

## Análisis exploratorio y preprocesamiento de los datos

Comenzaremos, por la exploración de los valores de las columnas:

La información mostrada aquí se basa en observaciones hechas al dataframe train.csv ejecutando el código **exploracionColumnas.py** y, parcialmente, en la información recolectada a través de las siguientes páginas web:

- <https://www.dtonomy.com/will-your-machine-be-hit-by-a-malware-soon/#ds>
- <https://www.kaggle.com/competitions/microsoft-malware-prediction/data>

### Columnas del dataframe dado por train.csv

1. **'MachineIdentifier'**: Identificador único de la máquina

```
*** Análisis para columna: MachineIdentifier ***
MachineIdentifier
fffff75ba4f33d938ccfdb148b8ea16      1.0
0000028988387b115f69f31a3bf04f09      1.0
000007535c3f730efa9ea0b7ef1bd645      1.0
000007905a28d863f6d0d597892cd692      1.0
00000b11598a75ea8ba1beea8459149f      1.0
...
000027c68b89acb49d4017763b043449      1.0
0000258d2b847c7549150cfec6464473      1.0
000024872c81cf03fa862aa8f99e0984      1.0
00001f26e9e5775277d6231fc6ac9e70      1.0
00001b924fcc6922321cfadbaf8a91a      1.0
Length: 8921483, dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> object
```

2. **ProductName**: Nombre del producto antivirus de Microsoft instalado (ej: `mse`, `wdav`, etc.). Usualmente: `Windows Defender Antivirus`.

```
*** Análisis para columna: ProductName ***
ProductName
win8defender      8826520.0
mse                94873.0
mseprerelease     53.0
scep               22.0
windowsintune     8.0
fep                7.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> object
```

3. **'EngineVersion'**: Versión del motor antivirus usado por Defender para detectar malware. Puede ayudar a identificar si se usaba una versión antigua o actualizada.

```
*** Análisis para columna: EngineVersion ***
EngineVersion
1.1.15200.1    3845067.0
1.1.15100.1    3675915.0
1.1.15000.2    265218.0
1.1.14901.4    212408.0
1.1.14600.4    160585.0
...
1.1.13301.0    2.0
1.1.11202.0    1.0
1.1.11104.0    1.0
1.1.10701.0    1.0
1.1.12802.0    1.0
Length: 70, dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> object
```

4. **'AppVersion'**: Versión de la interfaz gráfica del antivirus (GUI). Indica la versión del cliente que muestra el panel de Defender.

```

*** Análisis para columna: AppVersion ***
AppVersion
4.18.1807.18075      5139224.0
4.18.1806.18062      850929.0
4.12.16299.15         359871.0
4.10.209.0             272455.0
4.13.17134.1          257270.0
...
4.17.17686.1004       1.0
4.5.212.0               1.0
4.7.209.0               1.0
4.8.203.0               1.0
4.8.10240.17943       1.0
Length: 110, dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> object

```

5. `AvSigVersion`: Versión de la base de firmas (antivirus signature version). Muy útil, ya que si está desactualizada puede aumentar el riesgo de detección tardía.

```

*** Análisis para columna: AvSigVersion ***
AvSigVersion
1.273.1420.0      102317.0
1.263.48.0          98024.0
1.275.1140.0      97232.0
1.275.727.0          92448.0
1.273.371.0          86967.0
...
1.203.1253.0       1.0
1.203.125.0          1.0
1.203.961.0          1.0
1.203.933.0          1.0
1.197.1237.0       1.0
Length: 8531, dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> object

```

6. `IsBeta`: Si el antivirus es una versión beta (1 = sí, 0 = no). Puede influir si el usuario estaba probando versiones experimentales.

```
*** Análisis para columna: IsBeta ***
IsBeta
0    8921416.0
1      67.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> int64
```

7. `RtpStateBitfield`: Estado del \*Real-Time Protection\* como un bitfield (campo de bits). Algunos valores indican si la protección en tiempo real estaba encendida, desactivada, etc.

```
*** Análisis para columna: RtpStateBitfield ***
RtpStateBitfield
7.0    8651487.0
0.0    190701.0
NaN    32318.0
8.0    21974.0
5.0    20328.0
3.0    3029.0
1.0    1625.0
35.0    21.0
dtype: float64
La cantidad de valores nulos es 32318
El tipo de datos es -> float64
```

8. `IsSxsPassiveMode`: Si Defender estaba en modo pasivo porque había otro antivirus activo. 1 = sí, 0 = no. Si está activado, Defender no detecta amenazas directamente.

```
*** Análisis para columna: IsSxsPassiveMode ***
IsSxsPassiveMode
0    8766840.0
1    154643.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> int64
```

9. `DefaultBrowsersIdentifier`: Identificador hash del navegador por defecto del sistema. Puede correlacionar con comportamiento del usuario o exposición a riesgos (por ejemplo, usar IE vs. Chrome).

```
*** Análisis para columna: DefaultBrowsersIdentifier ***
DefaultBrowsersIdentifier
NaN      8488045.0
239.0    46056.0
3195.0   42692.0
1632.0   28751.0
3176.0   24220.0
...
33.0     1.0
29.0     1.0
27.0     1.0
25.0     1.0
6.0      1.0
Length: 2018, dtype: float64
La cantidad de valores nulos es 8488045
El tipo de datos es -> float64
```

10. `AVProductStatesIdentifier`: Código hash que representa el estado de todos los productos antivirus instalados (no solo Defender).

```
*** Análisis para columna: AVProductStatesIdentifier ***
AVProductStatesIdentifier
53447.0  5824565.0
7945.0   475897.0
47238.0  327656.0
62773.0  266764.0
46413.0  112878.0
...
73.0     1.0
72.0     1.0
71.0     1.0
68.0     1.0
66.0     1.0
Length: 28971, dtype: float64
La cantidad de valores nulos es 36221
El tipo de datos es -> float64
```

11. `AVProductsInstalled`: Número de productos antivirus instalados (por ejemplo, 1 si solo Defender, 2 si también McAfee, etc.).

```
*** Análisis para columna: AVProductsInstalled ***
AVProductsInstalled
1.0      6208893.0
2.0      2459008.0
3.0      208103.0
NaN       36221.0
4.0       8757.0
5.0        471.0
6.0         28.0
0.0         1.0
7.0         1.0
dtype: float64
La cantidad de valores nulos es 36221
El tipo de datos es -> float64
```

12. `AVProductsEnabled`: Cuántos antivirus están habilitados. Puede diferir de los instalados si alguno está desactivado.

```
*** Análisis para columna: AVProductsEnabled ***
AVProductsEnabled
1.0      8654101.0
2.0      198652.0
NaN       36221.0
0.0       25958.0
3.0       6075.0
4.0        453.0
5.0        23.0
dtype: float64
La cantidad de valores nulos es 36221
El tipo de datos es -> float64
```

13. `HasTpm`: Si el sistema tiene chip TPM (Trusted Platform Module). 1 = sí, 0 = no. TPM permite cifrado seguro (ej: BitLocker).

```
*** Análisis para columna: HasTpm ***
HasTpm
1    8814167.0
0    107316.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> int64
```

14. `CountryIdentifier`: Identificador numérico del país del usuario (no el código ISO estándar). Se necesita análisis cruzado para mapearlo.

```
*** Análisis para columna: CountryIdentifier ***
CountryIdentifier
43      397172.0
29      347991.0
141     333411.0
93      283625.0
171     280572.0
...
165      213.0
37       212.0
193     207.0
161     206.0
217     120.0
Length: 222, dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> int64
```

15. `CityIdentifier`: Id numérico de la ciudad (obscuro). Sirve más para clustering o distribución geográfica general.

```
*** Análisis para columna: CityIdentifier ***
CityIdentifier
NaN      325409.0
130775.0    94812.0
16668.0     84780.0
82373.0     83312.0
10222.0     71814.0
...
36.0        1.0
167918.0    1.0
167917.0    1.0
167915.0    1.0
167904.0    1.0
Length: 107367, dtype: float64
La cantidad de valores nulos es 325409
El tipo de datos es -> float64
```

16. `GeoNameIdentifier`: Identificador general de localización geográfica. Combina región, idioma y país.

```
*** Análisis para columna: GeoNameIdentifier ***
GeoNameIdentifier
277.0    1531929.0
211.0    423166.0
53.0     408807.0
89.0     360798.0
240.0    346568.0
...
14.0      1.0
279.0    1.0
278.0    1.0
55.0      1.0
197.0    1.0
Length: 293, dtype: float64
La cantidad de valores nulos es 213
El tipo de datos es -> float64
```

17. **OrganizationIdentifier**: ID de la organización a la que pertenece la máquina; este ID se asigna tanto a empresas específicas como sectores generales.

\*\*\* Análisis para columna: OrganizationIdentifier \*\*\*

OrganizationIdentifier

27.0	4196457.0
NaN	2751518.0
18.0	1764175.0
48.0	63845.0
50.0	45502.0
11.0	19436.0
37.0	19398.0
49.0	13627.0
46.0	10974.0
14.0	4713.0
32.0	4045.0
36.0	3909.0
52.0	3043.0
33.0	2896.0
2.0	2595.0
5.0	1990.0
40.0	1648.0
28.0	1591.0
4.0	1385.0
10.0	1083.0
51.0	917.0
20.0	915.0
1.0	893.0
8.0	723.0
22.0	418.0
39.0	413.0
6.0	412.0
31.0	398.0
21.0	397.0
47.0	385.0
3.0	331.0
16.0	242.0
19.0	172.0
26.0	160.0
44.0	150.0
29.0	135.0
42.0	132.0

```

    7.0          98.0
    41.0         77.0
    45.0         73.0
    30.0         64.0
    43.0         60.0
    35.0         32.0
    23.0         20.0
    15.0         13.0
    25.0         12.0
    12.0          7.0
    34.0          2.0
    17.0          1.0
    38.0          1.0
dtype: float64
La cantidad de valores nulos es 2751518
El tipo de datos es -> float64

```

18. `LocaleEnglishNameIdentifier`: Idioma del sistema operativo en inglés (ej: `en-US`, `fr-FR`).

```

*** Análisis para columna: LocaleEnglishNameIdentifier ***
LocaleEnglishNameIdentifier
75      2094585.0
182     450088.0
74      411056.0
42      409616.0
88      375223.0
...
39       2.0
258     2.0
191     1.0
203     1.0
259     1.0
Length: 276, dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> int64

```

19. `Platform`: Plataforma del sistema operativo (ej: `windows10`, `windows8`).

```
*** Análisis para columna: Platform ***
Platform
windows10      8618715.0
windows8       194508.0
windows7        93889.0
windows2016     14371.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> object
```

20. `Processor`: Tipo de procesador: `x64`, `x86`, `arm64`.

```
*** Análisis para columna: Processor ***
Processor
x64      8105435.0
x86      815702.0
arm64     346.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> object
```

21. `OsVer`: Versión del sistema operativo, ej: `10.0.0.0`. Muy correlacionada con el tiempo de instalación del sistema.

*** Análisis para columna: OsVer ***	
<b>OsVer</b>	
10.0.0.0	8632545.0
6.3.0.0	194447.0
6.1.1.0	93268.0
6.1.0.0	582.0
10.0.3.0	225.0
10.0.1.0	141.0
6.1.3.0	30.0
10.0.2.0	30.0
6.3.3.0	24.0
10.0.32.72	23.0
6.3.1.0	22.0
10.0.5.0	18.0
10.0.7.0	15.0
10.0.5.18	12.0
10.0.4.0	11.0
10.0.0.112	10.0
10.0.16.0	9.0
10.0.16.36	6.0
10.0.8.0	5.0
10.0.0.1	5.0
10.0.80.0	4.0
10.0.1.44	3.0
10.0.6.0	3.0
6.3.4.0	3.0
10.0.32.0	2.0
10.0.3.80	2.0
6.1.7.0	2.0
6.1.2.0	2.0
6.3.5.0	2.0
10.0.0.3	2.0
10.0.2.86	2.0
6.3.32.72	2.0
10.0.23.0	1.0
10.0.4.80	1.0
10.0.0.22	1.0
10.0.0.80	1.0
10.0.0.2	1.0
10.0.21.0	1.0
10.0.19.80	1.0
10.0.153.153	1.0

```
10.0.0.96           1.0
10.0.1.144          1.0
10.0.48.0           1.0
6.1.16.36           1.0
6.1.4.0             1.0
6.1.0.128           1.0
10.0.5.117          1.0
10.0.7.101          1.0
6.1.0.112           1.0
6.1.6.0             1.0
6.3.0.2             1.0
6.3.0.16            1.0
6.3.0.117           1.0
6.3.0.112           1.0
6.3.1.144           1.0
6.3.16.0            1.0
6.3.7.0             1.0
6.3.80.0            1.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> object
```

22. `OsBuild`: Número de compilación del sistema operativo Windows.

```
*** Análisis para columna: OsBuild ***
OsBuild
17134    3915521.0
16299    2503681.0
15063    780270.0
14393    730819.0
10586    411606.0
...
18226      1.0
18238      1.0
18236      1.0
18241      1.0
18244      1.0
Length: 76, dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> int64
```

23. `OsSuite`: Bitfield de las ediciones de Windows instaladas (por ejemplo, Home, Pro, Enterprise).

```
*** Análisis para columna: OsSuite ***
OsSuite
768      5560661.0
256      3346251.0
272      12092.0
400       793.0
16        731.0
305       662.0
784       198.0
274        39.0
144        34.0
49         17.0
307         2.0
18          1.0
528         1.0
402         1.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> int64
```

24. `OsPlatformSubRelease`: Versión específica del sistema (ej: `rs3`, `rs4`, `rs5`) relacionada con Redstone updates de Windows 10.

```
*** Análisis para columna: OsPlatformSubRelease ***
OsPlatformSubRelease
rs4        3915526.0
rs3        2503681.0
rs2        780270.0
rs1        730819.0
th2        411606.0
th1        270192.0
windows8.1  194508.0
windows7    93889.0
prers5     20992.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> object
```

25. `OsBuildLab`: Versión detallada de la build (ej: `17134.1.amd64fre.rs4\_release.180410-1804`)

```
*** Análisis para columna: OsBuildLab ***
17134.1.amd64fre.rs4_release.180410-1804          3658199.0
16299.431.amd64fre.rs3_release_svc_escrow.180502-1908    1252674.0
16299.15.amd64fre.rs3_release.170928-1534          961060.0
15063.0.amd64fre.rs2_release.170317-1834          718033.0
17134.1.x86fre.rs4_release.180410-1804          257074.0
...
10586.842.x86fre.th2_release_inmarket.170318-0600      1.0
10240.17943.amd64fre.th1.180724-1745      1.0
10240.17609.x86fre.th1.170904-1739      1.0
9600.18896.x86fre.winblue_ltsb_escrow.180108-1534    1.0
10240.16390.x86fre.th1_st1.150714-1601    1.0
Length: 664, dtype: float64
La cantidad de valores nulos es 21
El tipo de datos es -> object
```

26. `SkuEdition`: Tipo de licencia del sistema operativo (ej: `Home`, `Professional`, `Enterprise`, `Education`).

```
*** Análisis para columna: SkuEdition ***
SkuEdition
Home           5514341.0
Pro            3224164.0
Invalid        78054.0
Education       40694.0
Enterprise     34357.0
Enterprise LTSB 20702.0
Cloud           5589.0
Server          3582.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> object
```

27. `IsProtected`: Si el antivirus tiene protección en tiempo real activada. 1 = sí, 0 = no, NaN = desconocido.

```
*** Análisis para columna: IsProtected ***
IsProtected
1.0    8402282.0
0.0    483157.0
NaN    36044.0
dtype: float64
La cantidad de valores nulos es 36044
El tipo de datos es -> float64
```

28. **AutoSampleOptIn**: Si el usuario permite que Defender envíe automáticamente muestras sospechosas a Microsoft. 1 = sí, 0 = no. Afecta la capacidad de Microsoft para analizar nuevas amenazas.

```
*** Análisis para columna: AutoSampleOptIn ***
AutoSampleOptIn
0    8921225.0
1    258.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> int64
```

29. **PuaMode**: Indica si está activado el modo para detectar software potencialmente no deseado (PUA = Potentially Unwanted Application).

```
NaN    8919174.0
on    2307.0
audit    2.0
dtype: float64
La cantidad de valores nulos es 8919174
El tipo de datos es -> object
```

30. **SMode**: Este campo se establece como verdadero cuando se sabe que el dispositivo está en "Modo S", como en el modo S de Windows 10, donde solo se pueden instalar aplicaciones de Microsoft Store.

```
*** Análisis para columna: SMode ***
SMode
0.0    8379843.0
NaN    537759.0
1.0    3881.0
dtype: float64
La cantidad de valores nulos es 537759
El tipo de datos es -> float64
```

31. **IeVerIdentifier:** No disponible.

```
*** Análisis para columna: IeVerIdentifier ***
IeVerIdentifier
137.0    3885842.0
117.0    1767931.0
108.0    474390.0
111.0    467828.0
98.0    354411.0
...
32.0      1.0
29.0      1.0
381.0     1.0
404.0     1.0
402.0     1.0
Length: 304, dtype: float64
La cantidad de valores nulos es 58894
El tipo de datos es -> float64
```

32. **SmartScreen:** Este es el valor de la cadena "SmartScreen habilitado" del registro.

