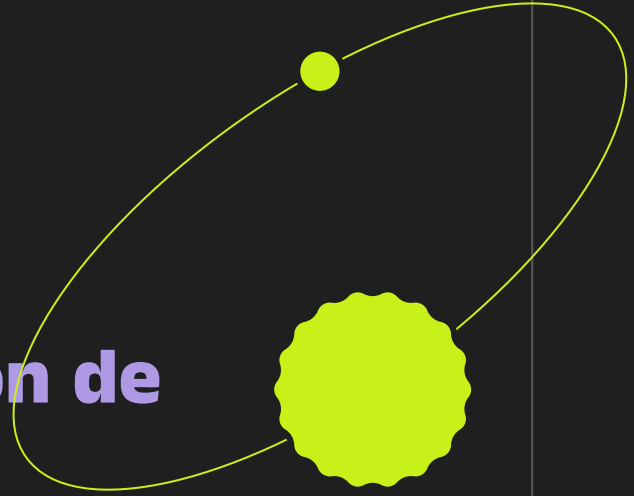


Fase 2

Procesamiento del lenguaje natural (NLP) para la detección de ataques de ingeniería social

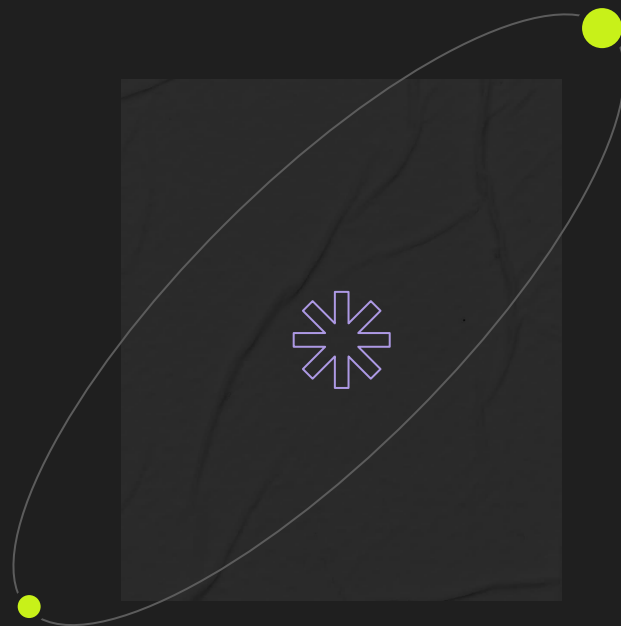
Roberto Castillo 18546
Hugo Roman 19199
Oscar de León 19298
Mirka Monzón 18139
Josué Sagastume 18173



