

*Ejercicio de Laboratorio 6: ETL – P1*

Instituto Politécnico Nacional.

Escuela Superior de Cómputo.

Licenciatura en ciencia de datos.

Bases de Datos Avanzadas

Emiliano López Méndez.

**Introduccion**

En esta práctica, trabajaremos con un archivo de datos en formato CSV que contiene información sobre ventas, y nuestro objetivo es aplicar técnicas de limpieza y manipulación de datos utilizando la biblioteca pandas en Python. En particular, nos enfocaremos en la eliminación de filas con valores nulos en la columna state y en el almacenamiento del DataFrame procesado en un archivo XML. Este ejercicio es fundamental para comprender los procesos de transformación de datos que se utilizan en proyectos de ciencia de datos y análisis de datos, donde la integridad y la organización de la información son claves para obtener resultados precisos y confiables.

A lo largo de esta práctica, no solo aplicaremos habilidades técnicas de manejo de archivos CSV y XML, sino que también comprenderemos la importancia de los procesos de ETL (Extract, Transform, Load), específicamente en la fase de transformación, en la que se asegura que los datos estén en el formato adecuado antes de ser utilizados en análisis posteriores o cargados en sistemas de almacenamiento de datos.

.

**Desarrollo de la Actividad:**

**Carga de los Datos**: En primer lugar, importamos la biblioteca pandas y utilizamos la función read\_csv para cargar el archivo sales\_data.csv en un DataFrame. Esta función nos permite leer el archivo CSV y tener la información en memoria en una estructura tabular, lo que facilita su manipulación.

**Limpieza de Datos**: Identificamos y eliminamos las filas en las que la columna state contiene valores nulos o vacíos. Para ello, utilizamos el método dropna, que elimina las filas con valores faltantes en una columna específica. Este paso es crucial, ya que los valores nulos pueden afectar el análisis de los datos y llevar a resultados incorrectos si no se manejan adecuadamente.

**Almacenamiento en Formato XML**: Una vez que los datos están limpios, guardamos el DataFrame resultante en un archivo XML utilizando el método to\_xml. Este archivo se almacena en una ubicación diferente a la del archivo original para mantener una organización clara de los datos procesados y para que el archivo limpio esté disponible para futuras consultas o análisis.

**Parte 1 Carga de archivo**

Pantalla de computadora con números

Descripción generada automáticamente con confianza mediaEn esta sección se procedemos a cargar el archivo csv como data frame para poder manejar con mucha facilidad y poder modifcarlo mejor

**Parte 2 Consultar espacios nulos en State y quitarlos**

En esta etapa, se realizan consultas sobre los datos nulos de la columna STATE y los vamos a quitar

Texto

Descripción generada automáticamente con confianza bajaTexto

Descripción generada automáticamente

**Parte 3 Guardamos el archivo como xml**

En esta parte del laboratorio cambiamos el archivo csv a xml y guardarlo en otra carpeta nueva

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Chat o mensaje de texto

Descripción generada automáticamente

**Conclusión**

En esta práctica, logramos procesar y limpiar un conjunto de datos utilizando herramientas fundamentales de pandas en Python, enfocándonos en la eliminación de valores nulos y el almacenamiento en distintos formatos. A través de la eliminación de las filas con valores nulos en la columna state, comprendimos la importancia de asegurar la calidad de los datos, ya que los valores faltantes pueden introducir errores en análisis y decisiones posteriores.

Además, el proceso de conversión del DataFrame a formato XML destacó la versatilidad de pandas y nos brindó una experiencia práctica en la manipulación y transformación de datos para prepararlos para almacenamiento o integración en otros sistemas. Resolver los problemas técnicos, como la codificación y la instalación de bibliotecas adicionales, nos permitió afianzar habilidades para manejar situaciones comunes al trabajar con archivos de distintos orígenes y formatos.