PROGETTO DI STATISTICA E ANALISI DEI DATI

Analisi statistica dei domini all'interno degli URL di Phishing

Autori
Daniele Fabiano
Mariantonietta Maselli
Università degli Studi di Salerno
A.A. 2024/2025

Indice

1 Introduzione

1.1 Contenuti del documento

Questo documento rappresenta un report che riassume i risultati prodotti in merito all'analisi statistica effettuata. Sulla base del dataset scelto, è stato realizzato uno studio statistico relativo alle caratteristiche dei domini presenti all'interno degli URL di phishing, utilizzando gli strumenti a nostra disposizione. Il documento è strutturato come segue:

- **Capitolo 1 - Introduzione**: In questo capitolo viene descritta la struttura del documento e sono ripresi alcuni concetti utili per comprendere al meglio il dominio di applicazione. Viene inoltre fornita una presentazione del dataset scelto e delle feature in esso contenute;
- **Capitolo 2 - Statistica Descrittiva**: In questo capitolo sono riportate tutte le informazioni utili, estratte dal dataset mediante l'utilizzo delle statistiche descrittive univariate e bivariate. Vengono inoltre utilizzati diversi grafici per mostrare le frequenze e le distribuzioni dei dati, e viene presentata una tecnica di regressione per effettuare la stima di una variabile;
- **Capitolo 3 - Statistica Inferenziale**: In questo capitolo si intende studiare la popolazione a cui fa riferimento il dataset iniziale, scegliendo un sample randomico di riferimento. Una volta individuata la distribuzione della popolazione, sono state effettuate delle stime per una variabile fissata;
- **Capitolo 4 - Generazione sintetica dei dati**: In questo capitolo sono descritti i passaggi effettuati per realizzare un prompt da fornire a un LLM, con l'obiettivo di ottenere un sample sintetico dei dati. Ai fini di analizzare il sample, sono state impostate alcune Research Question;
- **Capitolo 5 - Conclusioni**: In questo capitolo vengono brevemente discussi diversi aspetti per possibili miglioramenti futuri.

Al seguente ***repository***, sono presenti tutti gli artefatti prodotti per questo studio, che prevedono gli script in R per effettuare le varie analisi, oltre che le versioni del dataset ottenute in seguito alle relative operazioni/trasformazioni.

1.2 Che cos'è un URL

Un URL (*Uniform Resource Locator*) è l'indirizzo di una risorsa unica sulla rete internet e il principale meccanismo usato dai browser web per recuperare le risorse pubbliche [1].

1.2.1 Struttura di un URL

Un URL segue una particolare struttura che viene descritta di seguito, con il supporto di un'immagine:

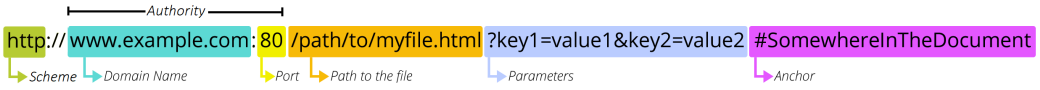

Figura 1: La struttura di un URL.

- **Scheme**: La prima parte dell'URL, indica quale protocollo viene usato dal client per recuperare la risorsa. In genere per i browser, il protocollo utilizzato è HTTP/HTTPS;
- **Domain Name**: Sostituibile anche da un indirizzo IP pubblico, questa parte dell'URL indica su quale Server deve essere inviata la richiesta per il recupero delle risorse. Anche il nome di un dominio segue a sua volta una particolare struttura[2]:
  - **Top Level Domain (TLD)**: Questa parte del dominio indica al client lo scopo generali dei servizi che vengono offerti. Nel nostro esempio il TLD è *.com*;
  - **Second Level Domain (SLD)**: Questa parte del dominio è l'etichetta principale del dominio, per la quale un'organizzazione acquisisce i diritti di utilizzo;
  - **Third Level Domain**: A partire dall'unione di TLD e SLD, possono essere aggiunte ulteriori etichette per ottenere un terzo livello del dominio. Questa parte viene anche denominata come subdomain.
- **Port**: Questa parte dell'URL indica un numero identificativo che permette ai client di interagire col processo sul server che fornisce il servizio. Nell'utilizzo dei browser web e dei protocolli HTTP/HTTPS, il numero di porta 80/443 viene solitamente omesso;
- **Path to resource**: Questa parte dell'URL definisce il percorso fisico/logico attraverso il quale un client può recuperare le risorse dal server;
- **Parameters**: Questa parte dell'URL permette di inviare dei parametri aggiuntivi al server per completare le operazioni (ad esempio delle query) di recupero delle risorse;
- **Anchor**: Questa parte dell'URL è tipicamente utilizzata per raggiungere in maniera più rapida delle risorse presenti su una pagina HTML. Questa parte dell'URL viene anche denominata come *fragment*.

[1] MDN Web Docs - What is a URL?
[2] MDN Web Docs - What is a Domain Name?

1.3 Il problema del phishing

Con il termine *phishing* si fa riferimento ad una famiglia di attacchi informatici che hanno lo scopo di rubare le informazioni dell'utente come le credenziali di login e coordinate bancarie. Un attacco di phishing prevede l'azione di un utente malevolo che tenta di camuffarsi in una fonte attendibile, ingannando un altro utente ad interagire al click di un link solitamente ricevuto via mail. La sequenza di azioni è mostrata nell'immagine seguente[3].

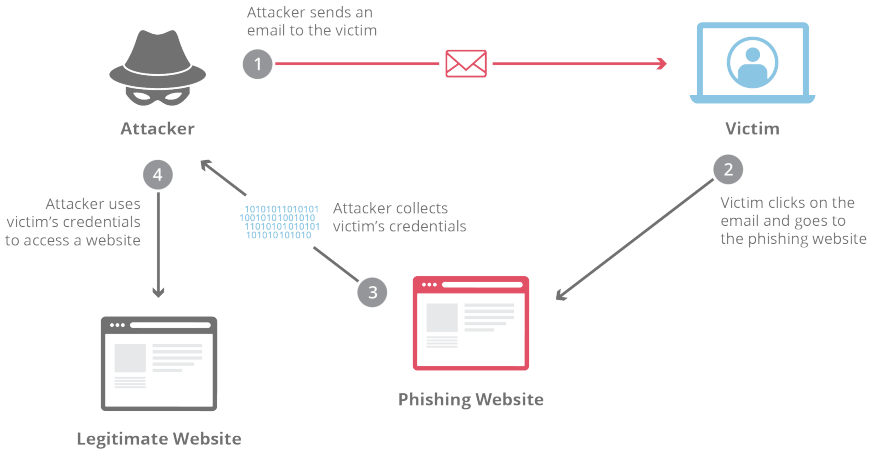

Figura 2: I passi in un attacco di phishing.

1.4 Entropia di Shannon

L'**entropia di Shannon** è una definizione matematica che permette di misurare la quantità di incertezza di una sorgente di informazione. Nel contesto dell'analisi di questo studio, l'unità di informazione è rappresentata dai caratteri che vengono combinati per formare gli url e i domini. Possiamo calcolare l'entropia E con la formula: $E = -\sum P_i * \log_2 P_i$, dove P_i indica la probabilità che ogni carattere sia presente nel dominio/url. Un'alta entropia indica che il dominio/url è composto da una distribuzione equilibrata e casuale di caratteri, trasportando una maggiore informazione. Viceversa, una bassa entropia suggerisce una struttura più prevedibile e meno variegata, con maggiore incertezza e minore informazione[4].

