

## GUÍA DE DATA MINING



### TEMA: INTRODUCCIÓN A LA MINERÍA DE DATOS

#### ARTÍCULOS:

- Moreno Salinas, José Gerardo. (2017). Científico de datos: codificando el valor oculto e intangible de los datos. Revista Digital Universitaria. Vol. 18, Núm. 7, septiembre-octubre 2017. Disponible en: <https://revista.unam.mx/vol.18/num7/art53/index.html>
- Sahu, Hemlata; Shirma, Shalini & Gondhalakar, Seema. (2011). A Brief Overview on Data Mining Survey. International Journal of Computer Technology and Electronics Engineering (IJCTEE). Volume 1, Issue 3. Disponible en: <https://es.scribd.com/document/456490549/A-Brief-Overview-on-Data-Mining-Survey-pdf#>

#### LIBROS:

- 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.

### ACTIVIDADES DEL ARTÍCULO DE MORENO SALINAS (2017)

#### 1) Definiciones:

Palabra	Definición
Big data	
Huella digital	
Minería de datos	
Científico de datos	
Visualización de datos	
Internet de Contenido	
Internet de las personas	
Internet de las cosas	
Internet de la ubicación	

2) Cantidad de información que producen en 60 segundos las siguientes plataformas, según el reporte del año 2017.

Plataforma	Cantidad-descripción
Youtube	
Email	
Facebook	
Google	
Instagram	
Twitter	
Wordpress	
WhatsApp	

3) ¿Por qué es importante valorar los datos, no subestimarlos? Ejemplifique.

4) Identificar las características de cada uno de los siguientes aspectos:

Científico de datos	Economía del conocimiento	Áreas de conocimiento del científico de datos	Actividades del científico de datos
-	-	-	-

5) Estimación de los tiempos en el desarrollo de actividades del científico de datos. Comente de dónde obtuvieron estos datos.

Porcentaje	Actividad

6) Describa y ejemplifique en qué consisten las Áreas que maneja el científico de datos.

a. Big Data

**b. Minería de datos**

**c. Visualización**

**7) Características de las relaciones**

**a. Relación 1. Big data y minería de datos**

**b. Relación 2. Big Data y visualización de datos**

**c. Relación 3. Minería de datos y visualización de datos**

**d. Relación 1, 2 y 3. Big Data, minería y visualización de datos**

**8) Conclusiones**

- 9) Describa cinco ejemplos en donde se aplique la minería de datos. Agregue la fuente de donde obtiene la información.


- 10) Busque información actualizada de la cantidad de información que producen en 60 segundos las siguientes plataformas. Agregue las fuentes de donde toma la información.

Plataforma	Cantidad-descripción
Youtube	
Email	
Facebook	
Google	
Instagram	
Twitter	
Wordpress	
WhatsApp	
Otra	

### ACTIVIDADES DEL ARTÍCULO DE Sahu, Hemlata; Shirma, Shalini & Gondhalakar, Seema (2011)

- 1) Definiciones:

