

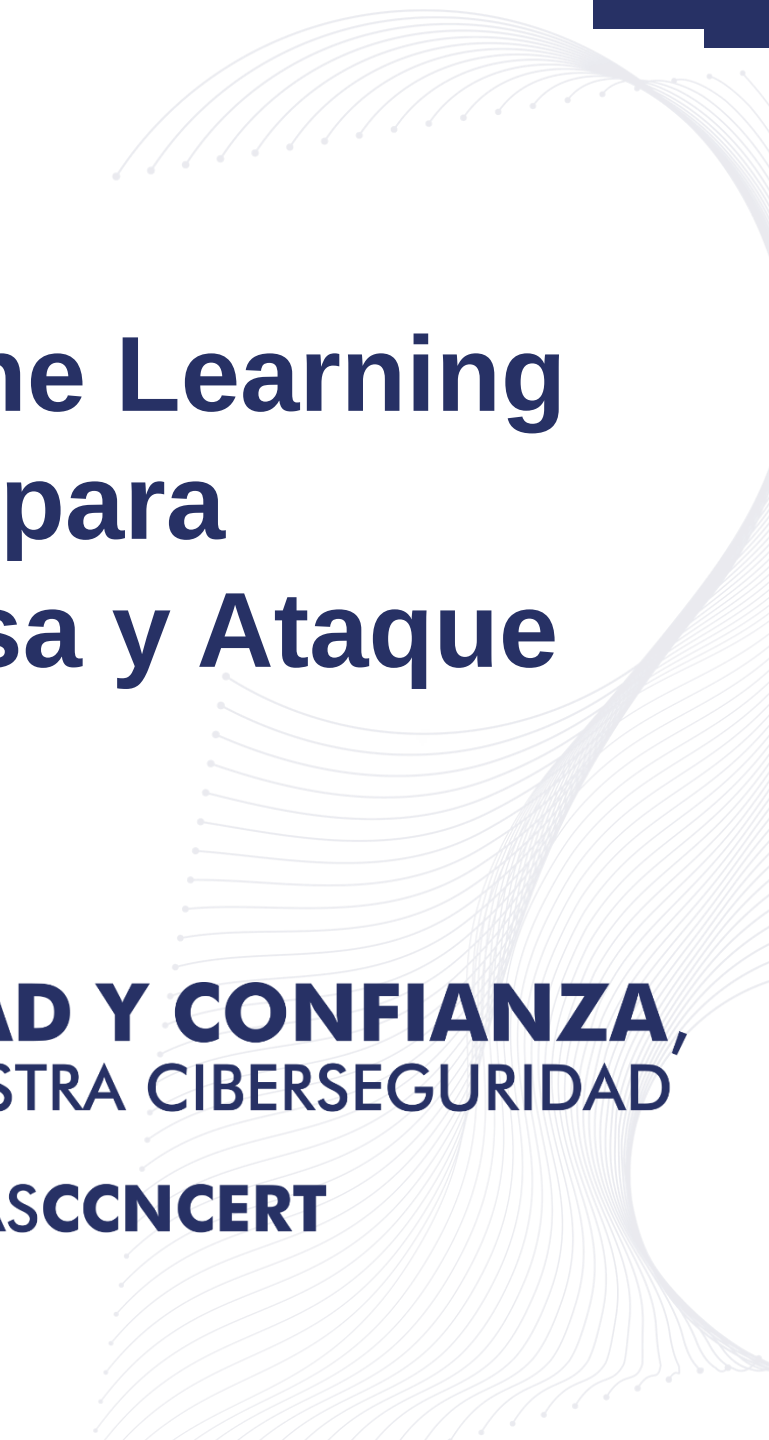
**XIII
JORNADAS
STIC
CCN-CERT**



**Machine Learning
para
Defensa y Ataque**

**COMUNIDAD Y CONFIANZA,
BASES DE NUESTRA CIBERSEGURIDAD**

#XIIIJORNADASCCNCERT





Jose Selvi

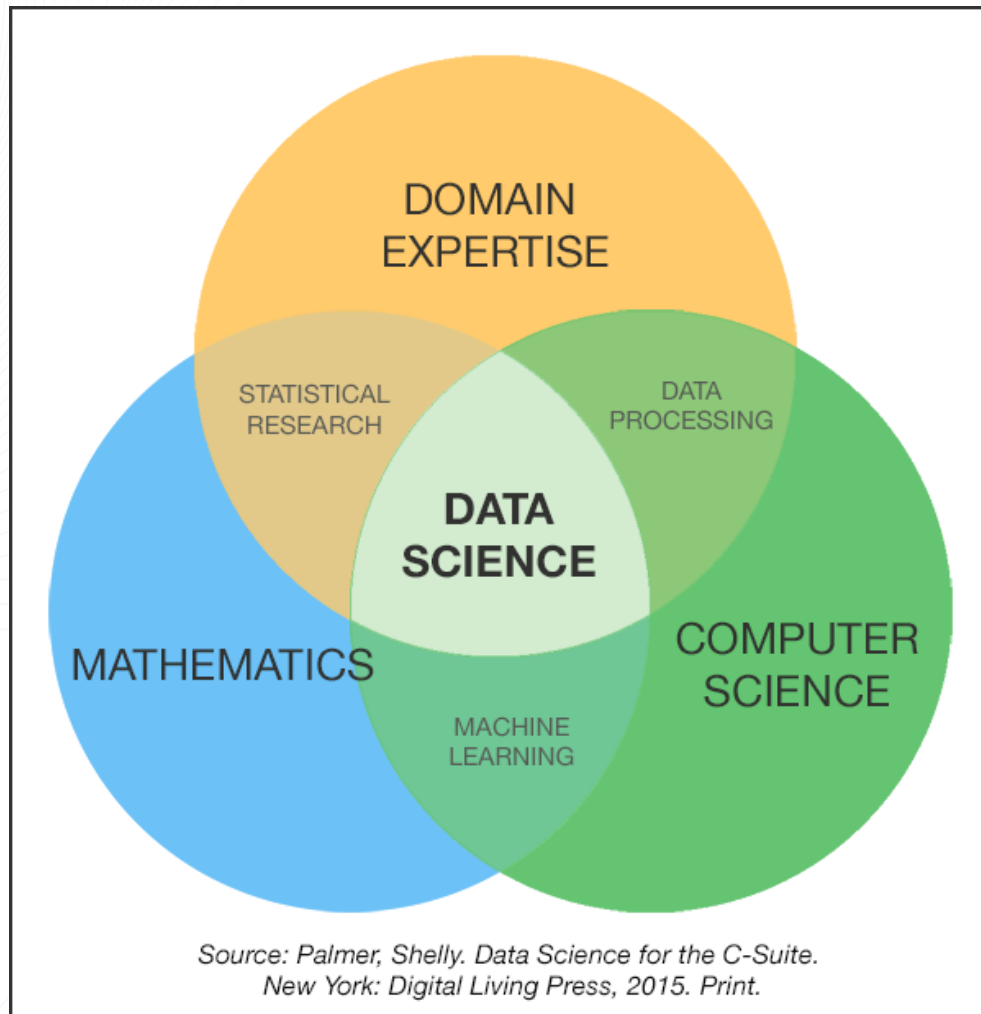
nccgroup[®]

jose.selvi@nccgroup.com

Índice

- 1. Introducción**
- 2. Detectando DGA Malware**
- 3. Project AVA & Outlier Detection**
- 4. Lecciones Aprendidas**
- 5. Preguntas**

¿Qué es el “Machine Learning”?



Machine Learning hasta en la sopa



1. Next Generation AntiVirus
2. Next Generation Firewall
3. Next Generation IDS
4. Next Generation SIEM
5. Next Generation Lo Que Sea

No es oro todo lo que reluce...

- "... y aquí tenemos las reglas para el motor de Machine Learning"
- $\text{Result} = w1 * \text{heur1} + w2 * \text{heur2} + w3 * \text{mlearn}$
- Usando solo unos pocos parámetros...
- Etc.



No es oro todo lo que reluce...

- "... y aquí tenemos las reglas para el motor de Machine Learning"
- $\text{Result} = w1 * \text{heur1} + w2 * \text{heur2} + w3 * \text{mlearn}$
- Usando solo unos pocos parámetros...
- Etc.



Tampoco es tan complicado como parece





Construcción de un modelo



```
func IsPaella(features []int) bool {  
    return true  
}
```

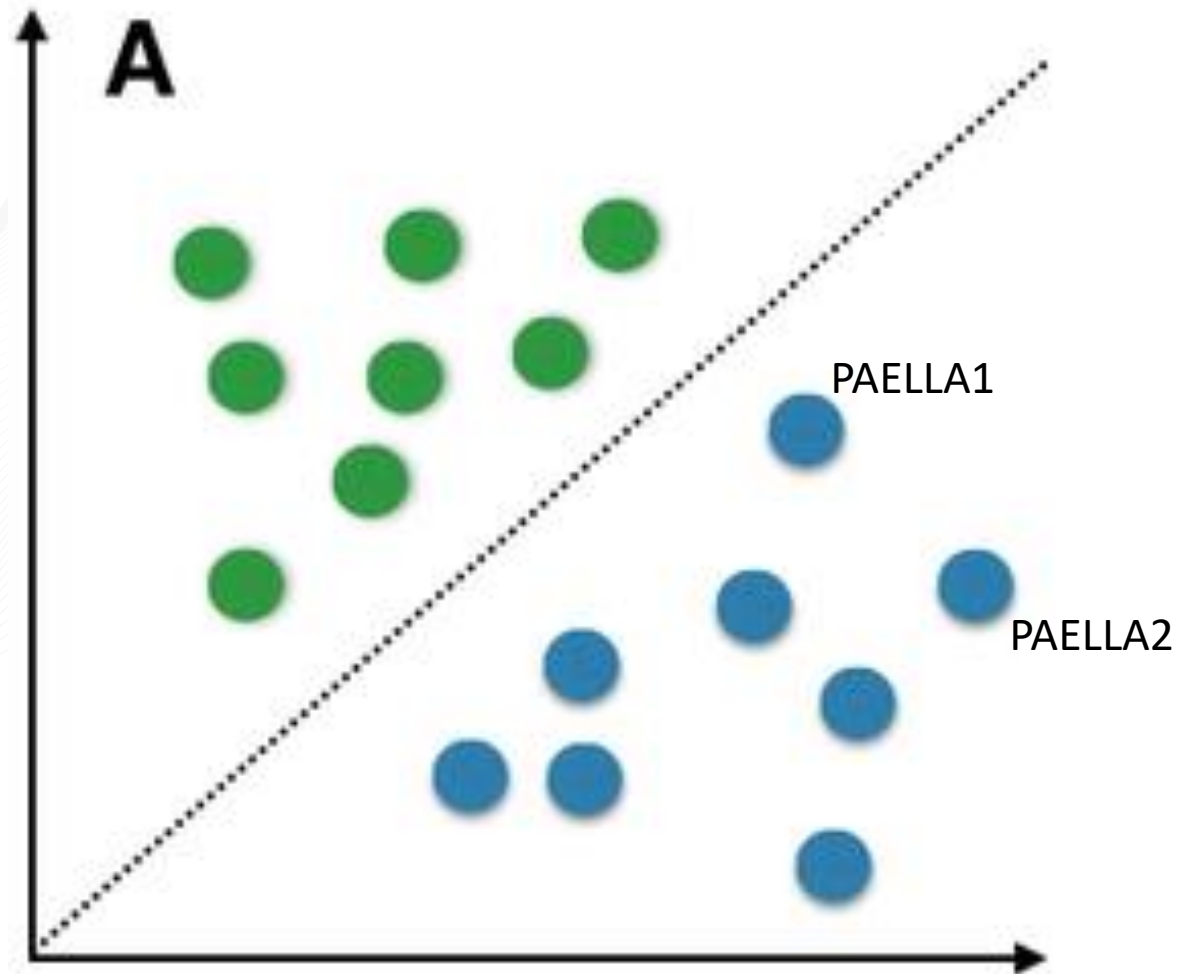


Extracción de Características



1. Cantidad de Arroz
2. Color del Arroz
3. “Melosidad” del arroz
4. Cantidad de Pollo
5. Cantidad de “Bachoqueta”
6. Cantidad de “Garrofó”
7. Cantidad de Guisantes
8. Cantidad de Pimiento Rojo

