Revisión de metodologías de Ciencias de Datos para proyectos de planificación territorial

Luciano Perdomo[1], Leo Ordinez[1]

[1]Laboratorio de Investigación en Informática (LINVI), Facultad de Ingeniería - UNPSJB, Bvd. Brown 3051, Puerto Madryn, Argentina

{lucianor.perdomo,leo.ordinez}@gmail.com

RESUMEN

En el presente artículo, se presenta una revisión de las principales metodologías utilizadas en ciencias de datos. Se plantea la realización del relevamiento y su posterior análisis de dichas metodologías. Se desea recabar las más utilizadas y más conocidas de la industria, con el objetivo de poder determinar la conveniencia de la utilización de cada una de ellas en el ámbito de la investigación e innovación, dependiendo de las características de cada caso particular. Se han seleccionado las siguientes metodologías: KDD, CRISP-DM, SEMMA, OSEMN, TDSP.

Palabras clave: Ciencias de datos, Minería de Datos, Machine Learning

CONTEXTO

La línea de I+D+I dentro de la que se inserta el proyecto llamado "Sistemas inteligentes para la planificación territorial en la región patagónica" (UNPSJB-PI 1770). En el mismo se plantea como objetivo general poder desarrollar estrategias, tanto metodológicas como técnicas y analíticas para poder capturar, generar, analizar y explotar de información, a fin de producir conocimiento que pueda asistir en la toma de decisiones complejas sobre la planificación territorial en la Patagonia. La línea de investigación en la que se inserta el proyecto, es la de Ciudades Inteligentes. Paradigma que utiliza y saca partido de la tecnología digital dentro de la planificación urbana [1].

1. INTRODUCCIÓN

Por la naturaleza del proyecto, surge la necesidad de lidiar con datos inherentemente complejos, que son de carácter espacio-temporales y que provienen de diversas fuentes. Se pretende analizar, las distintas metodologías de ciencias de datos para poder compararlas y sentar las bases de conocimiento sobre el tema para poder utilizarlo y aplicarlo dentro del proyecto.

Dentro de las ciencias de datos, existen diferentes enfoques sobre cómo enfrentar y llevar a cabo el ciclo de vida de un proyecto. Dentro de todos los disponibles, se han seleccionado las cinco metodologías que se consideraron las más conocidas y/o utilizadas en la industria. Las cuales son: KDD (Knowledge Discovery in Databases), CRISP-DM (Cross-Industry Standard Process for Data Mining), SEMMA (Sample, Explore, Modify, Model, Assess), OSEMN (Obtain, Scrub, Explore, Model, iNterpret) y TDSP (Team Data Science Process).

La primera metodología a investigar fue Knowledge Discovery in Databases (KDD) que es un proceso en el cual se realiza extracción de conocimiento de datos mediante la aplicación de técnicas de minería de datos. Es una metodología que está centrada en el descubrimiento de conocimiento dentro de grandes volúmenes de datos. Los autores Fayyad, Piatetsky-Shapiro y Smith en 1996 [2] definen a KDD como "El proceso no trivial de identificación de patrones válidos, novedosos, potencialmente útiles y fundamentalmente entendibles al usuario a partir de los datos". Posee cinco etapas, las cuales son: 1) Selección, 2) Preprocesamiento, 3) Transformación, 4) Minería de datos, y 5) Interpretación.
KDD tiene como ventajas: ser un proceso documentado, es decir, que se facilita la comunicación entre los miembros del equipo; además de simplificar y facilitar la comprensión de la metodología. Se puede utilizar en diversos proyectos de distinta naturaleza. Por ser un proceso iterativo y cíclico, permite realizar ajustes a lo largo del

proceso. Como desventajas, puede tornarse una metodología rígida y no encajar en todos los proyectos, sobre todo los complejos. Razón por la cual, al momento de implementarla puede tornarse costosa.