Se obtiene verificando en orden  
HKLM\SOFTWARE\Policies\Microsoft\Windows\System\SmartScreenEnabled y  
HKLM\SOFTWARE\Microsoft\Windows\CurrentVersion\Explorer\SmartScreenEnable  
d. Si el valor existe, pero está vacío, se envía el valor "ExistsNotSet" en la  
telemetría.

```
*** Análisis para columna: SmartScreen ***
RequireAdmin      4316183.0
NaN              3177011.0
ExistsNotSet     1046183.0
off              186553.0
Warn             135483.0
Prompt            34533.0
Block             22533.0
off                1350.0
On                 731.0
&#x02;              416.0
&#x01;              335.0
on                 147.0
requireadmin       10.0
OFF                4.0
0                  3.0
Promt              2.0
00000000          1.0
&#x03;              1.0
Enabled            1.0
prompt             1.0
requireAdmin        1.0
warn               1.0
dtype: float64
La cantidad de valores nulos es 3177011
El tipo de datos es -> object
```

33. **Firewall:** Este atributo es verdadero (1) para Windows 8.1 y versiones posteriores si el firewall de Windows está habilitado, según lo informado por el servicio. **UacLuaenable:** Este atributo indica si el tipo de usuario "administrador en modo de aprobación de administrador" está habilitado o deshabilitado en el Control de cuentas de usuario (UAC). El valor se obtiene leyendo la clave de registro HKLM\SOFTWARE\Microsoft\Windows\CurrentVersion\Policies\System\EnableLUA.

```
*** Análisis para columna: Firewall ***
Firewall
1.0    8641014.0
0.0    189119.0
NaN    91350.0
dtype: float64
La cantidad de valores nulos es 91350
El tipo de datos es -> float64
```

#### 34. UacLuaenable

```
*** Análisis para columna: UacLuaenable ***
UacLuaenable
1.0          8856517.0
0.0          53851.0
NaN          10838.0
48.0          206.0
2.0           30.0
49.0           17.0
6357062.0      13.0
3.0            6.0
5.0            2.0
255.0          1.0
7798884.0      1.0
16777216.0      1.0
dtype: float64
La cantidad de valores nulos es 10838
El tipo de datos es -> float64
```

35. **Census\_MDC2FormFactor:** Una agrupación basada en una combinación de características de hardware a nivel de censo del dispositivo. La lógica utilizada para definir el factor de forma se basa en estándares empresariales e industriales y se alinea con la percepción que los usuarios tienen de sus dispositivos. (Ejemplos: Smartphone, Tablet pequeña, Todo en uno, Convertible...)

```
*** Análisis para columna: Census_MDC2FormFactor ***
Census_MDC2FormFactor
Notebook      5723319.0
Desktop       1951086.0
Convertible    405378.0
Detachable     298233.0
AllInOne      292077.0
PCOther        139955.0
LargeTablet    67121.0
SmallTablet    31393.0
SmallServer     8630.0
MediumServer    3385.0
LargeServer     875.0
ServerOther     30.0
IoTOther        1.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> object
```

36. **Census\_DeviceFamily:** También conocido como DeviceClass. Indica el tipo de dispositivo al que está destinada una edición del sistema operativo. Valores de ejemplo: Windows.Desktop, Windows.Mobile e iOS.Phone

```
*** Análisis para columna: Census_DeviceFamily ***
Census_DeviceFamily
Windows.Desktop   8907053.0
Windows.Server     14410.0
Windows           20.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> object
```

37. **Census\_OEMNameIdentifier - NA**

```
*** Análisis para columna: Census_OEMNameIdentifier ***
Census_OEMNameIdentifier
2668.0    1287275.0
2102.0    1038567.0
1443.0    949531.0
2206.0    924349.0
585.0     895452.0
...
59.0      1.0
54.0      1.0
50.0      1.0
46.0      1.0
44.0      1.0
Length: 3833, dtype: float64
La cantidad de valores nulos es 95478
El tipo de datos es -> float64
```

### 38. **Census\_OEMModelIdentifier** - NA

```
*** Análisis para columna: Census_OEMModelIdentifier ***
Census_OEMModelIdentifier
313586.0   304782.0
242491.0   263382.0
317701.0   139035.0
317708.0   115257.0
NaN        102233.0
...
118.0      1.0
116.0      1.0
112.0      1.0
110.0      1.0
15.0       1.0
Length: 175366, dtype: float64
La cantidad de valores nulos es 102233
El tipo de datos es -> float64
```

### 39. **Census\_ProcessorCoreCount** - Número de núcleos lógicos del procesador

\*\*\* Análisis para columna: Census\_ProcessorCoreCount \*\*\*

Census\_ProcessorCoreCount

4.0	5430193.0
2.0	2311969.0
8.0	865004.0
12.0	92702.0
1.0	70390.0
6.0	69910.0
NaN	41306.0
16.0	18551.0
3.0	13580.0
32.0	2136.0
24.0	1847.0
20.0	1781.0
40.0	506.0
36.0	287.0
28.0	271.0
48.0	235.0
5.0	216.0
56.0	132.0
10.0	98.0
64.0	93.0
7.0	92.0
72.0	39.0
88.0	23.0
14.0	22.0
80.0	20.0
44.0	16.0
30.0	10.0
9.0	7.0
96.0	6.0
112.0	6.0
18.0	5.0
11.0	4.0
22.0	4.0
52.0	3.0
46.0	3.0
128.0	3.0
104.0	2.0
15.0	2.0
26.0	2.0

```
13.0          1.0
25.0          1.0
54.0          1.0
50.0          1.0
120.0         1.0
144.0         1.0
192.0         1.0
dtype: float64
La cantidad de valores nulos es 41306
El tipo de datos es -> float64
```

#### 40. Census\_ProcessorManufacturerIdentifier - NA

```
*** Análisis para columna: Census_ProcessorManufacturerIdentifier ***
Census_ProcessorManufacturerIdentifier
5.0      7839318.0
1.0      1040292.0
NaN      41313.0
10.0     339.0
3.0      218.0
4.0      1.0
9.0      1.0
7.0      1.0
dtype: float64
La cantidad de valores nulos es 41313
El tipo de datos es -> float64
```

#### 41. Census\_ProcessorModelIdentifier - NA

```

*** Análisis para columna: Census_ProcessorModelIdentifier ***
Census_ProcessorModelIdentifier
2697.0    289283.0
1998.0    267397.0
2660.0    191392.0
2373.0    175407.0
1992.0    171728.0
...
1045.0      1.0
4377.0      1.0
4379.0      1.0
4384.0      1.0
4386.0      1.0
Length: 3429, dtype: float64
La cantidad de valores nulos es 41343
El tipo de datos es -> float64

```

42. **Census\_ProcessorClass** - Clasificación de procesadores en alto/medio/bajo.  
 Inicialmente utilizado para SKU de nivel de precios. Ya no se mantiene ni actualiza

```

*** Análisis para columna: Census_ProcessorClass ***
c:\Users\cjcho\OneDrive\Documentos\AlgebraLinealAplicada\PCA.py:28: DtypeWarning: Columns
(41) have mixed types. Specify dtype option on import or set low_memory=False.
  for chunk in pd.read_csv("train.csv", usecols=[col], chunksize=500_000):
c:\Users\cjcho\OneDrive\Documentos\AlgebraLinealAplicada\PCA.py:28: DtypeWarning: Columns
(41) have mixed types. Specify dtype option on import or set low_memory=False.
  for chunk in pd.read_csv("train.csv", usecols=[col], chunksize=500_000):
c:\Users\cjcho\OneDrive\Documentos\AlgebraLinealAplicada\PCA.py:28: DtypeWarning: Columns
(41) have mixed types. Specify dtype option on import or set low_memory=False.
  for chunk in pd.read_csv("train.csv", usecols=[col], chunksize=500_000):
c:\Users\cjcho\OneDrive\Documentos\AlgebraLinealAplicada\PCA.py:28: DtypeWarning: Columns
(41) have mixed types. Specify dtype option on import or set low_memory=False.
  for chunk in pd.read_csv("train.csv", usecols=[col], chunksize=500_000):
c:\Users\cjcho\OneDrive\Documentos\AlgebraLinealAplicada\PCA.py:28: DtypeWarning: Columns
(41) have mixed types. Specify dtype option on import or set low_memory=False.
  for chunk in pd.read_csv("train.csv", usecols=[col], chunksize=500_000):
c:\Users\cjcho\OneDrive\Documentos\AlgebraLinealAplicada\PCA.py:28: DtypeWarning: Columns
(41) have mixed types. Specify dtype option on import or set low_memory=False.
  for chunk in pd.read_csv("train.csv", usecols=[col], chunksize=500_000):
NaN      8884852.0
mid      20914.0
low       9621.0
high      6096.0
dtype: float64
La cantidad de valores nulos es 8884852
El tipo de datos es -> object

```

43. **Census\_PrimaryDiskTotalCapacity**: Cantidad de espacio en disco en el disco principal de la máquina (MB).

```
*** Análisis para columna: Census_PrimaryDiskTotalCapacity ***
Census_PrimaryDiskTotalCapacity
476940.0      2841530.0
953869.0      2175780.0
305245.0      474616.0
122104.0      469060.0
244198.0      452284.0
...
16384000.0      1.0
17167872.0      1.0
10328.0        1.0
10354.0        1.0
10469.0        1.0
Length: 5736, dtype: float64
La cantidad de valores nulos es 53016
El tipo de datos es -> float64
```

44. **Census\_PrimaryDiskTypeName:** Nombre descriptivo del tipo de disco principal (HDD o SSD).

```
*** Análisis para columna: Census_PrimaryDiskTypeName ***
HDD            5806804.0
SSD            2466808.0
UNKNOWN        358251.0
Unspecified    276776.0
NaN            12844.0
dtype: float64
La cantidad de valores nulos es 12844
El tipo de datos es -> object
```

45. **Census\_SystemVolumeTotalCapacity:** Tamaño de la partición donde está instalado el volumen del sistema (MB).

```
*** Análisis para columna: Census_SystemVolumeTotalCapacity ***
Census_SystemVolumeTotalCapacity
NaN            53002.0
28542.0        51998.0
926992.0       50430.0
476389.0       44435.0
953253.0       41572.0
...
7667.0          1.0
7641.0          1.0
7640.0          1.0
7639.0          1.0
7385.0          1.0
Length: 536849, dtype: float64
La cantidad de valores nulos es 53002
El tipo de datos es -> float64
```

46. **Census\_HasOpticalDiskDrive**: El valor "true" indica que la máquina tiene una unidad de disco óptico (CD/DVD).

```
*** Análisis para columna: Census_HasOpticalDiskDrive ***
Census_HasOpticalDiskDrive
0    8232858.0
1    688625.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> int64
```

47. **Census\_TotalPhysicalRAM**: Obtiene la RAM física (MB).

```
*** Análisis para columna: Census_TotalPhysicalRAM ***
Census_TotalPhysicalRAM
4096.0      4094512.0
8192.0      2196505.0
2048.0      1097474.0
16384.0     531558.0
6144.0      398671.0
...
557056.0      1.0
360448.0      1.0
327680.0      1.0
311296.0      1.0
294912.0      1.0
Length: 3447, dtype: float64
La cantidad de valores nulos es 80533
El tipo de datos es -> float64
```

48. **Census\_ChassisTypeName**: Obtiene una representación numérica del tipo de chasis de la máquina. Un valor de 0 significa xx.

*** Análisis para columna: Census_ChassisTypeName ***	
Notebook	5248812.0
Desktop	1872125.0
Laptop	685581.0
Portable	360903.0
AllinOne	204295.0
MiniTower	85127.0
Convertible	84472.0
Other	75782.0
UNKNOWN	67212.0
Detachable	51466.0
LowProfileDesktop	50072.0
HandHeld	46009.0
SpaceSaving	29070.0
Tablet	13630.0
Tower	12549.0
Unknown	10011.0
MainServerChassis	9545.0
MiniPC	4433.0
LunchBox	3971.0
RackMountChassis	3410.0
SubNotebook	807.0
BusExpansionChassis	720.0
NaN	623.0
30	243.0
StickPC	142.0
0	133.0
MultisystemChassis	61.0
Blade	52.0
35	50.0
PizzaBox	46.0
SealedCasePC	39.0
SubChassis	16.0
ExpansionChassis	12.0
31	11.0
88	8.0
32	8.0
127	7.0
25	6.0
44	4.0

```
..  
36           3.0  
81           2.0  
DockingStation      2.0  
CompactPCI          2.0  
BladeEnclosure      2.0  
112          1.0  
76           1.0  
49           1.0  
82           1.0  
45           1.0  
39           1.0  
28           1.0  
IoTGateway        1.0  
EmbeddedPC         1.0  
dtype: float64  
La cantidad de valores nulos es 623  
El tipo de datos es -> object
```

49. **Census\_InternalPrimaryDiagonalDisplaySizeInInches:** Obtiene la longitud diagonal física en pulgadas de la pantalla principal.

```
*** Análisis para columna: Census_InternalPrimaryDiagonalDisplaySizeInInches ***  
Census_InternalPrimaryDiagonalDisplaySizeInInches  
15.5      3047431.0  
13.9      952078.0  
14.0      542450.0  
11.6      319376.0  
21.5      275337.0  
...  
102.1     1.0  
106.5     1.0  
108.4     1.0  
109.4     1.0  
109.7     1.0  
Length: 786, dtype: float64  
La cantidad de valores nulos es 47134  
El tipo de datos es -> float64
```

50. **Census\_InternalPrimaryDisplayResolutionHorizontal:** Obtiene el número de píxeles horizontales de la pantalla interna.

```
*** Análisis para columna: Census_InternalPrimaryDisplayResolutionHorizontal ***
Census_InternalPrimaryDisplayResolutionHorizontal
1366.0    4515064.0
1920.0    2220648.0
1280.0    527430.0
1600.0    501288.0
1024.0    342620.0
...
5873.0      1.0
5876.0      1.0
5920.0      1.0
540.0       1.0
5900.0      1.0
Length: 2181, dtype: float64
La cantidad de valores nulos es 46986
El tipo de datos es -> float64
```

51. **Census\_InternalPrimaryDisplayResolutionVertical:** Obtiene el número de píxeles en dirección vertical de la pantalla interna.

```
*** Análisis para columna: Census_InternalPrimaryDisplayResolutionVertical ***
Census_InternalPrimaryDisplayResolutionVertical
768.0    4973621.0
1080.0   2148402.0
900.0    655155.0
800.0    262058.0
1024.0   186322.0
...
394.0      1.0
390.0      1.0
373.0      1.0
358.0      1.0
285.0      1.0
Length: 1561, dtype: float64
La cantidad de valores nulos es 46986
El tipo de datos es -> float64
```

52. **Census\_PowerPlatformRoleName:** Indica el perfil de administración de energía preferido por el OEM. Este valor ayuda a identificar el factor de forma básico del dispositivo.

```
*** Análisis para columna: Census_PowerPlatformRoleName ***
Mobile           6182908.0
Desktop          2066620.0
Slate            492537.0
Workstation       109683.0
SOHOserver        37841.0
UNKNOWN           20628.0
EnterpriseServer  7094.0
AppliancePC      4015.0
PerformanceServer 97.0
NaN              55.0
Unspecified       5.0
dtype: float64
La cantidad de valores nulos es 55
El tipo de datos es -> object
```

53. **Census\_InternalBatteryType:** NA.

```
*** Análisis para columna: Census_InternalBatteryType ***
NaN      6338429.0
lion     2028256.0
li-i     245617.0
#       183998.0
lip      62099.0
...
p-sn     1.0
sail     1.0
sams     1.0
÷ÿöö     1.0
í♥-i     1.0
Length: 79, dtype: float64
La cantidad de valores nulos es 6338429
El tipo de datos es -> object
```

54. **Census\_InternalBatteryNumberOfCharges:** NA.

```
*** Análisis para columna: Census_InternalBatteryNumberOfCharges ***
Census_InternalBatteryNumberOfCharges
0.000000e+00    5053404.0
4.294967e+09    2252338.0
NaN              268755.0
1.000000e+00    53810.0
2.000000e+00    28128.0
...
6.534300e+04    1.0
6.534000e+04    1.0
6.533600e+04    1.0
6.533500e+04    1.0
6.551900e+04    1.0
Length: 41089, dtype: float64
La cantidad de valores nulos es 268755
El tipo de datos es -> float64
```

55. **Census\_OSVersions:** Versión numérica del SO. Ejemplo: 10.0.10130.0.

```
*** Análisis para columna: Census_OSVersions ***
Census_OSVersions
10.0.17134.228    1413627.0
10.0.17134.165    899711.0
10.0.16299.431    546546.0
10.0.17134.285    470280.0
10.0.16299.547    346853.0
...
10.0.18238.1000    1.0
10.0.18241.1000    1.0
10.0.18244.1000    1.0
10.0.7600.112      1.0
10.0.10240.16391    1.0
Length: 469, dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> object
```

56. **Census\_OSSArchitecture:** Arquitectura en la que se basa el SO. Derivada de OSVersionFull. Ejemplo: amd64.

```
*** Análisis para columna: Census_OSSArchitecture ***
Census_OSSArchitecture
amd64    8105885.0
x86      815252.0
arm64     346.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> object
```

57. **Census\_OSBranch:** Rama del SO extraída de OsVersionFull. Ejemplo: OsB= fbl\_partner\_eeap where OsVersion = 6.4.9813.0.amd64fre.fbl\_partner\_eeap.140810-0005

```

*** Análisis para columna: Census_OSBranch ***
Census_OSBranch
rs4_release           4009158.0
rs3_release            1237321.0
rs3_release_svc_escrow 1199767.0
rs2_release             797066.0
rs1_release              785534.0
th2_release             326655.0
th2_release_sec         266882.0
th1_st1                  195840.0
th1                      75764.0
rs5_release              15324.0
rs3_release_svc_escrow_im 6181.0
rs_prerelease             3171.0
rs_prerelease_flt        2714.0
rs5_release_sigma          62.0
rs1_release_srvmedia       10.0
winblue_ltsb_escrow        8.0
winblue_ltsb                 3.0
win8_gdr                   3.0
win7sp1_ldr                  3.0
win7sp1_ldr_escrow          2.0
rs5_release_edge              2.0
rs5_release_sigma_dev        2.0
rs_xbox                     2.0
rs1_release_sec                1.0
rs5_release_sign                1.0
Khmer OS                      1.0
rs1_release_svc                  1.0
rs3_release_svc                  1.0
rs_shell                      1.0
rs_onecore_base_cobalt          1.0
rs_onecore_stack_per1          1.0
win8_ldr                      1.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> object

```

58. **Census\_OSBuildNumber:** Número de compilación del SO extraído de OsVersionFull. Ejemplo: OsBuildNumber = 10512 o 10240

```
*** Análisis para columna: Census_OSBUILDNumber ***
Census_OSBUILDNumber
17134    4008881.0
16299    2443249.0
15063    797049.0
14393    785450.0
10586    593527.0
...
18226      1.0
18238      1.0
18236      1.0
18241      1.0
18244      1.0
Length: 165, dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> int64
```

59. **Census\_OSBUILDRevision:** Revisión de la compilación del SO extraída de OsVersionFull. Ejemplo: OsBuildRevision = 1000 o 16458

```
*** Análisis para columna: Census_OSBUILDRevision ***
Census_OSBUILDRevision
228      1413633.0
165      899712.0
431      546548.0
285      470280.0
547      346853.0
...
18756      1.0
23418      1.0
21703      1.0
24214      1.0
24149      1.0
Length: 285, dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> int64
```

60. **Census\_OSEdition:** Edición del SO actual. Obtenido de HKLM\Software\Microsoft\Windows NT\CurrentVersion@EditionID en el registro. Ejemplo: Enterprise

```

*** Análisis para columna: Census_OSEdition ***
Census_OSEdition
Core                               3469991.0
Professional                         3130566.0
CoreSingleLanguage                   1945461.0
CoreCountrySpecific                  166100.0
ProfessionalEducation                56698.0
Education                            40704.0
Enterprise                           35603.0
ProfessionalN                        28341.0
Enterprises                          20020.0
ServerStandard                       10128.0
Cloud                                6275.0
CoreN                                4790.0
ServerStandardEval                   2751.0
EducationN                           932.0
EnterpriseSN                         878.0
ServerDatacenterEval                 829.0
ServerSolution                        683.0
EnterpriseN                          351.0
ProfessionalEducationN               192.0
ProfessionalWorkstation                128.0
ServerDatacenter                      15.0
ProfessionalWorkstationN              13.0
CloudN                               8.0
ProfessionalCountrySpecific           5.0
Home                                 4.0
ServerRdsh                            4.0
Ultimate                             4.0
ProfessionalSingleLanguage            3.0
HomePremium                          2.0
Pro                                  1.0
Enterprise 2015 LTSB                 1.0
ServerDatacenterACor                 1.0
professional                          1.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> object