Vector de Características N-Dimensional



1. Cantidad de Arroz
2. Color del Arroz
3. “Melosidad” del arroz
4. Cantidad de Pollo
5. Cantidad de “Bachoqueta”
6. Cantidad de “Garrofó”
7. Cantidad de Guisantes
8. Cantidad de Pimiento Rojo

Direcciones IP

213.242.93.237



82.165.229.138



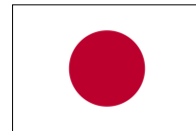
213.73.89.123



216.58.211.196

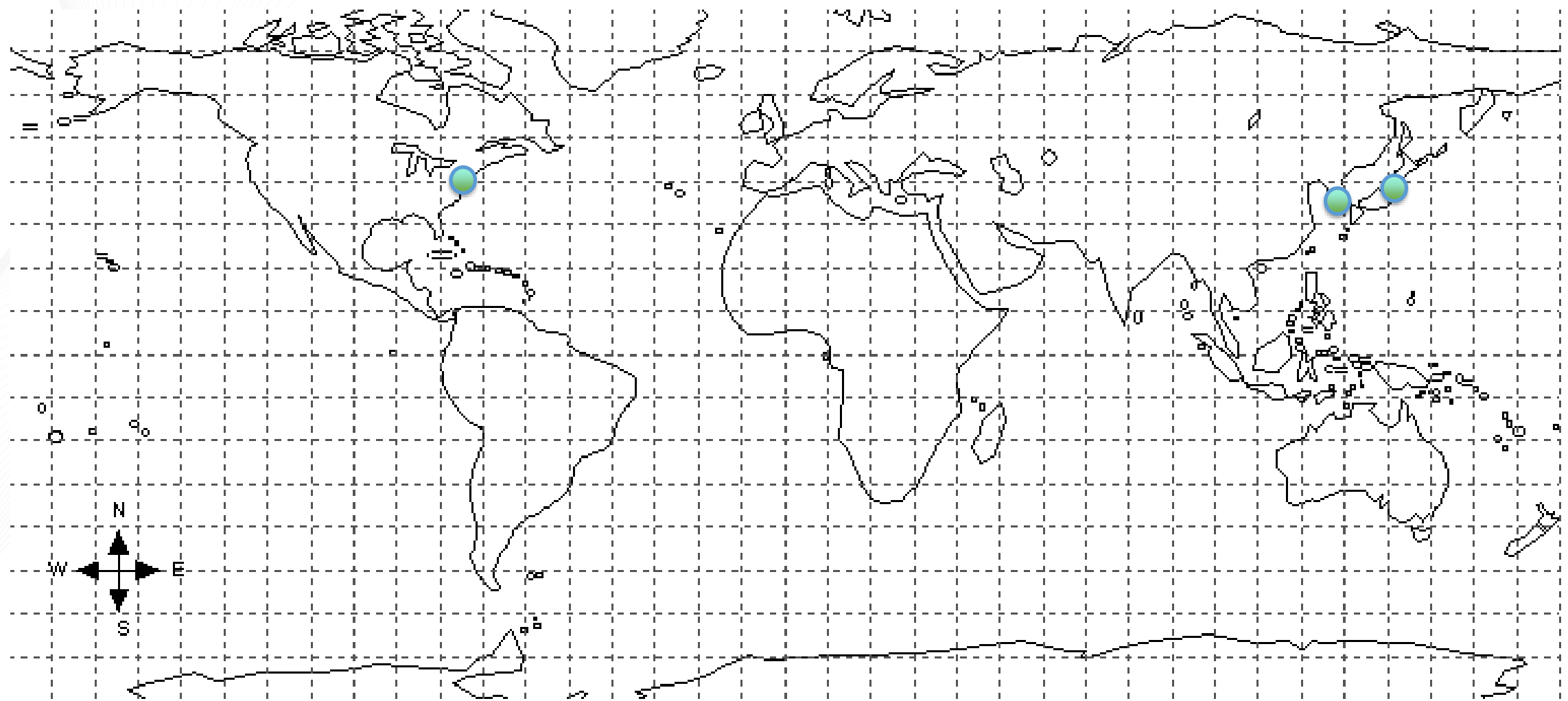


133.237.48.124

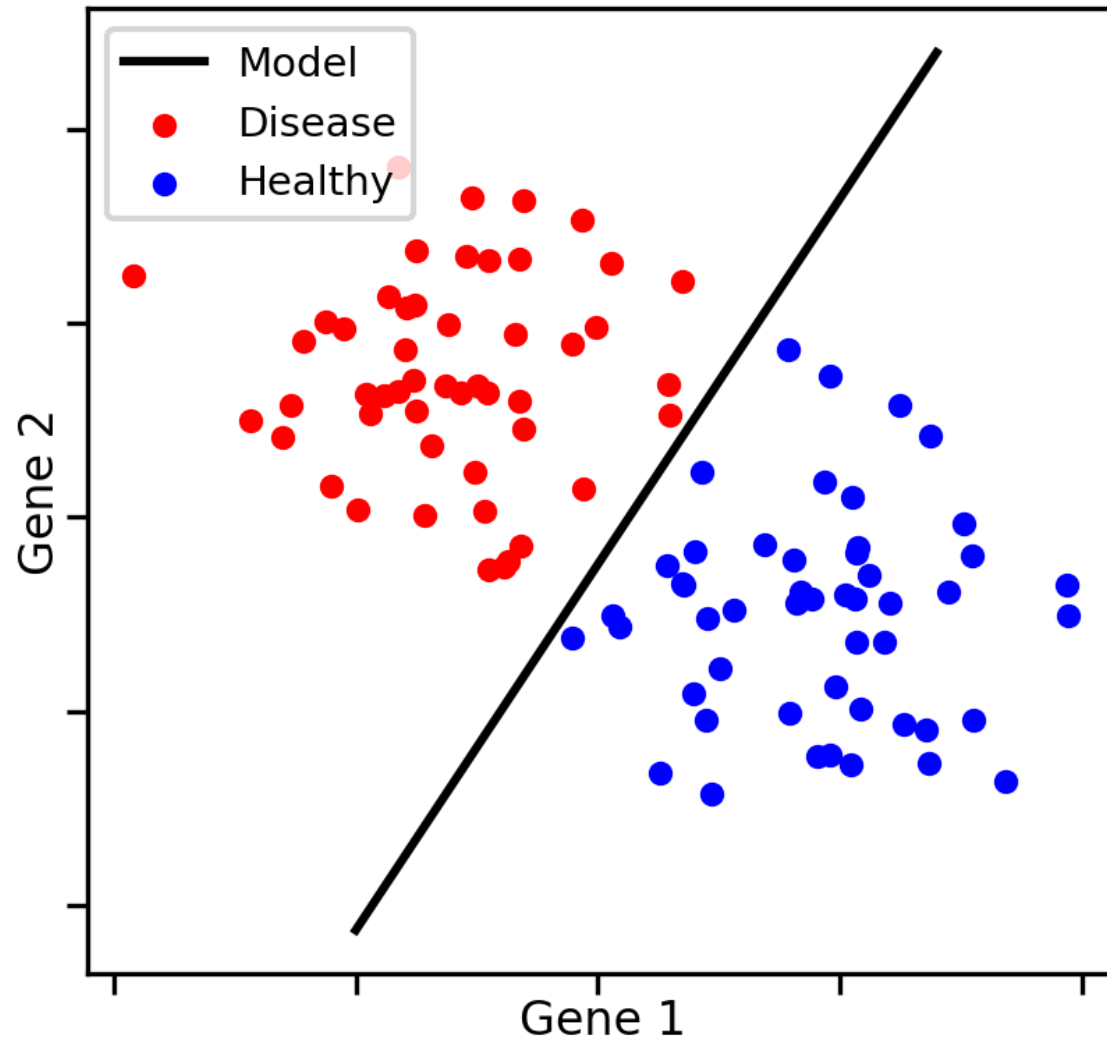


202.179.177.22

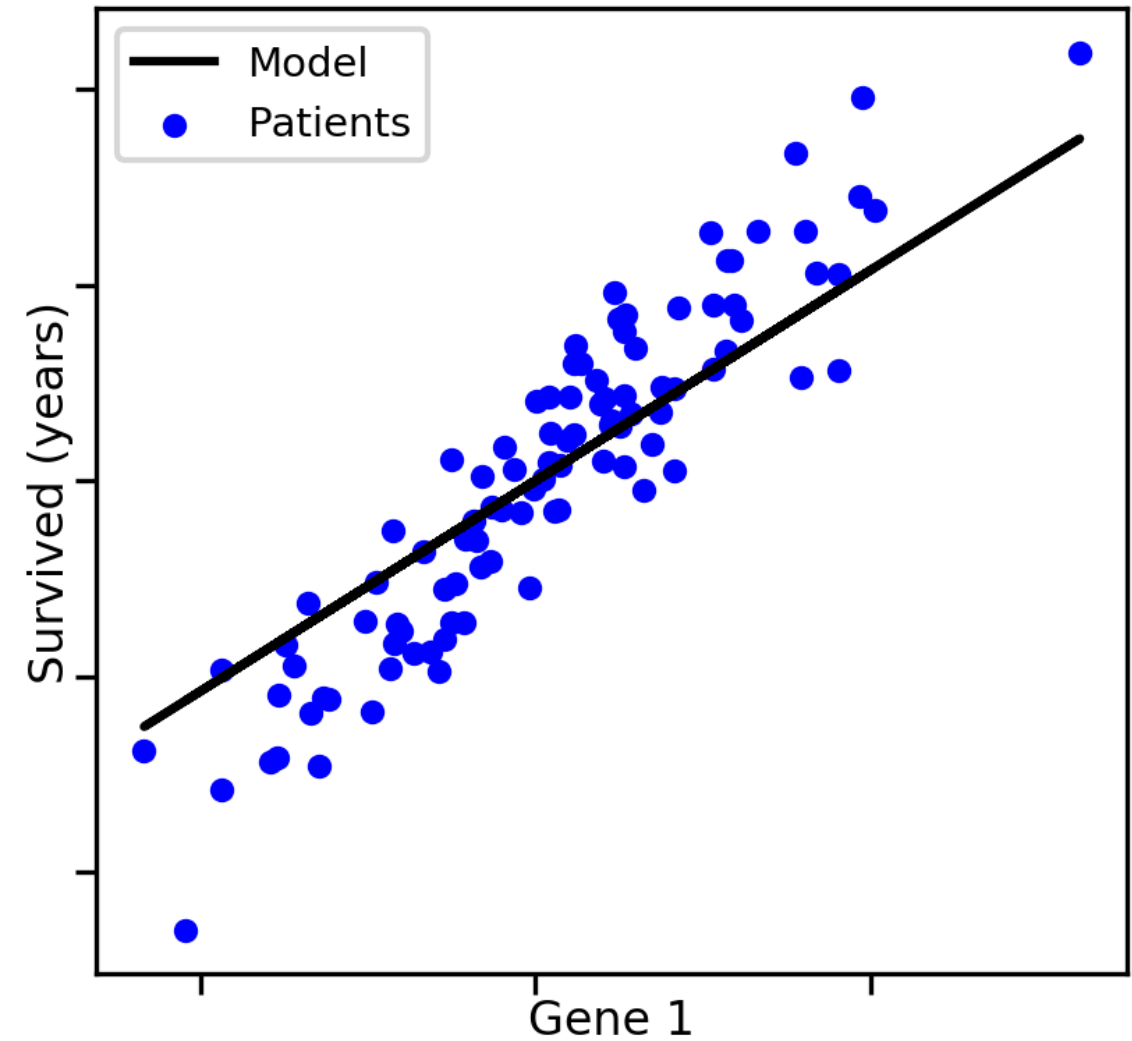




Classification

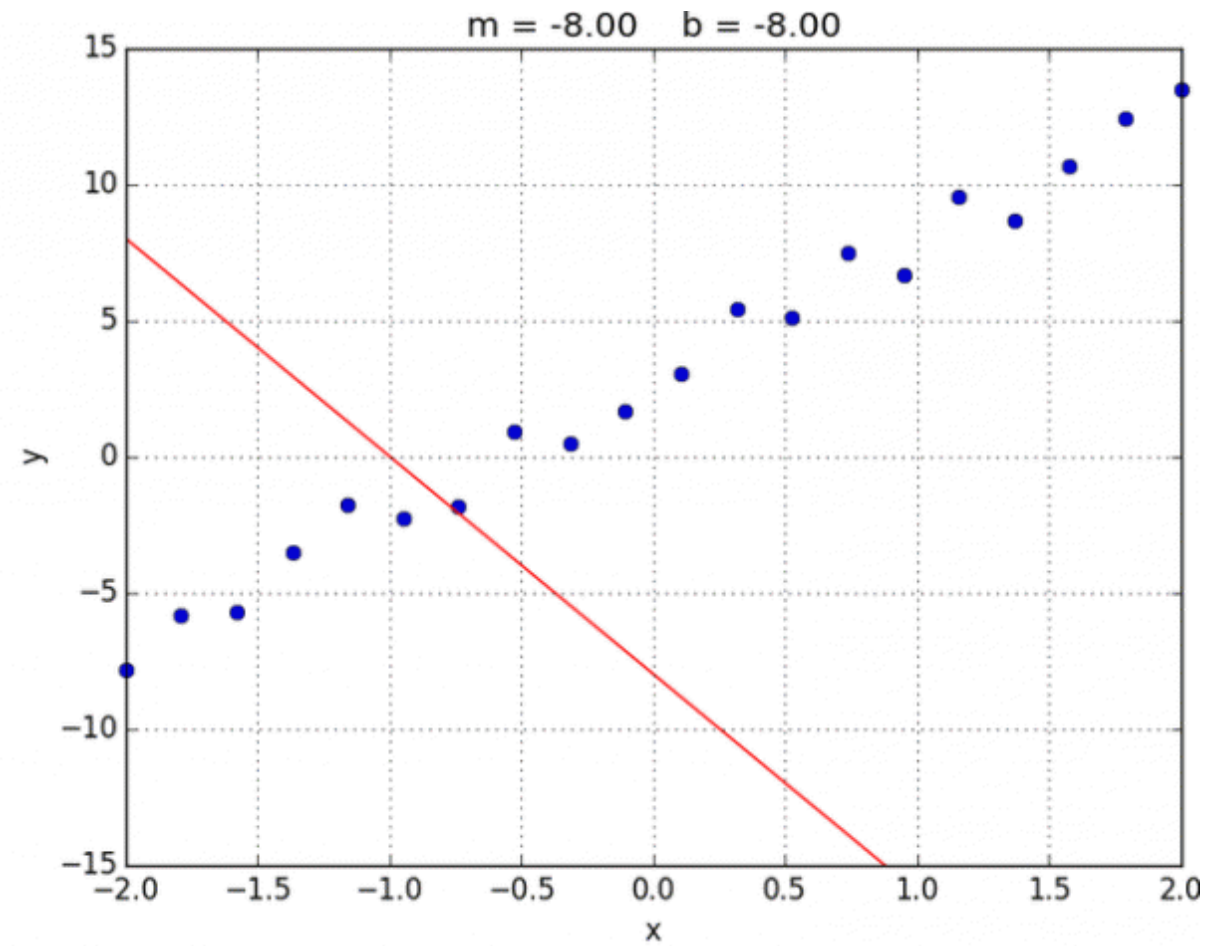
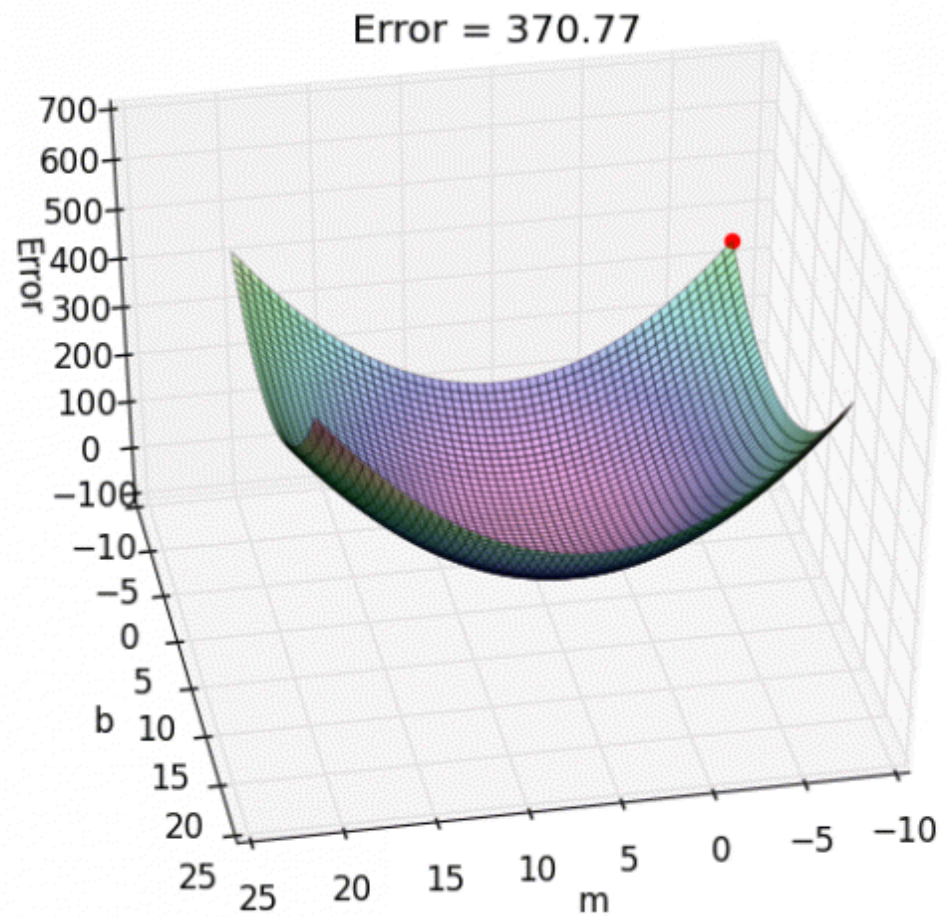


