# 機器學習期末報告

主題: PhiUSIIL 釣魚網站 URL 的準確預測與開發

組員:葉宇倫、楊書暐

# 摘要

本研究使用 UCI machine learning repository 上的 PhiUSIIL Phishing URL (Website)資料集。第一部分是資料的探索式分析,第二部分是使用不同變數與整個資料集的建模預測,以用來與最後一個部分的演算法建模預測做比較。

# 第一章 緒論

#### 第一節 研究動機與目的

隨著新冠疫情的爆發與全球大流行,各國政府、企業與個人紛紛加速採用數位轉型來應對社會運作與經濟活動的中斷。這一過程促使線上交易、遠距工作、虛擬會議與電子商務等需求激增,網路空間的重要性與依賴性也隨之大幅提高。然而,隨著網路活動的增加,釣魚網址的攻擊頻率亦呈現驚人的上升趨勢,對於個人隱私、財務安全與企業機密構成了前所未有的威脅。尤其是在疫情期間,犯罪者利用大眾對疫情資訊、政府補助或疫苗接種等緊急需求的渴望,進一步提高了釣魚網站的成功率,甚至針對弱勢族群和初次使用數位工具的用戶進行攻擊,造成了廣泛的社會與經濟損失。因此,為有效防範此類網路威脅,本研究旨在深入探討並尋找釣魚網站篩選的有效特徵以及探討各類特徵對釣魚網站的影響,以提升偵測準確率,從而降低網路安全風險。本研究的成果將為未來網路安全防護措施的制定提供重要參考,進一步保障數位時代中用戶的安全與信任。

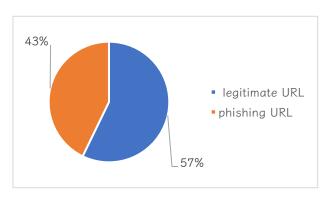
### 第二節 研究流程

我們先進行探索式資料分析找出可能是重要的特徵去建模,再建立非人造變 數資料集與人造變數資料集(原資料集)各兩個模型,共三個模型去做預測比較, 最後與論文內演算法建模做比較,並總結一些成果,提出進一步的建議。

# 第二章 探索式資料分析

#### 第一節 變數介紹與總結

本資料集有 235798 筆觀察值,56 個特徵,以及目標二元變數 label,0 表示釣魚 URL,1 代表非釣魚 URL。該資料集無缺失值,且 label 大致上是均衡的,見下方圖一。



圖一、label 分布圓餅圖

變數名稱	說明	變數名稱	說明	變數名稱	說明	變數名稱	說明
HasObfusc	是否存在混	HasPassw	是否存在密	NoOfObfu	混淆字元的	label	是否為釣魚
ation	淆	ordField	碼欄位	scatedCha	數量		網站
				r			
HasFavico	是否有網站	HasCopyri	是否包含版	NoOfAmpe	URL 中的	NoOfOther	URL 中其
n	圖標	ghtInfo	權資訊	rsandInUR	「&」符號	SpecialCh	他特殊字元
	(Favicon)			L	數量	arsInURL	的數量
HasDescri	是否包含描	HasTitle	是否有標題	NoOfLette	URL 中字	NoOfURLR	URL 的重
ption	述			rsInURL	母的數量	edirect	導次數
HasExtern	是否使用外	IsDomainI	是否使用	NoOfDegit	URL 中數	NoOfSelfR	自我重導的
alFormSub	部表單提交	P	IP 當作域	sInURL	字的數量	edirect	次數
mit			名				
HasSocial	是否與社交	IsHTTPS	是否使用	NoOfEqual	URL 中等	NoOfEmpt	空引用的數
Net	網絡相關		HTTPS	sInURL	號(=)的數	yRef	量
			100 3 3		量		
HasSubmit	是否有提交	IsResponsi	是否響應式	NoOfSelfR	自引用的次	NoOfPopu	彈出視窗的
Button	按鈕	ve		ef	數	р	數量
HasHidden	是否存在隱	NoOfSubD	子域名的數	NoOfQMar	URL 中間	NoOfiFram	iFrame 的
Fields	藏欄位	omain	量	kInURL	號(?)的數	е	數量
N. 007	同日典日	1	<b>工</b> /	11010: 11	量	C . T	14. 4. 4. 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
NoOfImag	圖片數量	TLD	頂級域名	URLSimilar	URL 相似	DomainTitl	域名與標題
е			( Top	ityIndex	性指標	eMatchSc	的匹配分數
			Level			ore	
NoOfCSS	CCC 操安	Title	Domain) 標題	CharConti	字符延續率	URLTitleM	URL 與標
1001033	CSS 檔案 的數量	Title	<b>小示</b>	CharConti nuationRat	子们 是順平	atchScore	題的匹配分
	可致里					atchiscore	數
NoOfJS	JavaScript	URLLength	URL 的長	e TLDLegiti	頂級域名的	Obfuscatio	混淆比例
1400100	檔案的數量	OKLLENGTH	度	mateProb	合法概率	nRatio	101/4 NO [/]
NoOfExter	外部引用的	DomainLe	域名的長度	URLCharP	URL 字符	Bank	與銀行相關
nalRef	數量	ngth	/\ _ = \ \ \ /\ \	rob	的概率		的標記
FILENAME	文件名稱	TLDLength	頂級域名的	LetterRati	URL 中字	Pay	是否與支付
		3	長度	oInURL	母比例		相關
URL	URL 本身	LineOfCod	代碼行數	DegitRatio	URL 中數	Crypto	與加密相關
		е		InURL	字比例		的標記
Domain	域名	LargestLin	最長代碼行	SpacialCh	URL 中特	Robots	是否與機器
		eLength	的長度	arRatioInU	殊字符比例		人文件相關
				RL			

