Segunda entrega de proyecto

Por:

María Paula Rojas Ortega

Juan Camilo Castañeda Ospina

Materia:

Introducción a la inteligencia artificial

Profesor:

Raúl Ramos Pollan

Universidad de Antioquia

Facultad de Ingeniería

Medellín

2023

Índice

1. Introducción

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 softwares 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. 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 (“Receiver operating characteristic”, 2023). 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.

2. Exploración descriptiva del 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 (por sus

siglas en inglés, comma-separated values) para calibrar el algoritmo (train.csv) y probar el modelo (test.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 test.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 sí la máquina fue infectada o no. Sin embargo, para simplificar el análisis se decidió reducir el número de filas a 800.000, provenientes del archivo train.csv, ya que el test.csv no contiene el resultado de la variable objetivo, por lo que no era posible obtener la métrica de evaluación.

3. Descripción del progreso alcanzado

3.1. Exploración de datos

Consta de los siguientes apartados:

- Descripción del dataset: Este incluye el tamaño del dataframe, los tipos de datos que contiene cada columna, se encontraron 45 columnas tipo flotante, 8 de tipo entero y 30 de tipo objeto; y una revisión de las columnas numéricas, en las que se encontraron 19 variables cuyo valor oscila entre 0 y 1.
- Descripción de datos faltantes: Debido a la gran cantidad de columnas se realizó a través de una gráfica, para por medio de esta establecer las columnas con mayor cantidad de datos faltantes.
- Distribución de la variable objetivo ("HasDetections")
- Distribución de las otras variables con respecto a la variable objetivo.
- Correlación entre variables.

3.2. Preprocesado

Del total de columnas se hizo una depuración bajo las siguientes condiciones:

1. No hay información sobre el contenido de la columna (Marcadas en el apartado "Data" de la competencia de Kaggle como NA, Columna no disponible o autodocumentada),
2. La cantidad de datos faltantes en la columna es mayor al 80%, y no provee información relevante con respecto a la variable objetivo.
3. Columnas que contienen un único valor para todas las filas
4. Columnas repetidas o con información similar a otras.

En cuanto a los datos faltantes, se encontró que en general la calidad de los datos de las 60 columnas seleccionadas, después del proceso de

depuración, es muy buena, es decir, los datos faltantes representan el 1,48 % de las entradas. Por lo que, para cumplir con el ejercicio académico, se generó el 3,6% de datos faltantes aleatoriamente.

Posteriormente, se realizan pruebas con el fin de determinar el método de llenado de datos faltantes más adecuado para los tipos de datos en el dataset, por lo anterior se realizan pruebas independientemente para las variables categóricas y las numéricas, y así mismo se establecen como se llevará a cabo la evaluación de los métodos. Esta última está en proceso de construcción, por lo que no se adjunta como parte del preprocesado, sino como archivos adjuntos.

4. 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