UNIVERSIDAD DE LAS FUERZAS ARMADAS ESPE

Departamento de Ciencias de la Computación

Lectura y Escritura de Textos Académicos

Tema:

Propuesta de Desarrollo y Solución

Autor:

Mateo Medranda
Bryan Quispe

NRC: 29765

Ecuador 2025-12-01

Índice de Contenido

1. Identificación de Materiales Utilizados

1.1. Datos

Los datos utilizados vienen del dataset de la University of New Brunswick denominado "IDS 2018 Intrusion CSVs (CSE-CIC-IDS2018)", dentro del cual se puede encontrar 80 columnas con diferentes variables representando el tráfico de red, pero la documentación del dataset, sugiere una mayor importancia en 6 variables, las cuales son:

Dst Port (Destination port):

Se refiere al puerto expuesto por la víctima, siendo un punto importante al ser una vulnerabilidad que puede permitir realizar una intrusión.

Sus valores varían en un rango de 0 – 65535 dependiendo de cual está expuesto.

Protocol:

Sus valores varían entre 0, 6 y 17, representando el protocolo usado en la comunicación, se puede observar una clara clasificación entre estos 3 tipos donde cada uno se define como:

0: HOPOPT (usado para tráfico no definido)

6: TCP (usado en conexiones web)

17: UDP (usado en streaming, DNS o VoIP)

Flow Duration:

Dentro de la variable se identifican valores muy bajos o altos dependiendo del tiempo que dure el flujo de información, generalmente TCP finaliza mediante un paquete FIN que interrumpe la conexión, mientras que UDP finaliza tras pasar un tiempo de espera.

Tot Fwd Pkts (Total forward packets):

Representa el número total de paquetes enviados desde el cliente hasta el servidor, teniendo valores variados, pero cabe destacar que una cantidad muy grande de paquetes enviados pueden indicar ataques o escaneos de vulnerabilidades.

Tot Bwd Pkts:

Representa el número total de paquetes enviados desde el servidor hasta el cliente, en respuesta a una solicitud, valores muy altos de estos paquetes pueden representar fallas en el servidor o un escaneo automático.

Label:

Es la variable categórica más importante del dataset, donde se define si el tráfico es benigno o maligno, en caso de que sea maligno, el dataset otorga una clasificación de diferentes tipos de intrusiones.

Sus valores pueden ser:

- Benign
- Bruteforce attack (FTP – Patator)
- Bruteforce attack (SSH – Patator)
- DoS attack (Hulk)
- DoS attack (GoldenEye)
- DoS attack (Slowloris)
- DoS attack (Slowhttptest)
- DoS attack (Heartleech)
- Web attack (DVWA – XSS y Brute-force)
- Infiltration attack (Dropbox download, Nmap, Portscan)
- Botnet attack (Ares RAT)
- DDoS attack (LOIC – UDP, TCP, HTTP) + PortScan

1.2. Software

Como software, se plantea el uso de python con FastAPI como lenguaje de programación de backend para el preprocesamiento, modelado y evaluación, mientras que para una capa de frontend se utilizará React con TypeScript. También se considera el uso de TensorFlow para modelos más complejos como CNN que va a ser utilizado en comparación con Random Forest.

1.3. Hardware

El hardware que se va a utilizar contiene las siguientes características:

- Laptop Acer Nitro 5
- Componentes:
- Procesador: I7-11800H
- RAM: Sodim – ddr4 64 GB
- Tarjeta Gráfica: Rtx 3050Ti
- Almacenamiento: 512Gb nvme + 1TB SSD Sata

1.4. Documentación

El dataset CSE-CIC-IDS2018 se basa en el artículo "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization" de Iman Sharafaldin,

Información bibliográfica :

- **Autores**: Iman Sharafaldin, Arash Habibi Lashkari, Ali A. Ghorbani
- **Institución**: Canadian Institute for Cybersecurity (CIC), University of New Brunswick (UNB), Canadá
- **Conferencia**: ICISSP 2018
- **Páginas**: 108-116
- **ISBN**: 978-989-758-282-0
- **DOI**: 10.5220/0006639801080116
- **Editorial**: SCITEPRESS

2. Identificación de la Metodología/Framework

Como metodología se utiliza CRISP-DM que significa Cross Industry Standard Process for Data Mining, permitiendo llevar al cabo un procedimiento estructurado, siguiendo las siguientes fases:

- Comprensión del negocio
- Comprensión de los datos
- Preparación de los datos
- Modelado
- Evaluación
- Despliegue

El modelo que se va a utilizar corresponde a Random Forest, pero también se va a comparar con un modelo de redes neuronales convolucionales o CNN por sus siglas en inglés.

Para la evaluación del modelo se toma en cuenta las siguientes métricas:

- Accuracy
- F1-Score
- Recall
- Curva Roc

3. Diseño de la Arquitectura