciertos contextos, la comprensibilidad se puede estimar a través de la simplicidad (por ejemplo, el número de bits necesarios para describir un patrón). Una noción importante, llamado “grado de interés”, generalmente se toma como una medida general del valor patrón, combinando validez, novedad, utilidad y simplicidad. Intereses funciones se pueden definir de forma explícita o implícitamente se pueden manifestar a través de un pedido colocado por el sistema KDD en los patrones o modelos descubiertos.

La minería de datos es un paso en el proceso KDD que consiste en una enumeración de patrones (o modelos) sobre los datos, sujeto a algunas limitaciones de cálculo de eficiencia aceptables. Dado que los patrones enumerables sobre cualquier conjunto de datos finito son potencialmente infinitos, y porque la enumeración de patrones implica alguna forma de búsqueda en un espacio grande, las limitaciones computacionales ponen límites severos sobre el subespacio que puede ser explorado por un algoritmo de minería de datos.

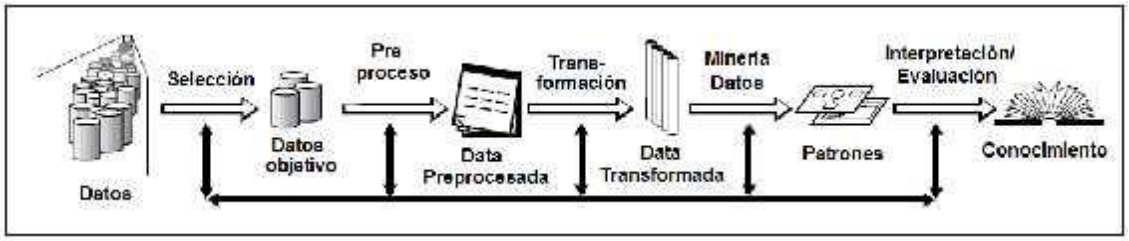

Figura 03: Pasos en el proceso KDD
Fuente: Fayyad et al, 1999.

El proceso KDD se describe en la Figura 03 es interactivo e iterativo (con muchas decisiones tomadas por el usuario), que incluyeron numerosas medidas, que se resumen como.:

1. Aprender el dominio de aplicación: incluye conocimientos previos relevantes y los objetivos de la aplicación
2. Creación de un objetivo conjunto de datos: incluye la selección de un conjunto de datos o concentrarse en un subconjunto de variables o muestras de datos en la que el descubrimiento se va a realizar
3. Limpieza de datos y preprocesamiento: incluye las operaciones básicas, tales como la eliminación de ruido o valores atípicos en su caso, la recogida de la información necesaria para modelar o cuenta para el ruido, la determinación de estrategias para el manejo de campos de datos que faltan, y la contabilidad de información de la secuencia del tiempo y los cambios conocidos, así como de decidir cuestiones de DBMS, tales como tipos de datos, esquemas, y la cartografía de los desaparecidos y los valores desconocidos
4. La reducción de datos y proyección: incluye la búsqueda de características útiles para representar los datos, dependiendo del objetivo de la tarea, y el uso de métodos de reducción de dimensionalidad o de transformación para reducir el número de efectivos

de las variables en estudio o para encontrar representaciones invariantes para los datos

5. La elección de la función de minería de datos: incluye decidir el propósito del modelo derivado por el algoritmo de minería de datos (por ejemplo, el resumen, la clasificación, la regresión, y la agrupación)
6. Elegir el algoritmo de minería de datos (s): incluye la selección de método (s) que se utilizará para la búsqueda de patrones en los datos, como por ejemplo decidir qué modelos y parámetros puede ser apropiada (por ejemplo, modelos para datos categóricos son diferentes de modelos en los vectores más reales) y juego un método de minería de datos en particular con los criterios generales del proceso de KDD (por ejemplo, el usuario pueden estar más interesados en la comprensión del modelo que en sus capacidades de predicción)
7. La minería de datos: incluye la búsqueda de patrones de interés en una forma de representación en particular o un conjunto de tales representaciones, incluyendo las reglas de clasificación o árboles, regresión, clustering, modelado de secuencia, la dependencia, y el análisis de la línea
8. Interpretación: incluye la interpretación de los patrones descubiertos y posiblemente volver a cualquiera de los pasos anteriores, así como la posible visualización de los patrones extraídos, la eliminación de patrones redundantes o irrelevantes, y la traducción de los útiles en términos comprensibles por los usuarios
9. Utilizando el conocimiento descubierto: incluye la incorporación de estos conocimientos en el sistema de rendimiento, la adopción de medidas basadas en el conocimiento, o simplemente documentarlo y notificarlo a los interesados, así como la revisión y resolución de los posibles conflictos con creía anteriormente (o extraído) conocimiento .

La mayor parte del trabajo previo sobre KDD se centró principalmente en la etapa de extracción de datos. Sin embargo, los otros pasos igualmente, si no son más importantes para la aplicación exitosa de KDD en la práctica. Ahora nos centramos en el componente de minería de datos, que ha recibido con mucho, la mayor atención en la literatura.

6. MINERÍA DE DATOS

La minería de datos es un término empleado para describir la búsqueda de conocimiento proveniente de grandes cantidades de datos. Técnicamente hablando, la minería de datos es un proceso que emplea técnicas provenientes de la estadística, matemática y la inteligencia artificial para identificar y extraer información útil y el conocimiento subsecuente desde grandes cantidades de data. Este conocimiento o patrón de datos puede ser de la forma de reglas de negocio, afinidades, correlaciones, tendencias, o modelos de predicción.

6.1 ¿Cómo trabaja la Minería de Datos?

Usando la data existente y relevante, la minería de datos construye modelos para identificar patrones entre los atributos de una base de datos. Los modelos son representaciones matemáticas (relaciones lineales simples y/o relaciones no lineales

altamente complejas), que identifica los patrones entre los atributos de los objetos (clientes, por ejemplo), descritos en la base de datos. Algunos de estos patrones son explicativos (explican las relaciones y afinidades entre los atributos de los objetos), mientras que otros son predictivos (pronosticando futuros valores de ciertos atributos del objeto). En general la minería de datos busca identificar cuatro tipos de patrones:

1. **Asociaciones**. Encuentra la ocurrencia recurrente común de grupos de objetos, tales como que la cerveza y embutidos van juntos en el carrito de compras.
2. **Predicciones**, que indican la posible futura ocurrencia de ciertos eventos basado en lo que ocurrió en el pasado, tales como la temperatura del día o el nivel de ventas de un producto.
3. **Clusters o conglomerados**, donde se identifica la agrupación natural de los objetos basados en sus características conocidas, tal como asignar clientes en diferentes segmentos en sus características demográficas y en su comportamiento de compra.
4. **Relaciones secuenciales**, que descubre los eventos relacionados en el tiempo, tales como descubrir la secuencia que los clientes que abren una cuenta de ahorro terminan abriendo un fondo de inversión en el transcurso del año.

Estos tipos de patrones ya habìan sido descubiertos manualmente en el pasado de la data disponible en esos tiempoos, pero el incremento del volumen de nuestros dìas ha originado la necesidad de nuevoas soluciones y automatismos. Como los conjuntos de datos han aumentado en tamaño y complejidad tambièn lo ha hecho el automatimo de sus procesos, generando metodologìas, algoritmos y procesos sofisticados. La manifestaciòn moderna de estos automatismos es lo que se conoce como minerìa de datos.

6.2 Procesos de Minería de Datos:

En miras de aplicarlo exitosamente, un proyecto de minería de datos debe ser observado como un proceso que sigue una metodología estándar en vez de solo el uso de un conjunto de herramientas tecnológicas. Ante la necesidad existente en el mercado de una aproximación sistemática para la realización de los proyectos de minería de datos, diversas empresas y consultorías han especificado un proceso de modelado diseñado para guiar al usuario a través de una sucesión de pasos que le dirijan a obtener buenos resultados. Así SAS propone la utilización de la metodología SEMMA (Sample, Explore, Modify, Model, Assess). En 1999 un importante consorcio de empresas europeas, NCR (Dinamarca), AG(Alemania), SPSS (Inglaterra) y OHRA (Holanda), unieron sus recursos para el desarrollo de la metodología de libre distribución CRISP-DM (Cross- Industry Standard Processfor Data Mining). Esta

metodología, junto con la metodología SEMMA, son las dos principales metodologías utilizadas por los analistas en los proyectos de Minería de Datos.

6.2.1 **CRISP-DM.**

Esta metodología inicialmente fue desarrollada por tres empresas que iniciaron sus investigaciones en el tema de la Minería de Datos: DaimlerChrysler (luego conocido como DaimlerBenz) quien siempre implementó principiosy técnicas de minería de datos en sus negocios, SPSS quien provee servicios basados en Minería de Datos desde 1990, y NCR.

- **Como metodología**, incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas.
- **Como modelo de proceso**, CRISP-DM ofrece un resumen del ciclo vital de minería de datos.