Palabra	Definición
Minería de datos	
KDD	

Diferencia entre minería de datos y el análisis de datos tradicional

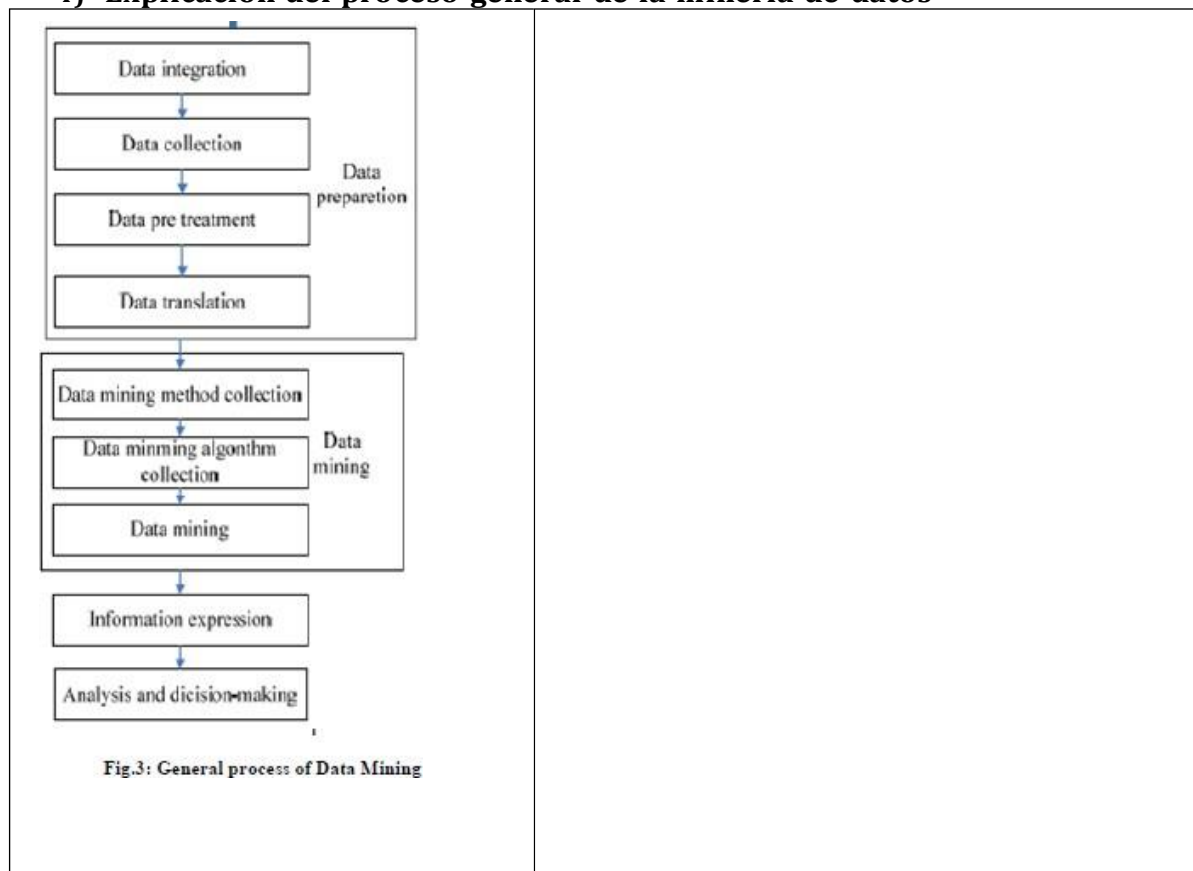
2) Describa cada una de las etapas del KDD.

Etapa	Descripción
Limpieza de datos	
Integración de datos	
Selección de datos	
Transformación de datos	
Minería de datos	
Evaluación de patrones	
Representación del conocimiento	