Regression

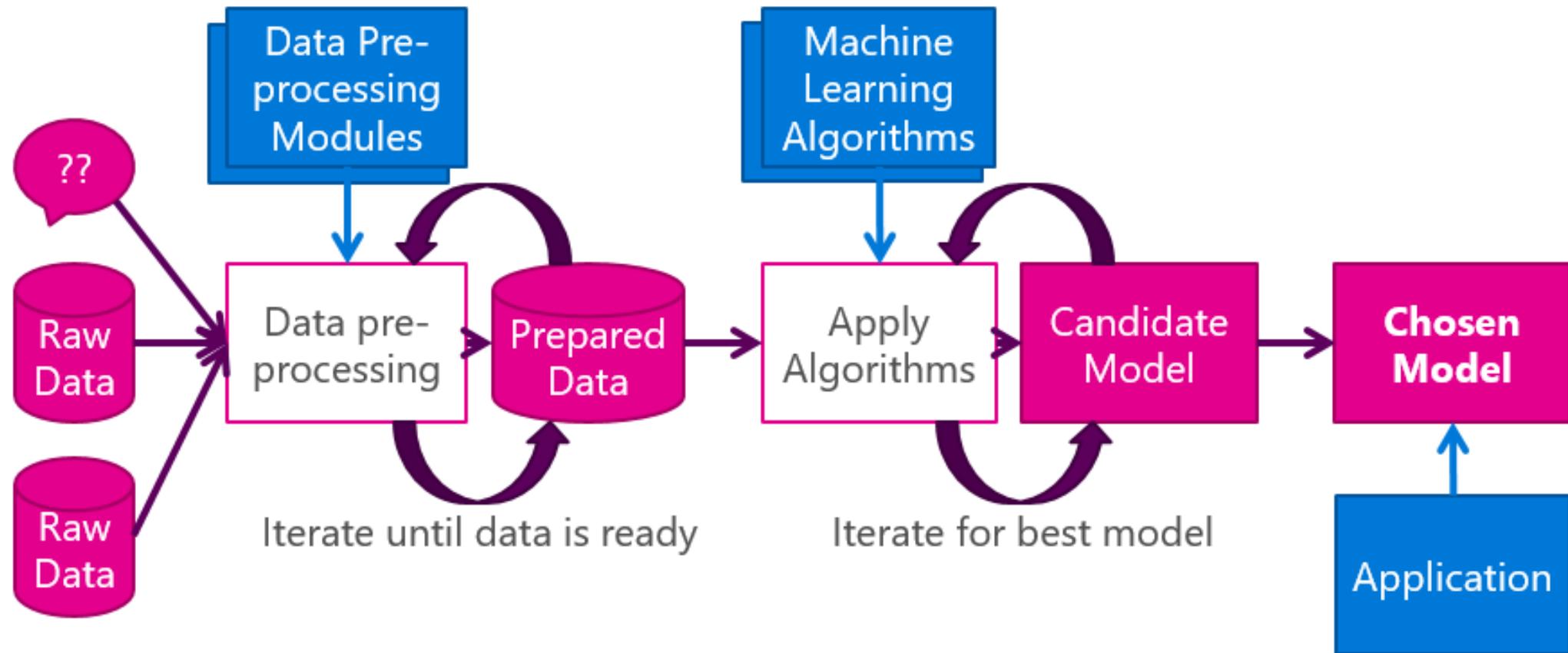


<https://github.com/ramrathi/IECSE-ML-Winter18/wiki/Basic-Logistic-Regression>

Descenso por Gradiente



Proceso de Aprendizaje



Índice

~~1. Introducción~~

2. Detectando DGA Malware

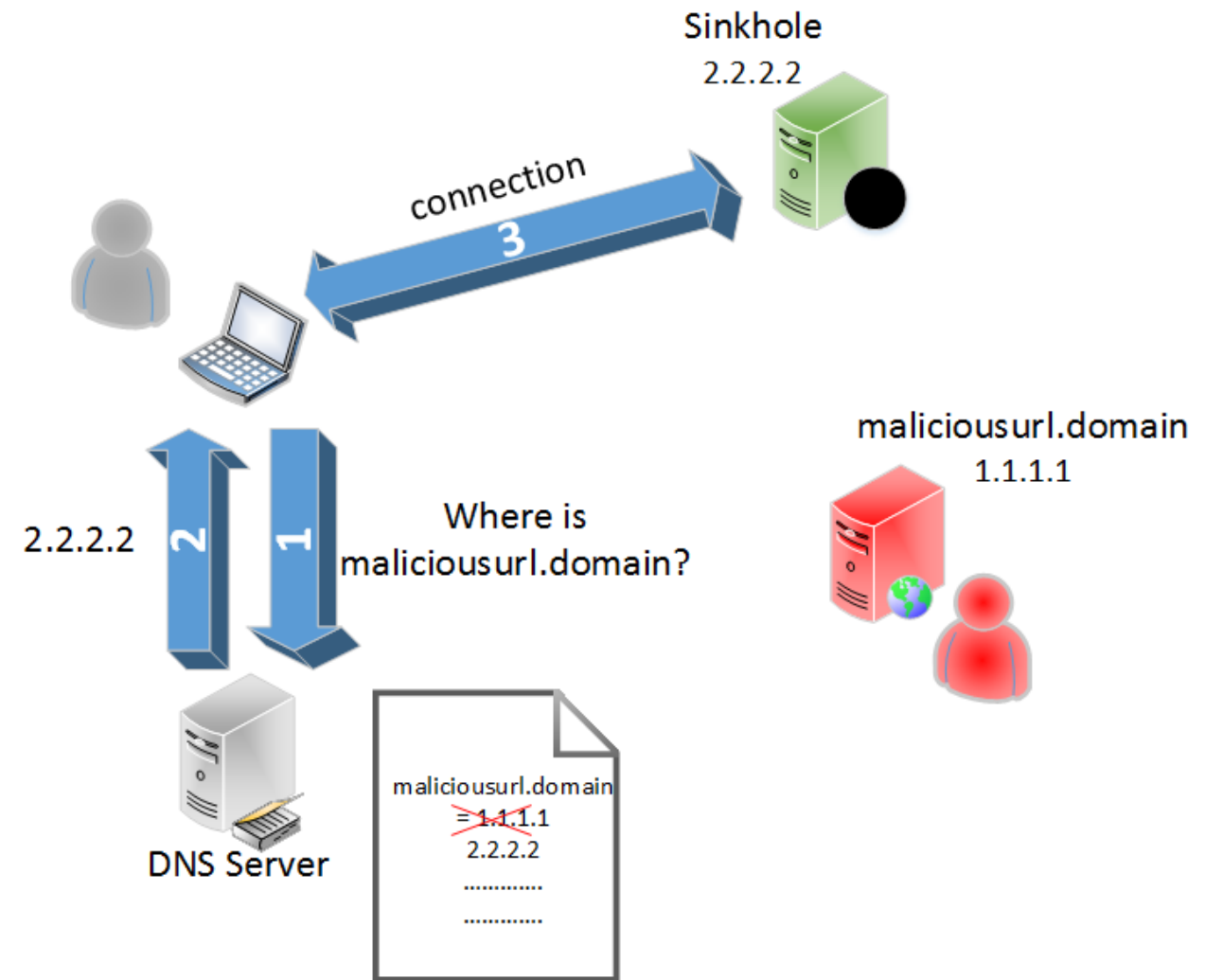
3. Project AVA & Outlier Detection

4. Lecciones Aprendidas

5. Preguntas

Detección usando DNS

nxlya47huo61czerb18o51e11d30i55gycwe31lx.ru
 jwdzptm69p62izcve41f22k37oyj16g63fqote11.com
 p42p52nvd50izkqazaqe21lvo21pycqotp22e61.net
 b28n40i25b68gte41o61dwc19htc29jwgxiqfzbr.org
 ktirhsn50kzc49b58cyf32fwh14h64dzgxiqcz.info
 bre41hvc29kri15ewpwwsazjyn40p52kwe21gw.biz
 n30mwhsoxfqe51j56lunsg13o11hyd60ewf52nu.ru
 hvcsjxd20mzm29d40nznunta27c29kyi55fun50.com
 nzosg13oymzg63ntpxaro51btkvfyoshrk27.info
 czfsn20exg53nzcqcrq43exf62b28p22pyd50lu.org
 e51h64muh24g13a37i35l68brc59fweqe61pwewbx.biz
 msdvlzirl48jzlroypvh14exo51gsntasjq.com
 pqhtjzmqdza27lsh34p62gsh14aqm19mvdqd20.ru
 mviriypxb38nriyjqgrksm49cxfzmsdsk.com
 Iso11ozpxnwk57m49kqnmwi45a27c69puoyi25.net
 b28pubyg13mzftf42czewguhsm19azd60prjx.biz
 n10o11ntetmykqn30a37l48iyfwctmulra27nv.info
 atewmunqbwjqfxnxja27ptczl58bvgzm69.com
 bqj26l38k27k67o41gqk37gum39hyc69n10l18kqa67.ru
 lshwb18bvp62m39fqfrfue61oym19gqgzlsmt.com
 mta37duhrpqpuc69asj46hxkvk57cvizo41d60.biz



Da Luz, Botnet Detection Using Passive DNS, 2013

	ID	Designation
Lexical	F1-3	1-gram (mean, variance and standard deviation)
	F4-6	2-gram (mean, variance and standard deviation)
	F7-9	3-gram (mean, variance and standard deviation)
	F10-12	4-gram (mean, variance and standard deviation)
	F13-14	Shannon entropy in 2LD and 3LD
	F15	Number of different characters
	F16	Number of digits / domain name length
	F17	Number of consonants / domain name length
	F18	Number of consonants / number of vowels
Network	F19-21	TTL, Window (first seen - last seen), Ratio (window / count)
	F22-24	Number of different first seen, answers and TTL
	F25-27	Maximum Ratio, TTL and Window
	F28-30	Mean Ratio, TTL and Window
	F31-33	Variance of Ratio, TTL and Window
	F34-36	Number of IP subnetworks

¿Por qué estadísticas de los N-Grams?

F	A	C	E	B	O	O	K	.	C	O	M
---	---	---	---	---	---	---	---	---	---	---	---

N	Grams	#
1	A, B, C, D, E, F, G, ... , X, Y, Z	40
2	AA, AB, AC, ... , BA, BB, BC, ... , ZZ	1600
3	AAA, AAB, AAC, ... , ZZW, ZZX, ZZZ	64k
4	AAAA, AAAB, ... , ZZZX, ZZZZ	2.56M

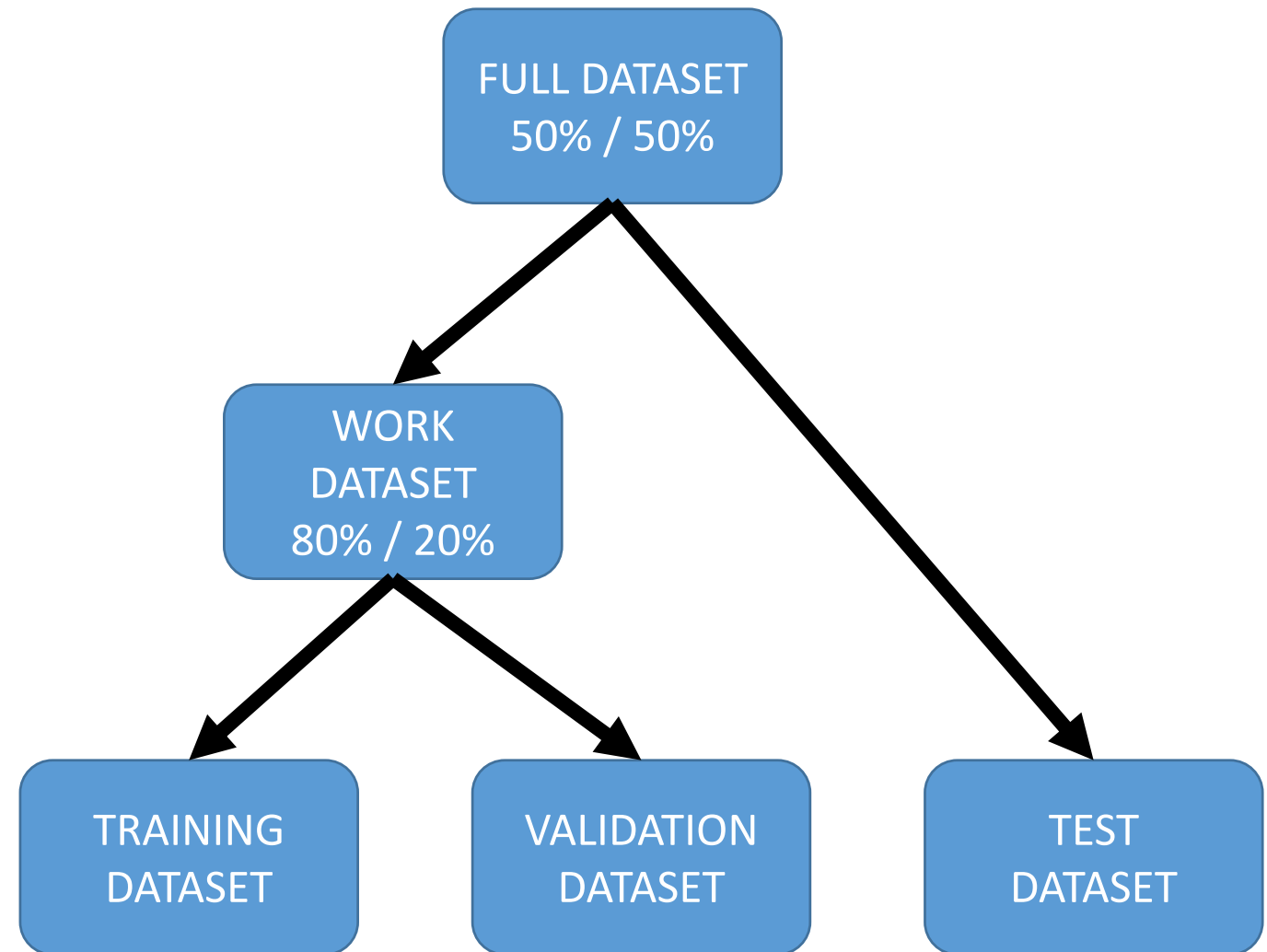
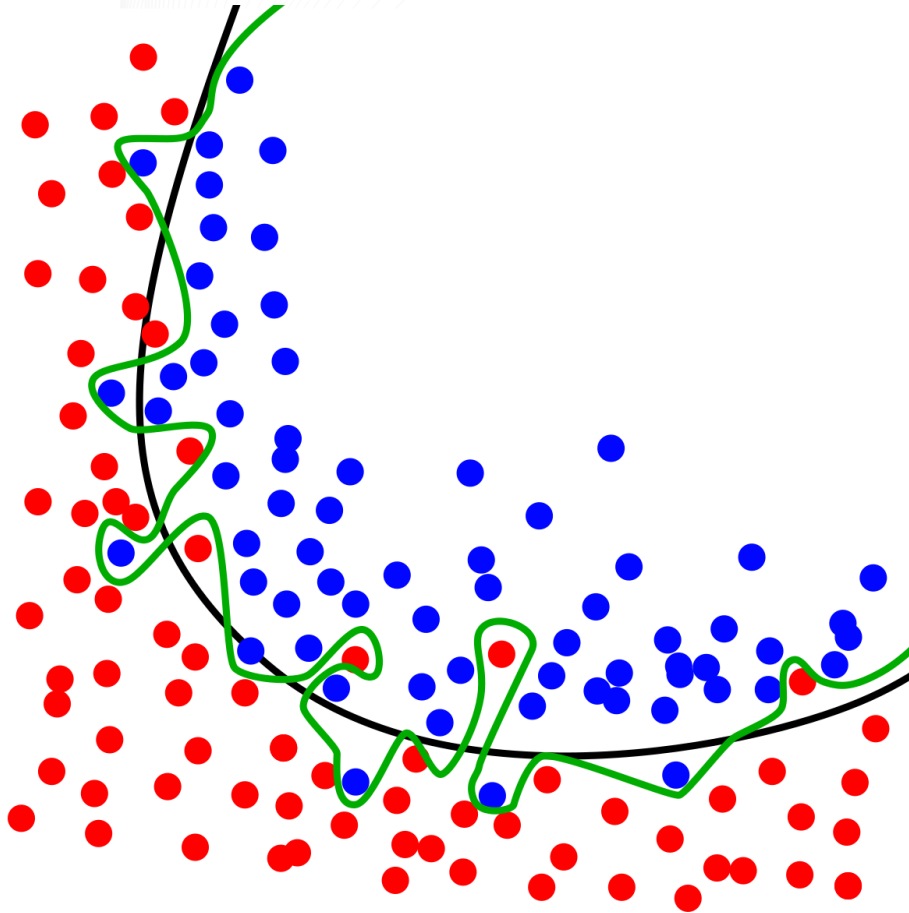
N-Grams Enmascarados

