ANALISI STATISTICA SUL DATASET PHISHING 10

Statistica e Analisi dei Dati

Authors
Ferrara Francesco (0522501959)
Hida Eljon (0522501890)
UNISA
2024-2025

Contents

1 Introduzione

1.1 Contesto del problema

Il phishing è uno dei vari tipi di attacchi informatici che esistono. Esso mira a ingannare gli utenti e indurli a fornire credenziali o dati sensibili.
Questa tipologia di attacchi, molto spesso, sfrutta: email, messaggi o URL malevoli che sembrano provenire da fonti affidabili e sicure. Tra le tante conseguenze del phishing abbiamo quelle finanziarie dove:

- Perdite economiche o furto di identità, nel caso di utenti;
- Perdita di fiducia da parte dei clienti, nel caso di aziende.

1.2 Obiettivi dell'analisi

Uno degli obiettivi dell'analisi sarà quello di investigare in qual modo la lunghezza degli URL e la presenza di link a riferimenti esterni influenzano la probabilità che un determinato URL sia classificato come URL di phishing.
Attraverso un'analisi attenta, puntiamo a identificare eventuali pattern utili per migliorare i sistemi di rilevamento per questi attacchi.
L'altro obiettivo sarà quello di far generare ad un LLM dati sintetici che verranno poi confrontati con i dati reali per trarne conclusioni in base alle RQ preidentificate.

1.3 Definizione della Research Question

Le Research Questions che andremo a studiare e, infine, a rispondere sono le seguenti:

🔍 **RQ$_1$.** *Qual è l'impatto della lunghezza degli URL sulla classificazione di questi ultimi come URL di phishing?*

🔍 **RQ$_2$.** *Qual è l'impatto dato dal numero di link esterni in un URL sulla classificazione di questi ultimi come URL di phishing?*

🔍 **RQ$_3$.** *Quali sono le differenze principali tra dati sintetici generati e dati reali in un contesto di regressione e descrittivo?*

1.4 Tecnologie utilizzate

Verranno utilizzati il linguaggio **R** ed **R Markdown** per l'analisi del dataset e la visualizzazione dei risultati. Tramite **Overleaf** verrà redatta la documentazione.

2 Analisi Esplorativa del Dataset

2.1 Introduzione al Dataset

Lo studio si occupa di analizzare un dataset che cataloga vari URL definendo se questi siano di phishing o legittimi. Il dataset originale che contiene dati estratti per lo più nel 2023 e disponibile al link Dataset Completo. Una parte di quel dataset è stato utilizzato in questo documento per analizzare il problema del Phishing.
All'interno del file Google Fogli Data Dictionary è presente l'elenco delle feature del dataset sul phishing, con la loro descrizione e il tipo di valore rappresentato. Il dataset presenta 23575 righe e 56 colonne, di cui 54 feature. In R, usando la funzione describe della libreria skimr, è possibile vedere che il dataset non presenta tuple mancanti. Per quanto riguarda l'analisi di eventuali valori duplicati, è possibile vedere tramite la funzione duplicated che non ce ne sono. Il dataset risulta quindi essere completo.

2.2 Analisi delle Feature

Oltre alla specifica delle feature, descritta nel Data Dictionary precedentemente menzionato, ci sono dei forti legami tra le stesse che vanno analizzati. Le feature URL e Domain rappresentano gli stessi concetti ma rappresentati in modo differente, ossia il dominio non è altro che il sito vero e proprio, eliminando la parte relativa all'http / https. Questa correlazione si riflette anche sulle due feature che rappresentano la loro lunghezza, ossia URLLength e DomainLength: la seconda ha un valore sempre minore rispetto alla prima, i cui caratteri in aggiunta sono molto spesso https:// oppure http:// (senza la s di secure). Poi le tre feature HasObfuscation, NoOfExternalRef e ObfuscationRatio hanno sempre valore 0 tranne quando è presente l'offuscamento: in quel caso la prima variabile assume valore 1, la seconda specifica il numero di caratteri offuscati, mentre la terza indica la percentuale. Va poi analizzato il rapporto che c'è tra DomainTitleMatchScore e URLTitleMatchScore: queste due colonne presentano sempre gli stessi valori, tranne 739 volte, che è un valore piccolo rispetto alla totalità delle righe presenti. Questo indica che più del 96% delle volte, i due valori coincidono, ossia di fatto, rappresentano la stessa variabile.

Per approfondire lo studio, sono stati selezionati due indicatori principali:

- **URLLenght**: indica la lunghezza complessiva dell'URL. Questa variabile è importante perché gli URL di phishing tendono spesso ad essere più lunghi, includendo sottodomini o percorsi complessi per mascherare la loro vera identità.
- **NoOfExternalRef**: la variabile NoOfExternalRef rappresenta il numero di riferimenti esterni presenti nella pagina web. Questa variabile conta quante risorse esterne (come immagini, script, CSS o link ad altri domini) sono referenziate all'interno della pagina web. Un alto numero di riferimenti esterni potrebbe essere un segnale di un sito sospetto.

Inoltre, durante tutto il processo di analisi e studio terremo conto anche della variabile **Label**, la quale classifica ciascun URL come phishing (0) o legittimo (1).

3 Analisi Descrittiva Univariata

La statistica descrittiva univariata costituisce un pilastro essenziale dell'analisi statistica, concentrando l'attenzione sull'analisi e l'interpretazione delle caratteristiche di una singola variabile quantitativa all'interno di un insieme di dati. La sua applicazione permette di indagare in maniera efficace le proprietà e la distribuzione della variabile, offrendo un fondamento empirico solido per le successive analisi e deduzioni.

3.1 Analisi di URLLength

3.1.1 Indici di centralità

Di seguito viene mostrata la tabella relativa agli indici di centralità, ossia media, mediana e moda, a cui sono stati aggiunti gli indici che indicano i valori minimo e massimo della distribuzione considerata. L'analisi tiene però conto anche del rapporto che intercorre tra URLLength e la label.
In particolare abbiamo che:

- **"Completo"** rappresenta l'analisi esclusivamente riferita sulla colonna URLLength, ossia senza distinzioni all'interno del dataset;
- **"Phishing(label=0)"** rappresenta il calcolo degli indici sulla variabile URLLength quando la label assume valore 0 (sito di phishing);
- **"NonPhishing(label=1)"** rappresenta il calcolo degli indici sulla variabile URLLength quando la label assume valore 1 (sito non di phishing).

	MIN	**MAX**	**MEDIA**	**MEDIANA**	**MODA**
Completo	14	1427	34.84	28.00	26
Phishing(label=0)	14	1427	46.31	34.00	26
NonPhishing(label=1)	16	51	26.16	26.00	25

Table 1: Indici di centralità di URLLength rispetto alla label

Andando ad analizzare la tabella scopriamo che:

- **Considerazioni di MIN e MAX:** I valori risultanti indicano che il range tra minimo e massimo, quando il sito è di Phishing, è molto più ampio rispetto a quando non lo è. Questo dato è confermato sia dall'andamento del valore assunto dalla media in corrispondenza delle due variabili, sia dal valore assunto dalla mediana. Questo indica che abbiamo una probabilità maggiore di trovarci di fronte a un sito di phishing se la lunghezza dell'URL è molto lunga.
- **Considerazioni media:** Gli URL di phishing hanno una lunghezza media e una variabilità maggiori rispetto agli URL legittimi. Questo potrebbe essere dovuto a tecniche mirate come l'aggiunta di sottodomini, path lunghi o stringhe casuali per confondere l'utente.
- **Considerazioni mediana:** I valori assunti dalla mediana sono direttamente proporzionali ai valori assunti dalla media, ed in particolare abbiamo che nel caso del phishing la lunghezza mediana della stringa è di 34 caratteri.

- **Considerazioni moda:** La moda, invece, presenta una variabilità minore. Abbiamo infatti un valore di 26 per il caso "Completo" e per il caso "Phishing", e 25 nel caso "NonPhishing". La sola analisi di questo indice quindi non ci consente di analizzare eventuali differenze tra i vari scenari, come invece accade nel caso della media e della mediana. Questo però potrebbe indicare che per entrambi i gruppi esiste una sovrapposizione tra phishing e non phishing per alcune lunghezze, il che potrebbe significare che ci sono anche URL di phishing con lunghezze comuni a quelli legittimi.

3.1.2 Indici di Dispersione

Nella tabella seguente sono indicati gli indici di dispersione, che servono a comprendere la variabilità dei dati a disposizione. Sono stati considerati il coefficiente di variazione(CV), lo scarto interquartile e la deviazione standard di URLLength, rispetto al caso completo, al caso Phishing e a quello NonPhishing.

	CV	**Scarto Interquartile**	**Deviazione Standard**
Completo	0.97	10	33.83
Phishing(label=0)	1.06	23	48.96
NonPhishing(label=1)	0.18	6	4.75

Table 2: Indici di Dispersione di URLLength

I valori della tabella indicano rispettivamente:

- **Coefficienti di Variazione(CV):** il coefficiente di variazione fornisce una misura relativa della dispersione dei dati rispetto alla media. Si può notare che è alquanto elevato nel caso degli URL dei siti di phishing (1.06), questo ci fa capire che per questa categoria la varianza della lunghezza degli URL è molto vasta. Per la categoria dei siti legittimi i valori sono contenuti facendo intuire una certa omogeneità nella lunghezza degli URL.

- **Scarto Interquartile:** è la differenza tra il terzo quartile (Q3) e il primo(Q1). Nel caso "Phishing" la differenza tra questi due valori è molto marcata rispetto agli altri due casi, e questo indica che la variabilità della lunghezza è molto elevata. Significativa è la differenza con il caso "NonPhishing", in cui abbiamo un valore di 6, che rappresenta come nel caso di URL legittimi la variabilità sia molto più ridotta. Nel caso "Completo" le differenze si appianano, il che non ci consente di comprendere in maniera efficace le caratteristiche proprie del problema.

- **Deviazione Standard:** indica la dispersione dei dati rispetto alla media nella stessa unità di misura dei valori originali, e ci consente di comprendere bene quanta variabilità c'è all'interno della distribuzione. La differenza tra il caso "Phishing" e il caso "NonPhishing" è molto accentuata, ossia ci troviamo di fronte ad una variabilità molto elevata nel primo caso, mentre molto ridotta nel secondo, cioè quasi piatta. Questo ci permette di capire come nel caso di URL legittimi, a differenza di quelli non legittimi, la loro lunghezza non sia molto grande e di conseguenza anche la loro variabilità si riduce significativamente.

