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*

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

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/

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).

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).

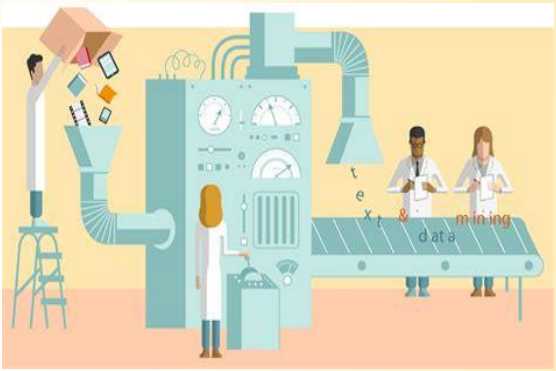


Figura Creative Commons. Tomada de: https://ruedimumenthaler.ch/2015/06/09/trend-und-herausforderung-text-and-data-mining/

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

7

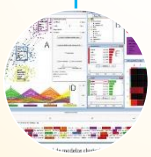
Minería de datos

que enfrenta

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

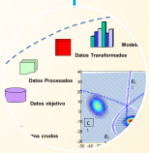
Diversos tipos de datos con ruido, ausentes, sucios, etc.

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.



Formas de evaluar los resultados y expresión gráfica de los resultados

8

Tipos de fuentes de datos

Otras bases de datos

- **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.

Bases de datos relacionales

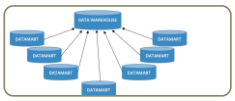
Imagen formada considerando lo expuesto en Hernández, Ramírez y Ferri (2004).

9 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, Shrma, & Gondhalakar, 2011).

10 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.

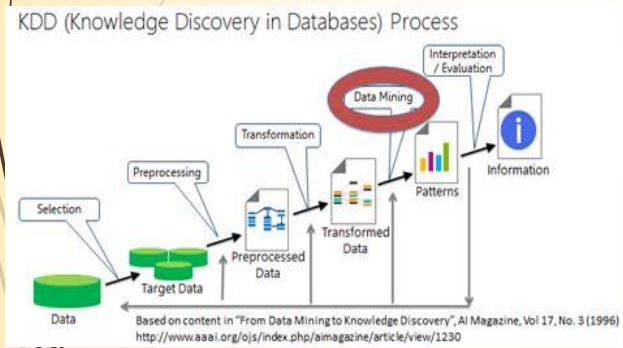


Figura Creative Commons. Tomada de: https://www.actuaries.digital/2016/07/28/dat203x-data-science-and-machine-learning/

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).

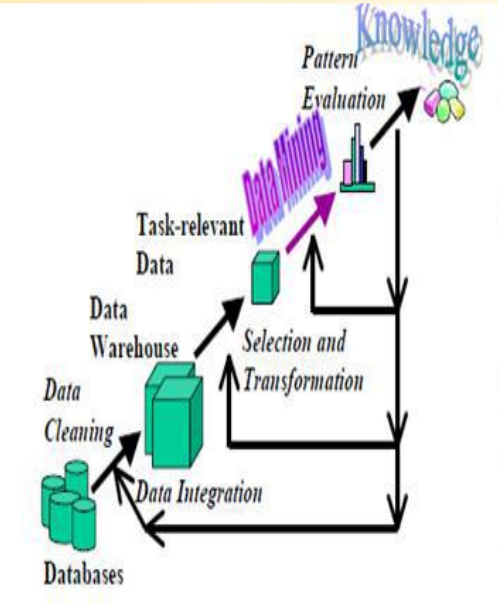


Imagen tomada de Sahu, Shrma, & Gondhalakar (2011).

El proceso iterativo consiste de los siguientes pasos:

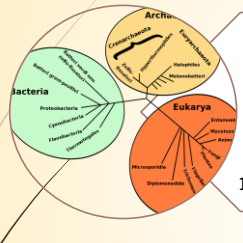
- **Limpieza de datos**. Se eliminan los datos con ruido o sucios y los datos irrelevantes que se encuentran en la colección de datos.
- **Integración de datos**. Se integran múltiples fuentes de datos comúnmente heterogéneas en una fuente de datos en común.
- **Selección de datos**. Se seleccionan de 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.

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

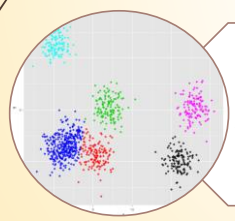
Paso	Proceso de KDD (Fayyad, 1996, citado en Joyanes, 2019).	Proceso de minería de datos Joyanes (2019)
1	Selección de datos. - Fuentes de datos. - Tipo de información.	Selección del conjunto de datos. - Variables objetivo, variables independientes y selección de registros.
2	Preprocesamiento. - Preparación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

13

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.

