PROCESO DE LIMPIEZA DE DATOS

Revisión del dataset:

El dataset original presentaba muy poca incongruencia de datos, para ser más específico, de las columnas que forman parte de este proyecto solo la columna ind_evp presentaba algunas anomalías.

Análisis del dataset:

Los registro que vamos a utilizar, refieren a los porcentajes de incidencia de cada concepto, y los mismos son cargados en el archivo en formato decimal. Por tal motivo todos los valores de las variables independientes se encontraran dentro del rango de 0 a 1.

Identificación de anomalías:

En el caso de la columna ind_evp, el error consistía en la carencia del 0, al comienzo de algunos números, lo que creaba cifras de magnitudes enormes (por ej. 10000425796) si consideramos que los valores de nuestros datos se encuentran entre 0 y 1.

Limpieza de datos:

La limpieza de datos, al ser sencilla, se realizó directamente en excel
Se procedió a generar una copia del archivo original y sobre ella se trabajó.
Se colocaron filtros en los encabezados de las columnas que íbamos a utilizar para el proyecto y se observaron los registros de los filtros, poniéndose en evidencia como mencionamos anteriormente que los valores se encontraban entre 0 y 1 excepto algunos casos en la columna ind_evp. Para esos casos se filtraron específicamente esos datos anómalos y luego se creó una nueva columna, en la que se aplicó la siguiente formula *="0,"&celda correspondiente (ej I2628)*, se extendió la formula y el resultado obtenido fue el numero anómalo antecedido por un 0, .Luego se copiaron los valores ya transformados y se pegaron en formato numero en la columna en cuestión (ind_evp). Con estos simpes pasos realizamos la limpieza de los datos.