Primera 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. Dataset

El dataset lo subió Microsoft al sitio de Kaggle para una competición llamada "Microsoft Malware Prediction". El dataset fue recopilado usando la herramienta Windows Defender, la cuál está activa en la mayoría de computadores con el sistema operativo Windows. Cada fila de datos corresponde a una máquina, identificada por la variable *MachineIdentifier*. La variable que dice si un computador ha sido infectado o no por un malware se llama *HasDetections*. Como el dataset cuenta con más de ocho millones de filas, se hace una submuestra de los datos, tomando solo cien mil filas.

Las columnas del dataset están descritas de está manera:

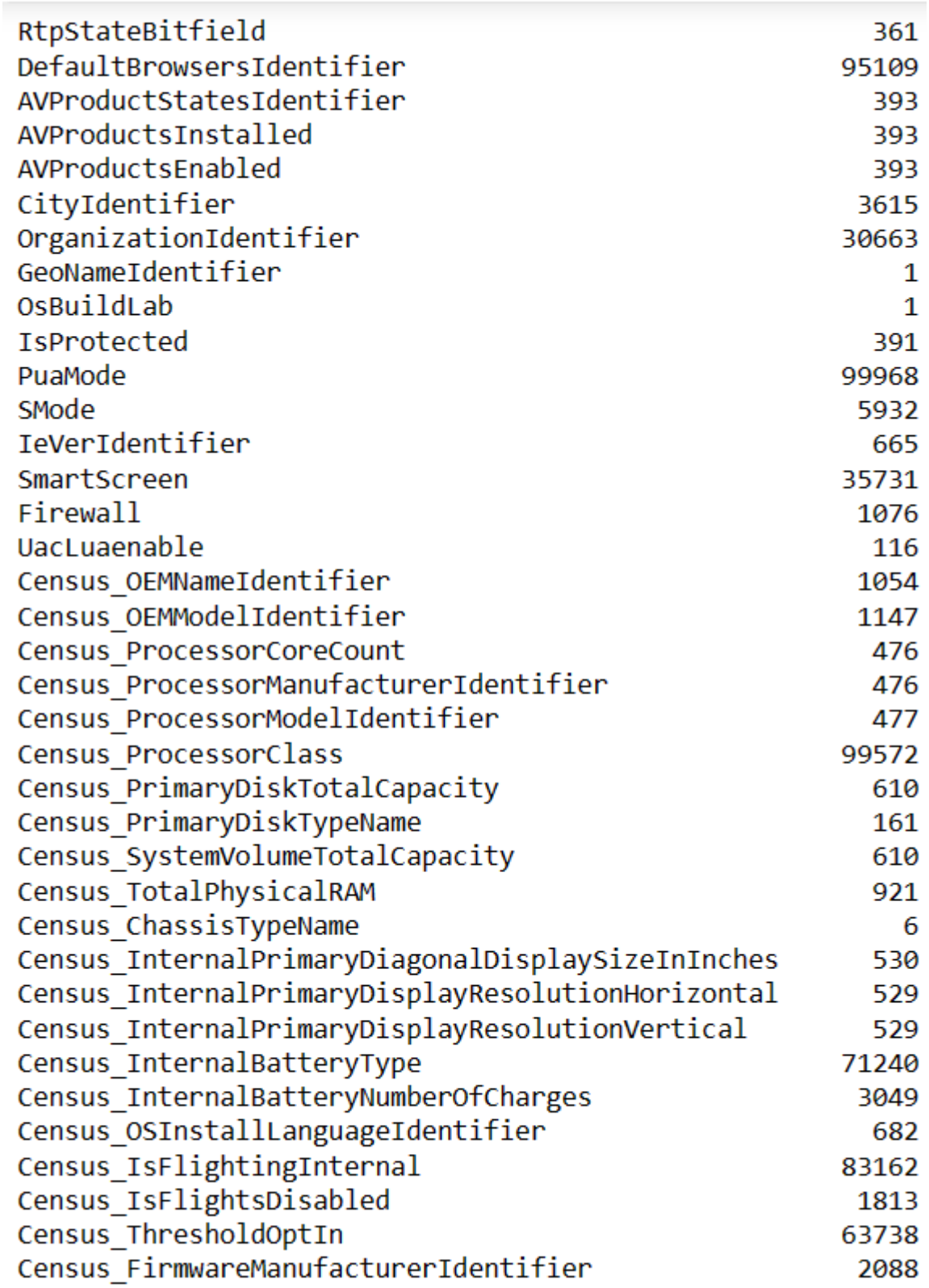
#	Column	Non-Null Count	Dtype
0	MachineIdentifier	99999 non-null	object
1	ProductName	99999 non-null	object
2	EngineVersion	99999 non-null	object
3	AppVersion	99999 non-null	object
4	AvSigVersion	99999 non-null	object
5	IsBeta	99999 non-null	int64
6	RtpStateBitfield	99638 non-null	float64
7	IsSxsPassiveMode	99999 non-null	int64
8	DefaultBrowsersIdentifier	4890 non-null	float64
9	AVProductStatesIdentifier	99606 non-null	float64
10	AVProductsInstalled	99606 non-null	float64
11	AVProductsEnabled	99606 non-null	float64
12	HasTpm	99999 non-null	int64
13	CountryIdentifier	99999 non-null	int64
14	CityIdentifier	96384 non-null	float64
15	OrganizationIdentifier	69336 non-null	float64
16	GeoNameIdentifier	99998 non-null	float64
17	LocaleEnglishNameIdentifier	99999 non-null	int64
18	Platform	99999 non-null	object
19	Processor	99999 non-null	object
20	OsVer	99999 non-null	object
21	OsBuild	99999 non-null	int64
22	OsSuite	99999 non-null	int64
23	OsPlatformSubRelease	99999 non-null	object
24	OsBuildLab	99998 non-null	object
25	SkuEdition	99999 non-null	object
26	IsProtected	99608 non-null	float64
27	AutoSampleOptIn	99999 non-null	int64
28	PuaMode	31 non-null	object
29	SMode	94067 non-null	float64
30	IeVerIdentifier	99334 non-null	float64
31	SmartScreen	64268 non-null	object
32	Firewall	98923 non-null	float64
33	UacLuaenable	99883 non-null	float64
34	Census_MDC2FormFactor	99999 non-null	object

```
35  Census_DeviceFamily                                99999 non-null  object
36  Census_OEMNameIdentifier                           98945 non-null  float64
37  Census_OEMModelIdentifier                          98852 non-null  float64
38  Census_ProcessorCoreCount                          99523 non-null  float64
39  Census_ProcessorManufacturerIdentifier             99523 non-null  float64
40  Census_ProcessorModelIdentifier                    99522 non-null  float64
41  Census_ProcessorClass                              427 non-null    object
42  Census_PrimaryDiskTotalCapacity                    99389 non-null  float64
43  Census_PrimaryDiskTypeName                         99838 non-null  object
44  Census_SystemVolumeTotalCapacity                   99389 non-null  float64
45  Census_HasOpticalDiskDrive                         99999 non-null  int64
