Leonardo Franco Calpa

Bryan Camilo Grueso

1. **Recopilación de información necesaria**

Descubrir conocimiento en grandes datos.

Teniendo en cuenta que la empresa Allers Group ha ido almacenando información durante quince años de sus ventas, clientes, proveedores, entre otros, ayudándose en un software de gestión empresarial (SAP) y que a toda esa información almacenada durante todo ese tiempo se desea sacarle provecho para impulsar la empresa, hay uno puntos a tener en cuenta. Tamaño de la información recaudada; forma en la que está almacenada la información: formato, estructura, organización; si hay que reestructurarla para un posterior modelamiento; qué herramientas son mejores de acuerdo con el volumen de la información…

Como lo dio a conocer Oscar Rendón, CTO de Allers Group, se estima que se hacen aproximadamente entre 300 a 600 registros de ventas por día, sin contar la información generada por otras áreas de la empresa. Con base en lo anterior, podría decirse que la información anualmente puede alcanzar registros del orden de 15x104, y contando desde 15 años atrás, el total de la información puede estar del orden de 2x106 registros, los cuales pueden presentar muchas dimensiones en la base de datos donde están alojados.

Como gran parte de los datos hacen referencia a la información de los clientes, proveedores, fechas, cantidades, es muy probable que se encuentre en formato de texto, para lo cual habría que darle un tratamiento a la información para estructurarla y así poder clasificarla. También como la información ha sido recolectada por medio programa de gestión empresarial SAP puede que no haya información duplicada que aumente innecesariamente el tamaño de la base de datos.

Para lograr sacar utilidad de toda la información es necesario llevar a cabo metodologías que permitan transformar simples datos en indicadores para realizar predicciones, explicaciones y clasificaciones se requiere de metodologías de análisis de datos, KDD (*Knowledge Discovery in Databases*).

La metodología de KDD consiste principalmente en un proceso para identificar patrones en los datos, que sean de gran utilidad para finalidad que se desea alcanzar, para de ese modo convertir información de bajo nivel el conocimiento de alto nivel. Como lo explican en el artículo “Minería de Datos: Conceptos y tendencias” el KDD consta de:

1.Comprender el dominio de aplicación: este paso incluye el conocimiento relevante previo y las metas de la aplicación.

2.Extraer la base de datos objetivo: recogida de los datos, evaluar la calidad de los datos y utilizar análisis exploratorio de los datos para familiarizarse con ellos. 3.Preparar los datos: incluye limpieza, transformación, integración y reducción de datos. Se intenta mejorar la calidad de los datos a la vez que disminuir el tiempo requerido por el algoritmo de aprendizaje aplicado posteriormente.

4.Minería de datos: como se ha señalado anteriormente, este es la fase fundamental del proceso. Está constituido por una o más de las siguientes funciones, clasificación, regresión, clustering, resumen, recuperación de imágenes, extracción de reglas, etc.

5.Interpretación: explicar los patrones descubiertos, así como la posibilidad de visualizarlos.

6.Utilizar el conocimiento descubierto: hacer uso del modelo creado

(Riquelme, Ruiz, Gilbert, 2006)

Minería de Datos.

En la minería de datos se intenta hallar patrones con mucha significancia en los conjuntos de datos para poder construir modelos predictivos, descriptivos y clasificadores. Para eso se utiliza técnicas de aprendizaje automático (*Machine Learning*) y programación en bases de datos. En ***ML*** existen dos grande s grupos de algoritmos que se usan para abstraer la información*: aprendizaje supervisado* y *no supervisado*, donde en ambos reciben como entrada un *data set* y producen como salida un modelo de los ya mencionados.

Para a generación de un modelo basado en ***MD*** puede seguirse una serie de pasos que se describen en el articulo *Introducción a la minería de datos y el data warehousing* por Coria:

La generación de un modelo basado en MD consiste en los siguientes pasos: 1) definición del objetivo del modelo, 2) selección de datos para análisis y modelación y de sus fuentes, 3) recolección, limpieza y pre-procesamiento de datos, 4) análisis estadísticos básicos, 5) selección y aplicación de algoritmos de aprendizaje automático, 6) reporte y evaluación de hallazgos con el experto de dominio, 7) explotación de los hallazgos.

(Coría, 2016)