FASE 2: RECOPILACIÓN DE LA INFORMACIÓN NECESARIA

Marco teórico

I. DataMine

El Data Mining es un conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera **automática** o **semiautomática**, con el objetivo de encontrar patrones repetitivos que expliquen el comportamiento de estos datos.

Su principal finalidad es explorar, mediante la utilización de distintas técnicas y tecnologías, **bases de datos** enormes con el objetivo de encontrar **patrones repetitivos**, tendencias o reglas que expliquen el comportamiento de los datos que se han ido recopilando con el tiempo. Estos patrones pueden encontrarse utilizando estadísticas o algoritmos de búsqueda próximos a la Inteligencia Artificial y a las redes neuronales. El objetivo general del proceso de minería de datos consiste en **extraer información** de un conjunto de datos y transformarla en una **estructura comprensible** para su uso posterior.

Los mineros o exploradores de datos a la hora de llevar a cabo un análisis de Data Mining, deberán realizar los siguientes pasos:

- **Selección del conjunto de datos:** tanto en lo que se refiere a las variables objetivo (aquellas que se quiere predecir, calcular o inferir), como a las variables independientes(las que sirven para hacer el cálculo o proceso), como posiblemente al muestreo de los registros disponibles.
- **Análisis de las propiedades de los datos:** elaboración de histogramas, diagramas de dispersión, presencia de valores atípicos y ausencia de datos (valores nulos).
- **Transformación del conjunto de datos de entrada:** se realiza con el objetivo de prepararlo para aplicar la técnica de minería de datos que mejor se adapte a los datos y al problema, a este paso también se le conoce como preprocesamiento de los datos.
- **Seleccionar y aplicar la técnica de minería de datos**: se construye el modelo predictivo, de clasificación o segmentación.
- **Extracción de conocimiento**: mediante una técnica de minería de datos, 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. También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un preprocesado diferente de los datos.
- **Interpretación y evaluación de datos:** una vez obtenido el modelo, se debe proceder a su validación comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias.

La tarea de minería de datos real es el análisis de grandes cantidades de datos para extraer patrones interesantes hasta ahora desconocidos, como los grupos de **registros de datos** (análisis clúster), **registros poco usuales** (la detección de anomalías) y **dependencias** (minería por reglas de asociación). Esto generalmente implica el uso de técnicas de bases de datos como los índices espaciales. Estos patrones pueden entonces ser vistos como una especie de resumen de los datos de entrada, y pueden ser utilizados en el análisis adicional.

II. Negocios y minería de datos

Análisis de la cesta de la compra:
Este análisis permite aumentar la venta de ciertos productos al conocer en qué épocas, dias , eventos o que tipo de consumidores buscan un determinado producto promoviendo la **venta compulsiva**. Un estudio muy citado detectó que los viernes había una cantidad inusualmente elevada de clientes que adquirían a la vez pañales y cerveza. Se detectó que se debía a que dicho día solían acudir al supermercado padres jóvenes cuya perspectiva para el fin de semana consistía en quedarse en casa cuidando de su hijo y viendo la televisión con una cerveza en la mano. El supermercado pudo incrementar sus ventas de cerveza colocándolas próximas a los pañales.

Patrones de fuga:
La minería de datos ayuda a determinar qué clientes son los que tienen mayor probabilidad a darse de baja (Cambiarse a la competencia) estudiando sus patrones de comportamiento y comparándolos con muestras de clientes que, efectivamente, se dieron de baja en el pasado. En muchas industrias existe un comprensible interés en detectar cuanto antes aquellos clientes para retenerlos con ofertas personalizadas,promociones especiales, etc.

Fraudes:
Generalmente, estas las prácticas fraudulentas o ilegales suelen seguir patrones característicos que permiten, con cierto grado de probabilidad, distinguirlas de las legítimas y desarrollar así mecanismos para tomar medidas rápidas frente a ellas.

Recursos humanos:
La minería de datos también puede ser útil para los departamentos de recursos humanos en la identificación de las características de sus empleados de mayor éxito. La información obtenida puede ayudar a la contratación de personal, centrándose en los esfuerzos de sus empleados y los resultados obtenidos por estos.

III. Técnicas de minería de datos

Las técnicas de la minería de datos provienen de la **inteligencia artificial** y de la **estadística**, las técnicas más comunes son las siguientes:

Redes neuronale:
Son un **paradigma de aprendizaje** y procesamiento automático inspirado en la forma en que funciona el **sistema nervioso** de los animales. Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida.

