**Temática:** Clase 6.

**Institución ofertante:** BA – Emprende D.

**Ponente:** Mg. Ing. Layla Scheli.



**Fundamentos Generales**

Como consecuencia de la generación masiva de datos, originados por el volumen asociado al Big Data, nos encontramos frente a un problema: la *infoxicación*, disponemos de tanta información, que a veces es imposible organizarla de manera efectiva. Por ello, la clave está en descubrir patrones o algoritmos para sacarle el máximo partido posible a los datos, y aquí es donde entra en juego el Data Mining (DM) o la Minería de Datos.

Pero entonces ¿qué es?, podemos definir a la Minería de Datos como un conjunto de técnicas de extracción de información orientada al negocio, vinculada al campo de estudio de la estadística y las ciencias de la computación.

Su sustento principal radica en el uso de diversos métodos asociados al campo de la inteligencia artificial, aprendizaje automático y sistemas de bases de datos, con la finalidad central de encontrar conocimiento *valioso* para el negocio dentro de ese *“mar inmenso de datos”*.

En términos generales podemos definir, que la Minería de Datos usa conceptos y herramientas de:

* Bases de datos.
* Carga y limpieza de información.
* Toma de decisiones.
* Aprendizaje automático.
* Probabilidad y estadísticas.
* Técnicas de visualización.
* Procesamiento en Paralelo.

Por su parte, todo *proyecto de Minería de Datos* suele cubrir las siguientes fases:

1. Filtrado de Datos.
2. Selección de Variables.
3. Extracción de Conocimiento.
4. Interpretación y Evaluación.

Filtrado de Datos: El formato de los datos contenidos en las fuentes de datos, nunca es el idóneo y la mayoría de las veces no es posible utilizar ningún algoritmo de Data Mining, de manera directa en un primer intento. Por lo tanto en esta etapa, se realiza una especie de preproceso, en el cual se filtran los datos (se eliminan valores incorrectos, no válidos, desconocidos, etc.), se obtienen muestras de los mismos (mayor velocidad de respuesta del proceso), o se reducen el número de valores posibles (mediante redondeo, agrupamiento, entre otros).

Selección de Variables: Aún después de haber filtrado los datos, es muy común seguir teniendo una cantidad ingente de datos. Por lo tanto, tenemos que elegir cuales son las variables más influyentes en el problema que estamos intentando de analizar, tratando siempre de sacrificar en la menor medida posible, la calidad del modelo de conocimiento obtenido del proceso de minería. Los métodos para la selección de características son dos:

1. Los basados en la elección de los mejores atributos del problema.
2. Los que buscan variables independientes mediante tests de sensibilidad, algoritmos de distancia o heurísticos.

Extracción de Conocimiento: Mediante una o distintas técnicas de Data Mining, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. Este paso consiste, como bien lo indica su nombre en obtener insights o inferencias relevantes sobre nuestro modelo.

Interpretación y Evaluación: Finalmente se procede a su validación, comprobando que las conclusiones son válidas y satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema. Si ninguno de los modelos alcanza los resultados esperados, se alterará alguno de los procesos anteriores en busca de nuevos modelos es decir entonces, que se genera un proceso de mejora continua.

**Técnicas de Minería de Datos**

Las técnicas de Data Mining pueden clasificarse como:

1) Descriptivas: Identifican patrones que explican o resumen aspectos de los datos, por ejemplo:

* + *Agrupamiento:* También conocido como clústering, es una técnica que consiste en la división de los datos, en grupos de objetos similares, llamados “clústers”.
  + *Reglas de asociación:* Son utilizadas para descubrir, hechos que ocurren en común dentro de un determinado conjunto de datos. Reflejan una implicación del tipo “si X entonces Y” (X⇒ Y).
  + *Correlaciones:* Nos permite determinar a través de un proceso de análisis, si dos variables se encuentran correlacionadas entre sí.

2) Predictivas: Tratan de inferir el comportamiento de una variable objetivo, a partir de una multiplicidad de atributos por ejemplo:

* + *Regresión*: Algoritmo utilizado, poder calcular el valor de una variable dependiente Y, en función de una variable independiente X. y = ax+b
  + *Árboles de decisión*: Representación gráfica en formato de árbol, de los posibles resultados de una serie de decisiones relacionadas.
  + *Redes neuronales:* Modelos computacionales, que procesan información imitando el funcionamiento de las neuronas biológicas de los seres humanos.

