JUSTIFICACIÓN DE LA METODOLOGÍA

Presentado por: GRUPO 1

CINDY JOHANNA ZAPATA ROMERO

HECTOR GEOVANY BELLO SANTAMARÍA

MARIO GUERRA GUALY

LEANDRO REYES JORDÁN

Docente:

CARLOS ISAAC ZAINEA MAYA

UNIDAD DE ESTUDIO:

GERENCIA DE PROYECTOS PARA CIENCIA DE DATOS

MAESTRÍA - GRUPO 1 - M1V - VIRTUAL - 2025

FACULTAD DE INGENIERÍA

BOGOTÁ, 12 DE FEBRERO DE 2025

UNIVERSIDAD EAN

Tabla de Contenido

COMPARACIÓN DE METODOLOGÍAS APLICABLES

En el desarrollo de proyectos de análisis de datos y modelado predictivo, existen diversas metodologías que proporcionan una estructura para la ejecución eficiente del proyecto. A continuación, se presentan las tres metodologías más ampliamente utilizadas en ciencia de datos: CRISP-DM, KDD y TDSP.

CRISP-DM (Cross Industry Standard Process for Data Mining)

CRISP-DM es una metodología estándar para proyectos de minería de datos que consta de seis fases:

- Comprensión del negocio
- Comprensión de los datos
- Preparación de los datos
- Modelado
- Evaluación
- Despliegue

Ventajas	Estructura clara y flexible.
	Enfocada en la comprensión de los datos y el negocio.
	Puede aplicarse a diferentes tipos de proyectos de ciencia de datos.
Desventajas	No está diseñada específicamente para entornos colaborativos y ágiles.
	No enfatiza directamente la automatización del pipeline de datos.

KDD (Knowledge Discovery in Databases)

El proceso KDD se enfoca en la extracción de conocimiento a partir de bases de datos y consta de las siguientes fases:

- Selección de datos
- Preprocesamiento
- Transformación
- Minería de datos
- Interpretación/Evaluación

Ventajas	Profundiza en la transformación de los datos antes del modelado.
	Ideal para exploración y descubrimiento de patrones.
Desventajas	No proporciona guías claras sobre la implementación y el despliegue.
	Menos estructura en la relación con el negocio.

TDSP (Team Data Science Process)

TDSP es un marco de trabajo desarrollado por Microsoft para la ciencia de datos en equipos, con énfasis en la colaboración, automatización y reutilización de código. Sus fases incluyen:

- Comprensión del problema
- Adquisición y exploración de datos
- Modelado
- Implementación
- Mantenimiento

Ventajas	Ideal para equipos grandes y proyectos con necesidades de despliegue.
	Enfocado en la automatización y reutilización de modelos.
Desventajas	Puede ser complejo para proyectos pequeños.
	Requiere una estructura organizativa definida.

JUSTIFICACIÓN DE LA METODOLOGÍA ELEGIDA

Al comparar las metodologías mas utilizadas en combinación con Scrum, se pueden observar algunas características relevantes que permiten escoger la mejor combinación, dada la naturaleza del proyecto.

CRISP-DM + Scrum	Se pueden organizar **sprints** basados en las fases de CRISP-DM.
	Ejemplo: un **Sprint 1** podría enfocarse en la "Comprensión del Negocio" y la "Comprensión de los Datos", mientras que un **Sprint 2** abarcaría la "Preparación de los Datos".
	Cada fase de CRISP-DM puede descomponerse en **historias de usuario** y tareas en el backlog.
KDD + Scrum	Como KDD es un proceso más general, puede servir como una guía para definir **épicas y tareas** en el backlog de Scrum.
	Ejemplo: las fases de "Selección de datos" y "Preprocesamiento" podrían ocupar los primeros sprints.
	La fase de "Minería de Datos" podría desarrollarse en varios sprints con iteraciones en los modelos.
TDSP + Scrum	TDSP ya está diseñado para trabajo en equipo, lo que lo hace compatible con Scrum.
	Se puede usar TDSP para definir entregables técnicos y Scrum para gestionar los sprints y revisiones periódicas.
	Las herramientas de TDSP (como repositorios estandarizados y pipelines de datos) pueden integrarse en el flujo de trabajo ágil.

Para el presente proyecto, se ha seleccionado CRISP-DM como la metodología principal de ejecución, complementada con elementos de Scrum para la gestión ágil del equipo. Esta elección se justifica por los siguientes motivos:

- **Estructura adaptable:** CRISP-DM proporciona una guía clara y flexible para el desarrollo de proyectos de ciencia de datos, lo que permite ajustarse a los requerimientos específicos del análisis de noticias y clustering.

- **Enfoque en los datos:** Dado que el proyecto se basa en la identificación de patrones en un corpus de noticias, la metodología enfatiza la comprensión y preparación de los datos, lo cual es fundamental para el éxito del modelo.

- **Iteración y evaluación:** CRISP-DM permite revisar y mejorar el modelo en cada fase, asegurando la optimización del clustering y la interpretabilidad de los resultados.

- **Compatibilidad con Scrum:** Dado que el proyecto se desarrolla en equipo y con un enfoque ágil, se incorpora Scrum para la gestión de tareas, reuniones de seguimiento y entrega iterativa de avances.

ADECUACIÓN A LA NATURALEZA DEL PROYECTO

El presente proyecto se centra en la clasificación y agrupación de noticias mediante Procesamiento de Lenguaje Natural (PLN), con la aplicación de técnicas como TF-IDF, PCA y K-Means.

La metodología elegida se justifica dado que la combinación de CRISP-DM y Scrum proporciona una estructura adecuada para garantizar el desarrollo exitoso del mismo, permitiendo un análisis eficiente de los datos y una gestión flexible del equipo.

Asimismo, la metodología elegida se ajusta a la naturaleza del proyecto de la siguiente manera

- **Exploración y preprocesamiento de datos:** CRISP-DM permite realizar un análisis profundo de los datos antes de aplicar técnicas de modelado, garantizando una mejor calidad del corpus de noticias.

- **Iteración en el modelado**: La posibilidad de ajustar el modelo a lo largo del ciclo de vida del proyecto es clave para optimizar la agrupación de noticias.

- **Adaptabilidad a cambios:** Con la incorporación de Scrum, el equipo puede responder rápidamente a ajustes en los requerimientos o cambios en los datos.