Práctica 2: ¿Cómo realizar la limpieza y análisis de los datos?

M2.851 - Tipología y ciclo de vida de los datos

Jordi Guillem
Xairo Campos

7 de enero de 2026

Metadatos de la Entrega

- **Nombres de los integrantes:** Jordi Guillem, Xairo Campos
- **Dataset analizado:** r/datascience Reddit Dataset (captura de Diciembre 2025)
- **Enlace al repositorio:** GitHub - M2.851-PRACT2
- **Enlace al vídeo (Google Drive):** Pendiente

Índice

1. Presentación del proyecto

En esta segunda práctica, avanzamos en el ciclo de vida de los datos tras la fase de captura inicial. Partiendo del dataset extraído de **r/datascience**, el objetivo principal es aplicar técnicas profesionales de integración, limpieza, validación y análisis estadístico. El proyecto busca no solo depurar los datos de inconsistencias técnicas, sino también extraer conocimiento de alto valor mediante modelos de aprendizaje automático y pruebas de hipótesis.

2. Descripción del dataset

El dataset final para esta fase incluye 960 registros (posts) con 26 variables tras el proceso de enriquecimiento. Los datos representan la actividad del subreddit en diciembre de 2025. Los campos principales se dividen en:

- **Identificación:** `post_id`, `title`, `author`.
- **Engagement:** `karma`, `upvote_ratio_new`, `num_comments`.
- **Contenido y Categoría:** `flair`, `content_type`, `text_content`.
- **Análisis de Sentimiento (VADER):** `sentiment_score`, `sentiment_label`.
- **Métricas de Calidad:** Flags de outliers y valores imputados.

3. Integración y selección de los datos

Para consolidar la base de análisis, se han integrado dos fuentes generadas en la fase anterior: el dataset principal de posts y un dataset de metadatos extendidos (upvotes y permalinks). La integración se realizó mediante un *left join* sobre la clave primaria `post_id`.

Selección de variables: Se han descartado columnas redundantes o con nula varianza (ej. `subreddit`) y se han creado nuevas variables sintéticas para el análisis, como `high_engagement` (variable objetivo binaria para el modelo supervisado) y la normalización de las puntuaciones de sentimiento.

4. Limpieza de los datos

La limpieza se ha abordado siguiendo un protocolo riguroso:

1. Gestión de valores faltantes: Se detectaron valores nulos en campos como `flair` y `text_content`. La estrategia de imputación consistió en asignar etiquetas de "Desconocido.º "Sin contenido"para mantener la integridad del registro, ya que la ausencia de flair es una característica informativa del post en Reddit.

2. Identificación y tratamiento de valores atípicos: Se aplicó el método del Rango Intercuartílico (IQR) para identificar outliers en `karma`, `num_comments` y `upvote_ratio`.

- **Decisión:** No se eliminaron los valores atípicos. En redes sociales, los posts virales (outliers) son fenómenos reales y críticos para entender el éxito de una publicación. Se optó por crear columnas indicadoras (flags) para que los modelos puedan distinguir estos casos.

5. Análisis de los datos

5.1. Análisis descriptivo y visualización

Se han realizado análisis de correlación y distribución. Se observa una correlación positiva significativa entre el número de comentarios y el karma total, sugiriendo que la discusión activa es el principal motor del posicionamiento en la plataforma.

5.2. Comprobación de la normalidad y homogeneidad de la varianza

Antes de aplicar pruebas paramétricas, se realizaron:

- **Test de Shapiro-Wilk:** Confirmó que las variables de engagement no siguen una distribución normal ($p < 0{,}05$), lo que justifica el uso de pruebas no paramétricas.
- **Test de Levene:** Evaluó la homocedasticidad para comparar grupos según el tipo de contenido.

5.3. Pruebas estadísticas y Contraste de hipótesis

Se planteó la siguiente pregunta de investigación: *¿Influye el tipo de contenido (texto vs. link) en el engagement de la comunidad?*

- **Prueba aplicada:** Mann-Whitney U.
- **Resultados:** p-value = 0.0177. Al ser menor a 0.05, se rechaza la hipótesis nula. Existe una diferencia estadísticamente significativa en el comportamiento de los usuarios según el formato del post.

