

MINERÍA DE DATOS, DE TEXTOS Y SENTIMIENTOS



Por: Ana María Polo Ahumada
Estudiante Maestría en Ingeniería Administrativa
Instituto Tecnológico de Orizaba
Fundamentos de Ingeniería Administrativa

Tabla de contenido

Introducción.....	3
¿Qué es minería de datos?	4
Historia de la minería de datos.....	4
Business Intelligence	5
Técnicas de minería de datos	8
¿Cómo se crea un modelo de minería de datos?	9
Definir el problema	10
Preparar los datos	11
Explorar los datos.....	12
Generar modelos.....	12
Explorar y validar los modelos.....	13
Implementar y actualizar los modelos.....	13
Aplicación del data mining	14
¿Qué es minería de textos?.....	15
¿Cómo se hace la minería de textos?	16
Aplicación de la minería de textos	16
¿Que es minería de sentimientos u opiniones?	17
Conclusión	18
Bibliografía.....	18
Agradecimientos	18

Introducción

La minería de datos podría ser definida como un proceso de descubrimiento de nuevas y significativas relaciones, tendencias y patrones al explorar grandes cantidades de datos. El tener a disposición una gran cantidad de información así como el uso de diversas herramientas informáticas ha dirigido el análisis de datos a la aplicación de diferentes técnicas especializadas enmarcadas en lo que se conoce como Data mining o minería de datos.

Las técnicas utilizadas en la minería de datos tienen como objetivo descubrir de manera automática el conocimiento que se encuentre almacenado de forma ordenada en la información contenido en una gran base de datos. El fin principal es encontrar patrones, perfiles y tendencias mediante el análisis de datos usando tecnologías que permitan reconocer patrones, redes neuronales, lógica difusa, algoritmos genéticos y muchas otras técnicas avanzadas del análisis de datos.

Hoy en día la minería de datos es utilizada en muchos campos de la ciencia, a nivel financiero y bancario, en el análisis de mercados y comercios, en el área de la salud tanto pública como privada, a nivel educativo, en procesos industriales, en medicina, biología y bioingeniería así en las telecomunicaciones y diferentes áreas. (Perez Lopez & Santín González, 2007)

En este escrito revisará el concepto de la minería de datos, su aplicación e importancia para la toma de decisiones en las organizaciones

¿Qué es minería de datos?

Se entiende como minería de datos a un grupo de técnicas utilizadas para extraer y obtener información válida, pormenorizada y de mucha utilidad que se encuentra en las diferentes bases de datos. Es muy útil para la toma de decisiones ya que contribuyen a predecir tendencias y comportamientos a futuro por lo cual es una herramienta poderosa para las organizaciones.

La minería de datos o Data mining como es conocida en inglés, relaciona los procedimientos utilizados en la industria minera para extraer los minerales de la tierra realizando explosiones que hagan que suba a la superficie. Siguiendo esta misma técnica, la minería de datos realiza explosiones en las bases de datos con el fin de extraer información oculta en ellas.

Mediante el uso de varias herramientas y técnicas algorítmicas, la minería de datos busca patrones de interés ocultos en las bases de datos con el fin de anticipar futuros y pronosticar situaciones con cierto grado de probabilidad. De esta manera se puede encontrar información predecible que cualquier persona por muy experta que sea no puede encontrar fácilmente. La minería de datos puede ser implementada en cualquier plataforma de hardware y software y puede ser integrada a sistemas de información on-line. (Reinosa & Maldonado, 2012)

Historia de la minería de datos

El data mining no es algo nuevo, se viene dando desde los años sesenta cuando en esos momentos los estadísticos utilizaban los términos data fishing, data mining o data archaeology; más tarde en los años ochenta se empezó a hablar de los del término KDD que por sus siglas en inglés significa proceso de extracción de conocimiento a partir de datos del cual la minería de datos forma parte. A partir de ese año se fueron creando varias empresas dedicadas a prestar servicios relacionados con la minería de datos y poco a poco fueron surgiendo muchas otras; hasta el momento existen más de cien empresas dedicadas a la minería de datos. (Felix, 2002)