[3] Cloudflare - What is a phishing attack?
[4] Medium - Information Theory Series: 1. Entropy and Shannon Entropy

1.5 Il Dataset scelto

Il dataset selezionato è composto da feature che descrivono gli URL nella loro interezza. Questa scelta nasce dal fatto che, invece di basarsi su elementi che si concentrano su contenuti quali messaggi di testo, CSS, loghi o immagini, è possibile focalizzare l'attenzione sugli URL stessi, i quali possono rivelare **schemi ricorrenti** e **anomalie** potenzialmente indicative di tentativi di phishing. Alcuni esempi includono sottodomini sospetti, estensioni insolite, utilizzo eccessivo di parametri o imitazioni di domini legittimi [1].Il dataset è stato sviluppato attraverso tre fasi principali:

1. **Acquisizione dei dati:** gli URL sono stati raccolti da fonti pubbliche affidabili.
2. **Pre-elaborazione:** gli URL sono stati scomposti in componenti strutturate dalle quali è stato generato un set di feature rilevanti. L'uso dell'**Optimal Feature Vectorization Algorithm (OFVA)**, sviluppato per identificare e ottimizzare le caratteristiche intra-URL più rilevanti, ha permesso:
   - L'adozione di 31 feature già note nella letteratura.
   - L'introduzione di 10 feature innovative, pensate per catturare dettagli non tradizionalmente analizzati.
3. **Pulizia e ottimizzazione:** sono stati eliminati duplicati e outlier, garantendo robustezza e qualità.

Il dataset finale risulta composto da **247.950 istanze**, suddivise in **119.409** URL di phishing e **128.541** URL legittimi. Contiene 41 feature che descrivono aspetti relativi agli URL, ai domini, ai sottodomini e ad altre componenti come query e fragments, oltre a una **variabile target** che indica se un URL è classificato come phishing o legittimo.

Figura 3: Distribuzione percentuale degli URL nel dataset

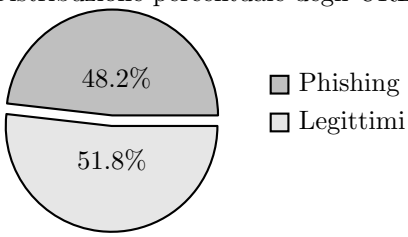

Tabella 1: Tabella che esplicita le feature del dataset [1]

FN	Feature	Descrizione	Tipo
F0	Type	Indica il tipo dell'URL, 0 rappresenta un URL legittimo e 1 rappresenta un URL di phishing.	Boolean
F1	url_length	Numero di caratteri in un URL.	Numeric
F2	number_of_dots_in_url	Numero di punti (`.`) nell'URL.	Numeric
F3	having_repeated_digits_in_url	Indica se l'URL contiene cifre ripetute (ad esempio, 2232).	Boolean
F4	number_of_digits_in_url	Numero di cifre (0–9) nell'URL.	Numeric
F5	number_of_special_char_in_url	Numero di caratteri speciali (ad esempio, `$`, `%`, `&`, `#`) nell'URL.	Numeric

Continua dalla pagina precedente			
FN	**Feature**	**Descrizione**	**Tipo**
F6	number_of_hyphens_in_url	Numero di trattini (-) nell'URL.	Numeric
F7	number_of_underline_in_url	Numero di trattini bassi (_) nell'URL.	Numeric
F8	number_of_slash_in_url	Numero di barre (/) o barre inverse (\) nell'URL.	Numeric
F9	number_of_questionmark_in_url	Numero di punti interrogativi (?) nell'URL.	Numeric
F10	number_of_equal_in_url	Numero di segni di uguale (=).	Numeric
F11	number_of_at_in_url	Numero di simboli @ nell'URL.	Numeric
F12	number_of_dollar_sign_in_url	Numero di segni di dollaro ($) nell'URL.	Numeric
F13	number_of_exclamation_in_url	Numero di punti esclamativi (!) nell'URL.	Numeric
F14	number_of_hashtag_in_url	Numero di hashtag (#) nell'URL.	Numeric
F15	number_of_percent_in_url	Numero di segni di percentuale (%) nell'URL.	Numeric
F16	domain_length	Lunghezza del nome di dominio nell'URL.	Numeric
F17	number_of_dots_in_domain	Numero di punti (.) nel nome del dominio.	Numeric
F18	number_of_hyphens_in_domain	Numero di trattini (-) nel dominio.	Numeric
F19	having_special_characters_in_domain	Indica se il dominio contiene caratteri speciali ($, %, &, #).	Boolean
F20	number_of_special_characters_in_domain	Numero di caratteri speciali nel nome del dominio.	Numeric
F21	having_digits_in_domain	Indica se il dominio contiene cifre (0–9).	Boolean
F22	number_of_digits_in_domain	Numero di cifre nel dominio.	Numeric
F23	having_repeated_digits_in_domain	Indica se il dominio contiene cifre ripetute (ad esempio, 223321).	Boolean
F24	number_of_subdomains	Indica il numero di sottodomini nell'URL.	Numeric
F25	having_dot_in_subdomain	Indica se il sottodominio contiene un punto (.).	Boolean
F26	having_hyphen_in_subdomain	Indica se il sottodominio contiene un trattino (-).	Boolean
F27	average_subdomain_length	Lunghezza media dei sottodomini nell'URL.	Continuous
F28	average_number_of_dots_in_subdomain	Numero medio di punti (.) nei sottodomini.	Continuous
F29	average_number_of_hyphens_in_subdomain	Numero medio di trattini (-) nei sottodomini.	Continuous
F30	having_special_characters_in_subdomain	Indica se il sottodominio contiene caratteri speciali ($, %, &, #).	Boolean
F31	number_of_special_characters_in_subdomain	Numero di caratteri speciali ($, %, &, #) nel sottodominio.	Numeric
F32	having_digits_in_subdomain	Indica se il sottodominio contiene cifre (0–9).	Boolean
F33	number_of_digits_in_subdomain	Numero di cifre nel sottodominio.	Numeric
F34	having_repeated_digits_in_subdomain	Indica se il sottodominio contiene cifre ripetute (ad esempio, 223342).	Boolean
F35	having_path	Indica se l'URL ha un percorso.	Boolean
F36	path_length	Lunghezza del percorso nell'URL.	Numeric
F37	having_query	Indica se l'URL ha una query.	Boolean
F38	having_fragment	Indica se l'URL ha un fragment.	Boolean
F39	having_anchor	Indica se l'URL ha un anchor.	Boolean
F40	entropy_of_url	Rappresenta l'entropia di Shannon dell'URL.	Continuous
F41	entropy_of_domain	Rappresenta l'entropia di Shannon del dominio.	Continuous

2 Statistica Descrittiva

2.1 Pre-analisi della dimensionalità

Prima di individuare le feature da analizzare, è stata realizzata una pre-analisi. In particolare, essendo che il dataset presenta un numero elevato di feature, pari a **42**, sono state effettuate varie considerazioni per scartare alcune feature e diminuire la dimensionalità dei nostri dati. Diverse feature sono state eliminate per le seguenti criticità:

- **Varianza bassa**: Quando una feature ha una varianza bassa, ha un potere informativo trascurabile;
- **Informazione ridondante**: All'interno di un dataset sono presenti varie feature, alcune di queste possono mantenere lo stesso tipo di informazione risultando ridondanti;
- **Informazione derivabile**: Un dataset può presentare delle feature che sono di supporto ad altre, cercando di sintetizzare una caratteristica della feature principale. In alcuni casi però, questa informazione aggiuntiva è facilmente derivabile e non giustifica l'aumento di dimensionalità del dataset.

Sulla base di queste criticità individuate, in maniera singola o incrociata, sono state scartate **18** feature e conservate le **24** rimanenti.

