Universidad
Alonso de Ojeda
UNIOJEDA

Revista ETHOS Venezolana

Vol. 9 No. 2, Julio - Diciembre 2017

ISSN: 1856-9862
Depósito legal: pp 200902ZU3258

Revista Ethos Venezolana
Vol. 9 No. 2, Julio-Diciembre 2017, 123 - 133
ISSN 1856-9862 ~ Depósito legal pp 200902ZU3258

Metodología para la ejecución de proyectos de minería de datos

Alfredo J. Díaz-Pérez*

Resumen

El objetivo del presente artículo es proponer una metodología para la ejecución de proyectos basados en minería de datos, especificando los pasos a seguir para obtener resultados que ayuden tanto a la solución de problemas como la toma de decisiones. La metodología se catalogó como un proyecto factible y documental; para lo cual se realizó una revisión de las teorías de Maimon y Rokach (2010), Korth *et al* (2011) y Pérez (2014). A tale fines se obtuvieron seis fases que parten de la comprensión de las reglas del negocio, formulación de la problemática, selección del conjunto de datos, diseño y construcción del modelo operativo, aplicación del proceso de minería de datos, análisis, interpretación y difusión de los resultados.

Palabras clave: metodología, minería de datos, ejecución de proyectos.

* Ingeniero en Informática (URBE, 2005), Magíster en Gerencia de Recursos Humanos (URBE, 2008). Doctor en Ciencias de la Educación (URBE, 2017). Contacto: alfredojodp@hotmail.com

Recibido: 20-04-2017 Aceptado: 28-07-2017

Methodology for making data mining projects

Abstract

The purpose of this article is to propose a methodology for the execution of data mining projects, specifying the steps to follow to get results that can help to problem solving and decision making. In order to define the methodology, a review of the theories of Maimon and Rokach (2010), Korth et al (2011) and Pérez (2014) was made and it was classified as a feasible project and documentary research. Thus it was structured by six (6) phases that include: understanding of business rules, problem formulation, data set selection, design and construction of the operating model, application of the data mining process, analysis, interpretation and diffusion of results.

Keywords: data mining, methodology, projects execution.

Introducción

En los últimos años, los sistemas de información han crecido notablemente. Están enfocados no solamente al procesamiento o automatización de transacciones rutinarias, sino que permiten generar salidas significativas como reportes o consultas que brindan apoyo a la gerencia de las organizaciones en la toma de decisiones. Sin embargo, existen casos donde se cuenta con grandes volúmenes de información, la cual puede cobrar mayor significado cuando se procesa con el propósito de obtener resultados de segundo nivel, tales como estadísticas, proyecciones o tendencias. Es allí cuando se habla de producir conocimiento a partir de los datos almacenados, lo que recibe el nombre de minería de datos.

En ese sentido, Korth *et al* (2011) exponen que la minería de datos (*data mining*) es el proceso de detectar la información procesable de los conjuntos grandes de datos. Así mismo, utiliza el análisis matemático y lógico para deducir los patrones de comportamiento o tendencias que existen en los mismos. Con base en lo anterior, esta actividad tiene como beneficio la utilización de los datos almacenados para lograr mayores niveles de significancia, elevando la confiabilidad y calidad de los resultados en virtud de apoyar una toma de decisiones asertiva o estimar comportamientos de fenómenos futuros.

Ahora bien, al momento de emprender proyectos de minería de datos, suele ocurrir que las organizaciones recurren a metodologías clásicas

para el desarrollo de sistemas o aplicaciones informáticas orientadas a la solución de problemas. Sin embargo, se generan situaciones donde no se profundiza en los procedimientos para realizar inferencias, cálculos o estimaciones en virtud de producir los resultados esperados. De este modo, se requiere de una secuencia lógica de pasos a seguir para lograr los objetivos propuestos, los cuales deben contemplar fases pertinentes con el estudio profundo de los datos, sin forzar procedimientos destinados a otros productos tecnológicos.