Es conocida una historia de éxito de la implementación de la minería de datos en 1992 de un empleado de la empresa NCR corporation que realizó un estudio para Osco Drugs de American Stores. Como resultado de este estudio se observó que durante el horario de 5 a 7 de la noche se producía la compra con más frecuencia de dos artículos que no estaban relacionados entre sí pero que eran llevados al mismo tiempo: los pañales y las cervezas. Esto concluyó entonces que muchos padres de familia que eran enviados a comprar pañales a esa hora terminaban llevando también unas latas de cervezas, situación que fue aprovechada por la tienda para colocar el refrigerador con las cervezas cerca al exhibidor de pañales para potencializar dicha compra bien sea por decisión propia o por impulso. Esto es una muestra de los resultados inesperados que se pueden encontrar al utilizar la minería de datos y de las decisiones que puede tomar la organización basada en dichos descubrimientos. Es muy importante que la organización sea ágil para utilizar los resultados obtenidos, por lo cual, la minería de datos por sí sola no es útil, a menos que se aproveche como debe ser. (Reinosa & Maldonado, 2012)

Business Intelligence

La minería de datos tiene su origen en los sistemas de información cuya finalidad era recopilar información sobre un tema determinado para tomar decisiones. Al surgir nuevos softwares y hardwares, las organizaciones se informatizaron y los sistemas de información pasaron a dar soporte a los procesos básicos de la empresa tales como en ventas, producción, recursos humanos y demás a los cuales se les llama Sistemas de información para la gestión. Con el tiempo y tras la necesidad de las empresas de tener una base que les ayudase a la toma de decisiones surgieron herramientas que suplieran estas necesidades llamadas DSS (Decision support system) tales como las EIS y OLAP así como las diferentes técnicas herramientas de la minería de datos.

Las EIS (Executive information systems) son un conjunto de herramientas y sistemas de información que le permiten a los ejecutivos de las empresas tener

acceso al estado de las actividades y su gestión. Permiten informar de manera inmediata cualquier cambio que se presente en la empresa, para ello analiza el estado diario de la organización mediante indicadores claves. El tipo de información que regularmente se pide suelen ser las ventas semanales, balances parciales y el nivel de stocks y a la vez es representada mediante gráficos en hojas de cálculo. (Perez Lopez & Santín González, 2007)

Las OLAP (On-line analytical processing) brindan facilidad para manejar y transformar los datos para producir nuevos datos. El objetivo de las OLAP es agilizar la consulta de grandes cantidades de datos.

Las herramientas de la minería de datos tienen como objetivo extraer patrones y tendencias con el fin de predecir comportamientos futuros. La minería de datos analiza los datos mientras que las OLAP e EIS facilitan el acceso a la información para que se puede hacer un análisis más efectivo, lo cual quiere decir que sirven de apoyo a la minería de datos.

El uso de cada herramienta dependerá del objetivo de la organización, para ello se debe partir de una cuestión básica, tal y como podemos ver en la siguiente tabla:¹ (Braga, Valencia, & Carvajal, 2009)

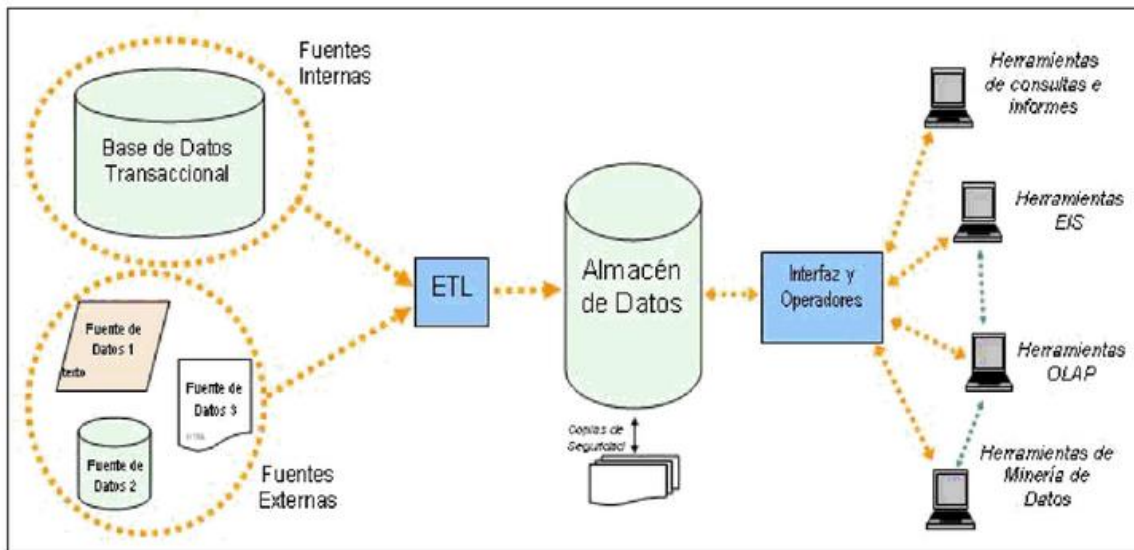
Tipo de Herramienta	Cuestión básica	Ejemplo de respuesta
Data Mining	¿Qué es interesante? ¿Qué puede suceder?	Tipos de clientes Predicción de ventas
OLAP	¿Qué sucedió y por qué?	Ventas mensuales versus variaciones de precios de los competidores
EIS/DSS	¿Qué necesito saber ahora ?	Cotizaciones diversas
Estudios y informes	¿Qué sucedió ?	Ventas del último mes

