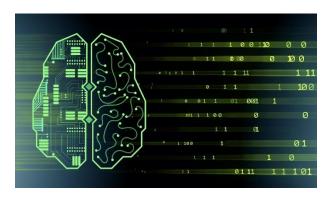
Ciberseguridad con Inteligencia Artificial



Dr. Vitali Herrera Semenets – CENATAV, La Habana, Cuba (<u>vherrera@cenatav.co.cu</u>)
MSc. Felipe Antonio Trujillo Fernández – IBERO, Ciudad de México, México (<u>felipe.trujillo@ibero.mx</u>)
MSc. Joshua Ismael Haase Hernández – IBERO, Ciudad de México, México (<u>joshua.haase@ibero.mx</u>)
Dr. Lázaro Bustio Martínez – IBERO, Ciudad de México, México (<u>lazaro.bustio@ibero.mx</u>)
Coordinación de Ciencia de Datos - Departamento de Estudios en Ingeniería para la Innovación – Ibero
Primavera 2024

Sesión 3

1. Introducción

La detección de anomalías y comportamientos maliciosos es un aspecto crucial en la ciberseguridad moderna. En un entorno digital cada vez más complejo y sofisticado, la capacidad de identificar actividades anómalas y potencialmente dañinas es fundamental para proteger la integridad y seguridad de sistemas y redes informáticas. En esta práctica, se explorarán técnicas y algoritmos de Aprendizaje Automático diseñados para detectar y mitigar amenazas cibernéticas, así como para identificar patrones de comportamiento sospechoso que podrían indicar actividades maliciosas.

2. Objetivo

Aplicar técnicas y algoritmos de detección de anomalías y comportamientos maliciosos en datos relacionados con la ciberseguridad.

3. Indicaciones

- a) Obtención de datos.
 - Descargue el dataset "conn250k.csv" del sitio web del taller. El dataset "conn250k.csv" contiene registros de conexiones de red, con cada registro identificado por un ID único. Incluye información sobre la duración de la conexión, así como la cantidad de bytes transferidos desde y hacia la

fuente y el destino respectivamente. Este conjunto de datos es útil para el análisis de patrones de tráfico de red y la detección de posibles anomalías o comportamientos maliciosos. Las columnas del dataset "conn250k.csv" se describen a continuación:

- record id: Identificador único para cada registro de conexión.
- duration: La duración de la conexión, medida en segundos y redondeada. Por ejemplo, una conexión de 0.17 segundos se registraría como 0 en este campo.
- src_bytes: Número de bytes de datos transferidos desde la fuente hasta el destino; es decir, la cantidad de bytes salientes desde el host.
- dst_bytes: Número de bytes de datos transferidos desde el destino hasta la fuente; es decir, la cantidad de bytes recibidos por el host.
- b) Realiza el Análisis Exploratorio de Datos para entender la naturaleza de los datos.
 - Cree una nueva columna llamada "diff_bytes" que contenga la diferencia entre los bytes enviados (src bytes) y los bytes recibidos (dst bytes).
 - Obtenga las estadísticas de los datos.
 - Represente la nueva columna "diff_bytes" mediante un histograma. Analice su comportamiento.
 - Represente visualmente mediante un scatter plot la relación entre las columnas dst bytes contra src bytes.
 - Represente visualmente mediante un scatter plot la relación entre las columnas src bytes y duration, y dst bytes y duration.
 - Obtenga la matriz de correlación entre las columnas de los datos. Represente la matriz de correlación mediante un mapa de calor. ¿Qué se puede concluir?
- c) Aplica un algoritmo de agrupamiento (por ejemplo, KMeans) para agrupar el tráfico de "conn250k.csv".
 - Visualiza los grupos obtenidos.
 - Entender la naturaleza de los grupos:
 - i. Analiza las características de los grupos obtenidos.
 - ii. Identifica patrones comunes en cada grupo.
 - El dataset "conn250k_anomaly.csv" contiene las etiquetas reales para cada transacción en "conn250k.csv". Cargue "conn250k_anomaly.csv" y compare las etiquetas asignadas (grupos) por el algoritmo de agrupamiento y las etiquetas reales. ¿Qué se puede concluir al respecto?
- d) Considera el dataset "conn250k_anomaly.csv" como etiquetas reales de los datos para un problema de clasificación y detección de anomalías. Divida el dataset "conn250k.csv" en set de entrenamiento y set de pruebas.

Taller de Ciberseguridad con Inteligencia Artificial.

Dr. Vitali Herrera Semenets – CENATAV, La Habana, Cuba.

Dr. Lazaro Bustio Martínez – IBERO, Ciudad de México, México.

MSc. Felipe Antonio Trujillo Fernández – IBERO, Ciudad de México, México.

- Entrene un modelo de clasificación para detectar anomalías. El algoritmo que entrenar puede ser de su preferencia.
- Evalúe el desempeño del modelo creado mediante las siguientes métricas:
 - i. Precision
 - ii. Recall
 - iii. F-Score
 - iv. AUC-ROC

Dado que el dataset "conn250k.csv" presenta un desbalanceo muy elevado, tenga en cuenta que la precisión puede ser engañosa. En estos casos se debe prestar especial atención a mejorar la puntuación de F-Score y AUC-ROC para obtener una evaluación más precisa del modelo.

e) Luego de realizar el ejercicio, ¿a qué conclusión se puede llegar?