Sobre la base de los postulados de Korth *et al* (2011), se puede afirmar que de continuar aplicando metodologías cuyo objetivo no está orientado hacia la minería de datos cuando los productos tecnológicos hacen uso de estas técnicas de obtención de información, se pudiesen producir resultados poco confiables, inexactos o sin la pertinencia requerida por las organizaciones a efectos de tomar decisiones estratégicas. De allí, la necesidad de un ordenamiento de fases conducentes a la ejecución de proyectos vinculados con la minería de datos.

Con base en lo anterior, esta investigación tiene como objetivo proponer una metodología orientada a la ejecución de proyectos de minería de datos que abarque las etapas iniciales de identificación de las reglas del negocio, selección del conjunto de datos, construcción de modelos o rutinas de análisis, obtención y difusión de los resultados. Para dar cumplimiento al objetivo general, se realiza una selección bibliográfica relacionada con la temática; luego, se efectúa una revisión de teorías preexistentes para encontrar aspectos convergentes o divergentes entre los distintos aportes y, finalmente, se formulan las etapas en las cuales está dividida la metodología.

Fundamentación teórica

El soporte teórico de la metodología propuesta, se apoya en Maimon y Rokach (2010), Korth *et al* (2011) y Pérez (2014), quienes han estudiado las aplicaciones de la minería de datos en distintos contextos. En ese sentido, exponen aspectos tanto conceptuales como procedimentales para el abordaje, análisis y resolución de problemas que involucran la minería de datos.

En primer lugar, Maimon y Rokach (2010) plantean una serie de fases que se deben cumplir para solucionar un problema de minería de datos, entre las que se encuentran: selección de un conjunto de datos, análisis de sus propiedades, transformación del conjunto de datos de entrada,

selección y aplicación de técnicas de minería, extracción de conocimiento e interpretación de los resultados. *En segundo lugar,* Korth *et al* (2011) expone que primeramente, deben conocerse las reglas del negocio, luego, realizar un diagnóstico del problema a solucionar, seleccionar el conjunto de datos que será objeto de estudio, diseñar rutinas de análisis e inferencia sobre los datos, codificación de programas, producción de resultados y presentación de los mismos.

Por otra parte, Pérez (2014) expone que al resolver un problema mediante minería de datos, se debe definir el mismo, considerando las variables implícitas dentro del negocio, establecer las salidas que se desean obtener (reportes, consultas, proyecciones, estimaciones), definir los procedimientos para compilar o descompilar la información, construir los programas requeridos para realizar la producción de resultados y, finalmente, realizar la corrida de estos a efectos de generar las salidas requeridas para su análisis por parte de los niveles estratégicos de la organización.

Aspectos metodológicos

En lo concerniente al aspecto metodológico y considerando la naturaleza del objetivo planteado, esta investigación se catalogó como un proyecto factible, de acuerdo con lo expuesto por Hernández *et al* (2014) quienes indican que estos buscan el desarrollo de un modelo operativo susceptible de ser implantado para satisfacer las necesidades de una organización. En ese sentido, se propone una metodología para la ejecución de proyectos basados en minería de datos que puede ser utilizada por analistas de sistemas, desarrolladores de *software* o cualquier profesional del área de informática en virtud de resolver problemas inherentes a la obtención de conocimiento a partir de los datos.

En ese propósito, el estudio se tipifica como documental, ya que se realiza una revisión bibliográfica de distintos autores, los cuales están relacionados con la minería de datos. Al respecto, Hernández *et al* (2014) establecen que estos se centran en estudiar hechos existentes en la realidad utilizando como fuentes de información todo tipo de documentos accesibles para el investigador. Como complemento a lo anterior, se indica que las unidades de análisis utilizadas para esta investigación fueron los postulados de Maimon y Rokach (2010), Korth *et al* (2011) y Pérez (2014) sobre metodologías o pasos a seguir para aplicar la minería de datos en la resolución de problemas.