F	A	C	E	B	O	O	K	.	C	O	M
---	---	---	---	---	---	---	---	---	---	---	---

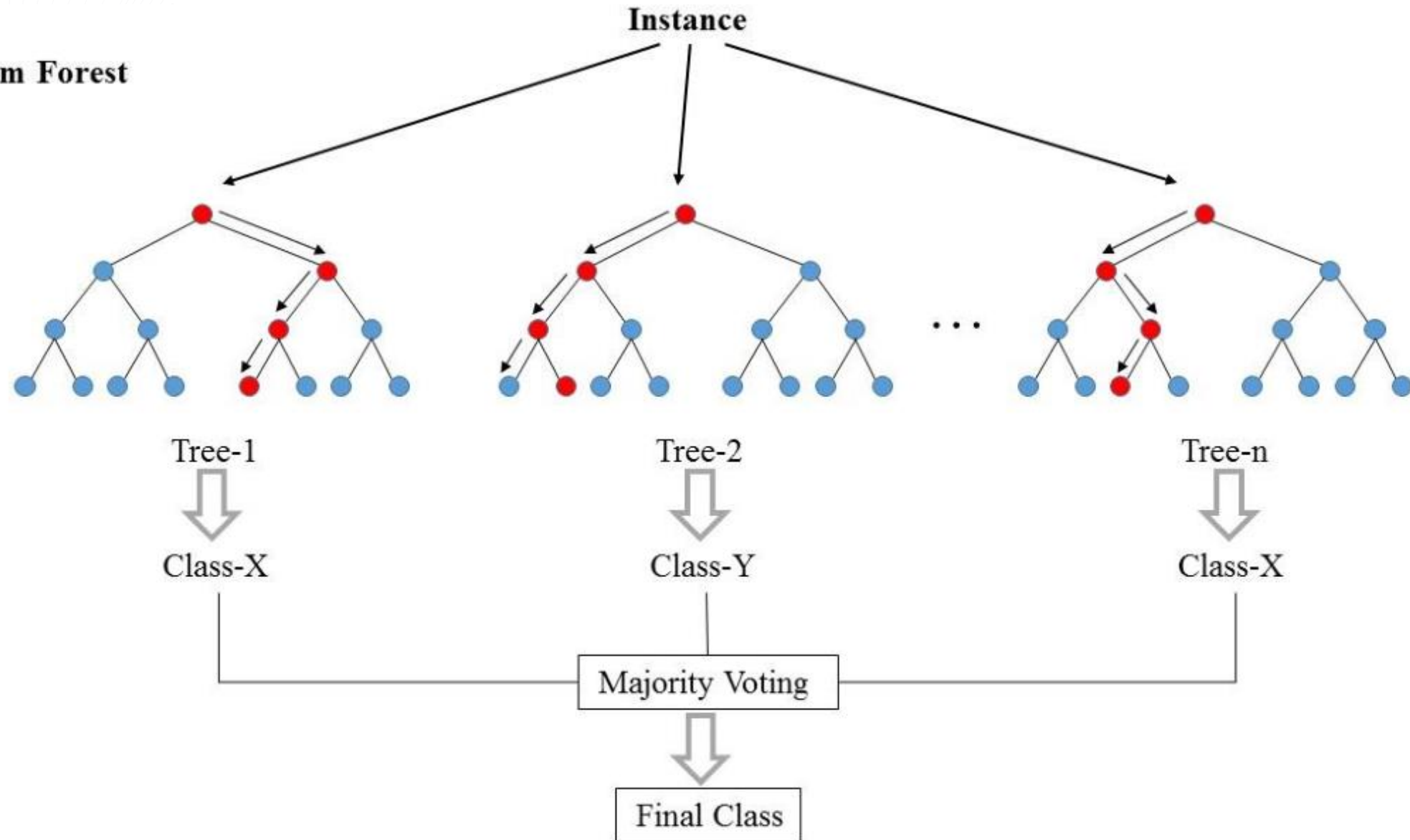
C	V	C	V	C	V	V	C	.	C	V	C
---	---	---	---	---	---	---	---	---	---	---	---

N	Grams	#
1	V, C, N, S	4
2	VV, VC, VS, ... , CV, CC, CN, ... , SS	16
3	VVV, VVC, VVS, ... , SSC, SSN, SSS	64
4	VVVV, VVVC, ... , SSSN, SSSS	256

Manejando el “Overfitting”



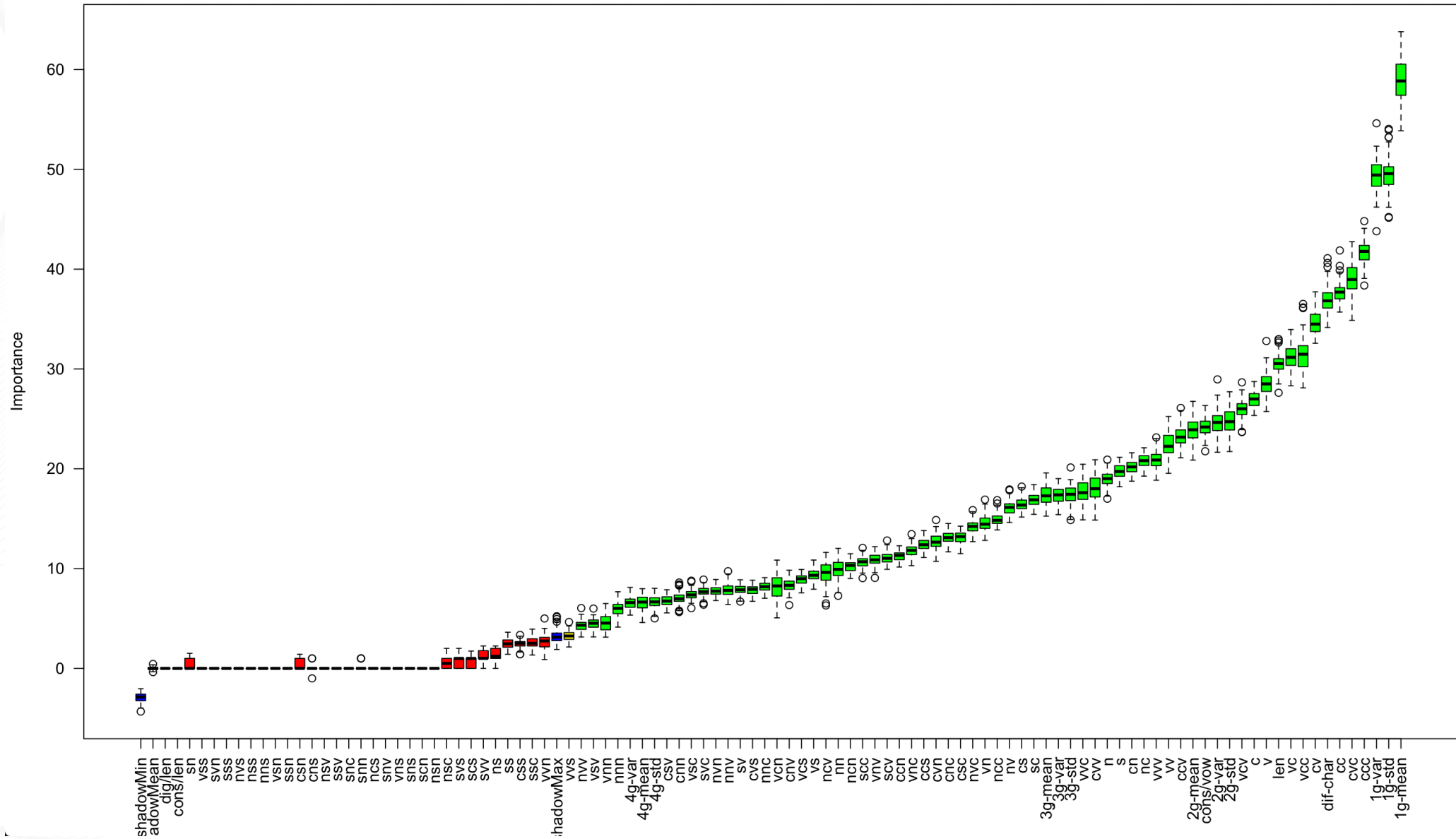
Random Forest



Resultados

	1-grams	2-grams	3-grams	4-grams
<i>No. of features</i>	22	34	82	274
<i>Training time</i>	0.65 h	1.21 h	4.21 h	47.74 h
<i>Testing time</i>	1 s	1 s	2 s	4 s
<i>Accuracy</i>	0.9890	0.9891	0.9879	0.9873
<i>Kappa</i>	0.9780	0.9783	0.9758	0.9747
<i>Sensitivity</i>	0.9896	0.9859	0.9857	0.9848
<i>Specificity</i>	0.9884	0.9924	0.9901	0.9899
<i>Clean prediction</i>	0.9884	0.9924	0.9900	0.9899
<i>DGA prediction value</i>	0.9896	0.9860	0.9858	0.9848

Variable Importance



Selección de Características

	2-grams	Top 25	Top 20	Top 15	Top 10
<i>Features</i>	34	25	20	15	10
<i>Training time</i>	1.21 h	1.18 h	0.82 h	0.63 h	0.44 h
<i>Accuracy</i>	0.9891	0.9865	0.9865	0.9873	0.9832
<i>Kappa</i>	0.9783	0.9730	0.9730	0.9747	0.9664
<i>Sensitivity</i>	0.9859	0.9851	0.9848	0.9859	0.9827
<i>Specificity</i>	0.9924	0.9880	0.9883	0.9888	0.9837
<i>Clean prediction</i>	0.9924	0.9879	0.9882	0.9888	0.9836
<i>DGA prediction</i>	0.9860	0.9851	0.9848	0.9859	0.9828

Mean (1-gram)	Variance (1-gram)	Standard deviation (1-gram)
Standard deviation (2-gram)	Different characters	String length
'ccc'	'cvc'	'vcc'
'vcv'	'cv'	'vc'
'cc'	'c'	'v'

