CYBER INCIDENTS MASTER

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. **Year**: Usar como columna base para corregir inconsistencias en **Date**.
4. **Target 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.
5. **Region**: No tomar en cuenta para el modelo ML. Según lo indicado, la región no aporta información útil adicional después de agrupar los países por continentes.
6. **Dnx Country**: Similar a **Target Country**, esto reduce la cardinalidad y permite al modelo capturar tendencias geográficas más amplias.
7. **Industry**: 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.

8. **Dnx Industry**: Caso similar que **Industry**.
9. **Victim**: Si se decide incluir, considerar **agrupar víctimas frecuentes o relevantes**.
10. **Type of Attack**: Caso similar que **Industry**.
11. **Threat Source**: Caso similar que **Industry**.
12. **Malware**: Caso similar que **Industry**.
13. **Impact**: Considerar su eliminación, todas las filas contienen `[]`, simbolizando una lista vacía, es decir, como si fueran valores nulos.
14. **Tisafe Score**: Considerar su eliminación, todas las filas contienen `[]`, simbolizando una lista vacía, es decir, como si fueran valores nulos. No aporta información al modelo.
15. **References**: Extracción de la información. Crear columnas nuevas con lo que se extraiga y examinarlas.
16. **Source Database**: En caso de que sea relevante en el estudio aplicar el uso de la codificación numérica, es decir, **Label Encorder** o **One-Hot Encoding**.
17. **Source Database Incident ID**: Similar a otros identificadores, no aporta información para el modelo.

18. **Date Uploaded**: La fecha en la que el incidente fue cargado al sistema no es relevante.
19. **Dnx ID**: La columna se utiliza como el identificador principal (ID) del DataFrame creado.