2.2 Scelta delle feature da analizzare

Una volta scartate le feature meno utili, bisogna passare alla scelta di quali feature voler analizzare. Come già descritto nella sezione **1.5**, le feature descrivono diversi aspetti di un URL. Per la nostra analisi, si è deciso di focalizzarci sulle feature che descrivono i domini (e i sottodomini). Questa scelta è dettata dalle seguenti motivazioni:

- **Importanza del dominio**: Il dominio è la parte più importante dell'URL, in quanto è la componente che definisce il mapping con l'indirizzo di rete. Questo permette all'utente di poter raggiungere la macchina che eroga il servizio direttamente tramite il nome del dominio;
- **Indipendenza del dominio dall'URL**: Il dominio è indipendente dalle altre componenti dell'URL. Prendendo ad esempio l'utilizzo dei moderni browser web, un utente può anche solamente digitare il nome del dominio per raggiungere la landing page di un servizio, interagendo successivamente con essa per recuperare le risorse necessarie;
- **Numero ristretto di feature rimaste**: **9** delle rimanenti feature sono legate al dominio. Con un numero minore di feature, l'analisi risulterà più mirata e dettagliata. Le feature sono: *domain_length, number_of_dots_in_domain, number_of_hyphens_in_domain, number_of_special_characters_in_domain, number_of_digits_in_domain, number_of_subdomains, average_subdomain_length, number_of_digits_in_subdomain, entropy_of_domain.*

2.3 Analisi Univariata

Per l'analisi univariata delle feature, le prossime sotto sezioni saranno così organizzate:

- **Rappresentazione grafiche dei dati**: Sono presentati due plot che permettono di visualizzare la distribuzione dei dati. I grafici realizzati sono:
  - **Istogramma**: Un grafico utilizzato per visualizzare la distribuzione dei dati mediante delle barre verticali che rappresentano degli intervalli numerici, dove ogni osservazione ricade in unico intervallo. Il numero di intervalli è stato definito di default mediante il metodo di Sturges ($\lceil \log_2 n + 1 \rceil$);
  - **Boxplot**: Un grafico a scatola utilizzato per individuare i quartili della distribuzione e i valori anomali che si posizionano oltre i limiti superiori e inferiori. Per una leggibilità migliore, il grafico è stato posizionato in senso orizzontale.
- **Tabella riassuntiva delle statistiche univariate**: Viene presentata una tabella che raccoglie le principali statistiche univariate;
- **Considerazioni**: Saranno aggiunte eventuali considerazioni sui grafici e sulle tabelle.

Le feature saranno analizzate sia prima che dopo la rimozione degli outlier, mostrando entrambi gli istogrammi e riportando sulle tabelle le differenze tra le statistiche. Tra le **9** feature rimanenti del dominio, saranno analizzate nello specifico **domain_length** e **entropy_of_domain**. Queste due feature saranno anche utilizzate nelle sezioni successive di questo studio. Considerando che il dataset risulta essere bilanciato e che siamo interessati alle osservazioni legate al fenomeno del phishing, è stata analizzata la relativa partizione del dataset dove la variabile target **Type** assumeva un valore pari a **1**.

2.3.1 Domain length

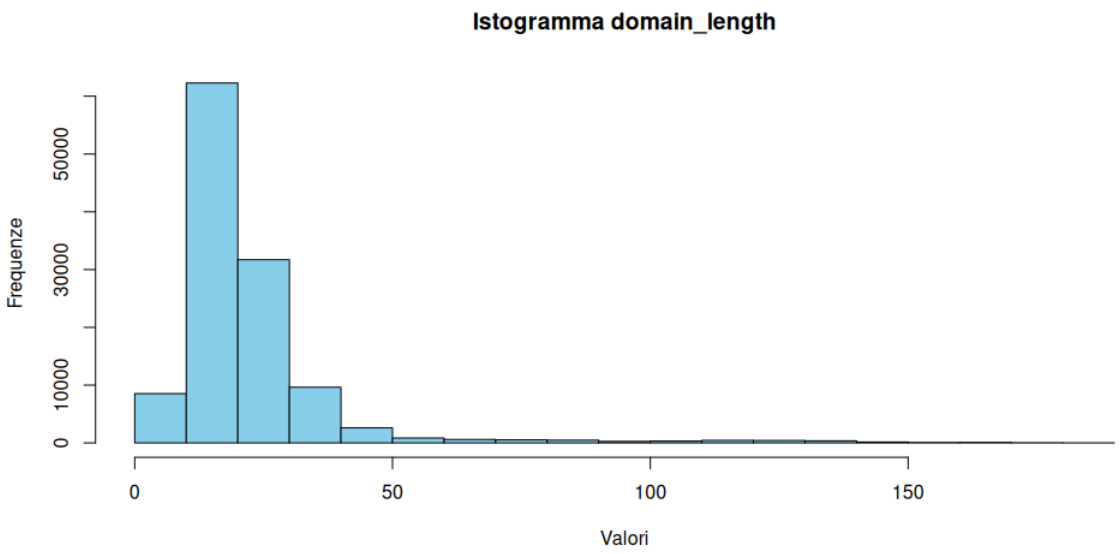

Figura 4: Istogramma di Domain Length

Questo grafico ci permette di osservare la concentrazione della distribuzione dei dati della lunghezza del dominio. Calcolando le frequenze di distribuzione, notiamo che circa il **96%** delle osservazioni sono minori uguali di **50**. La forma della distribuzione è allungata verso destra, risultante in un'**asimmetria positiva**. Nell'intervallo **(10, 20]** è presente il picco massimo della distribuzione, dove risiedono circa il **52%** delle osservazioni. Questo picco permette di definire la distribuzione come **leptocurtica**, ossia più piccata di una distribuzione normale.

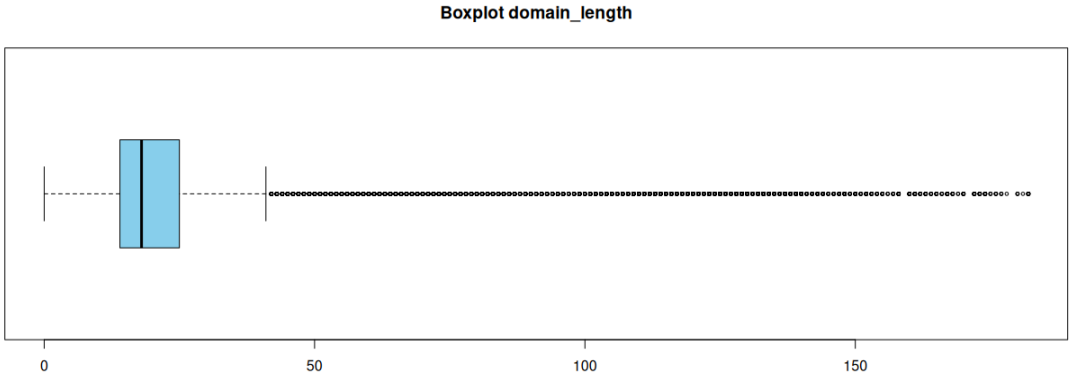

Figura 5: Boxplot di Domain Length

Questo grafico ci permette di osservare in maniera immediata la presenza di anomalie. Calcolando il limite inferiore, pari a **-2,5**, e superiore pari **41,5**, si evince come che circa il 4% delle osservazioni siano outlier.

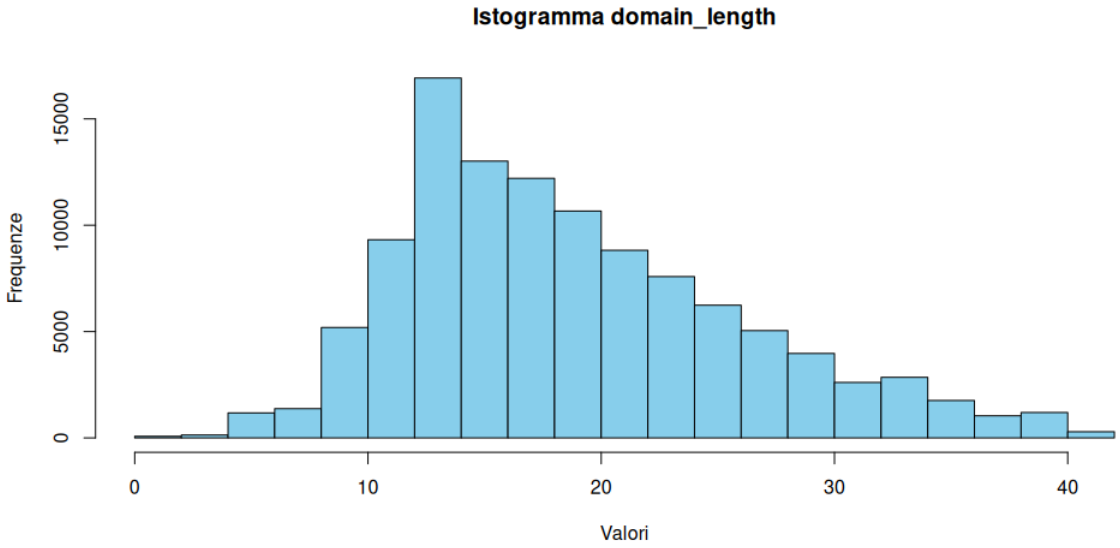

Figura 6: Istogramma di Domain Length (outlier rimossi)

Una volta rimossi gli outlier, l'istogramma risulta essere più centrato e la coda ristretta. Con la rimozione degli outlier, il **100%** delle osservazioni sono minori o uguali di **50**.