Para que puedan funcionar los sistemas arriba mencionados, es necesario que exista un almacén de datos o *Warehouse* el cual es una colección de datos

¹ Imagen tomada del libro Introducción a la minería de datos Luis Pablo Vieira Braga y otros. 2009

históricos internos o externos, que describen un contexto o área de estudio orientada hacia un dominio que permiten aplicar herramientas con el fin de describir, resumir y analizar los datos para ayudar en la toma de decisiones.

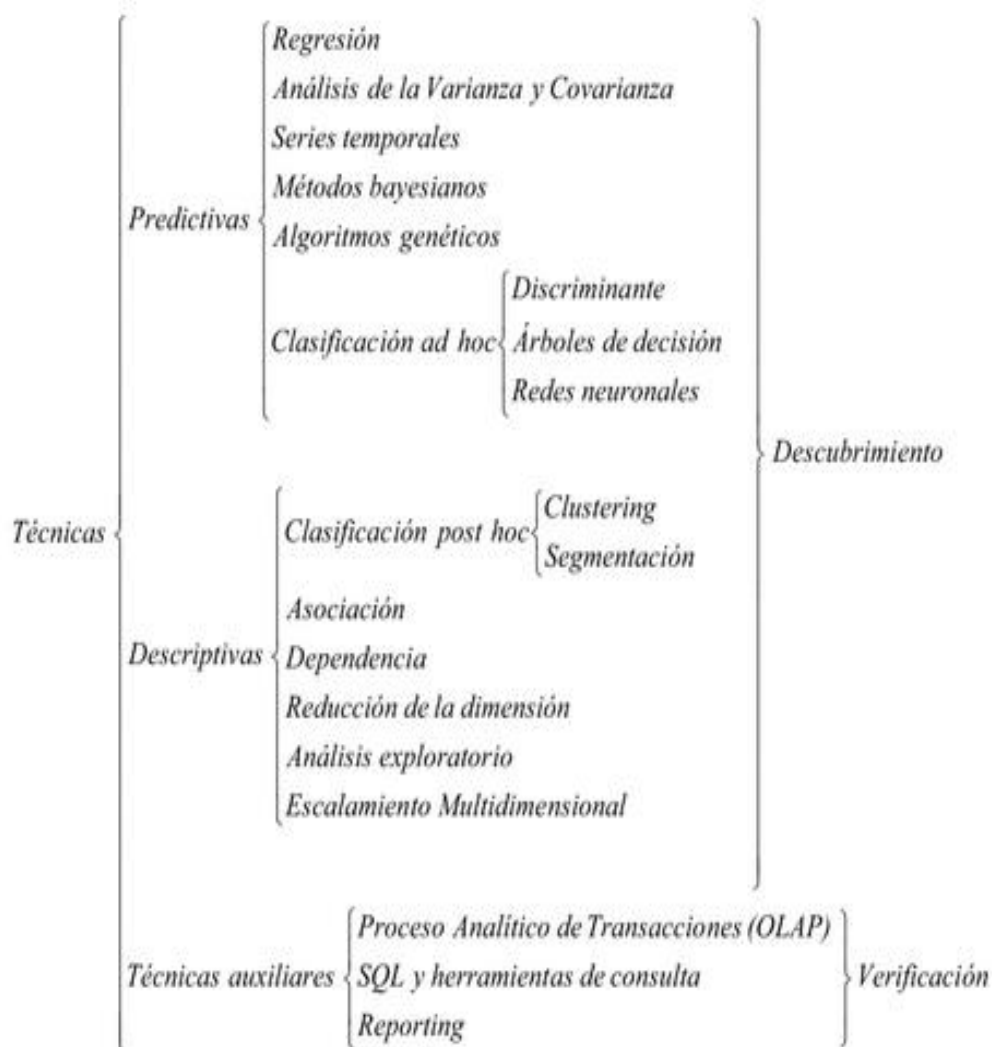
Para cargar o alimentar los datos se utiliza un sistema llamado ETL (Extraction, transformation, Load) que se encarga de la lectura de los datos , incorporación de nuevos datos, creación de claves etc. En la siguiente imagen se explica cómo funcionan estos sistemas ²