3.1.3 Quartili

Il calcolo dei quartili consente di suddividere la distribuzione dei dati in quattro parti, offrendo una visione dettagliata della dispersione e della centralità dei valori all'interno di ciascun gruppo. Attraverso l'identificazione di questi quartili è possibile determinare la mediana(rappresentata dal secondo quartile) e lo Scarto Interquartile(Q3-Q1), ottenendo così informazioni sull'eterogeneità dei dati e sulla loro tendenza. Grazie a questa metodologia riusciamo ad avere una buona esplorazione delle caratteristiche e potenziali anomalie delle variabili prese in considerazione.

La tabella seguente calcola gli indici basandosi sull'intervallo interquartile: escludendo la prima riga, che viene ripresa dalla tabella precedente, le altre calcolano gli indici basandosi solo sui valori corrispondenti ad ogni range.

	MIN	**MAX**	**NoValues**
Completo	14	1427	23575
0-Q1	14	24	7265
Q1-Q2	25	28	5720
Q2-Q3	29	34	4864
Q3-Q4	35	1427	5726

Table 3: Numero di valori presenti all'interno di URLLength, divisa per quartili

La distribuzione del numero di valori (NoValues) è abbastanza uniforme, nel senso che non ci sono intervalli con un numero di elementi molto più grande degli altri, tranne per il primo, in cui la differenza con gli altri è più accentuata. In generale si nota come la maggioranza degli URL presenti nel dataset abbia una lunghezza che non supera i 35 caratteri.

Di seguito viene visualizzato il boxplot relativo alla colonna URLLength, specificando i 3 casi. La visualizzazione è priva degli outlier.

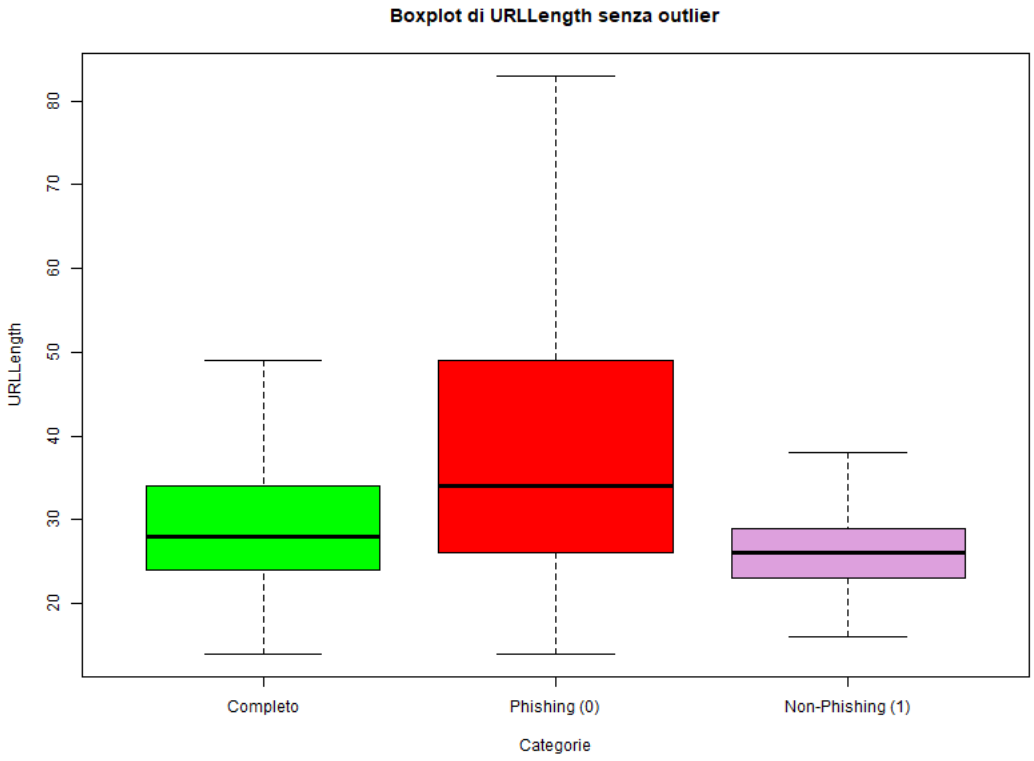

Figure 1: Boxplot di confronto di URLLength

Analizzando il boxplot, possiamo notare che:

- Nel caso **Completo**, la distribuzione risulta quasi simmetrica. Non lo è del tutto in quanto la mediana è leggermente più vicina a Q1 che a Q3 ed inoltre il baffo superiore ha un'ampiezza maggiore rispetto a quello inferiore.
- Nel caso del **Phishing**, osservando le distanze tra mediana e Q1 e mediana e Q3 si nota una vicinanza maggiore a Q1 indicando che la distribuzione è asimmetrica positiva. Inoltre, il boxplot è molto più alto degli altri casi facendoci intuire che non c'è una forte omogeneità nella lunghezza degli URL di phishing. Infine la lunghezza del baffo inferiore è nettamente minore rispetto alla lunghezza del baffo superiore, per cui si nota una forte asimmetria a destra, o asimmetria superiore, nella distribuzione.
- Nel caso del **NonPhishing** notiamo che è tutto molto più omogeneo rispetto al precedente. Infatti: osserviamo che il 50% dei dati intorno alla mediana risulta essere ben distribuito. Anche il baffo superiore e il baffo inferiore risultano simmetrici. Quindi, in questo caso, abbiano una variabilità molto più bassa rispetto al caso "Completo" e al caso "Phishing".

3.2 Analisi di NoOfExternalRef

3.2.1 Indici di Centralità

Come già visto in precedenza per URLLength per la tabella 1, di seguito viene mostrata la tabella relativa agli indici di centralità, con l'aggiunta dei valori di minimo e massimo. La tabella tiene conto anche del rapporto che intercorre tra NoOfExternalRef e la label.

	MIN	**MAX**	**MEDIA**	**MEDIANA**	**MODA**
Completo	0	27516	50.69	10	0
Phishing(label=0)	0	96	1.21	0	0
NonPhishing(label=1)	0	27516	88.11	45	13

Table 4: Indici di centralità di NoOfExternalRef rispetto alla label

Andando ad analizzare la tabella scopriamo che:

- **Considerazioni di MIN e MAX:** I valori risultanti indicano che il range tra minimo e massimo, quando il sito è legittimo, è molto più ampio rispetto a quando non lo è. Questo dato è confermato sia dall'andamento del valore assunto dalla media in corrispondenza delle due variabili, sia dal valore assunto dalla mediana. Questo indica che abbiamo una probabilità maggiore di trovarci di fronte a un sito legittimo se c'è una forte presenza di riferimenti esterni. Questo potrebbe essere dovuto al fatto che quando ci si trova su un sito di phishing, colui che l'ha creato tende a far rimanere l'utente sulla stessa pagina in modo da poter acquisire i suoi dati in maniera più semplice. Se l'utente si spostasse tra più riferimenti sarebbe più difficile riuscire in quest'intento.

- **Considerazioni media:** Il numero di riferimenti esterni è più ampio sia nel caso "Completo" che nel caso "NonPhishing". Nel caso di "Phishing", invece, questo valore è molto inferiore.

- **Considerazioni mediana:** I valori assunti dalla mediana sono direttamente proporzionali ai valori assunti dalla media, ed in particolare abbiamo che nel caso del phishing il suo valore è pari a 0.

- **Considerazioni moda:** il valore più frequente nel caso "Completo" e nel caso "Phishing" è 0, mentre nel caso di "NonPhishing" è 13. Questi valori sono in linea con le altre metriche e questo dimostra il fatto che quando ci si trova su siti di phishing, l'obiettivo dell'attaccante è quello di far muovere il meno possibile l'utente, in modo da prelevare i suoi dati in maniera più agevole.

3.2.2 Indici di Dispersione

Dall'analisi degli indici di dispersione di NoOfExternalRef, si ottiene la tabella seguente.

	CV	**Scarto Interquartile**	**Deviazione Standard**
Completo	5.66	55	287.25
Phishing(label=0)	2.58	1	3.14
NonPhishing(label=1)	4.27	89	376.38

Table 5: Indici di Dispersione di NoOfExternalRef

I valori della tabella indicano rispettivamente:

- **Coefficienti di Variazione(CV):** i valori ottenuti sono piuttosto elevati, e questo è dovuto ad un'alta variabilità dei dati e una media più contenuta;
- **Scarto Interquartile:** è la differenza tra il terzo quartile (Q3) e il primo(Q1). Nel caso "Completo" e nel caso "NonPhishing" lo scarto è molto grande, mentre nel caso "Phishing" tende a 0;
- **Deviazione Standard:** indica la dispersione dei dati rispetto alla media nella stessa unità di misura dei valori originali, e ci consente di comprendere bene quanta variabilità c'è all'interno della distribuzione. La variabilità dei dati è molto accentuata nel caso "Completo" e nel caso "NonPhishing", mentre quando ci troviamo nel caso "Phishing" questa è molto bassa, il che indica che i valori sono molto vicini alla media. Questi valori sono in linea con lo scarto interquartile.

3.2.3 Quartili

La tabella seguente calcola gli indici basandosi sull'intervallo interquartile: escludendo la prima riga, che viene ripresa dalla tabella precedente, le altre calcolano gli indici basandosi solo sui valori corrispondenti ad ogni range.

	MIN	MAX	NoValues
Completo	0	27516	23575
0-Q1	0	1	7824
Q1-Q2	2	10	4126
Q2-Q3	11	56	5761
Q3-Q4	57	27516	5864

Table 6: Indici di centralità di NoOfExternalRef, utilizzando i quartili

La distribuzione del numero di valori (NoValues) è abbastanza uniforme, nel senso che non ci sono intervalli con un numero di elementi molto più grande degli altri, tranne per il primo, in cui la differenza con gli altri è più accentuata. In generale si nota come il 75% dei riferimenti esterni è inferiore a 57.