類別變數										
變數名稱	唯一值									
FILENAME	235795									
URL	235370									
Domain	220086									
TLD	695									
Title	197874									

	數值	變數 及 二元	變數	
變數名稱	平均	標準差	最小值	最大值
URLLength	34.5731	41.3142	13	6097
DomainLength	21.4704	9.1508	4	110
IsDomainIP	0.0027	0.0519	0	1
URLSimilarityIndex	78.4308	28.9761	0.155574	100
CharContinuationRate	0.8455	0.2166	0	1
TLDLegitimateProb	0.2604	0.2516	0	0.522907
URLCharProb	0.0557	0.0106	0.001083	0.090824
TLDLength	2.7645	0.5997	2	13
NoOfSubDomain	1.1648	0.601	0	10
HasObfuscation	0.0021	0.0453	0	1
NoOfObfuscatedChar	0.0249	1.8762	0	447
ObfuscationRatio	0.0001	0.0038	0	0.348
NoOfLettersInURL	19.4289	29.0903	0	5191
LetterRatioInURL	0.5159	0.1233	0	0.926
NoOfDegitsInURL	1.881	11.8867	0	2011
DegitRatioInURL	0.0286	0.0709	0	0.684
NoOfEqualsInURL	0.0622	0.9347	0	176
NoOfQMarkInURL	0.0294	0.1935	0	4
NoOfAmpersandInURL	0.0251	0.8364	0	149
NoOfOtherSpecialCharsInU RL	2.3402	3.5276	0	499
SpacialCharRatioInURL	0.0633	0.0324	0	0.397
IsHTTPS	0.7826	0.4125	0	1
LineOfCode	1141.9	3419.951	2	442666
LargestLineLength	12789.53	152201.1	22	13975732

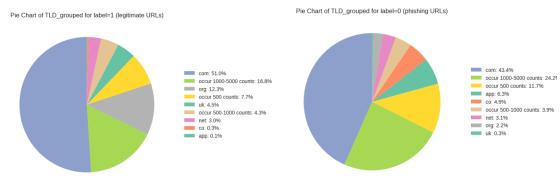
HasTitle	0.8613	0.3457	0	1
DomainTitleMatchScore	50.1314	49.677	0	100
URLTitleMatchScore	52.1221	49.6006	0	100
HasFavicon	0.3618	0.4805	0	1
Robots	0.2665	0.4422	0	1
IsResponsive	0.6245	0.4842	0	1
NoOfURLRedirect	0.1334	0.34	0	1
NoOfSelfRedirect	0.0401	0.1962	0	1
HasDescription	0.4402	0.4964	0	1
NoOfPopup	0.2218	3.8705	0	602
NoOfiFrame	1.5886	5.7626	0	1602
HasExternalFormSubmit	0.044	0.2051	0	1
HasSocialNet	0.4566	0.4981	0	1
HasSubmitButton	0.4143	0.4926	0	1
HasHiddenFields	0.3778	0.4848	0	1
HasPasswordField	0.1023	0.303	0	1
Bank	0.1271	0.3331	0	1
Pay	0.237	0.4252	0	1
Crypto	0.0235	0.1514	0	1
HasCopyrightInfo	0.4868	0.4998	0	1
NoOfImage	26.0757	79.4118	0	8956
NoOfCSS	6.3331	74.8663	0	35820
NoOfJS	10.5223	22.3122	0	6957
NoOfSelfRef	65.0711	176.6875	0	27397
NoOfEmptyRef	2.3776	17.6411	0	4887
NoOfExternalRef	49.2625	161.0274	0	27516
label	0.5719	0.4948	0	1