Misure di centralità			**Misure di dispersione**		
Outlier	**Con**	**Senza**	**Outlier**	**Con**	**Senza**
Media	22.611	19.231	Varianza	312.448	52.7
Mediana (Q2)	18	18	Dev. Standard	17.676	7.26
Moda	13	13	Range	182	41
Forma della distribuzione			**Summary**		
Outlier	**Con**	**Senza**	**Outlier**	**Con**	**Senza**
Skewness	4.289	0.691	Min.	0	0
Kurtosis	26.372	3.009	Q1	14	14
			Q3	25	24
			Max.	182	41

Tabella 2: Tabella statistiche Domain Length

Nella tabella sono state riassunte le principali statistiche calcolate prima e dopo la rimozione degli outlier. Possiamo notare come non ci siano stati cambiamenti per la mediana, questo evidenzia il suo essere robusta rispetto alla presenza di valori anomali. La varianza è invece crollata, a dimostrazione di come sia fortemente influenzata dagli outlier.

2.3.2 Entropy of domain

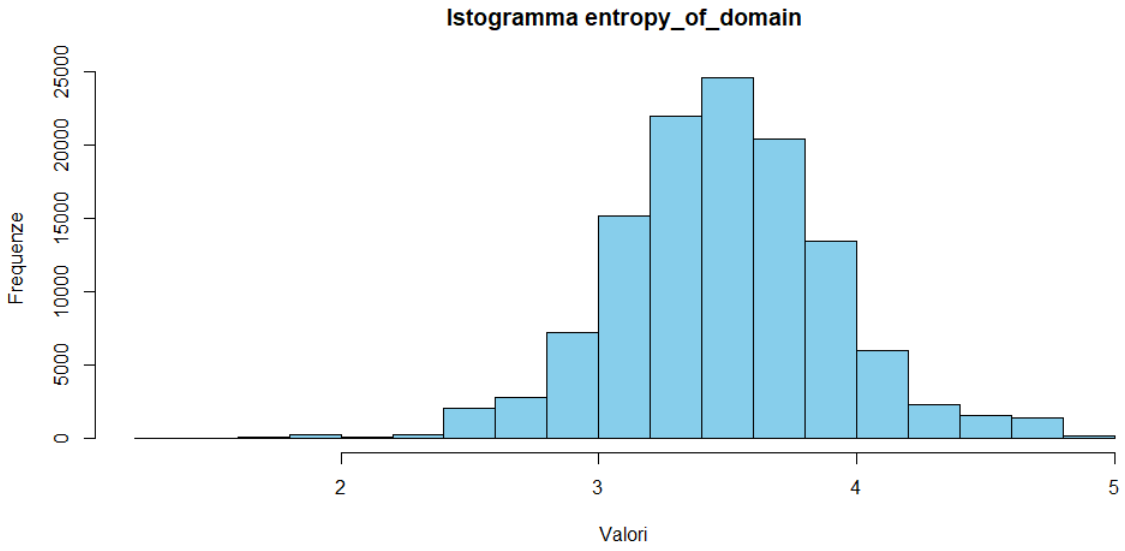

Figura 7: Istogramma di entropy of domain

L'istogramma evidenzia la concentrazione della distribuzione dei dati per l'entropia del dominio. Calcolando le frequenze di distribuzione viene evidenziato come circa il **90%** delle osservazioni totali sia minore uguale di **4**, con il picco di maggior concentrazione nell'intervallo **(3.4, 3.6]**. Il valore della skewness è leggermente superiore allo 0, come risultato si ha una distribuzione asimmetrica positiva. La distribuzione di frequenze ha una forma leptocurtica.

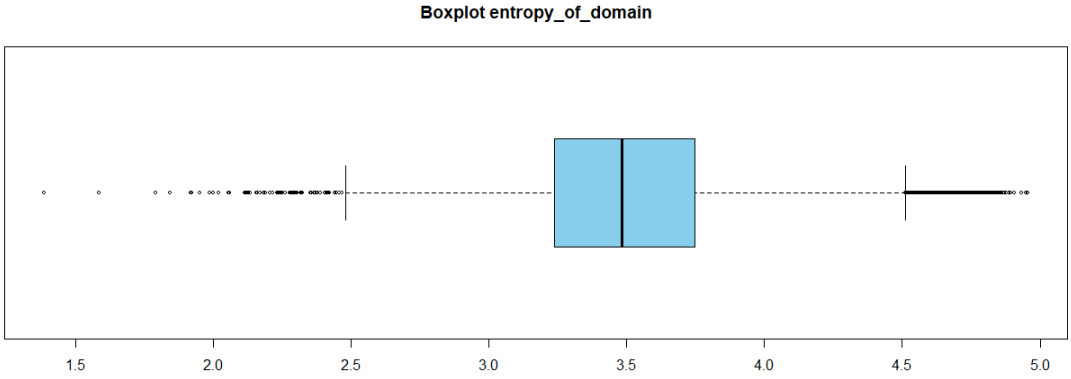

Figura 8: Boxplot di Entropy Of Domain

Il boxplot permette la visualizzazione delle anomalie presenti. Il limite inferiore è pari a **2,47**, quello superiore pari a **4,42**, escludendo dall'intervallo circa il **3%** delle osservazioni che risultano essere outlier.

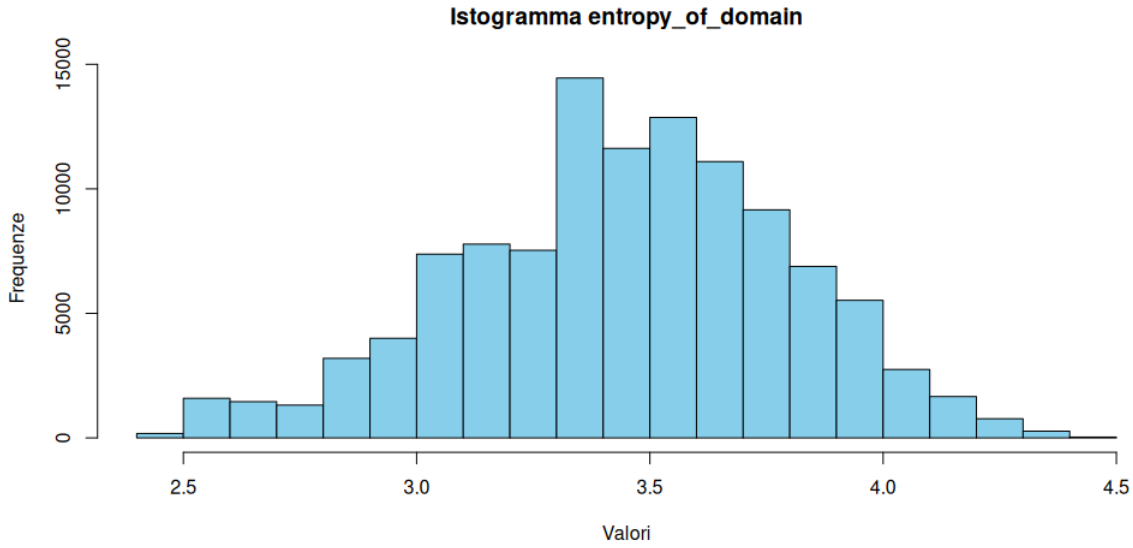

Figura 9: Istogramma di Entropy Of Domain (outlier rimossi)

Una volta rimossi gli outlier si osserva un cambiamento significativo nella skewness che scende sotto lo zero, trasformando la distribuzione in un'asimmetrica negativa. Anche il valore della curtosi scende, rendendo la distribuzione di frequenze più piatta di una normale.

