KONBRIEFING

Este documento redacta las acciones para cada columna antes de entrenar un modelo de Machine Learning en el caso de este CSV.

1. **Date**: Descomponer en las columnas: **Año**, **Mes**, **Día**. La fecha puede contener información importante sobre estacionalidad o tendencias temporales. Luego, elimina la columna original, ya que la fecha completa ya no será necesaria.
2. **Title**: Esta columna contiene texto libre, lo que puede ser difícil de utilizar directamente en modelos de ML. Si se desea utilizar, podrías convertirla en características numéricas a través de técnicas de procesamiento de lenguaje natural (NLP), como **TF-IDF** o **Word2Vec**.
3. **Description**: Esta columna contiene texto libre, lo que puede ser difícil de utilizar directamente en modelos de ML. Si se desea utilizar, podrías convertirla en características numéricas a través de técnicas de procesamiento de lenguaje natural (NLP), como **TF-IDF** o **Word2Vec**. Además, contiene la **región**, separar esta a una columna nueva llamada **Region**. Esta nueva columna **Region** se debe agrupar en continentes. Convertir la columna en variables **dummy** para representar continentes. Se debe usar **One-Hot Encoding** para los continentes, lo que permitirá al modelo aprender patrones geográficos sin complicarse con muchos países diferentes.
4. **References**: Extraer la información. Agregar dos nuevas columnas: una con el nombre de la página o fuente; otra con la **URL** del artículo.
5. **Date Uploaded**: La fecha en la que el incidente fue cargado al sistema no es relevante.

CASO DE VARIABLES CATEGÓRICAS:

Codificación de la columna a valores numéricos. Esta columna es categórica. Para utilizarla en un modelo de ML, se tiene que convertir en un formato numérico. Esto se puede lograr mediante **Label Encoding** si las categorías son pocas, o **One-Hot Encoding** si hay muchas categorías diferentes asociadas a las filas.

a. **Label Encoding**: Asigna un número único a cada categoría, convirtiendo las etiquetas en valores enteros.
b. **One-Hot Encoding**: Crea una columna binaria (0 o 1) para cada categoría, representando la presencia o ausencia de esa categoría.

Se recomienda primero una **agrupación** para reducir el número de valores únicos en la fila.