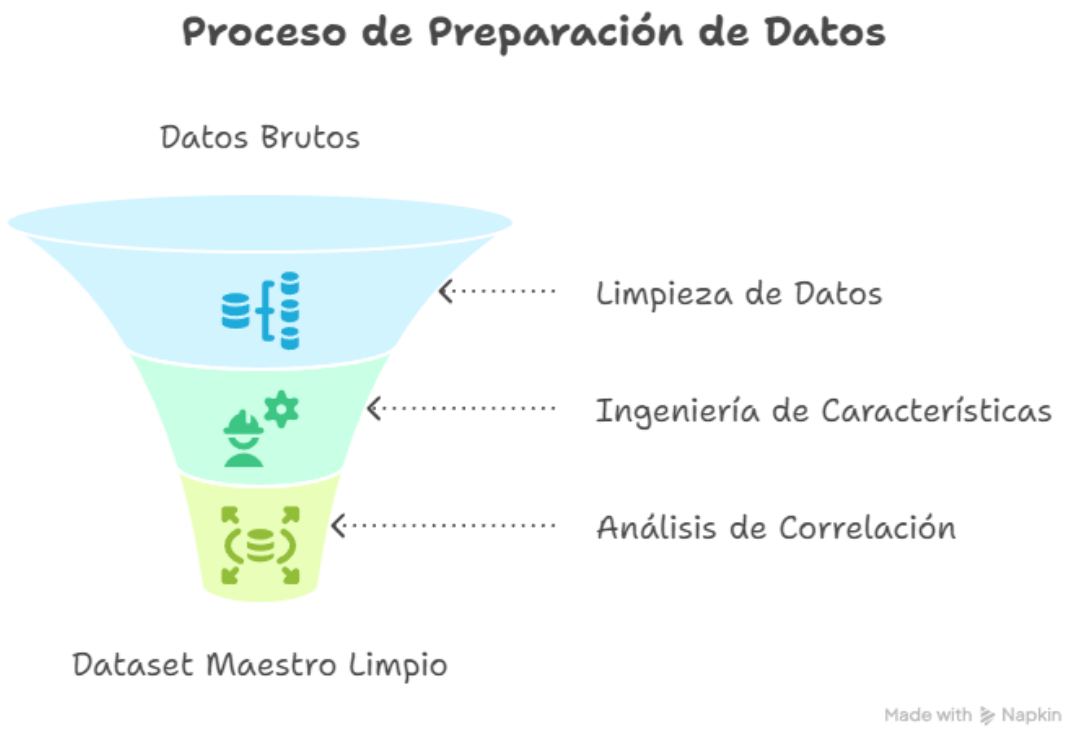
🔁 **Pipeline de Datos – Entregable 2: Exploración y Curación de Datos**

Deadline: 25/07

Este pipeline representa el flujo de trabajo lógico (workflow) del Entregable 2. El objetivo es profundizar el análisis, mejorar la calidad de los datos y preparar un dataset robusto, limpio y confiable, que será la base para construir nuestro modelo predictivo.



🔹 **1. Tratamiento de valores faltantes, duplicados e inconsistencias**

- **Fuente:** Notebook + Datasets resultantes del Entregable Nro. 1
- **Objetivo:** Revisión exhaustiva de los datasets obtenidos para identificar y aplicar estrategias de tratamiento adecuadas.
- **Tareas:**
    - Valores Faltantes: Imputación, eliminación o técnicas avanzadas según la naturaleza de los datos y el impacto en el análisis.
    - Registros Duplicados: Identificación y eliminación de entradas redundantes para asegurar la unicidad de los registros.
    - Inconsistencias: Corrección de errores de formato, tipográficos o lógicos que puedan afectar la calidad y la coherencia de los datos.

🔹 **2. Ingeniería de características (Feature Engineering) básica**

- **Objetivo:** Se aplicarán técnicas de curación, normalización y transformación de variables para optimizar el dataset para el modelado.
- **Tareas:**
    - Curación: Refinamiento de características existentes para mejorar su calidad y relevancia. Extracción y estandarización de categorías temáticas.
    - Normalización/Estandarización: Ajuste de las escalas de las variables numéricas para evitar que aquellas con rangos más amplios dominen el proceso de modelado.
    - Transformación: Aplicación de funciones matemáticas o lógicas para crear nuevas características a partir de las existentes, o para ajustar la distribución de las variables.

🔹 **3. Análisis más profundo de correlaciones y relaciones entre variables para seleccionar las más prometedoras**

- **Objetivo:** Se realizará un análisis más detallado de las interrelaciones entre las variables del dataset.
- **Tareas:**
    - Identificación de Correlaciones: Descubrir relaciones lineales y no lineales entre las características.
    - Selección de Variables: Elegir las características más prometedoras y relevantes para el modelo predictivo, descartando aquellas que aporten poco valor o introduzcan ruido.

🔹 **4. Generación de un dataset maestro (master dataset) limpio y bien documentado**

- **Objetivo:** Como resultado de las tareas anteriores, se consolidará un dataset final que será la base para la siguiente fase de modelado. Este dataset se caracterizará por ser:
    - Limpio: Libre de valores faltantes, duplicados e inconsistencias.
    - Robusto: Con características optimizadas y seleccionadas para el rendimiento del modelo.
    - Bien Documentado: Se proporcionará un diccionario de datos y una descripción clara de las transformaciones aplicadas a cada variable

---

🔹 **5. Documentación y Output**

- **Notebook Jupyter o Google Colab** con:
    - Notebook de curación con código y justificación de las decisiones tomadas
- **Documento técnico** con:

    - Documento técnico-funcional con hallazgos clave post-curación
- **Dataset maestro curado** (en formato CSV o similar), listo para el modelado final limpio

---

Anexo con características que debe cumplir el entregable:

- Desarrollar cada punto en la misma notebook donde se escriba el código (se puede trabajar con una notebook para cada entrega, o en una notebook integral). El pipeline debe contar con un índice, con sus diferentes apartados y el código debe ser fácil de leer, estar probado y comentado (esto último, en función de la necesidad).
- Enviar el link directo del archivo .ipynb ó alternativamente subir el entregable a un repositorio GitHub mediante la integración con Google Colab. Recordar que al compartir el notebook, queden habilitados los permisos de edición, para poder dejar comentarios/correcciones. Tener en cuenta que si bien, se pueden realizar diversos análisis y visualizaciones, se debe dejar en el entregable sólo aquello que sea relevante.

Repositorio Mentoría:

https://github.com/NoeliaFerrero/Proyecto_Mentoria_FAMAF_2025.git