Misure di centralità			**Misure di dispersione**		
Outlier	**Con**	**Senza**	**Outlier**	**Con**	**Senza**
Media	3.493	3.45	Varianza	0.173	0.123
Mediana (Q2)	3.484	3.456	Dev. Standard	0.416	0.351
Moda	3.927	3.392	Range	3.571	1.937
Forma della distribuzione			**Summary**		
Outlier	**Con**	**Senza**	**Outlier**	**Con**	**Senza**
Skewness	0.085	-0.234	Min.	1.386	2.481
Kurtosis	3.744	2.839	Q1	3.239	3.232
			Q3	3.748	3.689
			Max.	4.957	4.418

Tabella 3: Tabella statistiche Entropy of Domain

La tabella mostra le principali statistiche della feature, calcolate prima e dopo la rimozione degli outlier. La media e la mediana risultano leggermente diminuite, così come la varianza.

2.4 Analisi Bivariata

2.4.1 Analisi della correlazione

È possibile rappresentare in maniera chiara e intuitiva la correlazione tra le variabili mediante la seguente matrice. La scala di correlazione è compresa nell'intervallo **[-1, 1]** e corrisponde ai valori che il **coefficiente di correlazione campionario** può assumere. 0 indica l'assenza di correlazione tra le variabili, mentre valori maggiori o minori di 0 indicano rispettivamente una correlazione positiva o negativa.

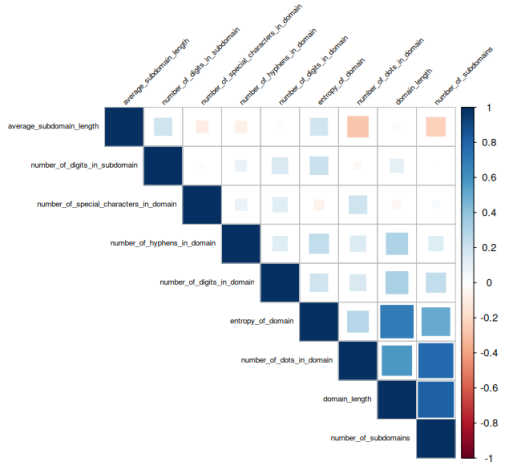

Figura 10: Matrice di correlazione delle feature del dominio pre rimozione degli outlier

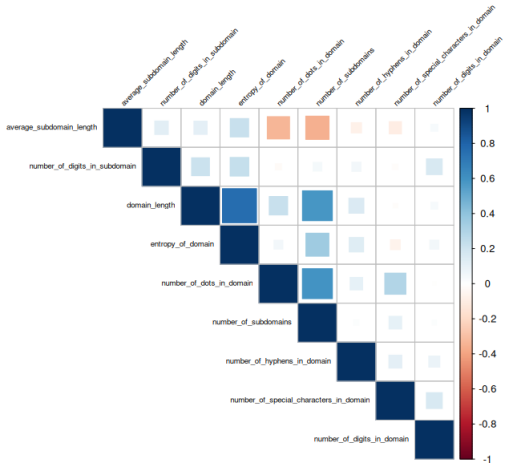

Figura 11: Matrice di correlazione delle feature del dominio dopo la rimozione degli outlier

Il grafico senza outlier mostra diverse variazioni rispetto al grafico precedente. La correlazione tra la lunghezza del dominio e l'entropia del dominio è aumentata da **0.698** a **0.766** dopo aver rimosso gli outlier, questo valore suggerisce una forte dipendenza lineare tra le due variabili e sarà per questo preso in considerazione per il modello di regressione lineare.

2.4.2 Modello di regressione lineare

La **regressione lineare** è una tecnica utilizzata per descrivere e modellare la relazione tra una variabile dipendente e una o più variabili indipendenti. Nel caso della regressione lineare semplice, il modello si basa sull'equazione di una retta che rappresenta al meglio la distribuzione dei punti nello scatterplot, minimizzando la distanza tra i punti osservati e la retta stessa rispetto a tutte le altre possibili rette.

Di seguito viene presentato un modello di regressione lineare per studiare la relazione tra l'**entropia del dominio** (variabile indipendente) e la **lunghezza del dominio** (variabile dipendente). Sono analizzati i risultati sia con che senza la presenza di outlier, al fine di comprendere l'impatto che i valori anomali hanno sul modello.

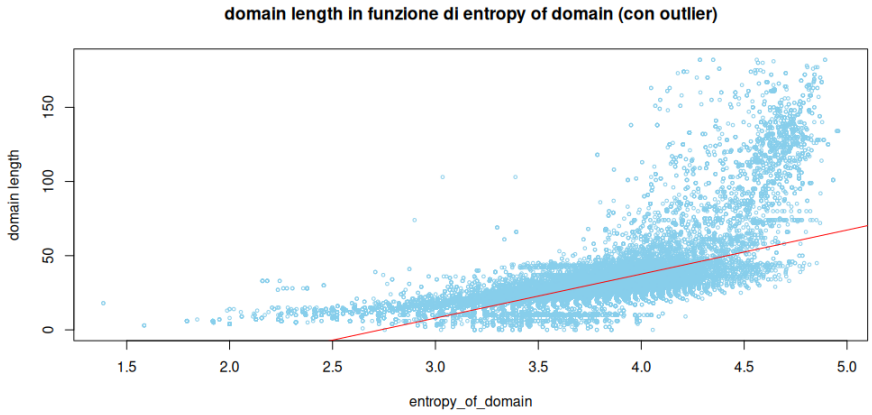

Figura 12: Regressione lineare con outlier

L'equazione della retta è $-81.063 + 29.673x$ e il coefficiente di determinazione è **0.488**.

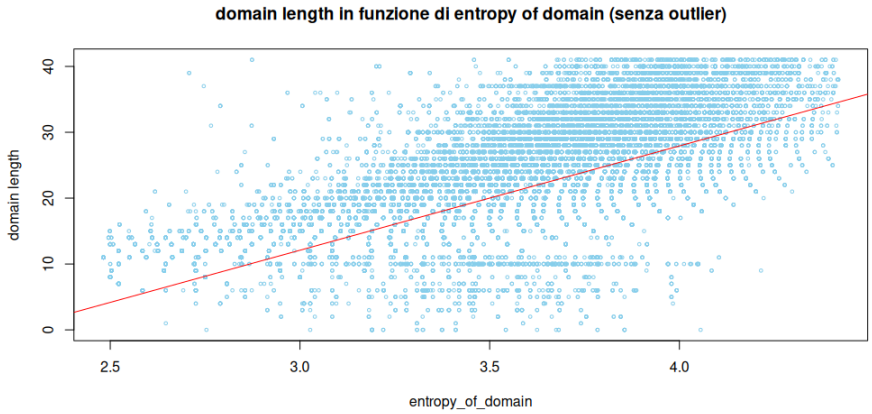

Figura 13: Regressione lineare senza outlier

L'equazione della retta è $-35.404 + 15.833x$ e il coefficiente di determinazione è **0.588**.

2.4.3 Analisi dei residui

I **residui** rappresentano la differenza tra i valori osservati nei dati reali e i valori stimati dal modello. Idealmente, dovrebbero essere distribuiti in modo casuale attorno allo zero, a dimostrazione del fatto che il modello è in grado di catturare correttamente la relazione tra le variabili considerate. Sarà di seguito effettuata un'analisi della loro distribuzione prima e dopo la rimozione degli outlier.

Min	1Q	Median	3Q	Max
-39.29	-6.61	-2.165	3.304	135.905

Tabella 4: Residui prima della rimozione degli outlier

Prima di rimuovere gli outlier, il valore negativo della mediana indica che il modello tende a sottostimare i valori rispetto ai dati reali. Il minimo e il massimo sono rispettivamente molto distanti da zero, il che segnala la presenza di outlier estremi che portano il modello a fare previsioni lontane dai valori reali per alcune delle osservazioni.

Min	1Q	Median	3Q	Max
-28.814	-3.159	-0.137	2.719	31.529

Tabella 5: Residui dopo la rimozione degli outlier

Dopo aver rimosso gli outlier, la distribuzione appare più simmetrica e la mediana risulta quasi prossima allo zero, a dimostrazione del fatto che il modello è migliorato e non tende più a sottostimare i dati. Anche il range si restringe notevolmente, a indicare che le previsioni del modello sono meno soggette a errori estremi.