表一、變數介紹表格

## 第二節 數據圖片與特徵發現

因為其他類別變數個數太多,因此只留下不同類別數僅有 695 的 TLD,再根據他們的 count 進行分類編碼,並使用卡方檢定得到這些類別非獨立,建模時會使用編碼過的 TLD 做成新特徵取代原本 TLD 變數。

見圖二與三、編碼後合法與釣魚網站 TLD 分布以及卡方檢定



Chi-squared inpendence test									
Chi-squared statistic	27578.6415								
P-value	0.0								
Degrees of freedom	8								

接著,觀看二元特徵的 count plot,可以觀察到,某些類別變數在預測釣魚網站時具有較高的區分能力,在建模部分可能會抓取一些特徵來建模。 見圖五、二元變數的 count plot



然後,我們觀看到各類數值特徵的相關係數圖,可以看出 URLLength 與 NoOfLettersInURL 有強線性關係,NO 家族關於 URL 的變數與 URLLength 彼此線性關係挺強的,可能有共線性的特徵我們只會擇一來建模。 見圖六、數值變數相關係數圖

	Correlation Coefficient Heatmap for not derived																				
URLLength	1.00	0.24			0.46	0.96		0.69	0.47	0.50											-0.06
DomainLength	0.24	1.00	0.15	0.40	0.01	0.28	0.13		0.06	0.02	0.27							-0.10	-0.10		-0.07
TLDLength	0.03	0.15	1.00	-0.16							-0.01										-0.03
NoOfSubDomain	0.07	0.40	-0.16	1.00							0.16										-0.00
NoOfObfuscatedChar	0.46				1.00	0.24	0.72				0.57										-0.00
NoOfLettersInURL	0.96	0.28			0.24	1.00	0.65	0.49	0.46	0.33	0.65							-0.09	-0.08		-0.07
NoOfDegitsInURL	0.84	0.13			0.72	0.65	1.00		0.32	0.66											-0.05
NoOfEqualsInURL	0.69					0.49	0.81	1.00	0.43	0.69	0.78										-0.02
NoOfQMarkInURL	0.47					0.46	0.32	0.43	1.00	0.11	0.47										-0.04
NoOfAmpersandInURL	0.50	0.02				0.33	0.66	0.69	0.11	1.00	0.54										-0.01
NoOfOtherSpecialCharsInURL	0.78	0.27		0.16	0.57	0.65			0.47	0.54	1.00	-0.10			-0.08	-0.10		-0.12	-0.11	-0.04	-0.09
LineOfCode	-0.06										-0.10	1.00	0.01		0.22	0.40	0.14	0.20	0.40	0.13	0.31
LargestLineLength	0.05												1.00	-0.00							-0.00
NoOfPopup	-0.01													1.00	0.02						0.06
NoOfiFrame	-0.04											0.22			1.00	0.24	0.03	0.16	0.13		0.09
NoOfImage	-0.06										-0.10	0.40			0.24	1.00	0.25	0.15	0.30	0.12	0.21
NoOfCSS	-0.01											0.14				0.25	1.00	0.06			0.03
NoOfJS	-0.08	-0.10									-0.12	0.20			0.16	0.15		1.00	0.18		0.15
NoOfSelfRef	-0.07	-0.10									-0.11	0.40			0.13	0.30		0.18	1.00	0.11	0.70
NoOfEmptyRef	-0.02											0.13				0.12			0.11	1.00	0.06
NoOfExternalRef	-0.06	-0.07	-0.03	-0.00	-0.00	-0.07	-0.05	-0.02	-0.04	-0.01	-0.09	0.31	-0.00	0.06	0.09	0.21	0.03	0.15	0.70	0.06	1.00

