UNIVERSIDAD DE LAS FUERZAS ARMADAS ESPE

Departamento de Ciencias de la Computación

Lectura y Escritura de Textos Académicos

Tema:

Análisis de dataset CSE-CIC-IDS2018

Autor:

Moisés Benalcázar
Mateo Medranda
Bryan Quispe

NRC: 29765

Ecuador 2025-10-12

Tabla de Contenidos

1. Introducción

El presente informe tiene como objetivo analizar el dataset CSE-CIC-IDS2018 generado por la University of New Brunswick (UNB), el cual recopila información sobre tráfico de red tanto normal como malicioso, incluyendo diferentes tipos de ciberataques.

Este dataset simula distintos escenarios mediante la generación de perfiles dividiendo en “B-profiles” y “M-profiles” dependiendo si simula un usuario en actividades normales o un atacante, de modo que permite analizar comportamientos típicos de la red, así como detectar patrones de intrusión. Los ciberataques incluidos en el dataset comprenden desde ataques de denegación de servicio (DDoS) hasta ataques dirigidos a aplicaciones web, fuerza bruta y explotación de vulnerabilidades conocidas, lo que lo convierte en una herramienta útil para la investigación en seguridad informática, o para la predicción de probabilidad de ciberataques según un perfil de red.

Dentro del dataset se exponen diferentes características en los ataques, así como en el tráfico normal, basado en este aspecto, se busca analizar y encontrar una posible variable objetivo o categorias para clasificación, lo cual permite identificar patrones o realizar predicciones.

2. Descripción del dataset

2.1. Origen del dataset

El dataset CSE-CIC-IDS2018, el cual es objeto de análisis en el presente informe, fue creado por la UNB o por sus siglas en ingles “University of New Brunswick” en Canadá. Los datos dentro del mismo fueron obtenidos de forma íntegra en el año 2018, dada la necesidad de conocer un conjunto de datos referentes a múltiples ciberataques dirigidos grandes empresas o instituciones.

El problema radica en que los datos que las empresas obtienen de los ataques que sufren, son clasificados como información sensible, y encontrar datasets de este tipo es dificil, por lo que en el presente dataset se trabaja generando un entorno realista para simular los ataques y obtener toda la información relacionada.

2.2. Estructura de archivos

Los archivos se encuentran en formato .csv, se encuentran clasificados principalmente por fechas, además categorizados en diferentes escenarios, por lo que se puede encontrar archivos donde el tráfico de datos fue normal, mientras que en otros casos se encuentran predominando los diferentes tipos de ciberataques.

La diferencia entre los archivos también distingue su utilidad, y un ejemplo es si se desea utilizar aquellos archivos con el 100% de tráfico normal, el cual será útil para entrenar

en detección de ataques, pero dada la categoría del archivo, este permitirá evitar falsos positivos en las detecciones.

2.3. Columnas importantes

Dentro del dataset, se encuentran 80 columnas y solo un archivo cuenta con 84 columnas, donde las 4 columnas de diferencia son: ip de la víctima, ip del atacante, puerto de origen y puerto de destino.

Pero pese a que todas las columnas brindan información, debemos tomar en cuenta la importancia de esta, y esto se reduce por recomendación de los creadores del dataset a 8 columnas las cuales son:

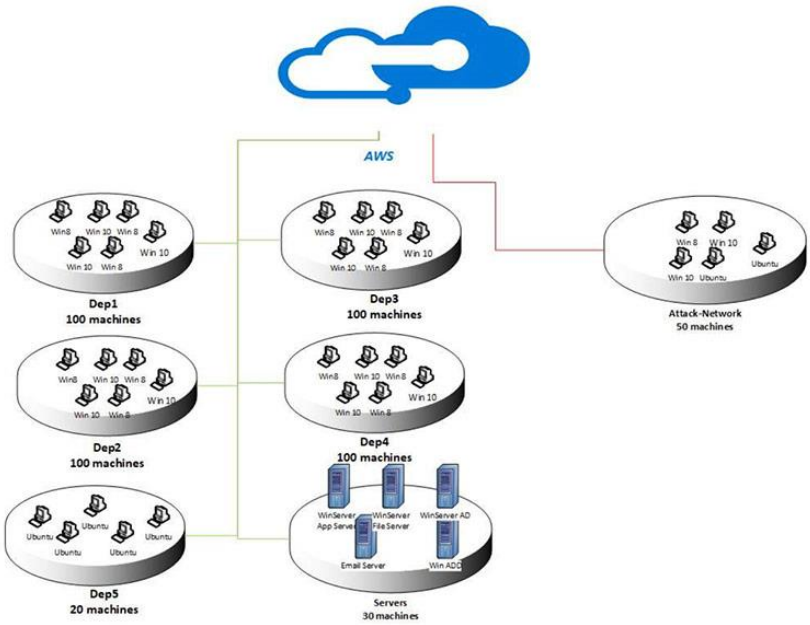
- Dst Port (Destination port)
- Protocol
- Flow Duration
- Tot Fwd Pkts (Total forward packets)
- Tot Bwd Pkts (Total backward packets)
- Label (Label)