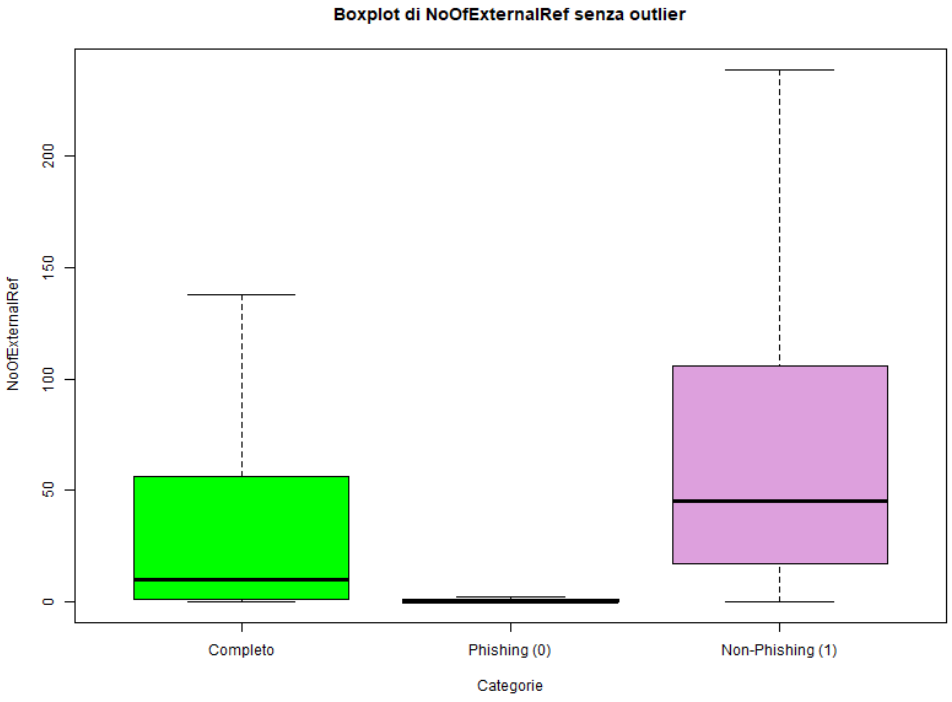

Figure 2: Boxplot di confronto di NoOfExternalRef

Analizzando il boxplot, possiamo notare che:

- Nel caso **Completo**, osservando le distanze tra mediana e Q1 e mediana e Q3 si nota una vicinanza maggiore a Q1 indicando che la distribuzione è asimmetrica positiva.La lunghezza del baffo inferiore è nettamente minore rispetto alla lunghezza del baffo superiore, per cui si nota una forte asimmetria a destra, o asimmetria superiore, nella distribuzione.

- Nel caso del **Phishing**, la distanza tra i vari quantili è praticamente nulla, e questo porta ad un boxplot piatto. Ciò indica che NoOfExternalRef, quando ha valori etichettati come Phishing, ha una variabilità assente.
- Nel caso del **NonPhishing** la distanza tra Q1 e mediana è minore rispetto alla distanza tra mediana e Q3, ma meno accentuata rispetto al caso Completo. Tuttavia la distribuzione è comunque asimmetrica positiva, anche se leggermente più bilanciata. Il baffo superiore presenta un range molto più ampio rispetto a quello inferiore, e in generale si ha la variabilità maggiore rispetto agli altri due casi.

3.3 Funzione di Distribuzione Empirica Continua

3.3.1 URLLength

Nell'ambito dell'analisi statistica dei fenomeni quantitativi continui, sono stati categorizzati i dati di URLLength in quattro gruppi attraverso una metodologia basata sui quantili. Tale approccio garantisce che ogni gruppo rappresenti circa il 25% del totale delle osservazioni. Queste categorie sono state delineate per abbracciare l'intera gamma di dati, estendendosi dal valore minimo al valore massimo registrato. La procedura successiva ha comportato il calcolo della Funzione di Distribuzione Empirica Continua (FDC) per ciascun intervallo considerato e attraverso l'utilizzo delle Frequenze Cumulate, come indicato nella tabella seguente.

Intervalli	**Frequenze Assolute**	**Frequenze Relative**	**Frequenze Cumulate**
[14,24)	7265	7265/23575	7265/23575
[24,28)	5720	5720/23575	12985/23575
[28,34)	4864	4864/23575	17849/23575
[34,1427]	5726	5726/23575	23575/23575

Table 7: Frequenze di URLLength in base all'intervallo

La FDC è stata calcolata per evidenziare come le osservazioni si accumulino progressivamente, ed è definita nel seguente modo:

$$F(x) = \begin{cases} 0 & \text{se } x < min \\ 0.25 & \text{se } min \leq x < Q1 \\ 0.49 & \text{se } Q1 \leq x < Q2 \\ 0.73 & \text{se } Q2 \leq x < Q3 \\ 0.99 & \text{se } Q3 \leq x < max \\ 1 & \text{se } x \geq max \end{cases}$$

I valori si muovono all'interno degli intervalli specificati, ed in particolare avremo che:

- i valori min e max sono rispettivamente 14 e 1427 ed indicano gli estremi del dominio di riferimento, ossia la variabile intera URLLength;
- i quartili sono le grandezze intermedie degli intervalli ed in particolare Q1 = 24, Q2 = 28, Q3 = 34.

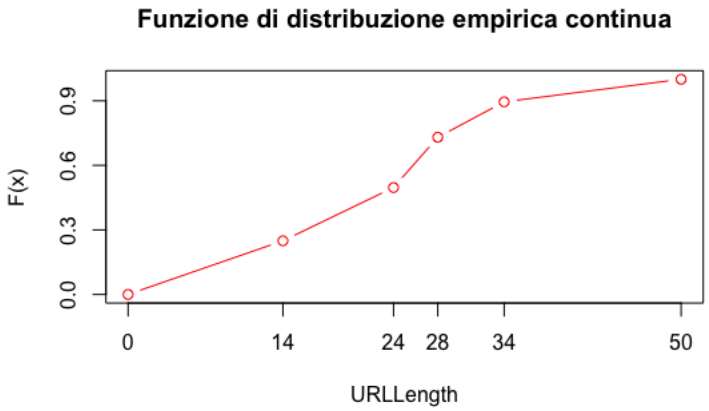

Figure 3: FdDEC per URLLength

Nel dataset, i valori della variabile **URLLength** variano da 14 a 1427. Tuttavia, per garantire un grafico chiaro ed equilibrato, abbiamo deciso di mostrare solo l'intervallo compreso tra 0 e 50. Questo consente di fornire comunque un'idea della distribuzione senza compromettere la leggibilità del grafico. Infatti, possiamo osservare:

- **Aumento graduale della distribuzione**: La funzione cresce in modo abbastanza regolare, indicando che i valori della lunghezza degli URL si distribuiscono in maniera progressiva senza grandi discontinuità nei primi 50 caratteri.
- **Presenza di intervalli meno densi**: Alcuni segmenti della funzione mostrano una crescita più lenta rispetto ad altri. Questo potrebbe indicare una minore densità di URL in quei range di lunghezza.
- **Crescita più ripida tra 24 e 34**: Si nota un incremento più marcato della distribuzione in questa fascia, il che suggerisce una maggiore concentrazione di URL con lunghezza compresa in questo intervallo.
- **Saturazione verso 50**: La curva si avvicina a 1, il che implica che una percentuale significativa di URL nel dataset si trovi sotto questa soglia. Tuttavia, dato che il dataset contiene valori fino a 1427, la distribuzione arriva fino ad 1.

3.3.2 NoOfExternalRef

La Funzione di Distribuzione Empirica Continua (FDC) per NoOfExternalRef, come per la variabile URLLenght, è stata calcolata per ciascun intervallo considerato nella tabella seguente.

Intervalli	Frequenze Assolute	Frequenze Relative	Frequenze Cumulate
[0,1]	7824	7824/23575	7824/23575
(1,10]	4126	4126/23575	11950/23575
(10,56]	5761	5761/23575	17711/23575
(56,27516]	5864	5864/23575	23575/23575

Table 8: Frequenze di NoOfExternalRef in base all'intervallo

La FDC è stata calcolata per evidenziare come le osservazioni si accumulino progressivamente, ed è definita nel seguente modo:

$$F(x) = \begin{cases} 0 & \text{se } x < min \\ 0.33 & \text{se } min \leq x < Q1 \\ 0.50 & \text{se } Q1 \leq x < Q2 \\ 0.75 & \text{se } Q2 \leq x < Q3 \\ 0.99 & \text{se } Q3 \leq x < max \\ 1 & \text{se } x \geq max \end{cases}$$

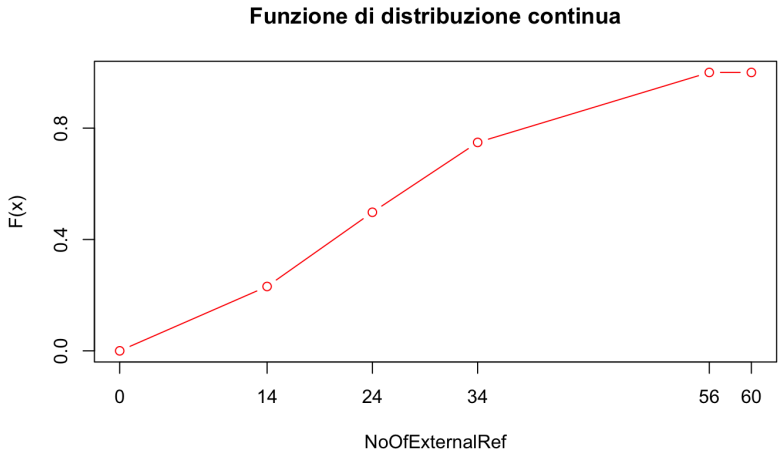

Figure 4: FdDEC per NoOfExternalRef

Nel dataset, i valori della variabile NoOfExternalRef variano da 0 a circa 27516. Tuttavia, per garantire un grafico chiaro ed equilibrato, il focus è su un intervallo rappresentativo che fornisce comunque un'idea della distribuzione senza compromettere la leggibilità del grafico.