14

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.

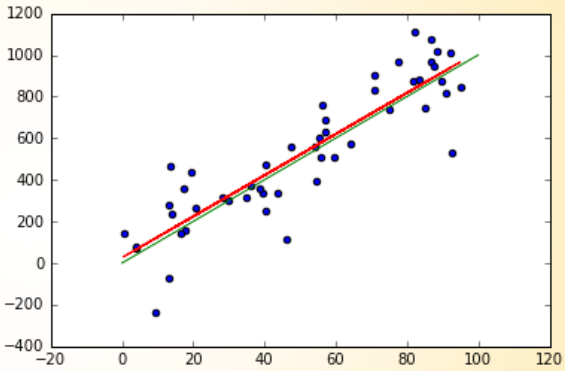


Figura Creative Commons. Tomada de: https://machinelearningparatodos.com/regresion-lineal-en-python/

15

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.

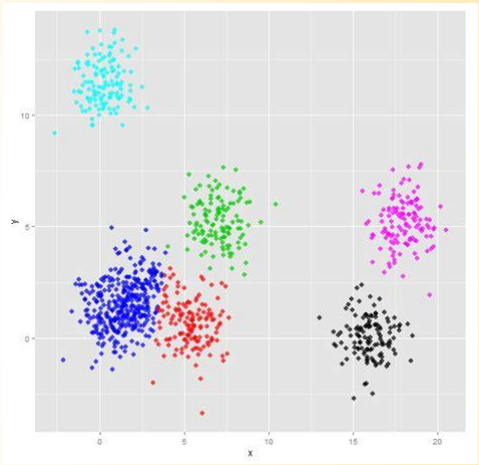


Figura Creative Commons. Tomada de: https://stats.stackexchange.com/questions/185441/meaning-of-this-cluster-analysis

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

16

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).

17

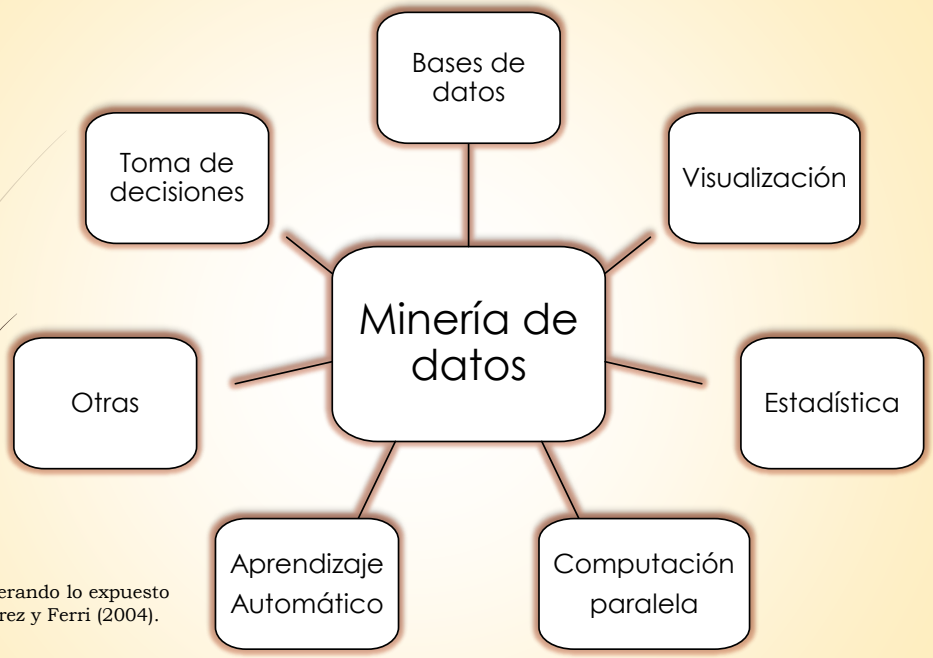


Figura creada considerando lo expuesto en: Hernández, Ramírez y Ferri (2004).