←

01

Correcciones de la fase 1



You can describe the topic of the section here

Se obtuvo un nuevo dataset, que cumple con los requerimientos planteados durante la fase 1.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W												
1	Name	[md5]	Machine	[SizeOfOptionalHeader]	Characteristics	[MajorLinkerVersion]	[MinorLinkerVersion]	[SizeOfCode]	[SizeOfInitializedData]	[SizeOfUninitializedData]	[AddressOfEntryPoint]	[BaseOfCode]	[BaseOfData]	[ImageBase]	[SectionAlignment]	[FileAlignment]																			
2	mement.exe	631ea355665728d4707448e442fb5b8	332	224	258	9	0	361984	115712	0	6135	4096	372736	4194304	4096	512	0	0	0	1	1036288	1024	485887	16	1024	1048576	4096	1048576	4096	0	16	8	5.7668065537	3.60742957555	17
3	ose.exe	9d10f99a6712e28f8acd5641e3a7ea6b	332	224	3330	9	0	130560	19968	0	81778	4096	143360	771751936	4096	512	5	1	0	5	1159744	1024	188943	2	33088	1048576	4096	1048576	4096	0	16	4	4.83968793753	2.37352509596	6.5
4	setup.exe	4d92f518527353c0db88a70fddcf390	332	224	3330	9	0	517120	621568	0	350896	4096	811008	771751936	4096	512	5	1	0	5	1150976	1024	1159817	2	32832	1048576	4096	1048576	4096	0	16	4	4.40955752803	4.885191068	19
5	DW20.EXE	a41e5248d45f0074fd07805f0c9b12	332	224	258	9	0	585728	369152	0	451258	4096	798720	771751936	4096	512	5	1	0	5	1962560	1024	867570	2	33088	1048576	4096	1048576	4096	0	16	4	6.64173122458	5.64256492784	1
6	dwtirig20.exe	c87e561258f2f8650cef999b643a731	332	224	258	9	0	294912	247296	0	217381	4096	536576	771751936	4096	512	5	1	0	5	1552960	1024	579287	2	33088	1048576	4096	1048576	4096	0	16	4	6.25268442524	4.18222664425	6.5
7	airinstall.exe	e6e5a0ab3b1a27127c5c4a29b23d823	332	224	258	9	0	512	46592	0	4488	4096	8192	4194304	4096	512	5	0	5	65536	1024	57436	2	34112	1048576	4096	1048576	4096	0	16	5	4.20766363972	1.93316084353	6.78	
8	AcroBroker.exe	dd7d901720f71e7a45fb13ec973d8e9	332	224	290	9	0	222720	67072	0	219331	4096	229376	4194304	4096	512	5	0	0	5	303104	1024	359472	2	33088	1048576	4096	1048576	4096	0	16	5	5.58670842941	4.8460964324	1
9	AcroRd32.exe	540c61844ccd78c121c3ef48f3a34f0e	332	224	290	9	0	823808	650240	0	587663	4096	831488	4194304	4096	512	5	0	0	5	1507328	1024	1495645	2	33088	1048576	4096	1048576	4096	0	16	5	5.26669400276	3.694051177	1
10	AcroRd32Info.exe	9afe3c62668f55b8433cde602258236e	332	224	290	9	0	4096	7168	0	6751	4096	8192	4194304	4096	512	5	0	0	5	24576	1024	28316	2	33088	1048576	4096	1048576	4096	0	16	5	4.14491201014	0.393689010804	5.97
11	AcroTextExtractor.exe	ba621a9e644f6558c08cf25b40cb1bd4	332	224	290	9	0	29696	12800	0	27055	4096	36864	4194304	4096	512	5	0	0	5	57344	1024	90988	2	33088	1048576	4096	1048576	4096	0	16	5	4.67104326306	2.834778301	1
12	AdobeCollabSync.exe	bff0a35c0efca6f505b9e346dfcbdb33	332	224	290	9	0	917504	316928	0	833800	4096	921600	4194304	4096	512	5	0	0	5	1257472	1024	125224	2	33088	1048576	4096	1048576	4096	0	16	5	4.9386928364	5.028	1
13	Eula.exe	1556a34d17a80bdc85a66d8ea4fbcf2	332	224	290	9	0	53248	34816	0	53601	4096	57344	4194304	4096	512	5	0	0	5	102400	1024	149705	2	33088	1048576	4096	1048576	4096	0	16	5	5.10746542349	3.97749945546	6.27260
14	LogTransport2.exe	c4005b6d3cf77068bce158ac8ef7c522b	332	224	258	9	0	206848	102400	0	110150	4096	212992	4194304	4096	512	5	0	0	5	323584	1024	318536	3	33088	1048576	4096	1048576	4096	0	16	5	5.42096858086	4.796372	1
15	reader_sl.exe	e595f220ed529885d8b0cfa2e455e4d	332	224	259	9	0	14848	14336	0	16529	4096	20480	4194304	4096	512	5	0	0	5	40960	1024	80575	2	32768	1048576	4096	1048576	4096	0	16	4	4.78114536034	3.47096785451	6.2267
16	AcrobatUpdater.exe	0e9dee95fd47d6195da804a0deeda5b	332	224	258	9	0	178688	134144	0	78084	4096	184320	4194304	4096	512	5	0	0	5	339968	1024	356142	2	33088	1048576	4096	1048576	4096	0	16	5	5.00170176244	3.014107	1
17	AdobeARM.exe	47c1de0a890613ffcc1d67648eedf90	332	224	258	9	0	413184	518144	0	160191	4096	417792	4194304	4096	512	5	0	0	5	962560	1024	963805	2	33088	1048576	4096	1048576	4096	0	16	5	4.9247739033	3.58100593085	1
18	armvsc.exe	11a52cf7b265631deeb24c6149309eff	332	224	258	9	0	37376	20992	0	30988	4096	45056	4194304	4096	512	5	0	0	5	73728	1024	10181	2	33088	1048576	4096	1048576	4096	0	16	5	4.6481637566	2.52060789631	5.9778
19	ReaderUpdater.exe	5ed9b78b308d302c702d44f4505b3f46	332	224	258	9	0	178688	134144	0	78084	4096	184320	4194304	4096	512	5	0	0	5	339968	1024	380588	2	33088	1048576	4096	1048576	4096	0	16	5	5.00161255348	3.0135732	1
20	Adobe AIR Application Installer.exe	2da201644a6912ca8a11bb3089d0f3453	332	224	258	9	0	32256	91136	0	5926	4096	36864	4194304	4096	512	5	0	0	5	139264	1024	177404	2	33088	1048576	4096	1048576	4096	0	16	5	4.77418114696		1
21	Adobe AIR Updater.exe	397ef02798d24bf192997b5f7d8ed8ca	332	224	258	9	0	31744	64512	0	5275	4096	36864	4194304	4096	512	5	0	0	5	114688	1024	105476	2	33088	1048576	4096	1048576	4096	0	16	5	4.86286118484	2.25446901	1
22	template.exe	86fdbb3c4793f2b2e85bcc000fab7b	332	224	258	9	0	32256	26624	0	5926	4096	36864	4194304	4096	512	5	0	0	5	77824	1024	64395	2	33088	1048576	4096	1048576	4096	0	16	5	4.6204223717	2.2602764607	6.569223
23	csscan.exe	104541302c404ec763384081346cbe0fb	332	224	258	8	0	34304	54784	0	30939	4096	40960	4194304	4096	512	4	0	4	5	110592	1024	131936	3	33088	1048576	4096	1048576	4096	0	16	6	4.38826762637	0.321014020682	6.36
24	dainstall.exe	444caa5fc9729ef77da6965bae11c184	332	224	258	8	0	37376	30720	0	39387	4096	45056	4194304	4096	512	4	0	4	5	86016	1024	122876	3	33088	1048576	4096	1048576	4096	0	16	6	4.03150247251	0.84815527362	6.39
25	entvutil.exe	5d0df5cbbfd37853e623565cd081e47b	332	224	258	8	0	13824	7168	0	6123	4096	20480	4194304	4096	512	4	0	4	5	36864	1024	53304	3	33088	1048576	4096	1048576	4096	0	16	5	4.75098336422	2.78718651592	6.4599400
26	fwinfo.exe	c3879b7359dd566ed8cd6d5846ae1bc	332	224	258	8	0	82944	95744	0	50929	4096	90112	4194304	4096	512	5	0	5	0	192512	1024	214241	3	33088	1048576	4096	1048576	4096	0	16	5	4.62186128216	3.45625063429	6.546