Infatti, possiamo osservare:

- **Aumento graduale della distribuzione**: La funzione cresce in modo progressivo, suggerendo che il numero di riferimenti esterni presenti negli URL è distribuito in maniera relativamente uniforme senza brusche variazioni.
- **Presenza di intervalli meno densi**: Alcuni tratti della funzione mostrano un'inclinazione meno marcata, indicando che vi sono fasce in cui la densità degli URL con un certo numero di riferimenti esterni è inferiore.
- **Crescita più ripida tra 14 e 34**: Si osserva un incremento più rapido della distribuzione in questa fascia, il che suggerisce una maggiore concentrazione di URL con un numero di riferimenti esterni compreso in questo intervallo.
- **Saturazione verso 56**: La curva si avvicina a 1 nei pressi di 56 riferimenti esterni, il che implica che una percentuale significativa di URL nel dataset si trovi sotto questa soglia.

4 Analisi Descrittiva Bivariata

4.1 Regressione Logistica

La **regressione logistica** è una tecnica statistica utilizzata per modellare la probabilità di appartenenza a una determinata classe in un problema di classificazione binaria. A differenza della regressione lineare, che prevede valori continui, la regressione logistica stima la probabilità che un'osservazione appartenga a una categoria specifica, utilizzando la funzione logit(trasformazione matematica utilizzata nella regressione logistica per convertire probabilità in un intervallo illimitato di valori reali permettendo di modellare la relazione tra le variabili indipendenti e la probabilità di un evento).

Nel nostro caso, la regressione logistica è stata applicata per analizzare la relazione tra alcune caratteristiche degli URL e la probabilità che un sito web sia phishing o legittimo. In particolare, sono state utilizzate le variabili **URLLength** e **NoOfExternalRef** come predittori della classe **label**.
L'obiettivo dell'analisi è identificare le variabili più influenti nella classificazione degli URL e comprendere come queste caratteristiche influenzino la probabilità di phishing.

	Estimate	**Std. Error**	**p-value**
URLLength	-0.160	0.005	$< 2e-16$
NoOfExternalRef	0.560	0.011	$< 2e-16$

Table 9: Regressione Logistica

Dai risultati della seguente tabella possiamo trarre le seguenti conclusioni:

- **Valori di Estimate**: questi risultati ci danno un'idea sulla direzione dell'effetto, cioè sull'effetto che ogni variabile indipendente ha sulla probabilità che un sito sia di phishing. Infatti, aver ottenuto un coefficiente negativo per quanto riguarda URLLength, ci fa capire che all'aumentare della lunghezza dell'URL diminuisce la probabilità che un sito sia di phishing. Mentre, per quanto riguarda il coefficiente positivo associato a NoOfExternalRef, capiamo che più riferimenti esterni vi sono, più è alta la probabilità che sia un sito di phishing.
- **Valori di Std. Error**: questo elemento sta ad indicarci la variabilità della stima del coefficiente(Estimate). Avere un Std. Error molto grande significa che il coefficiente è incerto, invece, un Std. Error piccolo indica un coefficiente stimato con precisione. Nel nostro caso i valori sono molto bassi per entrambe le variabili, di conseguenza abbiamo che il coefficiente è stimato in modo preciso.
- **Valori di p-value**: da questi risultati capiamo se le variabili indipendenti sono significative nel predire quella dipendente. Infatti, visto che il loro p-value è molto basso, possiamo affermare che entrambe le variabili URLLength e NoOfExternalRef sono altamente significative per predire label.

Oltre questi valori abbiamo ottenuto anche valori come la **devianza nulla** e la **devianza residua** che però abbiamo deciso di non tenere in considerazione in quanto, data la natura dei nostri dati, non avrebbe avuto molto senso considerare un modello nel caso in cui le variabili indipendenti avessero avuto un valore pari a 0(per niente realistico).

5 Analisi dei Cluster

5.1 Clustering non gerarchico

Il clustering non gerarchico è una tecnica di clustering che mira a raggruppare elementi simili di una distribuzione, senza creare la struttura ad albero del dendrogramma, come invece accade con il **clustering gerarchico**. La scelta di utilizzare questo tipo di tecnica, è stata presa in base alla distribuzione delle due variabili analizzate fino ad ora. Infatti, non avendo dati qualitativi, sarebbe impossibile ricreare una struttura ad albero comprensibile, poichè si avrebbero troppi valori da inserire come foglie. L'algoritmo utilizzato è il K-means.

L'algoritmo del K-means è un metodo non supervisionato che cerca di raggruppare un insieme di dati in **K** cluster distinti in modo che gli oggetti all'interno di ciascun cluster siano più simili tra loro rispetto agli oggetti appartenenti a cluster differenti. Esso si basa sull'iterazione della scelta dei **centroidi** cluster, ossia punti centrali di una sotto-distribuzione, finchè non viene raggiunta una condizione di convergenza, che generalmente si ottiene quando i centroidi rimangono costanti.

Per valutare la bontà del clustering svolto, sono state utilizzate alcune metriche e metodi di riferimento:

- **Elbow method**:È una tecnica utilizzata per determinare il numero ottimale di cluster in un dataset. Consiste nel tracciare la variazione della somma dei quadrati intra-cluster (WCSS) in funzione del numero di cluster. Il *"punto di gomito"* è il punto in cui la riduzione di WCSS diminuisce drasticamente, indicando che l'aggiunta di ulteriori cluster non porta a miglioramenti significativi nella compattezza dei cluster;
- **Silhouette method**: Questo metodo valuta quanto ogni punto sia ben assegnato al suo cluster rispetto ai punti degli altri cluster. L'indice di Silhouette varia da -1 a +1, dove valori vicini a +1 indicano che il punto è ben separato dal cluster vicino, valori vicini a 0 suggeriscono che il punto si trova nel "confine" tra due cluster, e valori negativi indicano che il punto potrebbe essere assegnato al cluster sbagliato. L'indice di Silhouette complessivo per un clustering è dato dalla media degli indici di silhouette di tutti i punti;
- **Within-cluster sum of squares (WCSS)**:Rappresenta la somma dei quadrati delle distanze tra ogni punto e il centroide del cluster a cui appartiene. L'obiettivo è minimizzare questa misura, poiché valori più bassi indicano una maggiore compattezza dei cluster, ovvero i punti sono più vicini al loro centroide, suggerendo un buon raggruppamento;
- **Between-cluster sum of squares (BCSS)**: Misura la somma dei quadrati delle distanze tra i centroidi dei cluster e il centroide complessivo (media di tutti i punti). L'obiettivo è massimizzare questa misura, poiché valori più alti indicano che i cluster sono ben separati, ossia le distanze tra i centroidi dei cluster sono ampie;

- **Calinski-Harabasz Index**: È un indice che calcola il rapporto tra la *somma dei quadrati inter-cluster (BCSS)* e la *somma dei quadrati intra-cluster (WCSS)*, con una normalizzazione in base al numero di osservazioni e cluster. Un valore più alto dell'indice di Calinski-Harabasz indica un clustering migliore, con una separazione netta tra i cluster e una compattezza interna elevata.

Infine i cluster risultanti saranno rapportati alla label, per analizzare come si comportano e per identificare eventuali pattern nascosti fino ad ora.

5.1.1 URLLength

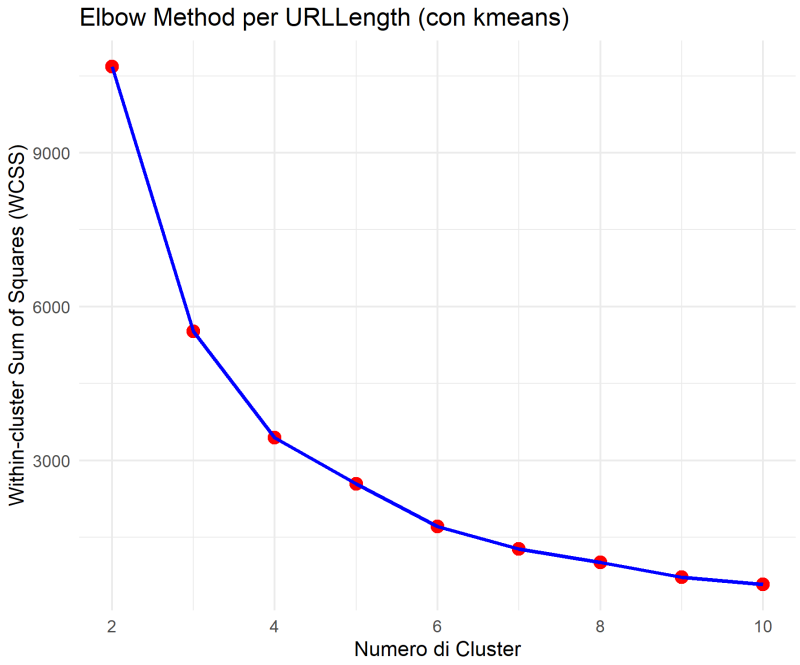

Figure 5: Elbow method normalizzato su URLLength

Analizzando il grafico 5 si vede come il punto di gomito sia situato tra 2 e 4, ossia i valori ottimali per condurre il clustering su URLLength. Aumentando il numero di cluster, la WCSS diminuisce sempre più, e sebbene questo sia un fatto positivo, dall'altra il vantaggio che si ottiene si riduce sempre più.

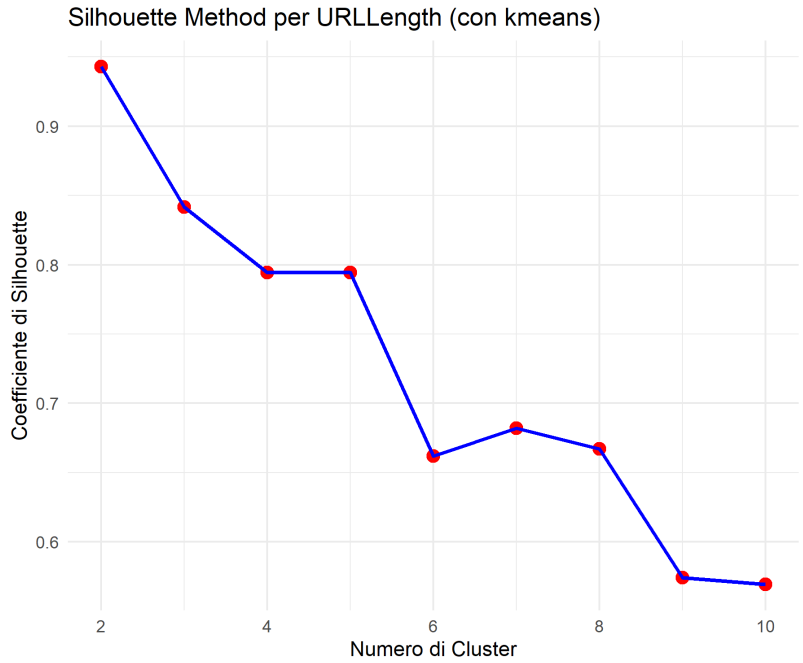

Figure 6: Silhouette method su URLLength

Per quanto riguarda il grafico 6, la silhouette identifica i valori ottimali per clusterizzare tra 2 e 4. In questo range, infatti, l'indice è compreso tra 0.94 (quando il numero di cluster è 2) e 0.79 (quando il numero di cluster è 4). Quando, invece, il numero di cluster è 3, il valore della silhouette è 0.84. Aumentando ulteriormente il numero di cluster, la curva cala, il che indica che il clustering è meno efficace nella divisione degli elementi. Tra 8 e 10 abbiamo i valori peggiori per la silhouette.