La metodología CRISP – DM, como lo muestra la Fig.01, está descrita en términos de un modelo de proceso jerárquico, que consiste en una serie de tareas descritas en cuatro niveles de abstracción (de lo general a lo específico):

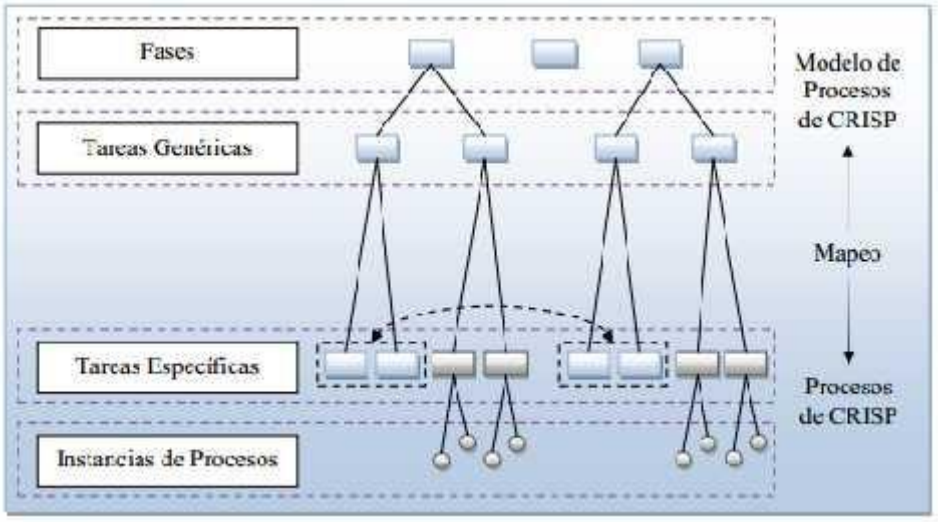

Figura 04: Modelo jerárquico CRISP-DM.
Fuente:

- **Fase:** Se le denomina fase al asunto o paso dentro del procesoCRISP-DM consta de 6 fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelación, evaluación y explotación.
- **Tarea genérica:** Cada fase está formada por tareas genéricas, o sea, la tarea genérica es la descripción de las actividades que se realizan dentro de cada fase. Por ejemplo, la tarea Limpiar los datos es una tarea genérica.
- **Tarea especializada:** La tarea especializada describe cómo se pueden llevar a cabo las tareas genéricas en situaciones específicas. Por

ejemplo, la tarea Limpiar los datos tiene tareas especializadas, como limpiar valores numéricos, y limpiar valores categóricos.

- **Instancias de proceso:** Las instancias de proceso son las acciones y resultados de las actividades realizadas dentro de cada fase del proyecto.

El ciclo vital del modelo contiene seis fases con flechas que indican las dependencias más importantes y frecuentes entre fases. La secuencia de las fases no es estricta. De hecho, la mayoría de los proyectos avanzan y retroceden entre fases si es necesario.

Figura 05: Fases de la Metodología CRISP-DM
Fuente:

Las fases del proyecto de Minería de acuerdo a lo establecido por la metodología CRISP-DM interactúan entre ellas de forma iterativa durante el desarrollo del proyecto. La secuencia de las fases no siempre es ordenada, o en ocasiones si se determina al realizar la evaluación que los objetivos del negocio no se cumplieron se debe regresar y buscar las causas del problema para redefinirlo.

1. **Análisis del negocio**, Esta fase inicial se centra en el entendimiento de los objetivos del proyecto y los requerimientos desde una perspectiva del negocio, para convertir este conocimiento en un problema de definición de minería de datos y un plan preliminar diseñado para alcanzar los objetivos.
2. **Análisis de los Datos**, Esta fase inicia con una colección inicial de datos y procede con actividades para familiarizarse con ellos, identificar problemas de calidad en los mismos, descubrir una primera idea de estos o detectar conjuntos interesantes que permitan formar hipótesis en la búsqueda de información escondida.

3. **Preparación de Datos**, Cubre todas las actividades para construir la base final de datos (datos que serán el alimento de las herramientas de modelado) desde una base en bruto. Es preferible que las tareas de preparación de datos se realicen varias veces y no en un orden preestablecido. Estas tareas incluyen tabulación, documentación y selección de atributos, también como transformación y limpieza de datos para las herramientas de modelado.
4. **Construcción de modelos de Minería de Datos**, Se seleccionan y aplican varias técnicas, y sus parámetros son calibrados a los valores óptimos. Por lo general hay varias técnicas para el mismo tipo de problema. Algunas técnicas tienen requerimientos específicos en la forma de los datos, por lo tanto será a menudo necesario devolverse a la fase de preparación de datos pruebas y evaluación.
5. **Evaluación**, Al llegar a esta fase se ha construido un modelo (o modelos) que aparentan tener una alta calidad desde la perspectiva del análisis de datos. Antes de proceder a la entrega final del modelo es importante evaluarlo más a fondo y revisar los pasos ejecutados para construirlo, de tal forma que este lo más cercano posible de alcanzar los objetivos del negocio. Un objetivo clave es determinar si hay algún evento importante del negocio que no haya sido considerado lo suficiente. Al final de esta fase, se debe tener una decisión sobre el uso de los resultados de minería de datos.
6. **Explotación**, La creación del modelo por lo general no es el final del proyecto. Incluso si el propósito del modelo es incrementar conocimiento sobre los datos, el conocimiento ganado necesitará ser organizado y presentado de una manera que el cliente lo pueda usar. A menudo implica aplicar modelos en vivo dentro del proceso de toma de decisiones de una organización, por ejemplo, en la personalización en tiempo real de las páginas web o la puntuación repetida en bases de datos de mercadeo. Sin embargo, dependiendo de los requerimientos, la fase de despliegue puede ser tan simple como generar un reporte o tan compleja como implementar un proceso repetible de minería de datos a través de la empresa. En muchos casos es el cliente, no el analista de datos, quien realiza los pasos de despliegue. Sin embargo, incluso si el analista no carga con el esfuerzo de despliegue, es importante que el cliente entienda que acciones deben ser llevadas a cabo para hacer uso de los modelos creados.

6.2.2 **SEMMA**

SAS Institute, desarrollador de esta metodología, la define como el proceso de selección, exploración y modelado de grandes cantidades de datos para

descubrir patrones de negocio desconocidos. El nombre de esta terminología es el acrónimo correspondiente a las cinco fases básicas del proceso (Sample-Explore-Modify-Model-Assess).

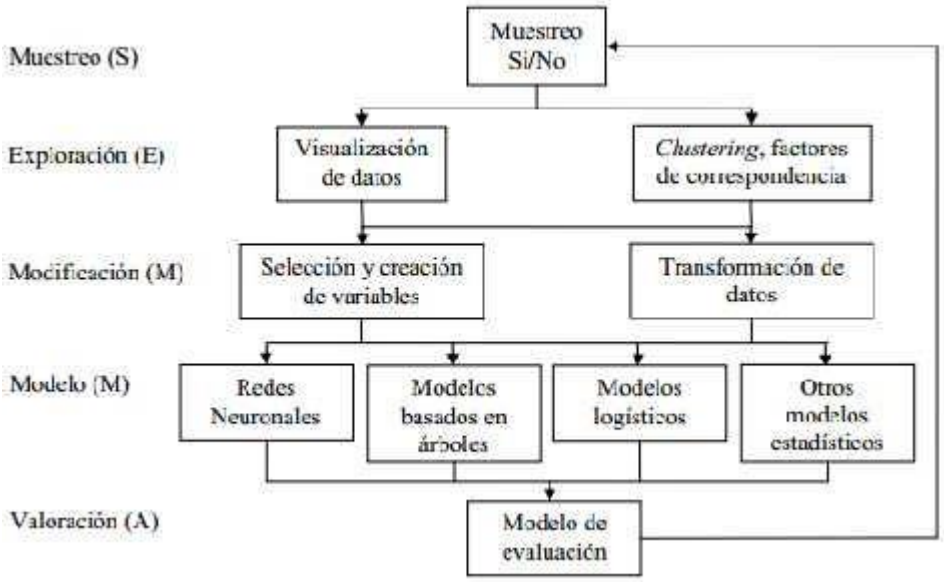

Figura 06: Fases del Proceso SEMMA
Fuente:

Empezando con una muestra estadística pequeña representativa de la data, SEMMA facilita aplicar análisis exploratorio de datos y técnicas de visualización, seleccionar y transformar las variables predictivas más significativas, modelar las variables para predecir las salidas y confirmar la exactitud del modelo

La explicación de las fases es la siguiente:

- **Muestreo**. Se busca extraer una porción de datos lo suficientemente grande para contener información significativa, pero reducida para manipularla rápidamente. Si los patrones generales aparecen en los datos en su conjunto, estos se pueden distinguir en una muestra representativa.
- **Explorar**, Se desea explorar los datos buscando tendencias y anomalías imprevistas para obtener una comprensión total de los mismos. Esta fase ayuda a refinar el proceso de descubrimiento. Si visualmente no hay un resultado claro se pueden tratar los datos por medio de técnicas estadísticas como el análisis factorial, de correspondencias y agrupaciones. A manera de ejemplo, en la minería de datos de campañas de correo directo, el agrupamiento podría revelar grupos de compradores con distintos patrones de ordenamiento, y sabiendo esto, se crea la oportunidad de generar correos personalizados o promociones.