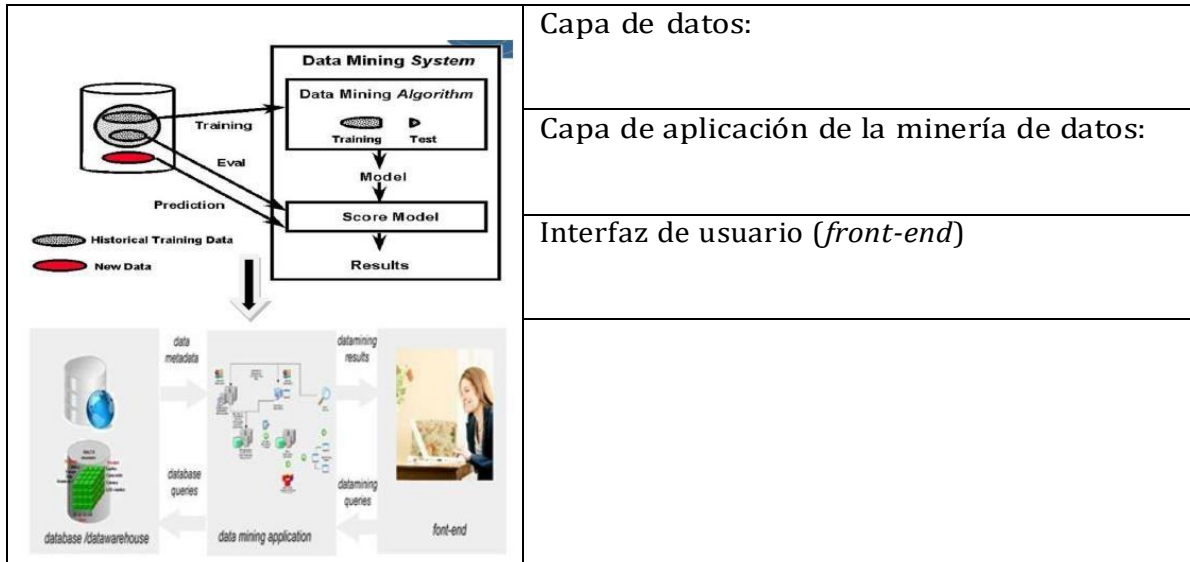
3) Tareas de la minería de datos:

Técnica	Descripción
Agrupamiento	
Clasificación	
Regresión	
Reglas de asociación	

4) Explicación del proceso general de la minería de datos



5) Explicación de cada una de las capas de la arquitectura



Capa de datos:

Capa de aplicación de la minería de datos:

Interfaz de usuario (*front-end*)

6) Técnicas de minería de datos:

Técnica	Descripción
Árboles de decisión	
Sistema de soporte de decisiones	
Redes neuronales	
Agrupamiento k-medias	

7) Explique seis aplicaciones de la minería de datos

1)	
2)	
3)	
4)	
5)	
6)	

8) Desventajas y ventajas de la minería de datos

Ventajas	Desventajas

**9) Retos de la minería de datos**

**10) Futuro de la minería de datos**

**11) Conclusiones**

**ACTIVIDADES DEL LIBRO: Hernández Orallo, José; Ramírez Quintana, M<sup>a</sup> José y Ferri Ramírez, César. (2004).**

- Lectura del capítulo 1 y 2
- Lectura de apoyo de los capítulos 8 y 9 del libro de Joyanes Aguilar, Luis. (2019)
- Diapositivas de la primera parte del curso (Introducción).



**TEMA: TRATAMIENTOS DE DATOS**

**TIPOS DE DATOS:**

- Nominales, ordinales, intervalo, razón
- Cualitativos y cuantitativos
- Transversales y series de tiempo
- Diferencia entre términos: estadísticas y estadística
- Estadísticos de localización
- Estadísticos de dispersión
- Gráficas y expresiones tabulares de representación de datos

**Bibliografía de consulta:**

- Anderson, Sweeney & Williams. (2008). Estadística para administración y economía, 10ª edición. Cengage Learning
- Bennet, Briggs & Triola (2011). Razonamiento estadístico. Pearson. México

**TRATAMIENTO DE DATOS:**

- Tratamiento de datos: imputación, transformación, normalización, recodificación, contenedores.

**Bibliografía de consulta:**

- Larose, T. Daniel & Larose, D. Chantal. (2015). Data Mining and Predictive Analytics. Second Edition. Wiley.



**TEMA: CLASIFICACIÓN**

**ARTÍCULO:** Bhumika Gupta, Aditya Rawat, Akshay Jain, Arpit Arora, Naresh Dhami. (2017). Analysis of Various Decision Tree Algorithms for Classification in Data Mining. International Journal of Computer Applications (0975-8887). Volume 163 – No 8, April 2017. Disponible en: [https://www.researchgate.net/publication/316173815\\_Analysis\\_of\\_Various\\_Decision\\_Tree\\_Algorithms\\_for\\_Classification\\_in\\_Data\\_Mining](https://www.researchgate.net/publication/316173815_Analysis_of_Various_Decision_Tree_Algorithms_for_Classification_in_Data_Mining)

**Ejercicio No. 1**

- a) Describa la minería de datos

- b) Explique las razones por las cuales se utilizan los árboles de decisión



c) ¿Cuál es la diferencia entre un árbol de clasificación y un árbol de regresión?