2.4.4 Correlation does not imply causation

Con la rimozione degli outlier abbiamo potuto constatare un aumento del coefficiente di determinazione, a dimostrazione della forte influenza che i valori estremi esercitano su di esso. La spiegabilità del modello rispetto ai dati è ancora migliorabile oltre l'attuale soglia del circa **60%**, valore che suggerisce la presenza di ulteriori fattori non ancora considerati che potrebbero contribuire a spiegare la variabilità dei dati. In primo luogo potrebbe essere utile considerare variabili aggiuntive per realizzare una regressione lineare multipla. In questo contesto, è utile ricordare che una correlazione non è sufficiente a stabilire una relazione causale tra l'aumento della lunghezza del dominio rispetto all'entropia. Altre variabili con una bassa correlazione potrebbero contribuire indirettamente all'aumento della lunghezza del dominio. Ad esempio:

- **Number of subdomains**: Quando la lunghezza del dominio aumenta, potrebbe voler dire che sono presenti anche dei sottodomini;
- **Number of dots in domain**: Se quindi sono presenti più sottodomini in un dominio, sono presenti anche più dots per effettuare la separazione tra le label.

3 Statistica Inferenziale

3.1 Criterio del chi quadrato

Il criterio del chi quadrato permette di stabilire se un campione è considerabile estratto da una popolazione descritta da una variabile aleatoria X con funzione di distribuzione $F_x(x)$.

Nel nostro caso, si vuole determinare una popolazione normale a partire da un sample randomico della partizione Phishing del dataset, pari all'1% delle osservazioni (1194), considerando come variabile l'entropia del dominio. La popolazione di riferimento è formata da tutti gli URL segnalati sulle piattaforme anti-phishing ***PhishTank*** e ***OpenPhish***.

Affinché si possa sfruttare questo criterio, bisogna definire delle ipotesi:

- H_0 - **Ipotesi Nulla**: X ha una funzione di distribuzione normale $N(\mu, \sigma^2)$.
- H_1 - **Ipotesi Alternativa**: X non ha una funzione di distribuzione normale.

Per poter accettare una delle ipotesi, bisogna condurre un test bilaterale. Un test bilaterale prevede che il risultato ottenuto dal criterio del chi-quadrato, possa ricadere in un intervallo delimitato da due quantili, che determinano l'ampiezza della regione di accettazione. Viene inoltre fissata la probabilità di errore di tipo I, ovvero di rifiutare erroneamente l'ipotesi nulla, pari a $\alpha = 0.05$. I parametri già fissati prevedono:

- **La probabilità rimanente**: A partire dalla p_i per cui i valori del campione ricadano in uno dei $r = 5$ intervalli, determinati mediante i quantili della normale, la rimanente probabilità può essere univocamente determinata;
- **Parametri della distribuzione**: Per la distribuzione normale sono previsti due parametri non noti, sostituiti dalle stime.

Ottenendo così 2 gradi di libertà, possiamo calcolare:

- **Limite inferiore della regione di accettazione**: 0.050635;
- **Limite superiore della regione di accettazione**: 7.377759;
- **Valore del chi-quadrato**: 6.251256.

Il chi-quadrato ricade nella regione di accettazione, possiamo quindi accettare l'ipotesi nulla e affermare che il campione è stato estratto da una popolazione descritta da una variabile aleatoria normale.

3.2 Stima puntuale

La stima è il processo mediante il quale si cerca di determinare il valore di un parametro incognito di una popolazione basandosi sui dati di un campione. Con la **stima puntuale** intendiamo fornire un singolo valore come approssimazione del parametro incognito.

I parametri incogniti ϑ che intendiamo stimare sono la media e la varianza campionaria. Utilizzando il metodo dei momenti, considerando un campione rappresentativo di una popolazione normale, si vuole determinare la stima del parametro μ mediante la media campionaria del campione. La stima del parametro σ^2 avviene mediante l'utilizzo della variabile aleatoria $\frac{n-1*S^2}{n}$ dove n è il numero di osservazioni del campione e S^2 la varianza campionaria.

Le stime puntuali ottenute sono le seguenti:

- **Stima del parametro** μ: 3.501435;
- **Stima del parametro** σ^2: 0.178302.

3.3 Stima intervallare

Spesso, al singolo valore, si preferisce indicare un **intervallo** di valori nel quale la stima è contenuta con un certo intervallo di confidenza. Attraverso il campione è possibile definire una stima dell'intervallo di confidenza per un parametro della popolazione, nel nostro caso, la media. L'intervallo viene individuato mediante il **metodo pivotale**, utilizzando una variabile pivot che dipende dal campione e dal parametro non noto da stimare, ma la cui funzione di distribuzione non contiene il parametro.

Nel nostro caso, per stimare il valore medio con varianza incognita è possibile utilizzare la variabile aleatoria di Student: $\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \sim T(n-1)$, dove $\overline{X}$ corrisponde alla media campionaria, μ corrisponde al parametro della media della popolazione da stimare, S corrisponde alla radice quadrata della varianza campionaria e n rappresenta il numero di osservazioni del campione. Avendo fissato il valore della media campionaria, i gradi di libertà sono $n-1$.

Conoscendo i valori della media campionaria pari a 3.501435 e della radice quadrata della varianza campionaria pari a 0.4224352, fissando il grado di confidenza pari a $1-\alpha = 0.99$, otteniamo che la stima del valore medio dell'entropia del dominio per la popolazione, ricade nell'intervallo **(3.469894; 3.532976)**.

4 Generazione sintetica dei dati

4.1 LLM e Prompt Engineering

Per la generazione dei dati è stato utilizzato un LLM, in particolare il modello *Qwen2.5-Plus*, accessibile dalla sua ***Web-UI***. La conversazione effettuata è disponibile nella sua interezza al seguente ***link***. Il prompt è stato gradualmente raffinato fino a ottenere la versione utilizzata nella conversazione citata, formata da più parti e nelle quali si è cercato di utilizzare diversi **Prompt Pattern** noti in letteratura: [2]

1. **Assegnazione del ruolo**: Nella prima parte del prompt è stato definito il ruolo dell'LLM, fornendo alcune indicazioni su come trattare i vincoli definiti nella parte successiva. Per questa parte del prompt, si fa utilizzo del **Persona Pattern**;

2. **Definizione del contesto**: Nella seconda parte del prompt, è stato definito il contesto dove sono state indicate le principali proprietà statistiche delle varie feature del sample di dati da generare;

3. **Esempi di righe del dataset**: Nella terza parte del prompt, sono stati forniti degli esempi delle righe del dataset, seguendo la tecnica del **Few-Shot Prompting**;

4. **Indicazioni finali**: Nell'ultima parte del prompt, sono state fornite le indicazioni finali, che ci hanno assicurato di ottenere direttamente delle righe sample evitando la generazione di script. Per questa parte del prompt, si fa in parte utilizzo dell'**Infinite Generation Pattern**. Per rispettare la finestra di token di output è stato richiesto di fornire le righe sample in batch di dati. Inoltre, per cercare di evitare il più possibile i duplicati, viene esplicitamente richiesto di cercare di evitarli all'interno dello stesso batch e tra i diversi batch.