- **Modificar**, Se modifican los datos por medio de la creación, selección y transformación de variables, para centrar el proceso de selección del modelo. Basado en los descubrimientos en la fase de exploración, puede haber la necesidad de manipular los datos para incluir información como la de agrupamiento de compradores y subgrupos significativos, o introducir nuevas variables. También puede ser necesario buscar valores extremos (bordes) y reducir el número de variables, para reducir a los más significativos.

  También puede ser necesario modificar datos cuando la información "minada" cambie. Debido a que la minería de datos es un proceso dinámico e iterativo, puede actualizar los métodos o los modelos cuando esté disponible nueva información.
- **Modelar**, Se modelan los datos permitiendo que el software busque automáticamente una combinación de datos que prediga con cierta certeza un resultado deseado. Las técnicas de modelado en minería de datos incluyen las redes neuronales, modelos de árboles de decisión, modelos lógicos y otros modelos estadísticos (como los análisis de serie de tiempo, razonamiento basado en memoria y componentes principales). Cada uno tiene sus fortalezas, y dependiendo de la información se debe aplicar el más adecuado según las situaciones concretas para el análisis con la minería de datos. Por ejemplo, las redes neuronales son muy buenas en la conexión de relaciones no lineales de gran complejidad.
- **Evaluar**, Se califican los datos mediante la evaluación de la utilidad y fiabilidad de los resultados del proceso de minería de datos. Una forma común de evaluación de un modelo es la de aplicar el modelo a una porción aparte de resultados obtenidos durante el muestreo. Si el modelo es válido, debería funcionar para esta muestra, así como para la muestra utilizada en la construcción del modelo. De manera similar, se puede probar el modelo nuevamente con los datos conocidos. Por ejemplo, si se sabe cuáles clientes tienen altas tasas de retención y su modelo predice la retención, puede probar si el modelo selecciona estos clientes acertadamente.

6.2.3 **Comparación de Metodologías de Minería de Datos**

Las metodologías SEMMA y CRISP-DM comparten la misma esencia, estructurando el proyecto de Minería de Datos en fases que se encuentran interrelacionadas entre sí, convirtiendo el proceso de Minería de Datos en un proceso iterativo e interactivo. La metodología SEMMA se centra más en las características técnicas del desarrollo del proceso, mientras que la metodología CRISP-DM, mantiene una perspectiva más amplia respecto a los objetivos

empresariales del proyecto. Esta diferencia se establece ya desde la primera fase del proyecto de Minería de Datos donde la metodología SEMMA comienza realizando un muestreo de datos, mientras que la metodología CRISP-DM comienza realizando un análisis del problema empresarial para su transformación en un problema técnico. Desde ese punto de vista más global se puede considerar que la metodología CRISP-DM está más cercana al concepto real de proyecto, pudiendo ser integrada con una Metodología de Gestión de Proyectos específica que completaría las tareas administrativas y técnicas.

Tabla 02: Comparativa Metodologías de proyectos minería de datos
Fuente: Camargo y Silva, 2011.

CRISP - DM	SEMMA
Abierta	Cerrada (Abierta en los aspectos generales únicamente)
Funciona en cualquier esquema que aplique minería de datos. Permite que cualquier sistema informático pueda seguir estos pasos	Funciona específicamente en SAS
Implica retroalimentación, es cíclica	Implica retroalimentación, es cíclica
Fases: Entendimiento del negocio, Entendimiento de los datos, Preparación de losdatos, Modelado, Evaluado,Despliegue	Fases: Muestreo, Explorar, Modificar, Modelar, Evaluar
Metodología	Secuencia Lógica
Permite aplicar cualquier modelo estadístico	Está obligado a los modelos estadísticos que tenga incorporados la herramienta Enterprise Miner
Enfocada a resultados empresariales	Enfocada a resultados del proceso
Sigue el esquema propuesto en KDD	Sigue el esquema propuesto en KDD
Libre distribución	Distribución en clientes SAS

Otra diferencia significativa entre la metodología SEMMA y la metodología CRISP-DM radica en su relación con herramientas comerciales. La metodología SEMMA sólo es abierta en sus aspectos generales ya que está muy ligada a los productos SAS donde se encuentra implementada. Por su parte la metodología CRISP-DM ha sido diseñada como una metodología neutra respecto a la herramienta que se utilice para el desarrollo del proyecto de Minería de Datos siendo su distribución libre y gratuita.

7. MODELOS DE MINERÍA DE DATOS.

Un modelo de minería de datos es un conjunto de datos, estadísticas y patrones que se pueden aplicar a los nuevos datos para generar predicciones y deducir relaciones. El modelado de datos se refiere a un grupo de procesos en los que se combinan varios conjuntos de datos y se analizan mediante el empleo de algoritmos para descubrir relaciones o patrones. El objetivo del modelado de datos es utilizar los datos del pasado para informar a los futuros esfuerzos.

En función de su propósito general los modelos pueden ser descriptivos o predictivos.

- **Modelos descriptivos** (describen el comportamiento de los datos de forma que sea interpretable por un usuario experto).

- **Modelos predictivos** (además de describir los datos, se utilizan para predecir el valor de algún atributo desconocido).

Existen numerosos modelos de datos cuya diferencia estriba tanto en el algoritmo empleado como en los objetivos propuestos. Se presenta como ejemplo los modelos más empleados, pero no son de ninguna manera los únicos.

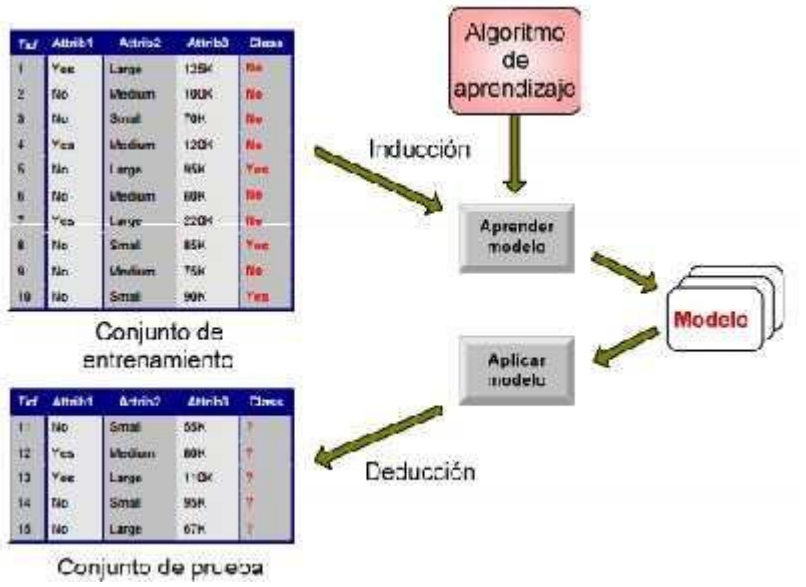

Figura 07: Creación de Modelo
Fuente: Microsoft, 2011

En la mayoría de los casos se emplea una muestra del conjunto de datos, llamado conjunto de entrenamiento, y se emplea un algoritmo de aprendizaje, tratando de encontrar la mejor configuración para detectar el patrón, estamos aprendiendo el modelo a utilizar. Una vez establecido el tipo de algoritmo, y por tanto el modelo, lo aplicamos en un conjunto de datos de prueba, tratando de probar las hipótesis de trabajo.

7.1 Clasificación.

Las observaciones 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. Se dice que este es un ***método supervisado***, porque se conoce cómo es la entrada y el resultado de la salida.

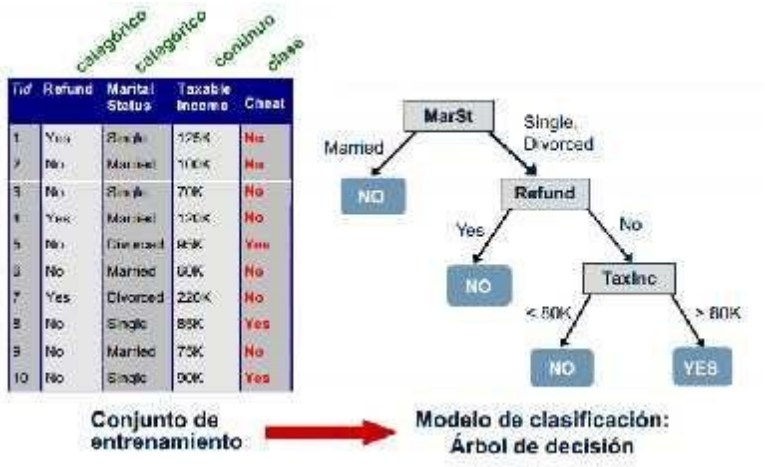

Figura 08: Árbol de Clasificación
Fuente: Microsoft, 2011.

En este caso se emplea el conjunto de entrenamiento y en base al algoritmo de clasificación se obtiene el árbol de decisión. (Figura 08)

7.2 Clusters o conglomerados.

Se construyen grupos de observaciones similares según un criterio prefijado. El proceso de clustering (conglomerado) 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.

