Obtención, Limpieza y Transformación de Datos con Numpy y Pandas

1. Justificación del uso de NumPy y Pandas

En este proyecto utilicé NumPy para generar y manejar datos numéricos de forma eficiente, ya que permite trabajar con grandes volúmenes de información de manera rápida y optimizada. Pandas fue fundamental para la manipulación de los datos, ya que su estructura de DataFrame facilita la exploración, limpieza, transformación y análisis de la información.

2. Descripción del dataset y fuentes externas

El dataset principal fue generado de forma ficticia utilizando NumPy, simulando información de clientes como edad, cantidad de compras y monto total gastado. Además, se integraron datos provenientes de archivos CSV, Excel y una tabla obtenida desde una página web utilizando read_html, lo que permitió trabajar con múltiples fuentes de datos.

3. Técnicas de limpieza y transformación

Se identificaron valores nulos en columnas clave y se aplicaron técnicas de imputación utilizando el promedio. También se detectaron y eliminaron outliers mediante el método del rango intercuartílico (IQR). Posteriormente, se realizaron transformaciones como eliminación de duplicados, creación de nuevas columnas, normalización de variables y conversión de tipos de datos.

4. Decisiones tomadas y desafíos encontrados

Uno de los principales desafíos fue la integración de datos desde diferentes formatos y la gestión de valores faltantes. Se decidió utilizar tipos de datos compatibles para evitar errores y asegurar la consistencia del dataset. Además, fue necesario validar que los datos unificados mantuvieran coherencia entre las distintas fuentes.

5. Resultados obtenidos y estado final del dataset

Como resultado final se obtuvo un dataset limpio, estructurado y enriquecido, listo para su análisis o uso en modelos de ciencia de datos. El proceso permitió transformar datos desordenados en información confiable y reutilizable.