

1

Instituto Politécnico Nacional  
Escuela Superior de Cómputo  
Secretaría Académica  
Departamento de Ingeniería en Sistemas Computacionales

## Minería de datos (*Data Mining*) Introducción

Profesora: Dra. Fabiola Ocampo Botello

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

2

## Introducción a la Minería de datos

### Los datos



#### Producto

Resultado  
histórico de los  
sistemas de  
información



#### Materia prima

Hay que  
explotar para  
un "producto  
elaborado"



#### Conocimiento

Toma de  
decisiones en  
contexto de  
los datos

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

## 3

- Información histórica que proviene de diversas fuentes de datos.
- Crecimiento importante de la cantidad de información.
- Son la memoria de la organización, la historia para entender el pasado, el presente y predecir el futuro.

*Resolver problemas actuales considerando los datos históricos es un aspecto importante de la minería de datos.*

*La minería de datos obtiene información descriptiva e inferencial útil para producir conocimiento.*



Fuente: Figura Creative Commons. En: <https://centro-documentacion-europea-ufv.eu/datos-abiertos-big-data/>

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

## 4

## ¿Qué es la Minería de datos?



Fuente: Figura Creative Commons. En: <http://fraterneo.blogspot.com/2010/11/5-programas-libres-para-data-mining.html>

Se define la minería de datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos (Witten & Frank 2000 citado en Hernández, Ramírez y Ferri, 2004).

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

5

La minería de datos es un campo multidisciplinario que integra trabajo de diversas áreas como tecnología de bases de datos, aprendizaje automático (*machine learning*), estadística, reconocimiento de patrones, recuperación de información, redes neuronales, sistemas basados en conocimiento, inteligencia artificial, cómputo de alto rendimiento y visualización de datos (Sahu, Shrma, & Gondhalakar, 2011).

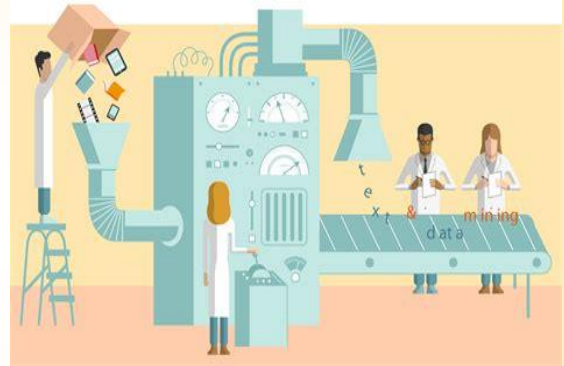


Figura Creative Commons. Tomada de: <https://ruedimumenthaler.ch/2015/06/09/trend-und-herausforderung-text-and-data-mining/>

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

6

La minería de datos es un proceso para extraer información y conocimiento implícitos, potencialmente útiles y que son desconocidos por las personas, los cuales se encuentran en datos masivos, incompletos, difusos y aleatorios (Sahu, Shrma, & Gondhalakar, 2011).

La minería de datos ha sido comúnmente definida como encontrar información en una base de datos, ha sido llamada análisis de datos exploratorio, descubrimiento conducido por datos y aprendizaje deductivo (Dunham, M. H., 2002).



Figura Creative Commons. Tomada de: [https://commons.wikimedia.org/wiki/File:Data\\_Mining.svg](https://commons.wikimedia.org/wiki/File:Data_Mining.svg)

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

7

**Ejercicio de clase:** Escriba el objetivo de cada uno de los siguientes tipos de problemas de la minería de datos.

Área	Objetivo
Análisis de créditos bancarios	
Análisis de la cesta de la compra	
Determinar las ventas de un producto	
Determinar grupos diferenciados de empleados	

Ejemplos tomados de: Hernández, Ramírez y Ferri, 2004:6-9).

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

8



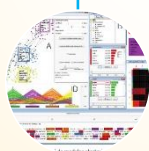
## Minería de datos

### Retos

que enfrenta



Diversos tipos de datos con ruido, ausentes, sucios, etc.  
Se desconocen los orígenes de los datos.

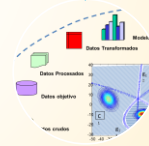


Conocimiento de las técnicas de minería de datos a aplicar para extraer información útil.

Conocimiento de la naturaleza de los datos.