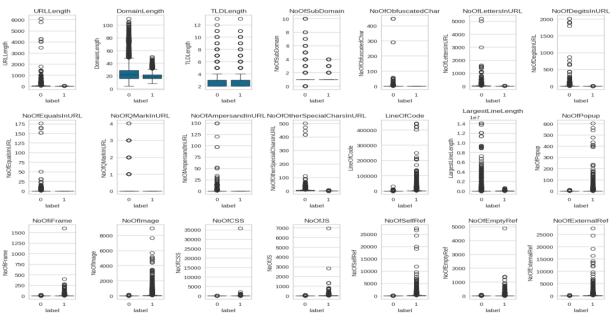
接著,我們觀察數值變數 spearman 相關係數,可以看出 URLLength 與 NoOfLettersInURL 有強非線性關係,URLLength 與 DomainLength 及其他關於 code、網站框架跟 Ref 類型的變數彼此非線性關係蠻強的,可能考慮將他們做一些變換之後建模。

見圖七、數值類變數 spearman 相關係數圖

	Spearman Correlation Coefficient Heatmap for not derived																				
URLLength	1.00	0.80	0.11	0.10	0.07	0.97	0.46	0.25	0.26	0.10	0.62		-0.04	-0.07		-0.30			-0.32		-0.28
DomainLength		1.00	0.17	0.31	0.01		0.22	0.01	0.01	0.01	0.36	-0.13	-0.00	-0.03	-0.10		-0.14			-0.12	-0.13
TLDLength	0.11	0.17	1.00		-0.00	0.14	0.05	0.01	0.00	-0.00		-0.06	0.03	0.00	0.01	-0.06	-0.04	-0.02	-0.06	-0.06	-0.05
NoOfSubDomain	0.10	0.31	-0.24	1.00	0.00	0.02	0.04	-0.02	-0.03	0.03	0.28	0.02	-0.02	0.01	0.01	0.02	0.02	0.00	0.02	0.02	0.02
NoOfObfuscatedChar	0.07	0.01	-0.00	0.00	1.00	0.07	0.10	0.13	0.13	0.74	0.08	-0.02	-0.00	-0.01	-0.02	-0.03	-0.02	-0.03	-0.04	-0.02	-0.04
NoOfLettersInURL	0.97	0.77	0.14	0.02	0.07	1.00	0.36	0.24	0.26	0.10	0.57	-0.29	-0.07	-0.07		-0.32	-0.28	-0.30	-0.34	-0.21	-0.30
NoOfDegitsInURL	0.46	0.22	0.05	0.04	0.10	0.36	1.00	0.31	0.32	0.14	0.59	-0.39	-0.13	-0.10		-0.42	-0.39	-0.41	-0.43		-0.41
NoOfEqualsInURL	0.25	0.01	0.01	-0.02	0.13	0.24	0.31	1.00	0.91	0.26	0.26	-0.13	-0.03	-0.03	-0.06	-0.12	-0.11	-0.12	-0.13	-0.08	-0.14
NoOfQMarkInURL	0.26	0.01	0.00	-0.03	0.13	0.26	0.32		1.00	0.24	0.28	-0.14	-0.04	-0.04	-0.07	-0.13	-0.12	-0.13		-0.08	-0.15
NoOfAmpersandInURL	0.10	0.01	-0.00	0.03	0.74	0.10	0.14	0.26	0.24	1.00	0.11	-0.04	0.01	-0.01	-0.01	-0.04	-0.04	-0.05	-0.05	-0.03	-0.05
NoOfOtherSpecialCharsInURL	0.62	0.36		0.28	0.08	0.57	0.59	0.26	0.28	0.11	1.00	-0.45		-0.12	-0.32	-0.48	-0.44	-0.48	-0.49	-0.30	-0.47
LineOfCode		-0.13	-0.06	0.02	-0.02	-0.29	-0.39	-0.13	-0.14	-0.04	-0.45	1.00		0.20	0.59				0.84	0.54	0.84
LargestLineLength	-0.04	-0.00	0.03	-0.02	-0.00	-0.07	-0.13	-0.03	-0.04	0.01	-0.21	0.67	1.00	0.13	0.44	0.59	0.61	0.65	0.60	0.35	0.67
NoOfPopup	-0.07	-0.03	0.00	0.01	-0.01	-0.07	-0.10	-0.03	-0.04	-0.01	-0.12	0.20	0.13	1.00	0.14	0.18	0.20	0.19	0.18	0.12	0.20
NoOfiFrame		-0.10	0.01	0.01	-0.02			-0.06	-0.07	-0.01	-0.32	0.59	0.44	0.14	1.00	0.55	0.48	0.56	0.53	0.34	0.50
NoOfImage	-0.30		-0.06	0.02	-0.03	-0.32	-0.42	-0.12	-0.13	-0.04	-0.48	0.87	0.59	0.18	0.55	1.00	0.76	0.80	0.82	0.53	0.82
NoOfCSS		-0.14	-0.04	0.02	-0.02	-0.28	-0.39	-0.11	-0.12	-0.04	-0.44	0.77	0.61	0.20	0.48	0.76	1.00			0.52	0.79
NoOfJS			-0.02	0.00	-0.03	-0.30	-0.41	-0.12	-0.13	-0.05	-0.48		0.65	0.19	0.56			1.00		0.54	0.80
NoOfSelfRef	-0.32		-0.06	0.02	-0.04	-0.34	-0.43	-0.13		-0.05	-0.49		0.60	0.18	0.53			0.80	1.00	0.51	0.84
NoOfEmptyRef		-0.12	-0.06	0.02	-0.02		-0.27	-0.08	-0.08	-0.03	-0.30	0.54	0.35	0.12	0.34	0.53	0.52	0.54	0.51	1.00	0.51
NoOfExternalRef	-0.28	-0.13	-0.05	0.02	-0.04	-0.30	-0.41	-0.14	-0.15	-0.05	-0.47	0.84	0.67	0.20	0.50	0.82	0.79	0.80	0.84	0.51	1.00

