Primera entrega de proyecto

Por:

María Paula Rojas Ortega

Juan Camilo Castañeda Ospina

Daniela Gómez Correa

Materia:

Introducción a la inteligencia artificial

Profesor:

Raul Ramos Pollan

Universidad de Antioquia

Facultad de Ingeniería

Medellín

2023

1. Planteamiento del problema

Se ha observado un aumento considerable en el uso de software maliciosos o malware debido al crecimiento de la oferta de servicios digitales. Una vez que un ordenador está infectado por malware, los delincuentes pueden perjudicar a consumidores y empresas de muchas formas, tales como corromper datos, robar o secuestrar información, y otros ciberdelitos mediados por este tipo de software. La industria del malware sigue siendo un mercado bien organizado y financiado dedicado a eludir las medidas de seguridad tradicionales. Estos software pueden afectar un gran número de sistemas en poco tiempo, por lo que su detección debe darse lo más rápido posible, sin embargo, usan diferentes métodos de ocultación y evasión haciendo más difícil su detección. Por lo anterior se hace necesario mejorar las técnicas de detección para que esta se dé más eficazmente (Aslan, Ö. A., & Samet, R., 2020; Aboaoja, F. A. et.al., 2022).

El objetivo es predecir la probabilidad de que una máquina Windows sea infectada por varias familias de malware, basándose en diferentes propiedades de dicha máquina.

2. Dataset

El dataset a utilizar proviene de una competencia de kaggle en la cual se proporcionan datos de máquinas con ciertas características que tuvieron un reporte de amenaza de Windows Defender y si fueron infectadas o no. El dataset está compuesto por dos robustos conjuntos de archivos .csv para probar (test.csv) y para calibrar el algoritmo (train.csv)

Cada fila representa una sola máquina y las columnas tienen información sobre las características de esta. El dataset original contiene 7 853 253 entradas para el test.csv y 8 921 483 entradas para el train.csv y un total de 82 columnas, más una en train.csv, "*HasDetection",* que es con la que se calibra el modelo, pues tiene información de si la máquina fue infectada o no.

Entre las características más relevantes de las máquinas se encuentran:

- MachineIdentifier – ID de la máquina.
- EngineVersion – Versión del motor. e.g. 1.1.12603.0
- AppVersion – Versión de la aplicación. e.g. 4.9.10586.0
- IsBeta – Si la versión es beta o no. e.g. false

- AVProductStatesIdentifier - ID para la configuración específica del software antivirus.
- CountryIdentifier - ID del país donde se encuentra la máquina.
- CityIdentifier - ID de la ciudad donde se encuentra la máquina.
- GeoNameIdentifier - ID de la región geográfica en la que se encuentra una máquina
- Platform - Nombre de sistema operativo.
- Processor - Arquitectura del sistema operativo.
- OsBuild - Compilación del sistema operativo.
- OsPlatformSubRelease - Devuelve la sub-release de la plataforma OS
- OsBuildLab - Laboratorio de compilación que generó el OS. Example: 9600.17630.amd64fre.winblue_r7.150109-2022
- SkuEdition – Edición de antivirus.
- SmartScreen - Indica si el SmartScreen está habilitado y de qué forma.
- UacLuaenable - Este atributo informa de si el tipo de usuario "administrador en modo de aprobación de administrador" está deshabilitado o habilitado.
- Census_MDC2FormFactor - Factor de forma. La lógica utilizada para definir el factor de forma se basa en estándares empresariales e industriales y se alinea con la forma en que la gente piensa en su dispositivo. (Ejemplos: smartphone, tableta pequeña, todo en uno, convertible...)
- Census_ProcessorCoreCount - Número de núcleos lógicos del procesador.
- Census_PrimaryDiskTotalCapacity - Cantidad de espacio en disco en el disco primario de la máquina en MB
- Census_PrimaryDiskTypeName - Nombre del tipo de disco - HDD or SSD
- Census_SystemVolumeTotalCapacity - El tamaño de la partición en la que está instalado el volumen de Sistema en MB.
- Census_TotalPhysicalRAM - RAM física en MB.

- Census_InternalPrimaryDiagonalDisplaySizeInInches - Talla de la pantalla ppal.

