**Instituto Tecnológico y de Estudios Superiores de Monterrey**

**Analítica de datos y herramientas de inteligencia artificial II (Gpo 101)**

# Limpieza de Bases de Datos


Amairany Rodríguez Huerta | A01702927
Luis Pablo Padilla Barbosa | A00572040
Renata Emilia Chávez Martínez | A01351716


21 abr 2023

## Datos de Facturación

El archivo cuenta con datos nulos en las siguientes columnas:

- CVE_VEND (48)
- FECHA_ENT (2)
- FECHA_CANCELA (10,537)

Para el caso de la clave del vendedor, se aplicó la función unique() para ver el tipo de datos que se encontraban en esa columna y con ello se logró apreciar que se encontraban de forma sucesiva los números del 1 al 12, pero faltaba el valor de 8, por lo que se concluyó que los datos faltantes correspondían a esa clave del vendedor. Por esta razón es que se aplicó la técnica en donde se asigna un valor numérico en concreto a todos los valores nulos.

En la columna de fecha de entrada, se consideró que la mejor manera de llenar los datos nulos era usando el valor que se encontraba en la columna de fecha del documento, puesto que en las reuniones con el socio formador se comentó que era la columna de fecha más importante dentro de ese dataset. Por ello es que por medio de un isnull() se obtuvieron los índices de las variables con datos nulas y posteriormente con la función iloc se completaron con los valores de la columna de fecha del documento correspondientes a esos mismos índices.

Por último, en el caso de la columna de fecha de facturación, se había mencionado en las reuniones con el socio formador que las filas que no tenían un valor en fecha de cancelación era porque se encontraban vigentes, así que se decidió, por medio de la técnica para asignar un string en concreto, asignarles una la fecha del 1 de enero del 2050, con el objetivo de que se pueda identificar fácilmente que es una fecha irreal y que no se debe considerar en caso de que se quiera analizar esa columna.

## Detalle precios y productos fabricados 2022

- El archivo solo cuenta con dos datos nulos en la columna de “NOMBRE_VENDEDOR”.
- Se sustituye estos datos nulos por “no existe”
- Finalmente ya no contamos con ningún valor nulo

Se sustituyeron estos valores nulos en dicha columna con un “no existe” debido a que es una columna con datos string. Además esta columna menciona el nombre del vendedor; por lo que, si no lo contiene se supone que es un vendedor que no existe o no esta activo dentro de la empresa.

## Gastos y costos 20-23

Considerando la estructura del documento se descargó la base de datos y al momento de hacerlo se eliminaron 5 filas considerando que estas no tenían los títulos de cada una de las columnas. Las columnas que se tienen datos nulos son:

- FOLIO (189)
- GASTO (2502)
- TC (391)
- IMPORTE (34)
- IVA (268)
- TIPO (1)
- PÓLIZA (3321)

Una vez que se localizaron los datos nulos se detectaron dos columnas en las cuales los datos nulos eran mayores al 80% de la muestra por lo cual se considera adecuado eliminar las columnas dado que no se tiene la información necesaria para hacer un estudio de las mismas. Una vez considerando esto es que se decidió depende de la información contenida en la columna es que se cambiaría por un código específico, una mediana de la muestra o media de la misma.

Para la columna de TIPO, se tomó el valor anterior considerando que no se tenía esa información pero podemos asumir que la variable se comporta similar a su vecino. En cuestiones de IVA se tomó el 13% respecto al TOTAL SAT que se tiene registrado de cada una de las compras para poder tener una idea respecto de cuanto es lo que realmente se cobra en cada uno de los productos.

En cuestión al IMPORTE se tomó el promedio considerando que lo que se gana en esas ventas es el promedio de lo que se ganó de manera global en 3 años. FOLIO se asignó un folio de 0 a esos folios que no se tienen registrados en la matrices de ventas y en TC se realizó algo similar asignando un 1 considerando que la mayoría se tiene como 1.