Regresión lineal:
En estadística la regresión lineal o ajuste linea**l** es un **modelo matemático** usado para aproximar la relación de dependencia entre una variable dependiente Y, las variables independientes X_i y un término aleatorio ε.

Árboles de decisión:
Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la **inteligencia artificial** y el **análisis predictivo**, dada una base de datos se construyen estos diagramas de construcciones **lógicas**, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema.

-Algoritmo ID3.
-Algoritmo C4.5

Agrupamiento o Clustering:
Es un procedimiento de agrupación de una serie de **vectores** de acuerdo con un criterio. Esos criterios son por lo general **distancia** o **similitud**. La cercanía se define en términos de una determinada función de distancia. La medida más utilizada para medir la similitud entre los casos es la **matriz de correlación** entre los nxn casos. Sin embargo, también existen muchos algoritmos que se basan en la maximización de una propiedad estadística llamada **verosimilitud.** Algunos de los algoritmos de agrupamiento son:

-Algoritmo K-means
-Algoritmo K-medoids

Modelos estadísticos:
Un modelo estadístico es un tipo de **modelo matemático** que usa la **probabilidad**, y que incluye un conjunto de asunciones sobre la generación de algunos datos muestrales, de tal manera que asemejen a los datos de una población mayor.

IV. Base de datos

Una base de datos es un conjunto de datos pertenecientes a un mismo contexto y almacenados sistemáticamente para su posterior uso. En este sentido; una biblioteca puede considerarse una base de datos compuesta en su mayoría por documentos y textos impresos en papel e indexados para su consulta. Actualmente, y debido al desarrollo tecnológico de campos como la informática y la electrónica, la mayoría de las bases de datos están en formato digital, siendo este un componente electrónico, por tanto se ha desarrollado y se ofrece un amplio rango de soluciones al problema del almacenamiento de datos.
Desde el punto de vista informático, la base de datos es un sistema formado por un conjunto de datos almacenados en discos que permiten el acceso directo a ellos y un conjunto de programas que manipulen ese conjunto de datos.
Cada base de datos se compone de una o más tablas que guarda un conjunto de datos. Cada tabla tiene una o más **columnas** y **filas**. Las columnas guardan una parte de la información sobre cada elemento que queramos guardar en la tabla, cada fila de la tabla conforma un registro.

V. Gestion de datos

La definición oficial suministrada por la Data Management Association (DAMA) es "La Gestión de Datos es el desarrollo y ejecución de arquitecturas, políticas, prácticas y procedimientos que gestionan apropiadamente las necesidades del ciclo de vida completo de los datos de un estudio".

Un Sistema Gestor de Bases de Datos (SGBD) o DGBA (Data Base Management System) es un conjunto de **programas** no visibles que **administran y gestionan la información** que contiene una **base de datos**. A través de él se maneja todo acceso a la base de datos con el objetivo de servir de interfaz entre ésta, el usuario y las aplicaciones.

Gracias a este sistema de software invisible para el usuario final, compuesto por un lenguaje de definición de datos, un lenguaje de manipulación y de consulta, es posible gestionar los datos a distintos niveles. Tanto **almacenar, modificar y acceder** a la información como realizar **consultas y hacer análisis** para generar informes.

VI. Big Data

Cuando hablamos de Big Data nos referimos a conjuntos de datos o combinaciones de conjuntos de datos cuyo tamaño (volumen), complejidad (variabilidad) y velocidad de crecimiento (velocidad) dificultan su captura, gestión, procesamiento o análisis mediante tecnologías y herramientas convencionales, tales como bases de datos relacionales y estadísticas convencionales o paquetes de visualización, dentro del tiempo necesario para que sean útiles.

Bibliografía

---

[1] Algoritmos de minería de datos (Analysis Services: minería de datos). (2018). Recuperado de https://docs.microsoft.com
[2] Test Run - Asociación de aprendizaje de la regla. (2018). Recuperado de https://msdn.microsoft.com
[3] Algoritmo de árboles de decisión de Microsoft. (2018). Recuperado de https://msdn.microsoft.com
[4] Datamine. (2018). Recuperado de http://cursosgeomin.com
[5] Base de datos. (2018). Recuperado de https://es.wikipedia.org
[6] Minería de datos. (2018). Recuperado de https://es.wikipedia.org
[7] Minería de datos: cómo funciona, elementos y requisitos. (2018). Recuperado de https://blog.es.logicalis.com