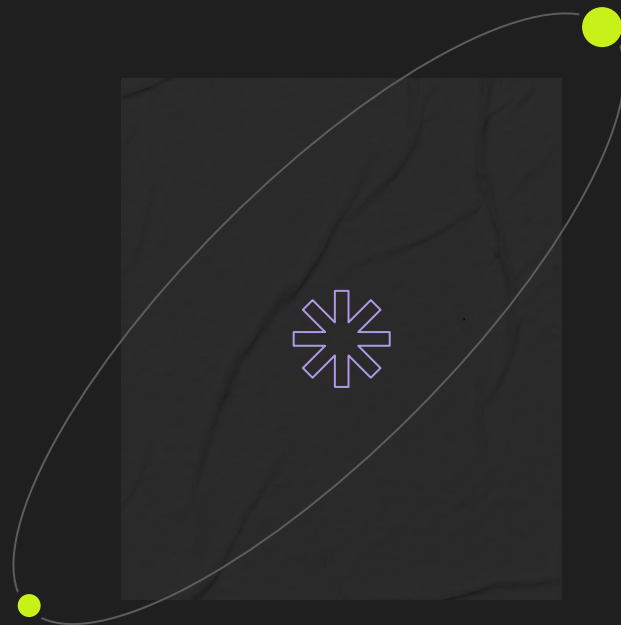
<https://github.com/muditmathur2020/RansomwareDetection/blob/master/Ransomware.csv>



←

02

Modelos ML/DL/RL



You can describe the topic of the section here

Regresión Logística

- Simplicidad
- Interpretabilidad
- Rápida convergencia y entrenamiento
- Menor riesgo de sobreajuste

Random Forest

- Robustez y precisión
- Manejo de características no lineales
- Importancia de las características

