CISSM

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

1. **ID**: La columna se utiliza como el identificador principal (ID) del DataFrame creado.
2. **Event 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**. Hacer esto en caso de que el modelo esté diseñado para procesar texto.
3. **Event 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.
4. **Actor**: 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.
5. **Actor Type**: Codificación de la columna a valores numéricos. Es categórica y debe ser convertida a un formato numérico para el modelo. Dado que tiene categorías limitadas, es ideal usar **Label Encoding**.
6. **Event Type**: Codificación de la columna a valores numéricos. Es una variable categórica. Como con las otras columnas categóricas, debes convertirla en un formato numérico para usarla en el modelo.
7. **Organization**: Esta columna podría ser categórica o tener demasiados valores únicos para utilizarla directamente en un modelo de ML. Si se decide incluirla, se podría hacer lo siguiente: **Agrupación por categorías**, **Frecuencia de aparición** o **Target Encoding** (Codificación basada en el objetivo). Si se considera poco relevante para el análisis, no incluir.
8. **Event Subtype**: Si es relevante para el análisis incluir en este.
9. **Motive**: Codificación de la columna a valores numéricos. Es una variable categórica. Decidir si hacer: **Label Encoding** u **One-Hot Encoding**.
10. **Event Source**: Extraer el dominio de la **URL**.
11. **Country**: 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.
12. **Industry**: No se haría nada ya que se tiene su codificación ordinal en la columna: **Industry Code**. Eso sí, se debe de mirar que en Industry hay un total de 21 variables únicas y en la otra hay un total de 25.
13. **Industry Code**: Asegurarse de que esta columna está en formato numérico.
14. **DNX ID**: No aporta información, esto es un identificador.
15. **Date Uploaded**: La fecha en la que el incidente fue cargado al sistema no es relevante.