--

## Ejercicio No. 2

Considerando los algoritmos: ID3, C4.5, CART y Random Forest. Realice un cuadro comparativo que considere los siguientes aspectos: descripción del algoritmo, criterio de partición que utiliza, si utiliza poda o no, tipo de datos que utiliza, ventajas y desventajas.

Criterio	ID3	C4.5	CART	Random Forest
Descripción				
Criterio partición				
Poda				
Tipos de datos que utiliza				

Criterio	ID3	C4.5	CART	Random Forest
Ventajas				
Desventajas				

## Ejercicio 3

Considerando los siguientes criterios de selección de atributo para particionamiento: *Entropy (Information Gain)*, *Gain Ratio* and *Gini Index*. Realice una descripción con sus propias palabras de cada uno de ellos.

Criterio	Descripción
Entropy (Information Gain)	
Gain Ratio	
Gini Index	

**LIBRO:** Tan, Pang-Ning, Steinbach, Michael & Kumar, Vipin. (2014).  
Introduction to data mining. Pearson.

**Capítulo 4. Clasificación: conceptos básicos, árboles de decisión y evaluación de modelos**

- 1) Explique qué es el Proceso de Clasificación
- 2) Describa los siguientes modelos
  - a. Modelo descriptivo
  - b. Modelo predictivo
- 3) Describa el enfoque general para resolver problemas de clasificación (Sección 4.2)
- 4) Describa la forma en que se resuelve un problema de clasificación (Sección 4.2.1)
- 5) Identifique los criterios considerados en la construcción de la figura 4.4 (Sección 4.3.1)
- 6) Describa los dos pasos del Algoritmo de Hunt para la construcción de árboles de decisión (Sección 4.3.2)
- 7) Explique los dos casos que requieren condiciones adicionales en la construcción de un árbol de decisión considerando la combinación de valores y etiquetas asociadas en los elementos (Página 154).
- 8) Responda las siguientes preguntas:
  - a. ¿Cómo se deben dividir los registros del conjunto de entrenamiento?
  - b. ¿Cómo debe terminar el procedimiento de división?
- 9) Explique los métodos para expresar la condición de particionamiento aplicable a cada uno de los siguientes tipos de atributo:
  - a. Binario
  - b. Nominales
  - c. Ordinales
  - d. Continuos
- 10) Explique el criterio que se utilice en las medidas para seleccionar la mejor forma de dividir los registros (Sección 4.3.4)
- 11) Justifique por qué los autores mencionan que el mejor atributo de división es el Tipo de Automóvil (Car Type) en el árbol de la figura 4.12 (Sección 4.3.4)
- 12) Explique en qué consiste el criterio de impureza de nodos tomado en cuenta para la elección de atributos de particionamiento (Sección 4.3.4).
- 13) Explique el significado de  $\Delta$  (Sección 4.3.4).