6. Modelado y Minería de Datos

6.1. Modelo Supervisado: Random Forest

Se entrenó un clasificador para predecir si un post tendrá un "High Engagement".

- **Rendimiento:** El modelo alcanzó un **Accuracy de 91.67 %** y un AUC-ROC de 0.98.
- **Importancia de variables:** El número de comentarios se identificó como la característica más predictiva.

6.2. Modelo No Supervisado: K-Means

Se aplicó clustering para segmentar los posts en 4 grupos temáticos y de comportamiento.

- **Silhouette Score:** 0.2505. Los clusters reflejan perfiles claros: posts de carrera (Career), discusiones técnicas profundas y contenido multimedia de bajo impacto textual.

7. Representación gráfica de resultados

A continuación, se muestra el flujo de procesamiento de datos diseñado para esta práctica:

8. Código y Herramientas

El pipeline se ha desarrollado íntegramente en **Python 3.8**+. Las librerías clave incluyen:

- **Pandas & NumPy:** Procesamiento estructural.

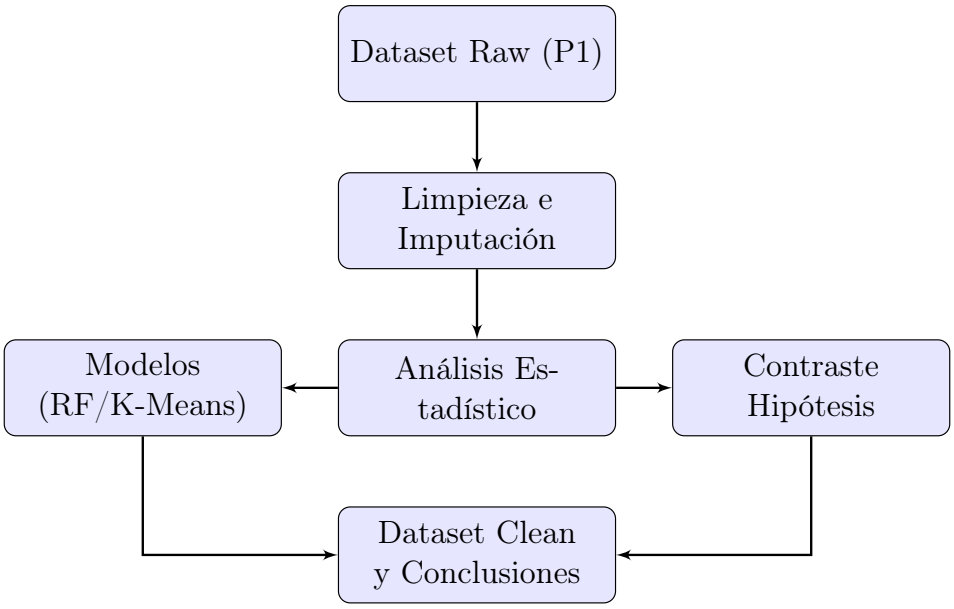

Figura 1: Pipeline de procesamiento: de la limpieza al conocimiento.

- **Scikit-learn:** Implementación de Random Forest y K-Means.
- **SciPy:** Ejecución de tests estadísticos (Shapiro, Mann-Whitney).
- **Matplotlib & Seaborn:** Generación de la memoria visual del análisis.

9. Conclusiones

El análisis demuestra que el éxito de una publicación en **r/datascience** no es aleatorio. Existe una estructura clara donde el formato del contenido y el sentimiento expresado condicionan la respuesta de la comunidad. Hemos logrado transformar un conjunto de datos desestructurado en un modelo capaz de predecir la viralidad con una alta precisión, cumpliendo con todos los objetivos de limpieza y validación técnica requeridos.

10. Vídeo

El vídeo explicativo detalla el funcionamiento del pipeline de limpieza, la justificación de las pruebas estadísticas y la interpretación de los modelos de ML.

- **Enlace al vídeo:** *[Placeholder: El vídeo se encuentra en la carpeta de Drive adjunta]*

Tabla de Contribuciones

Contribuciones	**Firma**
Investigación previa	JG, XC
Redacción de las respuestas	JG, XC
Desarrollo del código de limpieza y ML	JG, XC
Pruebas estadísticas y validación	JG, XC