La segunda metodología fue CRISP-DM (Cross-Industry Standard Process for Data Mining). Es un proceso iterativo e incremental, que permite interactuar entre sus seis fases [3] . Cada fase posee un conjunto de tareas y actividades genéricas, las cuales están descritas en un alto nivel [4]. En cada proyecto, éstas deben implementarse de acuerdo al contexto y necesidades pertinentes. Ésta metodología es independiente de cualquier herramienta utilizada para su implementación. Posee la etapas de: 1) Comprensión del negocio, 2) Comprensión de los datos, 3)Preparación de los datos, 4)Modelado, 5) Evaluación, y 6) Despliegue
Como ventajas, es una metodología iterativa y cíclica, lo que permite ajustar el proceso según sea necesario; es flexible y puede adaptarse a una variedad de proyectos. Por estar documentada, facilita la planificación y ejecución del proyecto; facilita la comunicación entre los miembros del equipo, incluyendo tanto al cliente como la coordinación de actividades.
Como desventajas, se tiene que es una metodología compleja, que puede requerir un tiempo y un esfuerzo de implementación. No es la más adecuada para los proyectos con objetivos difusos, o con calidad de datos insuficiente.

Luego, se investigó SEMMA (Sample, Explore, Modify, Model, Assess), metodología desarrollada para el software SAS por SAS Institute [4]. Es considerada de propósito general para aplicaciones de data mining. Consta de las fases: 1) Muestreo, 2) Exploración, 3) Modificación , 4) Modelo y 5)Evaluación.
Como ventajas, es simple y fácil de aprender e implementar, especialmente útil si se tienen conceptos básicos de data mining. Es útil en proyectos de pequeña escala, debido a que no requiere una gran cantidad de tiempo o recursos. Es flexible y puede adaptarse a una variedad de proyectos. Por ser iterativo permite obtener feedback e ir refinando y mejorando los modelos.
Como desventajas, es una metodología propietaria, donde sólo es abierta en sus aspectos generales, por estar fuertemente vinculada a los productos de SAS que la implementan. No es una metodología con una estructura completa que pueda ser utilizada en proyectos de datos de gran escala, como puede ser CRISP-DM. Es decir, que no posee fases de comprensión del negocio (que es de importancia para garantizar que el proyecto esté alineado a los objetivos del negocio) o Despliegue (que garantiza que los resultados del proyecto sean implementados y utilizados por los usuarios finales). Es por esto, que se incrementa la posibilidad de generar sesgos y suposiciones erróneas sobre la calidad de los datos, o en el alcance del proyecto. Es posible que la complejidad y el tiempo empleados aumente considerablemente durante las iteraciones, sobre todo en las etapas de modificación y modelado. Para mitigar los sesgos, se deben involucrar a las partes interesadas durante el ciclo de vida para garantizar que el proyecto esté alineado con sus objetivos, y obtener feedback que pueda ser utilizado para mejorar tanto el proceso como sus resultados. Además, documentar facilita la comprensión y la reproducción de lo realizado.

La siguiente es OSEMN (Obtain, Scrub, Explore, Model, iNterpret), creada por Hilary Mason y Chris Wiggins en un post en el 2010, el cual se llamaba "A Taxonomy of Data Science" [6]). Consta de 5 pasos: 1) Obtención, 2)Raspar/Limpiar, 3) Explorar, 4)Modelar, 5) Interpretar, que implica interpretar los resultados obtenidos y derivar conclusiones.
Tiene las ventajas de ser una metodología directa y bastante práctica que puede ser aprendida fácilmente por personas con un conocimiento básico de data mining; es flexible y puede adaptarse a proyectos de pequeña y mediana escala; hace énfasis en la exploración de los datos, lo que es importante para comprender los datos y identificar oportunidades de modelado
Tiene un enfoque altamente práctico y está centrada en usuarios con cierto nivel de expertise o conocimientos en programación y/o uso de herramientas informáticas. Al igual que SEMMA, no es una metodología con una estructura demasiado detallada; y también carece de fases de Comprensión del negocio o Despliegue. Por lo que tampoco es adecuada

para utilizarla en proyectos de datos de gran escala o con requisitos difusos.