Los modelos de agrupación en clústeres identifican las relaciones en un conjunto de datos que no se podrían derivar lógicamente a través de la observación casual. Por ejemplo, puede discernir lógicamente que las personas que se desplazan a sus trabajos en bicicleta no viven, por lo general, a gran distancia de sus centros de trabajo. Sin embargo, el algoritmo puede encontrar otras características que no son evidentes acerca de los trabajadores que se desplazan en bicicleta.

El algoritmo de clústeres de Microsoft identifica primero las relaciones de un conjunto de datos y genera una serie de clústeres basándose en ellas. Un gráfico de dispersión es una forma útil de representar visualmente el modo en que el algoritmo agrupa los datos, tal como se muestra en el siguiente diagrama. El gráfico de dispersión representa todos los casos del conjunto de datos; cada caso es un punto del gráfico. Los clústeres agrupan los puntos del gráfico e ilustran las relaciones que identifica el algoritmo.

En el siguiente diagrama, el clúster A representa los datos sobre las personas que suelen conducir hasta el trabajo, en tanto que el clúster B representa los datos sobre las personas que van hasta allí en bicicleta.

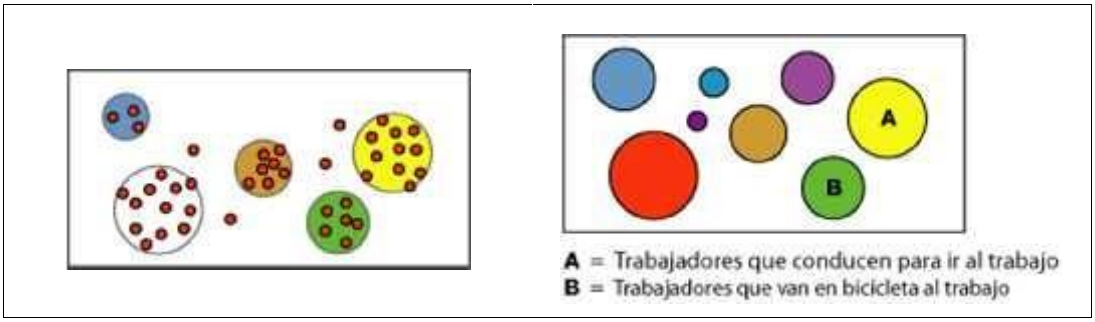

Figura 09: Modelo Cluster o aglomerados.
Fuente: Microsoft, 2011.

Después de definir los clústeres, el algoritmo calcula el grado de perfección con que los clústeres representan las agrupaciones de puntos y, a continuación, intenta volver a definir las agrupaciones para crear clústeres que representen mejor los datos. El algoritmo establece una iteración en este proceso hasta que ya no es posible mejorar los resultados mediante la redefinición de los clústeres.

El algoritmo de clústeres se diferencia de otros algoritmos de minería de datos, en que no se tiene que designar una columna de predicción para generar un modelo de agrupación en clústeres. El algoritmo de clústeres entrena el modelo de forma estricta a partir de las relaciones que existen en los datos y de los clústeres que identifica el algoritmo.

7.3 Asociaciones.

Las observaciones 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.

Dado un conjunto de transacciones, encontrar reglas que describen tendencias en los datos: Detectar cuándo la ocurrencia de un artículo está asociada a la ocurrencia de otros artículos en la misma transacción.

Figura 10: Creación de Asociación
Fuente: Microsoft, 2011

7.4 Patrones secuenciales.

Se trata de identificar patrones de comportamiento y tendencias. Un ejemplo sería intensidades de expresión en micro arreglos que permiten distinguir entre diferentes expresiones de genes para individuos con cáncer o sin él.

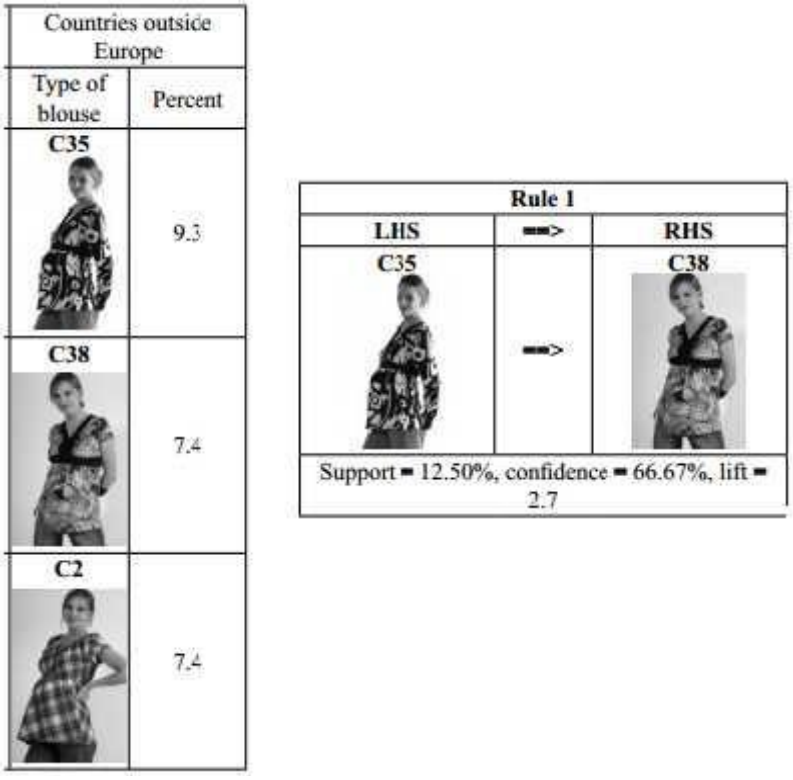

Figura 11: Creación de modelo de secuencia
Fuente:

8. APLICACIONES DE MINERÍA DE DATOS.

La minería de datos se ha convertido en una herramienta popular para solucionar retos complejos del negocio. Se ha comprobado con su empleo y al paso del tiempo su utilidad en muchas áreas, algunas de las cuales se muestran para ejemplo a continuación.

8.1 Negocios

La minería de datos puede contribuir significativamente en las aplicaciones de administración empresarial basada en la relación con el cliente. En lugar de contactar con el cliente de forma indiscriminada a través de un centro de llamadas o enviando cartas, sólo se contactará con aquellos que se perciba que tienen una mayor probabilidad de responder positivamente a una determinada oferta o promoción. Por lo general, las empresas que emplean minería de datos ven rápidamente el retorno de la inversión, pero también reconocen que el número de modelos predictivos desarrollados puede crecer muy rápidamente. En lugar de crear modelos para predecir qué clientes pueden cambiar, la empresa podría construir modelos separados para cada región y/o para cada tipo de cliente. También puede querer determinar qué clientes van a ser rentables durante una ventana de tiempo (una quincena, un mes, ...) y sólo enviar las ofertas a las personas que es probable que sean rentables. Para mantener esta cantidad de modelos, es necesario gestionar las versiones década modelo y pasar a una minería de datos lo más automatizada posible.

8.2 Hábitos de compra en supermercados

El ejemplo clásico de aplicación de la minería de datos tiene que ver con la detección de hábitos de compra en supermercados. Un estudio muy citado detectó que los viernes había una cantidad inusualmente elevada de clientes que adquirían a la vez pañales y cerveza. Se detectó que se debía a que dicho día solían acudir al supermercado padres jóvenes cuya perspectiva para el fin de semana consistía en quedarse en casa cuidando de su hijo y viendo la televisión con una cerveza en la mano. El supermercado pudo incrementar sus ventas de cerveza colocándolas próximas a los pañales para fomentar las ventas compulsivas.

8.3 Patrones de fuga

Un ejemplo más habitual es el de la detección de patrones de fuga. En muchas industrias — como la banca, las telecomunicaciones, etc. — existe un comprensible interés en detectar cuanto antes aquellos clientes que puedan estar pensando en rescindir sus contratos para, posiblemente, pasarse a la competencia. A estos clientes
— y en función de su valor — se les podrían hacer ofertas personalizadas, ofrecer promociones especiales, etc., con el objetivo último de retenerlos. La minería de datos ayuda a determinar qué clientes son los más proclives a darse de baja estudiando sus patrones de comportamiento y comparándolos con muestras de clientes que, efectivamente, se dieron de baja en el pasado

8.4 Fraudes

Un caso análogo es el de la detección de transacciones de blanqueo de dinero o de fraude en el uso de tarjetas de crédito o de servicios de telefonía móvil e, incluso, en la relación de los contribuyentes con el fisco. Generalmente, estas operaciones fraudulentas o ilegales suelen seguir patrones característicos que permiten, con cierto grado de probabilidad, distinguirlas de las legítimas y desarrollar así mecanismos para tomar medidas rápidas frente a ellas.

8.5 Comportamiento en Internet

También es un área en boga el del análisis del comportamiento de los visitantes sobre todo, cuando son clientes potenciales — en una página de Internet. O la utilización de la información — obtenida por medios más o menos legítimos sobre ellos para ofrecerles propaganda adaptada específicamente a su perfil. O para, una vez que adquieren un determinado producto, saber inmediatamente qué otro ofrecerle teniendo en cuenta la información histórica disponible acerca de los clientes que han comprado el primero.

