Revisión de metodologías de Ciencias de Datos para Proyectos de Planificación Territorial

Luciano Perdomo , Leo Ordinez
{lucianor.perdomo, leo.ordinez}@gmail.com
Laboratorio de Investigación en Informática (LINVI)
Facultad de Ingeniería - UNPSJB, Bvd. Brown 3051, Puerto Madryn, Argentina

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

Resumen

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.

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. Se abordaron distintas metodologías de ciencias de datos, con el objetivo de poder documentar lo aprendido, y permitir su posterior aplicación dentro del proyecto.

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. 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.

Formación de R.R.H.H.

El proyecto es interdisciplinario y posee especialistas en las Ciencias de la Computación y 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 Facultad de Ingeniería de la Universidad Nacional de la Patagonia San Juan Bosco. Participó un Doctor Ingeniero, especializado en la temática de Ciudades Inteligentes y un doctorando en las Ciencias de la Ingeniería que está realizando su investigación en el ámbito de la Informática en Salud Pública

Resultados Obtenidos / Esperados

Se abordaron distintas metolodogías para documentar lo aprendido y permitir su posterior uso. 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 en el mercado. Debido a que cada una posee sus propias características, la elección adecuada dependerá de cada caso. Se debe tener en cuenta los objetivos, la disponibilidad de recursos humanos, temporales y la complejidad inherente de los datos, entre otras. Además, los conocimientos obtenidos fueron empleados en un curso de Problemas de Datos dentro de una Diplomatura en Gestión y Análisis de Datos dictada por la Facultad de Ciencias Económicas (UNPSJB). Además se utilizó la metodología OSEMN para realizar un artículo en el que se utilizó machine learning para determinar si una persona posee diaberes, a partir de sus factores de riesgo

Comparación rápida

KDD

Etapas: 5
Ventajas:
- Proceso documentado (facilita la comunicacion y comprensión)
- Proyectos de distinta naturaleza
- Es iterativo y cíclico

Desventajas:
- Pude tornarse rígida
- Puede no encajar en todos los proyectos

CRISP-DM

Fases: 6
Ventajas:
- Iterativo, incremental y cíclica
- Flexible
- Documentada (Actividades y Tareas genéricas)
- Independencia de herramientas
- Variedad de proyectos

Desventajas:
- Complejidad
- Tiempo y Esfuerzo requeridos
- No es adecuda para proyectos con requisitos difusos o calidad de datos insuficiente

SEMMA

Fases: 5
Ventajas:
- Simple y fácil de aprender
- Iterativo
- Flexible
- Para proyectos pequeños

Desventajas:
- Propietaria
- No tiene una estructura tan completa como CRISP-DM
- No es adecuda para proyectos con requisitos difusos
- Puede crear sesgos y suposiciones erroneas sobre datos o alcance

TSP

Fases: 5
Ventajas:
- Es ágil e iterativa
- Proceso documentado (facilita la comunicacion y comprensión)
- Define objetivos, tareas, artefactos, entregas y como hacerlas
- Define los roles del proyecto
- Pensada para proyectos de gran tamaño (pero puede usarse en los de pequeña escala)

Desventajas:
- Pude tornarse demasiado compleja
- Requiere demasiados recursos (humanos, de tiempo y equipo)
- Puede no encajar en todos los proyectos
- Necesita cierto nivel de expretise para implementarla de forma efectiva

OSEMN

Pasos: 5
Ventajas:
- Directa
- Práctica
- Flexible
- Facil Aprender
- Proyectos pequeños / medianos

Desventajas:
- Necesita cierta expertise
- No tiene una estructura detallada ni tan completa
- No es adecuda para proyectos con requisitos difusos o de gran escala

Conclusiones:

- Cada metodología posee sus propias características con sus ventajas y desventajas
- La elección adecuada dependerá de las necesidades y particularidades de cada caso