Por último se analizó, TDSP (Team Data Science Process), desarrollada por Microsoft [7] para ser una metodología ágil e iterativa. Está pensada para proyectos de gran tamaño, sin embargo, según sus creadores, puede utilizarse para pequeños proyectos, omitiendo ciertos elementos. Posee cinco fases en las que se detallan los objetivos específicos, las tareas a realizar y los artefactos, incluyendo sus entregas y cómo hacerlas. Éstos están asociados con los siguientes roles del proyecto: a) Arquitecto de soluciones, b) Jefe de proyecto, c) Ingeniero de datos, d) Científico de datos, e) Desarrollador de aplicaciones, f) Responsable de proyecto. Las fases de TDSP son: 1) Comprensión del negocio, 2) Adquisición y comprensión de los datos, 3) Modelado, 4) Implementación y 5) Aceptación del cliente.
Como ventajas, TDSP proporciona un marco común para que los roles y miembros del equipo se comuniquen y colaboren entre sí. Mejorando la colaboración y el aprendizaje; facilitando la comunicación y la colaboración entre los miembros. Se proporciona un proceso estructurado para seguir, lo que puede ayudar a garantizar la completitud del proyecto. El enfoque sistemático para la selección, el entrenamiento y la evaluación de modelos, influye en la mejora de la calidad de los modelos.
Como desventajas, la metodología, puede tornarse demasiado compleja y requerir demasiado tiempo para proyectos simples o pequeños. Además los requerimientos en los recursos humanos y el equipo es una limitante, debido a la experticia necesaria para implementarlo de manera efectiva.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

El proyecto posee la característica de ser de carácter interdisciplinario, e involucrar distintas áreas de influencia de las ciudades inteligentes y la planificación territorial. Éstas áreas son: el sector de la economía del conocimiento, el pesquero y el de la salud pública. En el mismo, se propone investigar y aplicar la teoría para poder asistir a la captura de información compleja y el posterior desarrollo de artefactos y componentes como resultado de desarrollos experimentales.

3.RESULTADOS ESPERADOS / OBTENIDOS

Se abordaron distintas metodologías de ciencias de datos, con el objetivo de poder documentar lo aprendido, y permitir su posterior aplicación dentro del proyecto. Por lo tanto, es preciso conocer y comprender cada una de ellas y así determinar cuál utilizar en cada caso. Facilitando de esta manera, tanto los procesos y análisis a realizar, cómo los resultados y productos que se desean obtener y generar.

Durante el proceso, se obtuvo conocimiento documentado sobre las principales metodologías en Ciencias de Datos en el mercado. Debido a que cada una posee sus propias características con sus ventajas y desventajas, la elección adecuada dependerá tanto de las particularidades de cada subproyecto. Éstas particularidades pueden ser los objetivos, la disponibilidad de recursos humanos, temporales y la complejidad inherente de los datos, entre otras.

4. FORMACIÓN DE RECURSOS HUMANOS

El proyecto está integrado por especialistas de Ciencias de la Computación y de Ciencias Sociales. A fin de consolidar la investigación interdisciplinaria y poder abordar las diferentes temáticas desde distintas perspectivas. Además participan investigadores del Laboratorio de Investigación en Informática (LINVI) de la Universidad Nacional de la Patagonia San Juan Bosco (UNPSJB). Dentro del proyecto se incluyen alumnos de grado y de posgrado con el objetivo de formar recursos humanos en tales especialidades.

5. BIBLIOGRAFÍA

[1] Ratti, C., & Claudel, M. (2016). The city of tomorrow: Sensors, networks, hackers, and the future of urban life. Yale University Press.

[2] Fayyad, U., Piatestky-Shapiro, G. y Smyth, P. (1996). The kdd Process for Extracting Useful Knowledge from Volumes of Data. Communications of the acm, 39(11), 27-34.

[3] Sordo, M.A., López, J.A., García, A.J., & Gil, M.A. (2003). Análisis y modelado de datos para la predicción de fallos en aerogeneradores. [En línea]. Disponible en: https://www.aeipro.com/files/congresos/2003pamplona/ciip03_0257_0265.2134.pdf

[4] Espinoza Mina , M.A. (s.f.) .Retos y perspectivas de las tecnologías de información. Capítulo III. CRISP-DM: Conocimiento y comunicación de una metodología para minería de datos. [En línea]. Disponible en: https://libros.ecotec.edu.ec/index.php/editorial/catalog/download/2/2/29-1?inline=1

[5] SAS Institute Inc. (2023). Introduction to SEMMA. [En línea]. Disponible en: https://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jnj8bbjjm1a2.htm.

[6] Mason, H. (2010, septiembre). A Taxonomy of Data Science. [Entrada en blog]. [se quitó una URL no válida]. https://web.archive.org/web/20160220042455/dataists.com/2010/09/a-taxonomy-of-data-science/

[7] Microsoft. (n.d.). ¿Qué es el Proceso de ciencia de datos en equipo (TDSP)?. [Página web]. Microsoft Azure. https://learn.microsoft.com/es-es/azure/architecture/data-science-process/overview