Teorías que sustentan la formulación de la metodología propuesta

Primeramente, se ha considerado la metodología propuesta por Maimon y Rokach (2010), quienes plantean una fase de selección del conjunto de datos, y que consiste en elegir de acuerdo a criterios de inclusión establecidos por el investigador, cuáles son las variables objetivo, es decir, aquellas sobre las que se realizará inferencia o cálculo. Seguidamente, el autor expone una fase de análisis de las propiedades de los datos en la que se revisa la naturaleza de los mismos, sus tipos y metadatos a fin de determinar el procedimiento inferencial a seguir.

Sobre la base de lo anteriormente expuesto, Pérez (2014) indica que se debe partir de un análisis de las reglas del negocio para comprender el objetivo de la organización y, luego, comprender cuáles son los requerimientos del nivel estratégico, respondiendo a interrogantes sobre lo que se desea investigar y para qué se realizará este proceso. Posteriormente, se procede a realizar una revisión de los datos existentes para determinar si es factible el trabajo de minería de datos.

A la luz de lo anterior, se encuentran diferencias entre Maimon y Rokach (2010) y Pérez (2014), los primeros parten de la selección de un conjunto de datos, mientras el segundo se enfoca en comprender las reglas del negocio antes de comenzar a trabajar sobre la data. Sin embargo, coinciden en la importancia de la selección de grupos de datos adecuados, ya que serán la materia prima del proyecto. Por consiguiente, se fija posición al formular una etapa de reconocimiento de las reglas del negocio, objetivos y finalidad, así como el levantamiento de información para conocer los requerimientos a efectos de dar respuesta a las interrogantes mediante el modelo de minería de datos.

Posteriormente, Maimon y Rokach (2010), añaden una fase de transformación del conjunto de datos de entrada, en la que se organizan los mismos en grupos para su procesamiento y se establecen los procedimientos para la obtención de resultados. Es de hacer notar que pueden existir diversos procedimientos, rutinas o algoritmos que deban ejecutarse antes de obtener una salida comprensible.

Como complemento a lo anterior, Korth *et al* (2011) indican que para procesar grandes volúmenes de datos, es necesario agruparlos de acuerdo a criterios válidos tales como tipo o significado de los mismos,

homogeneidad o fuentes de obtención; luego, dependiendo de estos, deberán procesarse a efectos de adquirir significado y ayuden a la resolución de problemas en las organizaciones.

A la luz de las ideas expuestas, se encuentran similitudes entre Maimon y Rokach (2010) y Korth *et al* (2011), quienes coinciden en agrupar los datos de acuerdo con criterios como su naturaleza, el aplicativo a utilizar para la minería o el interés del investigador. Cabe destacar que, dentro de un mismo proyecto, pueden utilizarse distintas aplicaciones de *software* o programas, pero no todos reciben las entradas de la misma manera, de allí la importancia de su correcta agrupación.

Por ende, en virtud de todo lo anterior se puede proponer una fase de preparación donde los datos sean tratados para introducirlos dentro de los programas o rutinas destinadas a generar los resultados. De otra manera, la heterogeneidad o falta de estructuración de los mismos podría generar resultados inconsistentes o poco confiables.

En este orden de ideas, Maimon y Rokach (2010), exponen una fase de aplicación de técnicas de minería de datos, esto es, una vez que se han organizado los datos en grupos de entrada, se procede a diseñar y crear las rutinas o procedimientos para obtener los resultados. Esto debe responder a los modelos predictivos, inferenciales o de clasificación previamente elegidos.

Asimismo, Korth *et al* (2011) coinciden en desarrollar y probar programas o rutinas para el tratamiento de los datos. De hecho, existen iniciativas como el PMML (Lenguaje de marcación de modelos predictivos), que busca la interoperabilidad de los modelos entre distintas plataformas. En efecto, pueden utilizarse los lenguajes de programación estandarizados, pero también implementaciones de *applets* o complementos enfocados a la minería de datos.