Una volta completata la generazione e ottenute le righe richieste, è stato effettuato un ulteriore controllo sui duplicati. A partire dalle **1940** righe del sample, sono stati individuati e scartati **376** duplicati. Si ottengono in totale **1565** righe univoche, dalle quali è stato estratto un sample randomico di righe pari all'1% della partizione Phishing del dataset reale. Il prompt utilizzato viene mostrato nella pagina successiva.

```
Act as an AI Assistant that helps generate synthetic data with the same structure that I am
providing you next. Between each of your new answers, don't worry about the data count and
just focus to respect the constraints.

The following is useful context:
"Dataset of suspicious phishing url detection is a collection of different observations
about legitimate and phishing url features. In particular, focusing on phishing, there are
119.409 instances. This dataset has many features and there are 9 important features about
the domain:
- domain_lenght: A numerical discrete feature with a mean of 22.611, median of 18 and a
variance of 312.448. The values of this feature ranges from 0 to 182.
- number_of_dots_in_domain: A numerical discrete feature with a mean of 3.306, median of 3
and a variance of 4.893. The values of this feature ranges from 0 to 28.
- number_of_hyphens_in_domain: A numerical discrete feature with a mean of 0.676, median of
0 and a variance of 1.480. The values of this feature ranges from 0 to 23.
- number_of_special_characters_in_domain: A numerical discrete feature with a mean of 0.565,
median of 0 and a variance of 2.019. The values of this feature ranges from 0 to 47.
- number_of_digits_in_domain: A numerical discrete feature with a mean of 7.450, median of 1
and a variance of 162.586. The values of this feature ranges from 0 to 144.
- number_of_subdomains: A numerical discrete feature with a mean of 2.192, median of 2 and a
variance of 3.241. The values of this feature ranges from 0 to 27.
- average_subdomain_length: A numerical continuous feature with a mean of 7.617, median of 6
and a variance of 31.570. The values of this feature ranges from 0 to 110.
- number_of_digits_in_subdomain: A numerical discrete feature with a mean of 0.428, median
of 0 and a variance of 3.216. The values of this feature ranges from 0 to 44.
- entropy_of_domain: A numerical continuous feature with a mean of 3.493, median of 3.484
and a variance of 0.173. The values of this feature ranges from 1.386 to 4.957.

It's also important to consider that the correlation between domain_length and
entropy_of_domain is of 0.698."

The following are some examples of dataset rows in a csv format:
domain_length,number_of_dots_in_domain,number_of_hyphens_in_domain,
number_of_special_characters_in_domain,number_of_digits_in_domain,
number_of_subdomains,average_subdomain_length,number_of_digits_in_subdomain,
entropy_of_domain
11,2,1,0,0,1,7,0,2.913977073
26,5,0,0,0,2,3,0,3.532573258
42,5,0,2,11,5,3,0,4.220429781

Given these informations, from now on, I want you to generate me a synthetic data sample
without using any external script but giving back an output in a table format or csv format,
until you reach 1940 rows. You can also generate the data sample step by step using batches.
Please try your best to avoid duplicates between the different batches and inside the same
batch.
```

Figura 14: Prompt usato per la generazione sintetica dei dati

4.2 Research Question

4.2.1 RQ1

Research Question 4.1. I dati sintetici generati dal Large Language Model mantengono le proprietà statistiche descritte nel prompt?

Lo scopo di questa RQ è quello di stabilire se l'LLM sia stato in grado di soddisfare le richieste espresse nel prompt, il che ci permette anche di scoprire se ci sia similarità tra i dati reali e i dati generati in termini di statistiche. È importante tener conto del fatto che la quantità di dati generati corrisponda solo all'**1%** del dataset preso in considerazione, percentuale che costituisce una prima possibile limitazione in termini di efficacia dei risultati.

Viene mostrata una tabella riassuntiva per evidenziare le differenze delle caratteristiche dei dati rispetto a quelle che erano state richieste dal prompt.

domain_length			
	Richiesta	**Generata**	**Differenza**
Media	22.611	37.265	-14.654
Mediana	18	35	-17
Varianza	312.448	202.644	109.844
Min	0	10	-10
Max	182	85	97
entropy_of_domain			
	Richiesta	**Generata**	**Differenza**
Media	3.493	3.384	0.109
Mediana	3.484	3.431	0.053
Varianza	3.216	0.184	3.032
Min	1.386	2.890	-1.504
Max	4.957	4.120	0.837

Tabella 6: Tabella per confronto caratteristiche dei dati

La tabella ci permette di evidenziare come i dati generati, seppur senza esplicite richieste, si siano concentrati nel ridurre la dispersione dei dati dalla media e la presenza di valori anomali. La differenza tra i dati generati e quelli richiesti potrebbe essere spiegata da una forma di bias nel processo generativo, in cui il modello tende a privilegiare valori centrali e limita la generazione di valori estremi, come si nota dai valori massimi e minimi più stretti rispetto ai dati richiesti.

Di seguito viene riproposta l'analisi delle due variabili prese dettagliatamente in esame nella sezione **2.3**, si andranno quindi a visualizzare graficamente le distribuzioni e i valori anomali delle due feature.

Domain Length

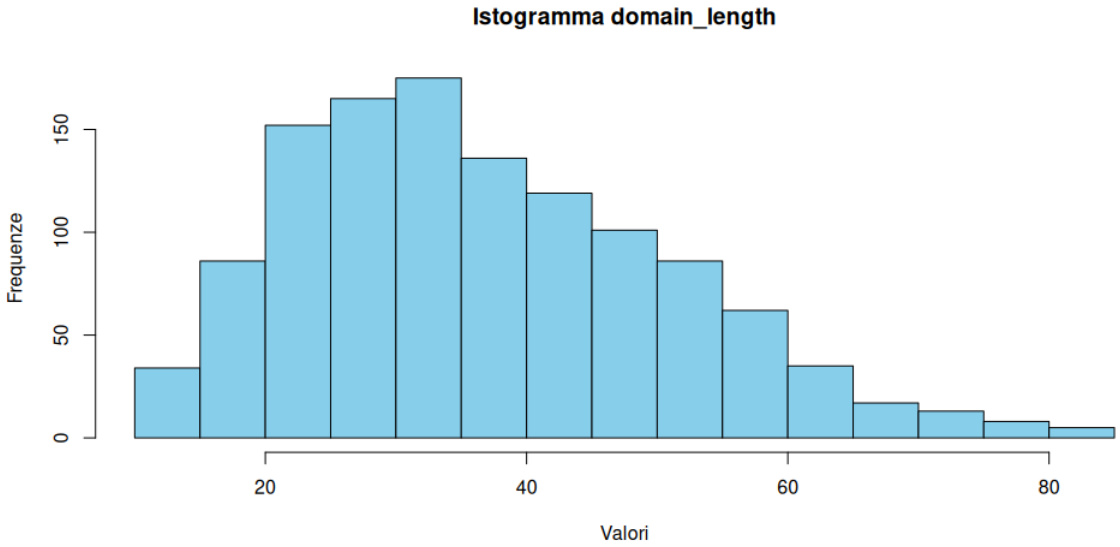

Figura 15: Istogramma di Domain Length (dati sintetici)

L'istogramma mostra come la distribuzione dei dati sia asimmetrica e allungata verso destra, con una forma che risulta più piatta di una normale. Calcolando le frequenze di distribuzione, notiamo che circa il **93%** delle osservazioni sono minori uguali di **60**.

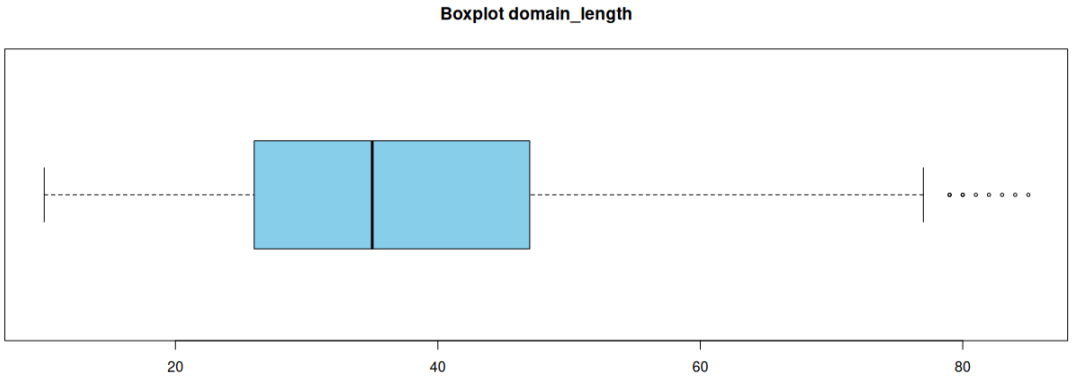

Figura 16: Boxplot di Domain Length (dati sintetici)

Il boxplot ci permette di visualizzare le osservazioni anomale che ricadono oltre il limite superiore fissato a **77**, che costituiscono circa il **6%** delle osservazioni totali.

