FASE 2: RECOPILACIÓN DE LA INFORMACIÓN NECESARIA

Marco teórico

Allers Group

Son una empresa Colombiana de origen Alemán fundada en el año 1955, especializada en importación, ventas al mayor y al detal; de equipos médicos, insumos hospitalarios, instrumental quirúrgico y medicamentos. Allers Group cuenta con marcas propias y exclusivas de compañías líderes en el mundo además de un amplio reconocimiento por más de 5 décadas en el sur Occidente Colombiano, Siendo Pioneros en la distribución de productos de salud.

I. Big Data

Es un concepto que hace referencia a un **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 aplicaciones informáticas tradicionales de procesamiento de datos.

Los macrodatos se pueden describir por las siguientes características:

- **Volumen:** la cantidad de datos generados y guardado. El tamaño de los datos determina el valor y entendimiento potencial, y si los puede considerar como auténticos macrodatos.
- **Variedad:** el tipo y naturaleza de los datos para ayudar a las personas a analizar los datos y usar los resultados de forma eficaz. Los macrodatos usan textos imágenes, audio y vídeo. También completan pedazos pedidos a través de la fusión de datos.
- **Velocidad:** en este contexto, la velocidad a la cual se generan y procesan los datos para cumplir las exigencias y desafíos de su análisis.
- **Veracidad:** la calidad de los datos capturados puede variar mucho y así afectar a los resultados del análisis.

El análisis de Big Data ayuda a las organizaciones a aprovechar sus datos y utilizarlos para identificar nuevas oportunidades. Eso, a su vez, conduce a movimientos de negocios más inteligentes, operaciones más eficientes, mayores ganancias y clientes más felices. Las empresas con más éxito con Big Data consiguen valor de las siguientes formas:

- **Reducción de coste**. Las grandes tecnologías de datos, como Hadoop y el análisis basado en la nube, aportan importantes ventajas en términos de costes cuando se trata de almacenar grandes cantidades de datos, además de identificar maneras más eficientes de hacer negocios.
- **Más rápido, mejor toma de decisiones**. Con la velocidad de Hadoop y la analítica en memoria, combinada con la capacidad de analizar nuevas fuentes de datos, las empresas pueden analizar la información inmediatamente y tomar decisiones basadas en lo que han aprendido.
- **Nuevos productos y servicios**. Con la capacidad de medir las necesidades de los clientes y la satisfacción a través de análisis viene el poder de dar a los clientes lo que quieren. Con la analítica de Big Data, más empresas están creando nuevos productos para satisfacer las necesidades de los clientes.

La recopilación de grandes cantidades de datos y la búsqueda de **tendencias** dentro de los datos permiten que las empresas se muevan mucho más rápidamente, sin problemas y de manera eficiente. También les permite eliminar las áreas problemáticas antes de que los problemas acaben con sus beneficios o su reputación.

II. Técnicas de análisis de datos

Teniendo los datos necesarios almacenados, es necesario utilizar distintas técnicas de análisis como:

- **Asociación**:permite encontrar relaciones entre diferentes variables. Bajo la premisa de causalidad, se pretende encontrar una predicción en el comportamiento de otras variables. Estas relaciones pueden ser los sistemas de ventas cruzadas en los comercios electrónicos.
- **Minería de datos (*data mining*)**:Tiene como objetivo encontrar comportamientos predictivos. Engloba el conjunto de técnicas que combina métodos estadísticos y de aprendizaje automático con almacenamiento en bases de datos.
- **Agrupación (*clustering*)**:el análisis de clústeres es un tipo de minería de datos que divide grandes grupos de individuos en grupos más pequeños de los cuales no conocíamos su parecido antes del análisis. El propósito es encontrar similitudes entre estos grupos, y el descubrimiento de nuevos, conociendo cuáles son las cualidades que lo definen.
- **Análisis de texto (*text analytics*):**gran parte de los datos generados por las personas son textos, como correos, búsquedas web o contenidos. Esta metodología permite extraer información de estos datos y así modelar temas y asuntos o predecir palabras.

III. DataMining

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 principal de la minería de datos 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.

IV. Negocios y minería de datos

Analizar la información reunida en una empresa es por lo menos tan importante como recolectarla. Por lo tanto, es importante tener inteligencia de negocios (BI). Esto es, la capacidad de transformar los datos en información y posteriormente en conocimiento. Es la mejor manera de optimizar el proceso de toma de decisiones en los negocios.

Los principales usos de la minería de datos en las empresas son:

- **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.

V. 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 lineal 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 se acerquen a los datos de una población mayor.

VI. Algoritmos Apriori

El algoritmo a priori es un algoritmo utilizado en minería de datos, sobre bases de datos transaccionales, que permite encontrar de forma eficiente "conjuntos de ítems frecuentes", los cuales sirven de base para generar reglas de asociación. Procede identificando los ítems individuales frecuentes en la base y extendiéndose a conjuntos de mayor tamaño siempre y cuando esos conjuntos de datos aparezcan suficientemente seguidos en dicha base de datos. Este algoritmo se ha aplicado grandemente en el análisis de transacciones comerciales y en problemas de predicción.

El proceso de generación de candidatos en el algoritmo apriori genera un número grande de subconjuntos. Exploración de conjuntos de forma bottom-up encuentra cualquier subconjunto maximal solo después de todos los (2^S)-1 de sus subconjuntos propios. Algoritmos posteriores como Max-Miner trata de identificar el conjunto maximal de ítems frecuentes sin enumerar sus subconjuntos, ejecutan "saltos" en el espacio de búsqueda en vez de una estrategia puramente bottom-up.

VII. Reglas de asociación

En minería de datos y aprendizaje automático, las reglas de asociación se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos. Se han investigado ampliamente diversos métodos para aprendizaje de reglas de asociación que han resultado ser muy interesantes para descubrir relaciones entre variables en grandes conjuntos de datos.

VIII. 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.

IX. 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.

X. ERP (Enterprise Resource Planning)

Es un conjunto de sistemas de información que permite la integración de ciertas operaciones de una empresa, especialmente las que tienen que ver con la producción, la logística, el inventario, los envíos y la contabilidad.

El propósito de un software ERP es apoyar a los clientes de la empresa, dar tiempos rápidos de respuesta a sus problemas, así como un eficiente manejo de información que permita la toma de decisiones y minimizar los costes.

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