Tecnológico de Monterrey

Analítica de Datos y Herramientas de Inteligencia Artificial

Actividad 2.2: Reporte escrito “Valores Nulos”

Alberto Rodriguez Porras | A01721497

Brenda Villa Campos | A01732238

Alexa Bustamante de la Cruz | A01173639

Melisa Hernández Cid | A01732341

21 de abril del 2023

Profesor: Alfredo García Suárez

En el proyecto que se ha trabajado a lo largo de los meses anteriores con Calor y Control se ha trabajando con múltiples bases de datos las cuales hemos limpiado anteriormente. Con los nuevos conocimientos que estamos obteniendo, hemos tomado estas bases de datos limpias con nulos para poder manipularlas con estas nuevas técnicas vistas en clase. Por lo tanto, estas bases de datos ya tiene una selección de columnas, limpieza de filas enteras duplicadas y filtrando únicamente los datos de interés de la investigación. Partiendo de aquí, avanzamos a la limpieza de datos nulos.

Archivos Gastos y costos 20-23:

En la base de datos del 2020 se presenta la columna “POLIZA” con un total 3321 datos nulos, seguido de “IVA” con 268 datos nulos. El resto de las columnas no superaban los 40 datos nulos por lo que para “FOLIO”,”GASTO”,”IMPORTE”,”IVA” y “TIPO” se utilizó el reemplazar los datos con “-” o en su defecto para los tipos de datos numéricos con “0”.

En cuanto a la columna de TC ya que estamos hablando de un tipo de cambio, nos guiamos de la columna “TOTAL_SAT” , que cuando tienen un tipo de cambio, es una cantidad menor. Por lo tanto, observamos los comportamientos de los datos por medio de la media y vemos que los valores son incluso mayores. Esto indica que el tipo de cambio usualmente, muy seguramente no son distintos a 1, pues si lo fuera se esperaría una media menor a la media del comportamiento de “TOTAL_SAT” entera (con y sin nulos). Por lo tanto, 1 sería una buen predicción en TC.

Finalmente para la columna de “POLIZA” se llevo a cabo una discusión en equipo acerca de la importancia de los datos para la finalidad del proyecto , resultando en que no se consideraban relevantes para el mismo, aunado a esto al contar con menos del 1% de los datos, decidimos eliminar como tal la columna.

La base de datos del 2021, no tiene ningún dato nulo, por lo que no se necesitó hacer una limpieza en este tema.

La base de datos del 2022 tuvo 314 valores nulos en la columna TC. TC es el tipo de cambio, que se multiplica con la cantidad de “TOTAL_SAT”. Por lo tanto los valores en “TOTAL_SAT”, cuando tienen un tipo de cambio, es una cantidad menor. Por lo tanto, observamos los comportamientos de los datos por medio de la media y vemos que los valores son incluso mayores. Se utiliza la misma predicción que para la primera base de datos en la columna “TC” y se ussa el valor “1” como predeterminado para rellenar los datos faltantes.

La base de datos del 2023 tuvo 8 datos nulos en la columna “TIPO GASTO”. Se observa que los valores de “TIPO GASTO” son categóricos y difícilmente se le podrá predecir a los valores faltantes exitosamente dado a la alta cantidad de variantes. Por lo tanto, se convertirán en una cadena que se logre identificar fácilmente que no están definidos. Por lo tanto se reemplazó con “--”.

Detalle Precios y Productos Fabricados 2022:
Esta base de datos, tras la limpieza mencionada previamente, no presenta ningún valor nulo.

Datos de Facturación
El archivo “Datos de Facturación” presentaba valores nulos sólo en tres columnas: “CVE_VEND”, clave del vendedor con 48 valores nulos, “FECHA_ENT”, fecha de entrega del producto con 2 valores nulos y “FECHA_CANCELA”, fecha de cancelación de orden el cuál tenía el mayor número de valores nulos con un total de 10537.

Para el proceso de eliminación de nulos se comenzó con la columna “CVE_VEND”, en el cual habían valores del 1-5 por lo que los valores nulos fueron sustituidos por el número “0”. Por otro lado, tanto en “FECHA_ENT” y “FECHA_CANCELA” no sería apropiado reemplazarlo por otra fecha o algo similar, por lo que en ambos casos se reemplazaron los valores nulos por “NA”.