Entropy of domain

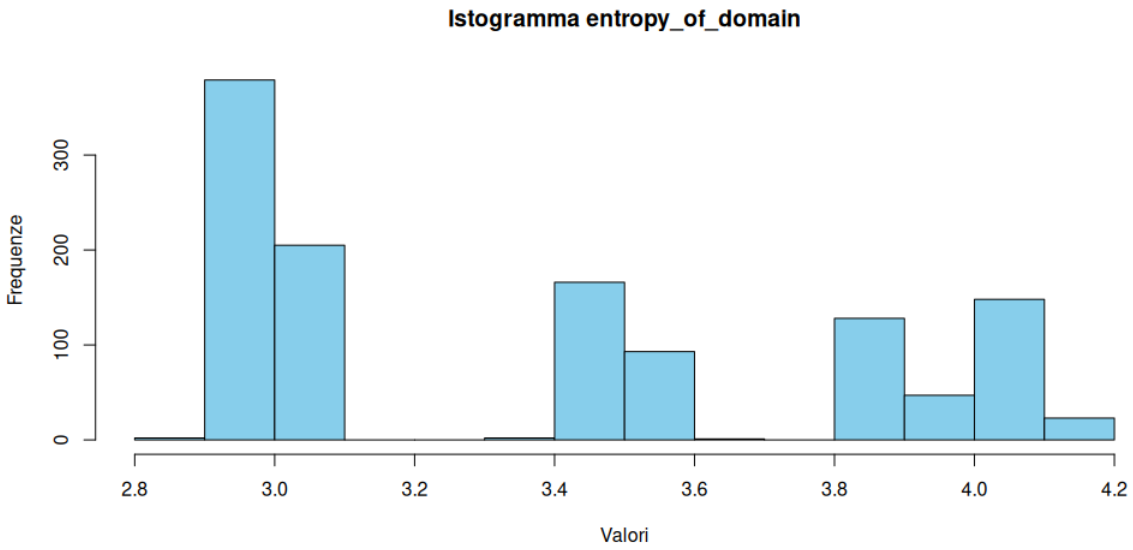

Figura 17: Istogramma di Entropy of Domain (dati sintetici)

In questo caso l'istogramma mostra dei salti della distribuzione dei dati. In particolare negli intervalli **(3.1, 3.2]**, **(3.2, 3.3]**, **(3.7,3.8]** non ricade nessuna osservazione. Il picco massimo della distribuzione è nell'intervallo **(2.9, 3.0]**, dove risiede circa il **31%** delle osservazioni totali. Una spiegazione plausibile per i tre picchi distinti evidenziati potrebbe essere data dagli esempi di righe che sono stati forniti al LLM. Nel prompt sono stati proposti tre valori di entropia distanti da loro (2.91, 3.53, 4.22), e sebbene sia stato fornito un valore di range, il LLM sembra essersi concentrato nel proporre risultati che non si sono distanziati molto dai singoli dati indicati, lasciando dei vuoti nella distribuzione.

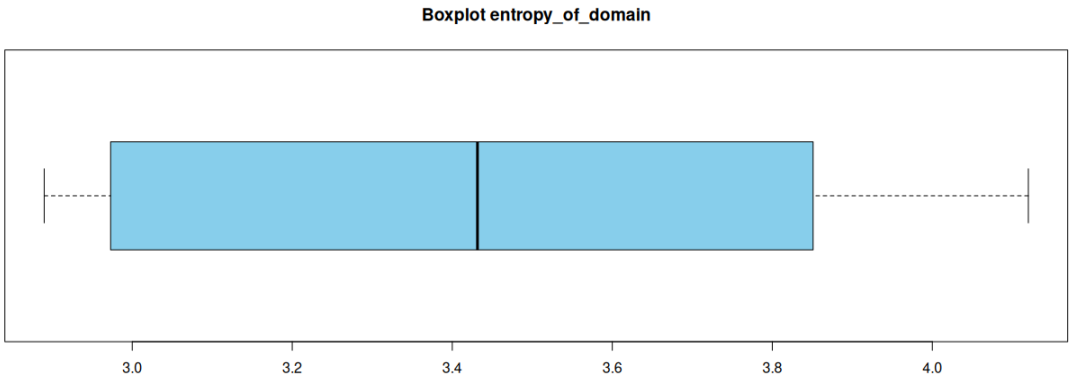

Figura 18: Boxplot di Entropy of Domain (dati sintetici)

Il boxplot per la feature analizzata non mostra alcun valore anomalo.

4.2.2 RQ2

Research Question 4.2. Quali sono le differenze principali tra dati sintetici generati e dati reali in contesti di regressione?

Prima di considerare un modello di regressione, bisogna verificare che anche nei dati sintetici ci sia una correlazione tra le feature di domain_length e di entropy_of_domain.

Il coefficiente di correlazione calcolato utilizzando i dati sintetici risulta pari a **0.654**, e in questo caso la rimozione degli outlier fa diminuire tale coefficiente a **0.648**. Pertanto, il nuovo modello di regressione lineare che si ottiene dal sample sintetico ha un coefficiente di determinazione di **0.427**. Questo valore evidenzia come il sample non sia sufficiente a costruire un modello con un'alta spiegabilità dei dati.

4.2.3 RQ3

Research Question 4.3. La feature d'interesse, tra quelle generate dall'LLM, può essere ricondotta a una distribuzione normale?

Avendo verificato con il criterio del chi quadrato che la feature dell'entropia del dominio rispetto al sample di 1% del dataset reale è estraibile da una popolazione descritta da una variabile aleatoria normale, si è proceduto col ripetere lo stesso test anche per il sample sintetico.

In questo caso, il valore del chi-quadrato ottenuto è pari a **286.9213**. Conoscendo già i limiti della regione accettazione, possiamo direttamente affermare che il sample sintetico **non** supera il test e che quindi non sia estraibile da una popolazione descritta da una variabile aleatoria normale. Non potendo ricondurre un'uguale distribuzione tra la popolazione reale e quella sintetica, non è possibile effettuare un confronto tra le popolazioni.

5 Conclusioni

Giunti alla conclusione di questo studio, è possibile effettuare alcune considerazioni:

- **Dimensione del sample sintetico**: Il sample generato è pari solamente all'**1%** della partizione di phishing del dataset reale. In futuro si potrebbe tentare di generare un sample più grande e di ripetere le varie analisi, in modo da cercare di ottenere risultati migliori;
- **Migliorie al prompt**: Il fallimento del test del chi-quadrato e la bassa spiegabilità del modello di regressione lineare suggeriscono che il processo di generazione dei dati sintetici necessiti di miglioramenti significativi per garantire una maggiore aderenza alle proprietà statistiche del dataset reale. Per generare nuovi sample, potrebbe essere utile migliorare il prompt da fornire al LLM. Alcuni aspetti modificabili potrebbero essere relativi all'inserimento di informazioni aggiuntive e più dettagliate riguardo le caratteristiche dei dati reali, così come al fornire maggiori esempi di istanze di dati reali per tentare di colmare i vuoti evidenziati nelle distribuzioni;
- **Nuovo modello di regressione**: A partire dalla partizione phishing del dataset reale e rispetto alla feature di domain_length, l'unica variabile che manteneva con essa un'alta correlazione positiva, sia prima che dopo la rimozione degli outlier, era quella di entropy_of_domain. Si potrebbe quindi esplorare l'idea di coinvolgere altre variabili per cercare di creare un nuovo modello di regressione, valutando eventualmente l'adozione di modelli più complessi e non necessariamente lineari.

Sulla base di queste considerazioni, possiamo ritenerci soddisfatti del lavoro svolto e di aver applicato diversi concetti affrontati durante il corso.

Riferimenti bibliografici

[1] M. A. Tamal, M. K. Islam, T. Bhuiyan, and A. Sattar, "Dataset of suspicious phishing url detection," *Frontiers in Computer Science*, vol. 6, p. 1308634, 2024.

[2] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt, "A prompt pattern catalog to enhance prompt engineering with chatgpt," *arXiv preprint arXiv:2302.11382*, 2023.