Extraer información, identificar relación entre variables, modelos de datos, patrones de comportamiento



Formas de evaluar los resultados y expresión gráfica de los resultados

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello



9

## Tipos de fuentes de datos

### Otras bases de datos

### Bases de datos relacionales

Desde el punto de vista de las técnicas de minería de datos más habituales nos interesa distinguir sólo entre dos tipos, numéricos y categóricos.

**Bases de datos temporales.** Almacena datos que pueden utilizarse para encontrar las características de la evolución o las tendencias del cambio de distintas medidas o valores de la base de datos.

**Bases de datos espaciales.** Información relacionada con el espacio físico. Datos geográficos, imágenes, información de tráfico.

**Bases de datos documentales.** Almacena descriptores de textos que van desde palabras clave a resúmenes.

**Bases de datos multimedia.** Almacena imágenes, audio y video.

Imagen formada considerando lo expuesto en Hernández, Ramírez y Ferri (2004).

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

## Características de la minería de datos



Surge como una tecnología para apoyar a comprender el contenido de una base de datos.



Es una etapa de un proceso llamado extracción de conocimiento en bases de datos (Knowledge Discovery in Databases, KDD).



Mediante la minería de datos se detectan relaciones entre los datos que no han sido identificadas, lo que permite conocer relaciones con sentido, patrones de comportamiento, secuencias, predicciones, agrupamiento que serán analizados para la toma de decisiones.



La minería de datos generalmente implica el análisis de los datos almacenados en un almacén de datos (datawarehouse). Tres de las principales técnicas de la minería de datos son: la regresión, la clasificación y el agrupamiento (Sahu, Shorma, & Gondhalakar, 2011).

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

11

## Extracción de conocimiento en bases de datos (Knowledge Discovery in Databases, KDD)

KDD es el proceso de extraer información y la minería de datos es el uso de algoritmos para extraer esa información y forma parte del KDD.

KDD (Knowledge Discovery in Databases) Process

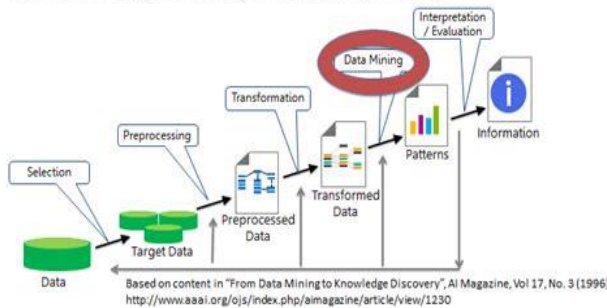
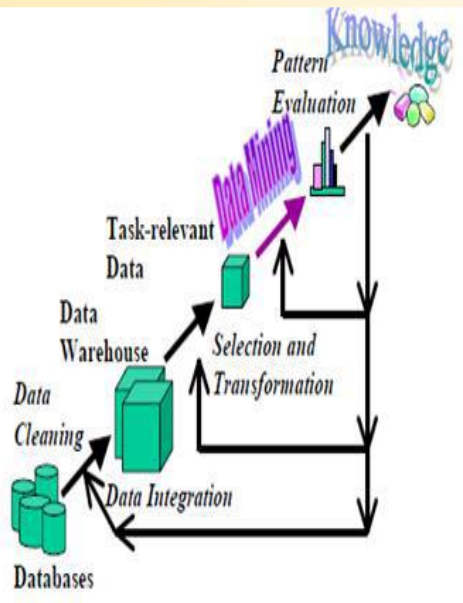


Figura Creative Commons. Tomada de:

<https://www.actuaries.digital/2016/07/28/dat203x-data-science-and-machine-learning/>

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

Es el proceso no trivial de descubrir conocimiento e información potencialmente útil dentro de los datos contenidos en algún repositorio de información. No es un proceso automático, es un proceso iterativo que, exhaustivamente explora volúmenes muy grandes de datos para determinar relaciones (U Fayyad et al 1996, citado en Joyanes, 2019).