Di seguito viene mostrata la tabella con gli indici di valutazione del clustering, quando questo viene effettuato con 2, 3 o 4 cluster.

Numero di Cluster	**WCSS**	**BCSS**	**Silhouette**	**CH index**
2	10680	12893	0.94	28454
3	5521	18052	0.84	38538
4	3439	20134	0.79	45993

Table 10: Indici relativi al numero di cluster per URLLength

Analizzando la tabella 10 si vede come al crescere del numero di cluster, tutti gli indici subiscono una netta variazione. In particolare si ha che:

- **WCSS e BCSS** sono misure inversamente proporzionali tra loro, ossia quando la prima diminuisce l'altra aumenta. WCSS subisce un dimezzamento quando si passa da 2 a 3 cluster, segno del fatto che la distanza tra gli elementi è diminuita, il che rappresenta un fatto positivo. BCSS ha una variazione più grande da 2 a 3 cluster che da 3 a 4, ossia la distanza inter cluster cresce ma non aumenta come prima, ossia i valori stanno convergendo;

- la **Silhouette** tende a diminuire con l'aumento dei cluster ma non degrada in maniera repentina, anche se scende sotto lo 0.8 già quando si hanno "solo" 4 cluster;
- il **CH Index** è una misura correlata a WCSS e BCSS, quindi se esse subiscono una variazione, anche l'indice cambia. Con l'aumento dei cluster, questa misura subisce un aumento, ma che tende a diminuire man mano.

Analizzando i valori risultanti si può affermare che la scelta ottimale del numero di cluster ricada su 3, in quanto è il valore che rappresenta meglio il bilanciamento tra le varie metriche utilizzate. Quando usiamo 2 cluster WCSS e BCSS hanno valori troppo vicini, mentre quando ne usiamo 4, il valore della silhouette risulta essere minore rispetto agli altri.

Di seguito vengono plottati i dati quando viene effettuata la divisione in 3 cluster.

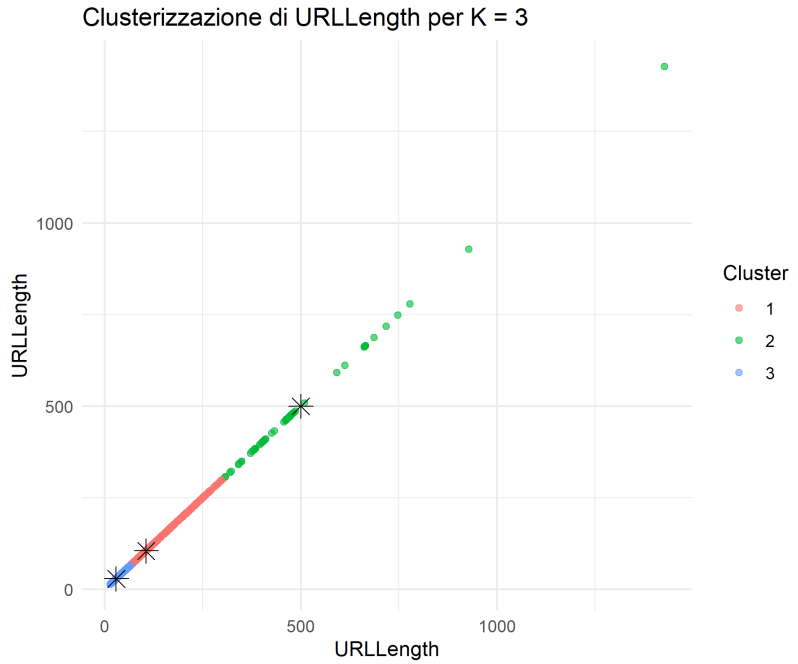

Figure 7: Visualizzazione dei cluster di URLLength

Il grafico mostra la divisione degli elementi di URLLength in cluster: nella visualizzazione, le stelline nere rappresentano i centroidi. I centroidi dei cluster 1 (arancione) e 3 (azzurro) sono molto ravvicinati tra loro, rispetto al centroide del cluster 2.

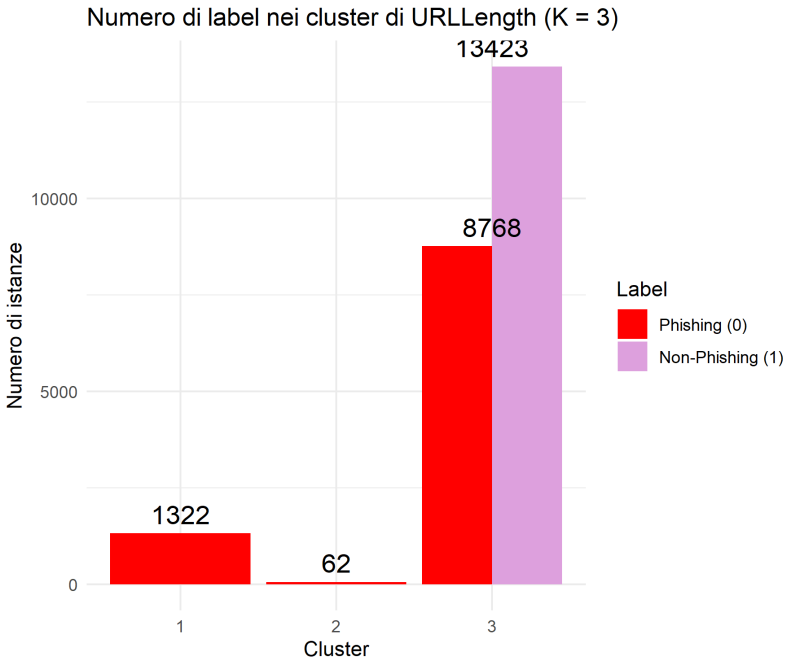

Figure 8: Distribuzione della label rispetto ai cluster di URLLength

Il grafico 8 a differenza del grafico 7 rappresenta meglio quella che è la distribuzione degli elementi nei cluster. Il terzo cluster è quello che contiene la quasi totalità degli elementi ed in particolare la totalità di elementi associati a dati Non Phishing. I cluster 1 e 2 sono molto meno popolati e la totalità dei loro elementi è composta da dati etichettati come Phishing. Il 1 e il 2 cluster vanno ad identificare gli outlier della distribuzione totale, ossia quei valori molto distanti dalla maggior parte degli elementi.

5.1.2 NoOfExternalRef

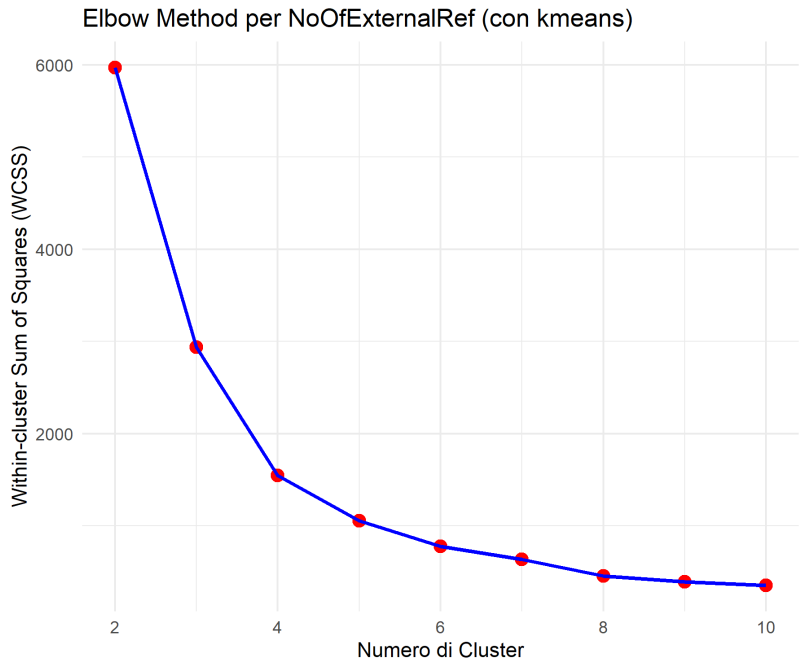

Figure 9: Elbow method normalizzato su NoOfExternalRef

Analizzando il grafico 9, si vede come anche in questo caso il punto di gomito sia situato tra 2 e 4. Tra i valori 5 e 10, i valori di WCSS sono più assottigliati, il che indica che non c'è un grosso miglioramento.

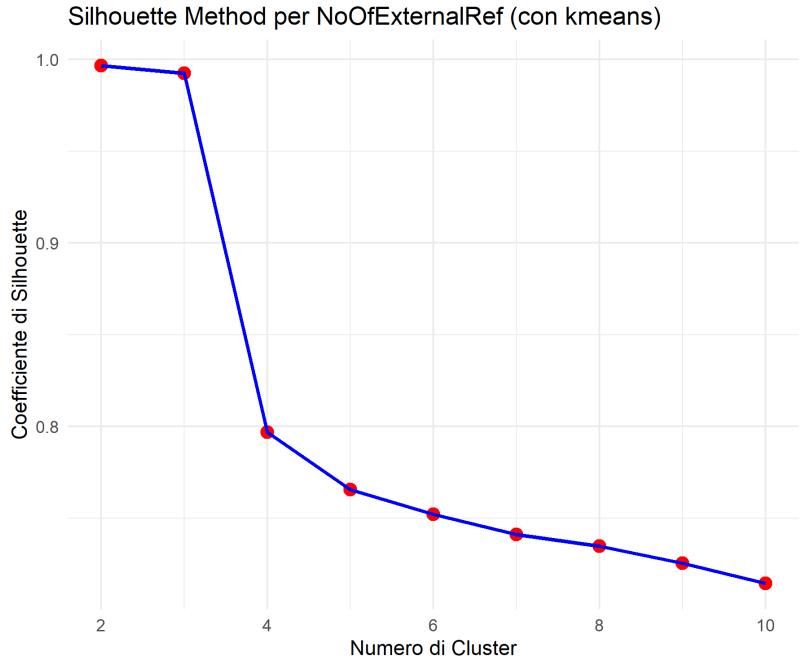

Figure 10: Silhouette method su NoOfExternalRef

Per quanto riguarda il grafico 10, i valori della silhouette sono compresi tra 0.7 (nel caso peggiore) ed 1.0 (nel caso migliore), quindi anche se scegliessimo il numero di cluster maggiore, la loro qualità sarebbe comunque sufficientemente elevata. Tuttavia, tra 2 e 4, i valori della silhouette sono più grandi, e questo indica che in quella fascia il clustering ha una qualità migliore.

