Tecnológico Nacional de México
Instituto Tecnológico de Tijuana

Subdirección Académica
Departamento de Sistemas y Computación

Semestre:
Febrero – Junio 2021

Carrera:
Ingeniería en Tecnologías de la Información y Comunicaciones

Materia y serie:
Minería de datos BDD-1703TI9A

Unidad a evaluar: Unidad I

Nombre de la Tarea:
What is data mining

Nombre del Alumno:
Hernández Negrete Juan Carlos 16212021
Manuel Javier Sifuentes Martinez 17212934

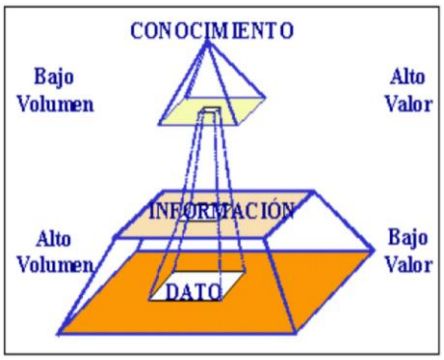
Nombre del docente:
José Christian Romero Hernández

Data Mining

Data mining is responsible for preparing, probing and exploring the data to extract the hidden and useful information in it. If the data is read and analyzed, they can provide, together, a true knowledge (future trends and behaviors) that helps in decision-making, since for the person in charge of a system, the data itself is not the most relevant, but the information that is enclosed in its relationships, fluctuations and dependencies.

It is known as data mining to a whole set of techniques responsible for the extraction of actionable knowledge through search programs and identification of global patterns and relationships, trends, deviations and other indicators, implicit in the databases, to discover, extract and store relevant information, thus making better decisions with greater knowledge. It is strongly linked to the supervision of industrial processes, as it is very useful to take advantage of the data stored in the databases.

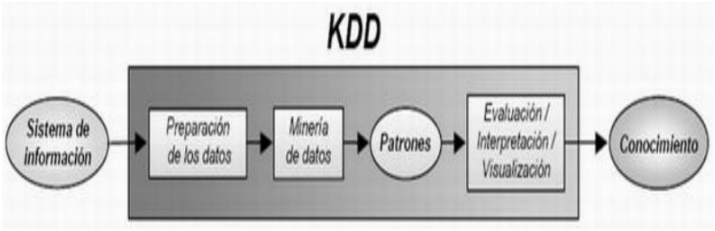
The bases of data mining are found in artificial intelligence, statistical analysis, Graphical Computing, Databases and Mass Processing. By using data mining techniques it is possible to solve problems of prediction, classification and segmentation.
Hierarchy that a database follows, between data and knowledge:



As the level of data is raised, the volume of data decreases, since the higher we are in the pyramid, the more specific and processed information is needed. Data mining works at higher levels looking for patterns, behaviors, sequences, trends or associations that can generate a model that allows us to better understand the business, through a combination of tasks such as: Data extraction, data cleaning, selection of characteristics, analysis of results, etc.

The term data mining is considered a stage within a larger process called the extraction or discovery of knowledge in databases (Knowledge Discovery in Databases or KDD), being located at the highest level of the evolution of the technological processes of data analysis.

Although some authors use the terms Data Mining and KDD interchangeably, as synonyms, there are clear differences between the two. KDD, as mentioned, is a process that consists of a set of phases, one of which is data mining, therefore the complete process that includes pre-processing, mining and post-processing of the data is called KDD.



Some of the most common tasks in KDD processes are classification and clustering, pattern recognition, predictions, and the detection of dependencies or relationships between data.

After the development previously shown about the concept of data mining and its function, the most common tasks related to this topic will be shown:

- **Classification:** classifies a data within one of the predefined categorical classes. It answers questions such as, What is the risk of granting a loan to this client? Given this new patient, what state of the disease does his analysis indicate?
- **Regression:** the purpose of this model is to match a data with a real value of a variable. Answer questions such as What is the sales forecast for next month? What does it depend on?
- **Clustering:** refers to the grouping of records, observations, or cases in similar object classes. A cluster is a collection of records that are similar to each other, and distinct from records in another cluster. How many types of customers come to a business? What profiles of needs exist in a certain group of patients?
- **Rule generation:** here rules are extracted or generated from the data. These rules refer to the discovery of association relationships and functional dependencies between the different attributes. How much must this indicator be worth in blood for a patient to be considered serious? If a customer of a hypermarket buys diapers, does he also buy beer?
- **Summary or summarization:** These models provide a compact description of a subset of data. What are the main characteristics of my clients?
- **Sequence analysis:** sequential patterns are modeled, such as time series analysis, gene sequences, and so on. The objective is to model the states of the process, or to extract and report deviation and trends over time. Is this month's electricity consumption similar to last year? Given the levels of air pollution for the last week, what is the forecast for the next 24 hours.

Data mining application areas:

- **Commerce and banking:** customer segmentation, sales forecast, risk analysis.
- **Medicine and Pharmacy:** diagnosis of diseases and the effectiveness of treatments.
- **Security and fraud detection:** facial recognition, biometric identifications, access to networks not allowed, etc.
- **Retrieval of non-numerical information:** text mining, web mining, image, video, voice and text search and identification from multimedia databases.
- **Astronomy:** identification of new stars and galaxies.
- **Geology, mining, agriculture and fishing:** identification of areas of use for different crops or fishing or mining exploitation in satellite image databases
- **Environmental Sciences:** identification of models of functioning of natural and / or artificial ecosystems (eg wastewater treatment plants) to improve their observation, management and / or control.
- **Social Sciences:** Study of the flows of public opinion. City planning: identify neighborhoods with conflict based on sociodemographic values.

Data mining is not statistics

Both terms are commonly confused, in fact, according to some, Data mining is the successor to statistics as it is currently used. Statistics and Data Mining have the same objective, which is to build compact and understandable "models" that account for the relationships established between the description of a situation and a result related to said description. However, there are clear differences between them.

Fundamentally, the difference between the two is that Data Mining techniques build the model automatically while "classical" statistical techniques need to be managed and guided by a professional statistician. Data Mining techniques allow you to gain both in performance and manageability and even in working time, the possibility of making your own models yourself without the need to contract or agree with a statistician, provides great freedom to professional users.

Data mining is not OLAP

OLAP tools allow you to quickly navigate through data, but no information is generated in the process. OLAP (On Line Analytical Processing) systems are those systems that must:

- Support complex analysis requirements
- Analyze data from different perspectives
- Withstand complex analysis against a huge volume of data

The functionality of OLAP systems is characterized by being a multidimensional analysis of data through user navigation through them in an assisted way.

References

Beltrán, B. (s.f) Minería de datos. March 10, 2021. Benemérita Universidad Autónoma de Puebla. Web site: http://bbeltran.cs.buap.mx/NotasMD.pdf

Riquelme, José., & Ruiz, R., & Gilbert, K. (2006) Minería de Datos: Conceptos y Tendencias. March 10, 2021. Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial, vol. 10, Web site: https://www.redalyc.org/pdf/925/92502902.pdf

Belichón, Y. (s.f) Minería de datos. March 10, 2021. Universidad Carlos III de Madrid, España. Web site: http://www.it.uc3m.es/jvillena/irc/practicas/10-11/15mem.pdf

Ángeles, M., & Santillán, A. (s.f) Minería de datos: Concepto, características, estructura y aplicaciones. March 10, 2021. Web site: http://www.ejournal.unam.mx/rca/190/RCA19007.pdf