El proceso iterativo consiste de los siguientes pasos:

- **Limpieza de datos.** Se eliminan los datos con ruido o sucios y los datos irrelevantes que se encuentran en la colección de datos.
- **Integración de datos.** Se integran múltiples fuentes de datos comúnmente heterogéneas en una fuente de datos en común.
- **Selección de datos.** Se seleccionan de la fuente de datos, los datos relevantes para el análisis.
- **Transformación de datos.** También conocida como consolidación de los datos. Los datos seleccionados se transforman a formas apropiadas para realizar el proceso de la minería.
- **Minería de datos.** Es el paso crucial en el cual se aplican técnicas inteligentes para la extracción de patrones de relación potencialmente útiles.
- **Evaluación de los patrones.** Se identifican patrones estrictamente interesantes que representen conocimiento sustentados en las medidas proporcionadas.
- **Representación del conocimiento.** El conocimiento descubierto es presentado al usuario de manera visual, las técnicas de representación visual permite a los usuarios interpretar y entender los resultados de la minería de datos.

Imagen tomada de Sahu, Shirma, & Gondhalakar (2011).

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

13

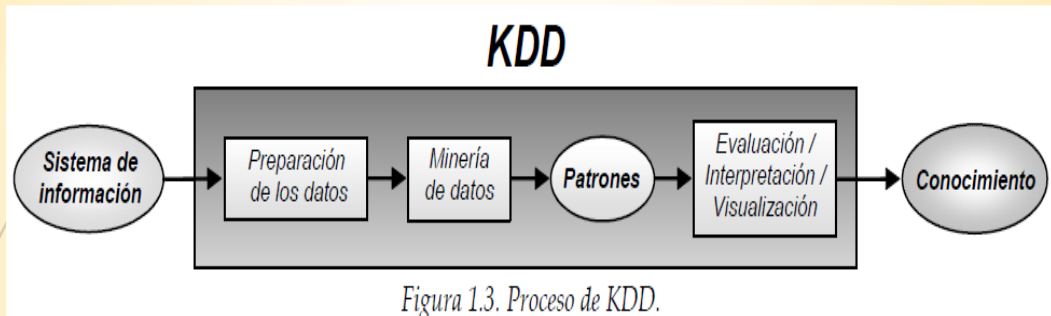


Imagen tomada de Hernández, Ramírez y Ferri (2004)

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

### Proceso de descubrimiento de conocimiento (KDD) y la minería de datos.

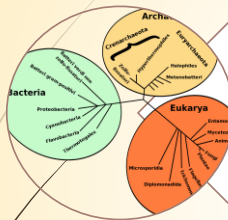
14

Paso	Proceso de KDD (Fayyad, 1996, citado en Joyanes, 2019).	Proceso de minería de datos Joyanes (2019)
1	<u>Selección de datos.</u> - Fuentes de datos. - Tipo de información.	<u>Selección del conjunto de datos.</u> - Variables objetivo, variables independientes y selección de registros.
2	<u>Preprocesamiento.</u> - Preparación y limpieza de los datos. - Datos faltantes, en blanco, atípicos, etc.	<u>Análisis de las propiedades de los datos.</u> - Ausencia de datos, datos atípicos, muestras gráficas.
3	<u>Transformación.</u> - Generación de nuevos datos. - Normalización.	<u>Transformación o preprocesamiento del conjunto de datos de entrada.</u> - Técnicas de tratamiento de los datos ausentes, atípicos o dudosos. - Elección de tipo.
4	<u>Minería de datos.</u> - Métodos para extraer patrones, relaciones, datos ocultos.	<u>Selección y aplicación de técnicas de minería de datos.</u> - Elección del modelo a utilizar.
5	<u>Interpretación y evaluación.</u> - Evaluación de resultados.	<u>Extracción de conocimiento.</u> - Identificación de patrones de comportamiento.
6		<u>Interpretación y evaluación de datos.</u> - Validación del modelo.

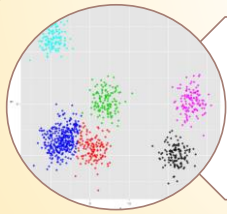
Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

15

Maimon & Rokach (2010) presentan la siguiente clasificación de los métodos de aprendizaje automático (*machine learning*):



Aprendizaje supervisado. Los métodos de aprendizaje supervisado son métodos que intentan descubrir la relación entre los atributos de entrada (variables independientes) y un atributo de referencia o fuente (variable dependiente). La relación descubierta se representa como un modelo.



