MEMORIA DEL PIPELINE DE PREPARACIÓN DE DATOS

Proyecto EDA – Suicide Rates (1985–2015)

1. Título

Pipeline de preparación, integración y análisis del dataset mundial de tasas de suicidio (1985–2015).

2. Resumen

Se desarrolló un pipeline integral para limpiar, integrar, imputar y preparar un dataset global compuesto por registros de suicidios de la OMS junto con indicadores socioeconómicos, demográficos y geográficos de países de todo el mundo. El resultado final es un dataset robusto —master_final_from_original.csv— optimizado para análisis exploratorio, visualizaciones avanzadas y modelado predictivo.

3. Origen y alcance de los datos

Ficheros de origen: master.csv y countries of the world.csv. Cada fila representa país, año (1985–2015), sexo, edad y generación, incluyendo variables como población, tasas de suicidio, PIBs, IDH y características estructurales.

4. Principios metodológicos del pipeline

- Conservar máxima información.
- Imputar por país y serie temporal.
- Documentar cada transformación.
- Minimizar sesgos.
- Asegurar reproducibilidad.

5. Etapas del pipeline

- Estandarización de encabezados.
- Limpieza estructural y filtrado temporal.
- Conversión y limpieza numérica.
- Imputación avanzada (HDI, PIBs).
- Winsorización 1%–99%.
- Relleno de nulos con mediana.
- Transformaciones logarítmicas.
- Codificación de categóricas.
- Generación de artefactos analíticos.

6. Validación del pipeline

Evaluación mediante comparación de conteos, pruebas de imputación, inspección visual de outliers y validaciones de interpolación vs. alternativas.

7. Limitaciones

• La interpolación no captura rupturas abruptas.
• El target encoding puede generar leakage si no se valida correctamente.
• Winsorización puede ocultar extremos relevantes.

8. Conclusión general

El pipeline produce un dataset científicamente defendible, adecuado para análisis descriptivo, causal y predictivo, manteniendo coherencia y robustez estadística.