ICSSTRIVE

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

1. **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**.
2. **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.
3. **Locations**: 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. **Estimated Cost:** Interpretar el texto, extraer los valores numéricos y asignar a los valores a **categorías cualitativas**.
5. **Victims:** En caso de que fuera una columna relevante, extraer **patrones clave** y agrupar en categorías comunes.
6. **Type of malware**: Extracción de dos nuevas columnas: **Malware title** con el nombre del malware y **Malware link URL** con el dominio del enlace asociado. En la columna con el **URL**, extraer el dominio. En la otra, codificación a valores numéricos (primero se recomienda agrupar) con **Label Encoder**.
7. **Threat source**: Se debe extraer la información en nuevas columnas y aplicar lo que se considere óptimo a cada una de ellas.
8. **Industries:** Se debe de limpiar ya que se encuentra en listas. 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.

9. **Impacts:** Caso similar a **Industries**.
10. **References**: Extraer el dominio de la **URL**.
11. **Data source link url**: La columna se utiliza como el identificador principal (ID) del DataFrame creado.