8.6 Genética

En el estudio de la genética humana, el objetivo principal es entender la relación cartográfica entre las partes y la variación individual en las secuencias del ADN humano y la variabilidad en la susceptibilidad a las enfermedades. En términos más llanos, se trata de saber cómo los cambios en la secuencia de ADN de un individuo afectan al riesgo de desarrollar enfermedades comunes (como por ejemplo el cáncer). Esto es muy importante para ayudar a mejorar el diagnóstico, prevención y tratamiento de las enfermedades. La técnica de minería de datos que se utiliza para realizar esta tarea se conoce como "reducción de dimensionalidad multifactorial".

9. APLICACIONES INFORMÁTICAS PARA LA MINERÍA DE DATOS.

El análisis predictivo y otras categorías de análisis avanzados se están convirtiendo en un factor importante en el mercado de análisis de la data. Por los últimos ocho años, Gartner (2014), ha publicado un Cuadrante Mágico de BI y análisis (aunque el nombre del informe ha evolucionado a lo largo de los años). Estos Cuadrantes Mágicos han evaluado los vendedores en todo el espectro de BI y las capacidades analíticas, pero se han centrado principalmente en la capacidad de los vendedores a ofrecer consulta tradicional y generación de informes (descriptivos). En los últimos años, la visualización de datos y análisis predictivo y prescriptivo se han vuelto más importantes para las organizaciones y esto se ha reflejado en los proveedores evaluados y sus posiciones en el Cuadrante Mágico.

En general, el mercado de analítica avanzada se estima en $ 2 mil millones (Gartner 2014), a través de una amplia variedad de industrias (servicios financieros, el comercio electrónico al por menor, y las comunicaciones son, probablemente, el más grande, aunque existen casos de

uso en casi todas las industrias) y geografías (América del Norte y Europa son los mercados más grandes, aunque Asia / Pacífico también está creciendo rápidamente).

La figura 12 muestra el cuadrante mágico con la posición de las empresas proveedoras de plataformas de análisis avanzadas que se utilizan para construir soluciones a partir de cero.

Los líderes son aquellos proveedores con un historial sólido y probado en el mercado que también puedan influir en el crecimiento general del mercado y de dirección. Los líderes son vendedores adecuados para la mayoría de las organizaciones a evaluar. No deben ser los únicos proveedores considerados en un proceso de evaluación, pero al menos dos o tres son susceptibles de ser incluidas en la lista restringida entre cinco y ocho vendedores que las organizaciones consideren.

Retadores tienden a caer en una de dos categorías. O bien son competidores en el mercado a largo plazo que necesitan para revitalizar su visión para estar al tanto de la evolución del mercado y se vuelven más ampliamente influyente, o son vendedores en mercados adyacentes que están entrando en este mercado bien establecido y tener soluciones que pueden ser razonablemente considerados por la mayoría de sus clientes. Como estos vendedores prueban que pueden influir en el mercado más amplio, pueden eventualmente convertirse en líderes.

Figura 12. Cuadrante Mágico para Plataformas AdvancedAnalytics

Fuente: Gartner (2014)

Los visionarios son los vendedores generalmente más pequeños que encarnan las tendencias que están dando forma, o darán forma, el mercado. Representan una oportunidad para que algunas organizaciones para saltar una generación en el uso de la tecnología en el mercado, o proporcionan cierta capacidad convincente que ofrecerá una ventaja competitiva como complemento o sustituto de las soluciones existentes. Como visionarios madurar y demostrar su capacidad de ejecución en el tiempo, pueden pasar a convertirse eventualmente en líderes. Jugadores de nicho también caen en una de dos categorías. Algunos jugadores de nicho son "Visionarios en espera;" que tienen algún grado de visión (a menudo generada internamente), pero están luchando para hacer esa visión convincente para el mercado o para desarrollar el historial de innovación continua que les permitirá avanzar a través del cuadrante de Visionarios. Otros jugadores de nicho son "Retadores en espera;" a menudo son los vendedores de los mercados adyacentes que aún están madurando sus soluciones en este ámbito. Su resistencia del producto y su trayectoria no es suficiente para que sean una opción por defecto seguro para sus clientes existentes (los atributos de un retador), pero si siguen para desarrollar el producto y demostrar el éxito que pueden llegar a ser retadores.

Los vendedores fueron evaluados por los analistas de Gartner, y en la encuesta a los clientes, en los siguientes 13 categorías de funcionalidad:

1. **Acceso a datos -** Código de libre integración de datos básicos; datos avanzados de integración; arquitectura orientada a servicios (SOA), integración de datos Web; extracción básica, transformación y carga (ETL) funcionalidad; funcionalidad ETL avanzada; aplicaciones empresariales de acceso; actualización de datos; (por ejemplo, multimedia) tipos soportados de datos; linaje de datos; datos geoespaciales y la integración de datos de los consumidores; geocodificación; limitaciones.
2. **Visualización y Exploración / Descubrimiento -** gráficos básicos; tipos de gráficos de visualización avanzadas; exportación de visualizaciones en informes y portales web; funciones avanzadas de visualización GUI; estadísticas univariadas y bivariadas; pruebas de significación estadística; procesamiento analítico en línea (OLAP), la interacción visual y la exploración.
3. **Filtrado de datos y la manipulación -** Agrupación y suavizado; generación de función de reducción de dimensionalidad y selección de características; de filtro y búsqueda, de rotación, de agregación y establecer operaciones; transformaciones; pre-procesamiento de la señal; mapeos personalizados; particionamiento de datos.
4. **Analítica avanzada descriptiva -** Clustering y mapas auto-organizativos; análisis de afinidad y gráfico; análisis conjunto y la encuesta; estimación de la densidad; métricas de similitud.
5. **Análisis predictivo -** modelos de regresión; análisis de series de tiempo; redes neuronales; árboles de clasificación y regresión; nuevas técnicas de inducción de reglas; apoyar a las máquinas de vectores; enfoques basados en instancia; Modelado

bayesiano; conjuntos y modelos jerárquicos; importación, la llamada y el desarrollo de otros modelos de predicción; medidas de ajuste; prueba de modelos predictivos.

6. **Optimización -** Solver; heurísticas; diseño de experimentos.
7. **Simulación -** eventos discretos, simulación de Monte Carlo; basado en agente de modelos.
8. **Más de Analítica Avanzada -** análisis de texto básicos; procesamiento de textos; vocabulario, el lenguaje y la gestión de la ontología; análisis de texto avanzados ;análisis multimedia; análisis geoespacial; modelado y la econometría financiera; procesamiento de señales y control.
9. **Analíticas de Negocios Casos de Uso -**marketing; ventas; gestión de riesgos y de gestión de calidad; otros.
10. **Entrega, Integración y Despliegue -**Integración; write-back; Gráficos / cuadros de mando de despliegue Web y la información; apoyo portal; entrega incrustado.
11. **Plataforma y Gestión de Proyectos -**Gestión de metadatos; modelo de gestión; problemas de licencia modelo; gestión de decisiones; scripting y automatización; objetos reutilizan; capacidades multiusuario; pruebas de depuración y la unidad; optimización de tiempo de ejecución; auditoría y registros; cifrado de datos; implementación del cliente; extensibilidad.
12. **La experiencia de usuario -** Facilidad de uso; documentación; orientación; asistentes y ayudas contextuales; comunidad de usuarios; aplicaciones de terceros.
13. **Rendimiento y Escalabilidad -** Big data, en memoria, en la base de datos-técnicas; escalabilidad volumen de datos; eficiencia algorítmica; datos en tiempo real y arroyos.

En la tabla 04, en la página siguiente, se describen las aplicaciones que se considera van a liderar el desarrollo tecnológico de la analítica de datos. Para la elaboración de esta tabla se realizaron 3,285 encuestas, donde cada encuestado brindó un promedio de 3.7 herramientas de herramientas de Minería de Datos empleada.

