Análisis de Datos y Modelado Predictivo

Leandro Reyes Jordan

Hector Geovany Bello Santamaría

Mario Guerra Gualy

Cindy Johanna Zapata Romero

Gerencia de Proyectos para Ciencia de Datos

Universidad EAN

2025

Plan de Proyecto: Análisis de Datos y Modelado Predictivo

1. **Identificación del Problema:**

El presente proyecto se enfoca en el análisis de datos y modelado clasificación mediante técnicas de Procesamiento de Lenguaje Natural (PLN). Se busca identificar patrones en un corpus de noticias y generar clusters de información relevante usando TF-IDF y Análisis de Componentes Principales (PCA).

2. **Objetivos del Proyecto (SMART):**

- Automatizar la clasificación de noticias en distintas categorías temáticas utilizando técnicas de Procesamiento de Lenguaje Natural (PLN) para mejorar la eficiencia en la organización de información textual**.**
- Implementar y evaluar un modelo de clustering en una base de datos de noticias, utilizando TF-IDF para la representación de datos, PCA para reducción de dimensionalidad y K-Means como algoritmo de agrupamiento, con el objetivo de optimizar la clasificación de noticias en clusters con alta cohesión y baja dispersión**.**
- Desarrollar una solución escalable mediante la integración de un pipeline automatizado de preprocesamiento y clustering de noticias con K-Means**,** asegurando procesamiento eficiente y bajo costo computacional**.**
- Completar la implementación del modelo y su evaluación final**,** asegurando entregables parciales en cada fase del proyecto y la documentación técnica correspondiente.

3. **Fases del Proyecto:**

Fase	Actividades	Duración
Fase 1: Definición	BD de noticias, definición de alcance y objetivos	
Fase 2: Preprocesamiento	Tokenización, limpieza de datos, aplicación de TF-IDF	
Fase 3: Modelado	Aplicación de PCA y clustering con K-Means	
Fase 4: Evaluación	Validación del modelo, métricas de desempeño	
Fase 5: Documentación y entrega	Generación de informes y presentación de resultados	

4. Entregables esperados:

- Documento de definición del problema y objetivos**.**
- Preprocesamiento de datos con implementación de TF-IDF.
- Modelo de clustering con K-Means y análisis de PCA**.**
- Evaluación del modelo con métricas establecidas.
- Informe final y presentación de resultados.

5. Responsables:

Actividad	Responsable
Preprocesamiento y NLP	
Desarrollo del modelo	
Evaluación del modelo	
Documentación y presentación	

6. Justificación de la Metodología:

Para este proyecto se ha seleccionado la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), ya que se alinea a la naturaleza del proyecto enfocado en análisis de datos y modelado predictivo**.**

- Ciclo de Vida del proyecto elegido: CRISP-DM (Cross Industry Standard Process for Data Mining)

Justificación de CRISP-DM:

- ✓ **Comprensión del negocio:** Se definen los objetivos del proyecto y la relevancia del análisis de noticias.
- ✓ **Comprensión de los datos:** Se explora la base de datos de noticias, aplicando técnicas de PLN y TF-IDF.
- ✓ **Preparación de los datos:** Se eliminan ruidos, se normaliza el texto y se aplican transformaciones adecuadas.
- ✓ **Modelado:** Se implementan técnicas de clustering con K-Means y PCA.
- ✓ **Evaluación:** Se aplican métricas para validar la calidad de los clusters.
- ✓ **Despliegue:** Se generan informes finales y visualizaciones

- Comparación con otras metodologías:

Metodología	Justificación
CRISP-DM	Enfoque estructurado para análisis de datos y modelado predictivo.
TDSP	No es ideal porque el equipo no está basado en entornos colaborativos empresariales.
OSEMN	Más adecuado para análisis exploratorio, pero este proyecto requiere modelado predictivo.
ASUM-DM	Enfoque más empresarial, menos útil para análisis de noticias.

7. **Adaptación del Proyecto:**

- El proyecto se ajustará de la siguiente manera:
  - ✓ docs/methodology → Contendrá la metodología utilizada y la justificación.
  - ✓ docs/tools_and_environment → Documentará las herramientas y tecnologías utilizadas.
  - ✓ project_management → Incluirá la planificación del proyecto, la evaluación de riesgos y el análisis de stakeholders.
  - ✓ reports → Almacenará los reportes de calidad de datos, evaluación del modelo y resumen ejecutivo.

- Incorporación de Documentación Requerida:

- ✓ Se elaborarán documentos detallados para reflejar la metodología, herramientas y procesos empleados.
- ✓ Se actualizarán los archivos existentes para asegurar coherencia con los entregables.

- Revisión de Roles:
  - ✓ Se adaptarán los roles según la metodología utilizada, considerando que el enfoque en modelado y evaluación ya está avanzado.

- Documentación de la Fase Final:
  - ✓ Se priorizará la documentación de las fases de evaluación, validación y resultados.
  - ✓ Se garantizará que los informes de rendimiento y calidad del modelo sean precisos y completos.