Sobre la base de lo expuesto, se encuentran similitudes entre Maimon y Rokach (2010) y Korth *et al* (2011), quienes indican que es necesario el desarrollo de rutinas o programas orientados al procesamiento de los conjuntos de datos para la obtención de resultados. Igualmente, coinciden en la aplicación de lenguajes estandarizados o complementos diseñados específicamente para labores de minería. En efecto, esto es vital al considerar una fase de diseño y construcción de rutinas para la aplicación de la minería de datos.

Seguidamente, Maimon y Rokach (2010), exponen una fase de extracción de conocimientos, que consiste en buscar patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre las mismas. Igualmente, pueden utilizarse técnicas variadas para generar diversos resultados, aunque depende de la naturaleza de estos, es decir, requieren de un preprocesamiento diferente dependiendo de cada caso. Igualmente, Pérez (2014) expone que, a partir de la ejecución de los procedimientos de minería de datos, se requiere realizar un análisis inferencial acerca de los patrones de comportamiento encontrados en la corrida de los programas.

En ese sentido, Maimon y Rokach (2010) presentan similitudes con las ideas de Pérez (2014) cuando plantean la búsqueda de patrones de comportamiento, lo cual está relacionado con la formulación de inferencias, es decir, primero se tratan de determinar algunas tendencias o esquemas generados por los datos y, posteriormente, se realizan deducciones a efectos de otorgar significado a los resultados. Sin embargo, un aspecto diferenciador entre los autores mencionados, es la necesidad de tratar la data de acuerdo con su naturaleza, ya que puede proceder de distintas fuentes y presentar distinta morfología.

Como última fase, Maimon y Rokach (2010) manifiestan la necesidad de interpretar y evaluar los resultados. En el caso de haber obtenido varios patrones o tipologías de hallazgos, se deben comparar en busca de aquellos que se ajusten mejor al problema. Si ninguno de los resultados ofrece un comportamiento adecuado, debe alterarse alguno de los pasos anteriores para volver a generarlos.

Finalmente, Pérez (2014) indica que la última fase en un proceso de minería de datos es la interpretación y difusión de los resultados, abarcando: la comprensión, elaboración de conclusiones y explicar los comportamientos encontrados a los entes interesados del negocio. Al punto, se proveerán las respuestas requeridas. En tal sentido, en concordancia con los criterios anteriores, será necesaria la divulgación de los resultados en virtud de proveer la información necesaria para la toma de decisiones.

Como complemento a lo anterior, tanto Pérez (2014) como Maimon y Rokach (2010) coinciden en que debe otorgarse significado a los hallazgos obtenidos, al punto de dar respuesta a las interrogantes planteadas al inicio del proyecto. Igualmente, deben ser comunicados a las partes interesadas a efectos de apoyar la toma de decisiones o resolver problemas.

Metodología propuesta

En este apartado se presenta la estructura de fases o etapas que conforman la metodología para la ejecución de proyectos de minería de datos, la cual surge de las propuestas de Maimon y Rokach (2010), Korth *et al.* (2011) y Pérez (2014).

Primera fase: comprensión de las reglas del negocio: Consiste en entender a profundidad la naturaleza de la organización, objetivos y actividades medulares con el fin de internalizar cuáles son los procesos que se ejecutan en virtud de familiarizarse con la misma. Abarca un diagnóstico de la situación actual para determinar cuáles son las factibilidades, fortalezas, oportunidades, debilidades y amenazas para emprender el proyecto de minería de datos. Para ello, se recomienda realizar mesas de trabajo, reuniones con las distintas áreas involucradas y levantamiento de procesos.

Segunda fase: formulación de la problemática: se basa en formular claramente cuál es el problema que se desea solucionar; determinar las variables a investigar y establecer las interrogantes por responder mediante el trabajo de minería. En tal sentido, todas las partes involucradas deben estar en cuenta de las tareas a realizar y, para ello, se pueden levantar pliegos de requerimientos tanto funcionales como no funcionales para clarificar lo que se desea resolver.