También hay que tomar en cuenta que los registros guardan información del flujo de ida y de vuelta, es decir al momento de realizar las peticiones y al momento de obtener respuesta del servidor, contando información como el número de paquetes, el tamaño de los paquetes y también medidas estadísticas como la desviación estándar, promedio, máximos, mínimos, etc.

La columna “Label” tiene una importancia en particular dado que dentro de la misma se determina si es un tráfico benigno o maligno determinando al mismo tiempo el tipo de ataque que es, por lo que puede ser útil en caso de querer predecir un ataque basados en las características presentadas por el resto de las columnas.

3. Construcción del dataset

3.1. Arquitectura de la red usada



La red usada en el dataset CICIDS 2018 está montada en AWS y simula una empresa con varios departamentos conectados entre sí. Cada departamento tiene diferentes computadoras con sistemas como Windows 7, Windows 10 y Ubuntu, para representar un entorno de trabajo real. Además, hay una sección de servidores donde se manejan servicios como correo, archivos, aplicaciones y Active Directory. Aparte de esto, existe una red de ataque con 50 máquinas que se usa para lanzar diferentes tipos de ciberataques. Esta arquitectura permite generar tráfico realista entre usuarios y atacantes, ayudando a probar y mejorar sistemas de detección de intrusos (IDS).

3.2. Tipos de ataques realizados

El conjunto de datos se introdujo en siete escenarios de ataque diferentes para modelar la penetración real y generar tráfico de red malicioso: fuerza brutal, frecuencia cardíaca, red de robots, denegación de servicios (DOS), denegación de servicios distribuidos (DDOS), ataques web (por ejemplo, información SQL. Estos ataques se realizan utilizando herramientas conocidas: por ejemplo, * Patator * usa ataques de fuerza brutal contra FTP y SSH; * Heartleech* para aprovechar la vulnerabilidad del corazón; * Zeus* y* ares* sobre las actividades de la red de robots; *Lowloris*,*hoic*y*lentamente,*ddos; Además de un ataque web automatizado usando *DVWA* con XSS, inyección SQL y fuerza brutal; Finalmente, para la

infiltración interna, se utiliza E -Past con archivo malicioso, vulnerabilidad, instalación por puerta trasera y escaneo interno con *nmap*. De estos ataques, los más comunes o notables de datos juntos son la fuerza brutal, Dos/DDos y los ataques web, ya que en el entorno real representan vectores frecuentes y permiten estudiar diferentes tipos de anomalías del tráfico. En particular, los ataques Dos/DDos tienden a destacar porque provocan grandes cantidades de tráfico malicioso que rápidamente se opone al tráfico benigno, lo que facilita la detección de anomalías. Además, los ataques web y la infiltración proporcionan métodos más complejos y sutiles que ayudan a evaluar la capacidad del sistema de determinación para determinar una penetración menos obvia.

4. Análisis final

Luego de un amplio reconocimiento de los datos del dataset, se ha llegado a la conclusión de que una posible variable objetivo es la columna “label” ya que esta se encuentra en función de los datos recolectados en el resto de las columnas, tales datos como el número de paquetes, el tamaño de estos, o el tiempo intervalo entre peticiones.

En conclusión, la variable objetivo “label” representa si una serie de peticiones a un servidor, se tratan de ataques o de tráfico normal, y es posible generar modelos de inteligencia artificial que permitan predecir la probabilidad de un ciberataque según el perfil de red, con mucha más precisión al tener datos de tráfico normal para control de falsos positivos, y datos de ciberataques, para clasificarlos.

Tema propuesto:

Predicción de probabilidad de ciberataques según perfil de red

Identificación Problema:

En las redes corporativas, los atacantes suelen realizar escaneos para identificar puertos abiertos y vulnerabilidades antes de lanzar un ataque. Analizando el perfil del tráfico de la red mediante variables como duración de flujos, número y tamaño de paquetes, tiempos entre paquetes y tasas de transferencia, es posible predecir la probabilidad de ciberataques y detectar patrones sospechosos de manera temprana, lo que permite anticiparse a incidentes y fortalecer la seguridad de los sistemas