HACKMAGEDDON

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

1. **Date Reported:** 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. **Date Occurred:** 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.
    a. Si se viera que hay muchos nulos, eliminar.
3. **Date Discovered**: 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.
    a. Si se viera que hay muchos nulos, eliminar.
4. **Author:** 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.

5. **Target:** La columna ya está agrupada y representada de manera más clara en la columna **Target Class**. Se utilizará esa en el modelo.
6. **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**.
7. **Attack:** La columna ya está agrupada y representada de manera más clara en la columna **Attack Class**. Se utilizará esa en el modelo.
8. **Target Class:** Caso similar a **Author**. Hacer caso
9. **Attack Class:** Caso similar a **Author**.
10. **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.
11. **Link**: Extraer el dominio de la **URL**.
12. **Tags**: Si se considera relevante, aplicar técnicas de **NLP** para extraer información de las etiquetas. Caso similar a **Description.**