Athena (Android Malware Detection)

Mateo Aguirre Duque

Departamento de ingeniería de sistemas

Universidad de Antioquia

Inteligencia Artificial para las Ciencias e Ingenierías

Raul Ramos P.

Marzo 05 de 2023

Descripción del problema

Hoy en día una gran cantidad de nuestra información personal, bancaria y/o profesional la tenemos en nuestros dispositivos móviles los cuales también usamos para estar en redes sociales, descargar diversos archivos e instalar aplicaciones. El problema surge cuando todos estos archivos y/o aplicaciones tienen una dudosa procedencia y ponen en riesgo la integridad de nuestro dispositivo.

La intención en este proyecto es crear una serie de modelos predictivos con la capacidad de detectar patrones en la red para identificar si en nuestro dispositivo existe alguna aplicación que pueda ser maliciosa o directamente contener un malware.

Dataset

El dataset que se usará para el proyecto es Android_Malware.csv que está disponible en la plataforma kaggle más concretamente en el siguiente link (https://www.kaggle.com/datasets/subhajournal/android-malware-detection).

Este dataset contiene registros de diversos comportamientos en la red de diferentes dispositivos y en su última columna llamada “Label” una clasificación del posible tipo de malware que está afectando el dispositivo, la idea del proyecto no es clasificar el posible malware, si no indicar si el dispositivo está infectado o no, por esto es necesario modificar la última columna para que ya no indique el tipo o si el comportamiento es normal sino más bien solo si esta infectado o no.

Dimensiones del Dataset

El dataset cuenta con 355630 registros o filas y con 85 columnas, en estas columnas se encuentran datos de red como direcciones de IP de origen y destino, protocolos de

comunicación, tamaño de los paquetes enviados, tamaño de las cabeceras que se envían, longitud total de todos los paquetes enviados, entre muchos otros valores de red. Otras columnas a destacar son, un identificador único del registro, una clasificación de tipo de malware o si no existe ninguno y un timestamp que nos indica el día y la hora exacta en que se produjo la transferencia de datos.

Métricas de desempeño

Para evaluar los modelos se planea usar dos funciones muy conocidas en los problemas de clasificación binaria, estas funciones serían:

1. Precision: es el porcentaje de resultados que son positivos dentro del dataset de prueba.

$$precision = \frac{True\ positives}{True\ positives + False\ Positives}$$

2. Recall: es el porcentaje de positivos reales que se identificaron correctamente.

$$recall = \frac{True\ positives}{True\ positives + False\ negatives}$$

La idea es que tanto el valor de *precision* como el valor de *recall* sean lo más alto posible para considerar que los modelos quedaron bien entrenados.

Desempeño deseado

Lo que se espera de los modelos resultantes de este proyecto es que sean capaz de encontrar y diagnosticar dispositivos comprometidos con gran facilidad y velocidad, de esta forma el usuario dueño del dispositivo móvil podrá tomar acción sobre lo que sucede, para esto el

modelo debería garantizar una baja tasa de falsos positivos y tener un rendimiento alto en los verdaderos positivos.

Bibliografía

1. Android Malware Detection, kaggle 2023
   https://www.kaggle.com/datasets/subhajournal/android-malware-detection
2. Evaluating the Performance of Machine Learning Models, Towards Data Science
   Abril 18 de 2020
   https://towardsdatascience.com/classifying-model-outcomes-true-false-positives-negatives-177c1e702810