Ejemplos

	1-Gram Mean	1-Gram Std	1-Gram Var	2-Gram Std
facebook.com	1.33	0.71	0.50	0.00
wxhyqqrbouru.pw	1.36	0.50	0.25	0.00

	Different #chars	Length
facebook.com	9	12
wxhyqqrbouru.pw	11	15

Domain name	Masked domain	ccc	cvc	cc	cv	vcc	vc	v	c	vcv
facebook.com	cvcvcvvc.cvc	0	3	0	4	0	4	5	6	2
wxhyqqrbouru.pw	cccccccvvcv.cc	6	0	8	2	0	1	3	11	1

Índice

~~1. Introducción~~

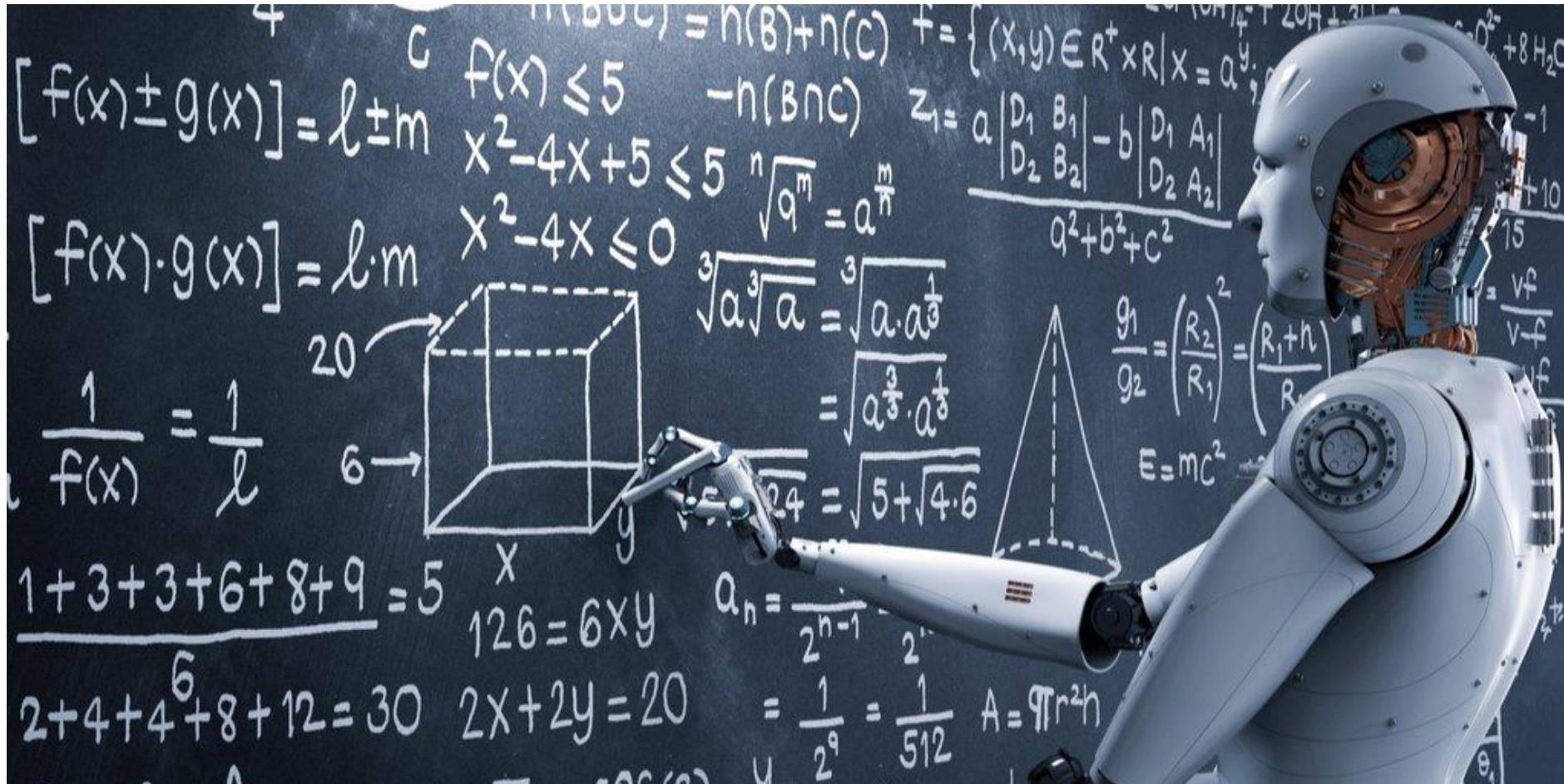
~~2. Detectando DGA Malware~~

3. Project AVA & Outlier Detection

4. Lecciones Aprendidas

5. Preguntas

Project AVA – Anomaly detection



<https://www.nccgroup.trust/uk/about-us/newsroom-and-events/blogs/2019/june/project-ava-on-the-matter-of-using-machine-learning-for-web-application-security-testing-part-1-understanding-the-basics-and-what-platforms-and-frameworks-are-available/>

Conjunto

Decoder Comparer
View Logs Filter Library Gre

Log From

☐ In scope items only

☐ All Tools

☐ Spider

☒ Intruder

☐ Scanner

☐ Repeater

☐ Sequencer

☐ Proxy

☐ Target

☐ Extender

Elastic Search

Address:

Cluster Name:

Index:

Upload Delay:

Include Request and Response

logger-facebook-lo...

Selected fields

? _source

Available fields

t _id

t _index

_score

t _type

t host

t method

t newcookies

t path

t protocol

t referrer

t requestbody

t requestcontentt...

requestlength

○ requesttime

t responsebody

t responsedelay

responselength

○ responsetime

_source

protocol: https method: POST host: www.facebook.com path: /login/device-based/regular/login/ requesttime: February 8th 2019, 18:09:55.000 responsetime: February 8th 2019, 18:09:55.000 responsedelay: 356ms status: 200
title: newcookies: act=deleted; fr=1R0VQNfxhXaLpZoUg..BcXWVL.2u.AAA.0.0.BcXanT.AWufz3r4; sentcookies: datr=S2VdXPUQ0iYVcjtd1kaI0Rz6; reg_fb_ref=https%3A%2F%2Fwww.facebook.com%2Flogin%2Fdevice-based%2Fregular%2Flogin%2F%3Flogin_attempt%3D1%26lwv%3D110; reg_fb_gate=https%3A%2F%2Fwww.facebook.com%2F; fr=1R0VQNfxhXaLpZoUg..BcXWVL.2u.AAA.0.0

Table JSON

[View single document](#)

t _id	VwXfzWgBantJRNaTabPR
t _index	logger-facebook-login-1
# _score	1
t _type	requestresponse
t host	www.facebook.com
t method	POST
t newcookies	act=deleted; fr=1R0VQNfxhXaLpZoUg..BcXWVL.2u.AAA.0.0.BcXanT.AWufz3r4;
t path	/login/device-based/regular/login/
t protocol	https
t referrer	https://www.facebook.com/login/device-based/regular/login/?login_attempt=1&lwv=110
t requestbody	POST /login/device-based/regular/login/?login_attempt=1&next=https%3A%2F%2Fwww.facebook.com%2F&lwv=120&lwv=120&lwv=120 HTTP/1.1 Host: www.facebook.com User-Agent: Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:65.0) Gecko/20100101 Firefox/65.0 Accept: text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8 Accept-Language: en-GB,en;q=0.5 Accept-Encoding: gzip, deflate Referer: https://www.facebook.com/login/device-based/regular/login/?login_attempt=1&lwv=110 Content-Type: application/x-www-form-urlencoded Content-Length: 580 Connection: close

Características

host

method

newcookies

path

protocol

referrer

requestbody

requestcontenttype

requestlength

requesttime

responsebody

responsedelay

responselength

responsetime

sentcookies

status

title

Petición Referencia

Request ▲	Payload	Status	Error	Timeout	Length	
0		200	☐	☐	4829	▲
1	'	200	☐	☐	463	
2	a' or 1=1--	200	☐	☐	461	
3	"a"" or 1=1--"	200	☐	☐	4770	
4	or a = a	200	☐	☐	4770	
5	a' or 'a' = 'a	200	☐	☐	5127	
6	1 or 1=1	200	☐	☐	4836	
7	a' waitfor delay '0:0:10'--	200	☐	☐	485	
8	1 waitfor delay '0:0:10'--	200	☐	☐	470	
9	declare @q nvarchar (200) sel...	200	☐	☐	4770	
10	declare @s varchar(200) select...	200	☐	☐	4770	
11	declare @q nvarchar (200) 0x7...	200	☐	☐	4770	
12	declare @s varchar (200) selec...	200	☐	☐	4770	
13	a'	200	☐	☐	464	
14	?	200	☐	☐	4770	
15	' or 1=1	200	☐	☐	461	▼

Ingeniería de Datos

	1xx		2xx		3xx		4xx		5xx	
200	0	0	1	0	0	0	0	0	0	0
302	0	0	0	0	1	2	0	0	0	0
404	0	0	0	0	0	0	1	4	0	0

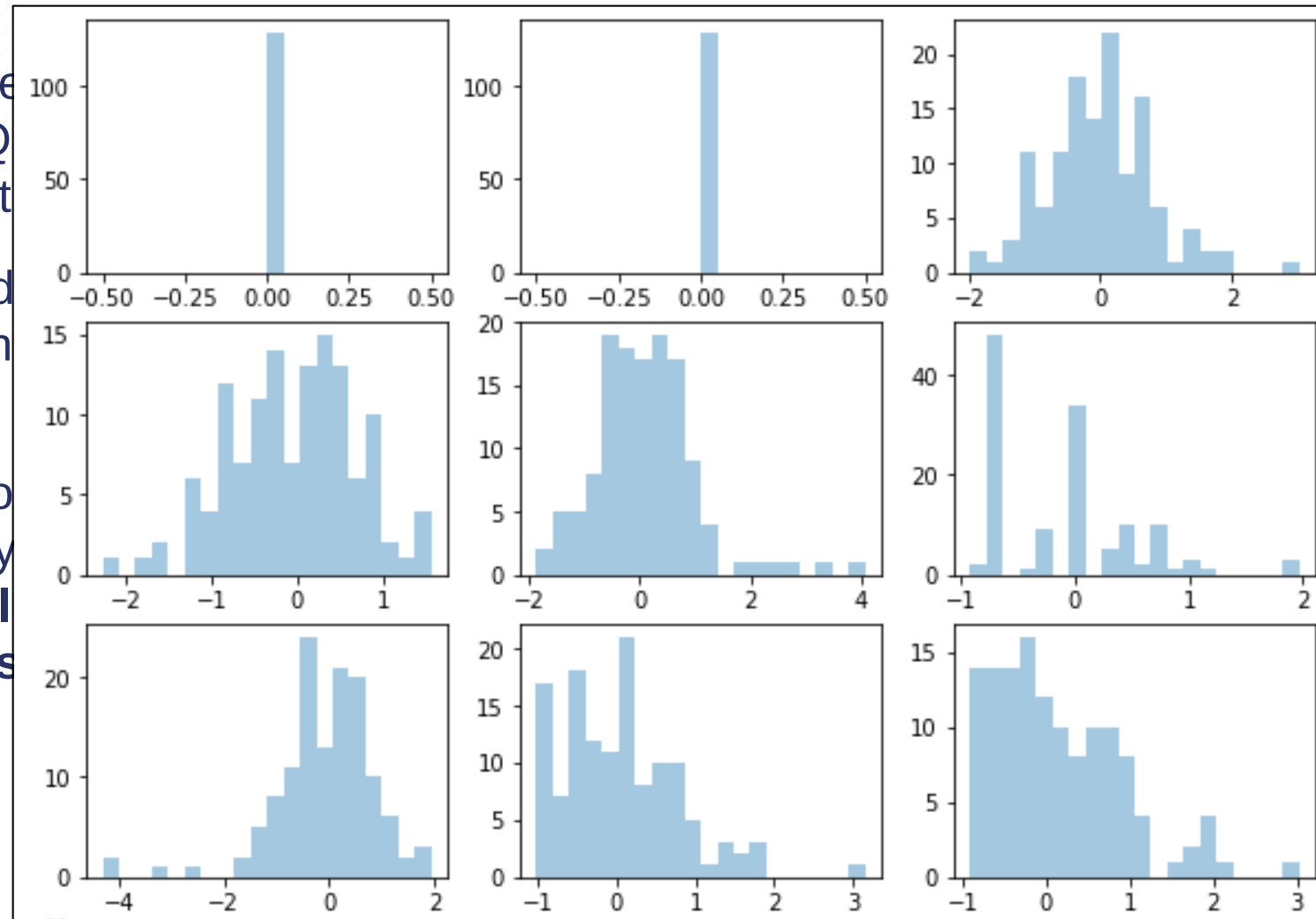
Status	Response Size	Lines in Resp Diff	Avg Line Length	Diff Resp Delay	Body Diff Statistics
Status Encoded	Response Size Diff	Biggest Line Length	Diff Avg Line Length	Response Body Stats	
Status Difference	Lines in Response	Diff Biggest Line Length	Response Delay	Diff Resp Body Stats	

