Segunda entrega de proyecto

POR:
Jhon Vásquez
Juan Felipe Santa

MATERIA:
Introducción a la Inteligencia Artificial

PROFESOR:
Raul Ramos Pollán

UNIVERSIDAD®
DE ANTIOQUIA

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

MEDELLÍN

2022

1. Planteamiento del Problema

El objetivo del proyecto es predecir la probabilidad de que una máquina con el sistema operativo Windows se infecte con distintas familias de malware, en función de diferentes propiedades que tenga dicha máquina. La detección temprana de malware es fundamental para garantizar la seguridad de los usuarios y sus datos, por lo que este problema tiene una gran relevancia práctica en el campo de la ciberseguridad. El reto en este caso es encontrar un modelo de machine learning capaz de aprender patrones sutiles en los datos que permitan diferenciar entre sistemas limpios y sistemas infectados con malware, y que pueda generalizar bien a nuevos datos no vistos.

2. Preprocesamiento de los datos

Se realizó una submuestra del conjunto de datos original, lo que resultó en un conjunto de datos con un total de 20,000 filas. Posteriormente, se verificó la presencia de valores faltantes en todas las columnas. Se encontró que la columna *"PuaMode"* tenía la mayor cantidad de valores faltantes, mientras que la columna *"Census_OEMNameIdentifier"* presentaba la menor cantidad de valores faltantes. Además, se verificó los tipos de datos de cada columna, obteniendo presencia de datos de tipo int64, float64 y object (ver figura 1). El porcentaje de valores faltantes se observa en la figura 2, donde el 6% de las columnas del dataset tiene un porcentaje de valores faltantes mayor al 70%; dichas columnas son eliminadas del dataset.

MachineIdentifier	object
ProductName	object
EngineVersion	object
AppVersion	object
AvSigVersion	object
IsBeta	int64
RtpStateBitfield	float64
IsSxsPassiveMode	int64
DefaultBrowsersIdentifier	float64
AVProductStatesIdentifier	float64
AVProductsInstalled	float64
AVProductsEnabled	float64
HasTpm	int64
CountryIdentifier	int64

Figura 1. Algunos columnas del dataset y sus tipos de datos

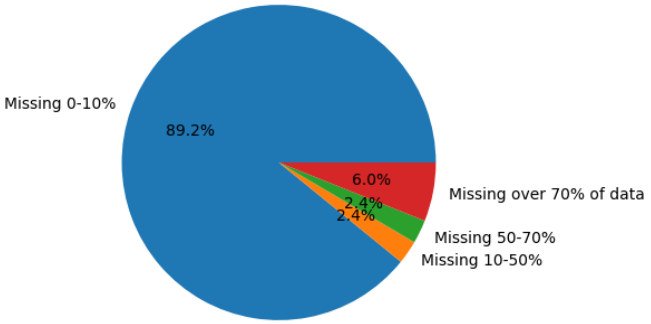

Figura 2. Porcentaje de columnas con un valor de datos faltantes

Se realizó la matriz de correlación, donde se observó que la mayoría de variables tienen correlación positiva débil (ver figura 3). Además, en la correlación respecto a la variable objetivo, se observa que hay variables que dan resultado NaN (Ver figura 4).

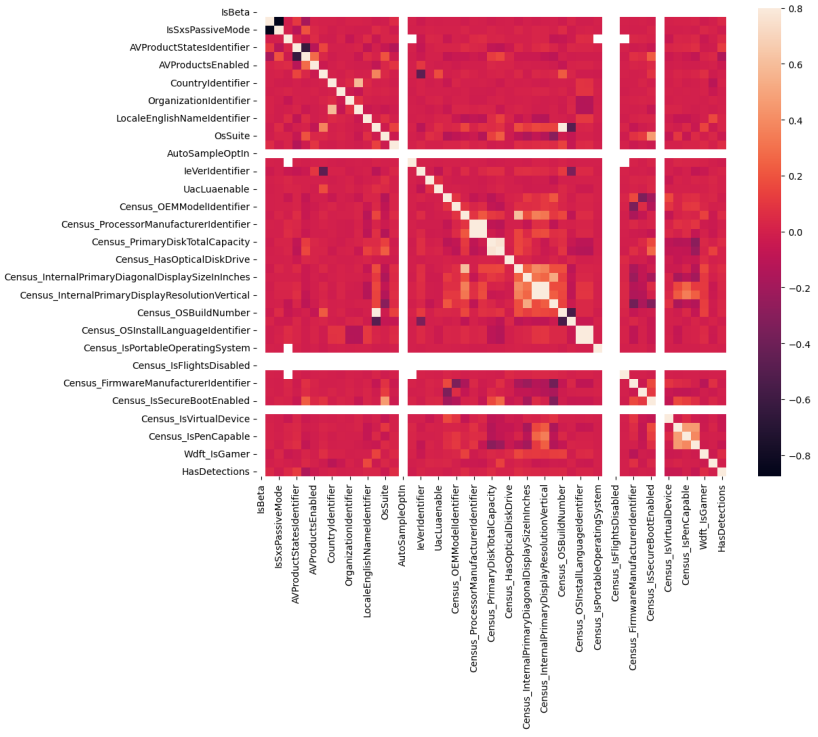

Figura 3. Matriz de Correlación

AVProductsInstalled	-0.145669
IsBeta	NaN
AutoSampleOptIn	NaN
Census_IsFlightingInternal	NaN
Census_IsFlightsDisabled	NaN
Census_IsWIMBootEnabled	NaN

Figura 4. Valores NaN en correlación

Para las variables categóricas se llenan los datos faltantes con la moda y se aplica la técnica de One-Hot Encoding, lo cuál hace que la dimensión del dataset aumente considerablemente, teniendo ahora 2953 columnas.

Por otra parte, se analiza la distribución de las clases, encontrando que el dataset se encuentra muy balanceado, teniendo 10019 de máquinas que han sido infectadas y 9980 que no (ver figura 5)

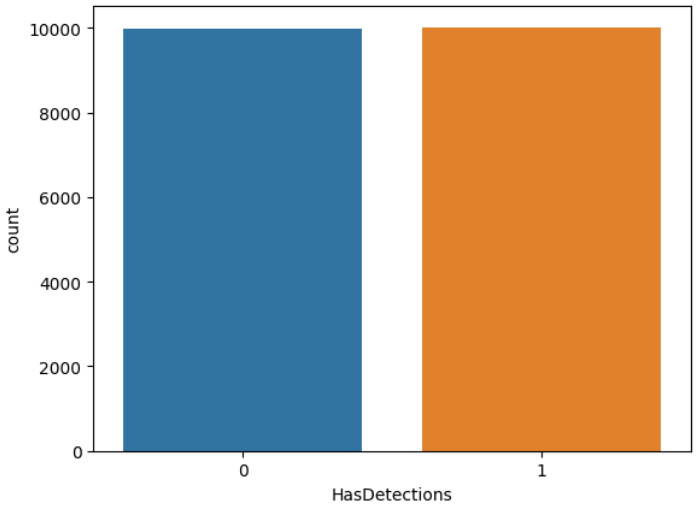

Figura 5. Valores NaN en correlación

3. Referencias

- Microsoft Malware Prediction | Kaggle. (s. f.-b). https://www.kaggle.com/competitions/microsoft-malware-prediction/