Tabla 03: Empleo de aplicaciones de Minería de Datos
Fuente:

RapidMiner	44.2%
R	38.5%
Excel	25.8%
SQL	25.3%
Python	19.5%
Weka	17.0%
Knime	15.0%
Hadoop	12.7%
SAS Base	10.9%
Microsoft SQL Server	10.5%

Tabla 04: Aplicaciones para la Analítica de Datos
Fuente: Gartner (2014)

	IBM (www.ibm.com), con sede en Chicago, Illinois, Estados Unidos, adquirió SPSS en 2009 y ha evolucionado su cartera de manera que el análisis predictivo es accesible para varios tipos de usuarios y niveles de habilidad. Mejor conocido por sus productos y soluciones de Estadísticas y Modeler (Minería de Datos), IBM SPSS resuelve una amplia gama de desafíos relacionados con la analítica de clientes, operaciones, amenazas y riesgos.
	KNIME (www.knime.com), con sede en Zurich, Suiza, ofrece una plataforma de código abierto libre, basada en el escritorio de analítica avanzada. También ofrece un comercial, un servidor basado en el lugar o solución en la nube del cliente que proporciona funcionalidad adicional de la empresa. KNIME tiene presencia a través de una variedad de sectores, pero con especial experiencia en ciencias de la vida, el gobierno, la educación y las comunicaciones.
	RapidMiner (www.rapidminer.com), anteriormente conocido como Rápido-I, está basada en Cambridge, Massachusetts, Estados Unidos RapidMiner es un código abierto, solución basada en cliente / servidor también disponible como una solución comercial con la capacidad de trabajar en grande conjuntos de datos y conectarse a más fuentes de datos. La plataforma deriva su extensibilidad a través de la disponibilidad y la integración de otras soluciones de código abierto (por ejemplo, R y Weka) de código fuente.
	SAS (www.sas.com) se basa en Cary, Carolina del Norte, EE.UU. Con más de 40.000 clientes y el mayor ecosistema de usuarios y socios, SAS ha sido tradicionalmente la opción segura para las organizaciones que buscan un entorno de análisis avanzado. SAS tiene fuerza en banca, seguros, servicios de negocios y el gobierno.
	Oracle (www.oracle.com) se basa en Redwood Shores, California, EE.UU. Su Opción AdvancedAnalytics (OAA), un componente opcional de la base de datos Oracle Enterprise Edition, se ha implementado en múltiples geografías e industrias diferentes y facilita una serie de despliegue opciones - desde en las instalaciones y acogido a y embebidos en aplicaciones basadas en la nube.
	Microsoft (www.microsoft.com) está basada en Seattle, Washington, Estados Unidos, y su capacidad de análisis predictivo está incrustado dentro de SQL Server. Estas capacidades se pueden acceder directamente a través de SQL Server oa través de un plug-in de Excel que actúa como interfaz de SQL Server.

10. CASOS DE EMPRESAS QUE EMPLEAN MINERÍA DE DATOS EN EL PERÚ.

En el Perú el trabajo de minería de datos es incipiente, pero algunos sectores como la banca, seguros y las telecomunicaciones vienen adquiriendo un desarrollo acelerado.

10.1 Banco de Crédito.

Cuenta con un Data Warehouse que sirve de repositorio central de información para todas las áreas del Banco. Es explotada mediante herramientas de Inteligencia de Negocios como Microstrategy y BussinesObjects, herramientas de consultas transaccionales como BI-Query, procesos de Minería de Datos con SAS y sirve como fuente de datos para diversos aplicativos CORE de atención al cliente y a la Red de usuarios de la Intranet Corporativa.

El Datawarehouse cuenta con información de Riesgos de Banca, Clientes, Canales, Cobranzas, Servicios y Comercial.

La información está en la base de datos Oracle 10g sobre un servidor AIX 5.3 con más de 4 Terabytes de información, que se actualiza de manera diaria, semanal y mensual mediante procesos de carga (ETL).

10.2 Seguros Pacífico.

Una compañía sólida y de gran trayectoria en el mercado asegurador. Cuenta con un DataMart de Asistencia Médica sobre la base de datos ACSEL/X que se actualiza mensualmente mediante procesos de carga ETL.

10.3 Telefónica I+D.

Empresa del Grupo Telefónica dedicada a la Innovación y el Desarrollo, nace en 1988 con la misión de contribuir a la competitividad y modernidad del Grupo Basada en la innovación y el desarrollo tecnológicos, y con la aplicación de nuevas ideas, conceptos y métodos, desarrolla productos y servicios avanzados.

Telefónica I+D es uno de los primeros centros privados de I+D en España en cuanto a actividad y recursos, y es la primera empresa del continente en número de proyectos europeos de investigación en los que participa. El principal activo de Telefónica I+D es su plantilla, integrada en un 97% por titulados universitarios de 18 nacionalidades.

Actualmente TID colabora con numerosos líderes tecnológicos y organizaciones de 40 países; entre ellas, con más de 150 universidades en todo el mundo. A su vez, participa en los principales foros internacionales de conocimiento tecnológico del

sector de las TIC, construyendo a su alrededor uno de los mayores ecosistemas de innovación europeos.

El laboratorio de Telefónica I&D trabaja en un amplio conjunto de áreas de investigación:

- Sistemas móviles
- Big Data
- Sistemas distribuidos
- Usermodeling& Machine Learning
- Interacción Humano Computador
- Análisis de Multimedia
- Seguridad y Privacidad
- Economía de redes

11. MITOS Y LIMITACIONES DE LA MINERÍA DE DATOS

El oficio de la minería de datos tiene muchos mitos y retos asociados, entre ellos ser enterrado bajo montañas de datos. Algunos riesgos son sólo mitos que necesitan ser desacreditado. Otros, sin embargo, son reales. En esta sección se presentan algunos de estos mitos y conceptos erróneos y luego se describen algunos de los retos y dificultades encontrados comúnmente cuando se realiza la minería de datos, junto con los pasos que se pueden tomar para protegerse de ellos.

Oviedo (2011), lista algunos de los mitos que rodean la Minería de Datos y que dificultan la adopción de la misma, desbaratándolos si son falsos y previniendo si resultan verdaderos.

1. El cliente debe tener implementado un Data Warehouse para ser considerado prospecto potencial.

   Falso. Las herramientas modernas existentes en el mercado nos permiten implementar Minería de Datos con datos provenientes de bases de datos, archivos de Excel, archivos planos etc.

2. La Minería de Datos es para grandes volúmenes de datos.

   Falso. El nivel de aprovechamiento de la información no depende de la cantidad de Gigabytes, Terabytes o Petabytes. Ciertamente es requerido que a la hora de hacer un análisis se cuente con la mayoría de datos relevantes al modelo, pero debemos tener claro que si la realidad de la organización es que maneja solamente cientos o miles de datos, estos son suficientes para identificar los patrones de comportamiento de los datos.

3. Se requiere la eliminación datos basura y datos faltantes.

   Falso. Si bien la limpieza de datos es lo más recomendado. Los algoritmos en los cuales se implementa la Minería de Datos están basados en estadísticas que asumen probabilidades y márgenes de aceptación. Por lo tanto, una cantidad relativamente pequeña de errores y datos faltantes no influye en el resultado del modelo.

4. Se requiere un alto nivel de conocimiento matemático y estadístico.
   Falso. Es cierto que los algoritmos usados en la Minería de Datos se basan en métodos como la inteligencia artificial, aprendizaje automático, estadística, ciencias matemáticas como la lógica, probabilidad, etc. Pero estos algoritmos ya están implementados, lo que se requiere es la comprensión de dichos algoritmos para saber cuál debemos implementar según las necesidades del negocio.
5. Un proyecto de Minería de Datos es complejo, costoso y lleva mucho tiempo.
   Falso. Contrario a lo que se cree, la complejidad de un proyecto de Minería de Datos no proviene de las herramientas, sino de la comprensión "real" del negocio. Típicamente los departamentos de Tecnologías de Información son dados a pensar que saben lo que los usuarios quieren, esto es lo que normalmente lleva un proyecto de este tipo al fracaso. El consumo en tiempo y costo va a radicar mayormente, en la habilidad de comprender el negocio para implementar modelos útiles.
6. No hay recurso humano técnico disponible.
   Verdadero. Lamentablemente, en el ámbito latinoamericano no se cuenta con amplia demanda de estos servicios, esto genera baja oferta tanto a nivel profesional como corporativo.

Khabaza Tom (2005), nos presenta los retos más desafiantes, indicando que algunos no lo son tanto y otros merecen toda nuestra atención.

- **Reto # 1: Enterrado bajo montañas de datos**
  La minería de datos debe ser un proceso interactivo e iterativo en el cual el analista aplica un conocimiento sustancial del negocio y está "comprometido" con los datos. Sin embargo, aquellos que sostienen que la minería de datos se trata de grandes cantidades de datos, a menudo suponen que este proceso debe ser aplicado a todos los datos disponibles.
  Esto puede llevar a intentos de volúmenes de datos para los que el hardware y el software disponible no pueden proporcionar una respuesta interactiva aceptable. En estas situaciones, el proceso de minería de datos se vuelve lento, y para cuando una pregunta se responde, el analista no puede recordar por qué se preguntó.
  La forma de evitar este escollo es emplear alguna forma de muestreo. Por ejemplo, si tenemos un millón de clientes y una deserción anual del 20 por ciento, no necesitamos trazar nuestros gráficos o construir nuestros modelos usando el millón de ejemplos, siendo necesaria solo una muestra de diez o veinte mil desertores y un número equivalente de los clientes leales es probable que sea suficiente para este análisis.
  Tenga en cuenta que esto no significa que los mineros de datos nunca se encontrará con la necesidad de construir modelos de millones de casos; sólo que no deben asumir que deben hacerlo, sólo porque los datos están disponibles.

- **Reto # 2: La misteriosa desaparición de los Terabyte**

  Este es un fenómeno común, pero no siempre una trampa. Se refiere al hecho de que, por un problema de minería de datos dado, la cantidad de datos disponibles y pertinentes puede ser mucho menor que suponía inicialmente.

  Considere el siguiente escenario: Usted es un consultor de minería de datos, y su cliente es un banco grande, que desea extraer datos de sus clientes para determinar el riesgo de crédito. El banco tiene terabytes de datos sobre sus clientes y le preocupa que los recursos informáticos disponibles pueden ser inadecuados para extraer este volumen de datos.