- 14) Considerando la figura 4.14, explique la conveniencia de la elección del atributo B para realizar el particionamiento (Sección 4.3.4).
- 15) Analice el efecto que tienen las divisiones binarias con respecto a la división múltiple de la figura 4.15 (Sección 4.3.4).
- 16) Explique cómo se realiza la división de atributos continuos (Sección 4.3.4).
- 17) Explique la forma en que se podrían tratar atributos de clave principal aplicando criterios de división (Gain Ratio Sección 4.3.4).
- 18) Explique por lo menos seis de las once características de la inducción del árbol de decisión (Sección 4.3.7)
- 19) Explique lo siguiente (Sección 4.4):
  - a. En qué consisten los errores de entrenamiento y los errores de generalización;
  - b. Las características que tiene un buen modelo
  - c. ¿Qué significa el sobreajuste?
  - d. ¿Qué significa el subajuste?
- 20) Explique el comportamiento de las tasas de error de entrenamiento y prueba expresados en la figura 4.23 (Sección 4.4)
  - a. Cuando el árbol es pequeño (pocos nodos)
  - b. Conforme aumenta la cantidad de nodos en el árbol
- 21) Describa de forma breve los siguientes casos de sobreajuste del modelo:
  - a. Sobreajuste por presencia de ruido
  - b. Sobreajuste debido a la falta de muestras representativas
- 22) Describa los siguientes procesos de evaluación del desempeño de un clasificador:
  - a. Método de retención (holdout) (Sección 4.5.1)
  - b. Submuestreo aleatorio (Sección 4.5.2)
  - c. Validación cruzada (Sección 4.5.3)
  - d. Bootstrap (Sección 4.5.4)



## Agrupamiento

4

### TEMA: AGRUPAMIENTO

**LIBRO:** Dunham, M. H. (2002). Data mining: introductory and advanced topics. Prentice Hall

Estudiar los temas:

- 5.1-5.4 – Teoría y ejemplos
- 5.5 – Teoría y ejemplos (K-Means, KNN)
- 5.6 – Teoría y ejemplos (5.7)

**LIBRO:** Larose, T. Daniel & Larose, D. Chantal. (2015). Data Mining and Predictive Analytics. Second Edition. Wiley. *Capítulo 22*

Estudiar los temas:

Medidas de bondad del Clúster (*Measuring Cluster Goodness*)

Características de diversos tipos de Algoritmos de Agrupamiento.

Materiales de curso.



## Predicción

5

### TEMA: PREDICCIÓN

Estudiar los temas:

- Los materiales del curso
- Correlación, mínimos cuadrados, Verificación de la ecuación de estimación, Suma de cuadrados debida al error, Suma total de cuadrados, Suma de cuadrados debida a la regresión, Consideraciones al aplicar la regresión, Coeficiente de determinación, Interpretación, Coeficiente de correlación
- El error estándar de la estimación, Intervalos de confianza para la estimación (o el valor esperado), Consideraciones de la aplicación de la regresión lineal, Residuales, gráficas de residuales
- Verificación de la ecuación de estimación, Definición de significancia estadística, Estimación de  $\sigma^2$ , SCE (Suma de cuadrados debida al error), El error cuadrado medio (ECM), Error estándar de estimación, Prueba t, Prueba F.

**Bibliografía de consulta:**

- Levín Rubín, Balderas, Del Valle y Gómez. 2004. Estadística para administración y economía Séptima Edición. Prentice Hall
- Mason, Lind Marshal. 2000. Estadística para administración y economía. Alfaomega. 10<sup>a</sup> edición
- Anderson, Sweeney Williams. 2008. Estadística para administración y economía, 10<sup>a</sup> edición Cengage Learning
- Bennet, Briggs Triola. 2011. Razonamiento estadístico Pearson México



## Reglas de Asociación

6

### TEMA: REGLAS DE ASOCIACIÓN

Estudiar los temas:

- Qué son las reglas de asociación (RA), Conceptos necesarios para la aplicación de las RA: Soporte, Confianza, Algoritmo A priori, Itemset, k-itemset, itemset frecuente, elementos frecuentes, reglas fuertes, propiedad a priori de las RA
- Comprensión de la aplicación de los conceptos en la identificación de RA
- Los tres pasos para la creación de las RA
- Pasos de la creación de conjuntos de elementos frecuentes
- Evaluación de las RA
- RA aplicadas en aprendizaje supervisado y aprendizaje no supervisado
- Modelos vs patrones

**Bibliografía de consulta:**

- Han, Jiawei Kamber Micheline Pei Jian 2012 Data Mining concepts and techniques Third edition Morgan Kaufman Series
- Larose, T. Daniel & Larose, D. Chantal. (2015). Data Mining and Predictive Analytics. Second Edition. Wiley