MÁSTER EN IA APLICADA A LA CIBERSEGURIDAD

MÓDULO 5. GESTIÓN DE DATOS Y PRIVACIDAD EN MACHINE LEARNING

TAREA 1

MÁSTER EN
INTELIGENCIA ARTIFICIAL APLICADA A LA CIBERSEGURIDAD

Campus Internacional CIBERSEGURIDAD

ENIIT INNOVA IT BUSINESS SCHOOL

UCAM Universidad Católica de Murcia

Introducción

En esta primera evaluación **vamos a construir un dataset desde cero**.

El arco argumental de la primera evaluación te llevara por todo el tema 2 y 3. Tocaremos algo del 5 pero de forma parcial.

La segunda evaluación nos induce a aplicar lo aprendido en el tema 4, 5 y 6.

Material

Tenemos un archivo comprimido con varios archivos PCAP (capturas de tráfico de red)

El archivo se llama: pcaps.zip

Estos pcaps son el tráfico capturado de una sandbox detonando malware real.

Es decir, la fuente de los datos en bruto es REAL, nada sintético. De hecho, no te recomiendo que reproduzcas el pcap dado que creará conexiones de red a infraestructura de malware.

Instrucciones

Puedes usar IA para generar código.

Deberás explicar y **justificar ampliamente** las decisiones tomadas.

Deberás hacer **capturas de pantalla** de todos los procesos que emplees, paso a paso y **comentarlas**.

Ejercicio que técnicamente esté bien pero no esté suficientemente justificado y documentado, no se evaluará.

Ejercicio

Contexto

Has recibido varios pcaps capturados durante detonaciones de malware en un laboratorio. Estos archivos contienen tráfico real, con ruido, duplicados y datos sensibles. Tu misión es transformar esa materia prima en un dataset utilizable para análisis de ciberseguridad, siguiendo las fases del ciclo de vida del dato.

Objetivos de aprendizaje

Comprender el ciclo de vida de los datos en ciberseguridad (Cap. 2).

Aplicar métodos de recopilación y creación de datasets (Cap. 3).

Realizar procesos básicos de anonimización (Cap. 5.4).

Preparar un dataset limpio y cargado en una base de datos relacional.

Tareas a realizar

Extracción: convierte el pcap en CSV/JSON con los campos:

timestamp, src_ip, dst_ip, protocol, src_port, dst_port, length.

Añade los siguientes campos si están presentes en el pcap:

dns_query, http_host y http_path, user_agent

Limpieza: elimina, duplicados, trata valores nulos y detecta outliers.

El dataset incluye mucho tráfico de red normal (ARP, SSDP, LLMNR, etc.) que no es relevante para ciberseguridad porque sabemos que ninguna de las muestras lo emplea.

Deberás filtrar ese ruido y quedarte solo con tráfico de interés para el análisis de malware: consultas DNS, conexiones HTTP/HTTPS, y flujos TCP/UDP útiles, etc.

Anonimización básica: aplica al menos una técnica sobre src_ip y dst_ip (hash, tokenización o enmascaramiento).

Carga en base de datos: diseña una tabla en SQLite/Postgres e inserta el dataset limpio.

Consultas que deberás hacer y mostrar una vez cargues los datos en la base de datos:

- Total de registros
- Top de direcciones IP de destino
- Dominios más consultados (posiblemente verás que los primeros puestos no son maliciosos)
- Puertos de destino más comunes
- Media y máximo de longitud de paquete (length)
- Distribución de protocolos

Entregables

Dataset final (CSV/JSON).

Script Python (extracción + limpieza + anonimización básica + carga en BD).

Informe en PDF.