Aprendizaje no supervisado. Se refiere principalmente técnicas que agrupan instancias sin un atributo dependiente pre especificado.

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

16

En la práctica, los modelos pueden ser de dos tipos : descriptivos o predictivos (Hernández, Ramírez y Ferri, 2004):

**Predictivos.** Los modelos predictivos pretenden estimar valores futuros o desconocidos de variables de interés, que se denominan *variables objetivo o dependientes*, usando otras variables o campos de la base de datos, a las que se refiere como *variables independientes o predictivas*. Por ejemplo, un modelo predictivo sería aquel que permite estimar la demanda de un nuevo producto en función del gasto en publicidad.

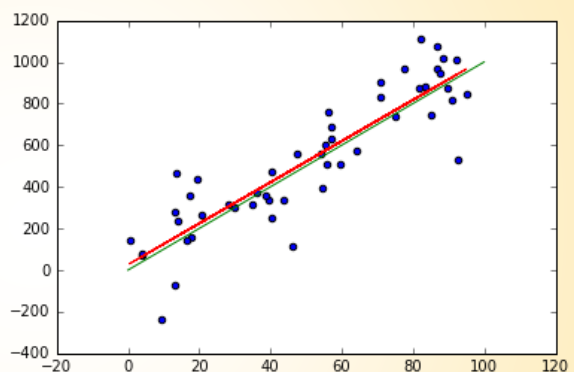


Figura Creative Commons. Tomada de: <https://machinelearningparatodos.com/regresion-lineal-en-python/>

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello



17

**Descriptivos.** Identifican patrones que explican o resumen los datos, es decir, sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos. Por ejemplo, una agencia de viaje desea identificar grupos de personas con unos mismos gustos, con el objeto de organizar diferentes ofertas para cada grupo y poder así remitirles esta información; para ello analiza los viajes que han realizado sus clientes e infiere un modelo descriptivo que caracteriza estos grupos.

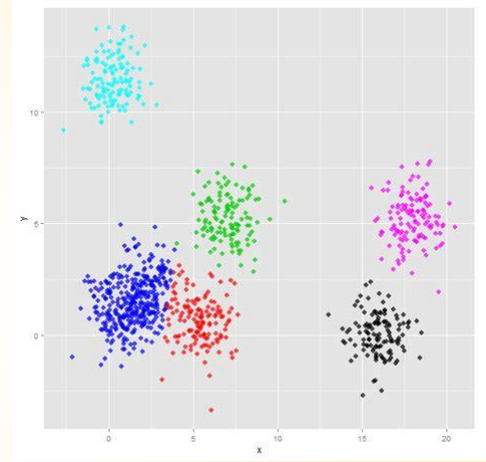


Figura Creative Commons. Tomada de: <https://stats.stackexchange.com/questions/185441/meaning-of-this-cluster-analysis>

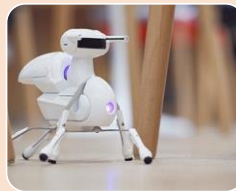
Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

Las propiedades del conocimiento extraído son (Hernández, Ramírez y Ferri, 2004):

18



**Válido:** hace referencia a que los patrones deben seguir siendo precisos para datos nuevos (con un cierto grado de certidumbre), y no sólo para aquellos que han sido usados en su obtención.



**Novedoso:** que aporte algo desconocido tanto para el sistema y preferiblemente para el usuario.



**Potencialmente útil:** la información debe conducir a acciones que reporten algún tipo de beneficio para el usuario.



**Comprensible:** la extracción de patrones no comprensibles dificulta o imposibilita su interpretación, revisión, validación y uso en la toma de decisiones. De hecho, una información incomprensible no proporciona conocimiento (al menos desde el punto de vista de su utilidad).