Tercera fase: selección del conjunto de datos: consiste en seleccionar el grupo o conjunto de datos sobre la cual aplicará el proceso de minería. Pueden estar integrados o separados. Implica un ordenamiento y clasificación de los mismos, ya que pueden proceder de fuentes distintas (hojas de cálculo, archivos separados por comas o similares). Es recomendable realizar un preprocesamiento para homogeneizarlos en caso de presentarse de forma heterogénea. También será necesario agrupar los datos de acuerdo a elementos de semejanza, proximidad u homogeneidad para construir lotes de entrada hacia los programas.

Cuarta fase: diseño y construcción del modelo operativo: comprende la maquetación, estructuración y codificación de las rutinas, programas o módulos necesarios para realizar el proceso de minería de datos. Estos se elaboran utilizando lenguajes de programación, motores de bases de datos o aplicativos especializados. Pueden realizarse modelos conceptuales, lógicos y físicos de datos; así como ensamblados de programas o componentes (*add-ons*).

Una práctica recomendada es la división modular de programas muy complejos. Al no ser posible la homologación de los datos, se podrán utilizar tantos lenguajes de programación como sean necesarios. Igualmente, la interfaz de usuario quedaría en segundo plano si el interés primario son los resultados en archivos o reportes; sin embargo, en el caso de aplicaciones que ameriten ser distribuidas entre usuarios, se considerarán pantallas de entrada, salida y retroalimentación.

Quinta fase: aplicación del proceso de minería de datos: implica la ejecución de los programas diseñados previamente para la obtención de resultados. Puede realizarse de manera iterativa hasta que se logre un producto entendible por el ente humano. En algunas ocasiones sucede que, en la primera ejecución, la información procesada carece de significado, debiendo ser reprocesada. Entonces, se deberá realizar un respaldo previo de la data original para poder retroceder en caso de no obtener los resultados esperados. De la misma manera, se recomienda realizar una ejecución de prueba para validar el comportamiento de los programas antes de realizar la corrida real.

Al profundizar en la aplicabilidad de los algoritmos de minería de datos, pueden emplearse con fines de clasificación para predecir el comportamiento de variables discretas, basándose en algún atributo del conjunto de datos, como tratar de determinar si un cliente particular adquirirá un producto. Se podrán utilizar algoritmos regresivos, orientados a predecir variables continuas en virtud de anticiparse a determinados eventos, tal es el caso de prever el comportamiento del flujo de efectivo en un período futuro. También, podrán aplicarse algoritmos de segmentación que dividen los datos en grupos con características similares para comprender su comportamiento, por ejemplo, si se divide la información de los clientes de distintas zonas para determinar los productos más comprados.

Por otra parte, se encuentran los algoritmos de asociación, que buscan correlaciones entre diferentes atributos de un conjunto de datos. Uno de los usos más comunes es la creación de reglas de asociación utilizadas en un análisis de cestas de compra. Por último, los algoritmos de análisis de secuencias, resumen secuencias o patrones en los datos, como un flujo de visitas en los sitios web.

Sexta fase: análisis, interpretación y difusión de los resultados: en esta fase, se procede a examinar cuidadosamente los resultados obtenidos, encontrando significados presentes en los patrones de comportamiento.

Se puede recurrir a la comparación y contrastación de los valores encontrados con el propósito de negar o afirmar alguna hipótesis previamente planteada. La interpretación debe realizarse de manera objetiva y, de ser posible, manejar diversos puntos de vista. Se elaborará el informe respectivo a partir de las inferencias o hallazgos encontrados y finalmente, se procede a la divulgación de los mismos.

Un aspecto interesante sobre la metodología propuesta es la capacidad iterativa, es decir, en cualquier momento del proceso, se puede revisar o regresar a las etapas anteriores para realizar ajustes, lo que permite refinar cada resultado parcial obtenido de la aplicación de las fases que la conforman, en otras palabras, tiene una estructura cíclica, capaz de ser repetida tantas veces se requiera.