46  Census_TotalPhysicalRAM                            99078 non-null  float64
47  Census_ChassisTypeName                             99993 non-null  object
48  Census_InternalPrimaryDiagonalDisplaySizeInInches  99469 non-null  float64
49  Census_InternalPrimaryDisplayResolutionHorizontal  99470 non-null  float64
50  Census_InternalPrimaryDisplayResolutionVertical    99470 non-null  float64
51  Census_PowerPlatformRoleName                       99999 non-null  object
52  Census_InternalBatteryType                         28759 non-null  object
53  Census_InternalBatteryNumberOfCharges              96950 non-null  float64
54  Census_OSVersion                                   99999 non-null  object
55  Census_OSArchitecture                              99999 non-null  object
56  Census_OSBranch                                    99999 non-null  object
57  Census_OSBuildNumber                               99999 non-null  int64
58  Census_OSBuildRevision                             99999 non-null  int64
59  Census_OSEdition                                   99999 non-null  object
60  Census_OSSkuName                                   99999 non-null  object
61  Census_OSInstallTypeName                           99999 non-null  object
62  Census_OSInstallLanguageIdentifier                 99317 non-null  float64
63  Census_OSUILocaleIdentifier                        99999 non-null  int64
64  Census_OSWUAutoUpdateOptionsName                   99999 non-null  object
65  Census_IsPortableOperatingSystem                   99999 non-null  int64
66  Census_GenuineStateName                            99999 non-null  object
67  Census_ActivationChannel                           99999 non-null  object
68  Census_IsFlightingInternal                         16837 non-null  float64
69  Census_IsFlightsDisabled                           98186 non-null  float64
70  Census_FlightRing                                  99999 non-null  object
71  Census_ThresholdOptIn                              36261 non-null  float64
```

```
72  Census_FirmwareManufacturerIdentifier              97911 non-null  float64
73  Census_FirmwareVersionIdentifier                   98166 non-null  float64
74  Census_IsSecureBootEnabled                         99999 non-null  int64
75  Census_IsWIMBootEnabled                            36340 non-null  float64
76  Census_IsVirtualDevice                             99814 non-null  float64
77  Census_IsTouchEnabled                              99999 non-null  int64
78  Census_IsPenCapable                                99999 non-null  int64
79  Census_IsAlwaysOnAlwaysConnectedCapable            99156 non-null  float64
80  Wdft_IsGamer                                       96584 non-null  float64
81  Wdft_RegionIdentifier                              96584 non-null  float64
82  HasDetections                                      99999 non-null  int64
dtypes: float64(36), int64(17), object(30)
memory usage: 63.3+ MB
None
```