Support Vector Machines

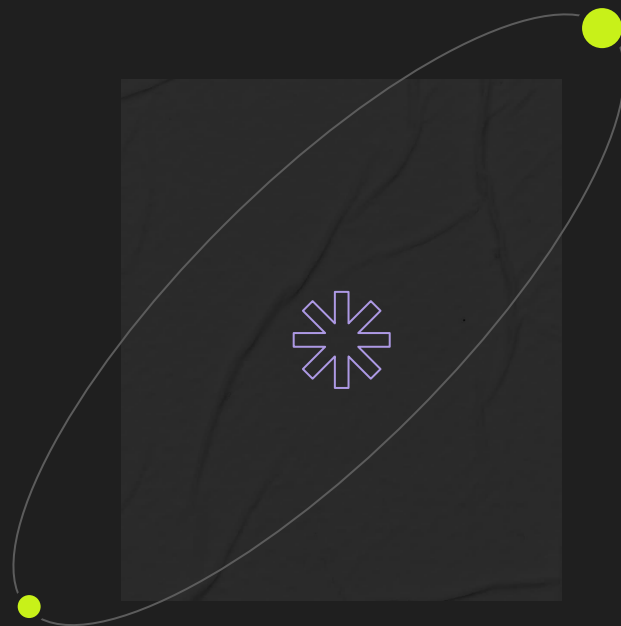
- Eficacia en espacios de alta dimensión
- Flexibilidad
- Robustez
- Generalización



←

03

Analisis exploratorio



You can describe the topic of the section here

	Name object	md5 object	Machine int64	SizeOfOptional...	Characteristics i...	MajorLinkerVer...	MinorLink
0	memtest.exe	631ea355665f28d4707448e442fbf...	332	224	258	9	
1	ose.exe	9d10f99a6712e28f8acd5641e3a7ea...	332	224	3330	9	
2	setup.exe	4d92f518527353c0db88a70fddcd...	332	224	3330	9	
3	DW20.EXE	a41e524f8d45f0074fd07805ff0c9b12	332	224	258	9	
4	dwtrig20.exe	c87e561258f2f8650cef999bf643a731	332	224	258	9	
5	airappinstaller.exe	e6e5a0ab3b1a27127c5c4a29b237d...	332	224	258	9	
6	AcroBroker.exe	dd7d901720f71e7e4f5fb13ec973d8e9	332	224	290	9	
7	AcroRd32.exe	540c61844ccd78c121c3ef48f3a34f...	332	224	290	9	
8	AcroRd32Info.exe	9afe3c62668f55b8433cde6022582...	332	224	290	9	
9	AcroTextExtractor.exe	ba621a96e44f6558c08cf25b40cb1...	332	224	290	9	

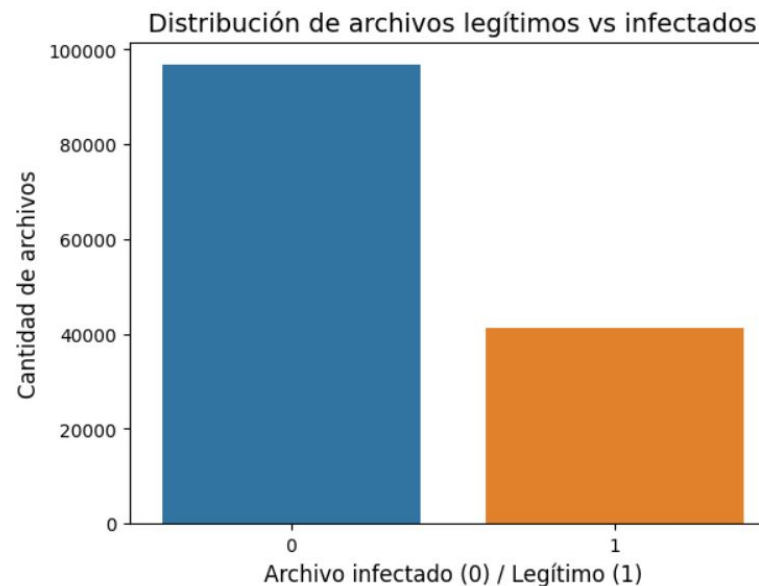
La cantidad de registros del dataset es: 138047

