Identificación de Patrones en URLs de Malware

Esteban Mayen Soto - 0212614
Ricardo A. Flores Peregrina - 0213358

Pregunta de Investigación

¿Los links que recibimos en el día a día con patrones anormales pueden ser confiables?

Trasfondo

En la actualidad, el utilizar links que contienen malware para infectar equipos, es una de las maneras más sencillas y versátiles de esparcir malware.

Objetivo del proyecto

Identificar patrones comunes de links que contienen malware en URLs

Variables dependientes e independientes

- Independientes
  - URL
  - Tipo
  - Dominio
- Dependiente
  - Categoría

Datasets

Durante nuestra investigación encontramos distintos varios datasets y concluimos:

- Un dataset si contenía mucha información, pero no era útil para nuestro proyecto, así que fue descartado.
- Los datos necesitaban limpieza y agrupación.
- Armamos nuestro propio dataset a partir de 3 dataset útiles.
- Montamos el dataset en un Kaggle para tener un manejo eficaz.

kaggle

URLhaus

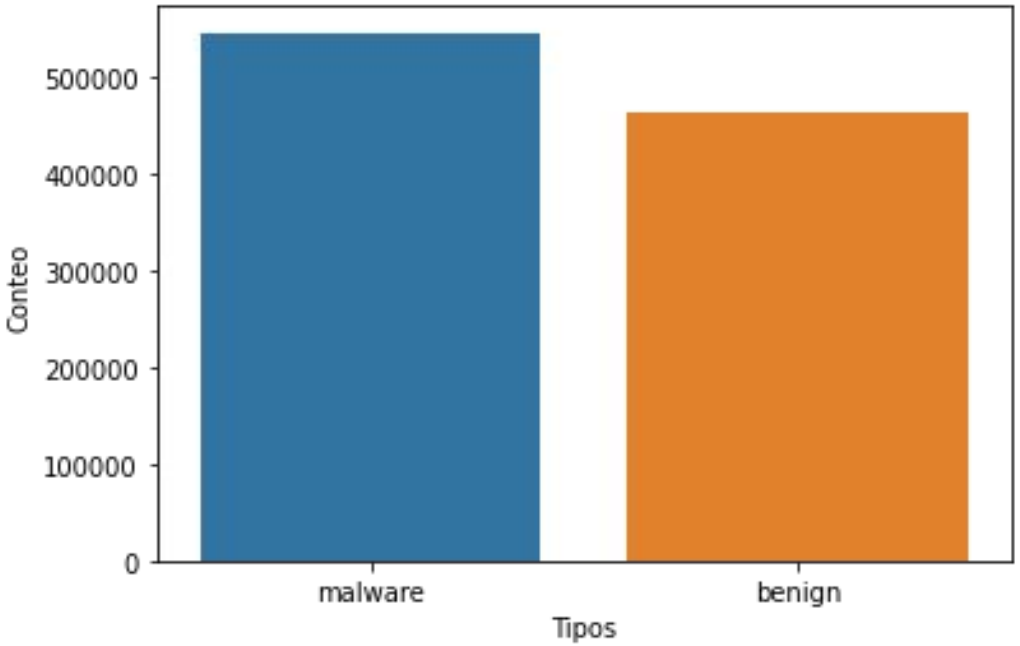
by ABUSE.ch

PhishTank

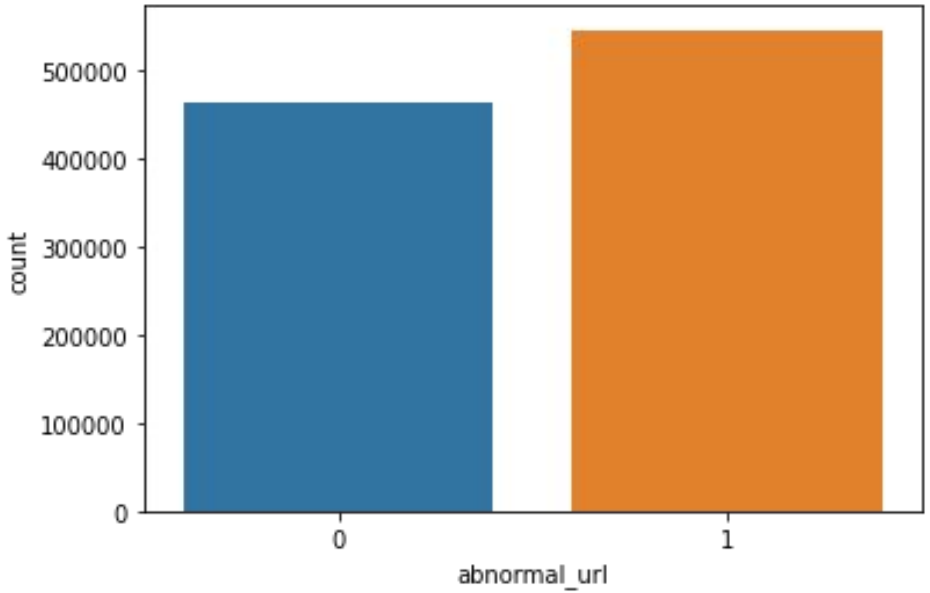
¿Cómo se ven los datos?