  Así es como la situación podría desarrollarse. Diferentes tipos de crédito (préstamos personales, préstamos comerciales, los descubiertos) presentan diferentes patrones de riesgo de crédito, por lo que cada proyecto de minería de datos se concentrará en un solo tipo de prestatario. Los expertos en el dominio del banco evaluarán una serie de factores a ser relevante, y el banco realizará la planificación anticipada, así comenzaron a recoger datos sobre estos factores hace aproximadamente 18 meses. Desde entonces, se han producido casi un millar de casos de mala deuda.

  Por lo tanto, los datos relevantes consisten en menos de mil casos de mala deuda más una muestra de un suministro abundante de casos de buenas deuda-digamos 3000 registros en total. De alguna manera, la necesidad de terabytes de datos ha desaparecido "misteriosamente".

- **Trampa # 3: la minería de datos desorganizado**

  La minería de datos puede de vez en cuando, a pesar de las mejores intenciones, llevará a cabo de manera ad hoc, sin objetivos claros y sin idea de cómo se utilizarán los resultados. Esto lleva a la pérdida de tiempo y los resultados no utilizables.

  Para producir resultados útiles, es fundamental contar con objetivos claramente definidos de negocios y de minería de datos, formuladas al principio del proyecto, y los planes de despliegue claramente articulados. Una manera simple de asegurar esto es usar un proceso estándar como la práctica estándar entre la industria de la minería de datos (CRISP-DM). Este proceso asegura la correcta preparación para la minería de datos y proporciona un lenguaje común para los métodos y resultados de la comunicación. Herramientas de minería de datos deben apoyar los modelos de procesos estándar.

- **Trampa # 4: Insuficiente conocimiento del negocio**

  En varias ocasiones este artículo ha mencionado el papel crucial que desempeña el conocimiento del negocio de la minería de datos. Sin ella, las organizaciones no pueden ni conseguir resultados útiles ni guiar el proceso de minería de datos hacia ellos. Se supone a veces que el usuario final puede decir razonablemente que el minero de datos: "Aquí están los datos, por favor, vaya, hacer su minería de datos, y se vuelve con las respuestas." Si esto llegara a suceder, el proyecto, en el mejor, tomar

muchas iteraciones largas y costosas para producir resultados útiles. En el peor, los resultados serían un galimatías, y el proyecto sería un fracaso. Esta trampa sólo puede evitarse mediante la participación, en todas las etapas del proceso de minería de datos, tanto para el usuario final y alguien con un conocimiento detallado de la empresa.

Lo ideal sería que el consultor minero de datos tenga el conocimiento del negocio. A falta de ello, el minero de datos debe sentarse literalmente al lado de alguien con el conocimiento empresarial necesario que entienda la cuestión examinada. Para que esto funcione de manera efectiva, se requiere un entorno de minería de datos altamente interactivo con un buen tiempo de respuesta.

- **Trampa # 5: Conocimiento Datos insuficientes**

  Para llevar a cabo la minería de datos, debemos ser capaces de responder a preguntas como "¿Qué significan los códigos en este campo significa?" y "¿Puede haber más de un registro por cliente en este cuadro?". En algunos casos, esta información es sorprendentemente difícil de conseguir. Podría ser que el experto de datos ha dejado la organización o trasladado a otro departamento o, en el caso de los sistemas de legado, no puede ser ningún experto datos en absoluto. Este problema se agrava cuando se subcontrata la gestión de base de datos o almacén de datos: el proveedor externo es mucho menos motivado que la organización de usuarios para mantener esta información "en caso de que podría ser necesario en el futuro."

  No hay ninguna solución sencilla a este problema. Los departamentos de TI deben ser conscientes de la necesidad de mantener la información acerca de las bases de datos de su organización. Además, cuando se propone un proyecto de minería de datos, minería de datos deben considerar cómo es el conocimiento de muchos datos disponibles y evaluar los riesgos provocados por su ausencia o escasez.

- **Trampa # 6: suposiciones erróneas, cortesía de los expertos**

  Expertos en negocios y de datos son recursos cruciales, pero esto no significa que el minero de datos debe aceptar incondicionalmente todas las declaraciones que hacen. La minería de datos debería tratar de confirmar la validez de las declaraciones de los expertos.

  Ejemplos típicos de declaraciones erróneas o engañosas pueden incluir:

  - Ningún cliente puede tener cuentas de estos dos tipos
  - Ningún caso incluirá más de un evento de este tipo
  - Sólo los siguientes códigos estarán presentes en este campo

  Los analistas de datos deben verificar las declaraciones como éstas mediante el examen de los datos. Esto es particularmente importante cuando el procesamiento de los datos dependerá de su exactitud. Idealmente, los errores en las suposiciones sobre los datos pueden ser vistos antes de que lleven a errores en el tratamiento de los datos. Herramientas de minería de datos deben hacer esto fácil de lograr.

- **Trampa # 7: Incompatibilidad de herramientas de minería de datos**

  El proceso de minería de datos requiere una amplia gama de capacidades, así que no es inusual que durante un solo proyecto puede utilizarse una amplia variedad de herramientas. Esto puede, sin embargo, conducir a costos fijos elevados, debido al tiempo y los recursos necesarios para cambiar contextos y convertir datos de un formato a otro. En el peor, esto puede conducir a la omisión de pasos necesarios en el proceso de minería de datos y puede interferir seriamente con el carácter exploratorio de la minería de datos.

  La mejor solución es utilizar un conjunto de herramientas de minería de datos que integra todas las capacidades requeridas. Sin embargo, ningún conjunto de herramientas proporcionará toda posible capacidad, especialmente cuando se toman las preferencias individuales de los analistas en cuenta, por lo que el conjunto de herramientas también deben ser "abiertas", es decir, capaz de interactuar fácilmente con otras herramientas disponibles y las opciones de otros fabricantes.

- **Trampa # 8: Encerrado en la cárcel de datos**

  Además de la apertura en materia de herramientas, soluciones de minería de datos también deben estar abiertos con respecto a los datos. Algunas herramientas de minería de datos requieren que los datos se mantienen en un formato propietario que no es compatible con los sistemas de bases de datos de uso común. (Esto se refiere a veces como la "cárcel de datos"). Esto puede resultar en elevados gastos generales, debido a la necesidad de la transferencia de datos en el formato requerido, y llevar a la dificultad en la implementación de los resultados en los sistemas operativos de una organización. Una buena herramienta de minería de datos se conectará con sus datos a través de las normas comunes.

12. BIG DATA

12.1 Crecimiento incontenible de la data

No se puede hablar del futuro en las técnicas de búsqueda de conocimiento y la evolución de la minería de datos si no se habla del BIG DATA. Empezaremos preguntándonos el crecimiento gigantesco que experimenta la Internet en cuanto contenidos. ¿De dónde proviene toda esa información? Los seres humanos estamos creando y almacenando información constantemente y cada vez más en cantidades astronómicas. Se podría decir que si todos los bits y bytes de datos del último año fueran guardados en CD's, se generaría una gran torre desde la Tierra hasta la Luna y de regreso. (IBM, 2014).

El concepto de Big Data aplica para toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales. Sin embargo, Big Data no se refiere a alguna cantidad en específico, ya que es usualmente utilizado cuando se habla en términos de petabytes y exabytes de datos. Entonces

¿Cuánto es demasiada información de manera que sea elegible para ser procesada y analizada utilizando Big Data? Analicemos primeramente en términos de bytes:

Megabyte = 10^6 = 1,000,000
Gigabyte = 10^9 = 1,000,000,000
Terabyte = 10^{12} = 1,000,000,000,000
Petabyte = 10^{15} = 1,000,000,000,000,000
Exabyte = 10^{18} = 1,000,000,000,000,000,000

Esta contribución a la acumulación masiva de datos la podemos encontrar en diversas industrias, las compañías mantienen grandes cantidades de datos transaccionales, reuniendo información acerca de sus clientes, proveedores, operaciones, etc., de la misma manera sucede con el sector público. En muchos países se administran enormes bases de datos que contienen datos de censo de población, registros médicos, impuestos, etc., y si a todo esto le añadimos transacciones financieras realizadas en línea o por dispositivos móviles, análisis de redes sociales (en Twitter son cerca de 12 Terabytes de tweets creados diariamente y Facebook almacena alrededor de 100 Petabytes de fotos y videos), ubicación geográfica mediante coordenadas GPS, en otras palabras, todas aquellas actividades que la mayoría de nosotros realizamos varias veces al día con nuestros "smartphones", estamos hablando de que se generan alrededor de 2.5 quintillones de bytes diariamente en el mundo.

1 quintillón = 10^{30} = 1,000,000,000,000,000,000,000,000,000,000

Además del gran ***volumen*** de información, esta existe en una gran ***variedad*** de datos que pueden ser representados de diversas maneras en todo el mundo, por ejemplo de dispositivos móviles, audio, video, sistemas GPS, incontables sensores digitales en equipos industriales, automóviles, medidores eléctricos, veletas, anemómetros, etc., los cuales pueden medir y comunicar el posicionamiento, movimiento, vibración, temperatura, humedad y hasta los cambios químicos que sufre el aire, de tal forma que las aplicaciones que analizan estos datos requieren que la ***velocidad de respuesta*** sea lo demasiado rápida para lograr obtener la información correcta en el momento preciso. Estas son las características principales de una oportunidad para Big Data.

