ÁREA 4. SOLUCIONES DE CÓMPUTO INTELIGENTE

SUBÁREA 4.2 MINERÍA DE DATOS

BIBLIOGRAFÍA DE LA GUÍA:

- Aggarwal, Charu. C. (2015). Data mining: the textbook. Springer.
- Bishop, Christopher. M. (2006). Pattern recognition and machine learning. Estados Unidos: Springer.
- Flach, Peter. (2012). Machine learning: the art and science of algorithms that make sense of data. Estados Unidos: Cambridge University Press.
- Gonzalez C., Rafael y Woods E. Richard. (2008). Digital image processing. 3a. ed Estados Unidos: Pearson.
- Goodfellow, Ian., Bengio, Yoshua y Courville, Aaron. (2016). Machine learning basics.
- Hand, David, Mannila, HMannila y Smyth, Padrhaic. (2001). Principles of data mining. Estados Unidos: A Bradford book, the MIT Press.
- Madhulatha, T. Soni. (2012). An overview on clustering methods. IOSR Journal of Engineering. Recuperado de: https://arxiv.org/ftp/arxiv/papers/1205/1205.1117.pdf
- Matousek, Jiri y Gärtner, Bernd. (2007). Understanding and Using Linear Programming (Universitext). Estados Unidos. Springer.
- Mohri, Mehryar, Rostamizadeh, Afshin y Talwalkar, Ameet. (2018). Foundations of machine learning. MIT press.
- Russell, Sturt y Norvig, Peter. (2011). Inteligencia Artificial Un Enfoque Moderno (3a ed.). Madrid, España: Pearson Prentice Hall.
- Zaki, Mohammed J. y Meira, Wagner. (2014). Data mining and analysis: fundamental concepts and algorithms. Estados Unidos: Cambridge University Press.

Acerca de la bibliografía: El libro que contiene todos los temas es el de Aggarwal, por si solo quieren basarse en un libro.

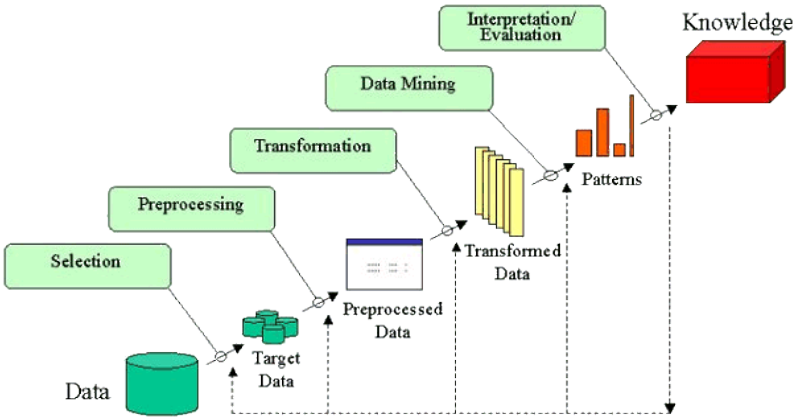
TEMAS IMPORTANTES:

Descubrimiento de Conocimiento en Bases de Datos:

La minería de datos es parte del proceso llamado **Knowledge Discovery in Databases (KDD)**, por lo que realmente no solo deben estudiar el tema de minería de datos por sí solo, sino en conjunto con todo el proceso. Lo primero es leer en general acerca del proceso:

http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html
https://www.geeksforgeeks.org/kdd-process-in-data-mining/
https://www.tamps.cinvestav.mx/~hmarin/Mineria/EC2.pdf

Es necesario entender y recordar todos los módulos, componentes y la lógica del proceso completo de KDD. El diagrama del proceso de KDD es:



Más información del proceso completo viene en el capítulo 1 del libro de Aggarwal.

Datos:

Lo primero que el proceso de KDD requiere es de los datos. Es importante diferenciar y conocer los diferentes tipos de datos que existen, ya que dependiendo del problema se requieren o se utilizan diferentes datos. Además, conocer los conceptos de dataset (conjunto de datos), datos crudos (raw data), database (bases de datos), etc.

Para comprender este tema se sugiere leer el capítulo 1 e inicio del 2 del libro de Aggarwal.