- Census_PowerPlatformRoleName - Indica el perfil de gestión de energía preferido. Este valor ayuda a identificar el factor de forma básico del dispositivo.

- Census_OSVersion - Versión numérica del SO Example - 10.0.10130.0

- Census_OSArchitecture - Arquitectura del SO. Example - amd64

- Census_OSBranch - Rama del SO extraída del OsVersionFull. Example - OsBranch = fbl_partner_eeap where OsVersion = 6.4.9813.0.amd64fre.fbl_partner_eeap.140810-0005

- Census_OSBuildNumber - Número de compilación del sistema operativo extraído de OsVersionFull. Example - OsBuildNumber = 10512 or 10240

- Census_OSEdition - Nombre de la edición OS (currently Windows only).

- Census_OSInstallTypeName -Descripción amigable de qué instalación se utilizó en la máquina. Ejemplo: Refresh.

- Census_OSWUAutoUpdateOptionsName - Nombre descriptivo de la configuración de actualización automática de WindowsUpdate en la máquina.

- Census_GenuineStateName - Nombre del estado de la licencia.

- Census_ActivationChannel - Clave de licencia por menor o clave de licencia por volumen para una máquina.

- Census_IsTouchEnabled - Indica si es un dispositivo táctil.

- Wdft_IsGamer - Indica si el dispositivo es usado por gamers o no.

Del total de columnas de los archivos originales se hizo una depuración a priori, con la cual se obtuvieron 35 columnas útiles. Los motivos para descartar columnas fueron: 1) no había información sobre el contenido de la columna (no metadata), 2) la información contenida era un único valor y 3) columnas repetidas o con información similar a otras.

En cuanto a los datos faltantes, encontramos que en general la calidad de los datos de las 35 columnas seleccionadas es muy buena, es decir, las columnas que tienen datos faltantes representa menos del 4% de las entradas, exceptuando la columna “SmartScreen” que tiene el 45%. Por lo

que para cumplir con el ejercicio académico, generaremos el 10 % de datos faltantes en otras 2 columnas de manera artificial.

3. Métricas

La métrica de evaluación principal para el modelo será el área bajo la curva ROC (Receiver operating characteristic) entre la probabilidad predicha y la observada. Esta curva se genera graficando la tasa de verdaderos positivos (TPR) frente a la tasa de falsos positivos (FPR) con distintos umbrales. La tasa de verdaderos positivos también se conoce como sensibilidad, memoria o probabilidad de detección, y la tasa de falsos positivos también se conoce como probabilidad de falsa alarma (“Receiver operating characteristic”, 2023).

Los cuales se calculan mediante la siguiente expresión:

sensitivity, recall, hit rate, or true positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = 1 - \mathrm{FNR}$$

fall-out or false positive rate (FPR)

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} = 1 - \mathrm{TNR}$$

(“Receiver operating characteristic”,2023).

De igual forma, se espera que el modelo aumente la inversión en antivirus y herramientas de detección de malware, y de soluciones personalizadas, tales como aplicaciones con un diseño más enfocado a la prevención de brechas que podrían permitir la entrada de malware al sistema.

4. Desempeño

Se espera que el modelo tenga un porcentaje de acierto de al menos un 95%, ya que el error en la detección de malwares en un mayor grado puede representar millonarias pérdidas en sistemas y datos almacenados. Con esta información se desearía obtener mejores análisis de la detección del malware y sus posibles relaciones con el hardware y software particular de cada ordenador, lo que permitiría crear soluciones personalizadas para los procesadores y equipos más vulnerables.

5. Bibliografía

- Aslan, Ö. A., & Samet, R. (2020). A comprehensive review on malware detection approaches. IEEE Access, 8, 6249-6271.

- Aboaoja, F. A., Zainal, A., Ghaleb, F. A., Al-rimy, B. A. S., Eisa, T. A. E., & Elnour, A. A. H. (2022). Malware detection issues, challenges, and future directions: A survey. Applied Sciences, 12(17), 8482.
- Microsoft Malware Prediction | Kaggle. (2023). Retrieved 2 March 2023, from https://www.kaggle.com/competitions/microsoft-malware-prediction/overview
- Receiver operating characteristic. (2023). Retrieved March 2, 2023, from https://en.wikipedia.org/wiki/Receiver_operating_characteristic