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

19

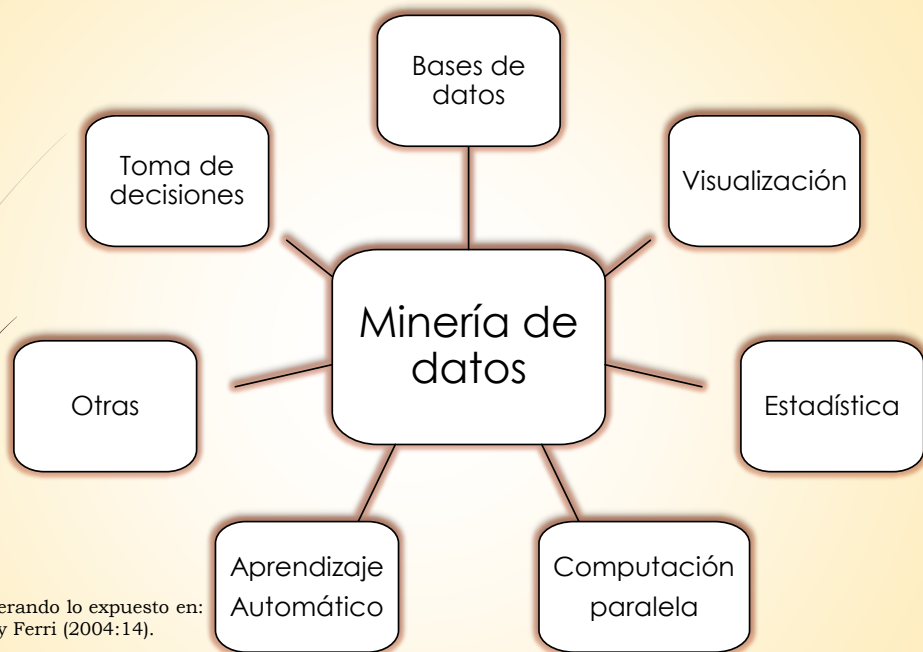


Figura creada considerando lo expuesto en:  
Hernández, Ramírez y Ferri (2004:14).

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

20



Figura creada considerando lo expuesto en:  
Hernández, Ramírez y Ferri (2004:16-17).

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

## 21

Clasificación de sistemas y herramientas considerando el modelo de datos que generan, el tipo de datos que minan, el tipo de técnica o el tipo de aplicación al que se pueden aplicar (Hernández, Ramírez y Ferri (2004)).

- **Tipo de base de datos minada:** teniendo en cuenta los diferentes modelos de datos se consideran: sistemas de minería de datos relacionales, multidimensionales, orientados a objetos, etc. Asimismo, atendiendo al tipo de datos manejados, se tienen sistemas textuales, multimedia, espaciales o web.
- **Tipo de conocimiento minado:** también pueden tenerse en cuenta los niveles de abstracción del conocimiento minado: conocimiento generalizado (alto nivel de abstracción), a nivel primitivo (a nivel de filas de datos), o conocimiento a múltiples niveles (de abstracción). Por último, se pueden distinguir entre los sistemas que buscan regularidades en los datos (patrones) frente a los que analizan las irregularidades (excepciones).

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

## 22

- **Tipo de funcionalidad y de técnica:** los sistemas de minería de datos se pueden clasificar basándose en su funcionalidad (clasificación, agrupamiento, etc.) o por los métodos de análisis de los datos empleados (técnicas estadísticas, redes neuronales, etc.).
- **Tipo de aplicación:** Se distinguen dos clases de sistemas según la aplicación para la que se usan: los sistemas de propósito general y los sistemas específicos (como los usados en aplicaciones financieras, web, e-mail, análisis de stocks, etc.).

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

23

## Metodología CRISP-DM de minería de datos

Joyanes (2019) establece que la metodología llamada Proceso estándar de la industria para la minería de datos (*Cross Industry Standard Process for Data Mining*, CRISP-DM) es una metodología de minería de datos abierta y no propietaria, se construyó sobre la base de experiencias reales y por empresas de gran prestigio. Consta de seis fases:

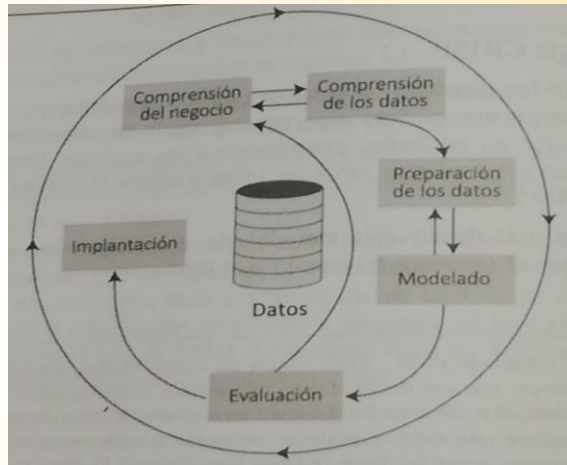


Figura tomada de Daza (2016).