18

19

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:

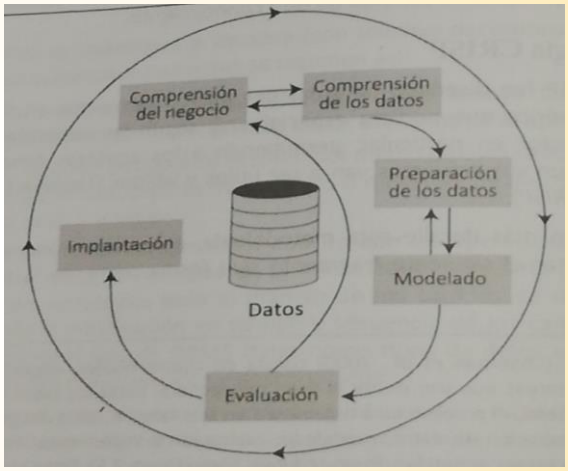


Figura tomada de Daza (2016).

20

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.

21

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.

22

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:

23

Las etapas son:

- **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.

24

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.

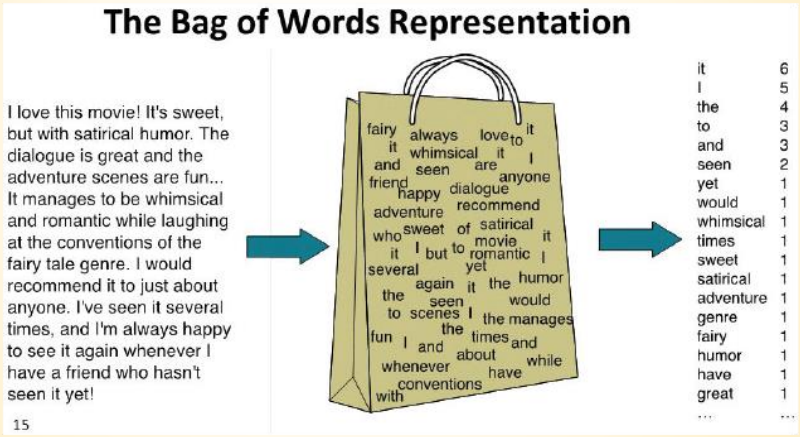
Minería de textos (*text mining*)

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

25

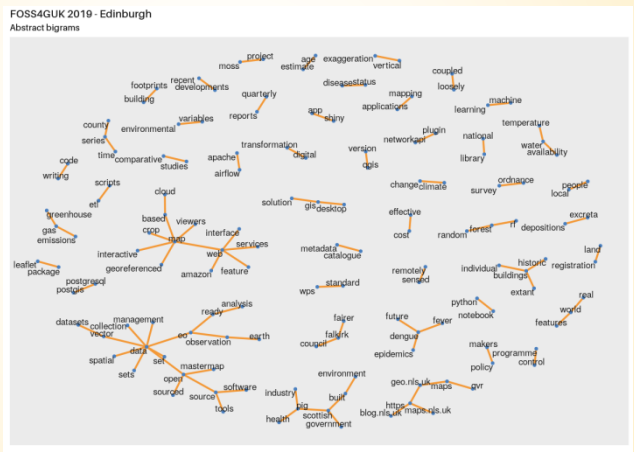
1) Bag of words. Cada palabra constituye una posición de un vector y el valor corresponde al número de veces que ha aparecido.



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

26

2) N-gramas o frases. Permite tener en cuenta el orden de las palabras. Trata mejor frases negativas como "…excepto…", "…pero no…", y considera las frases que le siguen como relevantes.



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

27

3) Representación relacional (primer orden). Permite detectar patrones más complejos (si la palabra X está a la izquierda de la palabra Y en la misma frase …).

4) Categorías de conceptos.

28

Minería de datos web (web mining)

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

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:

29

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).

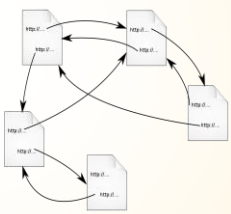
Esta foto de Autor desconocido está bajo licencia CC BY-SA

30

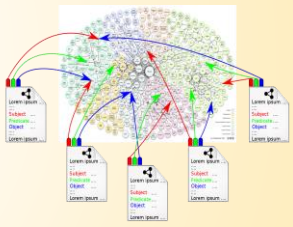
2) Multimedia mining. Para imágenes, audio, video, etc.

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

31

¿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).

32

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.

Referencias Bibliográficas

Daza Vergaray, Alfredo. (2016). Data Mining. Minería de datos. Alfomega-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, Mª 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 inteligence & Analytics. Alfaomega.

Maimon, O. & Rokach, L. (2010). Data Mining and Knowledge Discovery Handbook. Second Edition. Springer.

Sahu, Hemlata; Shrma, Shalini; Gondhalakar, Seema. (2011). A Brief Overview on Data Mining Survey. *International Journal of Computer Technology and Electronics Engineering (IJCTEE)*. Vol.1.