ABP MÓDULO 3

Resumen del Flujo de Trabajo: Procesamiento Integral de Datos con Python

Descripción breve

Informe de implementación en Python de un caso práctico de manipulación y preparación de datos reales.
Objetivo es demostrar capacidad para abordar proyectos de preparación de datos,
una de las habilidades más demandadas en la industria de datos y tecnología. Se destacan:

- Las técnicas aplicadas y librerías utilizadas.
- Las principales problemáticas encontradas y cómo fueron resueltas.
- Capturas del código, el dataset limpio y las visualizaciones de los resultados.

Garcy Valenzuela
garcyv@gmail.com

Contenido

Documento Resumen del Flujo de Trabajo: Procesamiento Integral de Datos con Python

1. Justificación del Uso de NumPy y Pandas

El proyecto requirió la implementación de un *pipeline* robusto y eficiente para la manipulación y preparación de datos. La elección de **NumPy** y **Pandas** como herramientas exclusivas responde a su condición de pilares fundamentales en el ecosistema de ciencia de datos de Python:

- **NumPy:** Se utilizó para la generación inicial de datos ficticios y para operaciones numéricas y estadísticas de alto rendimiento (ej. cálculo de Z-scores, medianas, sumas). Su eficiencia radica en las operaciones vectorizadas y su núcleo escrito en C, superando a las listas nativas de Python en velocidad y uso de memoria para datos numéricos.
- **Pandas:** Es la herramienta principal para la fase de ETL (Extracción, Transformación, Carga). Su estructura DataFrame permitió la gestión intuitiva de datos tabulares, facilitando tareas complejas como la unificación de fuentes (merge, concat), la limpieza (fillna, drop_duplicates) y la reestructuración (groupby, pivot_table).

2. Descripción del Dataset Generado y Fuentes Externas Integradas

El flujo de trabajo partió de datos generados sintéticamente y se enriqueció con fuentes externas:

Fuente Original	Tipo de Dato	Contenido Principal	Lección de Integración
NumPy Generado	NumPy Array (.npy)	Datos de Clientes (ID, Edad, Sexo) y Transacciones (ID, Monto, Tipo)	Lección 1 y 2
Archivo Excel	.xlsx	Datos Complementarios (Puntuación Crediticia, Nivel Socioeconómico)	Lección 3
Simulación Web	HTML/CSV	Mapeo de Regiones	Lección 3

El *dataset* final consolidado tiene aproximadamente 100 registros únicos de clientes combinados con 500 registros de transacciones, estructurados de manera relacional.

3. Técnicas Aplicadas para la Limpieza y Transformación

Se aplicaron diversas técnicas de *Data Cleaning* y *Data Wrangling* a lo largo de las lecciones 4 y 5:

- **Gestión de Nulos:** Se identificaron valores nulos (.isnull().sum()) y se trataron mediante:
    - **Imputación por Mediana:** Para la variable numérica Puntuacion_Crediticia.
    - **Imputación por Moda:** Para la variable categórica Nivel_Socioeconomico.
- **Detección de Outliers:** Se usaron métodos estadísticos (Z-score y Rango Intercuartílico - IQR) para identificar valores atípicos en la puntuación crediticia.
- **Estandarización y Mapeo:**
    - Los códigos numéricos de sexo (0/1) se mapearon a etiquetas de texto (Femenino/Masculino) usando .map().
    - Se aseguró la consistencia de tipos de datos (.astype()).
- **Creación de Variables (Feature Engineering):**
    - Se discretizó la edad en rangos (pd.cut()) para crear la variable Rango_Edad.
    - Se usó .apply() con una función personalizada para crear la Categoria_Crediticia (Regular, Bueno, Excelente).
- **Reestructuración:** Se aplicó groupby() para agregaciones, pivot_table() para crear tablas resumen y melt() para normalizar datos a formato largo.

4. Principales Decisiones Tomadas y Desafíos Encontrados

Desafíos:

- **Heterogeneidad de Fuentes:** El principal desafío fue la integración de diferentes formatos (Numpy arrays, CSVs, Excels) que requerían librerías y métodos de lectura específicos.
- **Inconsistencia de Esquema:** Las fuentes externas introdujeron valores nulos y formatos inconsistentes que debieron ser estandarizados antes del análisis.

Decisiones Clave:

- **Imputación Robusta:** Se eligió la mediana sobre la media para la imputación numérica para minimizar el impacto de posibles *outliers* en los datos imputados.
- **Retención de Outliers:** Se decidió identificar, pero no eliminar, los *outliers* de puntuación crediticia, ya que representan datos reales valiosos para un modelo predictivo, en lugar de errores de entrada.

5. Resultados Obtenidos y Estado Final del Dataset

El resultado del proyecto es un **proceso ETL automatizado y modularizado** que produce un conjunto de datos de alta calidad.

Estado Final del Dataset (dataset_final_analisis.csv):

- **Limpio:** Cero valores nulos y duplicados eliminados.
- **Consistente:** Tipos de datos correctos y formatos estandarizados (ej. etiquetas de sexo en texto).
- **Enriquecido:** Contiene nuevas variables calculadas (Rango_Edad, Categoria_Crediticia) listas para el análisis segmentado.
- **Estructurado:** Exportado en formatos CSV y Excel listos para ser consumidos por herramientas de *Business Intelligence* o modelos de *Machine Learning*, cumpliendo el objetivo principal del proyecto.