Normalización: Robust Scaler

This Scaler re
(defaults to IQ
quantile) and t

Centering and
statistics on th
to be used on

Standardization
estimators. Ty
However, **outl**
In such cases

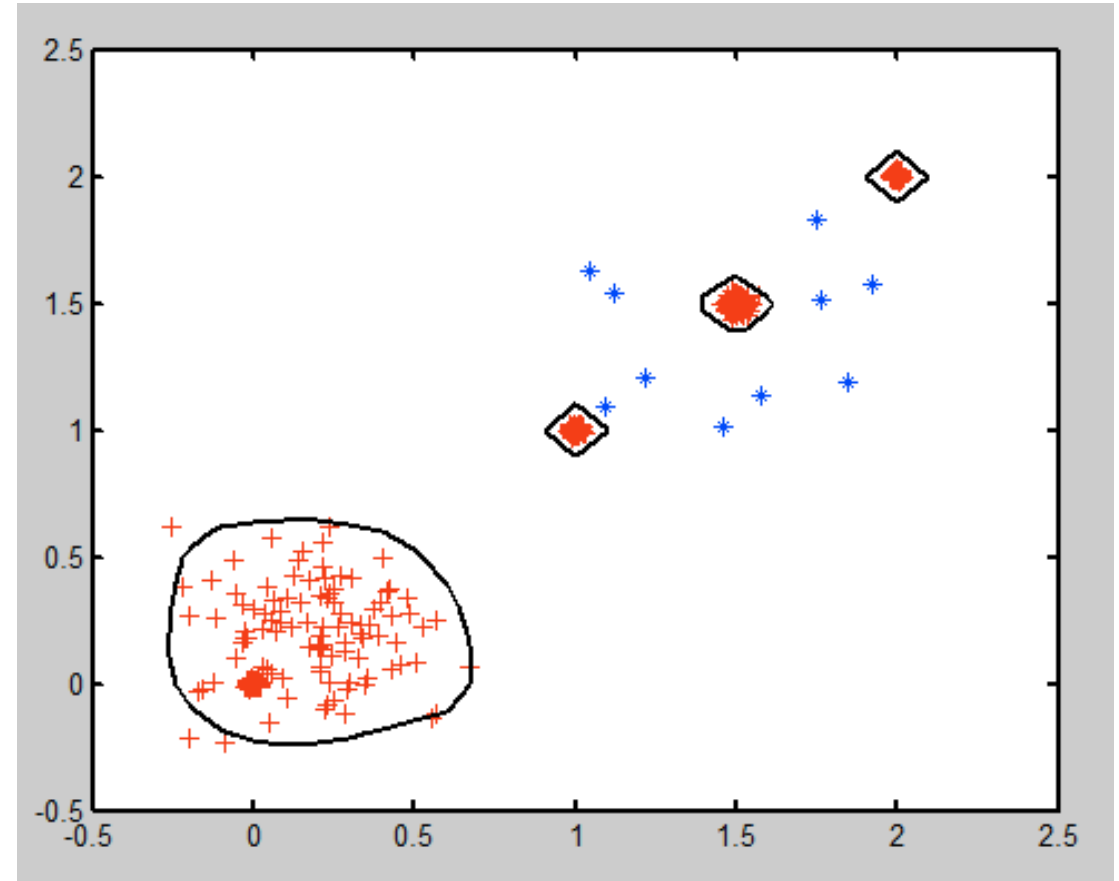
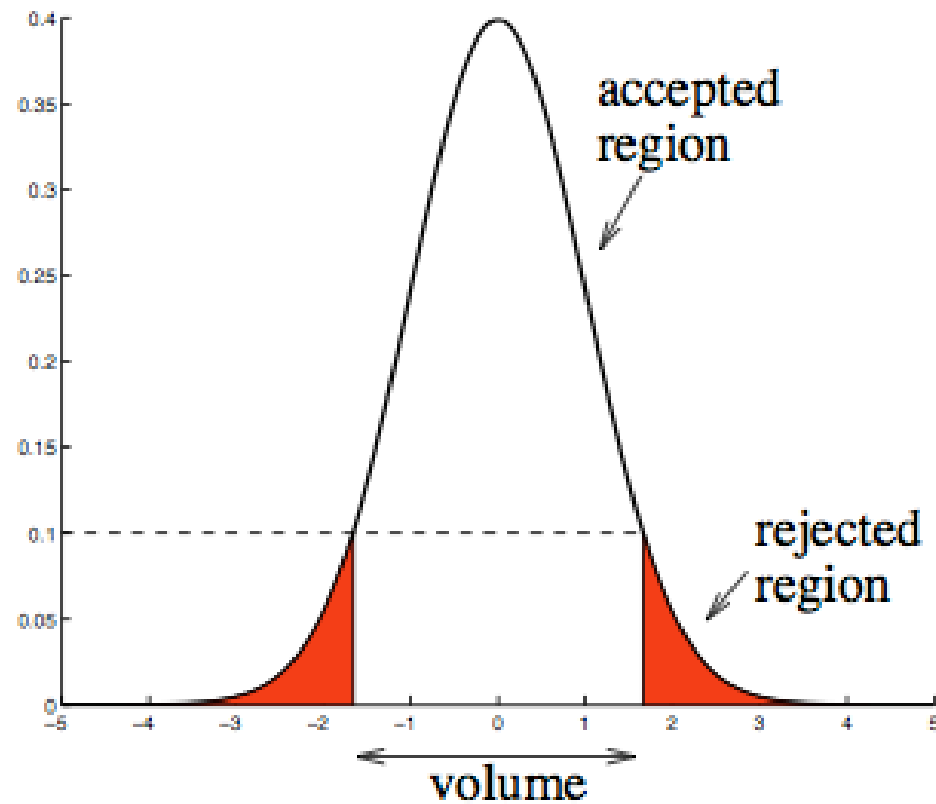


quantile range
quartile (25th

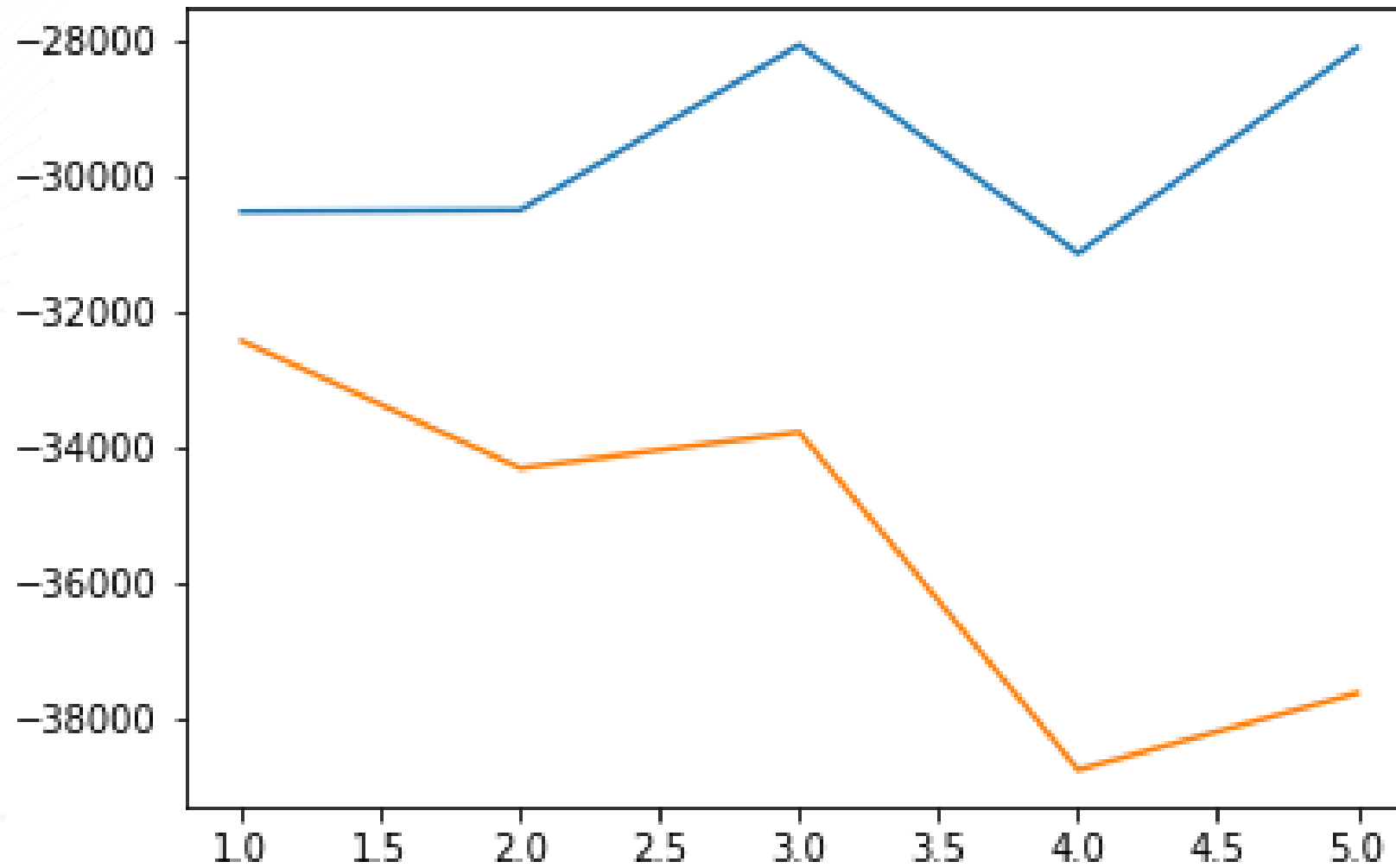
the relevant
e then stored

line learning
unit variance.
negative way.
results.