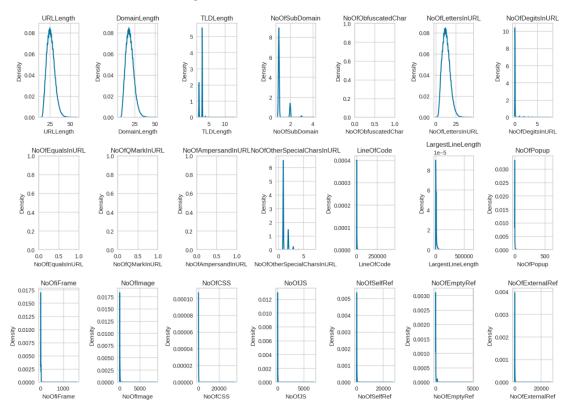
畫出以釣魚網站標籤 0 或 1 為分類的數值類變數盒鬚圖,可以發現大多數數值類變數有許多離群值,除了混淆字源數量、問號數量、連接符號(&)數量在非釣魚網站沒有,且除了 DomainLength 與 TLDLength 的其他數值類變數分配都集中在 0 附近。

見圖八、釣魚網站標籤0或1為分類的數值類變數盒鬚圖



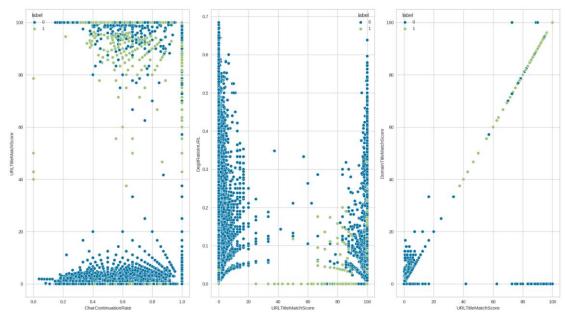
繪製數值類變數兩兩的 histogram,可以看到大多數數值類變數基本上看不出分配形式,或許標準化能減少 outlier 影響。

### 見圖九、數值類變數 histogram



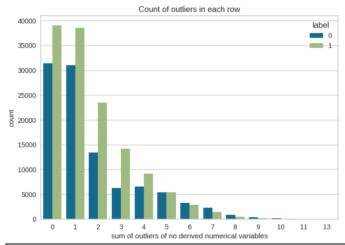
繪製 scatter plot,可以看出某些數值類變數散佈圖受到異常值的扭曲,這 使得線性關係變得不那麼明顯。

### 見圖十、可能重要數值類變數 scatter plot



現在將有 outlier 的每一列對應的行設置成 1,否則 0。將每一列數值類變數有 outlier 的 column 相加並畫出 countplot,以標籤做分類,能看出每列 outliers 比較少的,是合法網站占比更大,最後進行 Mann-Whitney 檢定,說明 釣魚網站與合法網站的數值類變數 outlier 數量分佈有顯著差異。

見圖十一與十二、每列 outliers 數量分類圖以及 Mann-Whitney 檢定結果



Mann-Whitney U statistic i	independence test					
Mann-Whitney U statistic	6681537497					
P-value	4.12*1 <i>e</i> -15					
Degrees of freedom	15					

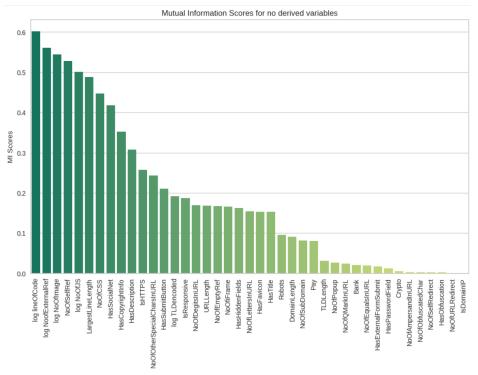
# 第三章 建立模型與變數選擇

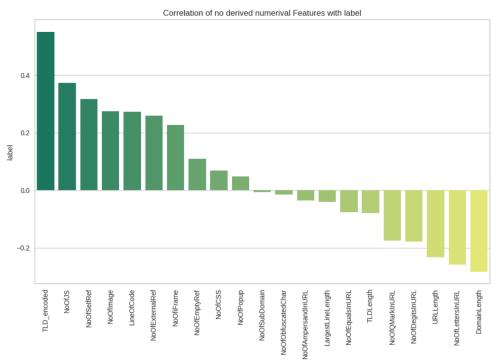
## 第一節 資料處理與變數選取

先將有與 label 有強非線性關係變數 Log 轉換,例如:LineOfCode,然後使用 mutual information 方法來選擇變數。

見圖十三與十四、

非人造變數 mutual information 以及人造變數 mutual information

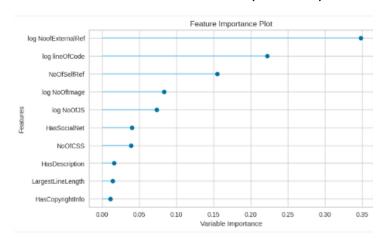




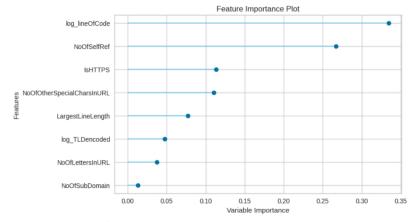
## 第二節 模型建立(3種方法,(TRAIN\_SIZE, TEST\_SIZE)=(0.8,0.2))

我們使用 pycaret 套件標準化特徵後自動選取特徵建模並進行預測比較 見下方圖十五、十六與十七、各模型的特徵重要程度圖

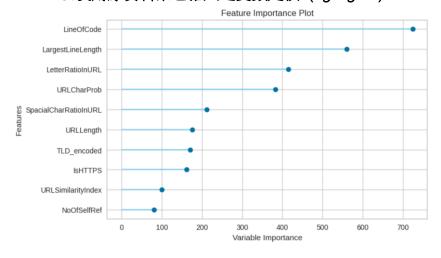
### 1. 使用選擇的 10 個變數建模(隨機森林)



## 2. 使用所有非人造變數加上 URLSimilarityIndex(隨機森林)



## 3.使用原資料集包括人造變數建模 (lightgbm)



### 第三節 小節與表格比較

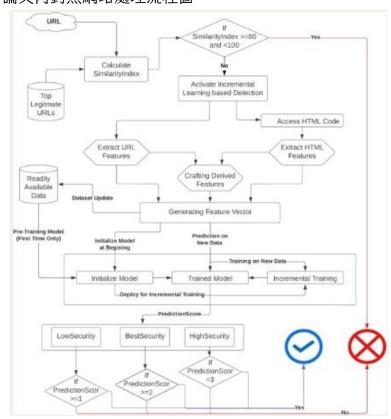
可以發現原本第一種方法所建模的交叉驗證準確度就能達到 99.7%,但第二種方法能提高 0.1%準確度,說明我們找到的 10 個特徵並不是最佳組合。且使用原資料集的人造變數可使得準確度到達 1,說明 URLSimilartyIndex 解釋了更多變數間關係。code 越長以及是 https 相關網站有很高可能性是合法網站。URL 相似性指標確實與論文內容說得一樣是關鍵變數。

見圖十八、三種模型表格比較

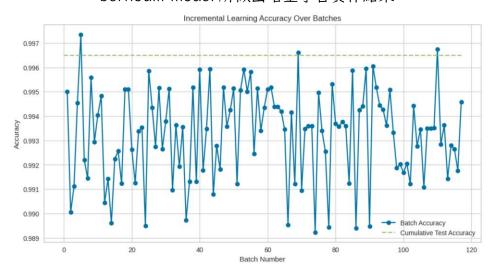
Model	Accuracy	AUC	Recall	Prec.	F1	Карра	MCC	CV	備註
1.隨機森林 分類器	0.9967	0.9999	0.9977	0.9966	0.9971	0.9933	0.9933	0.997	
2.隨機森林 分類器	0.9996	1.0000	0.9996	0.9996	0.9996	0.9991	0.9991	0.999	多了 IsHTTPS
3.LightGBM	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.000	

# 第四章 論文實作

見圖十九、論文內釣魚網站處理流程圖



見圖二十、將原資料集每 2000 筆 data 切成一個 batch,使用 naive bayes bernoulli model 所做出增量學習實作結果



# 第五章 總結

本研究通過對釣魚網站檢測的深入探討,驗證了 URL 相似性指標(URL Similarity Index)的有效性。結果顯示,該指標不僅能顯著提高模型的準確 性,還可以更全面地捕捉釣魚網站與合法網站之間的潛在差異。此外,研究還 發現 HTTPS 的使用和程式碼的長度與合法網站的可能性密切相關,說明這些 特徵在區分釣魚網站和合法網站時具有重要的參考價值。 在模型性能方面,本 研究的模型在靜態資料集上表現卓越,最佳模型達到了 100% 的準確度,展現 了研究框架的理論有效性。然而,當面臨快速變化的網路環境時,模型的適應 性不足成為一個挑戰。靜態學習框架無法及時處理不斷增長的動態數據,因此 每次模型的重新訓練既耗時又資源密集,無法滿足實際應用需求。 針對上述挑 戰,增量學習被認為是未來應對此問題的關鍵技術之一。與傳統的靜態學習方 法不同,增量學習能夠在新的數據到達時進行局部更新,而無需重新訓練整個 模型。這不僅提高了模型的運行效率,還能確保模型在動態網路環境中保持高 準確性。增量學習還適合於處理超大數據集和資源受限場景,尤其是在需要頻 繁更新數據的實務應用中。 未來,研究可以進一步探討增量學習框架的優化, 並結合深度學習技術來提升檢測性能。此外,還可以嘗試應用其他高效的數據 處理技術,例如在線學習和流數據處理,以進一步增強系統的靈活性和實用 性。同時,考慮到釣魚網站攻擊手法的多樣性,未來模型的開發可以融入更多 語義分析、網頁結構特徵等多維度特徵,以提高釣魚網站檢測的全面性和穩健 性。 綜上所述,本研究不僅為釣魚網站檢測提供了有效的框架,還為未來的網 路安全技術發展提供了重要啟示。模型的高準確度和增量學習的潛在應用將在 實務中發揮關鍵作用,特別是在當前網路安全威脅不斷增長的背景下,為實現 更智能、更高效的網路安全防護奠定基礎。

# Reference

[1]PhiUSIIL: A diverse security profile empowered phishing URL detection framework based on similarity index and incremental learning 資料集網址:

https://archive.ics.uci.edu/dataset/967/phiusiil+phishing+url+dataset