Di seguito viene mostrata la tabella con gli indici di valutazione del clustering, quando questo viene effettuato con 2, 3 o 4 cluster.

Numero di Cluster	**WCSS**	**BCSS**	**Silhouette**	**CH Index**
2	5971	17602	0.996	69490
3	2938	20635	0.992	82756
4	1544	22029	0.79	112031

Table 11: Indici relativi al numero di cluster per NoOfExternalRef

Analizzando la tabella 11 si vede come al crescere del numero di cluster, tutti gli indici subiscono una netta variazione, tranne la silhouette. In particolare si ha che:

- **WCSS e BCSS** hanno valori molto distanti già dall'inizio, a differenza di URLLength. Questo indica che già con 2 cluster la divisione è piuttosto ottimale;
- la **Silhouette** tra 2 e 3 cluster è praticamente costante, mentre cala maggiormente con 4 cluster;
- il **CH Index** subisce un un grande aumento in tutte e 3 i casi osservati.

Anche in questo caso la scelta ottimale del numero di cluster ricade su 3. La silhouette è altissima, ma lo sono anche gli altri indici considerati.

Di seguito vengono plottati i dati quando viene effettuata la divisione in 3 cluster.

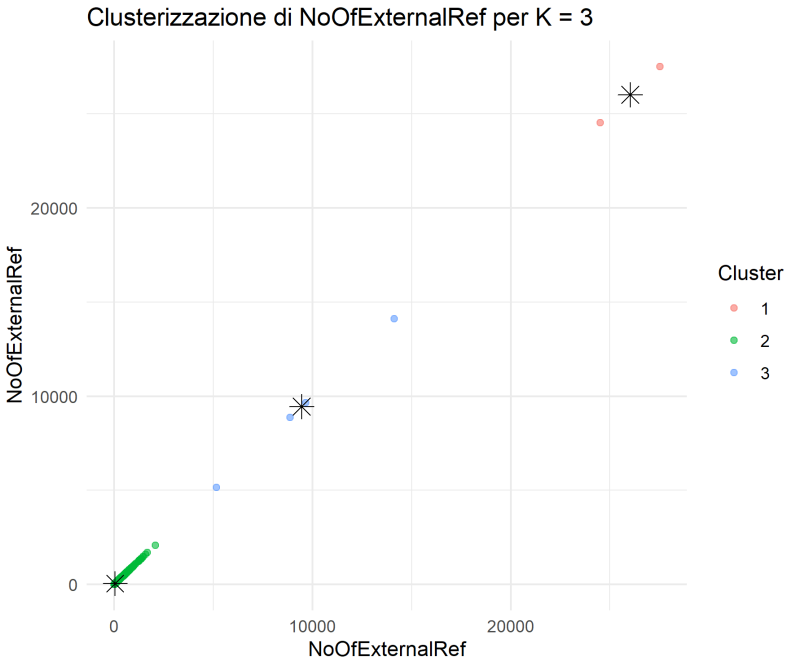

Figure 11: Visualizzazione dei cluster di NoOfExternalRef

In questo caso, rispetto ad URLLength, la divisione dei cluster appare più immediata poichè i valori sono disposti in maniera tale da formare naturalmente dei gruppi separati. Rispetto alla situazione di URLLength, dove il grafico 7 era meno esplicito nella visualizzazione del numero di elementi, qui come tutti i valori siano compresi nel cluster 2 (verde). Questo dato sarà confermato anche dal grafico successivo.

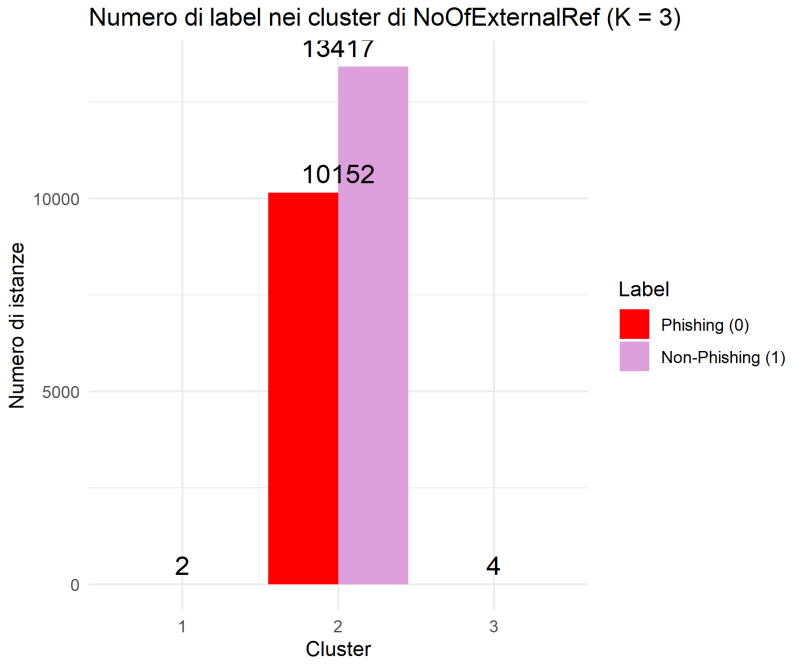

Figure 12: Distribuzione della label rispetto ai cluster di NoOfExternalRef

In questo grafico, infatti, la distinzione tra cluster per numero di elementi è ancora più netta.Come visto nella Figura 8, anche qui un cluster domina sugli altri per numero di elementi.
Mentre la totalità degli elementi è racchiusa nel cluster 2, i cluster 1 e 3 includono gli outlier della distribuzione e in particolare solo elementi etichettati come label = 1, ossia Non Phishing.

5.2 Conclusioni

Dopo aver analizzato il clustering prodotto per entrambe le variabili in esame possiamo dire che:

- **URLLength**: dato che la quasi totalità dei valori è inclusa in un unico cluster è difficile riuscire a fare una distinzione tra dati phishing e non, questo perchè si trovano nello stesso gruppo. Ciò che emerge però, è che i cluster con meno elementi, e quindi con più outlier, contengono solo valori etichettati come phishing. Questo conferma, anche se solo in parte visto che il numero di elementi totali nei cluster 1 e 2 sono pochi, che aumentando la lunghezza dell'URL, la probabilità di trovarsi di fronte ad un sito di phishing aumenta.

- **NoOfExternalRef**: anche in questo caso non abbiamo una netta separazione in cluster tra le due label, dato che la totalità degli elementi è contenuta nel cluster 2. Nei cluster 1 e 2, però, gli elementi sono tutti etichettati come Non Phishing. Si potrebbe quindi pensare che maggiore sia il numero di riferimenti esterni e maggiore è la probabilità di trovarsi in un sito sicuro, ma visto che si parla di solo 6 elementi questa supposizione ha meno validità rispetto a quella fatta per URLLength.

6 Inferenza statistica

6.1 Verifica della normalità delle distribuzioni

Nell'analisi inferenziale, è fondamentale verificare se le variabili seguono una distribuzione normale, poiché la scelta dei test statistici dipende dalla natura dei dati. Alcuni test parametrici, come il test t di Student, richiedono che i dati siano normalmente distribuiti, mentre test non parametrici, come il test di Mann-Whitney U, sono utilizzati quando la normalità non è rispettata. L'obiettivo di questa fase è determinare se le variabili di interesse, URLLength e NoOfExternalRef, seguono una distribuzione normale nei due gruppi di studio: siti phishing e siti legittimi.

6.1.1 Shapiro - Wilk

Useremo il test Shapiro-Wilk per verificare se una variabile segue una distribuzione normale. Le ipotesi che andremo a considerare sia per la variabile URLLength che per NoOfExternalRef sono le seguenti:

- Ipotesi H_0: La variabile **segue** una distribuzione normale.
- Ipotesi H_A: La variabile **non segue** una distribuzione normale.

Inoltre, visto che il dataset è composto da 23575 righe e che il test di Shapiro-Wilk è sensibile al numero di osservazioni, si è deciso di procedere con un campionamento di 500 individui per entrambe le variabili in analisi. Dall'applicazione del test sul campionamento abbiamo ottenuto i seguenti risultati:

	W	**p-value**
URLLength	0.275	$< 2.2e-16$
NoOfExternalRef	0.040	$< 2.2e-16$

Table 12: Shapiro-Wilk applicato su un campione di 500 elementi

Dai risultati ottenuti, possiamo notare che sia W che il p-value sono estremamente piccoli per entrambe le variabili. Poiché W assume valori tra 0 e 1 e, quando è vicino a 0, indica una deviazione dalla normalità, e dato che un p-value ¡ 0.05 porta a rifiutare l'ipotesi di normalità, possiamo concludere che la distribuzione non è normale. Permettendoci quindi di rifiutare H_0 e accettare l'ipotesi alternativa.