Conclusiones

Primeramente, se evidenció la necesidad de proporcionar a los ingenieros en informática o sistemas, una metodología para la ejecución de proyectos que aplican la minería de datos en virtud de dar solución a interrogantes o problemas partiendo de variables explícitamente definidas. En ese propósito, la investigación se fundamentó en los aportes de Maimon y Rokach (2010), Korth *et al* (2011) y Pérez (2014), quienes, desde sus distintas perspectivas, proporcionaron insumos o herramientas que sirvieron para formalizar las etapas de la metodología.

Hechas las consideraciones anteriores, Maimon y Rokach (2010) enfatizan la importancia de la calidad de los insumos, mientras Korth *et al* (2011) otorgan mayor jerarquía a las reglas del negocio previo al análisis de los datos. Por su parte, Pérez (2014) establece elementos de diseño y construcción (rutinas, programas, software especializado), así como aplicativos focalizados en el proceso de minería de datos, tales como: *RapidMiner, Weka o Knime.* También, abarca los algoritmos predictivos o incluso, complementos para aplicaciones conocidas, a saber: *Data Mining Add-on for Excel o JHepWork.*

En ese propósito, se pueden aplicar las fases propuestas a proyectos de *software*, ya que plantean un conjunto de pasos por usuarios noveles o con poca experiencia en el área. Igualmente, son de fácil comprensión para todo profesional de la ingeniería en informática o carreras afines. Esto representa una ventaja, dadas sus similitudes con metodologías existentes para el desarrollo de productos tecnológicos como sistemas de

información o software de gestión, ya que parten de un diagnóstico de la situación actual, abarcan una fase de diseño-desarrollo y culmina con la obtención de resultados.

Con base en lo anterior, se concluye que la metodología propuesta es ecléctica, por cuanto toma aspectos relevantes de los autores previamente mencionados para enriquecerse, estructurarse y vincularse dentro del campo de aplicación para el cual ha sido creada. Por consiguiente, podrá ser fácilmente utilizada por estudiantes, especialistas y desarrolladores de distintos niveles, ya que las fases presentan algunas similitudes con otras metodologías dirigidas al desarrollo de productos tecnológicos e ingeniería del software.

Sobre la base de las consideraciones anteriores, la metodología resultante quedó conformada por seis fases, las cuales son: comprensión de las reglas del negocio, formulación de la problemática, selección del conjunto de datos, diseño y construcción del modelo operativo, aplicación del proceso de minería de datos, análisis, interpretación y difusión de los resultados.

Finalmente, se deja abierta la posibilidad de utilizar rutinas o programas desarrollados específicamente para determinados trabajos de minería de datos o aplicar cualquier herramienta preexistente, ya que pueden existir diversos casos donde sea necesario aplicar estas técnicas. De hecho, para un mismo estudio se puede emplear *software* de modelado de datos, simulación, lenguajes de consulta, generadores de código, procedimientos almacenados dentro de las bases de datos entre muchos otros. La idea es proveer flexibilidad a la presente metodología en virtud de ser actualizada a lo largo del tiempo.

Referencias bibliográficas

Hernández, Roberto; Fernández Carlos y Baptista, Pilar (2014). **Metodología de la Investigación.** Sexta Edición. Editorial Mc. Graw Hill Education. México, D.F, México.

Maimon, Oded y Rokach Lior (2010). ***Data Mining and Knowledge Discovery Handbook*.** Springer, New York.

Korth Henry, Silberschatz, Abraham, Sudarshan, S. (2011). ***Database System Concepts.*** Mc. Graw Hill, Pp. 1349.

Pérez, María (2014). **Minería de datos a través de ejemplos.** Editorial RC Libros, Madrid, España.

REVISTA ETHOS VENEZOLANA Vol. 9 Nº 2 Julio-Diciembre 2017
Se terminó de imprimir en diciembre de 2017
en los talleres gráficos de Ediciones Astro Data S.A.
Telf: 0261-7511905 / Fax: 0261-7831345
Correo electrónico: edicionesastrodata@gmail.com
Maracaibo, Venezuela

Revista Ethos Venezolana
Vol. 9 No.2, Julio-Diciembre 2017
ISSN 1856-9862 ~ Depósito legal pp 200902ZU3258

Contenido

ETHOS