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

24

### Fase 1. Comprensión del negocio.

Se enfoca en la comprensión de los objetivos del proyecto y en la definición de las necesidades del cliente, lo cual se transforma en un problema de minería de datos, se diseña un plan para lograr los objetivos. Todo desde una perspectiva de negocio y no técnica.

### Fase 2. Comprensión de los datos.

Hay que recopilar y comprender los datos, descubrir conocimiento preliminar de los datos mediante la realización de actividades, identificar problemas de la calidad y analizar las primeras potencialidades y/o descubrir subconjuntos interesantes para formular hipótesis sobre la información oculta.

### Fase 3. Preparación de datos.

Implica la construcción del conjunto final de los datos, es decir, los datos que se utilizarán en las herramientas de modelado a partir de los datos iniciales. Incluye: selección de tablas, registros y atributos, así como la transformación y limpieza de los datos para las herramientas que se van a utilizar para el análisis.

Se crea la vista “minable”. Se realiza la selección de datos, se limpian, se generan nuevas variables, se integran diferentes conjuntos de datos y los cambios de formato.

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello



25

**Fase 4. Modelado de datos.**

Se seleccionan y aplican las técnicas de modelado acordes al problema, hay varias técnicas para el mismo tipo de problema, algunas técnicas requieren aspectos específicos de los datos, por lo que en cualquier proyecto se acaba volviendo a la fase de preparación de datos.

**Fase 5. Evaluación.**

En esta etapa se han construido uno o varios modelos, antes de proceder al despliegue final del modelo, es importante evaluarlo con los objetivos del negocio.

**Fase 6. Despliegue.**

El objetivo de esta etapa es la distribución o desarrollo (despliegue) y la puesta en producción.

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

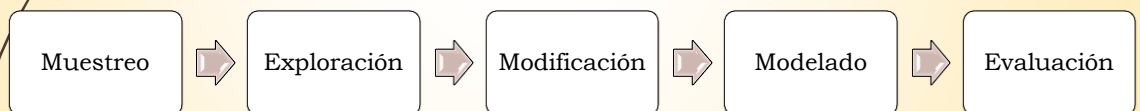
26

**Metodología SEMMA de minería de datos**

Joyanes (2019) presenta la metodología SEMMA, la cual es la abreviatura de *Sample* (muestreo), *Explore* (exploración), *Modify* (modificación), *Model* (modelado) y *Asses* (valoración).

Se puede definir como: “el proceso de selección, exploración y modelado de grandes volúmenes de datos para descubrir patrones de negocio desconocidos”.

El gráfico de SEMMA es:



Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

Las etapas son:

27

- **Muestreo.** Genera una muestra representativa de datos. Se identifican los datos: entradas de datos, ejemplos, partición de datos.
- **Exploración.** Visualización y descripción básica de los datos. Se exploran los conjuntos de datos para identificar relaciones y patrones, las variables importantes y las asociaciones.
- **Modificación.** Se seleccionan las variables y la transformación de las mismas, se preparan para el análisis considerando la transformación de variables, los datos fuera de rango, agrupamiento, ruido.
- **Modelado.** Se utilizan diversas técnicas estadísticas y modelos de aprendizaje automático mediante regresión, árboles, redes neuronales, etc.
- **Evaluación (Valoración).** Se evalúan la precisión y la utilidad de los modelos mediante reportes, medidas, evaluación.

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

28

García y otros (2018) presentan dos tipos recientes de minería de datos: minería de textos y minería de datos web. Los cuales se presentan a continuación.

## Minería de textos (*text mining*)

Surge de la necesidad de extraer automáticamente información de masas de textos, de datos no estructurados. Existen varias representaciones de la información no estructurada.



Esta foto de Autor desconocido está bajo licencia [CC BY-SA-NC](#)

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

29

Joyanes (2019:271) establece que la minería de textos se refiere a:

- El concepto de Minería de Textos o Descubrimiento del Conocimiento en textos, KDT (Knowledge Discovery in Texts ) fue utilizado por primera vez por Feldman et al en 1995. Ha ganado gran atención en los últimos años debido a la gran cantidad de datos de texto que se genera cada día y que se crea en una gran variedad de formatos, tales como datos de redes sociales
- La Minería de Textos es un enfoque de la Minería de Datos, tiene el mismo propósito y utiliza los mismos procesos, pero con la Minería de Textos la entrada al proceso es una colección de archivos de datos no estructurados o semiestructurados, tales como documentos Word, archivos PDF, resúmenes de textos, archivos XML, etcétera.