```

61. **Census\_OSSkuName:** Nombre descriptivo de la edición del SO (actualmente solo Windows)

```

*** Análisis para columna: Census_OSSkuName ***
Census_OSSkuName
CORE                               3469869.0
PROFESSIONAL                      3187913.0
CORE_SINGLELANGUAGE                 1945133.0
CORE_COUNTRYSPECIFIC                165886.0
EDUCATION                          40827.0
ENTERPRISE                         35602.0
PROFESSIONAL_N                     28522.0
ENTERPRISE_S                        20022.0
STANDARD_SERVER                     10128.0
CLOUD                              6167.0
CORE_N                             4787.0
STANDARD_EVALUATION_SERVER          2755.0
EDUCATION_N                        927.0
ENTERPRISE_S_N                      881.0
DATACENTER_EVALUATION_SERVER        829.0
SB SOLUTION SERVER                  684.0
ENTERPRISE_N                        356.0
PRO_WORKSTATION                     124.0
UNLICENSED                          17.0
DATACENTER_SERVER                   14.0
PRO_WORKSTATION_N                   12.0
CLOUDN                            7.0
PRO_CHINA                           5.0
SERVERRDSH                         4.0
ULTIMATE                            4.0
PRO_FOR_EDUCATION                  3.0
PRO_SINGLE_LANGUAGE                 2.0
ENTERPRISEG                         1.0
STARTER                            1.0
UNDEFINED                           1.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> object

```

62. **Census\_OSInstallTypeName:** Descripción descriptiva de la instalación utilizada en el equipo (p. ej., limpia)

```
*** Análisis para columna: Census_OSInstallTypeName ***
Census_OSInstallTypeName
UUPUpgrade           2608037.0
IBSClean             1650733.0
Update               1593308.0
Upgrade              1251559.0
Other                840121.0
Reset                649201.0
Refresh              205842.0
Clean                69073.0
CleanPCRefresh        53609.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> object
```

63. **Census\_OSInstallLanguageIdentifier:** NA

\*\*\* Análisis para columna: Census\_OSInstallLanguageIdentifier \*\*\*  
Census\_OSInstallLanguageIdentifier

8.0	3179262.0
9.0	1034201.0
7.0	512753.0
29.0	492267.0
14.0	432503.0
37.0	403190.0
10.0	366636.0
26.0	334766.0
5.0	252887.0
35.0	204832.0
39.0	201525.0
18.0	190828.0
20.0	169059.0
24.0	142175.0
25.0	132408.0
27.0	108176.0
19.0	84177.0
17.0	83445.0
1.0	79777.0
3.0	72370.0
NaN	60084.0
6.0	50489.0
33.0	48800.0
15.0	41514.0
4.0	35407.0
30.0	31956.0
23.0	29496.0
31.0	22149.0
12.0	19906.0
2.0	18270.0
16.0	15831.0
36.0	14585.0
28.0	12696.0
13.0	10869.0
34.0	10647.0
21.0	6865.0
32.0	4559.0
38.0	3737.0
11.0	3219.0

```
22.0      3167.0
dtype: float64
La cantidad de valores nulos es 60084
El tipo de datos es -> float64
```

64. **Census\_OSUILocaleIdentifier:** NA

```
*** Análisis para columna: Census_OSUILocaleIdentifier ***
Census_OSUILocaleIdentifier
31      3170824.0
34      1040042.0
30      513995.0
125     498236.0
49      436691.0
...
108     1.0
104     1.0
144     1.0
147     1.0
155     1.0
Length: 147, dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> int64
```

65. **Census\_OSWUAutoUpdateOptionsName:** Nombre descriptivo de la configuración de actualización automática de WindowsUpdate en el equipo.

```
*** Análisis para columna: Census_OSWUAutoUpdateOptionsName ***
Census_OSWUAutoUpdateOptionsName
FullAuto          3954497.0
UNKNOWN           2519925.0
Notify            2034254.0
AutoInstallAndRebootAtMaintenanceTime 371475.0
Off               26961.0
DownloadNotify    14371.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> object
```

66. **Census\_IsPortableOperatingSystem:** Indica si el sistema operativo se inicia y se ejecuta mediante Windows-To-Go en una memoria USB.

```
*** Análisis para columna: Census_IsPortableOperatingSystem ***
Census_IsPortableOperatingSystem
0    8916619.0
1     4864.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> int64
```

67. **Census\_GenuineStateName:** Nombre descriptivo de OSGenuineStateID. 0 = Genuino

```
*** Análisis para columna: Census_GenuineStateName ***
Census_GenuineStateName
IS_GENUINE          7877597.0
INVALID_LICENSE      801692.0
OFFLINE              228366.0
UNKNOWN              13826.0
TAMPERED              2.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> object
```

68. **Census\_ActivationChannel:** Clave de licencia minorista o clave de licencia por volumen para una máquina.

```
*** Análisis para columna: Census_ActivationChannel ***
Census_ActivationChannel
Retail          4727589.0
OEM:DM          3413350.0
Volume:GVLK      450954.0
OEM:NONSLP        317980.0
Volume:MAK         8028.0
Retail:TB:Eval      3582.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> object
```

69. **Census\_IsFlightingInternal:** NA

```
*** Análisis para columna: Census_IsFlightingInternal ***
Census_IsFlightingInternal
NaN      7408759.0
0.0     1512703.0
1.0        21.0
dtype: float64
La cantidad de valores nulos es 7408759
El tipo de datos es -> float64
```

70. **Census\_IsFlightsDisabled:** Indica si la máquina participa en la gestión de vuelos.

```
*** Análisis para columna: Census_IsFlightsDisabled ***
Census_IsFlightsDisabled
0.0     8760872.0
NaN     160523.0
1.0       88.0
dtype: float64
La cantidad de valores nulos es 160523
El tipo de datos es -> float64
```

71. **Census\_FlightRing:** El anillo para el que el usuario del dispositivo desea recibir vuelos. Este puede ser diferente del anillo del sistema operativo instalado actualmente si el usuario cambia el anillo después de obtener un vuelo de otro anillo.

```
*** Análisis para columna: Census_FlightRing ***
Census_FlightRing
Retail      8355679.0
NOT_SET    287803.0
Unknown    243438.0
WIS         10648.0
WIF         10322.0
RP          9860.0
Disabled    3722.0
OSG          7.0
Canary       3.0
Invalid      1.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> object
```

72. **Census\_ThresholdOptIn** - NA

```
*** Análisis para columna: Census_ThresholdOptIn ***
Census_ThresholdOptIn
NaN      5667325.0
0.0     3253342.0
1.0       816.0
dtype: float64
La cantidad de valores nulos es 5667325
El tipo de datos es -> float64
```

#### 73. **Census\_FirmwareManufacturerIdentifier** - NA

```
*** Análisis para columna: Census_FirmwareManufacturerIdentifier ***
Census_FirmwareManufacturerIdentifier
142.0    2699078.0
628.0    1229140.0
554.0    1175137.0
355.0    941793.0
556.0    800536.0
...
1052.0      1.0
1050.0      1.0
27.0       1.0
25.0       1.0
19.0       1.0
Length: 713, dtype: float64
La cantidad de valores nulos es 183257
El tipo de datos es -> float64
```

#### 74. **Census\_FirmwareVersionIdentifier** - NA

```
*** Análisis para columna: Census_FirmwareVersionIdentifier ***
Census_FirmwareVersionIdentifier
NaN      160133.0
33105.0   89611.0
33111.0   61583.0
33054.0   56626.0
33108.0   55040.0
...
55491.0      1.0
55490.0      1.0
55488.0      1.0
55529.0      1.0
55525.0      1.0
Length: 50495, dtype: float64
La cantidad de valores nulos es 160133
El tipo de datos es -> float64
```

#### 75. **Census\_IsSecureBootEnabled**: Indica si el modo de arranque seguro está habilitado.

```
*** Análisis para columna: Census_IsSecureBootEnabled ***
Census_IsSecureBootEnabled
0    4585438.0
1    4336045.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> int64
```

76. **Census\_IsWIMBootEnabled:** NA

```
*** Análisis para columna: Census_IsWIMBootEnabled ***
Census_IsWIMBootEnabled
NaN    5659703.0
0.0    3261779.0
1.0      1.0
dtype: float64
La cantidad de valores nulos es 5659703
El tipo de datos es -> float64
```

77. **Census\_IsVirtualDevice:** Identifica una máquina virtual (modelo de aprendizaje automático).

```
*** Análisis para columna: Census_IsVirtualDevice ***
Census_IsVirtualDevice
0.0    8842840.0
1.0    62690.0
NaN    15953.0
dtype: float64
La cantidad de valores nulos es 15953
El tipo de datos es -> float64
```

78. **Census\_IsTouchEnabled:** ¿Es un dispositivo táctil?

```
*** Análisis para columna: Census_IsTouchEnabled ***
Census_IsTouchEnabled
0    7801452.0
1    1120031.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> int64
```

79. **Census\_IsPenCapable:** ¿El dispositivo admite entrada de lápiz?

```
*** Análisis para columna: Census_IsPenCapable ***
Census_IsPenCapable
0    8581834.0
1    339649.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> int64
```

80. **Census\_IsAlwaysOnAlwaysConnectedCapable:** Obtiene información sobre si la batería permite que el dispositivo esté siempre conectado.

```
*** Análisis para columna: Census_IsAlwaysOnAlwaysConnectedCapable ***
Census_IsAlwaysOnAlwaysConnectedCapable
0.0    8341972.0
1.0    508168.0
NaN    71343.0
dtype: float64
La cantidad de valores nulos es 71343
El tipo de datos es -> float64
```

81. **Wdft\_IsGamer:** Indica si el dispositivo es para juegos según su combinación de hardware.

```
*** Análisis para columna: Wdft_IsGamer ***
Wdft_IsGamer
0.0    6174143.0
1.0    2443889.0
NaN    303451.0
dtype: float64
La cantidad de valores nulos es 303451
El tipo de datos es -> float64
```

82. **Wdft\_RegionIdentifier:** NA

```
*** Análisis para columna: Wdft_RegionIdentifier ***
Wdft_RegionIdentifier
10.0    1800105.0
11.0    1347828.0
3.0     1295892.0
1.0     1232258.0
15.0    1017591.0
7.0     597297.0
NaN     303451.0
8.0     276029.0
13.0    225130.0
5.0     205372.0
12.0    163711.0
6.0     158163.0
4.0     135567.0
9.0     79882.0
2.0     79385.0
14.0    3822.0
dtype: float64
La cantidad de valores nulos es 303451
El tipo de datos es -> float64
```

83. **HasDetections**: Columna que indica si se ha detectado malware en esta máquina (0 si no y 1 en caso contrario).

```
*** Análisis para columna: HasDetections ***
HasDetections
0     4462591.0
1     4458892.0
dtype: float64
La cantidad de valores nulos es 0
El tipo de datos es -> int64
```

## Selección de variables numéricas significativas

Para preparar un PCA (Análisis de Componentes Principales), necesitamos identificar qué variables son numéricas, ya que PCA no funciona directamente con variables categóricas. Además, debe evaluarse la necesidad de incluir identificadores; por ejemplo si una variable que refleja un ID codifica de alguna manera una característica significativa para el resultado, entonces podría usarse como entrada. Por ejemplo, si los ID son un indicador de la ubicación y hora de origen, el tipo de entidad, etc., y estos son importantes para el resultado, el modelo podría aprender a realizar predicciones acertadas dado un nuevo ID desconocido.

Sin embargo, si se trata simplemente de un ID aleatorio que sirve como marcador de otras características significativas almacenadas en otro lugar, el modelo no tiene datos con los que pueda trabajar. Por lo tanto, no funcionará.

. La siguiente tabla muestra las variables y sus tipos. Se marca en rojo aquellas que no se incluirán al momento de desarrollar el modelo y se resaltan en verde oscuro aquellas que se consideran de gran relevancia para PCA (por su naturaleza numérica y “comparable”):

Variable	Tipo
MachinelIdentifier	ID única
ProductName	Categórica (texto)
EngineVersion	Categórica (versión, tipo texto)
AppVersion	Categórica (versión, tipo texto)
AvSigVersion	Categórica (versión, tipo texto)
IsBeta	Booleana (0/1)
RtpStateBitfield	Identificador codificado (entero, pero representa estados combinados)
IsSxsPassiveMode	Booleana (0/1)
DefaultBrowsersIdentifier	Identificador codificado (entero)
AVProductStatesIdentifier	Identificador codificado (entero)
AVProductsInstalled	Numérica (entera discreta)
AVProductsEnabled	Numérica (entera discreta)
HasTpm	Booleana (0/1)
CountryIdentifier	Categórica codificada (ID)
CityIdentifier	Categórica codificada (ID)
OrganizationIdentifier	Categórica codificada (ID)
GeoNameIdentifier	Categórica codificada (ID)
LocaleEnglishNameIdentifier	Categórica codificada (ID de idioma)
Platform	Categórica (texto)
Processor	Categórica (texto)
OsVer	Categórica (versión en string)

OsBuild	Numérica (entera)
OsSuite	Categórica codificada (enteros que representan combinaciones de componentes del SO)
OsPlatformSubRelease	Categórica (texto)
OsBuildLab	Categórica (texto)
SkuEdition	Categórica (texto)
IsProtected	Booleana (0/1)
AutoSampleOptIn	Booleana (0/1)
PuaMode	Categórica (texto)
SMode	Booleana (0/1)
leVerIdentifier	Categórica codificada (ID)
SmartScreen	Categórica (texto)
Firewall	Booleana (0/1)
UacLuaenable	Booleana (0/1)
Census_MDC2FormFactor	Categórica (tipo de dispositivo)
Census_DeviceFamily	Categórica
Census_OEMNameIdentifier	Categórica codificada (ID)
Census_OEMModelIdentifier	Categórica codificada (ID)
Census_ProcessorCoreCount	Numérica (entera)
Census_ProcessorManufacturerIdentifier	Categórica codificada
Census_ProcessorModelIdentifier	Categórica codificada
Census_ProcessorClass	Categórica
Census_PrimaryDiskTotalCapacity	Numérica (entera)
Census_PrimaryDiskTypeName	Categórica
Census_SystemVolumeTotalCapacity	Numérica
Census_HasOpticalDiskDrive	Booleana
Census_TotalPhysicalRAM	Numérica
Census_ChassisTypeName	Categórica
Census_InternalPrimaryDiagonalDisplaySize	Numérica (float)

eInInches	
Census_InternalPrimaryDisplayResolutionHorizontal	Numérica (entera)
Census_InternalPrimaryDisplayResolutionVertical	Numérica (entera)
Census_PowerPlatformRoleName	Categórica
Census_InternalBatteryType	Categórica
Census_InternalBatteryNumberOfCharges	Numérica (entera, puede tener NaNs)
Census_OSVersion	Categórica
Census_OSSubFamily	Categórica (como 'x64', 'x86')
Census_OSBranch	Categórica
Census_OSBUILD	Numérica
Census_OSBUILDRevision	Numérica
Census_OSEdition	Categórica
Census_OSSkuName	Categórica
Census_OSIInstallTypeName	Categórica
Census_OSIInstallLanguageIdentifier	Categórica codificada
Census_OSUILocaleIdentifier	Categórica codificada
Census_OSWUAutoUpdateOptionsName	Categórica
Census_IsPortableOperatingSystem	Booleana
Census_GenuineStateName	Categórica
Census_ActivationChannel	Categórica
Census_IsFlightingInternal	Booleana
Census_IsFlightsDisabled	Booleana
Census_FlightRing	Categórica
Census_ThresholdOptIn	Booleana
Census_FirmwareManufacturerIdentifier	Categórica codificada
Census_FirmwareVersionIdentifier	Categórica codificada
Census_IsSecureBootEnabled	Booleana
Census_IsWIMBootEnabled	Booleana

Census_IsVirtualDevice	Booleana
Census_IsTouchEnabled	Booleana
Census_IsPenCapable	Booleana
Census_IsAlwaysOnAlwaysConnectedCapable	Booleana
Wdft_IsGamer	Booleana
Wdft_RegionIdentifier	Categórica codificada
HasDetections	Booleana (target binario)

## Descripción del programa *PCA\_Small\_Sample.py*:

### Principales funciones:

1. `tomar_muestra_csv(ruta_csv, n_muestras)`

```
# ----- 1. Carga la muestra aleatoria -----
def tomar_muestra_csv(ruta_csv, n_muestras):
    """Devuelve un DataFrame con una muestra aleatoria de filas del archivo csv."""
    with open(ruta_csv, 'r', encoding='utf-8') as f:
        n_filas = sum(1 for _ in f) - 1
    if n_muestras > n_filas:
        n_muestras = n_filas
    filas_a_ignorar = random.sample(range(1, n_filas + 1), n_filas - n_muestras)
    filas_a_saltar = sorted(filas_a_ignorar)
    print("Estas fueron las filas ignoradas -> ", filas_a_ignorar)
    df = pd.read_csv(ruta_csv, skiprows=filas_a_saltar)
    return df
```

### Argumentos:

- `ruta_csv`: Ruta del csv para crear el DataFrame
- `n_muestras`: Cantidad aleatoria de muestras que se desea tomar

Este método toma una muestra aleatoria de un CSV sin cargar todo el archivo a memoria. Primero calcula cuántas filas tiene el archivo (`n_filas`), luego se seleccionan aleatoriamente las filas que no se van a usar (`filas_a_ignorar`) y luego se cargan todas las demás (es decir, las que sí se usarán como muestra).

Retornará el dataframe con una cantidad de `n_muestras` aleatorias del original.

2. `codificar_categoricas(X, codificadores= None)`

### Parámetros:

- `X`: DataFrame de pandas

- codificadores: Diccionario opcional con codificaciones previas por columnas. Si es None, se genera uno nuevo.
- Retorna:**
- X: DataFrame con columnas categóricas convertidas a valores numéricos.
  - codificadores: Diccionario con las codificaciones usadas por columna.

Este método identifica todas las columnas categóricas del DataFrame X. Si no se proporcionan codificadores, genera un diccionario donde a cada valor único de cada columna se le asigna un número entero. Luego, reemplaza cada valor categórico en X por su correspondiente número según ese diccionario. Si encuentra columnas categóricas no previstas (por ejemplo, en datos de test), las deja en None.

```
# ----- 2. Codifica las variables categóricas -----
def codificar_categoricas(X, codificadores=None):
    categoricas = X.select_dtypes(include='object').columns
    if codificadores is None:
        codificadores = {}
    for col in categoricas:
        valores_unicos = X[col].dropna().unique()
        codificadores[col] = {val: idx for idx, val in enumerate(valores_unicos)}

    for col in categoricas:
        if col in codificadores:
            X[col] = X[col].map(codificadores[col])
        else:
            # Para evitar columnas inesperadas al tratar con test.csv:
            print(f"Columna inesperada en test: {col} – ignorando o asignando NaN.")
            X[col] = None # o X.drop(columns=[col], inplace=True) si prefieres eliminarla

    return X, codificadores
```

## 2. preprocessar\_df(df, incluir\_target=True)

**Parámetros:**

- df: DataFrame con los datos.
- incluir\_target: bool. Si es True, también retorna la columna objetivo (HasDetections).

**Retorna:**

- X: DataFrame con las características (sin la columna HasDetections ni MachinelIdentifier).
- y: Serie con la columna HasDetections si incluir\_target=True; de lo contrario, None.