Gaussian Mixture Model (GMM)



BIC Vs AIC



Índice

~~1. Introducción~~

~~2. Detectando DGA Malware~~

~~3. Project AVA & Outlier Detection~~

4. Lecciones Aprendidas

5. Preguntas

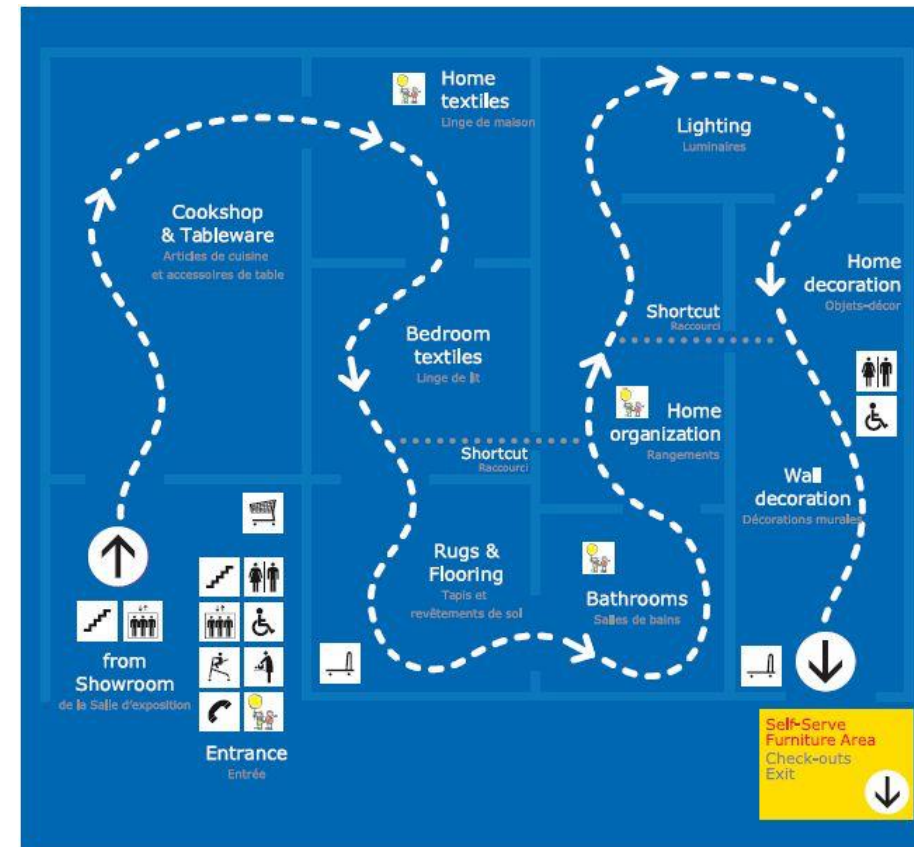
Céntrate en el Problema



Extracción de Características



- Length
- # Vowels
- # Consonants
- # Numbers



No siempre la mejor solución

G	O	O	G	L	E	.	C	O	M		
---	---	---	---	---	---	---	---	---	---	--	--

- Length
- # Vowels
- # Consonants
- # Numbers

if length < 15:
return CLEAN

if numbers > 5:
return MALWARE

$x = \text{vowels} / \text{consonants}$

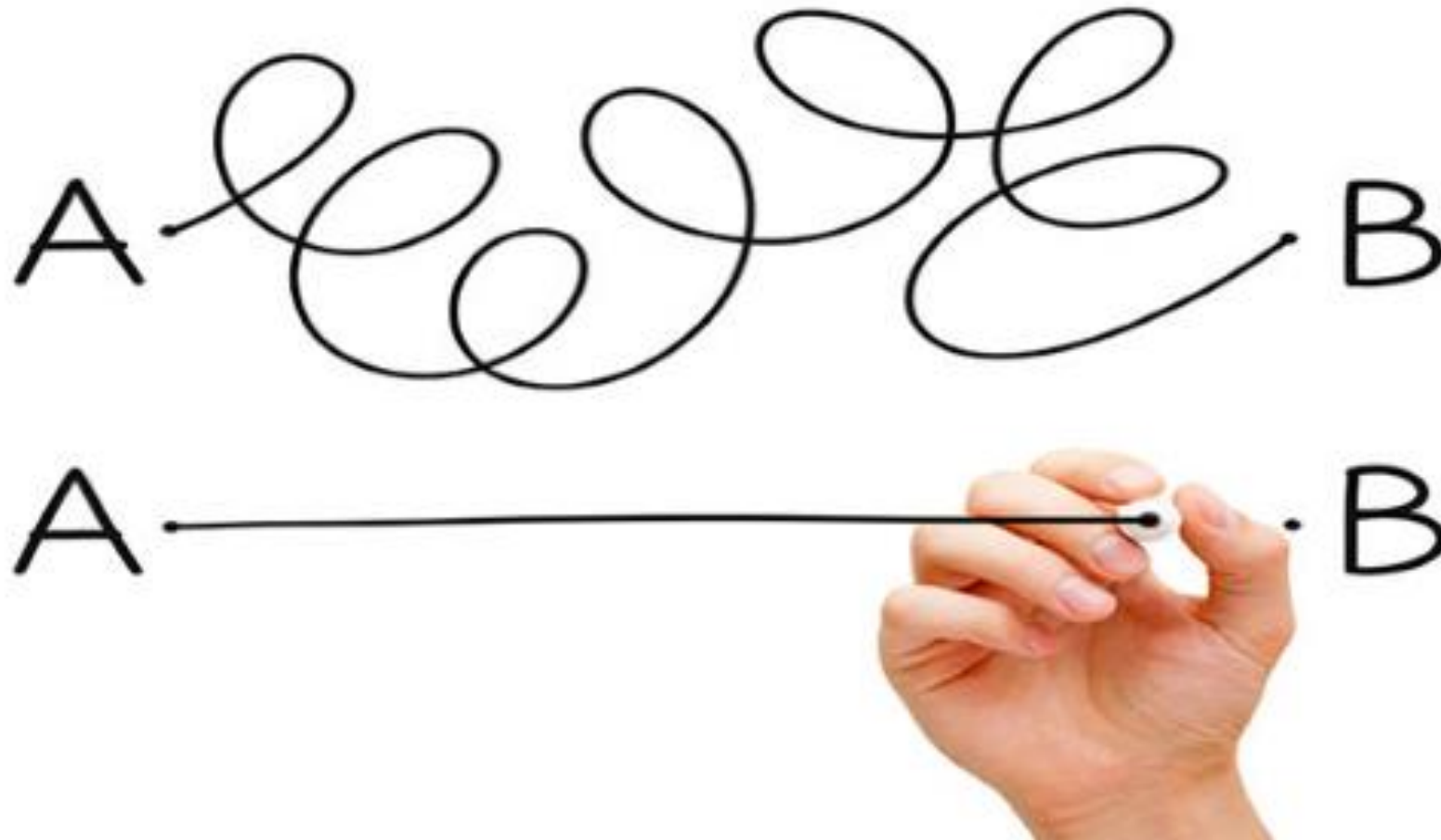
if $x < 0.2$:
return CLEAN

return MALWARE

Prepara los Datos



No te compliques



No te frustres

Me cago en su puta madre de Facebook
enseñándome anuncios para mujeres maduras.

Translated from Spanish by  bing

[Wrong translation?](#)

Fuck off Facebook, son of a bitch, showing me ads for mature women

LIKES

5



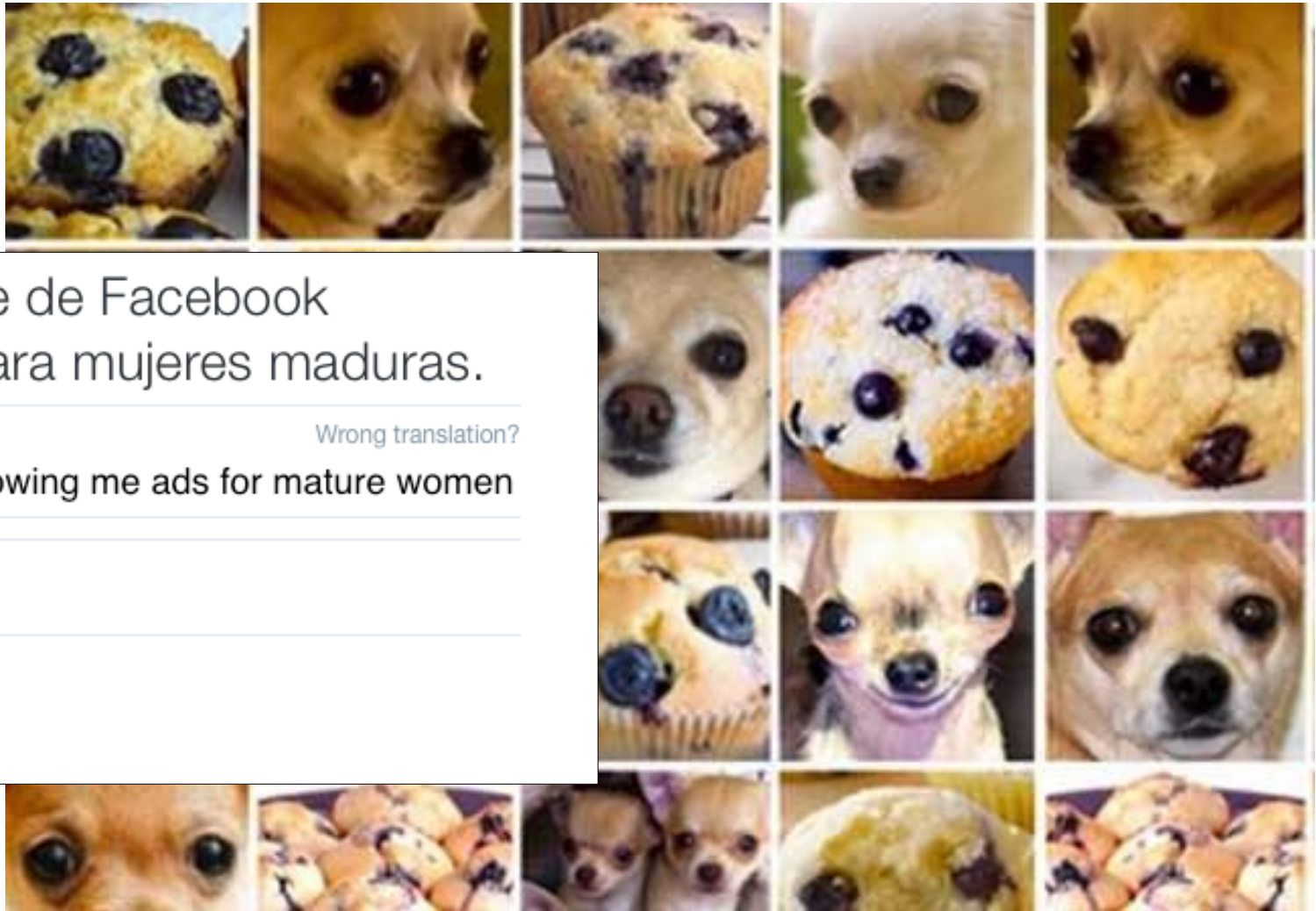
10:50 AM - 8 Sep 2015



2



5





jose.selvi@nccgroup.com



@JoseSelvi

¿PREGUNTAS?



COMUNIDAD Y CONFIANZA,
BASES DE NUESTRA CIBERSEGURIDAD

#XIIIJORNADASCNCERT

OC.CCN.CNI.ES

WWW.CCN.CNI.ES

WWW.CCN-CERT.CNI.ES

**XIII
JORNADAS
STIC
CCN-CERT**

ccn-cert
centro criptológico nacional