0 96724

1 41323

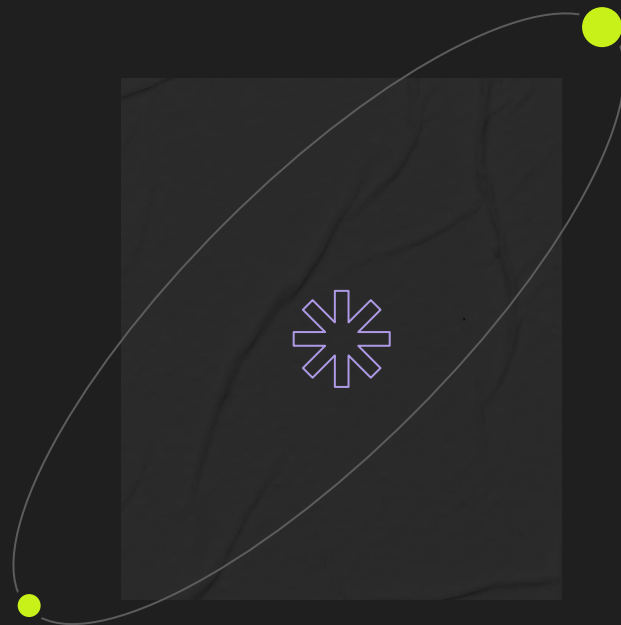
Name: legitimate, dtype: int64

Porcentaje de archivos legítimos: 29.93%
Porcentaje de archivos infectados: 70.07%





04

Implementación de los modelos



You can describe the topic of the section here



```
# Modelo 1: Regresión Logística
model1 = LogisticRegression()
model1.fit(X_train, y_train)

# Modelo 2: Random Forest
model2 = RandomForestClassifier()
model2.fit(X_train, y_train)

# Modelo 3: Support Vector Machines
model3 = SVC()
model3.fit(X_train, y_train)
```

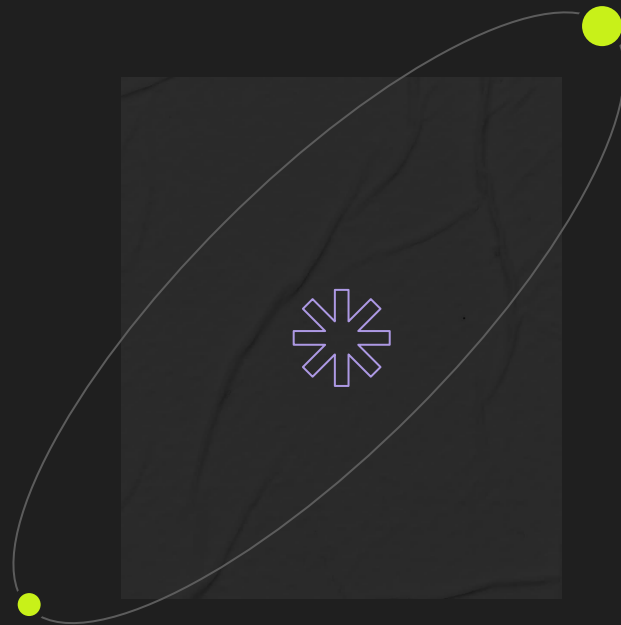


←

→

05

Métricas de evaluación



You can describe the topic of the section here



Regresión Logística

Accuracy: 0.9816486866791745
Precision: 0.9623708325209592
Recall: 0.9762658227848101
F1: 0.96926853215513
Confusion matrix: [11807, 193]
[120, 4936]

Random Forest

Accuracy: 0.9949577861163227
Precision: 0.9893658920834975
Recall: 0.9936708860759493
F1: 0.9915137162028813
Confusion matrix: [11946, 54]
[32, 5024]

Support Vector Machines

Accuracy: 0.9893878986866792
Precision: 0.9782224838140082
Recall: 0.9861550632911392
F1: 0.9821727568206441
Confusion matrix: [11889, 111]
[70, 4986]





Validación Cruzada

Regresión Logística

Accuracy: 0.9784353004423192
Precision: 0.9803188575077314
Recall: 0.9739671360741458
F1: 0.9771307024177563

Random Forest

Accuracy: 0.9961519525226235
Precision: 0.9945077081580076
Recall: 0.9969760250494717
F1: 0.995894845996788

Support Vector Machines

Accuracy: 0.9880593527880693
Precision: 0.9889209501771783
Recall: 0.9858002995045533
F1: 0.9873573649750609