	URL	TIPO	Categoria	URL_LEN	DOMAIN	@	?	-	=	.	#	%	+	$	!	*	,	//
0	http://66.208.203.190:36841/malware.a	malware	1	37	None	0	0	0	0	4	0	0	0	0	0	0	0	1
1	http://58.255.129.35:53862/malware.a	malware	1	36	None	0	0	0	0	4	0	0	0	0	0	0	0	1
2	http://60.25.156.155:47183/malware.m	malware	1	36	None	0	0	0	0	4	0	0	0	0	0	0	0	1
3	http://192.72.17.236:35284/malware.a	malware	1	36	None	0	0	0	0	4	0	0	0	0	0	0	0	1
4	http://27.41.38.130:50541/malware.m	malware	1	35	None	0	0	0	0	4	0	0	0	0	0	0	0	1

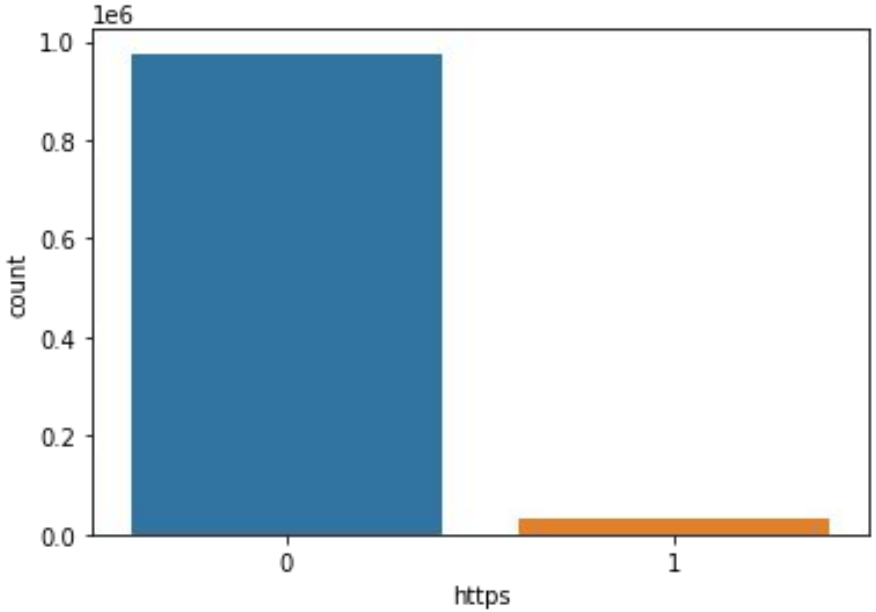
Conteo de URLs con malware y benignos



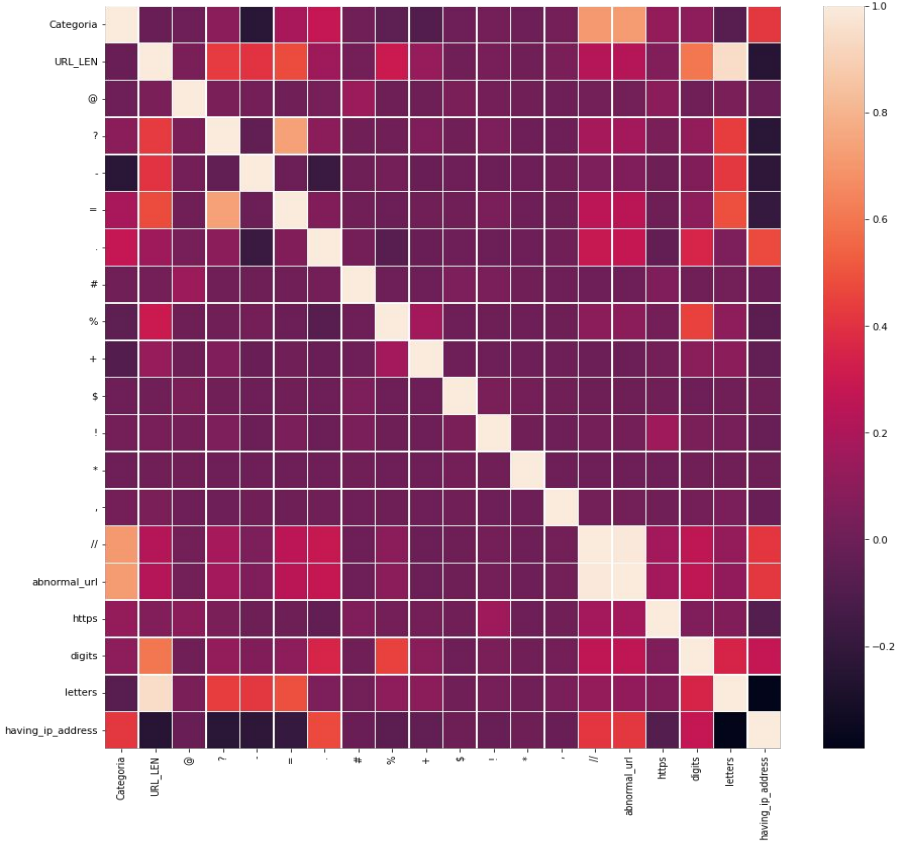
Conteo de URLs anormales



Conteo de dominios asegurados



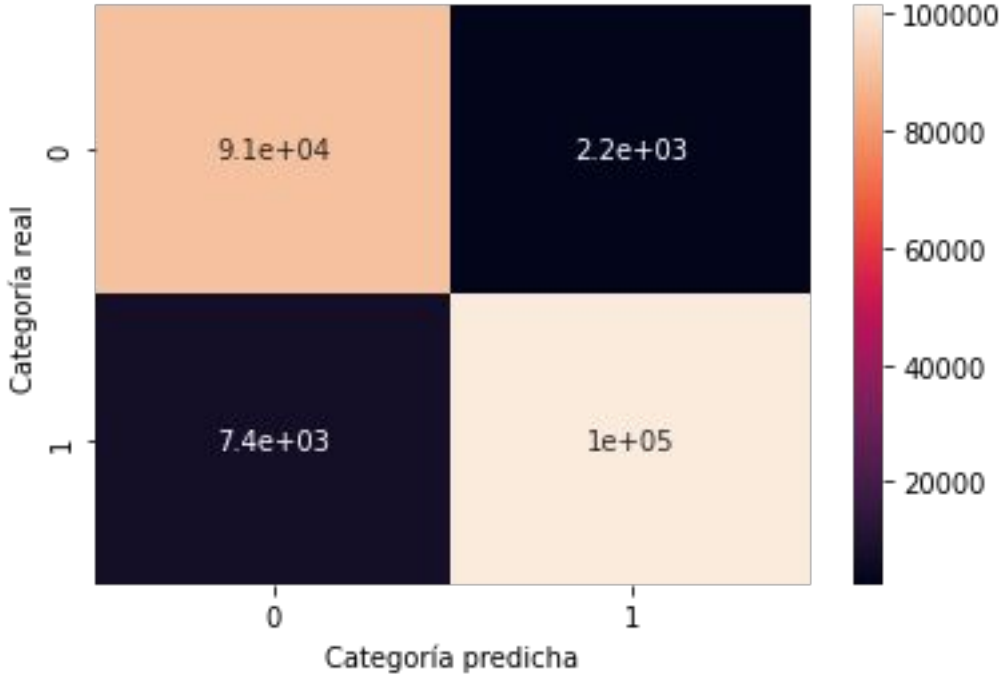
Correlograma



Modelo

Modelo supervisado de Árbol de Decisión

Matriz de Confusión




```
array([[ 90549,   2235],
       [  7432, 101558]])
```

Conclusión

Los URLs con malware comparten cierto tipo de características anormales y de longitud en muchos casos

Gracias!