6.1.2 Q-Q plot

Per dare anche una rappresentazione e conferma grafica dei risultati ottenuti dal Shapiro-Wilk abbiamo deciso di rappresentare questa distribuzione con il Q-Q plot:

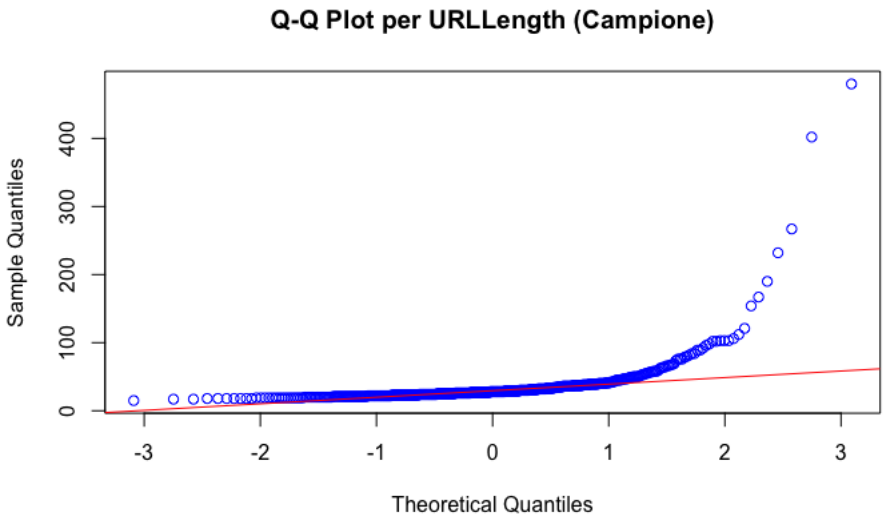

Figure 13: Q-Qplot per URLLength (campione)

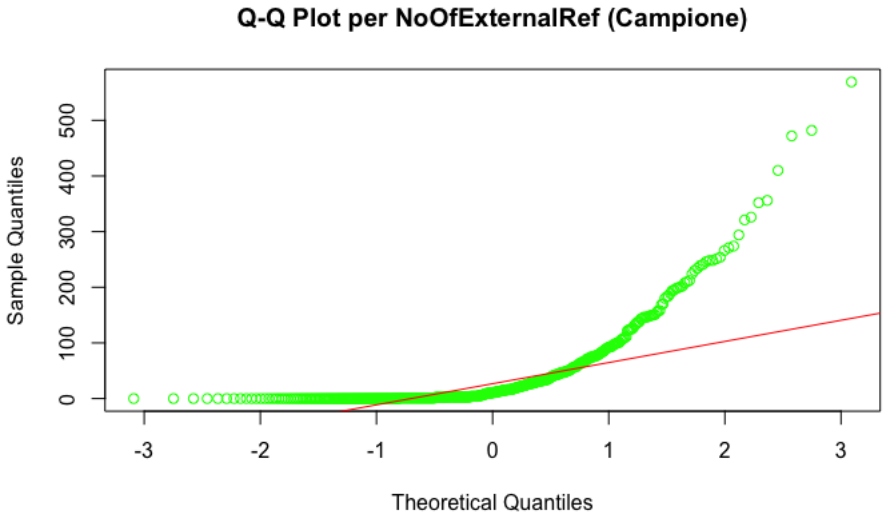

Figure 14: Q-Qplot per NoOfExternalRef (campione)

Come possiamo notare, i grafici confermano quanto detto dal Shapiro-Wilk visto che in entrambi i grafici i punti deviano fortemente dalla linea diagonale. Se la distribuzione fosse stata normale, i punti avrebbero dovuto seguire quest'ultima, ma dato che qui sono presenti code lunghe e deviazioni significative, abbiamo conferma che le variabili non seguono una distribuzione normale.

6.2 Test di significatività

Dopo aver trattato la normalità dei dati possiamo passare al test di significatività. Per il nostro caso specifico abbiamo dovuto applicare il test di Mann-Whitney U in quanto stiamo trattando una distribuzione non normale.

6.2.1 Mann-Whitney U (se non normale)

Il test di Mann-Whitney U è un test non parametrico(quindi senza fare assunzioni sulla distribuzione dei dati) che confronta due gruppi indipendenti(siti di phishing, siti legittimi) per determinare se ci sono differenze significative tra le loro distribuzioni. Le ipotesi considerate in questo caso sono le seguenti

- Ipotesi H_0: Le distribuzioni di URLLength (o NoOfExternalRef) sono **uguali** tra i siti phishing e legittimi.
- Ipotesi H_A: Le distribuzioni di URLLength (o NoOfExternalRef) sono **diverse** tra i siti phishing e legittimi.

Di seguito sono riportati i risultati ottenuti nell'applicazione del Mann-Whitney U e nel calcolo dell'effetto delle dimensioni.

	No.Obs.Phishing	**No.Obs.non-Phishing**	**statistic U**	**p-value**
URLLength	10152	13423	101808634	0
NoOfExternalRed	10152	13423	1753022	0

Table 13: Mann-Whitney U per URLLength e NoOfExternalRef

	effsize	**magnitude**
URLLength	0.4242659	moderate
NoOfExternalRef	0.8413069	large

Table 14: Effetto della dimensione per URLLength e NoOfExternalRef

Dai risultati possiamo notare che:

- Risultati test Mann-Whitney U:
  - **URLLength**: Per quanto riguarda i risultati di URLLength possiamo notare subito che lo statistic U presenta un valore molto alto, questo è dovuto anche alla grossa mole di osservazioni presenti nel dataset. Questo valore sta ad indicare quante volte i valori di un gruppo superano quelli di un altro, questo significa che più alto è il valore di U e maggiore sarà la differenza tra i due gruppi. Questo vi viene confermato anche grazie al p-value che, in questo caso, assume valore 0 indicandoci che la differenza è statisticamente significativa.
  - **NoOfExternalRef**: Anche per questa variabile lo statistic U presenta un valore molto alto e un p-value pari a 0, indicando un grossa differenza tra i gruppi di phishing e non-phishing.

- Risultati effetto della dimensione: Tuttavia, questi valori ottenuti dal Mann-Whitney U da soli non bastano. Infatti, abbiamo calcolato l'effetto della dimensione per entrambe le variabili per capire se questa differenza è effettivamente rilevante.
  - **URLLength**: Dai risultati ottenuti per URLLength notiamo che l'effsize ci indica che questi valori sono rilevanti in modo **moderato** in quanto il suo valore è 0.3 ¡ effsize ¡ 0.5.
  - **NoOfExternalRef**: I risultati per NoOfExternalRef, invece, mostrano che questa variabile è statisticamente significativa e rilevante in quanto il suo effsize è molto vicino ad 1.

Il parametro **magnitude** non è altro che la descrizione qualitativa dell'effsize.

In conclusione, poiché l'ipotesi H_0 è stata rifiutata per entrambe le variabili(in quanto la differenza tra i due gruppi è elevata) e l'effetto della dimensione è grande per NoOfExternalRef e moderato per URLLength, possiamo confermare che il numero di riferimenti esterni è significativamente più alto nei siti phishing rispetto ai siti legittimi, mentre la lunghezza dell'url non è così determinante per avere una corretta distinzione. NoOfExternalRef potrebbe quindi essere un fattore discriminante utile nell'identificazione di siti malevoli.

6.3 Intervalli di Confidenza

Nella fase precedente, abbiamo determinato che NoOfExternalRef non segue una distribuzione normale e che è l'unica tra le due variabili considerate a poter essere presa in considerazione come fattore per l'individuazione di URL malevoli. Per questo motivo, adottiamo un approccio non parametrico per costruire gli intervalli di confidenza (proprio perché non segue una distribuzione normale). Abbiamo scelto un grado di confidenza pari a 1 - $\alpha = 0.95$, con $\alpha = 0.05$, che rappresenta un compromesso ottimale tra affidabilità e precisione delle stime.

6.3.1 Metodo basato sui Quantili

Poiché non possiamo assumere che la popolazione segua una distribuzione normale, stimiamo l'intervallo di confidenza attraverso i quantili empirici della distribuzione osservata. Vogliamo determinare l'intervallo di fiducia per la mediana della popolazione, considerando le seguenti casistiche:

(i) Intervallo di confidenza per la mediana di NoOfExternalRef nei siti phishing (label=0). L'intervallo di confidenza è determinato come:

$$[Q_{2.5\%}, Q_{97.5\%}]$$

Dove Q2.5% e Q97.5% sono rispettivamente 2.5 e 97.5 percentile dei dati osservati nel gruppo phishing.

(ii) Intervallo di confidenza per la mediana di NoOfExternalRef nei siti legittimi (label=1). Anche per questo gruppo, l'intervallo di confidenza è calcolato usando i quantili:

$$[Q_{2.5\%}, Q_{97.5\%}]$$

6.3.2 Risultati

I risultati ottenuti sono i seguenti: Dai risultati ottenuti notiamo che gli intervalli di

	Inter. Inferiore	**Inter. Superiore**
Phishing	0	6.00
non-Phishing	3	362.45

Table 15: Intervallo di Confidenza per NoOfExternalRef

confidenza si sovrappongono poiché $U_{Phishing} = 6$ e $L_{non-Phishing} = 3$. Questo ci fa capire che:

- La differenza tra phishing e legittimi esiste, ma non è netta.
- Ci sono siti phishing che hanno un numero di riferimenti esterni simile a quelli dei siti legittimi.
- Questo indica che NoOfExternalRef da solo non è sufficiente per separare completamente i due gruppi.

I siti legittimi hanno una maggiore variabilità nel numero di riferimenti esterni (range enorme: da 3 a 362). I siti phishing hanno valori più concentrati tra 0 e 6, suggerendo che in media hanno meno riferimenti esterni. Questo ci fa capire che NoOfExternalRef è un buon indicatore ma non sufficiente da solo a distinguere in modo netto phishing da siti legittimi.

7 Confronto tra dataset originale e dataset generato

Per la generazione del dataset sintetico, è stato utilizzato **ChatGPT 4o**. Questo capitolo, infatti, mostra le fasi che hanno condotto alla generazione del dataset e l'analisi della statistica descrittiva per confrontare i dati a disposizione.

