Instrucciones Miniproyecto 1:

Procesamiento y consolidación de datos. Práctica ETL (Extract, Transform and Load).

Descripción

El caso que se presenta a continuación consiste en poblar una dimensión de clientes de una tarjeta de crédito en Taiwán, con información de transacciones entre abril de 2005 a septiembre de 2005. Esta empresa posee 50 clientes identificados con ID de cliente, límite de crédito otorgado, sexo, nivel educacional, estado civil, edad y estado de pagos mensuales. Ahora, esta empresa tiene 4 fuentes de datos donde se complementa la información, por lo que ninguna es suficiente por sí misma.

Las fuentes de datos corresponden a 4 archivos csv, que se encuentran en el sitio web del curso. Lo que se pide en este miniproyecto es que se realice una conexión a las 4 fuentes de datos, se limpien los datos, se consolide la información y se obtenga una base de datos única. Finalmente, se procederá a exportar un archivo csv o Excel con el resultado. Hay que considerar si efectivamente todas las bases de datos entregan información relevante para la consolidación del set final.

Se espera como entregable un archivo .ipynb y el archivo de salida en formato csv o Excel.

Trabajo a realizar

Usando Python en jupyter notebook o Google Colab, los pasos para realizar este miniproyecto son:

1. Cargar y revisar cada fuente de datos, ver su contenido, características y definir si será utilizado para complementar la información. Debe justificar la toma de decisión.
2. Filtrar los atributos que deben ser utilizados: ID cliente, límite de crédito otorgado, sexo, nivel educacional, estado civil, edad y estado de pagos por mes.
3. Unir las fuentes de información a partir de un atributo único.
4. Limpiar y depurar la base de datos unida, verificando datos faltantes, datos Nan, desconocidos, erróneos, duplicados y datos redundantes. Decidir qué hacer con los datos donde no hay información, ya sea imputar el dato o eliminar el cliente. Debe justificar esta toma de decisión.
5. Homogeneizar y normalizar datos. Dejar los datos con la estructura indicada en la tabla 1. Respetar la estructura y los espacios señalados. Finalmente normalizar, justificando el tipo de normalización escogida.
6. Generar un archivo de salida csv o Excel.