Imagen 1. Columnas del dataset

Los datos faltantes se pueden observar en la siguiente imagen:



RtpStateBitfield	361
DefaultBrowsersIdentifier	95109
AVProductStatesIdentifier	393
AVProductsInstalled	393
AVProductsEnabled	393
CityIdentifier	3615
OrganizationIdentifier	30663
GeoNameIdentifier	1
OsBuildLab	1
IsProtected	391
PuaMode	99968
SMode	5932
IeVerIdentifier	665
SmartScreen	35731
Firewall	1076
UacLuaenable	116
Census_OEMNameIdentifier	1054
Census_OEMModelIdentifier	1147
Census_ProcessorCoreCount	476
Census_ProcessorManufacturerIdentifier	476
Census_ProcessorModelIdentifier	477
Census_ProcessorClass	99572
Census_PrimaryDiskTotalCapacity	610
Census_PrimaryDiskTypeName	161
Census_SystemVolumeTotalCapacity	610
Census_TotalPhysicalRAM	921
Census_ChassisTypeName	6
Census_InternalPrimaryDiagonalDisplaySizeInInches	530
Census_InternalPrimaryDisplayResolutionHorizontal	529
Census_InternalPrimaryDisplayResolutionVertical	529
Census_InternalBatteryType	71240
Census_InternalBatteryNumberOfCharges	3049
Census_OSInstallLanguageIdentifier	682
Census_IsFlightingInternal	83162
Census_IsFlightsDisabled	1813
Census_ThresholdOptIn	63738
Census_FirmwareManufacturerIdentifier	2088

```
Census_FirmwareVersionIdentifier                1833
Census_IsWIMBootEnabled                        63659
Census_IsVirtualDevice                           185
Census_IsAlwaysOnAlwaysConnectedCapable          843
Wdft_IsGamer                                    3415
Wdft_RegionIdentifier                           3415
dtype: int64
```

Imagen 2. Datos Faltantes

3. Métricas

Para evaluar el sistema se utilizarán accuracy, f1 score y recall. El recall porque mide la proporción de predicciones positivas que son correctas en relación con el total de predicciones positivas realizadas por el modelo; f1-score combina la precisión y el recall en una sola métrica que proporciona una visión general del rendimiento del modelo y el accuracy es la métrica comúnmente utilizada para evaluar el rendimiento de los modelos de clasificación.

Se debe tener una predicción con un buen desempeño (más del 85% de precisión), ya que para la empresa es fundamental tener un modelo con un alto acierto en la clasificación, para así ofrecer mejor seguridad a los usuarios y por tanto, crear en éstos una mayor confianza respecto a sus productos.

4. Desempeño deseable en producción

El desempeño deseable en producción de un modelo hecho con el conjunto de datos de Microsoft Malware Prediction dependerá de varios factores. Sin embargo, a modo de referencia, se puede considerar el siguiente criterio general para el desempeño deseable:

- Un accuracy superior al 95%: esto indica que el modelo es capaz de clasificar correctamente la gran mayoría de los dispositivos de prueba y, por lo tanto, es eficaz para identificar la presencia o ausencia de malware.

- Un recall superior al 90%: esto indica que el modelo es capaz de detectar la gran mayoría de los dispositivos infectados y, por lo tanto, es efectivo para prevenir la infección de malware y reducir el riesgo para los usuarios.

- Una tasa de falsos positivos inferior al 5%: esto indica que el modelo produce una cantidad mínima de alertas de seguridad falsas, lo que minimiza la interrupción innecesaria del trabajo y la productividad del usuario.

- Un tiempo de respuesta promedio a la detección inferior a 1 hora: esto indica que el modelo es capaz de detectar y prevenir rápidamente la infección de malware, minimizando el impacto en el dispositivo y en la red.

5. Referencias

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