**Ventajas y Desventajas del Data Mining**



Los análisis de datos mediante técnicas de Data Mining, pueden aportar numerosas ventajas a las empresas para la optimización de sus procesos de gestión, tiempo, incluso hasta su productividad. De igual modo, podemos mencionar las mejoras que se pueden alcanzar para la captación y fidelización de clientes, que se traduce de manera directa en el incremento de las ventas. Algunas de las ventajas más relevantes a mencionar son:

* Permite descubrir información oculta que no esperábamos obtener, gracias a la implementación de diversos algoritmos que realizan múltiples combinaciones de datos distintas.
* Cualquier cantidad de datos puede ser analizados mediante la tecnología de Data Mining.
* Permite encontrar, atraer y retener clientes.
* La información obtenida, ayuda a la hora de tomar de decisiones tácticas y estratégicas dentro de la empresa u organización.
* Da a las empresas la posibilidad, de ofrecer a sus clientes los productos o servicios que necesitan de una manera mucho más personalizada.
* Las predicciones que se obtienen son confiables y válidas, ya que se fundamentan en análisis estadísticos basados en gran cantidad de datos.
* Ahorra costos empresariales y abre nuevas oportunidades de negocio.

Sin embargo, también puede aparecer algún inconveniente a la hora de utilizar técnicas de Data Mining, por ejemplo podemos mencionar que dependiendo del tipo de datos que se quiera recopilar, nos puede llevar mucho trabajo o esfuerzo durante el proceso de recolección, o a veces la inversión inicial para obtener las tecnologías necesarias para la recopilación de datos, puede tener un coste significativamente elevado.

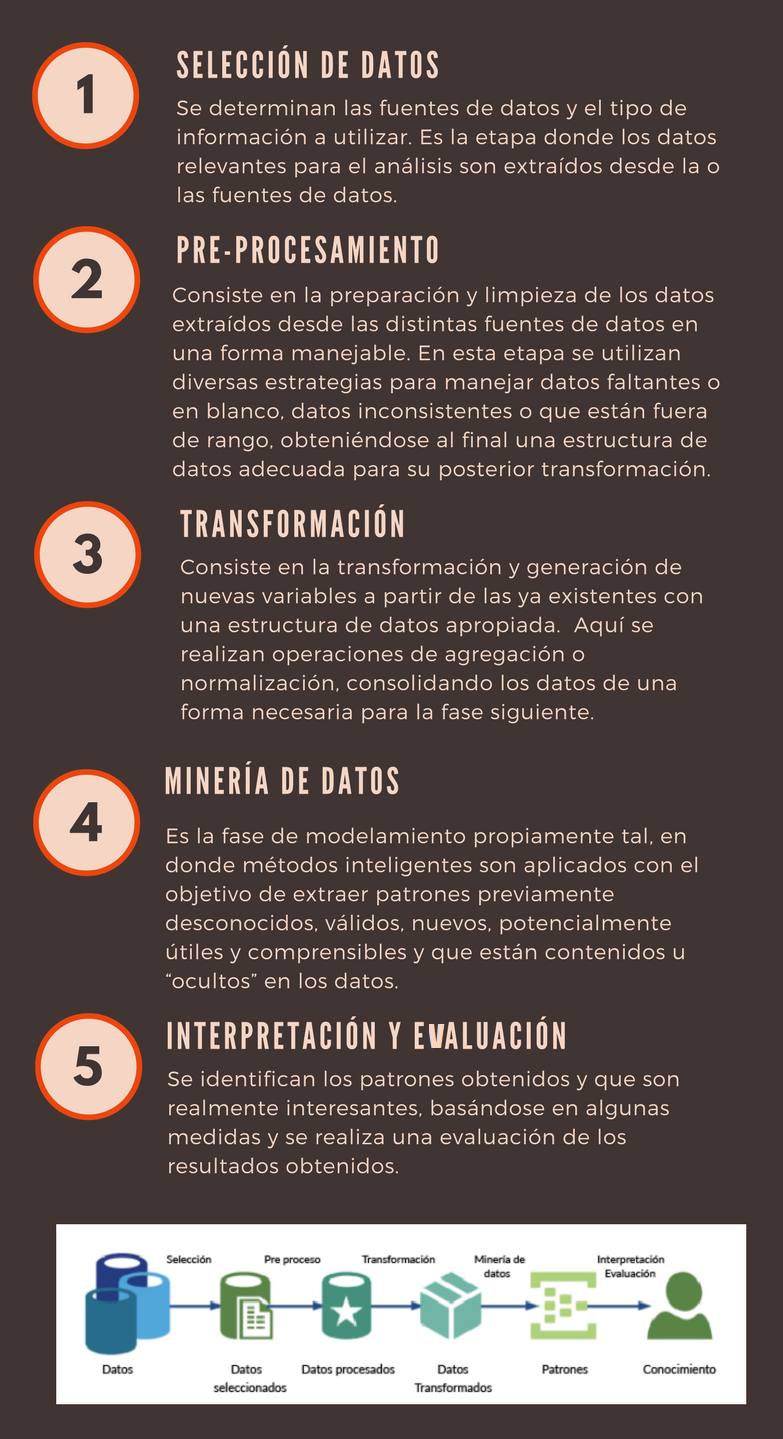
**Knowledge Discovery**

Lo primero que deberíamos preguntarnos es: *¿Qué es el conocimiento?,* Podemos pensarlo como la capacidad para describir con reglas simples fenómenos complejos. En el campo que nos ocupa los fenómenos complejos, se encuentran descriptos en grandes conjuntos de datos. El primer paso para descubrir reglas simples, es aplicar los criterios de la estadística descriptiva, promedios, dispersiones, percentiles, etc. Esto nos puede ayudar a sacar conclusiones muy elementales usando sólo el método de observación.

Una vez agotados estos medios y, en especial, frente a problemas donde el número de variables relevantes supera las dos o tres dimensiones, necesitamos una ayuda adicional para poder correlacionar distintas variables. En este punto, surge y adquieren total relevancia las herramientas informáticas, que nos permiten a través de técnicas de Minería de Datos encontrar una serie de “reglas” de diverso tipo, para alcanzar una mejor comprensión del problema a tratar. Uno de los problemas más comunes en el mundo de la Minería de Datos y también en el de Machine Learning, es que las herramientas utilizadas nos lleven a *reglas* que son en el fondo, un ajuste casual a la muestra de datos objeto de estudio, generando lo que se conoce en la jerga como "sobre ajuste" (overfitting) es decir, cuando el modelo que diseñamos “aprende” demasiado de los datos y no es extrapolable a otro conjunto de datos.

No obstante, resulta relevante mencionar que para luchar contra el “sobre ajuste”, tenemos dos remedios para esto. El primero es trabajar con varias muestras, inferir sobre una y luego asegurarse que los resultados se ajustan también a las otras. El segundo es más delicado ya que pasa por exigir que la regla tenga sentido dentro del marco teórico del problema. Este último tipo de verificación requiere que el analista entienda las dependencias entre las variables y no sea ajeno a la problemática de negocios que está abordando. Esto implica necesariamente, tener la mente abierta a la posibilidad de encontrar resultados que contradigan el sentido común, siendo justamente estos tipos de resultados los que suelen agregar más valor al negocio cuando pueden obviamente justificarse de manera adecuada.

**Fases del Proceso de KDD**



Como podemos observar en la imagen anterior, el proceso de Minería de Datos es un paso más dentro del proceso de Knowledge Discovery.

**Herramientas para implementar Data Mining**



Tal como nos podemos imaginar, existen múltiples alternativas disponibles en el mercado para realizar proyectos de Minería de Datos, entre las más relevantes podemos mencionar:



RapidMiner, antes conocida como YALE, siglas de “Yet Another Learning Environment”, es una herramienta open source muy popular en el sector. Destaca por permitir el acceso gratuito y por su fácil manejo dado que no requiere grandes conocimientos de programación. Está escrita en Java y contiene más de 500 operadores con diferentes enfoques para mostrar las conexiones en los datos: hay opciones para Data Mining, text mining o web mining, pero también análisis de sentimiento o minería de opinión. Asimismo, el programa puede importar tablas Excel, archivos SPSS y masas de datos de diferentes bases de datos e integra los programas de Data Mining como ser: WEKA y R.

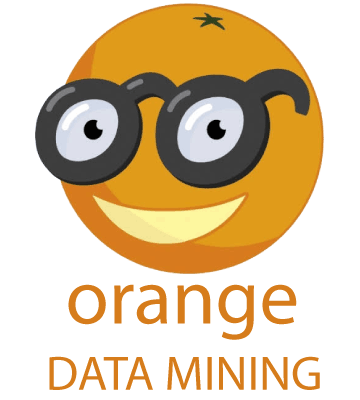


Es un software de código abierto basado en Java, compatible con Windows, MacOS y Linux y ofrece conexión a bases de datos SQL.

Las funciones que realiza son:

* Clasificación de datos, para lo que usa redes de neuronas artificiales.
* Análisis de clústeres.
* Árboles de decisión.
* Análisis de correlación o regresión.
* Algoritmos ID3 o C4.5.

Como desventaja WEKA no es tan fuerte en el análisis de clústeres.



El software de Data Mining [Orange](https://orange.biolab.si/), existe desde hace más de 20 años como proyecto de la Universidad de Liubliana. El núcleo del software se escribió en C++, para luego ampliarse al lenguaje de programación Python.Por su parte, las operaciones más complejas se llevan a cabo en C++, trabaja con operadores para la clasificación, regresión y clustering e integra una programación visual. Como ventaja principal, podemos mencionar que existe un sin número de tutoriales sobre la herramienta.



El software [KNIME](https://www.knime.com/knime-analytics-platform) fue desarrollado por la universidad de Constanza, originariamente se creó con un principio comercial, sin embargo se puso a disposición de los usuarios como software de código abierto y en la actualidad KNIME se considera una herramienta de gran popularidad entre la comunidad internacional de programadores. Como ventaja principal, esta herramienta permite descubrir estructuras ocultas de datos. Además, se pueden ampliar sus módulos con otras soluciones adicionales de pago.

Entre todas las funciones destaca el análisis de datos integrativo, también es muy eficiente en el proceso de extracción, transformación y carga de datos (ETL), así como en su tratamiento previo. Finalmente, es muy utilizado en campos de investigación farmacéutica, inteligencia empresarial y en el [sector financiero](https://www.gestion.org/prestamos-y-finanzas-online-la-mejor-forma-de-gestionar-el-dinero-es-a-traves-de-internet/).



SAS Enterprise Miner: Es un producto de SAS Institute, una de las mayores empresas privadas de software en todo el mundo. SAS, constituye el Data Mining tool principal en el análisis en el sector de los negocios y de hecho, se considera como el programa más adecuado para grandes empresas, aunque claramente con un costo significativamente mayor.

Podemos mencionar, entre sus principales ventajas que permite crear visualizaciones interactivas de datos y su tecnología es de las mejores en el mercado en lo que respecta a la realización de pronósticos. Además, también es un software escalable que permite aumentar progresivamente su eficiencia aumentando los recursos de hardware o de cualquier otro tipo.

**Cuadro comparativo entre Herramientas**



**Ejemplos de aplicación de Data Mining**

* Gobierno - El FBI analiza bases de datos comerciales para detectar posibles terroristas.
* Empresas - Detección de fraudes en las tarjetas de crédito.
* Empresas - Migración de clientes entre distintas compañías de comunicación móvil.
* Empresas - Predicción del tamaño de las audiencias televisivas para un programa determinado.
* Universidad - Conocer si los recién titulados llevan a cabo actividades profesionales relacionadas con sus estudios.
* Deporte - Sistema inteligente para prevenir lesiones en los jugadores, entre otros.

 **Conclusiones**

A modo de conclusión de esta charla, podemos decir entonces que la capacidad para almacenar datos ha crecido en los últimos años a velocidades exponenciales. En el otro extremo, la realidad es que nuestra capacidad para procesar esta enorme cantidad de datos para por utilizarlos eficazmente, no ha ido a la par sino todo lo contrario, cada vez tenemos más y más datos sin explotar de manera estratégica.

Por este motivo, el Data Mining se presenta como una tecnología de apoyo para explorar, analizar, comprender y aplicar el conocimiento obtenido, usando cantidades elevadas de datos. Dentro de este contexto, no podemos olvidarnos de destacar, que descubrir nuevos caminos que nos ayuden en la identificación de estructuras ocultas en los datos, es una de las tareas fundamentales del Data Mining.

Desde una perspectiva comercial, resulta interesante encontrar patrones ocultos de consumo en nuestros clientes, para poder explorar nuevos horizontes de desarrollo. Por ejemplo, predecir el comportamiento de un futuro cliente, basándose en los datos históricos de clientes que presentaron el mismo perfil, ayuda a poder retenerlo durante el mayor tiempo posible.

Las herramientas comerciales de Data Mining que existen actualmente en el mercado son muy variadas y excelentes en diversas aplicaciones. Las hay orientadas al estudio del web o al análisis de documentos o de clientes, mientras que otras son de uso más general. Su correcta elección depende de la necesidad de la empresa y de los objetivos a corto y largo plazo que se pretenda alcanzar. La decisión de seleccionar una solución de Data Mining, no es una tarea simple, es necesario siempre consultar a expertos en el área para seleccionar la herramienta más adecuada para el problema de la organización.

En resumen, el Data Mining se presenta como una tecnología emergente, con varias ventajas: por un lado, resulta un buen punto de encuentro entre los investigadores y las personas de negocios y por otro lado, ayuda a las empresas a ahorrar grandes sumas de dinero si es utilizada de manera estratégica.