² Imagen tomada del libro Minería de datos: técnicas y herramientas de César Perez López y Daniel Santín González

Técnicas de minería de datos

Las técnicas de la minería de datos se clasifican en predictivas, descriptivas y auxiliares y se organizan como aparecen en la siguiente imagen³



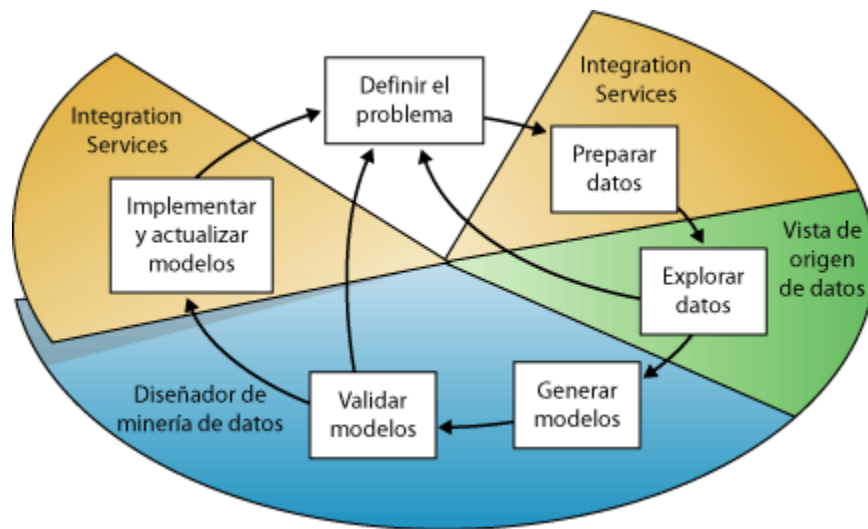
³ Imagen tomada del libro Minería de datos: técnicas y herramientas de César Perez López y Daniel Santín González

¿Cómo se crea un modelo de minería de datos?

Para aplicar la minería de datos se pueden seguir los siguientes seis pasos:

- Definición del problema
- Preparación de los datos
- Exploración de los datos
- Generación de los modelos
- Exploración y validación los modelos
- Implementación y actualización los modelos

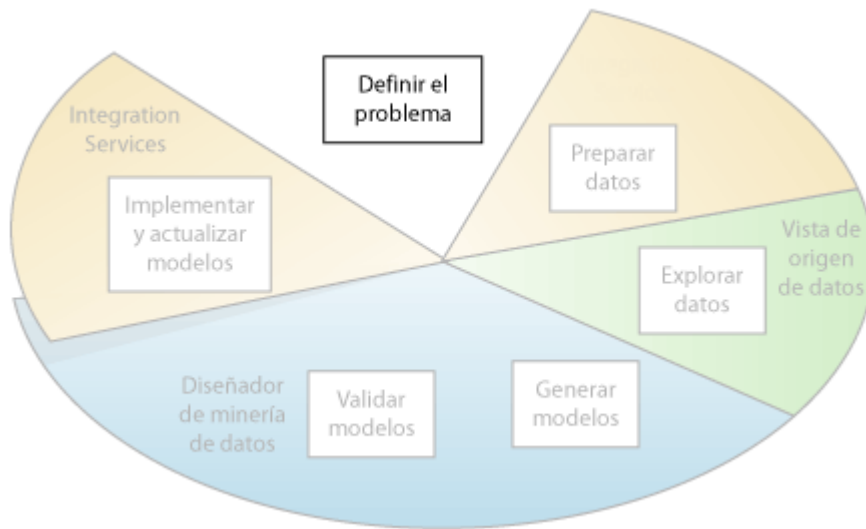
En la siguiente figura se puede observar estos pasos



Como se puede observar este es un proceso cíclico lo cual quiere decir que si los datos encontrados no son suficientes para la creación del modelo o los modelos no son los adecuados para los fines propuesto. Se deben entonces repetir los mismos pasos para crear un nuevo modelo.