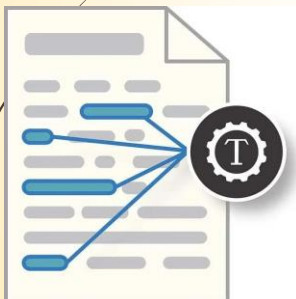


Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

30

Joyanes (2019:272) establece que la minería de textos se refiere a:

- La Minería de Textos se aplica en áreas donde existen o se generan grandes cantidades de datos, tales como bibliotecas, centros de investigación en universidades, finanzas, centros de salud (informes, recetas, análisis de enfermedades), medios de comunicación (revistas, periódicos), marketing— o mercadotecnia— (comentarios de clientes, socios, campañas de publicidad).



- En la actualidad, el correo electrónico, los sistemas de mensajería instantánea, sistemas de chats, comunicaciones electrónicas, entre otros, generan y distribuyen grandes cantidades de texto.

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

31

## Minería de datos web (web mining)



Es una tecnología que se utiliza para descubrir conocimiento en aspectos relacionados con la web. Dependiendo del tipo de información que se desea extraer, ésta se puede clasificar en tres conjuntos no disjuntos:

Esta foto de Autor desconocido está bajo licencia [CC-BY-SA](#)

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

32

Joyanes (2019:270) establece que la minería web se refiere a:

- La Web es una colección de archivos interrelacionados en uno o más servidores web. La Minería Web (*Web Mining*) es la aplicación de técnicas de Minería de Datos para extraer conocimiento de los datos de la Web.
- Los contenidos predominantes de la Web son datos de texto no estructurados, codificados en los lenguajes de programación HTML o XML, la infraestructura de la Web contiene información de hiperenlaces (conexiones o enlaces a otras páginas web) e información de uso (registros de interacciones de los visitantes en los sitios web)



Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello



33

Han y Kamber (2006, citado en Joyanes, 2019) mencionan los grandes retos que afrontaba la Web para el descubrimiento de un conocimiento efectivo y eficiente:

- La Web es demasiado grande para un minado de datos eficiente
- La Web es demasiado compleja
- La Web es demasiado dinámica
- La Web no es específica a un dominio
- La Web tiene de todo.



Minería Web es el uso de técnicas de minería de datos aplicadas a la búsqueda, extracción y evaluación automática de información para conseguir el descubrimiento del conocimiento de los recursos de la Web: documentos y servicios (Joyanes, 2019:278).

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

34

Joyanes (2019) establece que las técnicas de minería web se apoyan en técnicas de minería de datos, pero tienen características propias fundamentadas en los sitios y páginas web:

- La mayoría de los datos de la Web son no estructurados
- Los volúmenes de datos son muy grandes y crecen de modo exponencial, por lo que las técnicas de analítica de Big Data son imprescindibles
- Los datos a nivel de páginas están relacionados mediante vínculos o hipervínculos (*links*)
- Los datos tienen formatos muy variados, tales como HTML, PDF, imágenes, videos, audio, etcétera
- Los datos son de muy diferente fiabilidad: se mezclan datos muy fiables con datos de escasa fiabilidad, lo que produce inconsistencia.

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

35

## Categorías de la Minería Web (Joyanes, 2019:281)

**Minería Web de Contenido (Web Content Mining)**

- Se analiza texto no estructurado de las páginas web (comúnmente en formato HTML: texto, imágenes y registros).
- Se extrae información del contenido de los documentos en la Web.

**Minería Web de la Estructura (Web Structure Mining).**

- Análisis de los enlaces contenidos en las páginas web (URL, Uniform Resource Locator: hiperenlaces, tags, etcétera).
- Se intenta descubrir un modelo a partir de la topología de los enlaces de la red. Este modelo puede ser útil para clasificar o agrupar documentos.