Separa las variables independientes (X) de la variable objetivo (y). También elimina la columna MachinelIdentifier, que es un identificador irrelevante para el modelo.

```

# ----- 3. Preprocesa los dataframes -----
def preprocessar_df(df, incluir_target=True):
    if incluir_target:
        y = df['HasDetections']
        X = df.drop(columns=['HasDetections', 'MachineIdentifier'], errors='ignore')
    else:
        y = None
        X = df.drop(columns=['MachineIdentifier'], errors='ignore')

    return X, y

```

### 3. procesar\_train(df)

```

def procesar_train(df):
    df_copy = df.copy()
    # Eliminamos columnas con más del 50% de valores nulos
    porcentaje_nulos = df_copy.isnull().mean()
    columnas_filtradas = porcentaje_nulos[porcentaje_nulos <= 0.5].index.tolist()
    df_copy = df_copy[columnas_filtradas]
    # Separamos columnas numéricas y categóricas
    df_num = df_copy.select_dtypes(include=['number'])
    df_cat = df_copy.select_dtypes(include='object')
    # Vamos a mostrar la matriz de correlación y las variables altamente correlacionadas antes de PCA
    mostrar_correlacion(df_num)
    mostrar_altamente_correlacionadas(df_num, umbral=0.8)
    # Reemplazamos nulos en numéricas
    medianas = df_num.median()
    df_num = df_num.fillna(medianas)
    # Escalado y PCA
    scaler = StandardScaler()
    df_num_scaled = scaler.fit_transform(df_num)
    pca_model = PCA(n_components=0.95)
    df_num_pca = pca_model.fit_transform(df_num_scaled)
    # Mostramos las componentes principales
    print("\n[ Componentes principales del PCA (varianza explicada >= 95%) ]")
    componentes = pca_model.components_
    for i, comp in enumerate(componentes, start=1):
        pesos = [(col, round(peso, 3)) for col, peso in zip(df_num.columns, comp)]
        pesos_ordenados = sorted(pesos, key=lambda x: abs(x[1]), reverse=True)[:5] # top 5 por componente
        print(f"Componente {i}:")
        for variable, peso in pesos_ordenados:
            print(f"  {variable}: {peso}")
    # Codificamos variables categóricas
    df_cat_cod, codificadores = codificar_categoricas(df_cat)
    # Se concatenan las columnas transformadas numéricas (por PCA) y las categóricas codificadas en un solo array final, llamado X_train_final.
    X_train_final = np.hstack([df_num_pca, df_cat_cod.values])

    return X_train_final, medianas, scaler, pca_model, codificadores, columnas_filtradas

```

#### Parámetros:

- df: Un DataFrame de pandas que contiene los datos originales de entrenamiento, con columnas numéricas y categóricas, y posibles valores nulos.

#### Retorna:

- X\_train\_final: numpy.ndarray con las características finales (numéricas y categóricas preprocesadas).
- medianas: Medianas calculadas para imputar valores nulos numéricos.
- scaler: Objeto StandardScaler ya ajustado.
- pca\_model: Objeto PCA ajustado.
- codificadores: Diccionario de codificación categórica.
- columnas\_filtradas: Lista de columnas no eliminadas por exceso de nulos.

Este método tiene varias subfunciones. A nivel general, procesa los datos de entrenamiento siguiendo los pasos dados a continuación:

1. Se crea una copia para no modificar el DataFrame original
2. Elimina columnas con más del 50% de valores nulos.
3. Divide el DataFrame en variables numéricas (df\_num) y categóricas (df\_cat).
4. Muestra la matriz de correlación para df\_num y las variables numéricas altamente correlacionadas (antes de hacer PCA): Llama a la función *mostrar\_altamente\_correlacionadas* para imprimir pares de variables numéricas con correlación absoluta mayor o igual a 0.8. Esto sirve para detectar redundancias o multicolinealidad.
5. Imputa valores nulos en columnas numéricas usando la mediana.
6. Escala los datos numéricos con StandardScaler y aplica reducción de dimensionalidad usando PCA para conservar el 95% de la varianza.
7. Se imprimen las componentes principales que PCA encontró, mostrando para cada componente las 5 variables originales con mayor peso absoluto (influencia) en esa componente.
8. Codifica las columnas categóricas con codificar\_categoricals.
9. Une los datos transformados en un único arreglo numpy.

#### **4. graficar\_distribucion\_y(y, titulo='Distribución de HasDetections')**

##### **Parámetros:**

- y: Serie o columna objetivo (HasDetections).
- titulo: Título opcional para el gráfico.

##### **Retorna:**

- Nada (solo imprime y grafica).

Imprime en consola la proporción de clases (0 y 1) en la variable objetivo. Luego, genera un gráfico de torta para mostrar visualmente la distribución de clases.

```
# ===== 5. Visualización de y =====
def graficar_distribucion_y(y, titulo='Distribución de HasDetections'):
    print("\nDistribución de HasDetections:")
    proporcion = y.value_counts(normalize=True).rename('proportion')
    print(proporcion)

    plt.figure(figsize=(6, 6))
    counts = y.value_counts()
    plt.pie(counts, labels=['No Detection (0)', 'Has Detection (1)'],
            autopct='%1.1f%%', startangle=90, colors=['lightcoral', 'lightskyblue'])
    plt.axis('equal')
    plt.title(titulo)
    plt.show()
```

#### **5. mostrar\_correlacion(df)**

##### **Parámetros:**

- df: DataFrame con variables numéricas.

**Retorna:**

- Nada (solo genera el gráfico).

Este método selecciona las columnas numéricas del DataFrame y calcula la matriz de correlación de Pearson entre ellas. Luego, la visualiza como un mapa de calor (heatmap) usando seaborn, anotando los valores de correlación.

```
# ===== 6. Gráfico matriz de correlación =====
def mostrar_correlacion(df):
    numericas = df.select_dtypes(include=['number']).columns
    corr_matrix = df[numericas].corr()

    plt.figure(figsize=(12, 10))
    sns.heatmap(corr_matrix,
                annot=True, fmt='.2f', cmap='coolwarm', square=True,
                annot_kws={"size": 5})
    plt.xticks(fontsize=8, rotation=90)
    plt.yticks(fontsize=8)
    plt.title("Matriz de correlación - Variables numéricas", fontsize=12)
    plt.tight_layout()
    plt.show()
```

## 6. procesar\_test(df, columnas\_utilizadas, medianas, scaler, pca\_model, codificadores)

**Parámetros:**

- df: DataFrame de test sin procesar.
- columnas\_utilizadas: Columnas que se conservaron durante el preprocesamiento del entrenamiento.
- medianas: Medianas calculadas sobre el set de entrenamiento para imputación.
- scaler: Objeto StandardScaler ya ajustado.
- pca\_model: Objeto PCA ya entrenado.
- codificadores: Diccionario de codificación categórica entrenado.

**Retorna:**

- X\_test\_final: Arreglo numpy.ndarray con los datos del test procesados de forma compatible con el modelo entrenado.

De manera análoga a *procesar\_train* este método se encarga de procesar el DataFrame obtenido de test.csv, sólo que usando los valores que usaron o se obtuvieron durante el entrenamiento (Scaler, codificadores, pca\_model, etc).

1. Filtra el DataFrame para mantener solo las columnas válidas del entrenamiento.
2. Separa columnas numéricas y categóricas.
3. Imputa valores nulos de numéricas con las medianas calculadas del entrenamiento.
4. Escala y transforma las numéricas usando scaler y pca\_model.
5. Codifica las categóricas con los codificadores entrenados.
6. Combina ambas transformaciones y retorna el resultado.

```
# ===== 6. Aplicar preprocesamiento al test.csv =====

def procesar_test(df, columnas_utilizadas, medianas, scaler, pca_model, codificadores):
    df_copy = df.copy()

    # Usa solo las columnas filtradas desde train
    df_copy = df_copy[columnas_utilizadas]

    # Separa columnas numéricas y categóricas
    df_num = df_copy.select_dtypes(include=['number'])
    df_cat = df_copy.select_dtypes(include='object')

    # Reemplazamos nulos con medianas aprendidas
    df_num = df_num.fillna(medianas)

    # Escalado y PCA
    df_num_scaled = scaler.transform(df_num)
    df_num_pca = pca_model.transform(df_num_scaled)

    # Codificación categórica
    df_cat_cod, _ = codificar_categoricas(df_cat, codificadores)

    # Concatenación
    X_test_final = np.hstack([df_num_pca, df_cat_cod.values])

    return X_test_final
```

## 7. mostrar\_altamente\_correlacionadas

### Descripción del programa y del flujo principal

Dentro del main, primero se carga una muestra aleatoria del archivo CSV de entrenamiento. Debido a que el dataset original es muy grande, en lugar de cargarlo entero, el código toma una muestra aleatoria de 500 filas para trabajar más rápido y con menos memoria.

```
# Main -----
# --- Paso 1: Cargamos los datos
df_train = tomar_muestra_csv('train.csv', 500)
df_train = df_train.set_index('MachineIdentifier')
```

Setear los identificadores de las máquinas como índices nos permite seguir el rastro de las filas tras las operaciones que realizaremos más adelante

Seguidamente, se preprocesan los datos cargados, separando las variables predictoras (X) y la variable objetivo (y, que es HasDetections) y se eliminan las columnas innecesarias o identificadores (MachineIdentifier).

```
# --- Paso 2: Preprocesamiento inicial de los datos de entrenamiento
x, y = preprocesar_df(df_train, incluir_target=True)
```

Luego, se hace un preprocesamiento más detallado del conjunto de entrenamiento (X) con el método de *procesar\_train*:

```
x_train_final, medianas, scaler, pca_model, codificadores, columnas_filtradas = procesar_train(x)
```

En este punto se muestran gráficos de pastel para ver la proporción de casos con detección y sin detección, antes y después de la limpieza/preprocesamiento.

```
# --- Paso 3: Visualización y original
graficar_distribucion_y(y, 'Distribución ORIGINAL')
# --- Paso 4: Recuperamos correspondencia entre filas
y_tratado = y.loc[x_train_final.index]
# --- Paso 5: Visualizamos la distribución de la variable objetivo y después de limpieza
graficar_distribucion_y(y_tratado, 'Distribución TRAS LIMPIEZA')
```

En el paso siguiente, entrenamos un modelo de clasificación Random Forest con 50 árboles, usando los datos ya procesados.

```
# --- Paso 6: Entrenamos un modelo RandomForestClassifier simple
clf = RandomForestClassifier(n_estimators=50, random_state=42)
clf.fit(x_train_final, y_tratado)

# --- Paso 7: Cargamos y preparamos el dataset de test. Como es muy grande,
test_df = tomar_muestra_csv('test.csv', 500)
test_df = test_df.set_index('MachineIdentifier')

x_test_final = procesar_test(test_df, columnas_filtradas, medianas, scaler, | 
# Hacemos esta conversión sobre el dataset de test ya preparado para ver col
x_test_final = pd.DataFrame(x_test_final, index=test_df.index)

# Verificamos qué columnas hay en ambos dataframes
print("Lista X_train", list(x_train_final.columns))
print("Lista X_test", list(x_test_final.columns))
print(len(list(x_train_final.columns))==len(list(x_test_final.columns)))
print(clf.classes_) # Esto dice qué clases aprendió el modelo

# --- Paso 8: Hacemos la predicción sobre el conjunto de test ya procesado
predicciones = clf.predict(x_test_final)
```

Finalmente, hacemos la predicción de las etiquetas HasDetections para las muestras de test procesadas, usando el modelo entrenado y las guardamos en un archivo CSV que contiene el identificador de máquina y la predicción de detección para cada muestra del test.

```

# --- Paso 8: Hacemos la predicción sobre el conjunto de test ya procesado
prediccciones = clf.predict(x_test_final)

# --- Paso 9: Pasamos las predicciones hechas por el modelo a un archivo llamado resultados.csv
submission = pd.DataFrame({
    'MachineIdentifier': test_df.index,
    'HasDetections': prediccciones
})

submission.to_csv('resultados.csv', index=False)
print("\nArchivo 'resultados.csv' generado correctamente")

```

## Observaciones con respecto al código:

- 1) La imputación con la mediana se utiliza en este programa porque la mediana es una medida robusta de tendencia central que no se ve afectada por valores extremos o outliers, a diferencia de la media. Esto es especialmente útil cuando la distribución de los datos es sesgada o contiene valores atípicos. Básicamente, imputar con la mediana ayuda a mantener la integridad estadística de la variable sin distorsionar su distribución.
- 2) El conjunto de test debe ser transformado usando únicamente la información del conjunto de entrenamiento para evitar el data leakage (fuga de datos). Esto significa que cualquier cálculo (como la media, mediana o parámetros para normalización) debe basarse solo en el entrenamiento para simular cómo funcionaría el modelo en datos nuevos, no vistos antes. Esto garantiza una evaluación objetiva y realista del rendimiento del modelo.
- 3) Se descartan las columnas con más del 50% de valores faltantes porque suelen tener poca información útil y su imputación puede introducir mucho ruido o sesgo. Además, mantener columnas con tantos valores nulos puede dificultar el entrenamiento del modelo y degradar su rendimiento. Por eso, es común eliminar estas variables para simplificar el modelo y mejorar su calidad.
- 4) Los algoritmos de machine learning generalmente solo trabajan con datos numéricos, por lo que es necesario transformar variables categóricas en una representación numérica. La codificación permite que los modelos interpreten y usen la información contenida en variables categóricas.
- 5) Un umbral de correlación absoluto mayor que 0.8 (Statistics by Jim, s.f.) es un indicador justo de una relación lineal fuerte entre dos variables, por tal razón lo tomamos aquí para identificar las variables altamente correlacionadas.

## Análisis de datos basado en los resultados de cinco muestras aleatorias de train.csv

### Muestra 1:

Los resultados de la muestra 1 fueron los siguientes:

Se ignoraron 8920983 filas. Guardado en: FilasIgnoradas.txt

Variables altamente correlacionadas ( $>|0.8|$ ):

RtpStateBitfield <--> IsSxsPassiveMode = 0.89

OsBuild <--> Census\_OSBuildNumber = 0.93

Census\_ProcessorManufacturerIdentifier <--> Census\_ProcessorModelIdentifier = 0.81

Census\_InternalPrimaryDisplayResolutionHorizontal <-->

Census\_InternalPrimaryDisplayResolutionVertical = 0.88

Census\_OSInstallLanguageIdentifier <--> Census\_OSUILocaleIdentifier

= 0.97

[ Componentes principales del PCA (varianza explicada  $\geq 95\%$ ) ]

Componente 1:

Census\_InternalPrimaryDiagonalDisplaySizeInInches: 0.373

Census\_InternalPrimaryDisplayResolutionVertical: 0.342

Census\_InternalPrimaryDisplayResolutionHorizontal: 0.339

Census\_InternalBatteryNumberOfCharges: 0.311

Census\_TotalPhysicalRAM: 0.289

Componente 2:

Census\_SystemVolumeTotalCapacity: 0.376

AVProductsInstalled: 0.312

Census\_PrimaryDiskTotalCapacity: 0.303

IsSxsPassiveMode: 0.291

Census\_IsSecureBootEnabled: 0.283

Componente 3:

Census\_IsAlwaysOnAlwaysConnectedCapable: 0.37

Census\_IsPenCapable: 0.323

Census\_IsTouchEnabled: 0.315

Census\_PrimaryDiskTotalCapacity: -0.305

Census\_ProcessorManufacturerIdentifier: 0.261

Componente 4:

Census\_OSUILocaleIdentifier: 0.516

Census\_OSInstallLanguageIdentifier: 0.505

AVProductStatesIdentifier: -0.198

AVProductsInstalled: 0.198

Census\_IsTouchEnabled: -0.189

Componente 5:

Census\_OSBuildNumber: 0.366

OsBuild: 0.352

Census\_OSBuildRevision: -0.315

IsSxsPassiveMode: -0.248

RtpStateBitfield: 0.244

Componente 6:

Census\_ProcessorManufacturerIdentifier: 0.377

Census\_ProcessorModelIdentifier: 0.368

RtpStateBitfield: -0.366

IsSxsPassiveMode: 0.36

GeoNameIdentifier: 0.247

Componente 7:

IeVerIdentifier: 0.386

HasTpm: -0.342  
Census\_ProcessorModelIdentifier: 0.326  
Census\_ProcessorManufacturerIdentifier: 0.3  
CountryIdentifier: -0.29

Componente 8:

CountryIdentifier: 0.337  
Census\_ProcessorModelIdentifier: -0.334  
Census\_ProcessorManufacturerIdentifier: -0.328  
Census\_FirmwareVersionIdentifier: -0.273  
GeoNameIdentifier: 0.238

Componente 9:

HasTpm: 0.404  
IeVerIdentifier: -0.401  
Census\_OEMNameIdentifier: 0.342  
Census\_FirmwareVersionIdentifier: -0.338  
Census\_FirmwareManufacturerIdentifier: 0.272

Componente 10:

AVProductStatesIdentifier: 0.47  
AVProductsInstalled: -0.335  
AVProductsEnabled: -0.321  
Census\_ProcessorCoreCount: 0.214  
Census\_OEMNameIdentifier: -0.212

Componente 11:

CountryIdentifier: 0.371  
GeoNameIdentifier: 0.332  
Census\_FirmwareManufacturerIdentifier: 0.277  
Census\_PrimaryDiskTotalCapacity: 0.23  
Wdft\_RegionIdentifier: -0.227

Componente 12:

Firewall: 0.393  
Census\_IsVirtualDevice: 0.384  
LocaleEnglishNameIdentifier: -0.333  
Census\_FirmwareManufacturerIdentifier: 0.231  
Census\_PrimaryDiskTotalCapacity: -0.221

Componente 13:

OrganizationIdentifier: 0.378  
Census\_HasOpticalDiskDrive: 0.297  
Census\_IsVirtualDevice: -0.291  
Census\_OEMModelIdentifier: -0.262  
Census\_IsTouchEnabled: -0.242

Componente 14:

UacLuaenable: 0.615  
Firewall: 0.368  
Wdft\_IsGamer: 0.236  
GeoNameIdentifier: 0.219  
Wdft\_RegionIdentifier: 0.218

Componente 15:

Census\_HasOpticalDiskDrive: 0.424

Wdft\_RegionIdentifier: -0.402

Wdft\_IsGamer: 0.396

CityIdentifier: 0.39

AVProductsEnabled: -0.332

Componente 16:

Census\_HasOpticalDiskDrive: 0.449

Wdft\_RegionIdentifier: 0.402

LocaleEnglishNameIdentifier: 0.394

IsProtected: 0.289

Census\_IsVirtualDevice: 0.246

Componente 17:

IsProtected: 0.702

CityIdentifier: -0.257

Census\_OEMModelIdentifier: 0.236

AVProductsInstalled: -0.201

Census\_HasOpticalDiskDrive: -0.194

Componente 18:

CityIdentifier: 0.563

OrganizationIdentifier: 0.478

LocaleEnglishNameIdentifier: 0.286

Census\_HasOpticalDiskDrive: -0.223

Firewall: -0.187

Componente 19:

AVProductsEnabled: 0.528

Census\_OEMNameIdentifier: 0.332

IsProtected: -0.268

LocaleEnglishNameIdentifier: -0.254

OsSuite: 0.228

Componente 20:

Wdft\_IsGamer: 0.381

Census\_IsVirtualDevice: 0.379

Census\_FirmwareVersionIdentifier: 0.368

OrganizationIdentifier: 0.349

AVProductsEnabled: 0.222

Componente 21:

Wdft\_RegionIdentifier: 0.412

LocaleEnglishNameIdentifier: -0.394

CityIdentifier: 0.326

UacLuaenable: -0.324

Census\_HasOpticalDiskDrive: 0.305

Componente 22:

AVProductsEnabled: 0.517

CityIdentifier: 0.41

OrganizationIdentifier: -0.311

Census\_OEMModelIdentifier: -0.277

Census\_OEMNameIdentifier: -0.24

Componente 23:

UacLuaenable: 0.368

Firewall: -0.326

OrganizationIdentifier: -0.289

Census\_TotalPhysicalRAM: 0.278

LocaleEnglishNameIdentifier: -0.269

Componente 24:

Wdft\_IsGamer: 0.518

Census\_HasOpticalDiskDrive: -0.477

Wdft\_RegionIdentifier: 0.435

AVProductStatesIdentifier: -0.189

AVProductsInstalled: 0.161

Componente 25:

Firewall: 0.445

LocaleEnglishNameIdentifier: 0.401

UacLuaenable: -0.324

Census\_ProcessorCoreCount: 0.292

Census\_InternalBatteryNumberOfCharges: -0.267

Componente 26:

Census\_OEMModelIdentifier: 0.422

Census\_InternalBatteryNumberOfCharges: -0.378

Census\_FirmwareVersionIdentifier: -0.339

Census\_InternalPrimaryDisplayResolutionHorizontal: 0.3

Census\_InternalPrimaryDisplayResolutionVertical: 0.281

Componente 27:

Census\_FirmwareVersionIdentifier: 0.51

Census\_OEMNameIdentifier: 0.467

Census\_IsVirtualDevice: -0.44

Census\_FirmwareManufacturerIdentifier: 0.211

Census\_TotalPhysicalRAM: -0.193

Componente 28:

Census\_IsPenCapable: 0.619

Census\_IsAlwaysOnAlwaysConnectedCapable: -0.437

Census\_IsTouchEnabled: -0.292

HasTpm: 0.209

Census\_OEMModelIdentifier: 0.162

Componente 29:

Census\_IsAlwaysOnAlwaysConnectedCapable: 0.434

Census\_IsSecureBootEnabled: -0.405

Census\_OSBuildRevision: -0.396

HasTpm: 0.323

Census\_IsTouchEnabled: -0.286

Componente 30:

OsSuite: 0.571

Census\_IsSecureBootEnabled: -0.521

IeVerIdentifier: -0.209

Census\_ProcessorCoreCount: -0.208

HasTpm: -0.201

Componente 31:

Census\_IsTouchEnabled: 0.479

HasTpm: 0.368

Census\_IsAlwaysOnAlwaysConnectedCapable: -0.356

IeVerIdentifier: 0.333

Census\_IsPenCapable: -0.246

## Componente 32:

Census\_OSBUILDRevision: 0.571

IeVerIdentifier: 0.508

Census\_IsTouchEnabled: -0.291

Census\_IsAlwaysOnAlwaysConnectedCapable: 0.237

Census OSBuildNumber: 0.185

## Distribución de HasDetections:

## HasDetections

0 0.542

1 0.458

Name: proportion, dtype: float64

#### Distribución de HasDetections:

### HasDetections

0 0542

1 0.458

Name: proportion, dtype: float64

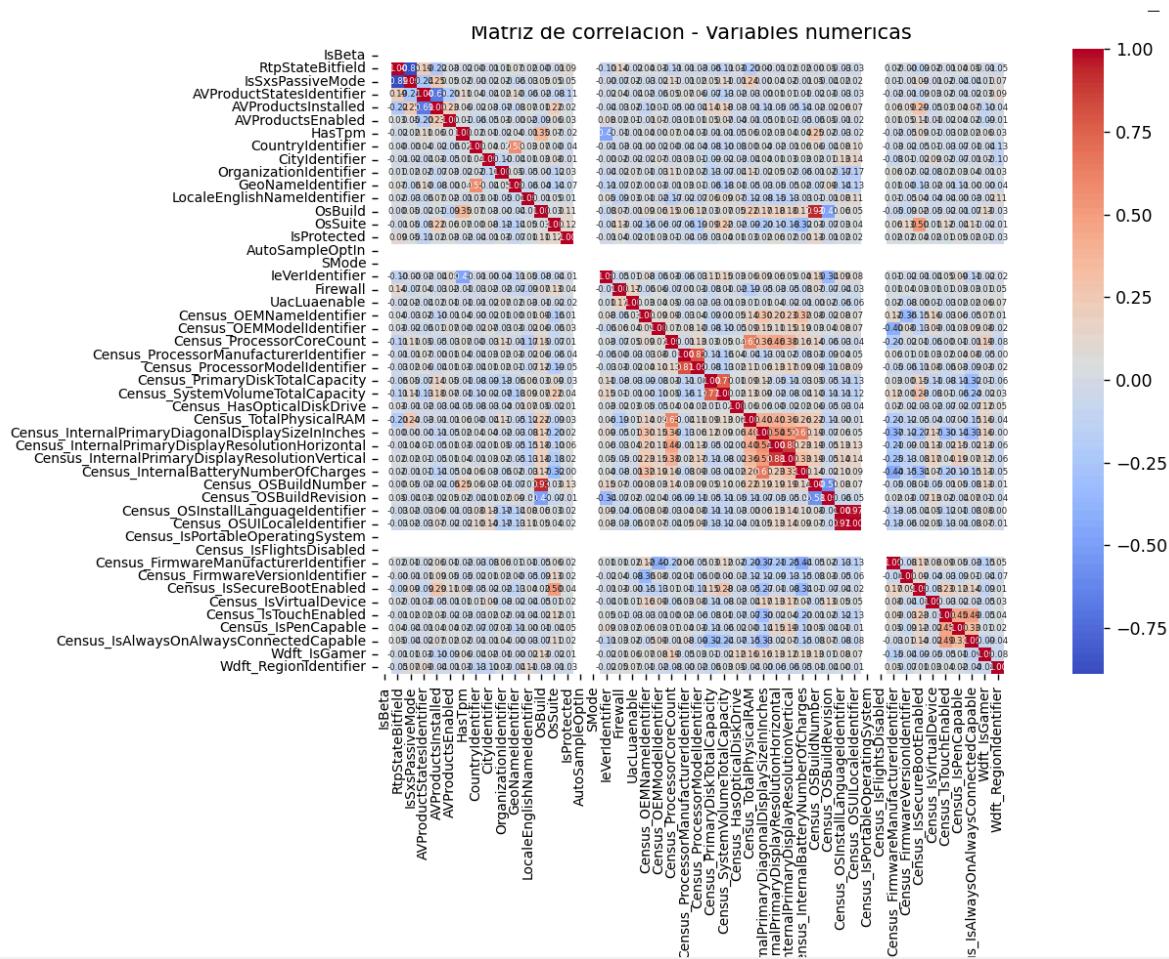
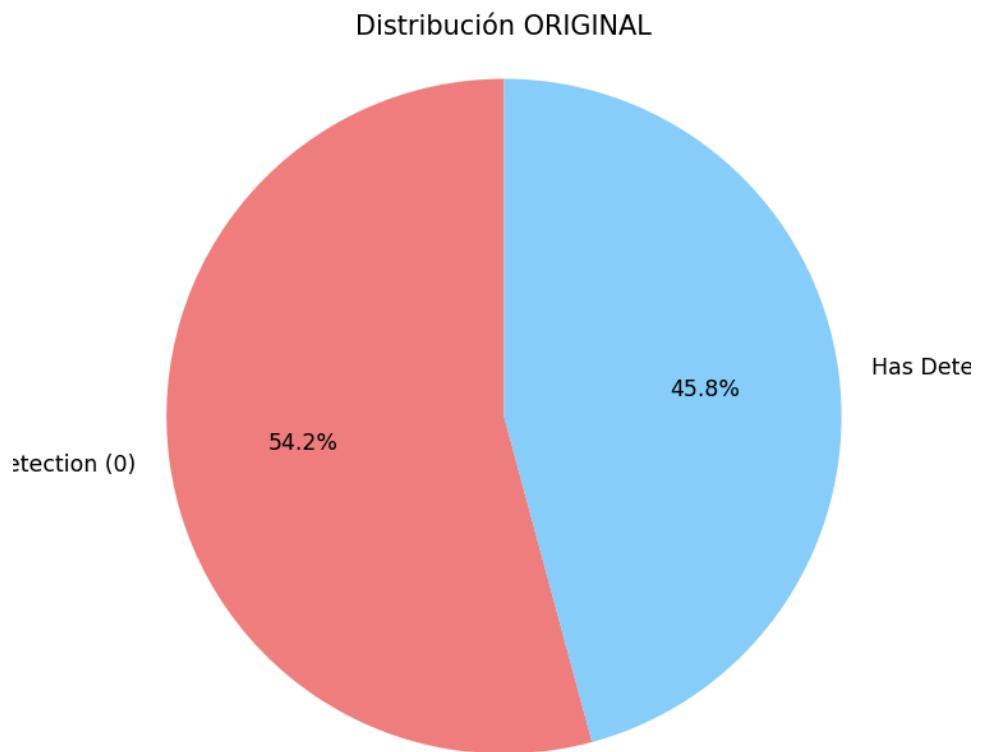


Figure 1

— □ ×



**Muestra 2:**

Los resultados de la muestra 2 fueron:

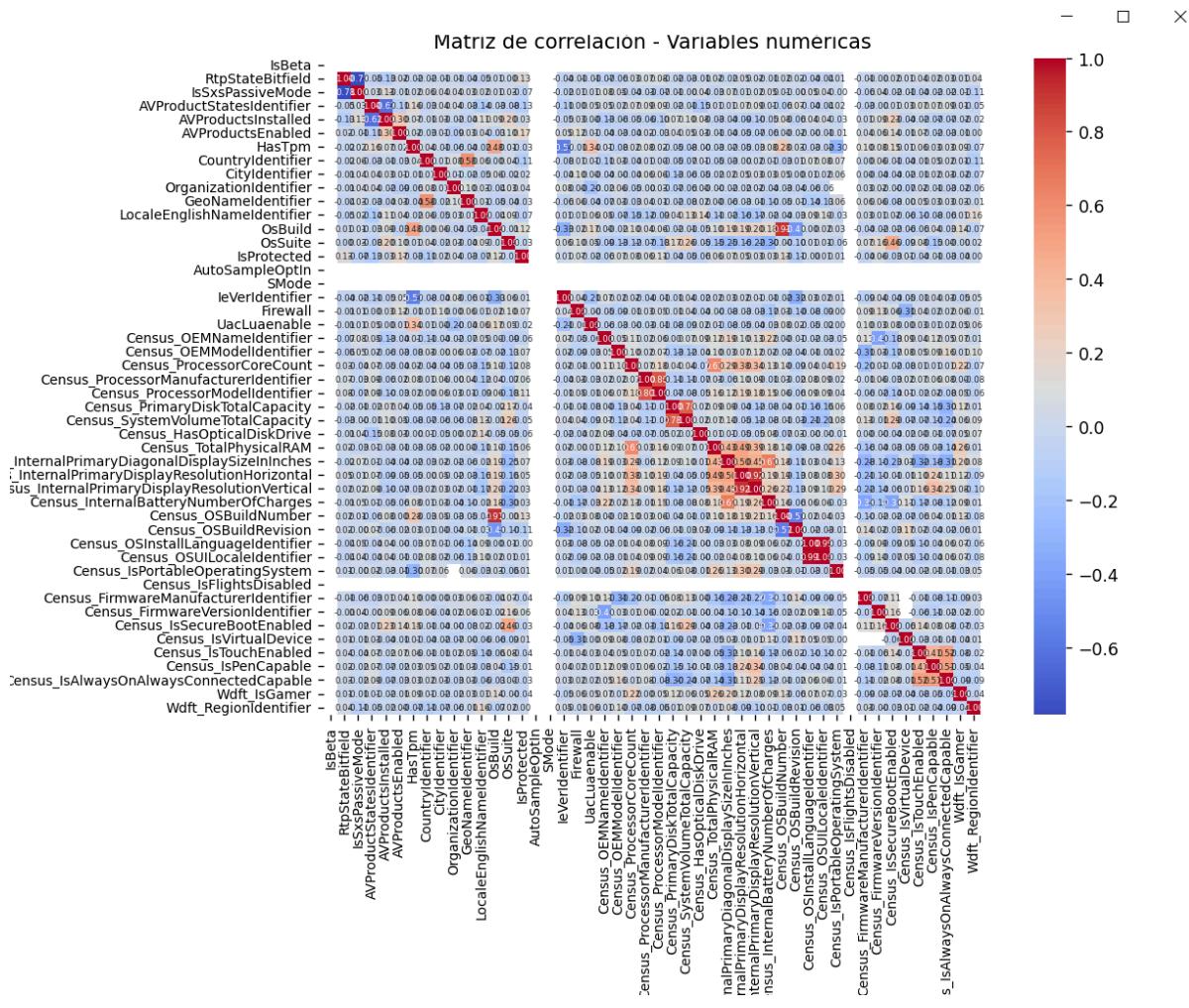
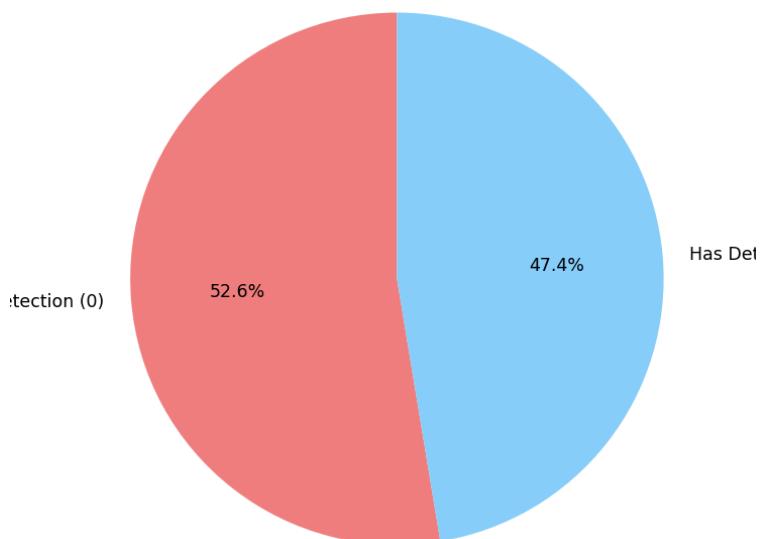


Figure 1

### Distribución ORIGINAL



Se ignoraron 8920983 filas. Guardado en: FilasIgnoradas.txt

Variables altamente correlacionadas ( $>|0.8|$ ):

OsBuild <--> Census\_OSBuildNumber = 0.91  
Census\_ProcessorManufacturerIdentifier <--> Census\_ProcessorModelIdentifier = 0.80  
Census\_InternalPrimaryDisplayResolutionHorizontal <-->  
Census\_InternalPrimaryDisplayResolutionVertical = 0.92  
Census\_OSInstallLanguageIdentifier <--> Census\_OSUILocaleIdentifier  
= 0.99

[ Componentes principales del PCA (varianza explicada  $\geq 95\%$ ) ]

Componente 1:

Census\_InternalPrimaryDisplayResolutionVertical: 0.376  
Census\_InternalPrimaryDisplayResolutionHorizontal: 0.373  
Census\_InternalPrimaryDiagonalDisplaySizeInInches: 0.323  
Census\_TotalPhysicalRAM: 0.288  
Census\_InternalBatteryNumberOfCharges: 0.268

Componente 2:

Census\_IsAlwaysOnAlwaysConnectedCapable: 0.39  
Census\_PrimaryDiskTotalCapacity: -0.357  
Census\_SystemVolumeTotalCapacity: -0.357  
Census\_IsPenCapable: 0.301  
Census\_IsTouchEnabled: 0.295

Componente 3:

OsBuild: 0.45  
Census\_OSBuildNumber: 0.399  
HasTpm: 0.396  
Census\_OSBuildRevision: -0.231  
IeVerIdentifier: -0.221

Componente 4:

Census\_OSInstallLanguageIdentifier: 0.419  
Census\_OSUILocaleIdentifier: 0.417  
Census\_IsTouchEnabled: -0.278  
Census\_SystemVolumeTotalCapacity: -0.256  
Census\_IsSecureBootEnabled: -0.242

Componente 5:

IeVerIdentifier: 0.359  
HasTpm: -0.339  
Census\_OSUILocaleIdentifier: 0.304  
Census\_OSInstallLanguageIdentifier: 0.303  
AVProductsInstalled: 0.267

Componente 6:

RtpStateBitfield: 0.481  
IsSxsPassiveMode: -0.476  
Census\_ProcessorModelIdentifier: 0.373  
Census\_ProcessorManufacturerIdentifier: 0.357  
IsProtected: 0.208

Componente 7:

CountryIdentifier: 0.303

IsSxsPassiveMode: 0.292  
RtpStateBitfield: -0.278  
GeoNameIdentifier: 0.252  
Census\_ProcessorManufacturerIdentifier: 0.241

Componente 8:

GeoNameIdentifier: 0.429  
CountryIdentifier: 0.316  
Census\_OSInstallLanguageIdentifier: -0.282  
Census\_OSUILocaleIdentifier: -0.28  
OrganizationIdentifier: 0.234

Componente 9:

CountryIdentifier: 0.328  
RtpStateBitfield: 0.307  
GeoNameIdentifier: 0.297  
Census\_ProcessorManufacturerIdentifier: -0.295  
IsSxsPassiveMode: -0.285

Componente 10:

AVProductsInstalled: 0.452  
AVProductStatesIdentifier: -0.379  
AVProductsEnabled: 0.345  
Census\_ProcessorManufacturerIdentifier: 0.296  
Census\_ProcessorModelIdentifier: 0.278

Componente 11:

Census\_FirmwareVersionIdentifier: 0.365  
Census\_OEMNameIdentifier: -0.36  
Census\_FirmwareManufacturerIdentifier: -0.313  
Census\_OEMModelIdentifier: 0.3  
CountryIdentifier: -0.283

Componente 12:

Census\_OEMModelIdentifier: 0.358  
CityIdentifier: -0.347  
Census\_FirmwareManufacturerIdentifier: -0.306  
Wdft\_RegionIdentifier: 0.283  
Census\_IsPortableOperatingSystem: -0.268

Componente 13:

Firewall: 0.527  
Census\_IsVirtualDevice: -0.455  
UacLuaenable: 0.242  
Census\_PrimaryDiskTotalCapacity: -0.228  
Census\_SystemVolumeTotalCapacity: -0.211

Componente 14:

Wdft\_RegionIdentifier: 0.509  
Census\_IsPortableOperatingSystem: 0.461  
UacLuaenable: 0.335  
Wdft\_IsGamer: -0.31  
LocaleEnglishNameIdentifier: 0.259

Componente 15:

Census\_FirmwareManufacturerIdentifier: 0.33

CityIdentifier: -0.313  
Census\_InternalBatteryNumberOfCharges: -0.283  
Census\_ProcessorCoreCount: 0.281  
LocaleEnglishNameIdentifier: -0.25

Componente 16:

OrganizationIdentifier: 0.611  
AVProductsEnabled: -0.396  
Census\_HasOpticalDiskDrive: 0.273  
CityIdentifier: -0.255  
Census\_PrimaryDiskTotalCapacity: -0.176

Componente 17:

IsProtected: 0.442  
AVProductsEnabled: 0.386  
OrganizationIdentifier: 0.309  
AVProductStatesIdentifier: 0.274  
CityIdentifier: -0.269

Componente 18:

Census\_OEMModelIdentifier: 0.467  
OrganizationIdentifier: 0.362  
CityIdentifier: 0.342  
leVerIdentifier: -0.315  
IsProtected: 0.258

Componente 19:

OsSuite: 0.452  
CityIdentifier: 0.373  
Census\_PrimaryDiskTotalCapacity: -0.291  
Census\_IsSecureBootEnabled: 0.262  
Census\_IsVirtualDevice: 0.249

Componente 20:

Census\_HasOpticalDiskDrive: 0.621  
Wdft\_IsGamer: 0.34  
CityIdentifier: 0.324  
Census\_IsVirtualDevice: 0.32  
Census\_FirmwareVersionIdentifier: 0.2

Componente 21:

Wdft\_RegionIdentifier: 0.408  
Wdft\_IsGamer: 0.346  
Census\_FirmwareVersionIdentifier: 0.269  
Census\_IsVirtualDevice: 0.266  
Census\_IsPortableOperatingSystem: -0.254

Componente 22:

Wdft\_IsGamer: 0.392  
AVProductsEnabled: 0.352  
Census\_IsPortableOperatingSystem: 0.323  
Census\_HasOpticalDiskDrive: -0.318  
CityIdentifier: -0.291

Componente 23:

UacLuaenable: 0.415

IsProtected: 0.413  
LocaleEnglishNameIdentifier: 0.384  
AVProductsEnabled: -0.325  
Wdft\_RegionIdentifier: -0.31

Componente 24:

LocaleEnglishNameIdentifier: 0.453  
OsSuite: -0.383  
Census\_ProcessorCoreCount: 0.343  
Wdft\_IsGamer: -0.271  
Census\_FirmwareVersionIdentifier: 0.269

Componente 25:

Census\_OEMModelIdentifier: 0.497  
Census\_IsTouchEnabled: -0.327  
AVProductsEnabled: 0.263  
IsProtected: -0.262  
UacLuaenable: 0.254

Componente 26:

Firewall: 0.391  
UacLuaenable: -0.287  
Census\_OEMNameIdentifier: 0.277  
Wdft\_RegionIdentifier: -0.27  
OrganizationIdentifier: -0.259

Componente 27:

UacLuaenable: 0.376  
Census\_FirmwareVersionIdentifier: 0.317  
Firewall: 0.316  
Census\_IsSecureBootEnabled: -0.288  
OrganizationIdentifier: 0.273

Componente 28:

Census\_IsPenCapable: 0.442  
Census\_IsTouchEnabled: -0.354  
Census\_IsSecureBootEnabled: 0.351  
Firewall: 0.31  
Census\_FirmwareManufacturerIdentifier: -0.265

Componente 29:

Census\_IsTouchEnabled: 0.382  
Census\_FirmwareVersionIdentifier: -0.355  
Census\_IsVirtualDevice: 0.355  
Census\_OEMNameIdentifier: -0.331  
Firewall: 0.31

Componente 30:

Census\_IsSecureBootEnabled: 0.364  
Census\_FirmwareManufacturerIdentifier: 0.349  
Census\_IsAlwaysOnAlwaysConnectedCapable: 0.334  
Census\_IsPenCapable: -0.32  
OsSuite: -0.294

Componente 31:

Census\_IsAlwaysOnAlwaysConnectedCapable: 0.477

GeoNameIdentifier: 0.414

CountryIdentifier: -0.392

Census\_IsTouchEnabled: -0.366

Census\_IsPenCapable: -0.221

Componente 32:

Census\_ProcessorCoreCount: 0.386

Census\_OEMNameIdentifier: -0.36

Census\_IsSecureBootEnabled: -0.306

OsSuite: 0.3

Census\_IsTouchEnabled: -0.275

Componente 33:

Census\_TotalPhysicalRAM: 0.659

Census\_ProcessorCoreCount: -0.403

Census\_InternalBatteryNumberOfCharges: 0.237

Census\_IsAlwaysOnAlwaysConnectedCapable: 0.236

OsSuite: 0.212

Distribución de HasDetections:

HasDetections

1 0.526

0 0.474

Name: proportion, dtype: float64

Distribución de HasDetections tras la limpieza:

HasDetections

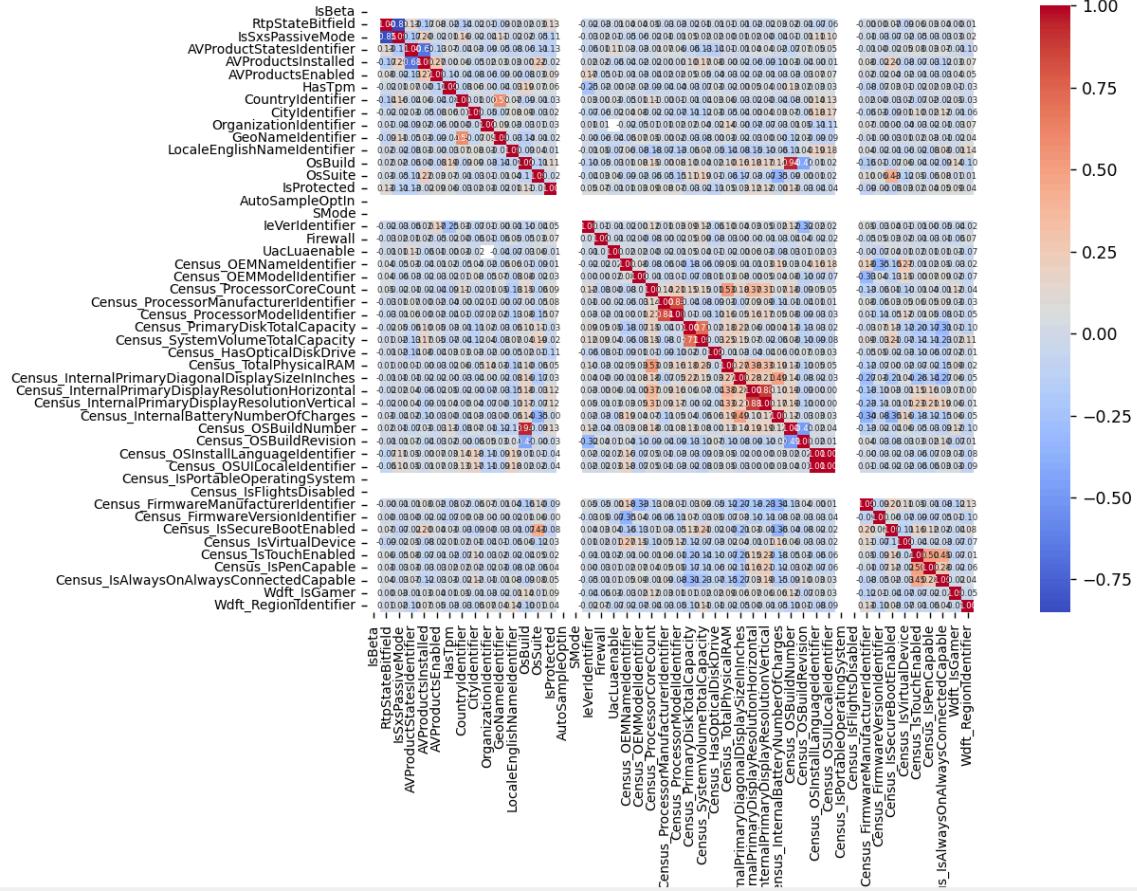
1 0.526

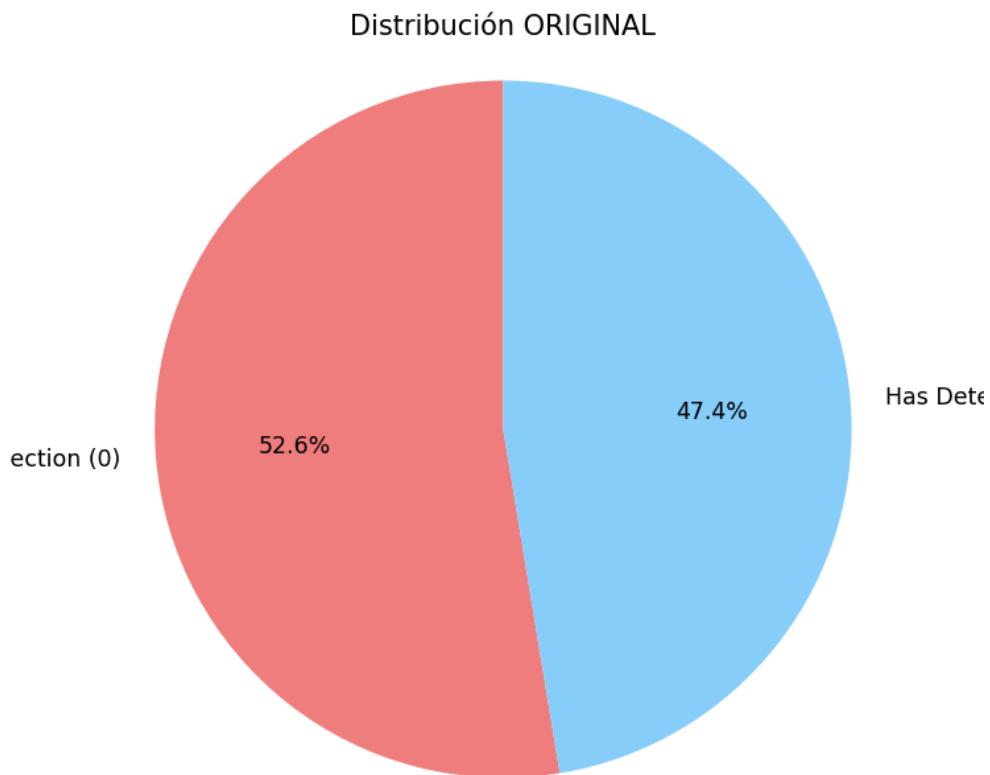
0 0.474

**Muestra 3:**

Los resultados de la muestra 3 fueron los siguientes:

Matriz de correlacion - Variables numericas





Se ignoraron 8920983 filas. Guardado en: FilasIgnoradas.txt

Variables altamente correlacionadas ( $>|0.8|$ ):

RtpStateBitfield <--> IsSxsPassiveMode = 0.85

OsBuild <--> Census\_OSBuildNumber = 0.94

Census\_ProcessorManufacturerIdentifier <--> Census\_ProcessorModelIdentifier = 0.84

Census\_InternalPrimaryDisplayResolutionHorizontal

<-->

Census\_InternalPrimaryDisplayResolutionVertical = 0.88

Census\_OSIInstallLanguageIdentifier <--> Census\_OSUILocaleIdentifier

= 1.00

[ Componentes principales del PCA (varianza explicada  $\geq 95\%$ ) ]

Componente 1:

Census\_OSBuildNumber: 0.342

Census\_InternalPrimaryDisplayResolutionHorizontal: 0.339

OsBuild: 0.33

Census\_InternalPrimaryDisplayResolutionVertical: 0.321

Census\_TotalPhysicalRAM: 0.296

Componente 2:

Census\_SystemVolumeTotalCapacity: 0.369

Census\_PrimaryDiskTotalCapacity: 0.367  
Census\_IsAlwaysOnAlwaysConnectedCapable: -0.324  
AVProductsInstalled: 0.301  
Census\_IsTouchEnabled: -0.3

Componente 3:

Census\_IsSecureBootEnabled: 0.336  
Census\_OSUILocaleIdentifier: -0.31  
Census\_OSInstallLanguageIdentifier: -0.309  
Census\_InternalBatteryNumberOfCharges: -0.287  
Census\_IsTouchEnabled: 0.28

Componente 4:

Census\_OSInstallLanguageIdentifier: 0.414  
Census\_OSUILocaleIdentifier: 0.41  
IsSxsPassiveMode: 0.356  
RtpStateBitfield: -0.333  
CountryIdentifier: 0.259

Componente 5:

GeoNameIdentifier: 0.334  
Census\_OSBuildNumber: -0.296  
OsBuild: -0.286  
Census\_OSUILocaleIdentifier: -0.269  
Census\_OSInstallLanguageIdentifier: -0.265

Componente 6:

IsSxsPassiveMode: 0.312  
RtpStateBitfield: -0.303  
AVProductStatesIdentifier: -0.287  
OsBuild: 0.277  
Census\_OSBuildNumber: 0.251

Componente 7:

Census\_ProcessorModelIdentifier: 0.465  
Census\_ProcessorManufacturerIdentifier: 0.463  
Census\_FirmwareManufacturerIdentifier: 0.277  
Census\_IsVirtualDevice: 0.23  
Census\_InternalPrimaryDisplayResolutionHorizontal: -0.221

Componente 8:

AVProductStatesIdentifier: 0.395  
Census\_OEMNameIdentifier: -0.338  
AVProductsInstalled: -0.316  
AVProductsEnabled: -0.305  
Census\_FirmwareVersionIdentifier: 0.294

Componente 9:

CountryIdentifier: 0.374  
GeoNameIdentifier: 0.341  
RtpStateBitfield: 0.303  
IsSxsPassiveMode: -0.277  
HasTpm: -0.244

Componente 10:

Census\_ProcessorManufacturerIdentifier: 0.292

GeoNameIdentifier: -0.284  
IeVerIdentifier: -0.284  
Census\_OSBuildRevision: 0.273  
CountryIdentifier: -0.272

Componente 11:  
IeVerIdentifier: 0.432  
HasTpm: -0.39  
GeoNameIdentifier: -0.283  
AVProductsEnabled: 0.242  
CountryIdentifier: -0.224

Componente 12:  
LocaleEnglishNameIdentifier: 0.436  
Wdft\_RegionIdentifier: 0.39  
Firewall: 0.353  
Census\_HasOpticalDiskDrive: -0.302  
UacLuaenable: -0.28

Componente 13:  
Census\_OEMModelIdentifier: 0.45  
Firewall: 0.419  
Census\_IsVirtualDevice: 0.309  
OsSuite: 0.281  
UacLuaenable: 0.225

Componente 14:  
Wdft\_IsGamer: 0.39  
UacLuaenable: 0.344  
CityIdentifier: -0.338  
AVProductsEnabled: 0.313  
OrganizationIdentifier: -0.279

Componente 15:  
OrganizationIdentifier: 0.415  
Wdft\_IsGamer: 0.387  
UacLuaenable: 0.292  
Wdft\_RegionIdentifier: -0.235  
Census\_IsSecureBootEnabled: -0.213

Componente 16:  
OrganizationIdentifier: 0.362  
Census\_FirmwareVersionIdentifier: 0.354  
Wdft\_RegionIdentifier: -0.35  
IsProtected: 0.346  
LocaleEnglishNameIdentifier: -0.283

Componente 17:  
Census\_IsPenCapable: 0.414  
Census\_PrimaryDiskTotalCapacity: 0.315  
Wdft\_RegionIdentifier: -0.285  
OrganizationIdentifier: -0.267  
OsSuite: -0.255

Componente 18:  
UacLuaenable: 0.559

Wdft\_IsGamer: -0.485  
Census\_HasOpticalDiskDrive: 0.303  
OrganizationIdentifier: 0.237  
Census\_TotalPhysicalRAM: -0.204

Componente 19:

Firewall: 0.551  
Census\_HasOpticalDiskDrive: 0.345  
CityIdentifier: -0.261  
Census\_ProcessorCoreCount: 0.256  
AVProductsEnabled: -0.252

Componente 20:

CityIdentifier: 0.478  
OrganizationIdentifier: -0.315  
Census\_FirmwareManufacturerIdentifier: 0.297  
Wdft\_IsGamer: 0.29  
Census\_IsVirtualDevice: 0.275

Componente 21:

OrganizationIdentifier: 0.351  
IsProtected: -0.349  
Wdft\_IsGamer: 0.344  
LocaleEnglishNameIdentifier: 0.323  
CityIdentifier: -0.299

Componente 22:

Census\_IsVirtualDevice: 0.411  
Census\_HasOpticalDiskDrive: 0.36  
Wdft\_RegionIdentifier: 0.322  
AVProductsEnabled: 0.285  
Census\_SystemVolumeTotalCapacity: 0.243

Componente 23:

IeVerIdentifier: 0.481  
Census\_HasOpticalDiskDrive: 0.406  
HasTpm: 0.357  
IsProtected: 0.288  
Census\_FirmwareVersionIdentifier: -0.263

Componente 24:

CityIdentifier: 0.489  
Firewall: 0.357  
LocaleEnglishNameIdentifier: -0.29  
Wdft\_IsGamer: 0.278  
AVProductsEnabled: 0.256

Componente 25:

Census\_IsPenCapable: 0.469  
Census\_IsAlwaysOnAlwaysConnectedCapable: -0.37  
AVProductsEnabled: -0.349  
Census\_InternalPrimaryDiagonalDisplaySizeInInches: 0.23  
Census\_OEMModelIdentifier: 0.206

Componente 26:

Census\_InternalBatteryNumberOfCharges: 0.457

Census\_OEMModelIdentifier: -0.323  
Census\_IsVirtualDevice: 0.303  
UacLuaenable: 0.267  
Census\_IsAlwaysOnAlwaysConnectedCapable: 0.262

Componente 27:

LocaleEnglishNameIdentifier: 0.489  
Census\_TotalPhysicalRAM: 0.336  
HasTpm: 0.3  
Census\_InternalPrimaryDiagonalDisplaySizeInInches: -0.287  
OsSuite: -0.238

Componente 28:

Census\_IsAlwaysOnAlwaysConnectedCapable: 0.44  
HasTpm: 0.396  
Census\_IsSecureBootEnabled: -0.353  
Census\_FirmwareManufacturerIdentifier: 0.331  
Census\_FirmwareVersionIdentifier: 0.298

Componente 29:

Census\_InternalPrimaryDiagonalDisplaySizeInInches: 0.609  
AVProductsEnabled: 0.27  
AVProductStatesIdentifier: 0.225  
Census\_InternalPrimaryDisplayResolutionVertical: -0.221  
OsSuite: 0.22

Componente 30:

Census\_OEMNameIdentifier: 0.495  
Census\_FirmwareVersionIdentifier: 0.412  
Census\_IsVirtualDevice: -0.34  
Census\_InternalBatteryNumberOfCharges: 0.275  
Census\_InternalPrimaryDiagonalDisplaySizeInInches: -0.272

Componente 31:

Census\_IsTouchEnabled: 0.42  
Census\_OSBuildRevision: -0.411  
Census\_IsSecureBootEnabled: -0.41  
Census\_OEMNameIdentifier: -0.29  
Census\_TotalPhysicalRAM: 0.289

Componente 32:

OsSuite: 0.547  
Census\_IsSecureBootEnabled: -0.496  
Census\_IsPenCapable: 0.407  
Census\_IsAlwaysOnAlwaysConnectedCapable: -0.181  
Wdft\_RegionIdentifier: 0.175

Componente 33:

Census\_IsTouchEnabled: 0.486  
Census\_OSBuildRevision: 0.479  
IeVerIdentifier: 0.319  
Census\_IsPenCapable: -0.287  
Census\_IsAlwaysOnAlwaysConnectedCapable: -0.263

Distribución de HasDetections:

## HasDetections

1 0.526

0 0.474

Name: proportion, dtype: float64

## Distribución de HasDetections:

### HasDetections

1 0.526

0 0.474

Name: proportion, dtype: float64

## Muestra 4:

Los resultados para la muestra 4 fueron los siguientes:

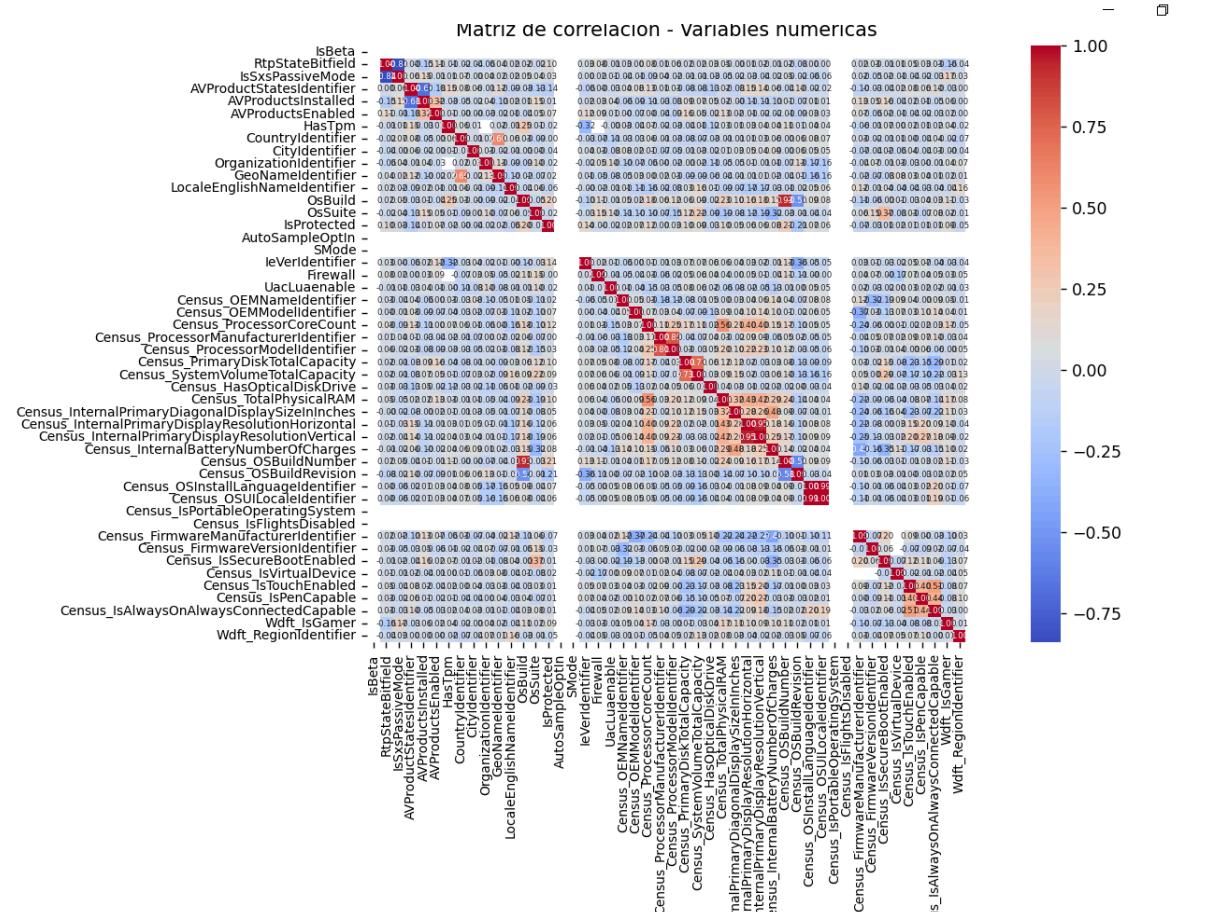
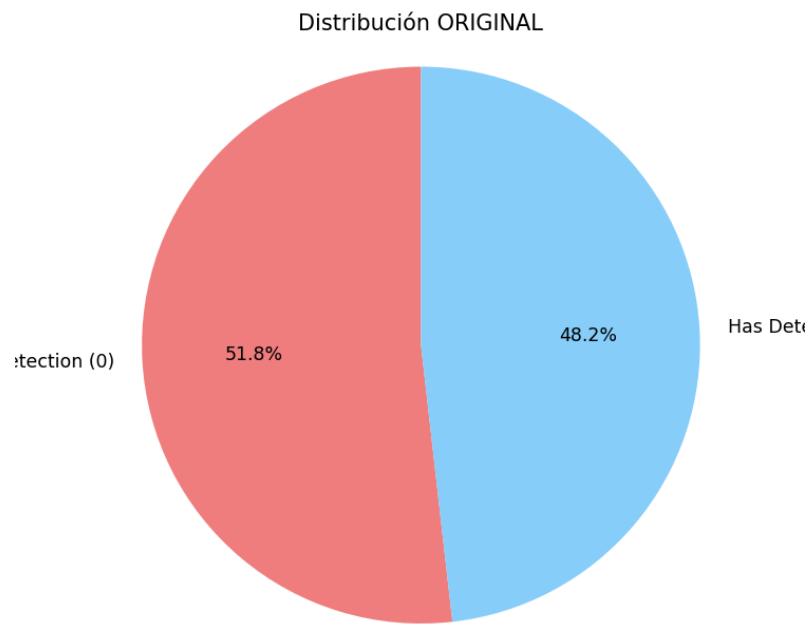


Figure 1

— □ ×



Se ignoraron 8920983 filas. Guardado en: FilasIgnoradas.txt

Variables altamente correlacionadas ( $>|0.8|$ ):

RtpStateBitfield <--> IsSxsPassiveMode = 0.84

OsBuild <--> Census\_OSBuildNumber = 0.93

Census\_ProcessorManufacturerIdentifier <--> Census\_ProcessorModelIdentifier = 0.80

Census\_InternalPrimaryDisplayResolutionHorizontal <-->

Census\_InternalPrimaryDisplayResolutionVertical = 0.95

Census\_OSIInstallLanguageIdentifier <--> Census\_OSUILocaleIdentifier

= 0.99

[ Componentes principales del PCA (varianza explicada  $\geq 95\%$ ) ]

Componente 1:

Census\_InternalPrimaryDisplayResolutionVertical: 0.371

Census\_InternalPrimaryDisplayResolutionHorizontal: 0.356

Census\_TotalPhysicalRAM: 0.326

Census\_ProcessorCoreCount: 0.307

OsBuild: 0.267

Componente 2:

Census\_PrimaryDiskTotalCapacity: 0.389

Census\_SystemVolumeTotalCapacity: 0.386

Census\_IsAlwaysOnAlwaysConnectedCapable: -0.321

Census\_IsTouchEnabled: -0.268

AVProductStatesIdentifier: -0.241

Componente 3:

Census\_OSIInstallLanguageIdentifier: 0.329

Census\_OSUILocaleIdentifier: 0.322  
Census\_IsAlwaysOnAlwaysConnectedCapable: 0.296  
Census\_OSBuildNumber: 0.268  
Census\_OSBuildRevision: -0.26

Componente 4:

Census\_OSUILocaleIdentifier: 0.353  
Census\_OSIInstallLanguageIdentifier: 0.35  
Census\_ProcessorManufacturerIdentifier: -0.332  
Census\_IsSecureBootEnabled: -0.307  
Census\_ProcessorModelIdentifier: -0.304

Componente 5:

IsSxsPassiveMode: 0.602  
RtpStateBitfield: -0.57  
OsBuild: 0.212  
Wdft\_IsGamer: 0.204  
Census\_OSBuildNumber: 0.192

Componente 6:

OsBuild: 0.328  
Census\_OSBuildNumber: 0.327  
GeoNameIdentifier: 0.292  
Census\_OSBuildRevision: -0.292  
AVProductsInstalled: -0.278

Componente 7:

Census\_ProcessorManufacturerIdentifier: 0.481  
Census\_ProcessorModelIdentifier: 0.454  
AVProductStatesIdentifier: -0.257  
Census\_SystemVolumeTotalCapacity: -0.243  
Census\_OEMNameIdentifier: -0.208

Componente 8:

leVerIdentifier: 0.342  
HasTpm: -0.32  
AVProductStatesIdentifier: -0.289  
AVProductsInstalled: 0.257  
Census\_FirmwareVersionIdentifier: -0.234

Componente 9:

CountryIdentifier: 0.577  
GeoNameIdentifier: 0.5  
AVProductsEnabled: 0.265  
AVProductsInstalled: 0.225  
Census\_OEMNameIdentifier: -0.203

Componente 10:

Census\_FirmwareManufacturerIdentifier: 0.385  
Census\_FirmwareVersionIdentifier: -0.373  
Census\_OEMModelIdentifier: -0.331  
LocaleEnglishNameIdentifier: 0.274  
Census\_OEMNameIdentifier: 0.256

Componente 11:

leVerIdentifier: 0.516

HasTpm: -0.426  
AVProductsInstalled: -0.297  
AVProductStatesIdentifier: 0.262  
IsSxsPassiveMode: 0.203  
Componente 12:  
Wdft\_RegionIdentifier: 0.37  
Census\_OEMModelIdentifier: 0.359  
Census\_IsVirtualDevice: 0.329  
Firewall: -0.323  
Census\_HasOpticalDiskDrive: -0.291

Componente 13:  
Firewall: 0.419  
UacLuaenable: -0.414  
Wdft\_RegionIdentifier: 0.391  
LocaleEnglishNameIdentifier: 0.341  
Census\_IsVirtualDevice: -0.317

Componente 14:  
UacLuaenable: 0.56  
OrganizationIdentifier: 0.448  
CityIdentifier: -0.316  
Wdft\_RegionIdentifier: 0.289  
Firewall: 0.256

Componente 15:  
Wdft\_IsGamer: 0.395  
IsProtected: 0.33  
Census\_FirmwareVersionIdentifier: -0.287  
Census\_OEMNameIdentifier: 0.26  
Census\_InternalPrimaryDisplayResolutionHorizontal: -0.258

Componente 16:  
CityIdentifier: 0.647  
OrganizationIdentifier: 0.342  
Wdft\_IsGamer: -0.238  
Census\_ProcessorCoreCount: -0.226  
Census\_TotalPhysicalRAM: -0.21

Componente 17:  
AVProductsEnabled: 0.546  
Census\_HasOpticalDiskDrive: -0.534  
AVProductStatesIdentifier: 0.243  
Census\_IsAlwaysOnAlwaysConnectedCapable: -0.188  
Census\_TotalPhysicalRAM: 0.174

Componente 18:  
Census\_IsVirtualDevice: 0.479  
OrganizationIdentifier: 0.334  
Wdft\_IsGamer: 0.319  
Wdft\_RegionIdentifier: 0.275  
UacLuaenable: -0.257

Componente 19:  
Census\_IsVirtualDevice: 0.435

LocaleEnglishNameIdentifier: -0.4  
Firewall: 0.379  
Census\_HasOpticalDiskDrive: 0.313  
Wdft\_IsGamer: -0.256

Componente 20:  
IsProtected: 0.537  
LocaleEnglishNameIdentifier: -0.35  
Census\_OEMModelIdentifier: -0.324  
OsSuite: -0.274  
Census\_InternalBatteryNumberOfCharges: 0.218

Componente 21:  
Firewall: 0.383  
CityIdentifier: -0.362  
Census\_InternalPrimaryDiagonalDisplaySizeInInches: 0.335  
OsSuite: 0.322  
Census\_TotalPhysicalRAM: -0.295

Componente 22:  
LocaleEnglishNameIdentifier: 0.351  
AVProductsEnabled: 0.338  
HasTpm: 0.305  
Census\_FirmwareVersionIdentifier: 0.267  
Census\_IsVirtualDevice: 0.253

Componente 23:  
OrganizationIdentifier: 0.394  
UacLuaenable: -0.326  
Wdft\_RegionIdentifier: -0.322  
Census\_InternalPrimaryDiagonalDisplaySizeInInches: -0.291  
LocaleEnglishNameIdentifier: 0.255

Componente 24:  
IsProtected: 0.407  
Census\_InternalPrimaryDisplayResolutionHorizontal: 0.292  
Census\_InternalPrimaryDiagonalDisplaySizeInInches: -0.255  
OsSuite: -0.25  
Census\_InternalPrimaryDisplayResolutionVertical: 0.248

Componente 25:  
Wdft\_IsGamer: 0.441  
Census\_ProcessorCoreCount: -0.434  
Census\_TotalPhysicalRAM: -0.356  
UacLuaenable: -0.278  
Census\_PrimaryDiskTotalCapacity: 0.208

Componente 26:  
Census\_FirmwareVersionIdentifier: 0.399  
Census\_HasOpticalDiskDrive: -0.39  
Census\_FirmwareManufacturerIdentifier: 0.314  
AVProductsEnabled: -0.276  
Census\_IsSecureBootEnabled: -0.253

Componente 27:  
Census\_OEMNameIdentifier: 0.398

HasTpm: 0.37  
Census\_IsVirtualDevice: -0.362  
Firewall: -0.27  
Census\_FirmwareVersionIdentifier: 0.268

Componente 28:  
HasTpm: 0.465  
Census\_OEMModelIdentifier: 0.302  
Firewall: 0.283  
leVerIdentifier: 0.281  
AVProductsEnabled: -0.269

Componente 29:  
Census\_IsTouchEnabled: 0.466  
Census\_IsPenCapable: -0.403  
Census\_FirmwareManufacturerIdentifier: -0.277  
Census\_InternalPrimaryDiagonalDisplaySizeInInches: -0.263  
Census\_FirmwareVersionIdentifier: -0.241

Componente 30:  
Census\_IsTouchEnabled: 0.473  
Census\_IsPenCapable: -0.397  
Census\_InternalPrimaryDiagonalDisplaySizeInInches: 0.344  
Census\_OEMModelIdentifier: 0.327  
Census\_FirmwareManufacturerIdentifier: 0.308

Componente 31:  
Census\_IsSecureBootEnabled: 0.418  
Census\_InternalBatteryNumberOfCharges: 0.412  
Census\_InternalPrimaryDiagonalDisplaySizeInInches: -0.363  
Census\_OEMNameIdentifier: 0.342  
Census\_FirmwareVersionIdentifier: 0.227

Componente 32:  
Census\_IsAlwaysOnAlwaysConnectedCapable: 0.431  
OsSuite: -0.408  
Census\_IsTouchEnabled: -0.347  
Census\_TotalPhysicalRAM: -0.338  
Census\_IsSecureBootEnabled: 0.282

Distribución de HasDetections:

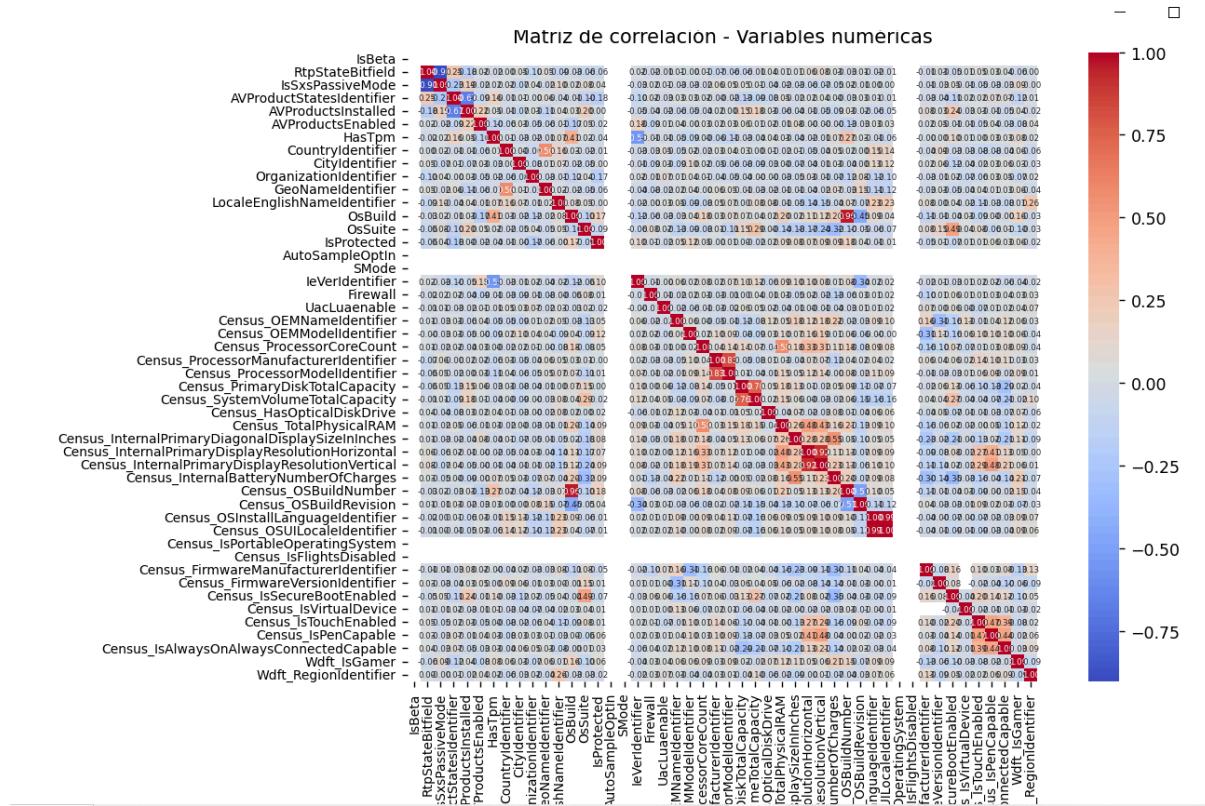
HasDetections  
0 0.518  
1 0.482  
Name: proportion, dtype: float64

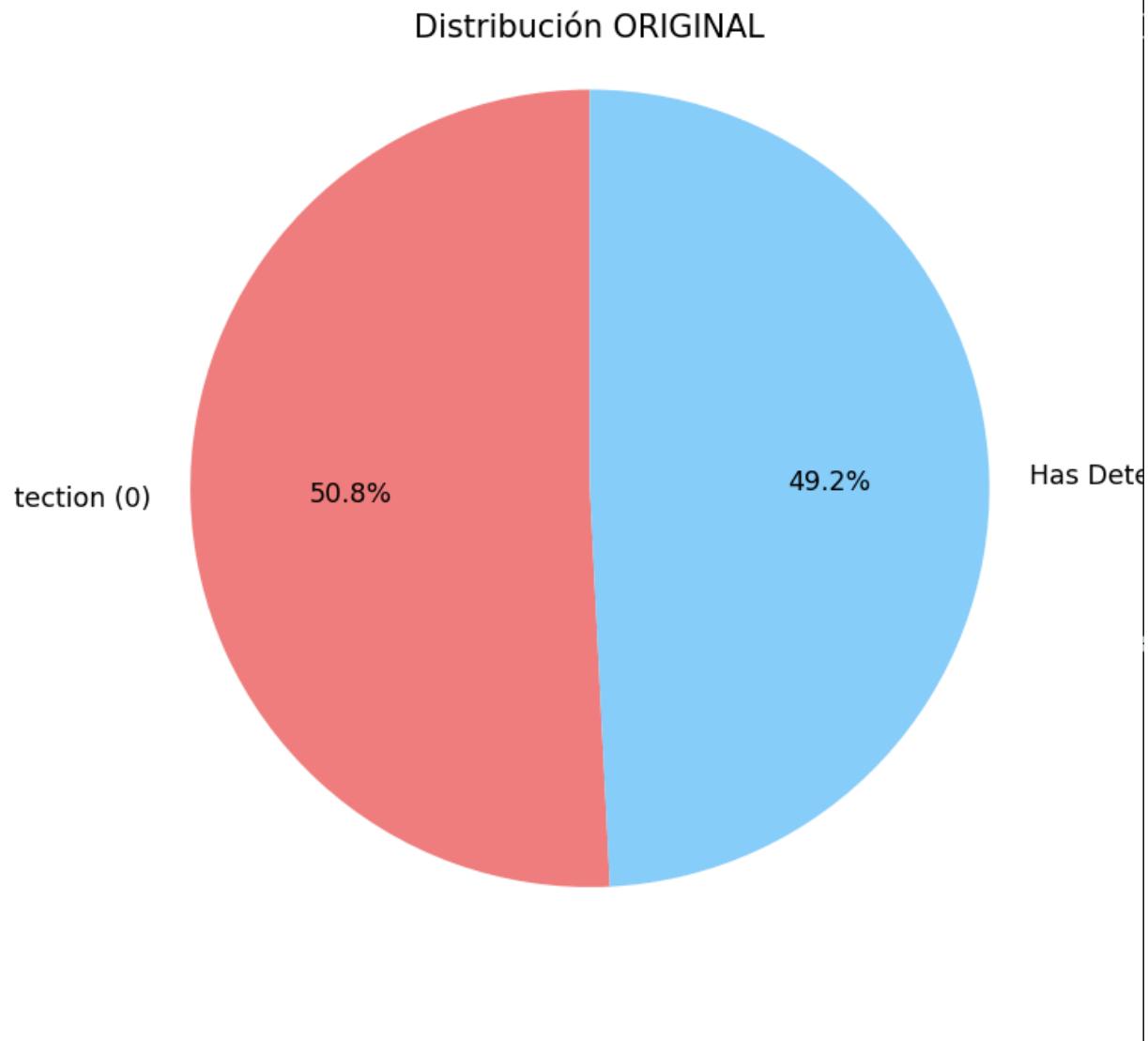
Distribución de HasDetections:

HasDetections  
0 0.518  
1 0.482

**Muestra 5:**

Los resultados para la muestra 5 fueron los siguientes:





Se ignoraron 8920983 filas. Guardado en: FilasIgnoradas.txt

Variables altamente correlacionadas ( $>|0.8|$ ):

RtpStateBitfield <--> IsSxsPassiveMode = 0.90

OsBuild <--> Census\_OSBuildNumber = 0.96

Census\_ProcessorManufacturerIdentifier <--> Census\_ProcessorModelIdentifier = 0.83

Census\_InternalPrimaryDisplayResolutionHorizontal

<-->

Census\_InternalPrimaryDisplayResolutionVertical = 0.92

Census\_OSIInstallLanguageIdentifier <--> Census\_OSUILocaleIdentifier

= 0.99

[ Componentes principales del PCA (varianza explicada  $\geq 95\%$ ) ]

Componente 1:

Census\_InternalPrimaryDisplayResolutionVertical: 0.381

Census\_InternalPrimaryDisplayResolutionHorizontal: 0.358

Census\_TotalPhysicalRAM: 0.314

Census\_OSBUILDNUMBER: 0.265

Census\_ProcessorCoreCount: 0.258

Componente 2:

Census\_IsAlwaysOnAlwaysConnectedCapable: 0.315

Census\_IsTouchEnabled: 0.307

Census\_IsPenCapable: 0.304

Census\_PrimaryDiskTotalCapacity: -0.293

Census\_OSBUILDNUMBER: -0.282

Componente 3:

Census\_IsSecureBootEnabled: 0.382

AVProductsInstalled: 0.289

Census\_SystemVolumeTotalCapacity: 0.285

OsSuite: 0.254

Census\_IsTouchEnabled: 0.249

Componente 4:

Census\_OSDisplayLanguageIdentifier: 0.344

Census\_OSUILocaleIdentifier: 0.319

Census\_ProcessorManufacturerIdentifier: 0.26

Census\_PrimaryDiskTotalCapacity: -0.259

Census\_InternalPrimaryDiagonalDisplaySizeInInches: -0.255

Componente 5:

HasTpm: 0.423

OsBuild: 0.389

Census\_OSBUILDNUMBER: 0.321

IeVerIdentifier: -0.285

Census\_OSUILocaleIdentifier: -0.275

Componente 6:

RtpStateBitfield: -0.392

IsSxsPassiveMode: 0.392

Census\_OSUILocaleIdentifier: -0.32

Census\_OSDisplayLanguageIdentifier: -0.315

Census\_SystemVolumeTotalCapacity: -0.2

Componente 7:

Census\_ProcessorManufacturerIdentifier: 0.472

Census\_ProcessorModelIdentifier: 0.462

GeoNameIdentifier: 0.375

CountryIdentifier: 0.317

Census\_OEMNameIdentifier: -0.238

Componente 8:

IeVerIdentifier: 0.388

Census\_OSBUILDRevision: -0.339

CountryIdentifier: -0.317

GeoNameIdentifier: -0.282

HasTpm: -0.261

Componente 9:

Census\_FirmwareVersionIdentifier: 0.427

Census\_FirmwareManufacturerIdentifier: -0.338

Census\_OEMModelIdentifier: 0.32  
Census\_OEMNameIdentifier: -0.303  
Wdft\_RegionIdentifier: -0.271

Componente 10:

GeoNameIdentifier: 0.335  
CountryIdentifier: 0.317  
AVProductsInstalled: 0.286  
AVProductStatesIdentifier: -0.284  
Firewall: -0.277

Componente 11:

IeVerIdentifier: 0.361  
GeoNameIdentifier: 0.321  
CountryIdentifier: 0.311  
AVProductsInstalled: -0.268  
HasTpm: -0.265

Componente 12:

Firewall: 0.416  
CityIdentifier: -0.372  
Census\_HasOpticalDiskDrive: 0.368  
Census\_IsVirtualDevice: 0.325  
Census\_OEMNameIdentifier: 0.307

Componente 13:

Wdft\_RegionIdentifier: 0.481  
LocaleEnglishNameIdentifier: 0.336  
Census\_OEMModelIdentifier: 0.257  
Census\_ProcessorCoreCount: -0.228  
Census\_IsPenCapable: 0.223

Componente 14:

IsProtected: 0.58  
OrganizationIdentifier: -0.534  
Census\_OSBuildRevision: 0.29  
Firewall: 0.241  
AVProductsEnabled: -0.191

Componente 15:

OrganizationIdentifier: 0.405  
Census\_IsVirtualDevice: -0.387  
Firewall: 0.323  
Wdft\_IsGamer: 0.301  
AVProductStatesIdentifier: -0.254

Componente 16:

UacLuaenable: 0.593  
Wdft\_RegionIdentifier: -0.344  
Wdft\_IsGamer: 0.313  
CityIdentifier: 0.268  
AVProductsEnabled: -0.221

Componente 17:

Census\_IsVirtualDevice: 0.597  
UacLuaenable: 0.454

Census\_OEMModelIdentifier: 0.284

Census\_IsTouchEnabled: -0.242

AVProductsEnabled: 0.201

Componente 18:

CityIdentifier: 0.504

UacLuaenable: -0.392

Census\_OEMModelIdentifier: 0.292

Census\_FirmwareVersionIdentifier: -0.245

Census\_OEMNameIdentifier: 0.241

Componente 19:

Wdft\_IsGamer: 0.405

Census\_HasOpticalDiskDrive: -0.357

IsProtected: -0.336

OrganizationIdentifier: -0.331

AVProductsEnabled: 0.298

Componente 20:

Census\_HasOpticalDiskDrive: 0.45

AVProductsEnabled: 0.446

Census\_ProcessorCoreCount: 0.333

Census\_IsVirtualDevice: -0.224

IsProtected: 0.207

Componente 21:

IsProtected: 0.342

Census\_IsAlwaysOnAlwaysConnectedCapable: 0.282

LocaleEnglishNameIdentifier: -0.25

Firewall: -0.25

Census\_FirmwareVersionIdentifier: -0.246

Componente 22:

OsSuite: 0.377

LocaleEnglishNameIdentifier: 0.363

Census\_IsSecureBootEnabled: 0.307

Census\_InternalPrimaryDiagonalDisplaySizeInInches: 0.306

AVProductsEnabled: -0.292

Componente 23:

Firewall: 0.468

CityIdentifier: 0.36

Wdft\_RegionIdentifier: -0.325

OrganizationIdentifier: 0.322

Wdft\_IsGamer: -0.298

Componente 24:

Census\_HasOpticalDiskDrive: 0.359

Census\_IsVirtualDevice: 0.355

Census\_FirmwareManufacturerIdentifier: -0.342

Census\_OEMNameIdentifier: -0.339

Census\_OEMModelIdentifier: -0.272

Componente 25:

Wdft\_IsGamer: 0.507

OrganizationIdentifier: 0.334

Census\_IsVirtualDevice: 0.265  
Census\_FirmwareVersionIdentifier: 0.247  
AVProductsEnabled: 0.229

Componente 26:

Census\_FirmwareVersionIdentifier: 0.474  
LocaleEnglishNameIdentifier: -0.368  
Wdft\_RegionIdentifier: 0.348  
Census\_IsAlwaysOnAlwaysConnectedCapable: 0.305  
Census\_FirmwareManufacturerIdentifier: 0.26

Componente 27:

LocaleEnglishNameIdentifier: 0.443  
Wdft\_RegionIdentifier: -0.36  
CityIdentifier: -0.261  
Census\_HasOpticalDiskDrive: -0.237  
OsSuite: -0.236

Componente 28:

Census\_IsTouchEnabled: 0.686  
Census\_IsAlwaysOnAlwaysConnectedCapable: -0.415  
Census\_TotalPhysicalRAM: 0.177  
Census\_InternalPrimaryDisplayResolutionVertical: -0.172  
Census\_InternalPrimaryDiagonalDisplaySizeInInches: 0.151

Componente 29:

Census\_OEMNameIdentifier: 0.384  
Census\_InternalPrimaryDiagonalDisplaySizeInInches: -0.37  
Census\_OSBuildRevision: 0.369  
CountryIdentifier: -0.313  
Census\_FirmwareVersionIdentifier: 0.272

Componente 30:

Census\_IsSecureBootEnabled: 0.649  
OsSuite: -0.54  
HasTpm: 0.206  
IeVerIdentifier: 0.2  
Census\_InternalPrimaryDisplayResolutionHorizontal: -0.198

Componente 31:

Census\_IsPenCapable: 0.556  
GeoNameIdentifier: 0.305  
CountryIdentifier: -0.287  
HasTpm: -0.248  
Census\_IsTouchEnabled: -0.239

Componente 32:

Census\_OSBuildRevision: 0.365  
Census\_IsPenCapable: -0.347  
Census\_FirmwareManufacturerIdentifier: 0.332  
Census\_TotalPhysicalRAM: -0.313  
Census\_OEMNameIdentifier: -0.27

Distribución de HasDetections:

HasDetections

```
0 0.508  
1 0.492  
Name: proportion, dtype: float64
```

Distribución de HasDetections:

```
HasDetections  
0 0.508  
1 0.492  
Name: proportion, dtype: float64
```

## Observaciones:

- Las siguientes variables mostraron correlaciones altas ( $>|0.8|$ ) en cada una de las muestras analizadas:

### Muestra 1

Variables altamente correlacionadas ( $>|0.8|$ ):  
~~RtpStateBitfield~~ <-> ~~IsSxsPassiveMode~~ = 0.89  
OsBuild <-> Census\_OSBuildNumber = 0.93  
Census\_ProcessorManufacturerIdentifier <-> Census\_ProcessorModelIdentifier = 0.81  
Census\_InternalPrimaryDisplayResolutionHorizontal <->  
Census\_InternalPrimaryDisplayResolutionVertical = 0.88  
Census\_OSInstallLanguageIdentifier <-> Census\_OSUILocaleIdentifier  
= 0.97

### Muestra 2

Variables altamente correlacionadas ( $>|0.8|$ ):  
OsBuild <-> Census\_OSBuildNumber = 0.91  
Census\_ProcessorManufacturerIdentifier <-> Census\_ProcessorModelIdentifier = 0.80  
Census\_InternalPrimaryDisplayResolutionHorizontal <->  
Census\_InternalPrimaryDisplayResolutionVertical = 0.92  
Census\_OSInstallLanguageIdentifier <-> Census\_OSUILocaleIdentifier  
= 0.99

### Muestra 3

Variables altamente correlacionadas ( $>|0.8|$ ):  
~~RtpStateBitfield~~ <-> ~~IsSxsPassiveMode~~ = 0.85  
OsBuild <-> Census\_OSBuildNumber = 0.94  
Census\_ProcessorManufacturerIdentifier <-> Census\_ProcessorModelIdentifier = 0.84  
Census\_InternalPrimaryDisplayResolutionHorizontal <->  
Census\_InternalPrimaryDisplayResolutionVertical = 0.88  
Census\_OSInstallLanguageIdentifier <-> Census\_OSUILocaleIdentifier  
= 1.00

### Muestra 4

Variables altamente correlacionadas ( $>|0.8|$ ):

```
RtpStateBitfield <-> IsSxsPassiveMode = 0.84
OsBuild <-> Census_OSBuildNumber = 0.93
Census_ProcessorManufacturerIdentifier <-> Census_ProcessorModelIdentifier = 0.80
Census_InternalPrimaryDisplayResolutionHorizontal <->
Census_InternalPrimaryDisplayResolutionVertical = 0.95
Census_OSIInstallLanguageIdentifier <-> Census_OSUILocaleIdentifier
= 0.99
```

### Muestra 5

Variables altamente correlacionadas ( $>|0.8|$ ):

```
RtpStateBitfield <-> IsSxsPassiveMode = 0.90
OsBuild <-> Census_OSBuildNumber = 0.96
Census_ProcessorManufacturerIdentifier <-> Census_ProcessorModelIdentifier = 0.83
Census_InternalPrimaryDisplayResolutionHorizontal <->
Census_InternalPrimaryDisplayResolutionVertical = 0.92
Census_OSIInstallLanguageIdentifier <-> Census_OSUILocaleIdentifier
= 0.99
```

En base a estos resultados, se puede concluir que ciertos pares de variables presentan una alta correlación de forma consistente a lo largo de todas las muestras:

- Ciertos pares de variables están altamente correlacionados en todas las muestras:

Par de variables	Rango de Correlación	Frecuencia
OsBuild ↔ Census_OSBuildNumber	0.91 – 0.96	5/5
Census_InternalPrimaryDisplayResolutionHorizontal ↔ Census_InternalPrimaryDisplayResolutionVertical	0.88 – 0.95	5/5
Census_ProcessorManufacturerIdentifier ↔ Census_ProcessorModelIdentifier	0.80 – 0.84	5/5
Census_OSIInstallLanguageIdentifier ↔ Census_OSUILocaleIdentifier	0.97 – 1.00	5/5
RtpStateBitfield ↔ IsSxsPassiveMode	0.84 – 0.90	4/5

- La varianza explicada  $\geq 95\%$  requirió aproximadamente 32 componentes, lo cual evidencia alta dimensionalidad y multicolinealidad en el conjunto de datos.
- Se detectaron agrupaciones coherentes:

- Componentes que representan características físicas del dispositivo (RAM, Display, CoreCount).
  - Componentes de seguridad y virtualización (SecureBoot, TPM, IsVirtualDevice, IsProtected, AVProductsEnabled).
  - Otros que agrupan identificadores geográficos o de región (GeoNameIdentifier, CountryIdentifier, CityIdentifier).
- 
- La variable objetivo (HasDetections) presenta una distribución lo suficientemente balanceada, por lo que no se consideran necesarias técnicas de remuestreo para equilibrar las clases.
  - IsSecureBootEnabled, IsProtected, HasTPM, Firewall, AVProductsInstalled, etc. se repiten en múltiples componentes, lo que indica que la protección del sistema tiene peso explicativo importante.
  - Dispositivos táctiles y virtuales (IsPenCapable, IsTouchEnabled, IsVirtualDevice) también aparecen agrupados, sugiriendo un patrón tecnológico que puede ser predictivo.
  - Geografía e idioma: Varios componentes cargan sobre LocaleEnglishNameIdentifier, CountryIdentifier, CityIdentifier, y eso puede sugerir factores regionales en la presencia de detecciones.

## **Conclusiones**

- Los datos contienen redundancia importante por lo que podría ser útil reducir variables para evitar colinealidad.
- PCA muestra una estructura lógica y utilizable para ingeniería de características.
- La variable objetivo tiende a verse balanceada sobre múltiples muestras.
- Los factores técnicos y geográficos están claramente diferenciados.
- Una parte significativa de los datos fue descartada en las muestras analizadas, lo cual podría influir en la confiabilidad de las estadísticas obtenidas. Aunque los resultados se basan en observaciones consistentes realizadas sobre varias muestras independientes, el porcentaje de datos eliminados en relación con el total podría contribuir a la aparición de posibles sesgos. En adelante, se buscará la forma de cargar y procesar el conjunto de datos completo, con el fin de obtener estadísticas más representativas y robustas del comportamiento general de las variables.

## Referencias

- DataCamp. (s.f.). *Mean vs median: What's the difference and when to use which?* <https://www.datacamp.com/es/tutorial/mean-vs-median>
- DataCamp. (s.f.). *Techniques to handle missing data values.* <https://www.datacamp.com/es/tutorial/techniques-to-handle-missing-data-values>
- Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., ... & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1), 27-46. <https://doi.org/10.1111/j.1600-0587.2012.07348.x>
- Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. Springer.
- Jadhav, S., & Thakare, V. (2019). Handling missing data in machine learning. *International Journal of Engineering and Advanced Technology (IJEAT)*, 9(3), 1344-1348. <https://doi.org/10.35940/ijeat.C6749.029319>
- Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4), Article 15. <https://doi.org/10.1145/2382577.2382589>
- Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. CRC Press.
- Statistics by Jim. (s.f.). Correlations: Simple explanation and how to interpret them. <https://statisticsbyjim.com/basics/correlations/>
- Semer Nahdi. (2023). *Data Mining Project*. Kaggle. <https://www.kaggle.com/code/semernahdi/data-mining-project>