Definir el problema

Lo primero que se debe hacer para crear un modelo de minería es definir el problema y considerar la forma que se pueden usar los datos para dar solución al mismo.



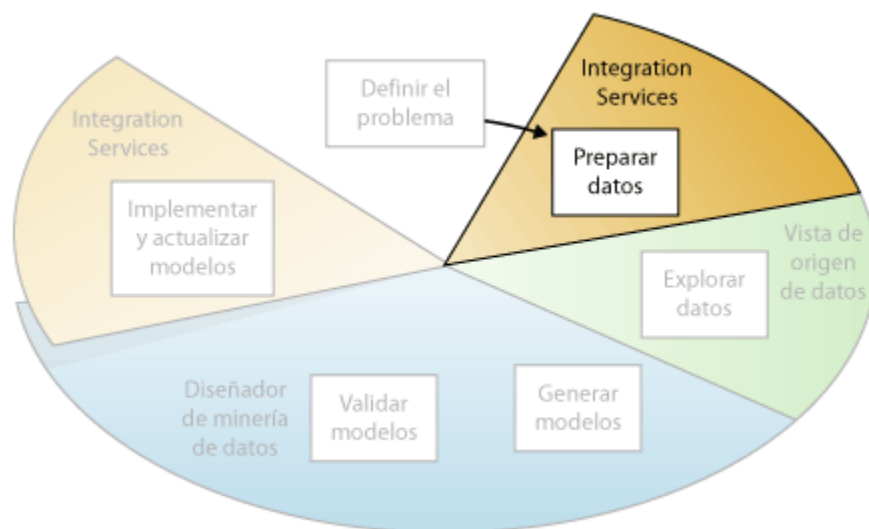
En este punto se analizan los requisitos empresariales, se define el ámbito del problema, la forma en que se evaluará el modelo así como establecer los objetivos concretos del proyecto de la minería de datos. Para ello se pueden hacer las siguientes preguntas:

- ¿Qué se busca? ¿Qué tipo de relaciones se intenta buscar?
- ¿El problema refleja resolver el problema que las directivas están intentando solucionar?
- ¿Que desea hacer a partir del modelo de la minería de datos?
¿Predicciones, buscar patrones interesantes o asociaciones?
- ¿Qué resultado se desea predecir?

- ¿Qué datos se tienen y que tipo de información hay en cada columna? Si hay tablas, ¿Cómo están relacionadas? ¿Se necesita limpiar, agregar o procesar los datos antes de ser utilizados?
- ¿Cómo están distribuidos los datos? ¿Son estacionales? ¿Representan de forma precisa los procesos de la empresa?

Preparar los datos

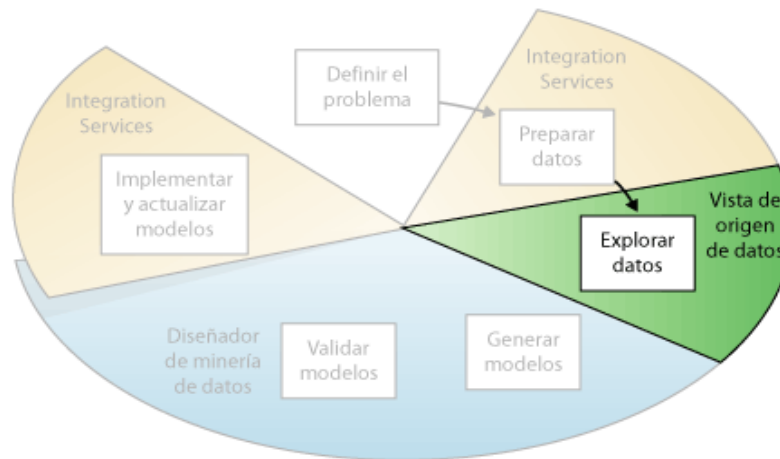
El siguiente paso consiste en consolidar y limpiar los datos identificados en el paso anterior. Estos datos pueden tener incoherencias o estar muy dispersos como por ejemplo que un cliente compró un producto antes de que saliera al mercado o que compra en una tienda ubicada a 20000 Km de su casa.



Esta limpieza no es únicamente quitar los datos que no sean válidos sino también buscar correlaciones que se ocultan en los datos, identificar el origen de los datos que son más preciso y determinar que columnas son las más adecuadas para el análisis.

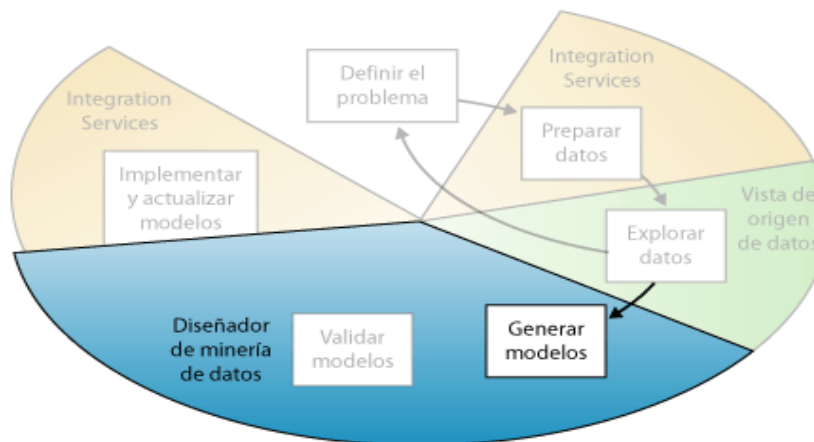
Explorar los datos

Se deben conocer los datos con el fin de tomar las mejorara decision a la hora de crear modelos de minería de datos, para ello se deben usar técnicas de exploración tales como calcular los valores mínimos y máximos, calcular la media y las desviaciones estándar y examinar la distribución de los datos.



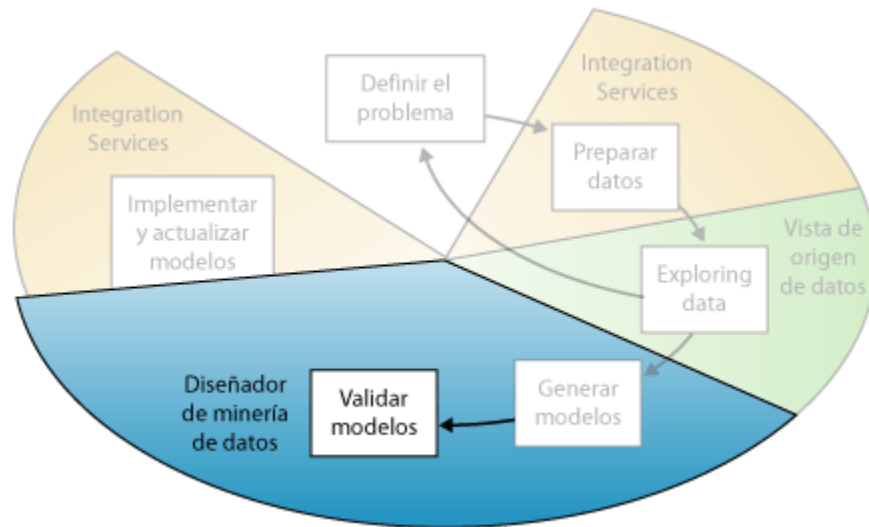
Generar modelos

En el cuarto paso de la minería de datos lo que se realiza es generar el modelo utilizando los conocimientos que se adquirieron en la exploración de los datos, para ello se deben definir qué columnas de datos se van a utilizar con el fin de crear una estructura de minería de datos.



Explorar y validar los modelos

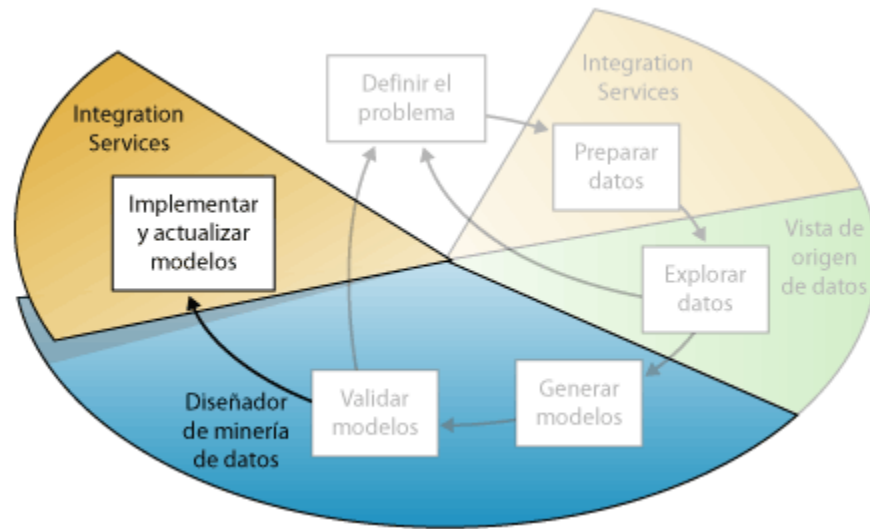
El siguiente paso del proceso de minería de datos es explorar los modelos obtenidos anteriormente y comprobar que sean eficaces previos a su implementación. Al probar los modelos se puede ver cual ofrece mejores resultados para el problema planteado inicialmente.



Si no funciona ninguno de los modelos que se han creado, se debe volver a los pasos anteriores para ya sea replantear el problema o volver a investigar los datos del conjunto original

Implementar y actualizar los modelos

Por último se debe implementar los modelos que funcionen mejor en el entorno de producción los cuales pueden llegar a realizar diferentes tareas de acuerdo a las necesidades de la empresa.



Entre las tareas que puede realizar el modelo son:

- Para hacer predicciones que se podrán utilizar luego para tomar decisiones comerciales
- Crear consultas de los contenidos para recuperar reglas, fórmulas y estadísticas del modelo.
- Incrustar la funcionalidad del modelo en una aplicación (Microsoft, 2014)

Aplicación del data mining

Actualmente el data mining puede ser utilizado en diversos campos dentro de los que se encuentran:

- Hacer análisis financieros: se aplica en el sector bancario o financiero y lo que se busca es proveer datos con los que sea posible hacer análisis sistemáticos confiables. Con ello se puede predecir el pago de préstamos, analizar las políticas de crédito de los clientes, clasificar y agrupar clientes para crear ofertas especializadas y detectar posibles fraudes y delitos financieros.
- En el sector retail : las tiendas dedicadas a este tipo de actividades recogen diariamente mucha información provenientes de las ventas, el historial de

compras y el transporte de las mercancías. Con estos datos se pueden hacer predicciones que le permitan a las tiendas ofrecer un mejor servicio y facilitar su retención. La minería de datos en estos casos puede hacer :

Análisis de las ventas, los clientes los productos, el tiempo y la región

Analizar la eficacia de las campañas de ventas

Recomendar productos de forma personalizada

- En las telecomunicaciones: en este sector se puede usar la minería de datos para identificar patrones de telecomunicaciones, ayuda a facilitar la detección de actividades fraudulentas y hacen posible uso mejor los recursos haciendo que de esta forma mejora la calidad del servicio. (Lantares, 2014)

¿Qué es minería de textos?

La minería de textos es la localización, análisis y organización de la información con el fin de crear una nueva información que no se puede de forma evidente al revisar los documentos. Es nueva información obtenido puede ser un patrón, una tendencia o una correlación que no se puede identificar sólo leyendo los documentos los cuales pueden ser páginas de internet, correos electrónicos, un campo en las bases de datos o un archivo de texto sin ningún formato.

El text mining o minería de textos comprende tres actividades fundamentales las cuales son:

- Recuperar la información : seleccionar los textos adecuados
- Extraer la información contenida en esos textos: datos claves, hechos y acontecimientos
- Utilizar la minería de datos para encontrar asociaciones entres esos textos claves (galeon.com, 2016)

¿Cómo se hace la minería de textos?

Para implementar la minería de datos se pueden seguir las siguientes cuatro etapas:

Primera etapa: Se establecen los objetivos con el fin de aclarar lo que se busca en la investigación y el fin de establecer los límites y acotar hasta que punto se desea profundizar.

Segunda etapa: Procesar los datos mediante la selección, el análisis y reducción de los textos o documentos de los cuales se extraerá la información. Esta es la etapa que consume la mayor cantidad de tiempo

Tercera etapa: Especificar qué modelo o técnica se va a utilizar, esto va a depender de los objetivos planteados y de las tareas que se van a realizar

Cuarta etapa: Se analizan los resultados con el fin de utilizar la información encontrada para tomar las decisiones que mejor se adapten a la organización. (galeon.com, 2016)

Aplicación de la minería de textos

- Para extraer información: puede ser usada para extraer información de grandes cantidades de texto que se encuentren en la web permitiendo de esta forma definir entidades y sus relaciones, revelar información significativa y facilitar la comprensión de los datos.
- Clasificar documentos: permite recuperar y navegar en documentos, especialmente en empresas que llevan un registro histórico de sus actividades y proyectos en documentos. Para ello se aplican algoritmos de minería de textos que agrupo los documentos y obtiene información descriptiva de cada uno de cada grupo para poder hacer una mayor comprensión de los mismos.

- **Elaboración de resúmenes:** se puede obtener una descripción de manera general de un conjunto de documentos con respecto a un tema específico. En este sentido estos métodos pueden ser clasificados en dos categorías la **sumarización extractiva** y la **sumarización abstracta**.
- **Extracción de conocimiento:** utilizando la minería de textos es posible realizar modelos de conocimiento a partir de la información extraída de los documentos.

¿Qué es minería de sentimientos u opiniones?

La minería de opiniones o de sentimientos puede definirse como la aplicación de una serie de técnicas del procesamiento del lenguaje natural, lingüística computacional y minería de textos cuyo objetivo es extraer información subjetiva a partir de información que publiquen las personas ya sea en blogs o reviews de productos por internet. A partir de este análisis se puede obtener información importante ya sea positiva o negativa.

Al hacer minería de opiniones o sentimientos se aplica la minería de textos y se puede hacer de dos formas:

- **Detección de la polaridad:** tiene como objetivo establecer si una opinión es negativa o positiva y a la vez tratar de obtener un valor numérico dentro de un rango establecido para obtener un rating asociado a una opinión determinada.
- **Análisis del sentimiento basado en características:** su objetivo es ser capaces de establecer cuáles son las características de un producto basándose en el review u opinión de los usuarios y con cada una de esas características obtener una polaridad. (Brainsins, 2015)

Conclusión

La minería de datos, de textos y de sentimientos brinda herramientas muy útiles para el análisis de los datos y textos que a la vez permiten identificar patrones de comportamiento que ayuden a la toma de decisiones. Son muchas las utilidades que se le pueden dar tanto a la minería de datos, de textos y de sentimientos, pero está en cada organización establecer qué tipo de técnica va a utilizar basada en el planteamiento inicial de un problema.

Bibliografía

- Braga, L. P., Valencia, L. I., & Carvajal, S. S. (2009). *Introducción a la minería de datos*. Sao Pablo: Sindicato nacional de editores.
- Brainsins. (2015). Obtenido de <http://www.brainsins.com/es/blog/mineria-opiniones/3555>
- Cesar Perez Lopez, D. S. (2007). *Minería de datos : técnicas y herramientas*. Madrid: International Thompson Ediciones Paraninfo s.a.
- Enrique Jose Reinoso, C. A. (2012). *Base de datos*. Mexico: Allfaomega.
- galeon.com. (02 de 04 de 2016). *Galeon.com*. Obtenido de <http://textmining.galeon.com/>
- Lantares. (2014). Obtenido de <http://www.lantares.com/blog/mineria-de-datos-aplicaciones-que-ya-son-una-realidad>
- Microsoft. (2014). Obtenido de <https://msdn.microsoft.com/es-mx/library/ms174949%28v=sql.120%29.aspx>
- Perez Lopez, C., & Santín González, D. (2007). *Minería de datos: Técnicas y Herramientas*. Madrid: Internacional Thomson Ediciones Paraninfo.

Agradecimientos

Al Instituto Tecnológico de Orizaba por darme la oportunidad de formarme profesionalmente y al profesor Fernando Aguirre y Hernández por todos los conocimientos que nos ha compartido en su materia Fundamentos de Ingeniería

Administrativa para aprender y perfeccionar mis capacidades de redactar artículos científicos de calidad.