**Minería Web del Uso (Web Usage Mining).**

- Descripción detallada de las visitas a un sitio Web (secuencia de clics por sesión de visitas: logs http, app server logs , etcétera).
- Se intenta extraer información (hábitos, preferencias, etcétera, de los usuarios).

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

36

**1) Minería del contenido de la web (web content mining).**

Extraer información del contenido de los documentos existentes en la web. Los cuales se clasifican en:

- Text mining. Documentos de texto, sin formato.
- Hypertext mining. Si los documentos contienen enlaces a sí mismo o a otros documentos.
- Markup mining. Si los documentos son semiestructurados (con marcas).



Esta foto de Autor desconocido está bajo licencia CC BY-SA

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

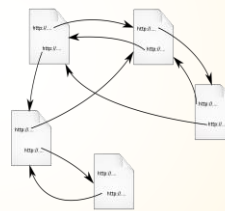
37

**2) Multimedia mining.** Para imágenes, audio, video, etc.

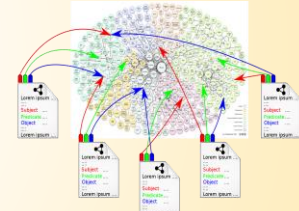


Esta foto de Autor desconocido está bajo licencia [CC BY-SA-NC](#)

**3) Minería de la estructura de la web (*web structure mining*).** Se intenta descubrir un modelo a partir de la tipología de enlaces de la red. Puede ser útil para agrupar o clasificar documentos.



The traditional web -  
A web of documents



The semantic web -  
A web of human and machine  
readable content employing linked data

Esta foto de Autor desconocido está bajo licencia [CC BY-SA](#)

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

38

## ¿Qué es el Big Data?

Según Joyanes (2019) Josep Curto establece que El Big Data es el conjunto de estrategias, tecnologías y sistemas para el almacenamiento, procesamiento, análisis y visualización de conjuntos de datos complejos, que frecuentemente, pero no siempre, viene definida por volumen, velocidad y variedad.



Esta foto de Autor desconocido está bajo licencia [CC BY-SA](#)

No es la cantidad de información lo que marca la diferencia, sino que se trata de nuestra capacidad para analizar series extensas y complejas de datos que van más allá de todo lo que hubiéramos podido hacer anteriormente (Joyanes, 2019:8).

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

39

Joyanes (2019) establece que el término Big Data se refiere al conjunto de datos de gran volumen y complejos que las herramientas tradicionales, como las bases de datos relacionales, son incapaces de procesar en un rango de tiempo o costos aceptables. El concepto de Big Data se utiliza para referirse a los conjuntos de datos voluminosos que exceden a la capacidad de manipulación de las herramientas tradicionales. Se alimenta de grandes volúmenes de datos que tienen diferente formato, no estructurados y semiestructurados.

#### Datos estructurados

- Almacenados en filas y columnas.

#### Datos Semiestructurados

- Usan marcadores para separar elementos. Ej. Documentos XML, HTML, datos de sensores, etc.

#### Datos no estructurados

- En formatos que no pueden ser manipulados fácilmente por las bases de datos relacionales. Ej. Documentos, multimedia, audio, voz, video, fotografías, correos electrónicos.

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

40

### Referencias Bibliográficas

- Daza Vergaray, Alfredo. (2016). Data Mining. Minería de datos. Alfaomega-Macro.
- Dunham, M. H. (2002). *Data mining: introductory and advanced topics*. Prentice Hall.
- García, Molina, Berlanga, Patricio, Bustamante y Padilla. (2018). Ciencia de datos. Técnicas analíticas y aprendizaje estadístico. Alfaomega.
- Hernández Orallo, José; Ramírez Quintana, M<sup>a</sup> José y Ferri Ramírez, César. (2004). Introducción a la Minería de datos. Editorial Pearson.
- Joyanes Aguilar, Luis. (2019). Inteligencia de negocios y analítica de datos. Una visión global de Business intelligence & Analytics. Alfaomega.
- Maimon, O. & Rokach, L. (2010). Data Mining and Knowledge Discovery Handbook. Second Edition. Springer.
- Sahu, Hemlata; Shirma, Shalini; Gondhalakar, Seema. (2011). A Brief Overview on Data Mining Survey. *International Journal of Computer Technology and Electronics Engineering (IJCTEE)*. Vol.1.

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello