Università degli Studi di Salerno

Dipartimento di Informatica

Corso di Laurea Magistrale in Informatica

Statistica ed Analisi dei Dati

Analisi e Rilevamento degli URL di Phishing

Autore

Simone D'Assisi

mat. 0522502038

Anno Accademico 2024/2025

Indice

1 Introduzione

Per *phishing* si intende un tipo di attacco informatico che consiste nel tentativo da parte di un malintenzionato di spingere un utente a rivelare informazioni sensibili (come password, dati bancari, ecc.) simulando una fonte affidabile. Rientra nella categoria delle tecniche di *social engineering*, utilizzate per la manipolazione della psicologia umana, in cui rientrano anche la falsificazione, il depistaggio o la menzogna.

Il phishing costituisce un metodo semplice, economico ed efficace. Gli indirizzi e-mail delle vittime sono facili da reperire e l'invio delle mail stesse non ha costi. Con poco sforzo e praticamente nessun costo, i cybercriminali possono entrare in possesso di dati sensibili. Chi cade vittima di phishing può subire il furto di identità, la perdita di dati, o venire infetto da malware, inclusi i ransomware.

I cybercriminali fanno leva su due fronti: la paura e il senso di urgenza. Una pratica molto frequente è, ad esempio, informare gli utenti che il loro account è stato bloccato o verrà sospeso se si ignora l'email. La paura porta gli utenti presi di mira a non far caso ai segnali rivelatori di un tentativo di phishing in atto, dimenticandosi di ciò che hanno imparato sul phishing. Nelle email di phishing, i cybercriminali utilizzano tre meccanismi principali per impossessarsi delle informazioni dell'utente: link malevoli, allegati dannosi e moduli di compilazione falsi.

2 Analisi del Dataset

2.1 Obiettivi della Prima Fase del Progetto

Pare chiaro che il problema del phishing costituisca un serio fattore di rischio durante la navigazione, che può causare una seria perdita dei propri dati personali. L'obiettivo della prima fase di questo progetto è quindi quello di analizzare un dataset contenente un elevato quantitativo di URL (sia di phishing che sicuri) per individuare pattern e relazioni tra feature al fine di costruire un modello in grado di prevedere se un URL è legittimo o di phishing. Le domande di ricerca che hanno dato la direzione al progetto sono le seguenti:

- **RQ1**: quali caratteristiche distinguono un URL legittimo da uno di phishing?
- **RQ2**: quali variabili hanno un maggiore valore predittivo?
- **RQ3**: è possibile creare un modello affidabile per la classificazione?

2.2 Descrizione del Dataset

Il dataset su cui verrà svolto il lavoro di analisi statistica è stato reperito dal sito **UC Irvine Machine Learning Repository**; nel dataset esaminato sono stati valutati 23.580 URL, ognuno dei quali è caratterizzato da 56 feature diverse (Tabella 1). Le feature sono suddivise in diverse categorie, ognuna delle quali analizza un aspetto specifico dell'URL o delle sue caratteristiche associate. Di seguito, una panoramica delle principali categorie di variabili presenti nel dataset:

- **Informazioni sull'URL**: feature come `URL`, `Domain`, `TLD` analizzano la struttura di base dell'URL e le sue componenti principali.
- **Caratteristiche Strutturali**: feature come `NoOfSubDomain`, `HasObfuscation`, `IsHTTPS` analizzano la complessità e la struttura dell'URL, rilevando schemi sospetti.
- **Probabilità e Indici di Similarità**: feature come `URLSimilarityIndex`, `TLDLegitimateProb`, `URLCharProb` misurano quanto un URL assomigli a quelli legittimi o noti.
- **Caratteristiche del Contenuto della Pagina**: feature come `LineOfCode`, `LargestLineLength`, `HasTitle` analizzano la struttura e i contenuti del sito web.
- **Interazioni e Riferimenti**: feature come `NoOfURLRedirect`, `NoOfExternalRef`, `NoOfPopup` misurano i collegamenti e i comportamenti di navigazione.
- **Indicatori di Sicurezza e Truffa**: feature come `HasExternalFormSubmit`, `HasPasswordField`, `Pay` si concentrano su elementi tipici delle truffe online.
- **Variabile Target**: `label` indica se l'URL analizzato risulta essere un URL di phishing o meno.

La prossima fase dell'analisi comporterà l'esplorazione dei dati e la selezione delle variabili significative per la costruzione di un modello predittivo.

Feature	Tipo di Dato	Descrizione
FILENAME	Categorico	Nome del file analizzato.
URL	Categorico	Indirizzo web analizzato.
URLLength	Discreto	Lunghezza totale dell'URL.
Domain	Categorico	Dominio dell'URL.
DomainLength	Discreto	Lunghezza del dominio.
IsDomainIP	Discreto	Indica se il dominio è un indirizzo IP (1) o testuale (0).
TLD	Categorico	Dominio di primo livello (es. .com, .net).
URLSimilarityIndex	Continuo	Indice di similarità con URL noti.
CharContinuationRate	Discreto	Tasso di continuità tra i caratteri.
TLDLegitimateProb	Continuo	Probabilità che il TLD sia legittimo.
URLCharProb	Continuo	Probabilità basata sulla distribuzione dei caratteri.
TLDLength	Discreto	Lunghezza del TLD.
NoOfSubDomain	Discreto	Numero di sottodomini presenti.
HasObfuscation	Discreto	Presenza di tecniche di offuscamento.
NoOfObfuscatedChar	Discreto	Numero di caratteri offuscati.
ObfuscationRatio	Discreto	Rapporto tra caratteri offuscati e lunghezza URL.
NoOfLettersInURL	Discreto	Numero di lettere presenti nell'URL.
LetterRatioInURL	Continuo	Numero di lettere rispetto alla lunghezza totale dell'URL.
NoOfDegitsInURL	Discreto	Numero di cifre presenti nell'URL.
DegitsRatioInURL	Discreto	Numero di cifre rispetto alla lunghezza totale dell'URL.
NoOfEqualsInURL	Discreto	Numero di simboli "=" nell'URL.
NoOfQMarkInURL	Discreto	Numero di simboli "?" nell'URL.
NoOfAmpersandInURL	Discreto	Numero di simboli "&" nell'URL.
NoOfOtherSpecialCharsInURL	Discreto	Numero degli altri caratteri speciali nell'URL.
SpacialCharRatioInURL	Continuo	Numero di caratteri speciali rispetto alla lunghezza totale dell'URL.
IsHTTPS	Discreto	Indica se l'URL utilizza il protocollo HTTPS (1) o meno (0).
LineOfCode	Discreto	Numero di righe nel codice sorgente.
LargestLineLength	Discreto	Lunghezza della riga di codice più lunga.
HasTitle	Discreto	Presenza del titolo nella pagina.
Title	Categorico	Titolo della pagina.
DomainTitleMatchScore	Discreto	Similarità tra dominio e titolo.
URLTitleMatchScore	Discreto	Similarità tra URL e titolo.
HasFavicon	Discreto	Presenza della favicon nel sito.
Robots	Discreto	Presenza del file robots.txt.
IsResponsive	Discreto	Indica se il sito è responsive.
NoOfURLRedirect	Discreto	Numero di redirect esterni.
NoOfSelfRedirect	Discreto	Numero di redirect interni.
HasDescription	Discreto	Indica se è presente descrizione
NoOfPopup	Discreto	Numero di popup presenti.
NoOfiFrame	Discreto	Numero di iframe presenti.
HasExternalFormSubmit	Discreto	Form che inviano dati a domini esterni.
HasSocialNet	Discreto	Presenza di link ai social network.
HasSubmitButton	Discreto	Presenza di bottoni di submit.
HasHiddenField	Discreto	Presenza di campi nascosti all'utente.
HasPasswordField	Discreto	Presenza di campi password.
Bank	Discreto	Presenza di parole chiave bancarie.
Pay	Discreto	Presenza di parole legate ai pagamenti.
Crypto	Discreto	Presenza di riferimenti alle criptovalute.
HasCopyrightInfo	Discreto	Presenza di informazioni sul copyright
oOfImage	Discreto	Numero di immagini nella pagina.
NoOfCSS	Discreto	Numero di file CSS inclusi.
NoOfJS	Discreto	Numero di file JavaScript inclusi.
NoOfSelfRef	Discreto	Numero di riferimenti interni.
NoOfEmptyRef	Discreto	Numero di riferimenti vuoti.
NoOfExternalRef	Discreto	Numero di riferimenti esterni.
label	Discreto	Variabile target: 0 (Legittimo), 1 (Phishing).

Tabella 1: Descrizione completa delle 56 feature del dataset analizzato.

2.3 Analisi delle Correlazioni con la Variabile Target

Dopo aver esaminato il dataset e le sue feature, è stato calcolato il coefficiente di correlazione tra ogni variabile numerica e la variabile target. La correlazione è una misura statistica che descrive la relazione tra due variabili, indicando quanto e in che modo due variabili tendono a variare insieme. Avremo una **Correlazione Positiva** quando un aumento in una variabile corrisponde a un aumento nell'altra variabile; avremo una **Correlazione Negativa** quando un aumento in una variabile corrisponde a una diminuzione nell'altra variabile; infine, avremo una **Assenza di Correlazione** quando i cambiamenti in una variabile non influenzano l'altra. Il calcolo della correlazione tra tutte le variabili numeriche e quella target ci permette di comprendere quali siano le feature maggiormente correlate a tale variabile e, di conseguenza, selezionare unicamente quelle più rilevanti.

L'eliminazione delle feature superflue può portare a numerosi vantaggi, come ad esempio può migliorare la velocità di apprendimento del modello, la sua precisione ed evitare l'overfitting[1].

Prima di procedere al calcolo del coefficiente di correlazione, è necessaria una fase preliminare che riguarda la trasformazione delle variabili categoriche in numeriche. Le categorie interessate sono due: `Domain`, che indica il dominio dell'URL (come ad esempio .com, .it, etc.), e `Title`, che indica il titolo della pagina a cui conduce l'URL. Le variabili sono state prima trasformate in fattori e successivamente convertite in valore numerico (**Label Encoding**). Ora che anche le variabili categoriche più significative sono state convertite in valori numerici, è stato possibile procedere al calcolo del coefficiente di correlazione tra ogni feature e la variabile target. Dall'analisi dei coefficienti di correlazione (Tabella 2) possiamo esaminare quali sono le variabili che presentano la correlazione più alta con `label` (verranno considerate le feature con un coefficiente di correlazione $\geq |0.6|$):

1. `URLSimilarityIndex` (correlazione di 0.856): la correlazione più alta, il che suggerisce che questa feature è fortemente legata alla variabile target e potrebbe essere cruciale per determinare se un URL è sicuro o sospetto.

2. `HasSocialNet` (correlazione di 0.778): Un'altra correlazione molto alta. Questo indica che la presenza di riferimenti ai social network in un URL potrebbe essere fortemente correlata con la classificazione del sito come phishing o legittimo.

3. `HasCopyrightInfo` (correlazione: 0.746): La presenza di informazioni sul copyright potrebbe essere un buon indicatore per distinguere tra siti legittimi e siti phishing, dato che i siti di phishing spesso mancano di questi dettagli.

4. `HasDescription` (correlazione: 0.683): La presenza di una descrizione potrebbe essere un buon indicatore, poiché i siti legittimi tendono a fornire una descrizione chiara delle loro pagine, mentre i siti phishing potrebbero mancare di questa.

5. `IsHTTPS` (correlazione: 0.646): La presenza di HTTPS è spesso associata a siti legittimi. La sua correlazione relativamente alta con la variabile target suggerisce che questa caratteristica può essere un buon discriminante.

[1] L'overfitting si verifica quando un modello impara troppo bene i dettagli del dataset di addestramento, inclusi i rumori e le anomalie, ma non riesce a generalizzare bene sui nuovi dati.

Feature	**Correlazione con "label"**
URLSimilarityIndex	0.856135403
HasSocialNet	0.778829467
HasCopyrightInfo	0.746076118
HasDescription	0.683989489
IsHTTPS	0.646885058
HasSubmitButton	0.586002993
DomainTitleMatchScore	0.570249492
IsResponsive	0.564786719
NoOfJS	0.541923412
URLTitleMatchScore	0.529816070
HasHiddenFields	0.516743503
HasFavicon	0.506485685
HasTitle	0.473294908
URLCharProb	0.454740398
Domain_numeric	0.451279836
CharContinuationRate	0.448866608
NoOfCSS	0.424377445
NoOfSelfRef	0.405329350
Robots	0.396208704
Pay	0.371572507
NoOfExternalRef	0.369758258
LineOfCode	0.351496677
Title_numeric	0.319122238
NoOfImage	0.317604924
NoOfiFrame	0.251850614
Bank	0.210757171
HasExternalFormSubmit	0.170393099
HasPasswordField	0.149780257
NoOfEmptyRef	0.135309322
Crypto	0.101112015
TLDLegitimateProb	0.097092703
NoOfPopup	0.046738726
NoOfSubDomain	-0.004390171
LargestLineLength	-0.034737240
NoOfURLRedirect	-0.038133810
NoOfObfuscatedChar	-0.040996807
ObfuscationRatio	-0.043464578
NoOfAmpersandInURL	-0.048959571
HasObfuscation	-0.050892372
IsDomainIP	-0.056391743
TLDLength	-0.075725589
NoOfSelfRedirect	-0.089700554
NoOfEqualsInURL	-0.111961552
NoOfQMarkInURL	-0.170552587
DomainLength	-0.262763580
NoOfDegitsInURL	-0.287961146
URLLength	-0.288858679
NoOfLettersInURL	-0.305340519
LetterRatioInURL	-0.329955655
NoOfOtherSpecialCharsInURL	-0.406440494
DegitRatioInURL	-0.419638475
SpacialCharRatioInURL	-0.517029402

Tabella 2: Correlazione tra le feature e la variabile target `label` (ordinato dalla più alta alla più bassa)

3 Distribuzioni di Frequenza

Una distribuzione di frequenza è una tabella dove, in corrispondenza delle modalità, viene riportato il numero di volte che quelle stesse modalità si sono verificate. Grazie alle distribuzioni di frequenza, siamo in grado di valutare i valori più comuni, l'eventuale simmetria o asimmetria dei dati e la distribuzione complessiva.

Le feature analizzate sono principalmente di natura qualitativa, in quanto, ad eccezione di `URLSimilarityIndex`, presentano solo valori discreti (0 o 1), che rappresentano una codifica binaria di tipo sì/no. `URLSimilarityIndex`, invece, è di natura quantitativa e presenta valori continui; per questo motivo, si procederà a dividere il dataset in classi considerando dieci intervalli di uguale ampiezza. Gli intervalli considerati sono i seguenti:

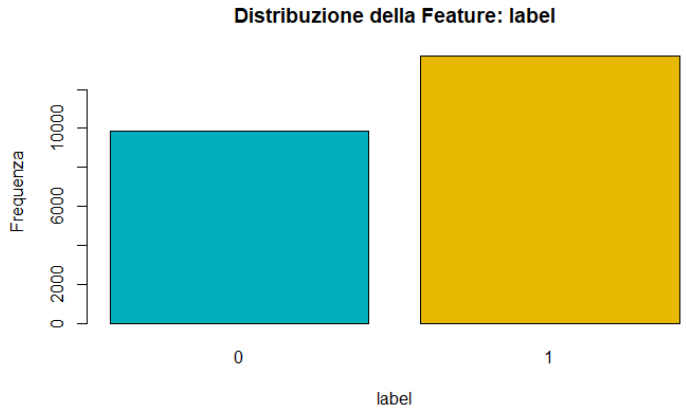
$$[0, 10), [10, 20), [20, 30), [30, 40), [40, 50), [50, 60), [60, 70), [70, 80), [80, 90), [90, 100)$$

3.1 Frequenze Assolute e Relative

Quando analizziamo le distribuzioni dei dati, è fondamentale fare una distinzione tra frequenza assoluta e frequenza relativa.
La **frequenza assoluta** indica il numero di volte in cui una determinata modalità o valore appare all'interno del nostro campione. In altre parole, è il conteggio effettivo delle occorrenze di un dato. La somma di tutte le frequenze assolute sarà sempre pari alla dimensione totale del campione, salvo che ci siano valori mancanti. Al contrario, la **frequenza relativa** rappresenta la proporzione di ciascun dato rispetto alla dimensione totale del campione. Viene calcolata dividendo la frequenza assoluta per la dimensione complessiva del campione. La somma delle frequenze relative sarà sempre uguale a 1, tranne nei casi in cui vi siano dati mancanti. Verrà considerata anche la **frequenza percentuale**, una forma di frequenza relativa che viene espressa come percentuale.

Per una visualizzazione efficace di tutte queste frequenze, è utile ricorrere ai **grafici a barre** (o barplot), che permettono di rappresentare chiaramente le distribuzioni e le proporzioni dei dati.

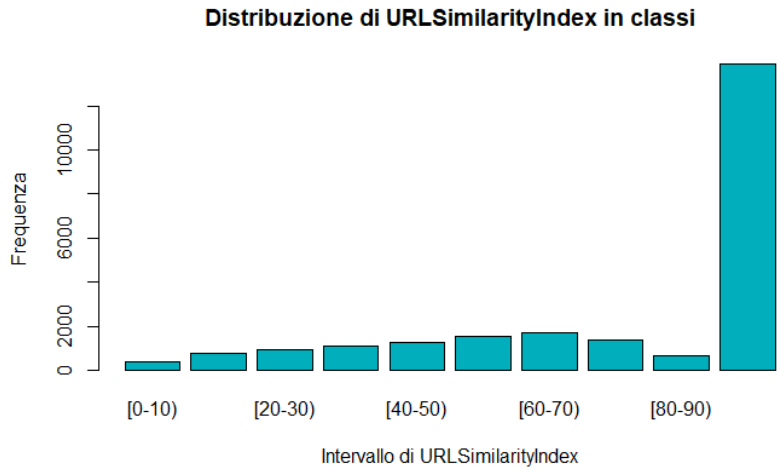
3.1.1 Distribuzione di `label`



Classe	Frequenza Assoluta	Frequenza Relativa	Frequenza Percentuale
0	9885	0.4192	41.92%
1	13695	0.5807	58.07%

Dai dati emerge che la distribuzione della variabile target è moderatamente sbilanciata, con la classe 1 che presenta circa il 16% di istanze in più rispetto alla classe 0. Tuttavia, poiché questo sbilanciamento non è particolarmente marcato, il bilanciamento della variabile target non è una priorità per l'addestramento del modello.

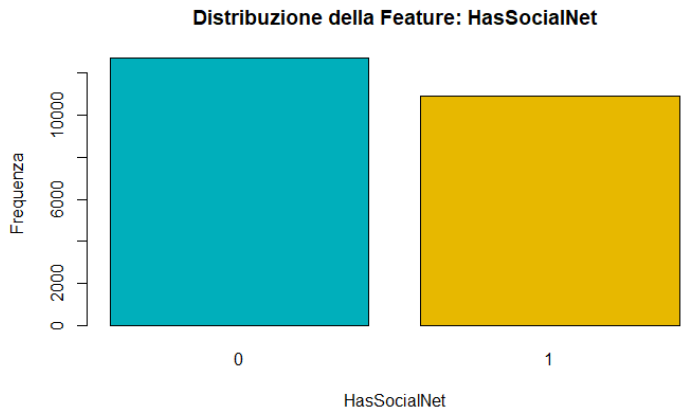
3.1.2 Distribuzione di `URLSimilarityIndex`



Intervallo	Frequenza Assoluta	Frequenza Relativa	Frequenza Percentuale
[0-10)	366	0.0155	1.55%
[10-20)	754	0.0320	3.20%
[20-30)	952	0.0404	4.04%
[30-40)	1083	0.0459	4.59%
[40-50)	1242	0.0527	5.27%
[50-60)	1540	0.0653	6.53%
[60-70)	1725	0.0732	7.32%
[70-80)	1352	0.0573	5.73%
[80-90)	675	0.0286	2.86%
[90-100)	13891	0.5891	58.91%

La distribuzione di `URLSimilarityIndex` risulta fortemente sbilanciata verso l'intervallo [90, 100), che rappresenta il 58,91% del totale, mentre gli altri intervalli contengono percentuali molto più basse, che vanno da una minima di 1.55% a una massima di 7.32%. Questo suggerisce che la maggior parte degli URL considerati sia relativamente simile a URL più noti.

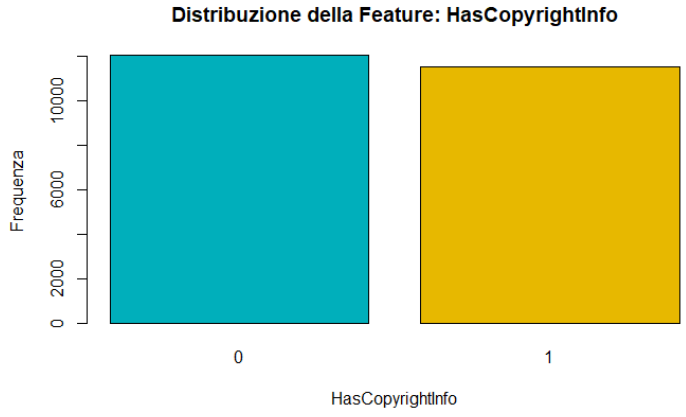
3.1.3 Distribuzione di `HasSocialNet`



Classe	Frequenza Assoluta	Frequenza Relativa	Frequenza Percentuale
0	12694	0.5383	53.83%
1	10886	0.4617	46.17%

La distribuzione di `HasSocialNet` è abbastanza equilibrata, con il 53.83% dei dati appartenenti alla classe 0 e il 46.17% appartenente alla classe 1. Questo significa che non c'è un forte sbilanciamento tra le classi, seppur la classe 0 sia leggermente più numerosa.

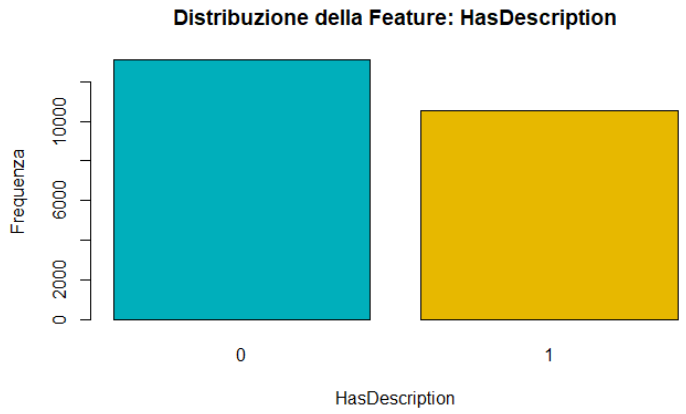
3.1.4 Distribuzione di `HasCopyrightInfo`



Classe	Frequenza Assoluta	Frequenza Relativa	Frequenza Percentuale
0	12053	0.5111	51.12%
1	11527	0.4888	48.88%

La distribuzione di `HasCopyrightInfo` mostra una differenza minima tra i dati che appartengono alla classe 0 rispetto a quelli appartenenti alla classe 1. Questo comporta l'assenza di problematiche relative allo sbilanciamento tra le classi in fase di addestramento del modello.

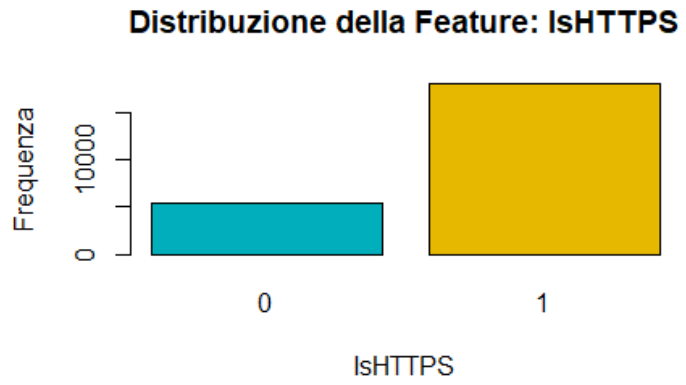
3.1.5 Distribuzione di `HasDescription`



Classe	Frequenza Assoluta	Frequenza Relativa	Frequenza Percentuale
0	13076	0.5545	55,45%
1	10504	0.4455	44.55%

La distribuzione di `HasDescription` risulta abbastanza bilanciata tra le due classi. Ciò significa che la loro differenza non è eccessiva, e che quindi non verrà causato sbilanciamento nel modello predittivo.

3.1.6 Distribuzione di `IsHTTPS`



Classe	Frequenza Assoluta	Frequenza Relativa	Frequenza Percentuale
0	5470	0.2319	23.20%
1	18110	0.7680	76.80%

La distribuzione di `IsHTTPS` evidenzia un forte sbilanciamento tra le due classi, con la classe 1 (76.8%) nettamente più popolosa della classe 0. L'alta percentuale di siti con HTTPS potrebbe influenzare i modelli di machine learning.

4 Statistica Descrittiva Univariata

La statistica descrittiva è costituita da un insieme di metodi di natura logica e matematica con lo scopo di raccogliere, elaborare, analizzare e a interpretare dati per riuscire a descrivere fenomeni collettivi e/o di estendere la descrizione di certi fenomeni osservati ad altri fenomeni dello stesso tipo non ancora osservati. Parliamo in particolare di statistica descrittiva univariata quando la descrizione della distribuzione riguarda una singola variabile.

4.1 Funzione di Distribuzione Empirica

Per i fenomeni quantitativi è spesso utile definire la **funzione di distribuzione empirica** (FDE), che ci permette di capire come si distribuiscono i dati osservati nel "fenomeno" per studiarne le "caratteristiche" ed il comportamento delle modalità dei caratteri del fenomeno osservato. Poiché le feature `HasSocialNet`, `HasCopyrightInfo`, `HasDescription` e `IsHTTPS` sono binarie, in quanto assumono unicamente valore 0 o 1, la loro distribuzione empirica è già descritta dalla frequenza relativa di ciascun valore. Diverso è il discorso per la variabile `URLSimilarityIndex` per la quale ha senso definire e analizzare la FDE. La variabile `URLSimilarityIndex`, essendo continua, può assumere un'ampia gamma di valori, e la FDE permette di descrivere come questi valori si distribuiscono all'interno del fenomeno osservato. In particolare, la FDE consente di calcolare la proporzione di osservazioni che sono minori o uguali a un dato valore x, offrendo così una rappresentazione cumulativa della distribuzione dei dati.

4.1.1 Calcolo della Funzione di Distribuzione Empirica Continua

La funzione di distribuzione empirica continua (FDEC) è una particolare funzione strutturata in classi. Come prima cosa organizziamo i dati numerici in k classi ovvero:

$$C_1 = [z_0, z_1),\ C_2 = [z_1, z_2),\ \ldots,\ C_k = [z_{k-1}, z_k), \quad \text{con } z_0 < z_1 < \cdots < z_{k-1} < z_k$$

Dove z_0 corrisponde al minimo delle osservazioni e z_k che corrisponde al massimo delle osservazioni. La funzione di distribuzione empirica continua è così definita:

$$F(x) = \begin{cases} 0, & x < z_0 \\ \vdots & \\ F_{i-1}, & x = z_{i-1} \\ \frac{F_i - F_{i-1}}{z_i - z_{i-1}} x + \frac{z_i F_{i-1} - z_{i-1} F_i}{z_i - z_{i-1}}, & z_{i-1} < x < z_i \\ F_i, & x = z_i \\ \vdots & \\ 1, & x \geq z_k \end{cases}$$

dove:

- F_i denota la frequenza relativa cumulativa della classe C_i, con $i = 1, 2, \ldots, k$.
- $F(x) = 0$ per $x < z_0$, e $F(x) = 1$ per $x \geq z_k$.

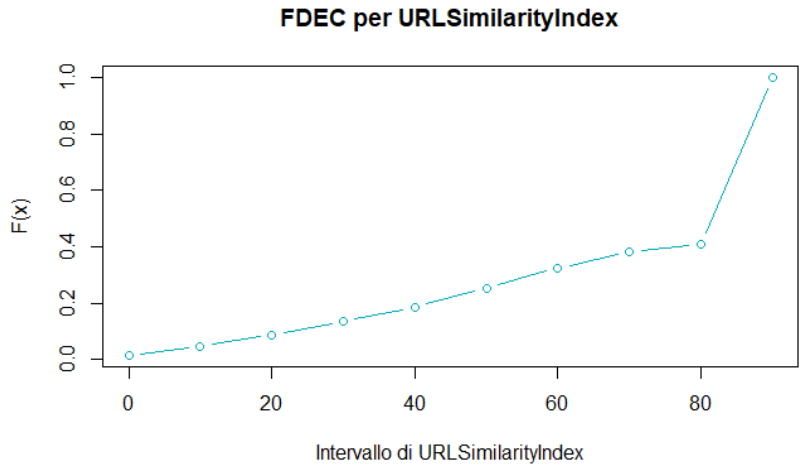
Consideriamo la feature `URLSimilarityIndex`. Sono state individuate $k = 10$ classi:

$$C_1 = [0, 10), \quad C_2 = [10, 20), \quad C_3 = [20, 30), \quad C_4 = [30, 40), \quad C_5 = [40, 50),$$
$$C_6 = [50, 60), \quad C_7 = [60, 70), \quad C_8 = [70, 80), \quad C_9 = [80, 90), \quad C_{10} = [90, 100)$$

La FDEC assumerà i seguenti valori:

Indice	C_i	n_i	f_i	F_i
1	[0-10)	366	$\frac{366}{23580}$	$\frac{366}{23580}$
2	[10-20)	754	$\frac{754}{23580}$	$\frac{1120}{23580}$
3	[20-30)	952	$\frac{952}{23580}$	$\frac{2072}{23580}$
4	[30-40)	1083	$\frac{1083}{23580}$	$\frac{3155}{23580}$
5	[40-50)	1242	$\frac{1242}{23580}$	$\frac{4397}{23580}$
6	[50-60)	1540	$\frac{1540}{23580}$	$\frac{5937}{23580}$
7	[60-70)	1725	$\frac{1725}{23580}$	$\frac{7662}{23580}$
8	[70-80)	1352	$\frac{1352}{23580}$	$\frac{9014}{23580}$
9	[80-90)	675	$\frac{675}{23580}$	$\frac{9689}{23580}$
10	[90-100)	13891	$\frac{13891}{23580}$	$\frac{23580}{23580}$

$$F(x) = \begin{cases} 0, & x < 0 \\ \frac{366}{23580} = 0.0155, & 0 \leq x < 10 \\ \frac{1120}{23580} = 0.0474, & 10 \leq x < 20 \\ \frac{2072}{23580} = 0.0879, & 20 \leq x < 30 \\ \frac{3155}{23580} = 0.1338, & 30 \leq x < 40 \\ \frac{4397}{23580} = 0.1865, & 40 \leq x < 50 \\ \frac{5937}{23580} = 0.2518, & 50 \leq x < 60 \\ \frac{7662}{23580} = 0.3250, & 60 \leq x < 70 \\ \frac{9689}{23580} = 0.3823, & 70 \leq x < 80 \\ \frac{23580}{23580} = 0.4110, & 80 \leq x < 90 \\ 1, & x \geq 100 \end{cases}$$



4.2 Introduzione agli Indici di Sintesi

Gli indici di sintesi sono strumenti statistici che aiutano a riassumere in un singolo valore (o in pochi valori) informazioni relative a un insieme di dati numerici. Sono necessari per sintetizzare e classificare specifiche osservazioni sui dati che abbiamo.

4.2.1 Misure di Centralità

Le misure di centralità forniscono informazioni riguardo ai valori attorno ai quali i dati tendono a concentrarsi.

Media Campionaria. Sia X un insieme di n valori numerici, cioè $X = \{x_1, x_2, \ldots, x_n\}$. La media campionaria è definita come la media aritmetica dei valori in X, ed è data dalla formula:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Per ogni valore x_i assunto da X siamo in grado di calcolare lo scostamento rispetto alla media stessa, cioè lo **scarto dalla media campionaria**, dato dalla formula:

$$s_i = x_i - \bar{x}$$

La somma algebrica degli scarti dalla media campionaria è sempre nulla.

Mediana Campionaria. Sia X un insieme di dati di ampiezza n, cioè $X = \{x_1, x_2, \ldots, x_n\}$, e siano $x_1 < x_2 < \cdots < x_n$ ordinati in ordine crescente. La mediana campionaria è definita come segue:

- Se n è dispari, la mediana è il valore in posizione $\frac{n+1}{2}$.
- Se n è pari, la mediana è la media aritmetica dei valori in posizione $\frac{n}{2}$ e $\frac{n}{2} + 1$.

La mediana campionaria bipartisce le osservazioni in due gruppi di uguale numerosità, in maniera tale che lo stesso numero di valori cada sia a sinistra che a destra della mediana stessa. Per descrivere la forma di una distribuzione si può confrontare la media campionaria e la mediana campionaria:

- se queste due misure sono uguali la distribuzione di frequenze tende ad essere simmetrica;
- se la media campionaria è sensibilmente maggiore della mediana campionaria, la distribuzione di frequenze è più sbilanciata verso destra;
- se invece la media campionaria è sensibilmente minore della mediana campionaria la distribuzione di frequenze è più sbilanciata verso sinistra.

Moda Campionaria. La moda campionaria di un insieme di dati, se esiste, è la modalità a cui è associata la frequenza (assoluta o relativa) più elevata. Se esistono più modalità con frequenza massima, ciascuna di esse è detta *valore modale*.

Quartili. I quartili sono dei particolari quantili che si ottengono dividendo l'insieme dei dati ordinati in quattro parti uguali. Aiutano a comprendere la distribuzione e la dispersione dei dati, specialmente quando si lavora con Dataset grandi o complessi. Sono definiti come segue: consideriamo un campione $(x_1, x_2, \dots, x_n)$ dei valori assunti da una variabile quantitativa X. Procediamo ad ordinare i valori del campione in ordine crescente:

- Q_0: corrisponde al minimo dei valori del campione;
- Q_1 (Primo Quartile): il valore per il quale il 25% dei dati sono alla sua sinistra;
- Q_2 (Secondo Quartile): il valore per il quale il 50% dei dati sono alla sua sinistra;
- Q_3 (Terzo Quartile): il valore per il quale il 75% dei dati sono alla sua sinistra;
- Q_4: corrisponde al massimo dei valori del campione.

Grazie ai quartili siamo in grado di suddividere i dati in parti uguali, andando ad ottenere informazioni sulla loro distribuzione. Inoltre, l'uso dei quartili consente di individuare i valori molto distanti dalla distribuzione centrale dei dati, cioè gli **outlier**; gli outlier sono definiti come i valori che si trovano al di fuori dell'intervallo interquartile (IQR), che è la differenza tra il terzo quartile (Q_3) e il primo quartile (Q_1). Infine, attraverso l'IQR si può avere una visione più affidabile della variabilità dei dati, in quanto non influenzata dagli outlier.

4.2.2 Misure di Dispersione

Le misure di dispersione sono statistiche che descrivono quanto i dati di un insieme siano distribuiti attorno alla media o al centro della distribuzione. Esse forniscono informazioni sulla variabilità dei dati e aiutano a comprendere quanto i valori si discostino dalla media.

Varianza. La varianza campionaria è una misura statistica che quantifica la dispersione dei dati attorno alla media campionaria. Indica, cioè, quanto i valori di un campione variano rispetto alla media del campione stesso. Assegnato un insieme di dati numerici $(x_1, x_2, \dots, x_n)$, si definisce varianza campionaria, e si denota con s^2, la quantità:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2, \quad (n = 2, 3, \dots)$$

dove $\bar{x}$ denota la media campionaria dei dati, n è la dimensione del campione e x_i il valore della i-esima osservazione nel campione.

Deviazione Standard. Si definisce deviazione standard campionaria la radice quadrata della varianza campionaria, ossia:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}, \quad (n = 2, 3, \dots)$$

La deviazione standard risolve il problema delle unità di misura legato alla varianza: essendo espressa nelle stesse unità dei dati originali, risulta più interpretabile e utile per l'analisi dei dati rispetto alla varianza.

4.2.3 Misure di Simmetria

Le misure di simmetria permettono di misurare quando una distribuzione di frequenze presenta simmetria o asimmetria, oppure se tale asimmetria è più o meno piccata.

Skewness Campionaria. Assegnato un insieme di dati numerici $x_1, x_2, \ldots, x_n$ si definisce skewness campionaria:

$$\gamma_1 = \frac{m_3}{m_2^{\frac{3}{2}}}$$

dove m_2 denota il momento campionario di ordine 2 ed m_3 denota il momento campionario di ordine 3. La skewness campionaria permette di misurare la simmetria di una distribuzione:

- $\gamma_1 = 0$: la distribuzione di frequenze è simmetrica;
- $\gamma_1 > 0$: la distribuzione di frequenze è asimmetrica positiva, ossia la distribuzione di frequenze ha la coda di destra più allungata;
- $\gamma_1 < 0$: la distribuzione di frequenze è asimmetrica negativa, ossia la distribuzione di frequenze ha la coda di sinistra più allungata.

Curtosi Campionaria. Assegnato un insieme di dati numerici $x_1, x_2, \ldots, x_n$ si definisce curtosi campionaria:

$$\gamma_2 = \beta_2 - 3$$

dove β_2, anche detto indice di Pearson, è dato dal rapporto tra il momento campionario di ordine 4 ed il quadrato del momento campionario di ordine 2, come segue:

$$\beta_2 = \frac{m_4}{m_2^2}$$

La curtosi campionaria permette di confrontare la distribuzione di frequenze dei dati con una densità di probabilità normale standard. In generale, se:

- $\beta_2 = 3$: la distribuzione di frequenze si definisce normocurtica o mesosurtica, ossia la distribuzione di frequenze è piatta come una normale;
- $\beta_2 < 3$: la distribuzione di frequenze si definisce platicurtica, ossia la distribuzione di frequenze è più piatta di una normale;
- $\beta_2 > 3$: la distribuzione di frequenze si definisce leptocurtica, ossia la distribuzione di frequenze è più piccata di una normale.

4.3 Indici di Sintesi per le Variabili Binarie

Poiché, come già detto, le feature `HasSocialNet`, `HasCopyrightInfo`, `HasDescription` e `IsHTTPS` sono variabili binarie, gli indici di sintesi non sono particolarmente informativi, in quanto molte informazioni possono essere dedotte a partire dalle distribuzioni di frequenza. Tuttavia, possiamo comunque calcolare:

- la media, che in questo caso rappresenta la proporzione di 1;
- la varianza, che misura la dispersione rispetto alla media, utile per capire il bilanciamento della variabile. Poichè in questo caso abbiamo a che fare con variabili binarie, la varianza assumerà valore compreso tra 0 e 0.25, il quale indica varianza massima;
- la deviazione standard, calcolata come la radice quadrata della varianza. Essendo una misura lineare, la deviazione standard fornisce una comprensione più immediata della dispersione dei dati;
- la moda, che corrisponde al valore più frequente.

Feature	**Media**	**Varianza**	**Deviazione Standard**	**Moda**
label	0.5808	0.2435	0.4934	1
HasSocialNet	0.4617	0.2485	0.4985	0
HasCopyrightInfo	0.4888	0.2499	0.4999	0
HasDescription	0.4455	0.247	0.497	0
IsHTTPS	0.768	0.1782	0.4221	1

Tabella 3: Indici di sintesi per le variabili binarie.

Analisi dei risultati. Come emerge dai risultati riportati nella Tabella 3, i risultati osservabili già a partire dall'osservazione delle distribuzioni di frequenza sono stati confermati: la classe `IsHTTPS` è fortemente sbilanciata verso la classe 1, ed ha una minore diversità nei dati, mentre le altre feature presentano una distribuzione più bilanciata.

4.4 Indici di Sintesi per la Variabile Continua

Data la natura continua della feature `URLSimilarityIndex`, gli indici di sintesi possono risultare più significativi rispetto alle variabili binarie analizzate precedentemente. Di seguito saranno riportati gli indici di sintesi e una loro interpretazione dettagliata.

4.4.1 Media, Moda e Mediana Campionaria

La media, la moda e la mediana, osservate congiuntamente (Tabella 4), possono fornire importanti indicazioni sulla simmetria della distribuzione, sulla presenza di valori predominanti e sull'eventuale presenza di asimmetrie. La mediana, pari a 100, rappresenta il valore centrale che separa i dati in due parti uguali. Poiché è maggiore della media (79.51), suggerisce che la distribuzione dei dati risulta asimmetrica, ed in particolare sbilanciata verso sinistra. La moda di 100 indica che è questo il valore che si verifica più frequentemente. Ciò conferma la presenza dell'asimmetria, sottolineando una concentrazione dei dati su quel punto, con pochi dati che si estendono verso i valori più bassi.

Feature	Media	Mediana	Moda
URLSimilarityIndex	79.5097	100	100

Tabella 4: Misure di centralità per la variabile continua.

4.4.2 Quartili

Grazie all'analisi dei quartili (Tabella 5), possiamo notare che i valori di mediana (Q_2), terzo quartile (Q_3) e massimo valore della distribuzione (Q_4) sono identici (100), indicando una forte concentrazione dei dati nel massimo. Conoscendo i quartili, possiamo inoltre procedere al calcolo dell'IQR per individuare eventuali outlier. Nel nostro caso, l'IQR è pari a:

$$IQR = Q_3 - Q_1 = 100 - 59.65332 = 40.34668$$

Gli outlier sono definiti come valori esterni all'intervallo $[Q_1 - 1.5 \cdot IQR, Q3 + 1.5 \cdot IQR]$, quindi gli outlier della variabile considerata dovranno essere ricercati all'esterno dell'intervallo:

$$[0.8667, 160.52002]$$

Essendo il limite inferiore dell'intervallo minore del valore minimo osservato, ed essendo il limite superiore dell'intervallo maggiore del valore massimo osservato possiamo dedurre che non esistono outlier per la feature `URLSimilarityIndex`.

Feature	Q_0	Q_1	Q_2	Q_3	Q_4
URLSimilarityIndex	1.292906	59.65332	100	100	100

Tabella 5: Quartili della variabile continua.

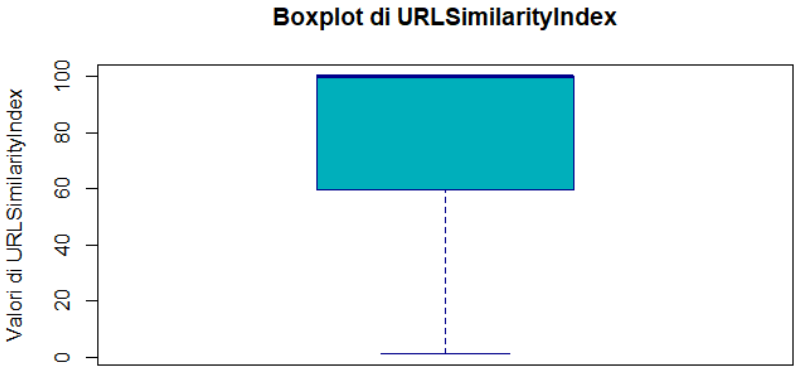


Figura 1: In questo boxplot relativo a `URLSimilarityIndex` è evidente la concentrazione di valori verso il valore massimo e l'asenza di outlier.

4.4.3 Varianza e Deviazione Standard

La varianza e la deviazione standard sono due indici fondamentali per misurare la dispersione dei dati rispetto alla loro media. La varianza di valore elevato (793.62) suggerisce che, sebbene la distribuzione di valori si concentri intorno a valori elevati (come confermano moda, mediana e quartili), tuttavia esistono valori lontani dalla media, contribuendo alla variabilità dei dati. Stesso dicasi per la deviazione standard.

Feature	**Varianza**	**Deviazione Standard**
URLSimilarityIndex	793.6245	28.1713

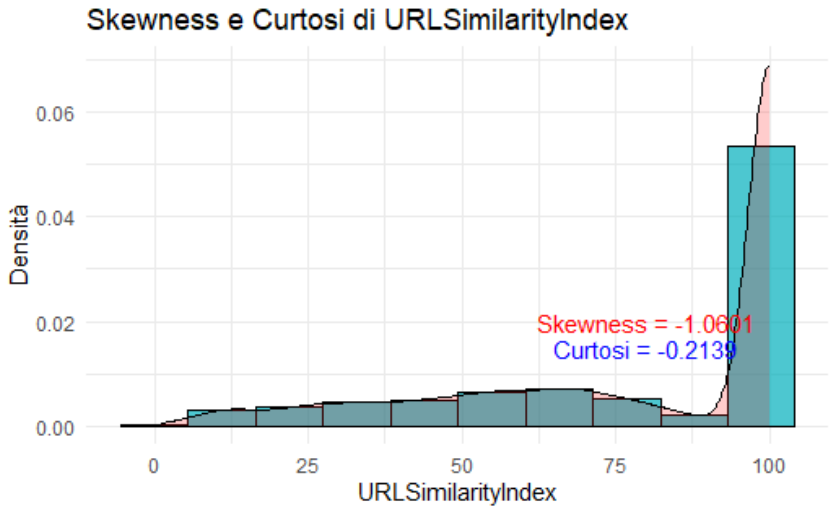
Tabella 6: Misure di Dispersione della variabile continua.

4.4.4 Skewness e Curtosi Campionaria

Poiché la skewness campionaria di `URLSimilarityIndex` è inferiore a zero, in particolare pari a -1.0601, la distribuzione di questa variabile risulta asimmetrica negativa. Ciò significa che la distribuzione ha una coda più lunga a sinistra, con una concentrazione maggiore di valori verso la parte alta della scala e pochi valori estremi più bassi. Per quanto riguarda la curtosi campionaria, un valore inferiore a 0, come in questo caso (-0.2139) indica una distribuzione platicurtica, ossia una distribuzione più piatta rispetto ad una distribuzione normale.

Feature	**Skewness**	**Curtosi**
URLSimilarityIndex	-1.0601	-0.2139

Tabella 7: Indici di Simmetria della variabile continua.



5 Statistica Descrittiva Bivariata

Indichiamo con Statistica Descrittiva Bivariata il ramo della statistica che si occupa dei metodi grafici e statistici atti a descrivere le relazioni che intercorrono tra due variabili. Le relazioni tra variabili quantitative possono essere rappresentate graficamente mediante diagrammi di dispersione (o **scatterplot**) in cui ogni coppia di osservazioni viene rappresentata sotto forma di un punto o di un cerchietto in un piano euclideo.

5.1 Covarianza Campionaria

Data una coppia (X, Y) di variabili quantitative. Consideriamo $C = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ con $n = |C|$, campione di n osservazioni. Definiamo la covarianza campionaria C_{xy} tra X e Y del campione C con la formula seguente:

$$C_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y})$$

Dove:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Da questa definizione notiamo che il prodotto interno alla sommatoria sarà positivo per le osservazioni (x_i, y_i) in cui le componenti della coppia sono o entrambe maggiori della media campionaria della variabile, negativo negli altri casi, cioè quando una parte della coppia risulta essere maggiore e l'altra minore.

Un'altra cosa da notare è che nella definizione la sommatoria viene divisa per $n - 1$ e questo viene fatto per normalizzarla, in quanto nel caso in cui le variabili x e y siano uguali si ottiene la varianza campionaria. A seconda del risultato ottenuto, possiamo dire che:

- Se $C_{xy} > 0$ consideriamo X e Y correlate positivamente;
- Se $C_{xy} < 0$ le consideriamo correlate negativamente;
- Se $C_{xy} = 0$ le consideriamo non correlate.

Nel primo caso ci possiamo aspettare che cambiamenti che troviamo nella prima variabile possono essere trovati anche nella seconda, mentre nel secondo caso no.

5.2 Coefficiente di Correlazione Campionario

È un metodo alternativo per ottenere una misura quantitativa della correlazione tra le variabili. Data una coppia (X, Y) di variabili quantitative. Consideriamo $C = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ con $n = |C|$, campione di n osservazioni. Siano:

- $\overline{x}$ la media campionaria dei valori assunti da X in $C(x_1, x_2, \ldots, x_n)$;
- S_x la deviazione standard campionaria di $(x_1, x_2, \ldots, x_n)$;
- $\overline{y}$ la media campionaria dei valori assunti da X in $C(y_1, y_2, \ldots, y_n)$;
- S_y la deviazione standard campionaria di $(y_1, y_2, \ldots, y_n)$;

Il coefficiente di correlazione campionario è definito dalla seguente formula:

$$r_{xy} = \frac{C_{xy}}{s_x s_y}$$

Esso, inoltre gode delle seguenti proprietà:

1. $-1 \leq r_{xy} \leq 1$;
2. presi $a, b \in \mathbb{R}$, con $a > 0$, tali che $y_i = ax_i + b$ per ogni $i = 1, 2, \ldots, n$, allora $r_{xy} = 1$;
3. presi $a, b \in \mathbb{R}$, con $a < 0$, tali che $y_i = ax_i + b$ per ogni $i = 1, 2, \ldots, n$, allora $r_{xy} = -1$;
4. se esistono quattro numeri reali a, b, c, d e se risulta $z_i = ax_i + b$ e $w_i = cy_i + d$ per $i = 1, 2, \ldots, n$, allora $r_{zw} = r_{xy}$ se $ac > 0$ e $r_{zw} = -r_{xy}$ se invece $ac < 0$.

La proprietà 1 afferma che il coefficiente di correlazione campionario è compreso nell'intervallo $[-1, 1]$. Le proprietà 2 e 3 mostrano che i valori limite -1 e $+1$ sono effettivamente raggiunti solo quando tra X e Y è presente una relazione lineare, ossia quando i punti dello scatterplot sono tutti su di una retta. Infine, la proprietà 4 afferma che il quadrato del coefficiente di correlazione non cambia se sommiamo costanti o moltiplichiamo per costanti tutti i valori di X e/o di Y. Ciò significa che il coefficiente di correlazione non dipende dalle unità di misura scelte per rappresentare i dati, e quindi è un indice adimensionale.

Dato che r_{xy} misura la forza del legame di natura lineare esistente tra due variabili quantitative, possiamo vedere graficamente, grazie al suo segno, la direzione della retta interpolante che indicherà una di queste situazioni:

- $r_{xy} = 1$: tutti i punti sono allineati su una retta ascendente (correlazione perfetta positiva);
- $0 < r_{xy} < 1$: i punti (x_i, y_i) sono posizionati in una nuvola attorno ad una linea retta interpolante ascendente, con x_i e y_i che tendono ad essere grandi e piccoli insieme (correlazione positiva);
- $r_{xy} = 0$: i punti sono completamente dispersi in una nuvola che non presenta alcuna evidente direzione di natura lineare (nessuna correlazione);
- $-1 < r_{xy} < 0$: i punti (x_i, y_i) sono posizionati in una nuvola attorno ad una linea retta interpolante discendente, con x_i grande e y_i piccolo o viceversa (correlazione negativa);
- $r_{xy} = -1$: tutti i punti sono allineati su una linea retta discendente (correlazione perfetta negativa).

5.3 Relazione delle feature considerate con la variabile target

Conoscendo il valore della variabile target `label`, risulta utile visualizzare la forza della sua relazione con le altre feature selezionate, cioè `URLSimilarityIndex`, `HasSocialNet`, `HasCopyrightInfo`, `HasDescription` e `IsHTTPS`. Nella creazione dei grafici, è stato effettuato un campionamento casuale di 100 osservazioni.

5.3.1 Covarianza Campionaria

Utilizziamo la covarianza campionaria per valutare la relazione tra la variabile target `label` e le feature selezionate. Una covarianza positiva indica che un aumento nei valori di una variabile tende a essere associato a un aumento nei valori dell'altra, mentre una covarianza negativa suggerisce una relazione inversa.

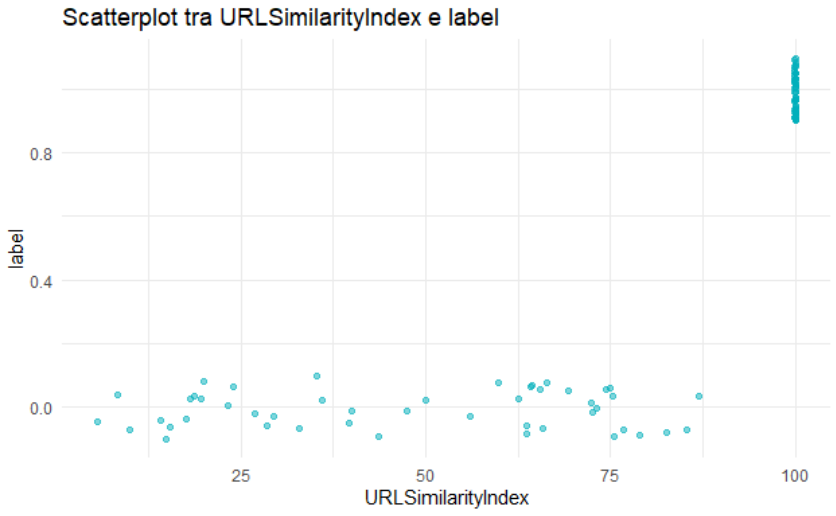
Feature	**Covarianza con label**
URLSimilarityIndex	11.901
HasSocialNet	0.1916
HasCopyrightInfo	0.184
HasDescription	0.1678
IsHTTPS	0.1347

Tabella 8: Covarianza Campionaria tra le feature selezionate e la variabile target.

URLSimilarityIndex e label. La covarianza tra `URLSimilarityIndex` e `label` è pari a 11.901, che è significativamente più alta rispetto alle altre feature. Questo indica una relazione forte e positiva tra questa variabile e la variabile target.
Sia dalla tabella di contingenza (Tabella 9) che dallo scatterplot emerge che i valori di `label` = 1 sono concentrati quando `URLSimilarityIndex` assume valore pari a 100, mentre i valori di `label` = 0 sono distribuiti su una gamma più ampia di valori, in particolare verso il basso (valori minori di 80). Questo suggerisce che `URLSimilarityIndex` ha un ruolo rilevante nel distinguere le due classi della variabile target, soprattutto per valori elevati della feature.

Label	**[0,10]**	**(10,20]**	**(20,30]**	**(30,40]**	**(40,50]**	**(50,60]**	**(60,70]**	**(70,80]**	**(80,90]**	**(90,100]**
0	366	754	952	1084	1242	1554	1710	1371	656	196
1	0	0	0	0	0	0	0	0	0	13695

Tabella 9: Tabella di contingenza tra `label` e `URLSimilarityIndex`

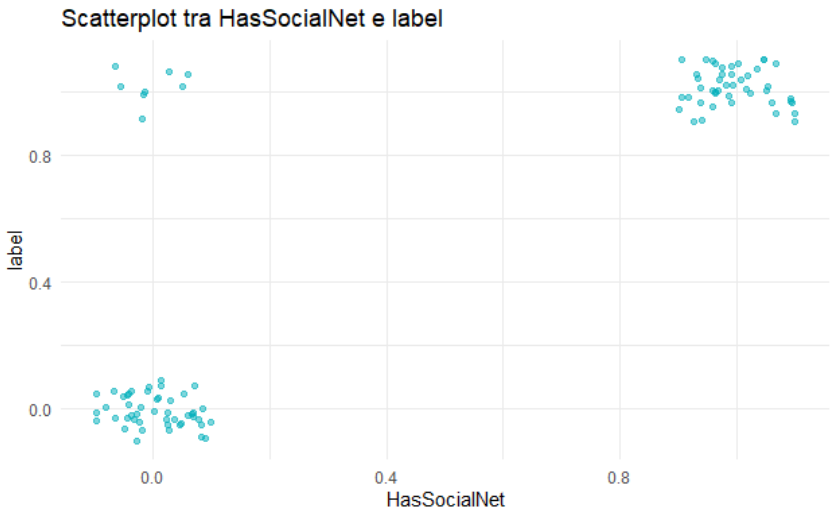


HasSocialNet e label. La covarianza tra `HasSocialNet` e `label` è pari a 0.1916, indicando una relazione moderata e positiva tra le due variabili.

Osservando sia la tabella 10 che il grafico a dispersione, si osserva che vi è una forte corrispondenza tra i valori di `HasSocialNet` e `label`. In particolare possiamo notare che esistono solo 46 tuple per cui ad un valore di `HasSocialNet` = 1 corrisponde un `label` = 0, andando ad indicare che quasi tutti i siti di phishing osservati presentano riferimenti a social network esterni.

Label	**HasSocialNet = 0**	**HasSocialNet = 1**
0	9839	46
1	2855	10840

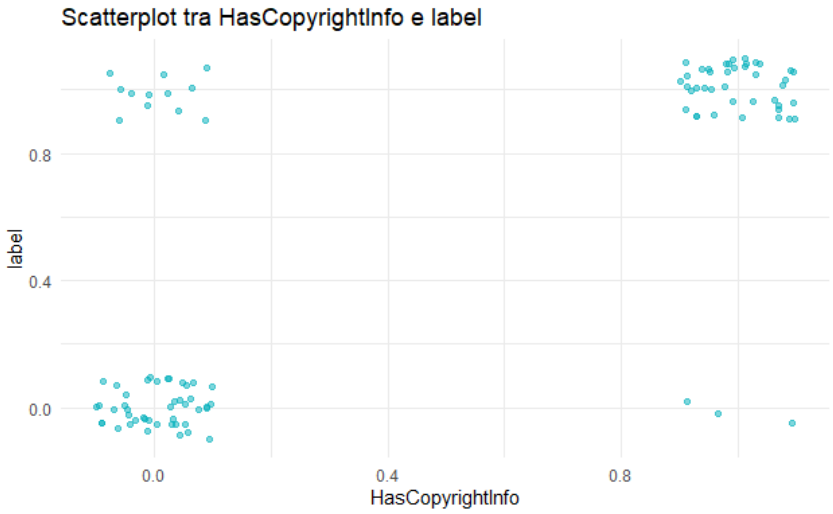
Tabella 10: Tabella di contingenza tra `label` e `HasSocialNet`



HasCopyrightInfo e label. La covarianza tra `HasCopyrightInfo` e `label` è pari a 0.184, suggerendo una relazione moderata e positiva tra questa feature e la variabile target.
Anche in questo caso, la tabella di contingenza (Tabella 11) e il diagramma di dispersione mostrano che i valori di `label` pari a 1 si associano a `HasCopyrightInfo` = 1. Ciò indica che la presenza di informazioni riguardo al Copyright ha un certo valore nel determinare se un URL considerato è di phishing.

Label	**HasCopyrightInfo = 0**	**HasCopyrightInfo = 1**
0	9392	493
1	2661	11034

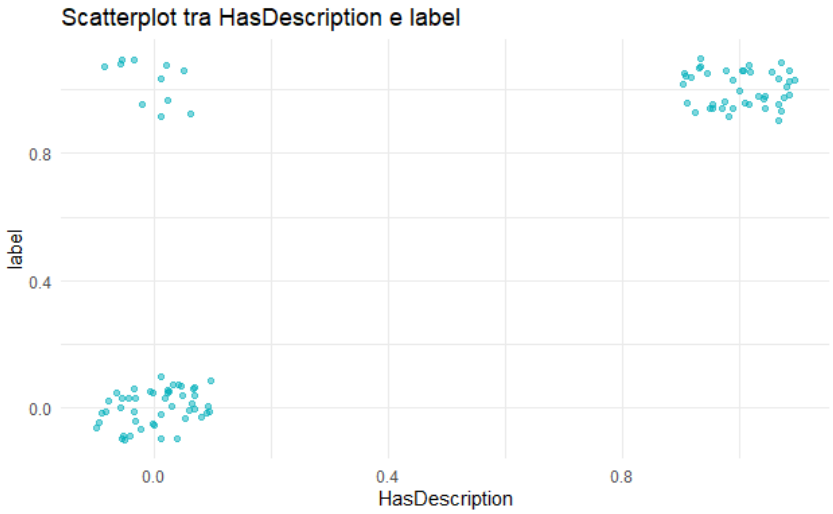
Tabella 11: Tabella di contingenza tra `label` e `HasCopyrightInfo`



HasDescription e label. La covarianza tra `HasDescription` e `label` è pari a 0.1678, indicando una relazione positiva moderata tra questa variabile e la variabile target.
Come nei casi precedenti, la tabella di contingenza 12 e lo scatterplot mostrano che i valori di `HasDescription` pari 1 si associano a `label` = 1. La presenza della Descrizione risulta un importante indicatore del fatto che un determinato indirizzo sia di phishing.

Label	**HasDescription = 0**	**HasDescription = 1**
0	9437	448
1	3639	10056

Tabella 12: Tabella di contingenza tra `label` e `HasDescription`

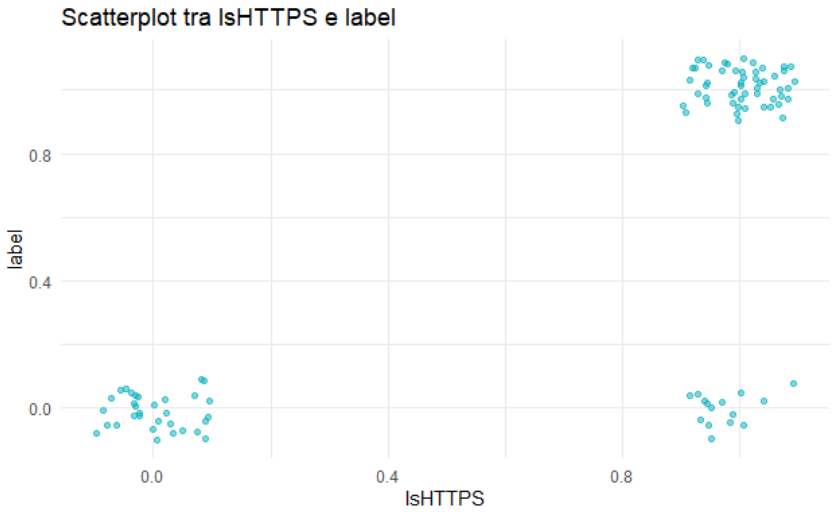


IsHTTPS e label. La covarianza tra `IsHTTPS` e `label` è pari a 0.1347, la più bassa tra le variabili analizzate. Questo indica che la relazione positiva tra `IsHTTPS` e `label` è presente, ma più debole rispetto alle altre feature.

Possiamo notare dall'osservazione della tabella 13 e del grafico sottostante che per IsHTTPS pari a 0 non esistono valori di `label` pari a 1. Ciò indica che ogni URL di phishing analizzato è associato al protocollo HTTPS. Per quanto riguarda `IsHTTPS` = 1, notiamo comunque una concentrazione di `label` verso il valore 1, ma in modo meno marcato rispetto alle altre variabili considerate. In ogni caso, anche la presenza del protocollo HTTPS può essere un indicatore della presenza di un URL di phishing.

Label	**IsHTTPS = 0**	**IsHTTPS = 1**
0	5470	4415
1	0	13695

Tabella 13: Tabella di contingenza tra `label` e `IsHTTPS`



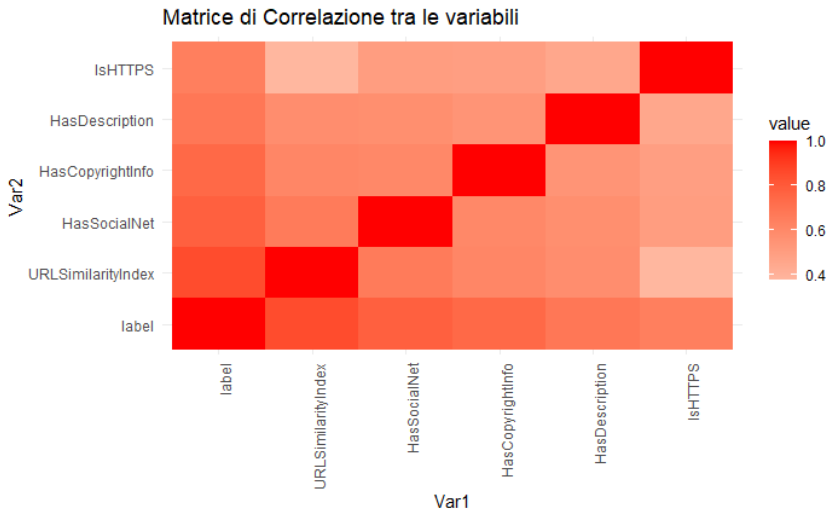
5.3.2 Coefficiente di Correlazione Campionario

Il calcolo dei coefficienti di correlazione tra le variabili considerate e la variabile target è stato sfruttato nel capitolo 2.2 per concentrarci sulle feature con il maggiore valore predittivo, piuttosto che sulle 56 feature iniziali. Saranno di seguito riportati i risultati ottenuti unicamente in riferimento alle feature considerate:

Feature	**Correlazione con label**
URLSimilarityIndex	0.856135403
HasSocialNet	0.778829467
HasCopyrightInfo	0.746076118
HasDescription	0.683989489
IsHTTPS	0.646885058

Tabella 14: Coefficienti di correlazione tra la variabile target e le feature considerate.

Risulta tuttavia interessante valutare le interrelazioni tra le feature considerate. La presenza di relazioni interne potrebbe causare ridondanze per il modello predittivo.



La matrice di correlazione riportata consente di osservare le relazioni lineari tra la variabile target `label` e le feature selezionate, nonché le interrelazioni tra le feature stesse. L'intensità delle relazioni è rappresentata cromaticamente: i valori prossimi a 1 (rosso intenso) indicano una forte correlazione positiva, mentre i valori prossimi a 0 (toni più chiari) indicano una relazione debole o nulla. Dall'analisi di tale matrice emerge che le feature selezionate (escludendo la variabile target) abbiano una correlazione perlopiù moderata ($\leq |0.6|$), e di conseguenza potrebbe esserci una collinearità (non troppo forte) nel dataset. Questo significa che alcune variabili condividono informazioni, ma non in modo ridondante.

6 Creazione del Modello

Conclusa la fase di analisi dei dati, si procede con la realizzazione di un modello predittivo, che dovrà essere in grado di prevedere se gli URL contenuti in un dataset sono URL di phishing. Prima di addestrare il modello, è necessaria una fase preliminare di preparazione dei dati, affinché il modello possa funzionare correttamente e minimizzare gli errori.

6.1 Data Preparation

La prima operazione da considerare è la **pulizia dei dati**. Spesso i dati grezzi contengono errori, duplicati o valori anomali che potrebbero compromettere le prestazioni del modello. Dall'analisi condotta emerge tuttavia che il dataset considerato non presenta valori mancanti o duplicati per nessuna delle feature osservate. Inoltre, non sono presenti outlier per nessuna delle feature osservate.

Una volta completata la fase di pulizia dei dati, il dataset viene suddiviso in due insiemi: il training set (70% dei dati) e il test set (30% dei dati). Questa suddivisione consente di addestrare il modello su un sottoinsieme dei dati e di valutarne le prestazioni su dati non visti. La funzione utilizzata per effettuare questa suddivisione è stata `sample.split()` dalla libreria `caTools` di R.

6.2 Addestramento del Modello

Divisi i dati in training set e test set, si procede con la creazione del modello. La fase preliminare nella costruzione di un modello predittivo riguarda la scelta del modello statistico da utilizzare per la classificazione. Successivamente, verranno analizzati i risultati del modello applicato al dataset suddiviso nei due gruppi definiti. Infine, si procederà con l'analisi delle performance del modello su un nuovo dataset di URL di phishing non ancora esplorato, per verificarne l'accuratezza su dati sconosciuti.

6.2.1 Regressione Logistica

La regressione logistica (o anche modello *logit*) è un modello statistico che stima la probabilità che si verifichi un evento, sulla base di un determinato set di dati di variabili indipendenti. Poiché il risultato è una probabilità, allora la variabile dipendente si trova nell'intervallo $[0, 1]$. È rappresentata dalle seguenti formule:

1. La funzione logit è data da:

$$\text{Logit}(p_i) = \ln\left(\frac{p_i}{1 - p_i}\right)$$

2. L'equazione del modello logit è:

$$\ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k$$

La prima formula descrive il logit come il logaritmo del rapporto tra la probabilità di successo e quella di insuccesso. La seconda formula esprime il logit come una combinazione lineare delle variabili indipendenti. In questa equazione di regressione logistica, $\text{Logit}(p_i)$ è la

variabile dipendente, o risposta, e x è la variabile indipendente. Il parametro β, o coefficiente, in questo modello viene comunemente stimato tramite la *stima di massima verosimiglianza (MLE, Maximum Likelihood Estimation)*. Questo metodo stima diversi valori di β attraverso più iterazioni per ottimizzare il miglior adattamento (*best fit*) della probabilità logaritmica. Ogni iterazione genera una funzione di verosimiglianza logaritmica e la regressione logistica cerca di massimizzare questa funzione per trovare la migliore stima dei parametri. Una volta trovato il valore ottimale del coefficiente (o dei coefficienti, nel caso di più variabili indipendenti), è possibile calcolare, registrare e sommare le probabilità condizionate per ciascuna osservazione per ottenere la probabilità prevista.

Odds Ratio. L'interpretazione della probabilità logaritmica può essere complessa. Per questa ragione, l'esponenziale delle stime β viene utilizzato per trasformare i risultati in un *Odds Ratio* (OR, o rapporto crociato), che rappresenta la probabilità che si verivichi un certo risultato in un determinato evento rispetto alla probabilità si verifichi in assenza di tale evento.

- $OR > 1$: l'evento è associato a una maggiore probabilità di generare un risultato specifico;
- $OR < 1$: l'evento è associato a una minore probabilità di generare tale risultato.

Differenze con Regressione Lineare. I modelli di regressione lineare vengono utilizzati per identificare la relazione tra una variabile dipendente continua e una o più variabili indipendenti. Quando c'è solo una variabile indipendente e una variabile dipendente, si parla di regressione lineare semplice, ma all'aumentare del numero di variabili indipendenti, si parla di regressione lineare multipla.

Anche la regressione logistica, invece, viene utilizzata per stimare la relazione tra una variabile dipendente e una o più variabili indipendenti, con la differenza che viene utilizzata specificamente per la previsione di variabili categoriche, come nel caso analizzato in cui deve essere previsto il valore della variabile `label`, che indica se un URL è di phishing o meno e che assume unicamente valori 0/1. L'unità di misura differisce anche dalla regressione lineare, poiché produce una probabilità, ma la funzione logit trasforma la curva S in linea retta.

6.2.2 Risultati del Modello

Il modello risulta avere degli ottimi risultati sui dati di test al termine dell'addestramento. I risultati sono sintetizzati all'interno della seguente matrice di confusione:

Prediction	**Istanza Negativa**	**Istanza Positiva**
Istanza Predetta come Negativa	(TN) 2963	(FN) 0
Istanza Predetta come Positiva	(FP) 2	(TP) 4109

Tabella 15: Matrice di confusione del modello

Per valutare la bontà del modello di classificazione, vengono utilizzate diverse metriche che permettono di misurare la sua capacità di predire correttamente le istanze in base alla matrice di confusione. Le principali metriche considerate sono le seguenti:

- **Accuratezza**: L'accuratezza rappresenta la proporzione di previsioni corrette rispetto al totale delle previsioni effettuate. È una metrica semplice e intuitiva che fornisce un'indicazione generale delle prestazioni del modello. Tuttavia, può essere meno utile in presenza di classi sbilanciate. È definita dalla seguente formula:

$$\text{Accuratezza} = \frac{TP + TN}{TP + TN + FP + FN}$$

  Il modello mostra avere un'accuratezza di **0.9997**, andando a classificare correttamente la quasi totalità delle osservazioni del test set.

- **Precisione**: La precisione misura la proporzione di veri positivi rispetto a tutte le istanze classificate come positive dal modello. È particolarmente utile quando l'errore di classificare un'istanza negativa come positiva è costoso, come nel caso di un falso positivo. È definita dalla seguente formula:

$$\text{Precisione} = \frac{TP}{TP + FP}$$

  La precisione del modello risulta pari a **0.9995**: il modello ha classificato correttamente quasi tutte le istanze positive del test set, con solo 2 istanze FP.

- **Richiamo**: Il richiamo, anche conosciuto come sensibilità, misura la proporzione di veri positivi correttamente identificati dal modello rispetto a tutte le istanze positive reali. È definita dalla seguente formula:

$$\text{Richiamo} = \frac{TP}{TP + FN}$$

  Il richiamo ottenuto è di **1**, il che indica che sono stati riconosciuti tutti gli URL di phishing.

- **Specificità**: La specificità misura la capacità del modello di identificare correttamente le istanze negative. È utile per comprendere quanto bene il modello riesca a distinguere tra le classi negative. È definita dalla seguente formula:

$$\text{Specificità} = \frac{TN}{TN + FP}$$

  Il modello ha una specificità di **0.9993**, riuscendo a predire correttamente la quasi totalità di istanze negative.

- **F1-Score**: L'F1-Score è la media armonica tra precisione e richiamo, e fornisce una misura bilanciata delle performance del modello. È particolarmente utile in presenza di un dataset sbilanciato, dove una metrica da sola potrebbe non riflettere correttamente le prestazioni. È definita dalla seguente formula:

$$F1\text{-Score} = 2 \cdot \frac{\text{Precisione} \cdot \text{Richiamo}}{\text{Precisione} + \text{Richiamo}}$$

  L'F1-Score registrato del modello è di **0.9997**, suggerendo un modello ben bilanciato e molto accurato.

In conclusione, il modello ha ottenuto ottimi risultati in tutte le metriche principali, con un'accuratezza e un F1-Score molto alti, un richiamo perfetto (1), e una precisione che mostra pochi falsi positivi. La bassa presenza di falsi positivi (FP) e falsi negativi (FN) suggerisce che il modello è molto robusto nel distinguere tra URL di phishing e URL legittimi. Il modello è ben addestrato e in grado di fare previsioni accurate e affidabili.

6.3 Applicazione del Modello a un Dataset Sconosciuto

Sebbene il modello abbia ottenuto ottimi risultati sul Dataset proposto, vi è comunque il rischio di overfitting, che potrebbe far sì che il modello risulti poco efficace con dati sconosciuti. È stato quindi analizzato un altro Dataset, sempre reperito dal sito UC Irvine Machine Learning Repository riguardo altri dati di phishing; il Dataset presenta altre 23580 tuple, le quali presentano le stesse feature di quelle analizzate. I risultati ottenuti sono sintetizzati nella seguente matrice di confusione:

Prediction	**Istanza Negativa**	**Istanza Positiva**
Istanza Predetta come Negativa	(TN) 9539	(FN) 0
Istanza Predetta come Positiva	(FP) 4	(TP) 14037

Tabella 16: Matrice di confusione del modello sui nuovi dati

Da una prima lettura della matrice, si osservano ottimi risultati nella classificazione degli URL di phishing anche su questo nuovo dataset. Verranno di seguito calcolate le metriche di bontà del modello anche su questi dati:

- **Accuratezza**: il modello mostra avere un'accuratezza di **0.9998**, andando a classificare correttamente 23476 tuple sulle 23480 che compongono il Dataset.
- **Precisione**: la precisione del modello risulta pari a **0.9997**: anche in questo caso, il modello ha classificato correttamente quasi tutte le istanze positive del test set, con solo 4 istanze FP.
- **Richiamo**: il richiamo ottenuto è di **1**, il che indica che sono stati nuovamente riconosciuti tutti gli URL di phishing.
- **Specificità**: il modello ha una specificità di **0.9995**, riuscendo a predire correttamente la quasi totalità di istanze negative anche per questo Dataset.
- **F1-Score**: l'F1-Score registrato del modello è di **0.9998**. Anche ion questo caso il modello risulta ben bilanciato e molto accurato.

7 Creazione e Valutazione di un Dataset Sintetico

Dall'addestramento e training del modello su dati reali, è stato appurato che lavori con incredibile accuratezza e precisione, e con un margine di errore infimo. In questa nuova sezione andremo a generare e valutare la creazione di un Dataset sintetico, che mantenga le stesse 56 feature del Dataset originario, ma con dati generati in modo casuale, seppur plausibile. Per la generazione del Dataset è stato utilizzato **GPT-4o**, che partendo da cinque righe casuali selezionate a partire dal Dataset originale ha generato un nuovo Dataset con 23580 osservazioni. L'obbiettivo di questo lavoro statistico è sintetizzato nelle seguenti Research Question:

- **RQ1**: i dati generali mantengano le stesse caratteristiche dei dati reali (media, mediana, varianza...)?
- **RQ2**: eventuali relazioni tra le variabili sono state mantenute in fase di generazione del Dataset?
- **RQ3**: le feature generate dai LLM possono essere ricondotte a distribuzioni statistiche note?

7.1 Analisi delle Correlazioni con la Variabile Target

Risulta particolarmente utile valutare il coefficiente di correlazione tra tutte le feature e la variabile target (label) per determinare se le feature che erano rilevanti per l'addestramento del modello sui dati reali mantengono una rilevanza predittiva anche nel dataset sintetico. Anche in questo caso è necessario convertire le variabili categoriche `Title` e `Domain` dovranno prima essere trasformate in fattori e poi in valore numerico attraverso il Label Encoding, in quanto il calcolo del coefficiente di correlazione funziona solo su variabili numeriche.

Analisi dei Risultati. Dalla tabella 17 è evidente che i coefficienti di correlazione tra la variabile target e le altre feature non sono stati mantenuti. In particolare possiamo notare che i valori vanno da un minimo di -0.0154567 a un massimo di 0.0117329, molto più bassi rispetto ai valori reali. L'LLM non ha quindi preservato l'intensità delle relazioni tra la variabile target e le altre variabili.

La mancata preservazione delle correlazioni suggerisce che i modelli addestrati su dati sintetici potrebbero non essere in grado di riprodurre con accuratezza gli stessi risultati ottenuti grazie a modelli addestrati su dati reali. Inoltre, i modelli addestrati su dati sintetici potrebbero non cogliere pattern rilevanti, utili per la classificazione. Verranno tuttavia analizzate altre caratteristiche del Dataset sintetico, per verificare se altre proprietà del Dataset reale siano state rispettate, come ad esempio le distribuzioni o gli indici di sintesi.

Feature	Correlazione (Sintetico)	Correlazione (Reale)
NoOfCSS	0.011732900	0.424377445
Robots	0.011381460	0.396208704
ObfuscationRatio	0.008386646	-0.043464578
IsResponsive	0.008206670	0.564786719
LetterRatioInURL	0.008130126	-0.329955655
SpacialCharRatioInURL	0.008049828	-0.517029402
NoOfDegitsInURL	0.007695651	-0.287961146
HasSocialNet	0.005938482	0.778829467
URLTitleMatchScore	0.005567498	0.529816070
NoOfImage	0.005286889	0.317604924
NoOfEmptyRef	0.005208557	0.135309322
NoOfExternalRef	0.004529260	0.369758258
TLDLength	0.004392434	-0.075725589
DegitRatioInURL	0.003879855	-0.419638475
IsDomainIP	0.003790634	-0.056391743
Pay	0.003739259	0.371572507
CharContinuationRate	0.002772281	0.448866608
NoOfJS	0.002435365	0.541923412
LargestLineLength	0.002097734	-0.034737240
NoOfPopup	0.002007492	0.046738726
URLCharProb	0.001717780	0.454740398
TLDLegitimateProb	0.001698498	0.097092703
NoOfURLRedirect	0.001162546	-0.038133810
NoOfAmpersandInURL	0.000713928	-0.048959571
IsHTTPS	-0.000085120	0.646885058
NoOfQMarkInURL	-0.000380544	-0.170552587
NoOfEqualsInURL	-0.000534872	-0.111961552
HasTitle	-0.000572285	0.473294908
HasPasswordField	-0.000903722	0.149780257
HasObfuscation	-0.002215555	-0.050892372
URLSimilarityIndex	-0.002815106	0.856135403
NormalizedURLSimilarityIndex	-0.002815106	0.856135403
Bank	-0.002978013	0.210757171
Domain_numeric	-0.003035555	0.451279836
HasFavicon	-0.003163430	0.506485685
LineOfCode	-0.003749212	0.351496677
Crypto	-0.003775081	0.101112015
NoOfSubDomain	-0.004893369	-0.004390171
HasSubmitButton	-0.005226314	0.586002993
DomainTitleMatchScore	-0.005676094	0.570249492
NoOfSelfRedirect	-0.005964507	-0.089700554
NoOfLettersInURL	-0.006541745	-0.305340519
DomainLength	-0.007160959	-0.262763580
HasHiddenFields	-0.007356669	0.516743503
HasCopyrightInfo	-0.007438808	0.746076118
Title_numeric	-0.007507694	0.319122238
NoOfSelfRef	-0.008037733	0.405329350
HasExternalFormSubmit	-0.009488803	0.170393099
HasDescription	-0.009697983	0.683989489
NoOfObfuscatedChar	-0.010413890	-0.040996807
NoOfiFrame	-0.011055290	0.251850614
URLLength	-0.011270490	-0.288858679
NoOfOtherSpecialCharsInURL	-0.015456700	-0.406440494

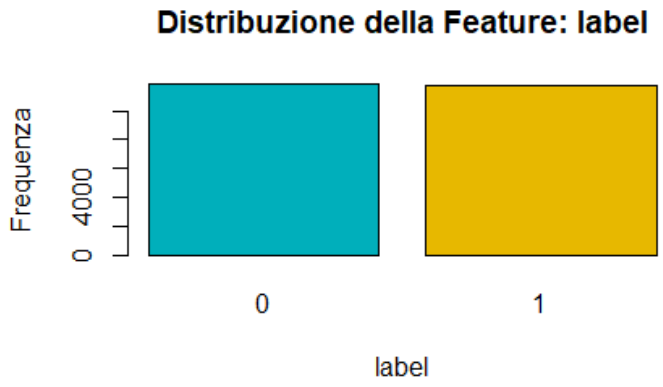
Tabella 17: Confronto tra le correlazioni tra feature e variabile target ”label” nei dataset sintetico e reale.

7.2 Distribuzioni di Frequenza

Verranno di seguito calcolate le distribuzioni di frequenza dei dati sintetici, e confrontate con le distribuzioni reali. Anche in questo caso, considereremo le feature `URLSimilarityIndex`, `HasSocialNet`, `HasCopyrightInfo`, `IsHTTPS` e la variabile target `label`, in quanto utilizzate in fase di addestramento del modello. Considereremo i seguenti intervalli di `URLSimilarityIndex`:

$$[0, 10), [10, 20), [20, 30), [30, 40), [40, 50), [50, 60), [60, 70), [70, 80), [80, 90), [90, 100)$$

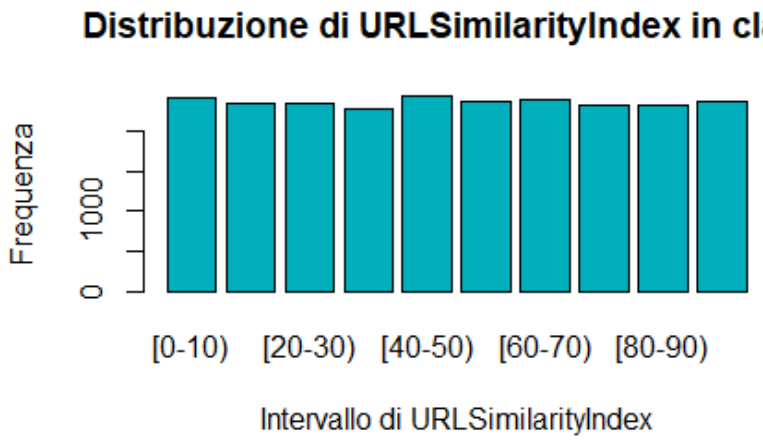
7.2.1 Distribuzione di `label`



Classe	Frequenza Assoluta	Frequenza Relativa	Frequenza Percentuale
syn_label = 0	11832	0.5018	50.18%
syn_label = 1	11748	0.4982	49.82%
real_label = 0	9885	0.4192	41.92%
real_label = 1	13695	0.5807	58.07%

Le etichette sintetiche non riflettono la distribuzione delle etichette reali: la distribuzione sintetica appare quasi perfettamente bilanciata, con una differenza di appena lo 0.36% tra le due classi, mentre la distribuzione reale ha una classe dominante, cioè `label` = 1, con una differenza del 16.15% rispetto alla classe `label` = 0.

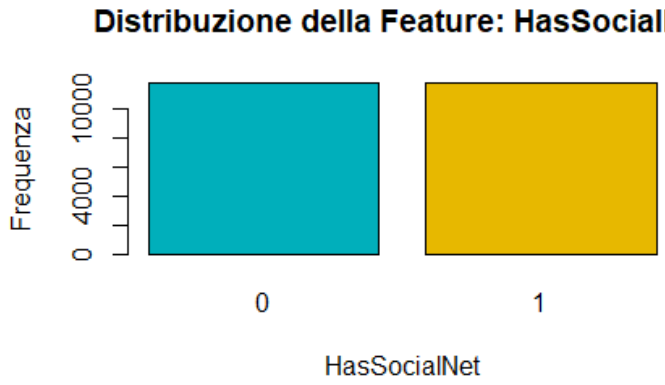
7.2.2 Distribuzione di `URLSimilarityIndex`



Intervallo	Frequenza Assoluta	Frequenza Relativa	Frequenza Percentuale
[0-10)	2413	0.1023	10.23%
[10-20)	2341	0.0992	9.92%
[20-30)	2347	0.0995	9.95%
[30-40)	2269	0.0962	9.62%
[40-50)	2435	0.1032	10.32%
[50-60)	2376	0.1007	10.07%
[60-70)	2384	0.1011	10.11%
[70-80)	2329	0.0987	9.87%
[80-90)	2322	0.0984	9.84%
[90-100)	2364	0.1002	10.02%

Mentre nei dati reali la distribuzione era fortemente sbilanciata verso l'intervallo $[90, 100)$, che rappresentava il 58.91% delle osservazioni totali, mentre gli altri intervalli avevano frequenze sotto il 10%, con i dati sintetici la distribuzione risulta quasi perfettamente bilanciata, con ciascun intervallo avente una frequenza percentuale intorno al 10%.

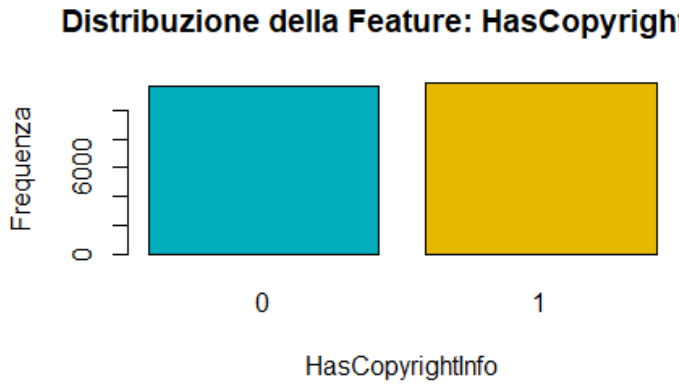
7.2.3 Distribuzione di `HasSocialNet`



Classe	Frequenza Assoluta	Frequenza Relativa	Frequenza Percentuale
syn_HasSocialNet = 0	11786	0.4998	49.98%
syn_HasSocialNet = 1	11794	0.5001	50.02%
real_HasSocialNet = 0	12694	0.5383	53.83%
real_HasSocialNet = 1	10886	0.4617	46.17%

La distribuzione di `HasSocialNet` è perfettamente bilanciata per quanto riguarda i dati sintetici, con una differenza di appena lo 0.04% tra la classe 0 e la classe 1, mentre per i dati reali, la differenza risulta del 7.66%, leggermente sbilanciata verso la classe 0.

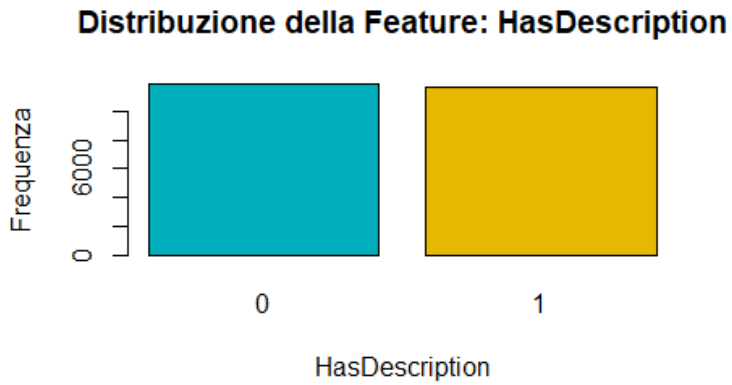
7.2.4 Distribuzione di `HasCopyrightInfo`



Classe	Frequenza Assoluta	Frequenza Relativa	Frequenza Percentuale
syn_HasCopyrightInfo = 0	11706	0.4964	49.64%
syn_HasCopyrightInfo = 1	11874	0.5036	50.36%
real_HasCopyrightInfo = 0	12053	0.5111	51.12%
real_HasCopyrightInfo = 1	11527	0.4888	48.88%

Anche nel caso di `HasCopyrightInfo`, la distribuzione dei dati generati è quasi perfettamente bilanciata, con una differenza del 0.72% tra le due classi. Tuttavia, nei dati reali non si osserva un forte sbilanciamento: con una differenza del 2.24%, i dati appaiono bilanciati e non molto distanti dai dati sintetici generati. Questa differenza potrebbe incidere sulle prestazioni del modello, ma non in modo così netto come per le altre feature osservate.

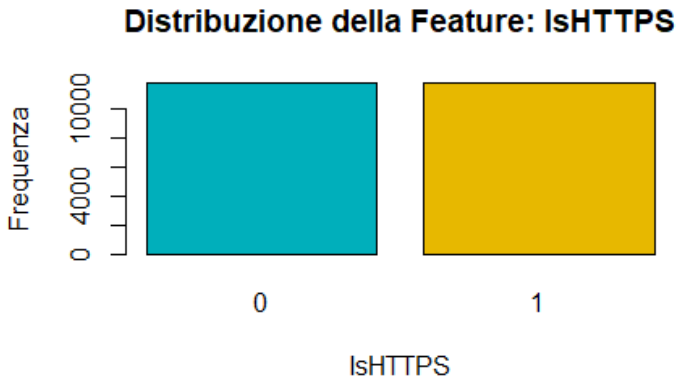
7.2.5 Distribuzione di `HasDescription`



Classe	Frequenza Assoluta	Frequenza Relativa	Frequenza Percentuale
syn_HasDesription = 0	11884	0.5040	50.40%
syn_HasDesription = 1	11696	0.4960	49.60%
real_HasDesription = 0	13076	0.5545	55.45%
real_HasDesription = 1	10504	0.4455	44.55%

La distribuzione di `HasDescription` sintetico mostra una distribuzione quasi perfetta tra le due classi, con una differenza dello 0.8%, mentre nei dati reali si osserva una discrepanza del 10.9%. Anche in questo caso, non sono state mantenute le stesse relazioni osservate sui dati reali.

7.2.6 Distribuzione di `IsHTTPS`



Classe	Frequenza Assoluta	Frequenza Relativa	Frequenza Percentuale
syn_IsHTTPS = 0	11791	0.5	50.0%
syn_IsHTTPS = 1	11789	0.5	50.0%
real_IsHTTPS = 0	5470	0.2319	23.20%
real_IsHTTPS = 1	18110	0.7680	76.80%

La distribuzione di `IsHTTPS` sintetico mostra un perfetto bilanciamento tra le due classi, indicando che il generatore ha creato una distribuzione equa tra i due valori. Nei dati reali si osserva un forte sbilanciamento, con la classe 1 avente frequenza percentuale pari al 76.80%. La distribuzione dei dati sintetici non è realistica e potrebbe influenzare negativamente le prestazioni del modello.

7.2.7 Analisi dei Risultati

Osservando le distribuzioni di frequenza delle feature prese in esame, appare evidente che l'LLM abbia dato molta importanza al bilanciamento delle variabili in fase di generazione dei dati sintetici. Questo, seppure permetta che ogni classe sia equamente rappresentata, tuttavia non riflette le relazioni reali che intercorrono tra le variabili e potrebbe portare a risultati inaccurati del modello addestrato sui dati reali.

7.3 Funzione di Distribuzione Empirica

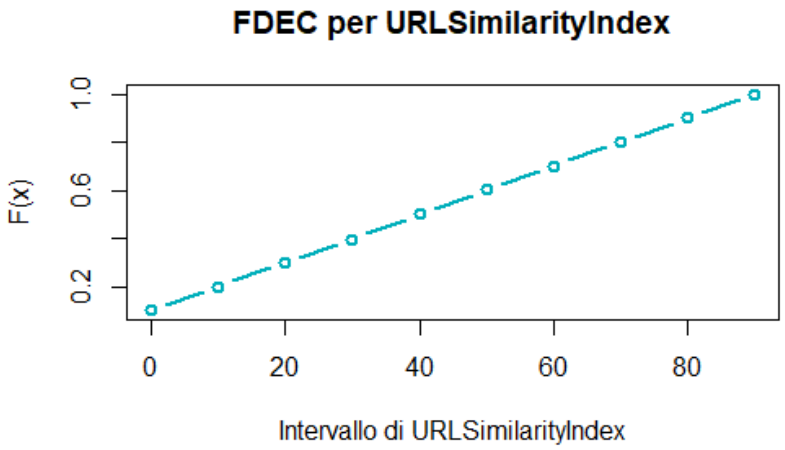
Consideriamo la feature `URLSimilarityIndex`. Sono state individuate $k = 10$ classi:

$$C_1 = [0, 10), \quad C_2 = [10, 20), \quad C_3 = [20, 30), \quad C_4 = [30, 40), \quad C_5 = [40, 50),$$
$$C_6 = [50, 60), \quad C_7 = [60, 70), \quad C_8 = [70, 80), \quad C_9 = [80, 90), \quad C_{10} = [90, 100)$$

La FDEC assumerà i seguenti valori:

Indice	C_i	n_i	f_i	F_i
1	[0-10)	366	$\frac{2413}{23580}$	$\frac{2413}{23580}$
2	[10-20)	754	$\frac{2341}{23580}$	$\frac{4754}{23580}$
3	[20-30)	952	$\frac{2347}{23580}$	$\frac{7101}{23580}$
4	[30-40)	1083	$\frac{2269}{23580}$	$\frac{9370}{23580}$
5	[40-50)	1242	$\frac{2435}{23580}$	$\frac{11805}{23580}$
6	[50-60)	1540	$\frac{2376}{23580}$	$\frac{14181}{23580}$
7	[60-70)	1725	$\frac{2384}{23580}$	$\frac{16565}{23580}$
8	[70-80)	1352	$\frac{2329}{23580}$	$\frac{18894}{23580}$
9	[80-90)	675	$\frac{2322}{23580}$	$\frac{21216}{23580}$
10	[90-100)	13891	$\frac{2364}{23580}$	$\frac{23580}{23580}$

$$F(x) = \begin{cases} 0, & x < 0 \\ \frac{2413}{23580} = 0.1023, & 0 \leq x < 10 \\ \frac{4754}{23580} = 0.2016, & 10 \leq x < 20 \\ \frac{7101}{23580} = 0.3011, & 20 \leq x < 30 \\ \frac{9370}{23580} = 0.3974, & 30 \leq x < 40 \\ \frac{11805}{23580} = 0.5006, & 40 \leq x < 50 \\ \frac{14181}{23580} = 0.6014, & 50 \leq x < 60 \\ \frac{16565}{23580} = 0.7025, & 60 \leq x < 70 \\ \frac{18894}{23580} = 0.8013, & 70 \leq x < 80 \\ \frac{21216}{23580} = 0.8997, & 80 \leq x < 90 \\ 1, & x \geq 100 \end{cases}$$



7.4 Indici di Sintesi

7.4.1 Variabili Binarie

Feature	Media	Varianza	Deviazione Standard	Moda
syn_label	0.4982	0.25	0.5	0
real_label	0.5808	0.2435	0.4934	1
syn_HasSocialNet	0.5002	0.25	0.5	1
real_HasSocialNet	0.4617	0.2485	0.4985	0
syn_HasCopyrightInfo	0.5036	0.25	0.5	0
real_HasCopyrightInfo	0.4888	0.2499	0.4999	0
syn_HasDescription	0.496	0.25	0.5	0
real_HasDescription	0.4455	0.247	0.497	0
syn_IsHTTPS	0.5	0.25	0.5	0
real_IsHTTPS	0.768	0.1782	0.4221	1

Tabella 18: Indici di Sintesi delle variabile binarie.

L'analisi degli indici di sintesi conferma che l'LLM selezionato dà una grande importanza al bilanciamento dei valori nelle classi. La media per ogni feature è poressoché 0.5, che unita alla varianza di 0.25 (massima per le variabili binarie) sottolinea che i valori di ogni classe sono equamente rappresentati. Nel caso delle variabili reali osserviamo dei leggeri sbilanciamenti nella maggior parte delle classi, con un forte sbilanciamento in `IsHTTPS`.

7.4.2 Variabile Continua

Media, Moda e Mediana Campionaria. Per quanto riguarda `URLSimilarityIndex` (Tabella 19), l'osservazione degli indici di sintesi risulta più interessante rispetto alle variabili binarie. Una media di 49.91 indica un valore centrale. Dall'analisi delle distribuzioni di frequenza è evidente che questo indice non sia influenzato da valori esterni, in quanto la variabile risulta bilanciata. La mediana di 49.96 suggerisce che la distribuzione dei dati sia molto simmetrica, a differenza dei valori reali, in cui la mediana corrisponde al valore massimo (100). La moda, che rappresenta il valore più frequente appare relativamente basso rispetto agli altri indici.

Feature	Media	Mediana	Moda
syn_URLSimilarityIndex	49.9101	49.957	13.54152
real_URLSimilarityIndex	79.5097	100	100

Tabella 19: Misure di Centralità della variabile continua.

Quartili. La distribuzione dei valori sintetici si mostra più equilibrata dei valori reali di `URLSimilarityIndex` anche per quanto riguarda i quartili. Lo scambio interquartile risulta avere valore pari a:

$$IQR = Q_3 - Q_1 = 74.7598 - 24.8091 = 49.9507$$

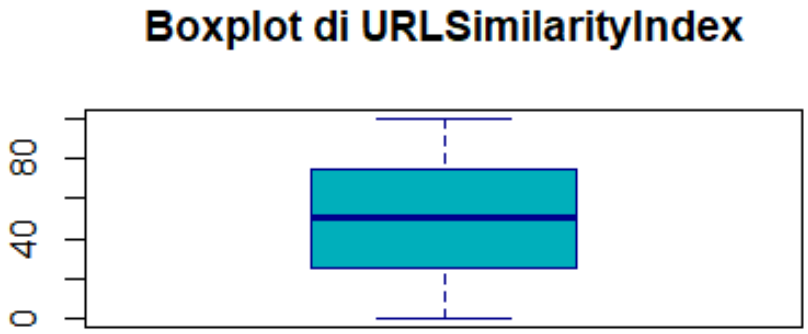
Gli outlier sono definiti come valori esterni all'intervallo $[Q_1 - 1.5 \cdot IQR, Q3 + 1.5 \cdot IQR]$, quindi gli outlier della variabile considerata dovranno essere ricercati all'esterno dell'intervallo:

$$[-50.11687, 149.6859]$$

Anche in questo caso non sono presenti outlier.

Feature	Q_0	Q_1	Q_2	Q_3	Q_4
syn_URLSimilarityIndex	0	24.8091	49.957	74.7598	100
real_URLSimilarityIndex	1.292906	59.65332	100	100	100

Tabella 20: Quartili della variabile continua.



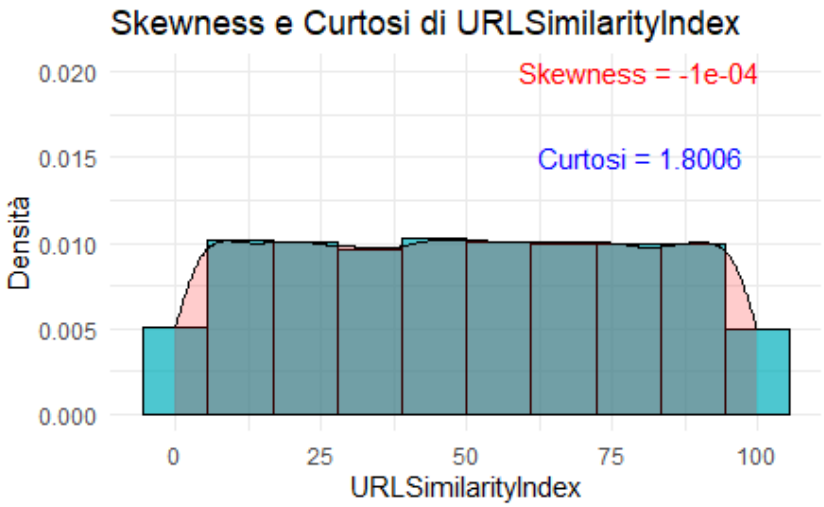
Varianza e Deviazione Standard. I valori di varianza e deviazione standard appaiono molto elevati, pari a rispettivamente 834.3124 e 28.8845, indicano la presenza di valori lontani dalla media, a conferma del fatto che il dataset generato risulta quasi perfettamente bilanciato.

Feature	**Varianza**	**Deviazione Standard**
syn_URLSimilarityIndex	834.3124	28.8845
real_URLSimilarityIndex	793.6245	28.1713

Tabella 21: Misure di Dispersione della variabile continua.

Skewness e Curtosi Campionaria. Poiché la skewness campionaria assume un valore pari a -0.0001, cioè molto vicino a 0, allora possiamo assumere che la distribuzione dei dati sia simmetrica rispetto alla media.
Un valore della curtosi pari a 1.8006 indica una distribuzione platicurtica, ciè con code più leggere rispetto ad una distribuzione normale.

Feature	**Skwewness**	**Curtosi**
syn_URLSimilarityIndex	-0.0001	1.8006
real_URLSimilarityIndex	-1.0601	-0.2139



7.5 Covarianza e Correlazione Campionaria

7.5.1 Covarianza delle feature sintetiche con label

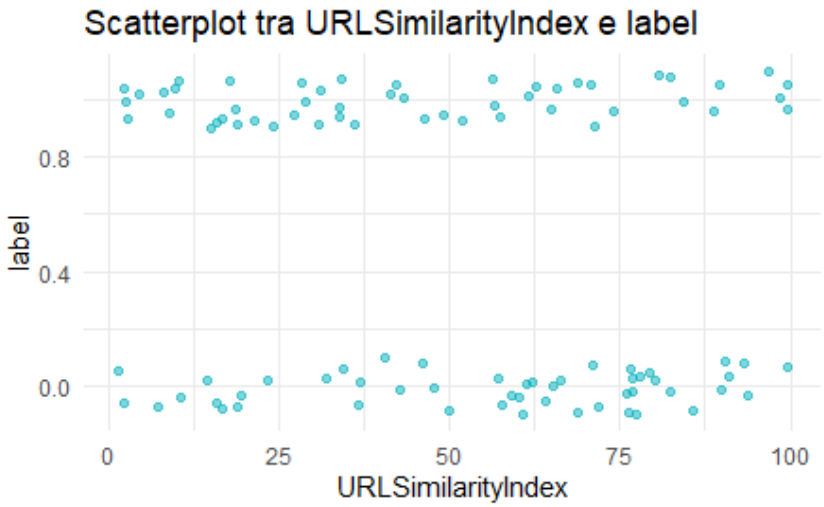
Utilizziamo la covarianza campionaria per valutare la relazione tra la variabile target `label` e le feature selezionate. Una covarianza positiva indica che un aumento nei valori di una variabile tende a essere associato a un aumento nei valori dell'altra, mentre una covarianza negativa suggerisce una relazione inversa.

Feature	**Covarianza con label**
URLSimilarityIndex	-0.0407
HasSocialNet	0.0015
HasCopyrightInfo	-0.0019
HasDescription	-0.0024
IsHTTPS	0

URLSimilarityIndex e label. La covarianza tra `URLSimilarityIndex` e `label` è pari a -0.0407, che è molto debole e negativa rispetto alla covarianza tra le due variabili registrata su dati reali (pari a 11.901).
Sia dalla tabella di contingenza (Tabella 22) che dallo scatterplot si osserva una distribuzione che appare abbastanza uniforme tra le due classi. Non emergono picchi significativi, il che significa che `URLSimilarityIndex` non discrimina in modo netto tra le due classi.

Label	**[0,10)**	**[10,20)**	**[20,30)**	**[30,40)**	**[40,50)**	**[50,60)**	**[60,70)**	**[70,80)**	**[80,90)**	**[90,100)**
0	1235	1147	1165	1135	1200	1229	1200	1170	1130	1221
1	1178	1194	1182	1134	1235	1147	1184	1159	1192	1143

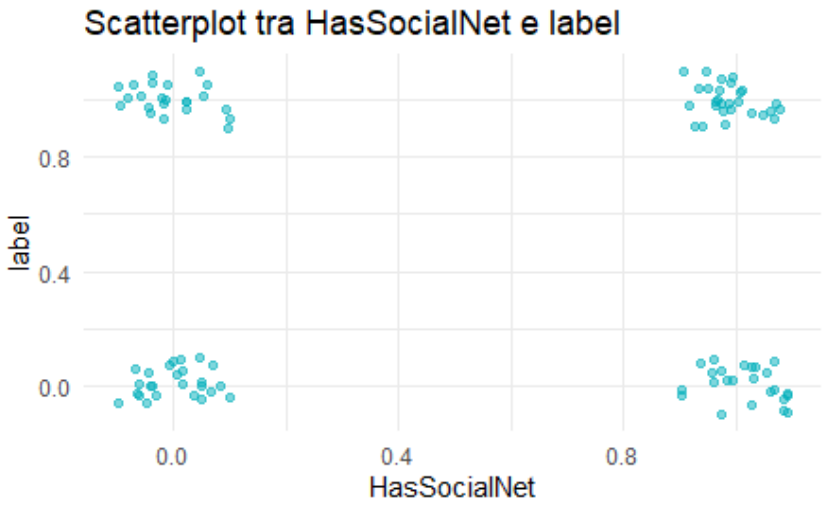
Tabella 22: Tabella di contingenza tra `label` e `URLSimilarityIndex`



HasSocialNet e label. La covarianza tra `HasSocialNet` e `label` è pari a 0.0015, indicando una relazione molto bassa e positiva tra le due variabili.
Osservando sia la tabella 23 che il grafico a dispersione, si osserva che non vi è corrispondenza tra i valori di HasSocialNet e label, indicando che `HasSocialNet` non è una variabile valida nella predizione dei valori di `label`.

Label	**HasSocialNet = 0**	**HasSocialNet = 1**
0	5949	5883
1	5837	5911

Tabella 23: Tabella di contingenza tra `label` e `HasSocialNet`

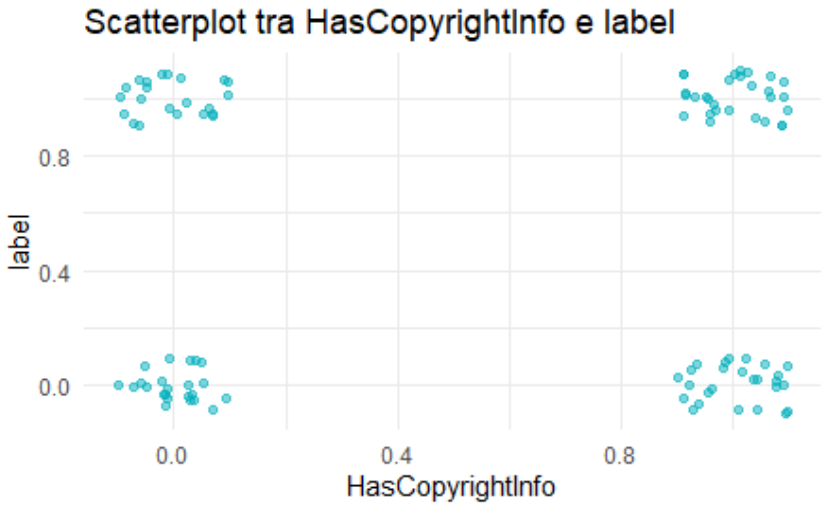


HasCopyrightInfo e label. La covarianza tra `HasCopyrightInfo` e `label` è pari a -0.0019, suggerendo una relazione molto bassa e negativa tra questa feature e la variabile target.

Anche in questo caso, la tabella di contingenza (Tabella 24) e il diagramma di dispersione mostrano che la variabile `HasCopyrightInfo` non è una valida discriminante per predire il valore di `label`.

Label	HasCopyrightInfo = 0	HasCopyrightInfo = 1
0	5830	6002
1	5876	5872

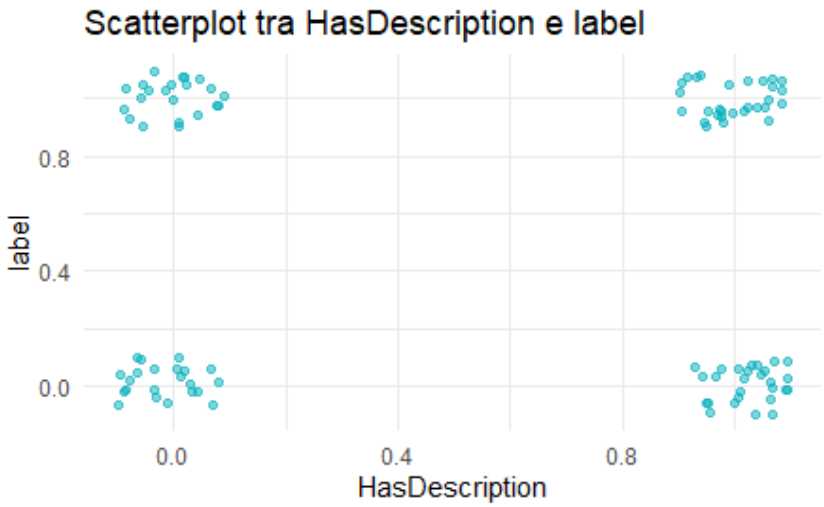
Tabella 24: Tabella di contingenza tra `label` e `HasCopyrightInfo`



HasDescription e label. La covarianza tra `HasDescription` e `label` è pari a -0.0024, indicando una relazione negativa molto bassa tra questa variabile e la variabile target.
Come nei casi precedenti, la tabella di contingenza 25 e lo scatterplot mostrano che i valori di HasDescription non possono essere utilizzati per determinare il valore di label in quanto le distribuzioni sono abbastanza bilanciate e non mostrano una differenza marcata tra i valori di `HasDescription` e le classi di `label`.

Label	HasDescription = 0	HasDescription = 1
0	5906	5926
1	5978	5770

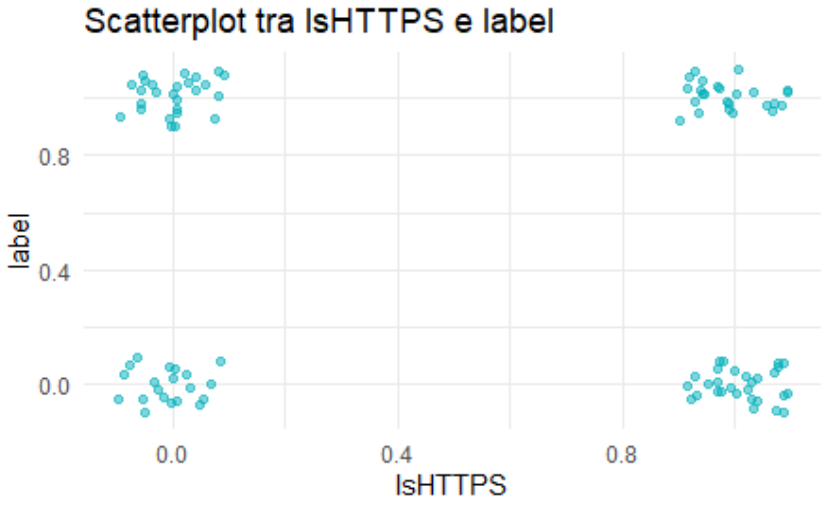
Tabella 25: Tabella di contingenza tra `label` e `HasDescription`



IsHTTPS e label. La covarianza tra `IsHTTPS` e `label` è pari a 0, valore che suggerisce che non vi è alcuna relazione lineare tra le due variabili. La variabile `IsHTTPS` risulta dunque inefficace nella predizione dei valori di `label` con i dati generati.

Label	**IsHTTPS = 0**	**IsHTTPS = 1**
0	5916	5916
1	5875	5873

Tabella 26: Tabella di contingenza tra `label` e `IsHTTPS`



7.5.2 Correlazione Campionaria

Si riporteranno di seguito i risultati ottenuti nella sezione 7.1, con un particolare focus sulle variabili che per i dati reali hanno un maggiore potere predittivo. Dalla tabella 27 appare chiaro che nessuna delle feature considerate ha una forte correlazione con la variabile target, andando a confermare i risultati ottenuti nella sezione precedente.

Feature	Correlazione con "label"
URLSimilarityIndex	-0.0028
HasSocialNet	0.0059
HasCopyrightInfo	-0.0074
HasDescription	-0.0097
IsHTTPS	0.0001

Tabella 27: Coefficienti di Correlazione per le variabili sintetiche.

Vale la pena valutare le interrelazioni tra le feature considerate, per vedere se le altre feature mantengono tra di loro, almeno in parte, le relazioni presenti trai dati reali. Tuttavia dall'analisi della matrice 2 emerge che le feature selezionate non abbiano correlazione.

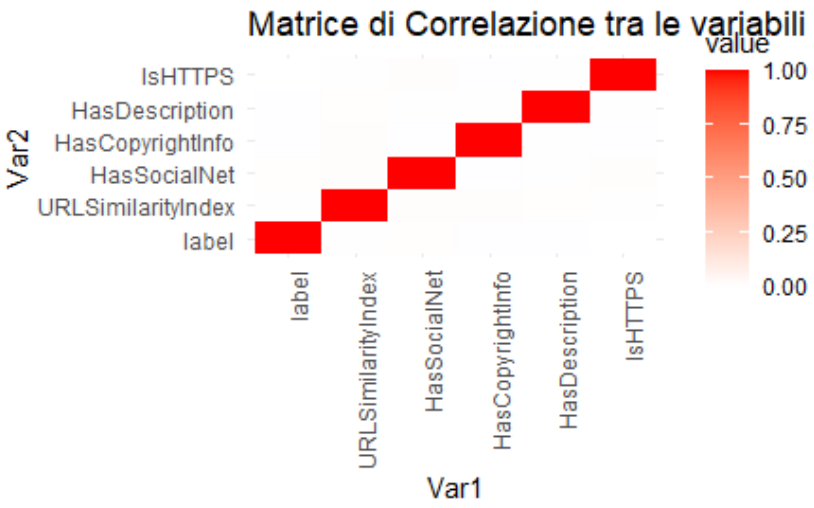


Figura 2: Matrice di Coprrelazione per le variabili sintetiche.

7.6 Verifica delle Ipotesi

7.6.1 Test del Chi-Quadrato

Con il test del chi-quadrato vogliamo verificare l'ipotesi che una certa popolazione, descritta da una variabile aleatoria X, sia caratterizzata da una certa funzione di distribuzione $F_X(x)$ con k parametri non noti da stimare. Denotiamo con H_0 l'ipotesi nulla soggetta a verifica, mentre con H_1 l'ipotesi alternativa. Il test considerato è bilaterale. Per un campione sufficientemente numeroso di ampiezza n, il test chi-quadrato bilaterale di misura α è il seguente:

- **Si accetti** l'ipotesi H_0 se:

$$\chi^2_{1-\alpha/2,r-k-1} < \chi^2 < \chi^2_{\alpha/2,r-k-1}$$

- **Si rifiuti** l'ipotesi H_0 se:

$$\chi^2 < \chi^2_{1-\alpha/2,r-k-1} \quad \text{oppure} \quad \chi^2 > \chi^2_{\alpha/2,r-k-1}$$

dove $\chi^2_{\alpha/2,r-k-1}$ e $\chi^2_{1-\alpha/2,r-k-1}$ sono le soluzioni di:

$$P(Q < \chi^2_{1-\alpha/2,r-k-1}) = \frac{\alpha}{2}, \quad P(Q < \chi^2_{\alpha/2,r-k-1}) = 1 - \frac{\alpha}{2}$$

7.6.2 P-Value

Il p-value è uno strumento fondamentale nei test di ipotesi statistici, utilizzato per valutare quanto i dati osservati siano compatibili con l'ipotesi nulla (H_0). Serve a quantificare l'evidenza contro H_0 e aiuta a decidere se rifiutare o meno H_0 in favore dell'ipotesi alternativa (H_1).

Sia l'ipotesi H_0 vera, il *p-value* è definito come la probabilità che la statistica del test $\hat{\xi}_n$ assuma un valore uguale o più estremo di quello effettivamente osservato $\hat{\xi}_{os}$. Poiché effettueremo un test bilaterale, allora il p-value sarà definito come:

$$p\text{-value} = 2P(\hat{\xi}_n \geq |\hat{\xi}_{os}| \mid H_0)$$

Se $p > \alpha$, l'ipotesi H_0 non può essere rifiutata, mentre se $p \leq \alpha$, l'ipotesi H_0 deve essere rifiutata.

7.6.3 Risultati sul Dataset Sintetico

Variabile Continua. Di seguito riportiamo i risultati ottenuti dal confronto della distribuzione della variabile continua `URLSimilarityIndex` con una distribuzione uniforme e una normale (Tabella 28).

Per quanto riguarda il confronto con la distribuzione uniforme, il test del chi-quadrato restituisce un p-value elevato ($0.4721 > 0.05$), il che indica che non abbiamo evidenze sufficienti per rifiutare l'ipotesi nulla. Possiamo dunque concludere che i dati sono compatibili con una distribuzione uniforme.

Analizzando invece il confronto con la distribuzione normale, osserviamo un p-value estremamente basso ($p < 2.2 \cdot 10^{-16}$), che suggerisce una forte deviazione rispetto a una distribuzione normale. Di conseguenza, rifiutiamo l'ipotesi nulla e concludiamo che i dati non seguono una distribuzione normale.

Test	X^2	p-value	Gradi di Libertà
Uniforme	8.6301	0.4721	9
Normale	4813.6	$2.2 \cdot 10^{-16}$	9

Tabella 28: Risultati del test del chi-quadrato per la variabile `URLSimilarityIndex`.

Variabili Binarie. Di seguito sarà verificato se le variabili binarie seguono una distribuzione binomiale, assumendo che le osservazioni siano indipendenti e seguano una sequenza di prove di Bernoulli.

Come mostrano i risultati riportati nella Tabella 29, tutte le variabili binarie presentano un p-value pari a 1. Questo indica che non ci sono evidenze sufficienti per rifiutare l'ipotesi nulla, ovvero che i dati siano compatibili con una distribuzione binomiale.

Feature	X^2	p-value	Gradi di Libertà
label	$2.7964 \cdot 10^{-28}$	1	1
HasSocialNet	$5.6128 \cdot 10^{-28}$	1	1
HasCopyrightInfo	0	1	1
HasDescription	0	1	1
IsHTTPS	0	1	1

Tabella 29: Risultati del test del chi-quadrato per le variabili binarie.

7.7 Risultati del Modello

Il modello applicato ai dati sintetici risulta avere dei risultati discreti, a differenza degli ottimi risultati riportati su dati reali.

Prediction	**Istanza Negativa**	**Istanza Positiva**
Istanza Predetta come Negativa	(TN) 11092	(FN) 11033
Istanza Predetta come Positiva	(FP) 740	(TP) 715

Tabella 30: Matrice di confusione del modello sui dati sintetici

Per valutare la bontà del modello di classificazione, verranno di seguito riportati i risultati ottenuti per le metriche considerate nel capitolo precedente.

- **Accuratezza**: il modello mostra avere un'accuratezza di **0.5007**, andando a classificare correttamente appena 11807 tuple delle 23480 che compongono il Dataset.
- **Precisione**: la precisione del modello risulta pari a **0.4914**, un valore che risulta dimezzato rispetto ai risultati ottenuti in precedenza.
- **Richiamo**: il richiamo ottenuto è di **0.06**, indice del fatto che su 11748 istanze appena 715 sono state correttamente classificate come negative.
- **Specificità**: il modello ha una specificità di **0.9374**, mostrandosi molto più efficace nel predire istanze negative rispetto alle positive.
- **F1-Score**: l'F1-Score registrato del modello è di **0.1069**, valore molto basso che indica che il modello ha una scarsa performance complessiva.

7.8 Conclusioni

In conclusione, è possibile affermare che i dati generati da GPT-4o non mantengono le stesse caratteristiche dei dati reali: infatti il LLM ha dato priorità al bilanciamento statistico dei dati (ad esempio, osserviamo distribuzioni uniformi tra le classi), andando tuttavia a compromettere le relazioni che esistevano tra le variabili considerate e la variabile target. Anche le relazioni che esistevano tra tutte le altre variabili non sono state mantenute. La perdita di relazioni tra le variabili suggerisce che i dati sintetici non riescono a catturare la struttura intrinseca dei dati reali. Di conseguenza, il modello addestrato su dati reali ha registrato pessime performance sui dati generati. Utilizzare i dati sintetici per addestrare nuovi modelli potrebbe non solo portare a errori significativi, ma anche introdurre bias sistematici.

8 Creazione e Valutazione di un Secondo Dataset Sintetico

È stato verificato che i dati generati da **GPT-4o** non abbiano mantenuto le relazioni originali tra le variabili, osservate nel dataset reale, né le stesse caratteristiche statistiche, come distribuzioni o indici, in quanto l'obiettivo dell'LLM è stato quello di generare dati bilanciati. È doveroso ricordare che il primo dataset sintetico è stato generato a partire da poche tuple selezionate in modo casuale, da cui l'LLM ha dedotto come generare l'intero dataset. Risulta quindi lecito domandarsi se GPT-4o riesca a produrre dati sintetici più fedeli al dataset reale nel caso in cui gli vengano fornite più informazioni di partenza. In quest'ultima fase di questo lavoro statistico si cercherà di rispondere alle seguenti domande:

- **RQ1**: la generazione di dati sintetici riesca a mantenere le stesse caratteristiche dei dati reali e le relazioni tra le variabili nel caso in cui all'LLM vengano fornite più informazioni al momento della generazione dei dati?
- **RQ2**: le feature generate in questo modo possono essere ricondotte a distribuzioni statistiche note?

8.1 Analisi delle Correlazioni con la Variabile Target

Anche in questo caso, il calcolo dei coefficienti di correlazione con la variabile target `label` può essere una valida risorsa per verificare se le feature che erano risultate rilevanti per la predizione della variabile target nei dati reali mantengano questa caratteristica anche su questi nuovi dati sintetici. Prima di procedere al calcolo dei coefficienti, le variabili categoriche `Title` e `Domain` dovranno prima essere trasformate in fattori e poi in valore numerico attraverso il Label Encoding.

Analisi dei Risultati. Come si evince dalla tabella 31, anche fornendo più informazioni in ingresso all'LLM, le relazioni tra la variabile target e le altre feature non sono state preservate. In particolare, si osserva che i valori di correlazione variano da un minimo di -0.0183172921 per `LetterRatioInURL` a un massimo di 0.0144971467 per `Crypto`, entrambi significativamente inferiori rispetto ai valori reali. Analogamente a quanto osservato nel caso del primo dataset sintetico, possiamo inferire che il modello addestrato sui dati sintetici non sarà in grado di ottenere gli stessi risultati quando applicato a dati reali. Pertanto, i dati generati dall'LLM non possono essere considerati validi per la fase di addestramento del modello. Successivamente, verranno esplorate ulteriori caratteristiche di questo secondo dataset sintetico per verificare se siano state mantenute altre proprietà tipiche del dataset reale, come ad esempio le distribuzioni delle variabili o gli indici di sintesi.

Feature	Correlazione (Sintetico)	Correlazione (Reale)
Crypto	0.0144971467	0.101112015
DomainTitleMatchScore	0.0136762357	0.570249492
Robots	0.0125443113	0.396208704
NoOfOtherSpecialCharsInURL	0.0107701053	-0.406440494
DomainLength	0.0103587266	-0.262763580
NoOfImage	0.0084280588	0.317604924
HasExternalFormSubmit	0.0070613672	0.170393099
Domain_numeric	0.0068290964	0.451279836
NoOfSelfRedirect	0.0061166207	-0.089700554
TLDLegitimateProb	0.0043800444	0.097092703
NoOfJS	0.0043087096	0.541923412
LineOfCode	0.0040748314	0.351496677
NoOfObfuscatedChar	0.0040716421	-0.040996807
SpacialCharRatioInURL	0.0040089659	-0.517029402
LargestLineLength	0.0038137081	-0.034737240
NoOfEmptyRef	0.0034368878	0.135309322
URLLength	0.0028222761	-0.288858679
NoOfPopup	0.0028036351	0.046738726
ObfuscationRatio	0.0026761605	-0.043464578
NoOfDegitsInURL	0.0026668200	-0.287961146
TLDLength	0.0025312434	-0.075725589
IsResponsive	0.0022937287	0.564786719
HasCopyrightInfo	0.0020139125	0.746076118
NoOfSubDomain	0.0016329581	-0.004390171
NoOfURLRedirect	0.0014527975	-0.038133810
IsDomainIP	-0.0002508300	-0.056391743
URLTitleMatchScore	-0.0004447631	0.529816070
HasPasswordField	-0.0004855510	0.149780257
IsHTTPS	-0.0005637395	0.646885058
URLCharProb	-0.0008225307	0.454740398
NoOfCSS	-0.0015202116	0.424377445
NoOfSelfRef	-0.0015320770	0.405329350
DegitRatioInURL	-0.0018442837	-0.419638475
NoOfiFrame	-0.0020154862	0.251850614
NoOfLettersInURL	-0.0020313342	-0.305340519
NoOfExternalRef	-0.0027221643	0.369758258
Bank	-0.0027857300	0.210757171
Title_numeric	-0.0028587683	0.319122238
URLSimilarityIndex	-0.0036939086	0.856135403
HasFavicon	-0.0042489935	0.506485685
HasTitle	-0.0045086981	0.473294908
NoOfAmpersandInURL	-0.0050196520	-0.048959571
HasHiddenFields	-0.0060184964	0.516743503
HasSubmitButton	-0.0067766366	0.586002993
Pay	-0.0070087433	0.371572507
HasObfuscation	-0.0070258742	-0.050892372
NoOfQMarkInURL	-0.0080600919	-0.170552587
HasDescription	-0.0080678568	0.683989489
NoOfEqualsInURL	-0.0088328087	-0.111961552
HasSocialNet	-0.0107812486	0.778829467
CharContinuationRate	-0.0134606639	0.448866608
LetterRatioInURL	-0.0183172921	-0.329955655

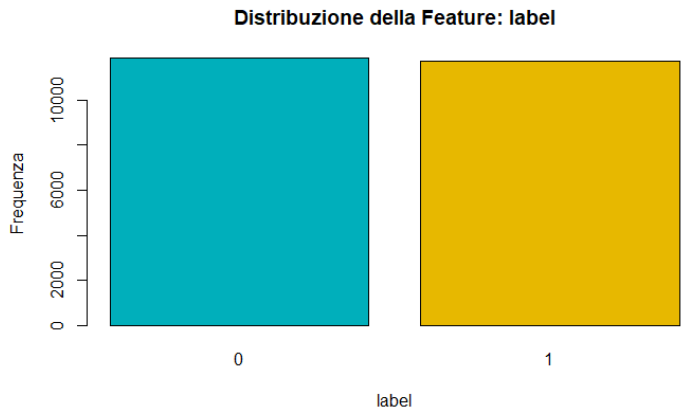
Tabella 31: Confronto tra le correlazioni tra feature e variabile target `label` tra il secondo dataset sintetico e quello reale.

8.2 Distribuzioni di Frequenza

Verranno di seguito calcolate le distribuzioni di frequenza dei dati sintetici e confrontate con le distribuzioni reali. Anche in questo caso, considereremo le feature `URLSimilarityIndex`, `HasSocialNet`, `HasCopyrightInfo`, `IsHTTPS` e la variabile target `label`, in quanto utilizzate in fase di addestramento del modello. Considereremo i seguenti intervalli di `URLSimilarityIndex`:

$$[0, 10), [10, 20), [20, 30), [30, 40), [40, 50), [50, 60), [60, 70), [70, 80), [80, 90), [90, 100)$$

8.2.1 Distribuzione della Variabile Target



Classe	Frequenza Assoluta	Frequenza Relativa	Frequenza Percentuale
syn_label = 0	11846	0.5024	50.24%
syn_label = 1	11734	0.4976	49.76%
real_label = 0	9885	0.4192	41.92%
real_label = 1	13695	0.5807	58.07%

Tabella 32: Distribuzioni di frequenza della variabile target.

Le etichette sintetiche non riflettono la distribuzione delle etichette reali: la distribuzione sintetica appare quasi perfettamente bilanciata, con una differenza di appena lo 0.64% tra le due classi, mentre la distribuzione reale ha una classe dominante, cioè label = 1, con una differenza del 16.15% rispetto alla classe label = 0.

8.2.2 Distribuzione della Variabile Continua

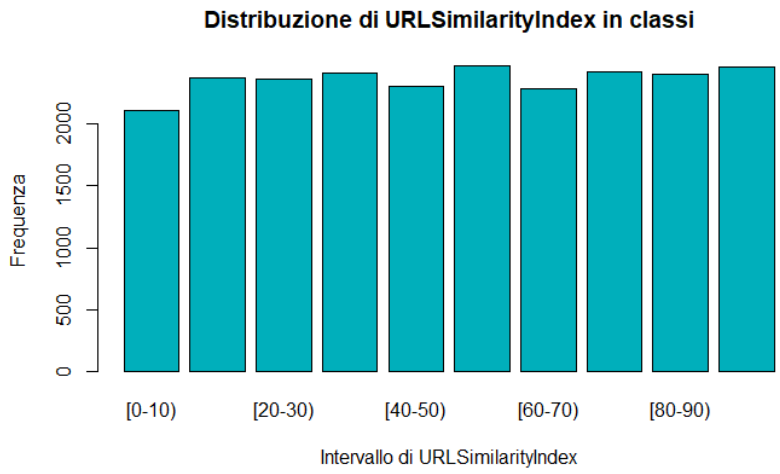


Figura 3: Distribuzioni di frequenza della variabile `URLSimilarityIndex`.

Intervallo	Frequenza Assoluta	Frequenza Relativa	Frequenza Percentuale
[0-10)	2107	0.0894	8.94%
[10-20)	2375	0.1007	10.07%
[20-30)	2356	0.0999	9.99%
[30-40)	2405	0.1020	10.20%
[40-50)	2306	0.0978	9.78%
[50-60)	2466	0.1046	10.46%
[60-70)	2284	0.0969	9.69%
[70-80)	2418	0.1025	10.25%
[80-90)	2400	0.1018	10.18%
[90-100)	2463	0.1045	10.45%

La distribuzione delle frequenze assolute per gli intervalli di valori dell'`URLSimilarityIndex` mostra una certa uniformità, con valori che oscillano tra 2107 e 2466 per ogni intervallo. Questo suggerisce che i dati sono distribuiti in modo piuttosto omogeneo, senza picchi marcati in nessun intervallo specifico. L'intervallo che mostra la frequenza assoluta più alta è il $[50 - 60)$ con 2466, mentre il più basso è il $[0 - 10)$ con 2107.

8.2.3 Distribuzione delle Variabili Binarie

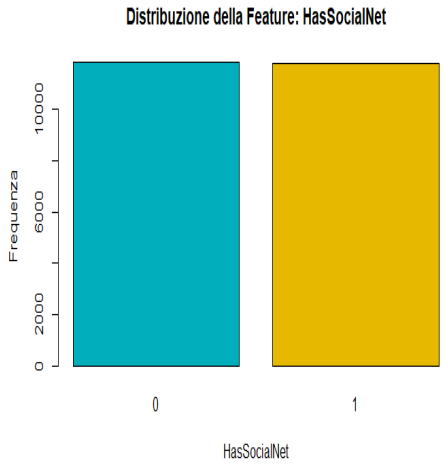


Figura 4: Distribuzione di `HasSocialNet`

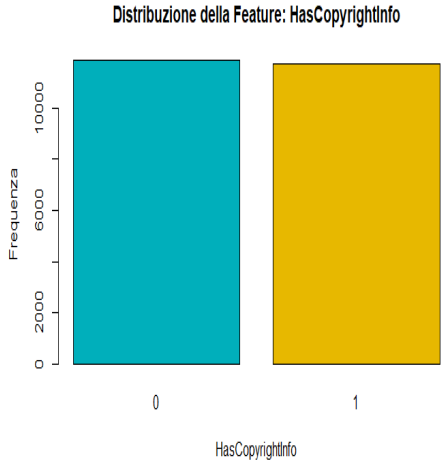


Figura 5: Distribuzione di `HasCopyrightInfo`

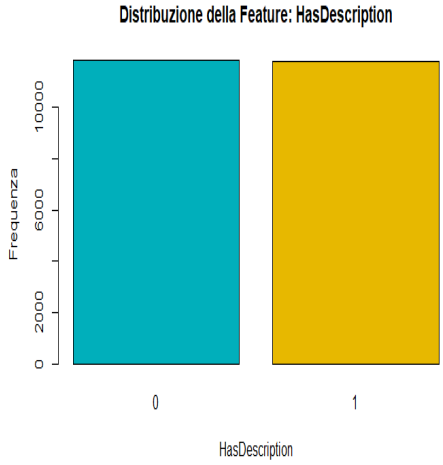


Figura 6: Distribuzione di `HasDescription`

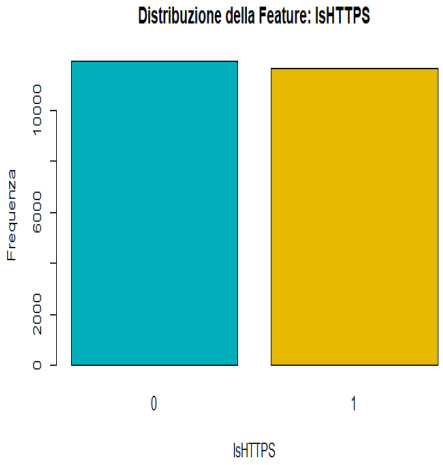


Figura 7: Distribuzione di `IsHTTPS`

Classe	Frequenza Assoluta	Frequenza Relativa	Frequenza Percentuale
HasSocialNet			
syn_HasSocialNet = 0	11813	0.5009	50.09%
syn_HasSocialNet = 1	11767	0.4991	49.91%
real_HasSocialNet = 0	12694	0.5383	53.83%
real_HasSocialNet = 1	10886	0.4617	46.17%
HasCopyrightInfo			
syn_HasCopyrightInfo = 0	11844	0.5023	50.23%
syn_HasCopyrightInfo = 1	11736	0.4977	49.77%
real_HasCopyrightInfo = 0	12053	0.5111	51.12%
real_HasCopyrightInfo = 1	11527	0.4888	48.88%
HasDesription			
syn_HasDesription = 0	11815	0.5010	50.10%
syn_HasDesription = 1	11765	0.4990	49.90%
real_HasDesription = 0	13076	0.5545	55.45%
real_HasDesription = 1	10504	0.4455	44.55%
IsHTTPS			
syn_IsHTTPS = 0	11926	0.5057	50.57%
syn_IsHTTPS = 1	11654	0.4943	50.43%
real_IsHTTPS = 0	5470	0.2319	23.20%
real_IsHTTPS = 1	18110	0.7680	76.80%

Così come per la variabile continua precedentemente analizzata, le distribuzioni delle variabili binarie considerate appaiono quasi perfettamente bilanciate. Anche in questo caso, l'LLM si è concentrato sulla generazione di un dataset bilanciato, senza mantenere le distribuzioni che si osservavano sulle variabili reali.

8.3 Funzione di Distribuzione Empirica

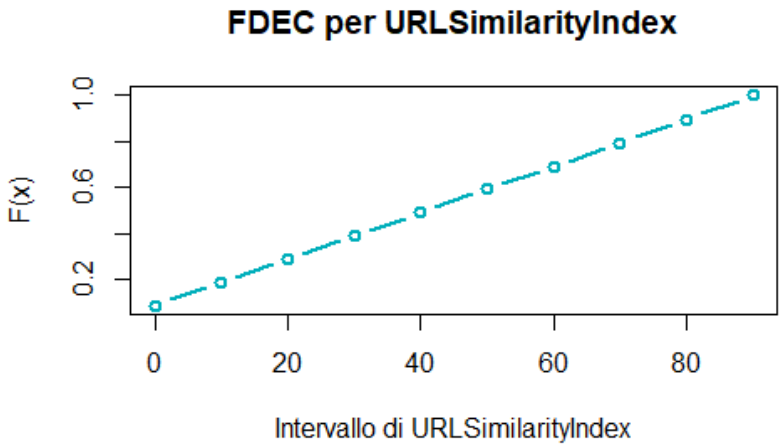
Consideriamo la feature `URLSimilarityIndex`. Sono state individuate $k = 10$ classi:

$$C_1 = [0, 10), \quad C_2 = [10, 20), \quad C_3 = [20, 30), \quad C_4 = [30, 40), \quad C_5 = [40, 50),$$
$$C_6 = [50, 60), \quad C_7 = [60, 70), \quad C_8 = [70, 80), \quad C_9 = [80, 90), \quad C_{10} = [90, 100)$$

La FDEC assumerà i seguenti valori:

Indice	C_i	n_i	f_i	F_i
1	[0-10)	366	$\frac{2107}{23580}$	$\frac{2107}{23580}$
2	[10-20)	754	$\frac{2375}{23580}$	$\frac{4482}{23580}$
3	[20-30)	952	$\frac{2356}{23580}$	$\frac{6838}{23580}$
4	[30-40)	1083	$\frac{2405}{23580}$	$\frac{9243}{23580}$
5	[40-50)	1242	$\frac{2306}{23580}$	$\frac{11549}{23580}$
6	[50-60)	1540	$\frac{2466}{23580}$	$\frac{14015}{23580}$
7	[60-70)	1725	$\frac{2284}{23580}$	$\frac{16299}{23580}$
8	[70-80)	1352	$\frac{2418}{23580}$	$\frac{18717}{23580}$
9	[80-90)	675	$\frac{2400}{23580}$	$\frac{21117}{23580}$
10	[90-100)	13891	$\frac{2463}{23580}$	$\frac{23580}{23580}$

$$F(x) = \begin{cases} 0, & x < 0 \\ \frac{2107}{23580} = 0.0893, & 0 \leq x < 10 \\ \frac{4482}{23580} = 0.1900, & 10 \leq x < 20 \\ \frac{6838}{23580} = 0.2899, & 20 \leq x < 30 \\ \frac{9243}{23580} = 0.3919, & 30 \leq x < 40 \\ \frac{11549}{23580} = 0.4897, & 40 \leq x < 50 \\ \frac{14015}{23580} = 0.5943, & 50 \leq x < 60 \\ \frac{16299}{23580} = 0.6912, & 60 \leq x < 70 \\ \frac{18717}{23580} = 0.7937, & 70 \leq x < 80 \\ \frac{21117}{23580} = 0.8955, & 80 \leq x < 90 \\ 1, & x \geq 100 \end{cases}$$



8.4 Indici di Sintesi

8.4.1 Variabili Binarie

Feature	**Media**	**Varianza**	**Deviazione Standard**	**Moda**
syn_label	0.4976	0.25	0.5	0
real_label	0.5808	0.2435	0.4934	1
syn_HasSocialNet	0.499	0.25	0.5	0
real_HasSocialNet	0.4617	0.2485	0.4985	0
syn_HasCopyrightInfo	0.4977	0.25	0.5	0
real_HasCopyrightInfo	0.4888	0.2499	0.4999	0
syn_HasDescription	0.4989	0.25	0.5	0
real_HasDescription	0.4455	0.247	0.497	0
syn_IsHTTPS	0.4942	0.25	0.5	0
real_IsHTTPS	0.768	0.1782	0.4221	1

L'importanza posta dall'LLM nel bilanciamento delle classi emerge dall'analisi degli indici di sintesi anche nel caso di questo nuovo dataset. La media per ogni feature, pari a 0.5, unita alla varianza di 0.25 (massima per le variabili binarie) sottolinea che i valori di ogni classe sono equamente rappresentati. Nel caso delle variabili reali osserviamo dei leggeri sbilanciamenti nella maggior parte delle classi, con un forte sbilanciamento in `IsHTTPS`.

8.4.2 Variabile Continua

Media, Moda e Mediana Campionaria. Per quanto riguarda `URLSimilarityIndex` (Tabella 33), l'osservazione degli indici di sintesi mostra una netta differenza trai valori reali e sintetici. La media del `URLSimilarityIndex` nei dati sintetici è 50.787, mentre nei dati reali è 79.5097. Questo indica che i valori medi nel dataset reale sono significativamente più alti rispetto a quelli sintetici, suggerendo che nei dati reali i URL tendano ad avere una maggiore similarità rispetto ai dati sintetici. La mediana è 50.8934, mentre per i dati reali è 100: ciò indica una distribuzione più uniforme, con valori che si distribuiscono meno concentrati attorno ai valori estremi. La moda del URLSimilarityIndex per i dati sintetici è 1.293383, mentre nei dati reali è 100. La moda è molto più bassa, indicando che non c'è una predominanza di valori di alta similarità.

Feature	**Media**	**Mediana**	**Moda**
syn_URLSimilarityIndex	50.787	50.8934	1.293383
real_URLSimilarityIndex	79.5097	100	100

Tabella 33: Indici di Sintesi della variabile continua sintetica.

Quartili. La distribuzione dei valori sintetici si mostra più equilibrata dei valori reali di `URLSimilarityIndex` anche per quanto riguarda i quartili. Lo scambio interquartile risulta avere valore pari a:

$$IQR = Q_3 - Q_1 = 75.80221 - 26.08193 = 49.72028$$

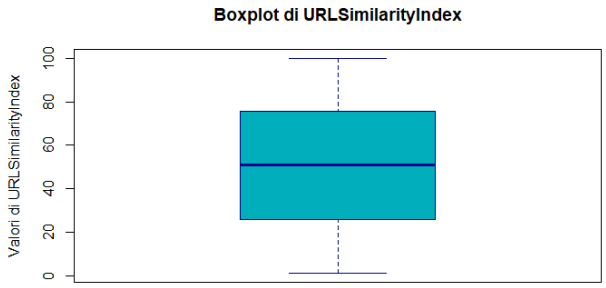
Gli outlier sono definiti come valori esterni all'intervallo $[Q_1 - 1.5 \cdot IQR, Q3 + 1.5 \cdot IQR]$, quindi gli outlier della variabile considerata dovranno essere ricercati all'esterno dell'intervallo:

$$[-48.49849, 150.3826]$$

Anche in questo caso non sono presenti outlier.

Feature	Q_0	Q_1	Q_2	Q_3	Q_4
syn_URLSimilarityIndex	1.293383	26.08193	50.8934	75.80221	99.99159
real_URLSimilarityIndex	1.292906	59.65332	100	100	100

Tabella 34: Quartili della variabile continua sintetica.



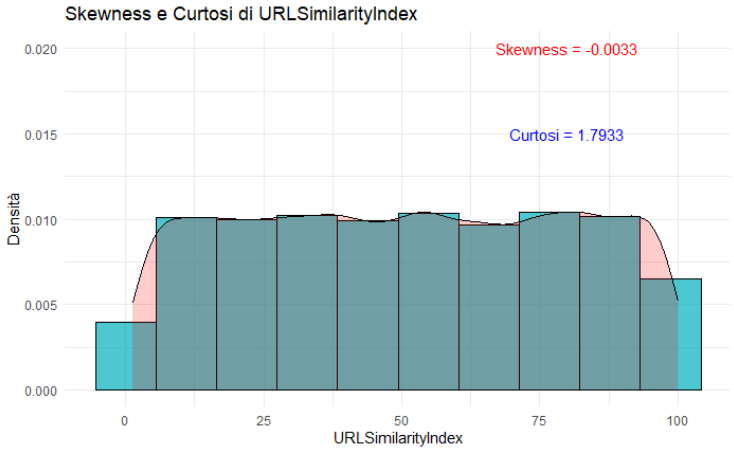
Varianza e Deviazione Standard. I valori di varianza e deviazione standard appaiono molto elevati, pari a rispettivamente 817.8945 e 28.5989, indicano la presenza di valori lontani dalla media, a conferma del fatto che il dataset generato risulta quasi perfettamente bilanciato.

Feature	Varianza	Deviazione Standard
syn_URLSimilarityIndex	817.8945	28.5989
real_URLSimilarityIndex	793.6245	28.1713

Tabella 35: Misure di Dispersione della variabile continua sintetica.

Skewness e Curtosi Campionaria. Poiché la skewness campionaria assume un valore pari a -0.0033, cioè molto vicino a 0, allora possiamo assumere che la distribuzione dei dati sia simmetrica rispetto alla media.
Un valore della curtosi pari a 1.7933 indica una distribuzione platicurtica, cioè con code più leggere rispetto ad una distribuzione normale.

Feature	Skewness	Curtosi
syn_URLSimilarityIndex	-0.0033	1.7933
real_URLSimilarityIndex	-1.0601	-0.2139



8.5 Covarianza e Correlazione Campionaria

8.5.1 Covarianza delle feature sintetiche con label

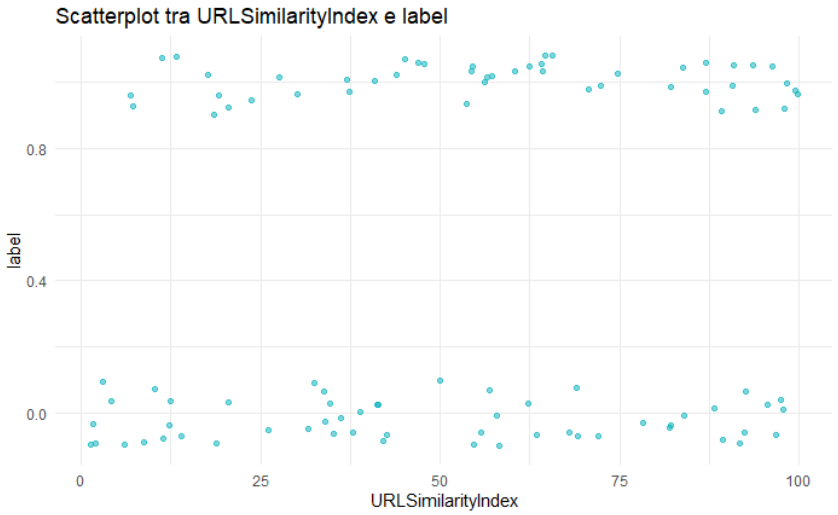
Utilizziamo la covarianza campionaria per valutare la relazione tra la variabile target `label` e le feature selezionate. Una covarianza positiva indica che un aumento nei valori di una variabile tende a essere associato a un aumento nei valori dell'altra, mentre una covarianza negativa suggerisce una relazione inversa.

Feature	**Covarianza con label**
URLSimilarityIndex	-0.0528
HasSocialNet	-0.0027
HasCopyrightInfo	-0.00004
HasDescription	-0.002
IsHTTPS	-0.0004

URLSimilarityIndex e label. La covarianza tra `URLSimilarityIndex` e `label` è pari a -0.0528, che è molto debole e negativa rispetto alla covarianza tra le due variabili registrata su dati reali (pari a 11.901).
Sia dalla tabella di contingenza (Tabella 36) che dallo scatterplot si osserva una distribuzione che appare abbastanza uniforme tra le due classi. Non emergono picchi significativi, il che significa che `URLSimilarityIndex` non discrimina in modo netto tra le due classi.

Label	**[0,10)**	**[10,20)**	**[20,30)**	**[30,40)**	**[40,50)**	**[50,60)**	**[60,70)**	**[70,80)**	**[80,90)**	**[90,100)**
0	1068	1176	1190	1189	1166	1229	1149	1211	1228	1240
1	1039	1199	1166	1216	1140	1237	1135	1207	1172	1223

Tabella 36: Tabella di contingenza tra *label* e *URLSimilarityIndex*

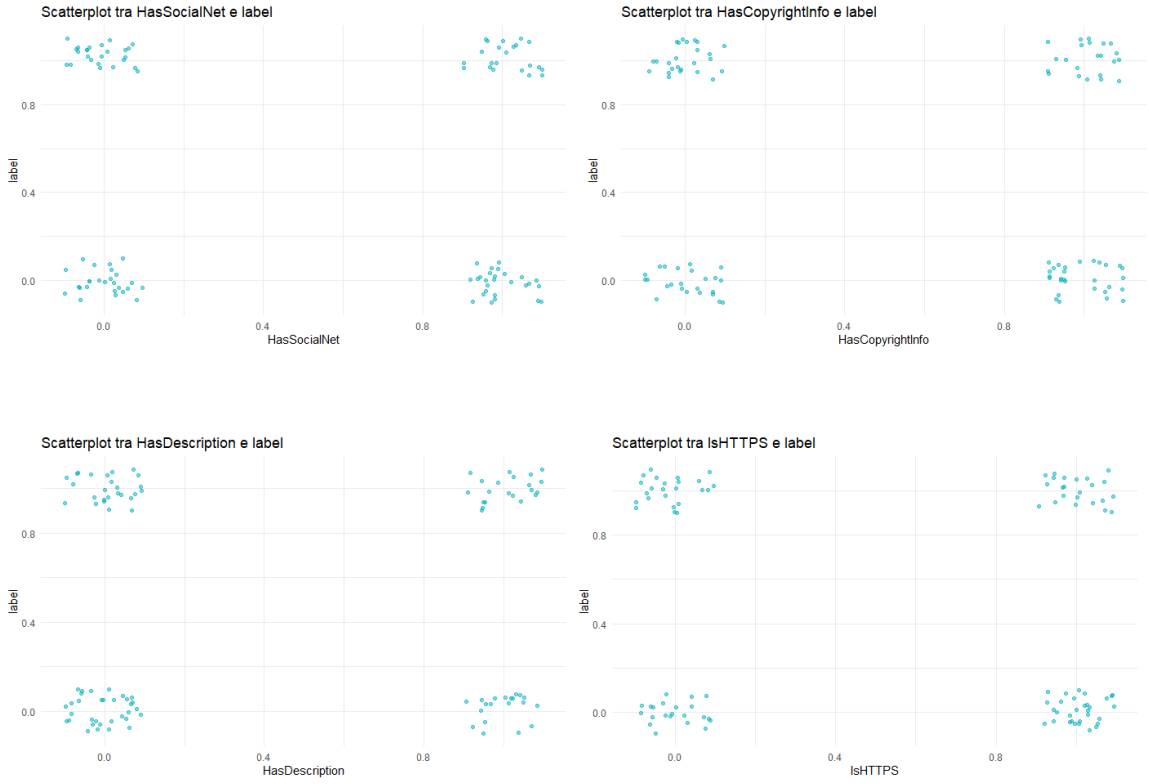


Variabili Binarie e label. La covarianza tra la variabile target `label` e tutte le altre variabili binarie appare molto bassa e negativa, con valori che vanno da un minimo di -0.0027 a un massimo di -0.00004.

Osservando sia la tabella 37 che gli scatterplot, si osserva che non vi è alcuna corrispondenza tra `label` e le altre variabili binarie.

Label	**0**	**1**
	HasSocialNet	
0	5871	5975
1	5942	5792
	HasCopyrightInfo	
0	5962	5884
1	5882	5852
	HasDescription	
0	5888	5958
1	5927	5807
	IsHTTPS	
0	5988	5858
1	5938	5796

Tabella 37: Tabella di contingenza tra *label* e le variabili binarie



8.5.2 Correlazione Campionaria

Osservando la tabella 38, possiamo notare che i coefficienti di correlazione sono tutti prossimi allo zero, indicando assenza di correlazione. Inoltre, dal grafico 8 è chiaro che le variabili

non abbiano nessuna interrelazione. Anche in questo caso, le relazioni presenti trai dati reali non sono state rispettate durante la generazione del dataset.

Feature	**Correlazione con `label`**
URLSimilarityIndex	-0.0037
HasSocialNet	-0.0108
HasCopyrightInfo	0.002
HasDescription	-0.0081
IsHTTPS	-0.0006

Tabella 38: Coefficienti di correlazione tra la variabile `label` e le feature considerate.

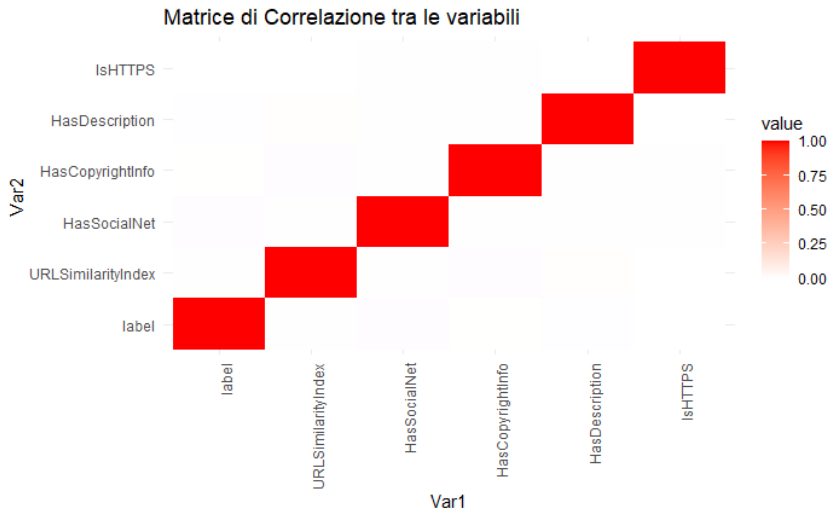


Figura 8: Correlazione tra tutte le feature.

8.6 Verifica delle Ipotesi

8.6.1 Risultati sul Secondo Dataset Sintetico

Variabile Continua. Di seguito riportiamo i risultati ottenuti dal confronto della distribuzione della variabile continua `URLSimilarityIndex` con una distribuzione uniforme e una normale (Tabella 39).

Per quanto riguarda il confronto con la distribuzione uniforme, il test del chi-quadrato restituisce un p-value pari a 0.1073, superiore rispetto alla soglia di 0.05, il che indica che non abbiamo evidenze sufficienti per rifiutare l'ipotesi nulla. Possiamo dunque concludere che i dati sono compatibili con una distribuzione uniforme.

Analizzando invece il confronto con la distribuzione normale, osserviamo un p-value estremamente basso ($p < 2.2 \cdot 10^{-16}$), che suggerisce una forte deviazione rispetto a una distribuzione normale. Di conseguenza, rifiutiamo l'ipotesi nulla e concludiamo che i dati non seguono una distribuzione normale.

I risultati ottenuti per questo secondo dataset sintetico ricalcano quindi quelli ottenuti nel primo dataset per la variabile continua.

Test	X^2	p-value	Gradi di Libertà
Uniforme	14.446	0.1073	9
Normale	4953.6	$2.2 \cdot 10^{-16}$	9

Tabella 39: Risultati del test del chi-quadrato per la variabile `URLSimilarityIndex`.

Variabili Binarie. Di seguito sarà verificato se le variabili binarie seguono una distribuzione binomiale, assumendo che le osservazioni siano indipendenti e seguano una sequenza di prove di Bernoulli.

Come mostrano i risultati riportati nella Tabella 40, tutte le variabili binarie presentano un p-value pari a 1. Questo indica che non ci sono evidenze sufficienti per rifiutare l'ipotesi nulla, ovvero che i dati siano compatibili con una distribuzione binomiale.

Anche in questo caso i risultati ottenuti sono congrui a quanto già osservato nel primo dataset sintetico generato.

Feature	X^2	p-value	Gradi di Libertà
label	0	1	1
HasSocialNet	0	1	1
HasCopyrightInfo	0	1	1
HasDescription	0	1	1
IsHTTPS	$2.7744 \cdot 10^{-28}$	1	1

Tabella 40: Risultati del test del chi-quadrato per le variabili binarie.

8.7 Risultati del Modello

Proprio come nel caso precedente, il modello risulta avere dei risultati discreti, a differenza degli ottimi risultati riportati su dati reali.

Prediction	**Istanza Negativa**	**Istanza Positiva**
Istanza Predetta come Negativa	(TN) 11139	(FN) 11021
Istanza Predetta come Positiva	(FP) 707	(TP) 713

Tabella 41: Matrice di confusione del modello sui dati sintetici

Per valutare la bontà del modello di classificazione, verranno di seguito riportati i risultati ottenuti per le metriche considerate nei precedenti capitoli:

- **Accuratezza**: il modello mostra avere un'accuratezza di **0.5026**, andando a classificare correttamente appena 11852 tuple delle 23480 che compongono il dataset.
- **Precisione**: la precisione del modello risulta pari a **0.5021**, un valore che risulta dimezzato rispetto ai risultati ottenuti su dati reali.
- **Richiamo**: il richiamo ottenuto è di **0.06**, un risultato simile a quello ottenuto sui precedenti dati.
- **Specificità**: il modello ha una specificità di **0.94**, mostrandosi molto più efficace nel predire istanze negative rispetto alle positive.

- **F1-Score**: l'F1-Score registrato del modello è di **0.0535**, valore molto basso che indica che il modello ha una scarsa performance complessiva.

8.8 Conclusioni

Anche per quanto riguarda questo secondo dataset sintetico, i dati generati da GPT-4 non mantengono le stesse caratteristiche dei dati reali. Le problematiche riscontrate sono simili a quelle osservate nel caso del primo dataset, ovvero che l'LLM tende a generare un dataset bilanciato, senza riuscire a cogliere informazioni rilevanti sulle relazioni tra le variabili o sulle distribuzioni dei dati, nemmeno quando gli vengono fornite molte informazioni sulla natura del dataset. Le scarse performance del modello sono quindi una conseguenza prevedibile, considerando anche i risultati ottenuti con il precedente dataset sintetico.