Es importante entender que las bases de datos convencionales son una parte importante y relevante para una solución analítica. De hecho, se vuelve mucho más vital cuando se usa en conjunto con la plataforma de Big Data. Pensemos en nuestras manos izquierda y derecha, cada una ofrece fortalezas individuales para cada tarea en específico. Por ejemplo, un beisbolista sabe que una de sus manos es mejor para lanzar la pelota y la otra para atraparla; puede ser que cada mano intente hacer la actividad de la otra, más sin embargo, el resultado no será el más óptimo.

12.2 Arquitectura BIG DATA

El gráfico que veremos a continuación es de un alcance muy amplio y en ciertos casos el lector querrá centrarse en algunos de sus aspectos que hacen al proyecto que pueda estar manejando en particular. Sin embargo, los expertos coinciden en que es muy importante comprender a todo el stack o conjunto si se quiere estar preparado para el futuro. En algún momento sin duda habrá que utilizar alguno de los elementos de este marco de trabajo para resolver algún problema.

Figura 12. Arquitectura de BIG DATA

Fuente: Big Data for Dummies, 2014.

En la capa cero de la figura están la infraestructura física compuesta por hardware, red y otros elementos. Es posible que una compañía ya tenga su centro de datos o haya realizado inversiones importantes en infraestructura física, por eso la mayoría contempla la posibilidad de aprovechar los activos existentes en un proyecto Big Data. Pero los proyectos Big Data tienen requerimientos muy específicos en relación a todos los elementos de la arquitectura de referencia y por eso será necesario examinar esas necesidades capa por capa si se quiere estar seguro de que la implementación funcionará y escalará según lo demande el negocio. En el momento mismo en que se comienza a trazar una estrategia para la implementación de Big Data, es importante establecer algunos principios generales a seguir y aplicar. Una lista de esos principios debería incluir definiciones acerca de los siguientes puntos:

- **Performance**. Cuál deberá ser la capacidad de respuesta que hace falta. Acá estamos hablando de latencia, un elemento que suele medirse a lo largo del sistema mediante una transacción o consulta. Los sistemas de alta performance y muy baja latencia, que son los más rápidos dicho en pocas palabras, requieren de una infraestructura que es bastante costosa.

- **Disponbilidad**. Grado de tiempo activo garantizado para un servicio adecuado. Establecer qué tanto tiempo puede esperar el negocio de la empresa sin sufrir pérdidas sensibles en el caso de que se produzca una falla o interrupción del servicio. Recordemos que las infraestructuras de alta disponibilidad son también muy costosas.
- **Escalabilidad**. Qué clase de infraestructura se necesitará, qué espacio en discos actualmente y en el futuro. Qué nivel de potencia de computación. Esos son algunos de los temas sobre los que habrá que decidir, estableciendo qué hace falta y prever algo más para escalar frente a problemas inesperados.
- **Flexibilidad**. Velocidad con los que se pueden agregar o reemplazar recursos en la infraestructura. Qué tan rápido se recuperará la infraestructura de sus fallos. Las infraestructuras más flexibles también son costosas. Aquí se puede decir que los servicios cloud pueden ayudar a controlar esos costos ya que se paga sólo por lo que se utiliza.
- **Costos**. Es bueno tener bien claro cuánto se puede gastar antes de comenzar el proyecto. Dado que esta infraestructura es un conjunto de componentes, uno puede decidir si comprará lo mejor en equipos de redes y ahorrará en almacenamiento, por ejemplo. Hará falta establecer los requerimientos en cada una de las áreas de infraestructura en el contexto de un presupuesto general donde luego habrá que hacer intercambios o sustituciones de ser necesarias.

Dado que Big Data tiene que ver con velocidad, alto volumen y gran variedad de datos, la infraestructura física será la que posibilite o imposibilite la implementación. La mayor parte de las implementaciones Big Data necesitan alta disponibilidad y por eso los servidores de redes y el almacenamiento físico deben ser resilentes o “resistentes” y redundantes, dos propiedades que se interrelacionan. Una infraestructura o un sistema son resilentes ante cambios o fallas cuando existen suficientes recursos redundantes y prestos a entrar en acción. En esencia, siempre habrá casos en los que hasta los más sofisticados y resilientes contextos puedan fallar debido a un mal funcionamiento de hardware. Es por eso que la redundancia será la que determine que un mal funcionamiento no cause una caída completa.

13. CONCLUSIONES

1. La minería de datos es un proceso de negocio, lo que requiere un amplio conocimiento del negocio. Es el más practicado por expertos en negocios o por expertos en minería de datos, en estrecha colaboración con expertos en negocios. La minería de datos utiliza una variedad de técnicas y no debe centrarse sólo en los algoritmos de modelado y su exactitud predictiva. Cada técnica puede desempeñar una variedad de papeles.
2. Durante el proceso de minería de datos, minería de datos interactúa y se involucra con los datos en forma iterativa. Un modelo de proceso de minería de datos estándar, como CRISP-DM o el SEMMA contribuye a garantizar la correcta preparación y uso de minería de datos. Los analistas de datos deben tomar decisiones inteligentes acerca de la cantidad de datos necesarios, asumiendo que no todos los datos de una organización serán relevantes ni que se requerirá que todos los datos disponibles.
3. Herramientas de minería de datos deben ser evaluados en función de su accesibilidad para los usuarios de negocios, su escalabilidad y facilidad de uso, y su apoyo a los procesos estándar. La Minería de datos eficaz requiere técnicas flexibles e interoperables. Este requisito es mejor recibido por, juegos de herramientas integrados y abiertos que pueden interactuar con los datos a través de estándares abiertos.
4. A medida que los volúmenes de información que almacenamos crece, mayor serán los requerimientos tecnológicos para analizar la data, dejando paso a técnicas que desarrolla el BIG DATA.

DATA MINING

BIG DATA

14. REFERENCIAS BIBLIOGRÁFICAS

1. **Camargo Hernando, Mario Silva (2011).**
   Dos caminos en la búsqueda de patrones por medio de Minería de Datos: SEMMA y CRISP.
   Rev. Tecnol. – Journal of Technology • Volumen 9 No. 1
   Descargado de: http://www.uelbosque.edu.co/sites/default/files/publicaciones/revistas/revista_tecnologia/volumen9_numero1/dos_caminos9-1.pdf

2. **Fayyad Usama, Piatetsky-Shapiro Gregory, Smyth Padhraic(1996).**
   The KDD process for extracting useful knowledge from volumes of data
   COMMUNICATIONS OF THE ACM November 1996/Vol. 39, No. 11 Pages 27-34.
   Descargado de: http://cacm.acm.org/magazines/1996/11/

3. **Gartner (2014).**
   Magic Quadrant for Advanced Analytics Platforms.
   Descargado de: http://www.gartner.com/technology/reprints.do?id=1-1QXWEQQ&ct=140219&st=sg

4. **Ibermática (2007).**
   Business Intelligence
   Descargado de: http://www.ibermatica.com/soluciones/bi

5. **IBM CORPORATION (2012)**
   ¿Qué es BIGDATA?
   Descargado de: http://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

6. **IBM CORPORATION (2014)**
   Manual CRISP-DM de IBM SPSS Modeler
   Descargado de: ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf

7. **Jones Don (2010)**
   The Shortcut Guide to Achieving Business Intelligence in Midsize Companies
   Realtime Publishers. San Francisco EEUU.
   Descargado de: http://www.lpa.com/files/8013/5596/9896/eBook_The_Shortcut_Guide_to_Achieving_Business_Intelligence_in_Mid_Sized_Companies.pdf

8. **Khabaza Tom (2005)**
   Hard Hats for Data Miners: Myths and Pitfalls of Data Mining SPSS
   Descargado de: http://www.spss.ch/upload/1113911541_data_mining_khabaza%20(3).pdf

9. **Microstrategy (2011).**
   Architecture for Enterprise Business Intelligence: an overview of the microstrategy platform architecture for Big Data, Cloud BI, and Mobile Applications.
   Descargado de: http://www.microstrategy.com/Strategy/media/downloads/white-papers/MicroStrategy-Architecture-for-Enterprise-BI.pdf

10. **Oviedo Blanco César (2011)**
    Rompiendo el Mito: Minería de Datos Una Perspectiva Latinoamericana
    Descargado de: http://businessintelligencelatam.com/wp-content/uploads/2013/01/Rompiendo-el-Mito-Data-Mining.pdf

11. **Piatetsky Gregory (2014).**
    KDnuggets 15th Annual Analytics, Data Mining, Data Science Software Poll
    Descargado de: http://www.kdnuggets.com/2014/06/kdnuggets-annual-software-poll-rapidminer-continues-lead.html

12. **Turban Efraim, et al. (2011).**
    Business Intelligence: A managerial approach. Editorial Pearson. SegundaEdición.

13. **Withee Ken (2010)**
    Microsoft® Business Intelligence for dummies. Editorial Wiley. PrimeraEdición.