Preprocesamiento y Transformación:

Una vez que se eligieron los datos específicos, se pueden realizar operaciones para el preprocesamiento de esos datos, como limpieza, remover ruido y outliers (puntos fuera de la mayoría de los datos), datos faltantes, etc. Este paso es importante para que el buen funcionamiento de los siguientes módulos.

https://sci2s.ugr.es/sites/default/files/ficherosPublicaciones/2133_Nv237-Digital-sramirez.pdf
https://www.researchgate.net/publication/319019536_Data_Pre-processing_for_knowledge_discovery

La transformación de datos puede ir ligada al preprocesamiento, la diferencia es que involucra métodos que cambian por completo los datos originales. Por ejemplo, en los datos originales podemos encontrar las calificaciones de cada materia de un conjunto alumnos, pero en lugar de usar todas las calificaciones, se podrían transformar y calcular solo el promedio de ellas. El dato original se transformó en otro que nos proporciona una información casi igual pero más compacta. Estos temas vienen en el capítulo 2 del libro de Aggarwal.

Selección:

En ciertos libros, la Selección de datos o características se coloca antes del preprocesamiento; pero también hay ciertas fuentes que lo colocan después. Realmente depende del problema, para que lo consideren. Los datos seleccionados se les conoce también como **features** o

características. Entonces, cuando se hable de Selección la pueden encontrar como Feature Selection.

Existen varios métodos y algoritmos que se usan para esta selección. Estos algoritmos pueden ser de Aprendizaje de Máquina (machine learning), por lo que puede resultar un poco confuso el usar estos métodos en esta parte, y luego también en la propia minería de los datos. El detalle es que podemos tener un dataset con muchas características, lo que puede hacer muy complicado usar todas esas características en la minería de datos, por lo que es mejor seleccionar un subconjunto de estas; particularmente las que ofrezcan más información.

https://aprendeia.com/metodos-de-seleccion-de-caracteristicas-machine-learning/

Similitud y Distancias:

Una de las principales tareas de la minería de datos es calcular la similitud entre diferentes datos (objetos, patrones, eventos, etc.) para poder agrupar aquellos que son similares entre sí, y separar los que no sean similares. Es importante entonces conocer algunas técnicas de cómo poder hacer esta medida de similitud, dicha medida es básicamente calcular la distancia que existe entre un dato y otro.

Entre las distancias más comunes se encuentra la Euclidiana, la de Manhattan, la Correlación de Pearson, entre otras. No es posible estudiar todas las distancias, pero al menos estudien las más simples, como la Euclidiana y la de Manhattan:

http://www.econ.upf.edu/~michael/stanford/maeb4.pdf
https://towardsdatascience.com/17-types-of-similarity-and-dissimilarity-measures-used-in-data-science-3eb914d2681
http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports/bhoenes/similarity.html

El capítulo 3 del libro de Aggarwal trata de este tema, pero es un tanto extenso. Es más recomendable usar las páginas web.

Minería de Datos:

Como pueden ver en este enlace:

http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/2_tasks.html

La minería de datos tiene como objetivo realizar una de las siguientes tareas: Clasificación, Regresión, Clustering, entre otras. Siendo la clasificación, regresión y el clustering las más comunes. Se recomienda que entiendan qué hace cada una de estas tareas.

Muchos métodos o algoritmos de clasificación se usan también para la regresión, por lo que se puede, hasta cierto punto, considerar los mismos algoritmos. Entre los métodos más comunes están las redes neuronales, árboles de decisión, naïve Bayes, k-vecinos más cercanos (k-nearest

neighbors), entre muchos otros. La subárea 4.1 Inteligencia Artificial seguramente también contiene temas relacionados a la clasificación y algoritmos de aprendizaje de máquina. Se les recomienda que estudien árboles de decisión, redes neuronales y k-vecinos:

https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/ml-decision-tree/tutorial/
https://www.simplilearn.com/tutorials/deep-learning-tutorial/neural-network
https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

En la bibliografía de la guía viene un documento específico de clustering, por lo que podemos asumir que además de clasificación, también vendrá algo de este tema:

https://arxiv.org/ftp/arxiv/papers/1205/1205.1117.pdf

Otro enlace que les puede ayudar es:

https://training.galaxyproject.org/training-material/topics/statistics/tutorials/clustering_machinelearning/tutorial.html

Uno de los métodos más comunes de clustering es el jerárquico (hierarchical clustering), y una de las representaciones más usadas en este método es de tipo árbol y los dendrogramas:

https://www.statisticshowto.com/hierarchical-clustering/

El otro método común de clustering es el k-means, que no hay que confundir con el k-vecinos:

https://pythonprogramminglanguage.com/how-is-the-k-nearest-neighbor-algorithm-different-from-k-means-clustering/

NOTAS ACERCA DE LOS CONCEPTOS: La minería de datos y el aprendizaje de máquina contienen muchos nombres y conceptos que se utilizan más en inglés que en español, por lo que es importante conocer los diferentes nombres por los cuáles se les puede llamar a diferentes conceptos.