

Discusión del Artículo: A Taxonomy of DataScienceProcessModels: Insights from Science and Practice

Integrantes

- Javier Chen - 22153
- Josué Say - 22801

Resumen elaborado por: Javier Chen

El artículo habla sobre la taxonomía estructurada para clasificar modelos de proceso en ciencia de datos (DSPM). Estos modelos son esenciales para guiar las etapas de proyectos de ciencia de datos, desde la recolección hasta la implementación de soluciones. A pesar de la existencia de modelos conocidos como CRISP-DM, KDD o SEMMA, su uso práctico es limitado, lo que revela una desconexión entre teoría y práctica.

La taxonomía desarrollada fue organizada en 35 modelos, 3 categorías, 13 dimensiones y 67 características, permitiendo comparar modelos según dominio de aplicación, roles involucrados, usabilidad, etc. Se utilizó el método ETDP, que combina enfoques deductivos e inductivos, en un proceso iterativo de cinco etapas apoyado por literatura científica y entrevistas con 12 expertos.

Los resultados muestran que muchos modelos carecen de entregables claros y herramientas prácticas, lo que reduce su uso en casos reales. Además, la mayoría siguen derivados de CRISP-DM, indicando que ha habido poca innovación. El artículo aporta una herramienta útil para académicos y profesionales, promoviendo una comprensión más profunda de los DSPM. Además, se destaca la necesidad futura de mejorar su usabilidad y valor económico para facilitar su uso en la industria.

Ideas principales del artículo

- La mayoría de los profesionales en ciencia de datos reconoce el valor de los modelos de proceso, muy pocos los utilizan formalmente, lo que evidencia una brecha entre la academia y su aplicación real.
- La taxonomía se desarrolló siguiendo un proceso estructurado y validado científicamente, integrando revisiones de literatura y retroalimentación de expertos.

Ideas más interesantes

- Una de las aportaciones más destacadas del artículo es la identificación de una debilidad común en los modelos existentes: su falta de orientación práctica. Ya que carecen de componentes esenciales como ejemplos aplicados, herramientas, guías detalladas, etc. Esta ausencia dificulta su implementación en contextos reales, donde se requiere no solo una estructura conceptual, sino apoyo operativo concreto.
- Gran parte de los modelos actuales siguen estando basados en estructuras antiguas como CRISP-DM. Aunque han surgido nuevas versiones y adaptaciones, la mayoría conserva una estructura básica similar, lo que evidencia una escasa innovación metodológica. Esta falta de evolución limita el ajuste de los modelos a los nuevos desafíos y necesidades de los proyectos modernos de ciencia de datos.

Expectativas de aprendizaje del curso

- Espero adquirir habilidades para interpretar datos de manera significativa, entendiendo patrones, relaciones y tendencias.
- Espero aprender a usar datos como soporte para justificar decisiones y proponer soluciones fundamentadas.

Resumen elaborado por: Josué Say

Este artículo se enfocó en organizar y clasificar los diferentes modelos de procesos en ciencia de datos (CRISP-DM o KDD), que son los pasos que se siguen para desarrollar proyectos de data science. Aunque hay muchos modelos, en la práctica casi no se usan de forma formal, lo cual se menciona que representa un problema, a lo cual se crean una taxonomía (una especie de “mapa organizado”) que agrupa estos modelos según varias características como:

- Qué tan detallados son
- Qué tipo de entregables producen
- Si están pensados para un uso general o específico, qué roles considera (científico de datos, ingeniero, etc.)
- Entre otras cosas

Analizaron 35 modelos distintos y entrevistaron a expertos para validar su clasificación. El resultado es una herramienta que ayuda tanto a investigadores como a personas que trabajan en proyectos reales, para elegir el modelo que más les convenga, comparar entre ellos, o simplemente entender mejor cómo funcionan. Es

decir que no se propone un nuevo modelo, sino que da una visión completa y estructurada de los que ya existen, ayudando a mejorar su uso en el mundo real.

Ideas principales del artículo

- Aunque se reconoce que los modelos ayudan a organizar bien los proyectos de data science, muy pocos profesionales los usan realmente. Solo el 18% sigue un proceso documentado, a pesar de que el 85% admite que ayudan mostrando una brecha entre la teoría y la práctica.
- El diseño realizado contiene 13 dimensiones y 67 características, lo que permite analizar en qué se parece o diferencia un modelo de otro, lo que puede ahorrar tiempo y evitar errores al elegir uno.

Ideas más interesantes

- La taxonomía no solo sirve para académicos, sino también para que empresas puedan elegir mejor cómo manejar sus proyectos de ciencia de datos según sus necesidades reales.
- Entre los hallazgos destaca que muchos modelos derivan de CRISP-DM, pero pocos definen entregables claros, roles específicos o herramientas prácticas, lo que limita su aplicabilidad. Solo un 14% evidencian rigor científico, y menos del 10% ofrecen usabilidad avanzada.

Expectativas de aprendizaje del curso

- Quiero entender cómo aplicar modelos como redes LSTM o redes convolucionales para usarlo en análisis de comportamiento en redes sociales.
- Me interesa mejorar en la parte de dashboards interactivos porque siento que a veces los datos que uno analiza se quedan en lo técnico, y no llegan a quienes toman decisiones. Me gustaría aprender a comunicar resultados de forma visual y clara.
- Tengo curiosidad por el análisis de sentimientos y la minería de texto porque son áreas que combinan programación con interpretación del lenguaje.