7.1 Analisi preliminare del dataset generato

Più precisamente, la generazione del dataset sintetico è stata realizzata utilizzando un approccio di **Prompt Engineering**, ovvero la formulazione strutturata di richieste testuali per guidare un modello di linguaggio nella produzione di dati con caratteristiche desiderate. L'obiettivo di questa fase era ottenere un dataset con caratteristiche simili a quelle del dataset reale,

7.1.1 Fase di Prompt Engineering

Il prompt è stato progettato in modo chiaro e dettagliato per specificare al modello le seguenti condizioni:

- **Formato Tabellare**: richiesta di un output ben strutturato come quello del dataset reale;
- **Range e Tipo di valori**;
- **Livello Randomicità**: per capire quanto i dati devono distanziarsi dai valori presenti nel dataset reale;

Di seguito è riportata la cronologia delle richieste effettuate per la generazione del dataset:

- **Prima Richiesta**:
  - Prompt: "Forniscimi tutte le informazioni di cui necessiti per la generazione di un dataset di dati sintetici."
  - Output: Le informazioni necessarie alla corretta generazione.
  - Problema: Nessuno.
- **Seconda Richiesta**:
  - Prompt: Elenco delle informazioni richieste come: Dataset reale, caratteristiche delle variabili (eventuali relazioni), volume dei dati e richiesta di un livello di randomizzazione medio.
  - Output: Dataset sintetico problematico.
  - Problema: Le variabili necessarie all'analisi assumevano valori negativi.
- **Terza richiesta**:
  - Prompt: Definizione dei limiti da rispettare e chiarimento delle caratteristiche delle variabili, ad es: "le variabili x, y e z non possono assupere valori negativi".
  - Output: Dataset sintetico finale.
  - Problema: Nessuno.

7.1.2 Struttura e descrizione

Il dataset generato presenta 23575 righe e 56 colonne, mantenendo il numero di valori totali del dataset originale. Non ci sono dati mancanti né dati ripetuti, il che consente un'analisi completa. Alcune colonne, tuttavia, presentano dei valori differenti rispetto a quelle originali. L'analisi di confronto, però, non sarà condotta su queste, bensì su quelle viste fino ad ora, ossia URLLength e NoOfExternalRef, che presentano dei valori in $\mathbb{N}$.

7.2 Statistica descrittiva del dataset generato

Sono stati utilizzati gli indici di centralità e gli indici di dispersione per comprendere analogie e differenze tra i due dataset. Per effettuare questi confronti viene utilizzata la stessa struttura vista nelle Tabelle 1 e 2.

7.2.1 URLLength

	MIN	**MAX**	**MEDIA**	**MEDIANA**	**MODA**
Completo	0	103	34.74	35	32
Phishing(label=0)	0	103	34.74	35	32
NonPhishing(label=1)	0	86	34.76	34	37

Table 16: Indici di centralità di URLLength (generato) rispetto alla label

Analizzando la tabella si ha che:

- **Il range tra MIN e MAX** in cui varia URLLength è nettamente inferiore rispetto al dataset originale, e questo comporta che i dati sono molto più vicini tra loro.
- **La media** è stabile in tutti e 3 i casi considerati, mentre prima era più alta per il caso di phishing e più bassa per il caso opposto. E' costante rispetto al caso completo originale.
- **La mediana** è intorno al valore 35 in tutti i casi, a differenza della tabella 3 dove i dati erano molto diversi tra loro.
- **La moda** è uguale nel caso completo e nel caso phishing, mentre varia nel caso in cui la label = 1. Questo accadeva anche nel dataset originale, ma la moda nel caso NonPhishing era quasi identica agli altri due casi, mentre qui subisce un aumento più netto.

	CV	Scarto Interquartile	Deviazione Standard
Completo	0.47	23	16.40
Phishing(label=0)	0.47	23	16.40
NonPhishing(label=1)	0.47	22	16.43

Table 17: Indici di Dispersione di URLLength (Generato)

Per quanto riguarda gli indici di dispersione, la situazione è la seguente:

- **Il Coefficiente di Variazione (CV)** è pari a 0.47 in tutti e 3 i casi. In precedenza questo valore era molto maggiore nel caso completo e nel caso Phishing, e minore nel caso NonPhishing;
- **Lo Scarto Interquartile** è pari a 23 nei primi due casi, mentre è 22 nel terzo. Nel dataset originale lo scarto era inferiore nei casi Completo e NonPhishing, mentre era lo stesso nel caso Phishing;
- **I valori della Deviazione Standard** sono praticamente identici nei tre casi, a differenza del dataset originale dove il caso Completo e il caso Phishing presentavano numeri più grandi, mentre il caso NonPhishing era molto inferiore;

Per riassumere, la variabile URLLength del dataset generato presenta una variabilità ridotta, un range di dati inferiore e valori quasi identici in tutti i casi. Questo significa che non c'è una netta separazione tra le 3 situazioni e in particolare quando la label è 0 e quando è ad 1.

7.2.2 NoOfExternalRef

Passando alla seconda variabile, i dati a disposizione sono riassunti nella tabella seguente.

	MIN	MAX	MEDIA	MEDIANA	MODA
Completo	0	583	121.49	103	1
Phishing(label=0)	0	583	121.65	103	1
NonPhishing(label=1)	0	489	117.47	104	72

Table 18: Indici di centralità di NoOfExternalRef (generato) rispetto alla label

Si ha che:

- **la distanza tra MIN e MAX** è piuttosto ampia in tutti e tre i contesti, e l'unica differenza la si nota nel caso NonPhishing dove i valori possibili sono leggermente di meno. Nel dataset originale si aveva un dominio di valori maggiore nel caso Completo e nel caso NonPhishing, mentre il caso Phishing aveva un range assai più ristretto di quello appena visto;
- **La media** è intorno al valore di 120, mentre in precedenza aveva valori molto minori. In particolare nel caso Phishing c'era una media pari a 1.21, molto inferiore rispetto a quella attuale;
- **la mediana** non assume valori diversi tra loro, a differenza del dataset originale, dove tutti erano univoci;

- **La moda** presenta valori uguali per il caso Completo e Phishing in entrambi i dataset (0 nell'originale e 1 in quello generato), ma quando la label = 1 in entrambi i dataset si ha un valore più alto, ma in quello generato questo incremento è molto maggiore.

La tabella seguente mostra gli indici di dispersione.

	CV	**Scarto Interquartile**	**Deviazione Standard**
Completo	0.75	125	91.72
Phishing(label=0)	0.75	126	91.87
NonPhishing(label=1)	0.74	120.5	87.83

Table 19: Indici di Dispersione di NoOfExternalRef (Generato)

I dati mostrano che:

- **Il CV** è molto inferiore al dataset originale, dove i valori erano sia più grandi ma anche diversi tra loro;
- **Lo Scarto Interquartile** presenta un grande aumento rispetto alla situazione di partenza. La differenza più grande però è rispetto al caso Phishing, dove abbiamo 1 per NoOfExternalRef originale e 126 per quella generata;
- **La Deviazione Standard** è ampia ma non come in precedenza, analizzando in particolare il caso Completo e il NonPhishing. Per il caso Phishing, invece, c'era una situazione molto più stabile, con un valore di 3.14, contro un valore attuale di 91.87.

In definitiva, anche NoOfExternalRef presenta la situazione già vista per URLLength. Entrambe le variabili generate hanno valori più ridotti, meno variabili e che non distinguono per niente le 3 situazioni di analisi.

7.3 Regressione Logistica Dataset Sintetico

Il passo successivo consiste nel confrontare i risultati della regressione del dataset reale con quelli ottenuti su un dataset sintetico generato. Questo confronto è fondamentale per comprendere eventuali differenze nelle prestazioni del modello.

	Estimate	**Std. Error**	**p-value**
URLLength	$6.615e-05$	$2.078e-03$	0.975
NoOfExternalRef	$-5.070e-04$	$3.796e-04$	0.182

Table 20: Regressione Logistica Dataset Sintetico

Dai risultati della tabella possiamo trarre le seguenti conclusioni:

- **Valori di Estimate**: i risultati ottenuti per **URLLength** indicano che all'aumentare della lunghezza dell'URL aumenta leggermente la probabilità che un sito sia legittimo, tuttavia è un valore troppo piccolo per essere significativo. Per **NoOfExternalRef**, invece, questo valore indica che all'aumentare del numero di riferimenti esterni aumenta leggermente la probabilità che un sito sia di phishing.
- **Valori di Std.Error**: questi valori sono piccoli per entrambe le variabili, assicurandoci che la stima dei coefficienti è precisa.
- **Valori di p-value**: nessuna delle due feature è significativa in quanto il loro p-value > 0.05.

7.4 Discussione sui limiti del dataset generato

I dati sintetici non mostrano un'associazione significativa tra le feature e la probabilità di phishing, a differenza di quanto ci si aspetterebbe dai dati reali. Questo potrebbe suggerire che la generazione dei dati sintetici non ha mantenuto le relazioni chiave, rendendo il dataset meno utile per la modellazione predittiva. Un miglioramento potrebbe derivare da un processo di prompt engineering più accurato, aumentando la specificità e riducendo la casualità. Tuttavia, questo rischierebbe di introdurre una rigidità eccessiva nella generazione dei dati, limitando il contributo dell'intelligenza artificiale e la sua capacità di apprendere e riprodurre pattern complessi in modo autonomo.

8 Conclusioni Finali

In questa sezione andremo a rispondere alle RQ facendo una ricapitolazione dei risultati finali ottenuti.

8.1 Risposta RQ1

RQ$_1$. *Qual è l'impatto della lunghezza degli URL sulla classificazione di questi ultimi come URL di phishing?*

L'analisi della variabile URLLength ha mostrato che all'aumentare della sua lunghezza cresce la probabilità che un sito sia di phishing. Tuttavia, la sola lunghezza dell'URL non è un fattore sufficiente per distinguere in modo certo tra siti legittimi e siti di phishing.

8.2 Risposta RQ2

RQ$_2$. *Qual è l'impatto dato dal numero di link esterni in un URL sulla classificazione di questi ultimi come URL di phishing?*

L'analisi effettuata su NoOfExternalRef mostra come il numero di riferimenti esterni sia inversamente proporzionale alla probabilità di trovarsi di fronte ad un sito di phishing. In un contesto del genere, l'attaccante cerca di mantenere l'utente su una determinata pagina piuttosto che lasciare che "scappi" in altre. Questo gli consente di prelevare dati sensibili in maniera più agevole. Anche in questo caso però abbiamo visto che questa variabile da sola non è sufficiente a distinguere i due casi.

8.3 Risposta RQ3

RQ$_3$. *Quali sono le differenze principali tra dati sintetici generati e dati reali in un contesto di regressione e descrittivo?*

I dati generati da ChatGPT mostrano scarsa capacità di distinguere i vari casi trattati e scarsa comprensione delle relazioni tra le variabili. Ciò non consente un'analisi profonda sul problema affrontato.