



UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento di Informatica

STATISTICA E ANALISI DEI DATI

Phishing Legitimate Analisys

DOCENTI

Prof. Stefano Cirillo

Prof. Luigi Di Biasi

AUTORI

Leopoldo Todisco

Matricola: 0522501795

Carlo Venditto

Matricola: 0522501796

Anno Accademico 2024-2025

Indice

1	Introduzione	1
1.1	Obiettivi del progetto	2
1.2	Descrizione del Dataset	3
1.2.1	Struttura e caratteristiche generali	3
1.2.2	Tipologia dei dati e intervalli di valore	4
1.2.3	La variabile target	5
1.2.4	Variabili principali e loro significato	5
2	Distribuzioni di frequenza	8
2.1	Iistogrammi	9
2.1.1	Distribuzione delle classi	9
2.1.2	Distribuzione di PctExtNullSelf	10
2.1.3	FrequentDomainNameMismatch	11
2.1.4	Frequenza per NumDash	12
2.1.5	Frequenza per SubmitInfo	13
2.1.6	Frequenza per PctNullSelf	14
2.2	Boxplot	15
2.2.1	Boxplot Multipli	15
2.2.2	Osservazioni dai Boxplot Multipli	16
2.3	Conclusioni	16

2.4	Kernel Density Plots	16
2.4.1	Kernel plot per NumDash	17
2.4.2	Kernel plot per UrlLength	19
2.4.3	Kernel plot per NumQueryParams	21
3	Statistica Descrittiva Univariata	24
3.1	Funzione di Distribuzione Empirica	24
3.2	Funzione di Distribuzione Empirica Continua	25
3.2.1	Distribuzione Empirica Continua - NumDash	25
3.2.2	Distribuzione Empirica Continua - NumQueryComponents	26
3.3	Indici di Sintesi	28
3.3.1	Misure di Centralità	28
3.3.2	Misure di Dispersione	32
3.3.3	Misure di Simmetria	36
4	Statistica descrittiva Bivariata	40
4.1	Correlazione fra le variabili	40
4.1.1	Covarianza Campionaria	40
4.1.2	Coefficiente di Correlazione Campionario	41
4.2	Analisi della Relazione tra le Feature e la Variabile Target	42
4.2.1	Covarianza e Correlazione	42
4.2.2	Interpretazione dei Risultati	42
4.2.3	Visualizzazione con Scatterplot	43
4.2.4	Analisi della Matrice di Correlazione	46
4.2.5	Interpretazione della Matrice di Correlazione	46
5	Creazione di un Modello predittivo	48
5.1	Regressione Logistica	48
5.1.1	Risultati della Regressione Logistica	49
5.1.2	Interpretazione dei Risultati	49
5.1.3	Metriche del Modello	50
5.1.4	Matrice di Confusione e Accuratezza	50
5.1.5	Analisi della Learning Curve	51

5.2 Interpretazione dei Risultati con SHAP	52
5.2.1 Valori SHAP per una Predizione Specifica	52
5.2.2 Grafico dei Valori SHAP	52
5.2.3 Conclusioni sulla Regressione Logistica	52
6 Generazione e Analisi di un Dataset Sintetico	54
6.1 Metodologia di Generazione	54
6.2 Obiettivi della Generazione Sintetica	58
6.3 RQ1: Misure di centralità del dataset sintetico	59
6.3.1 Scostamento delle Misure di Centralità	59
6.3.2 Conclusioni	60
6.4 RQ2: Metriche di Dispersione	61
6.4.1 Confronto Numerico	61
6.4.2 Grafico di confronto	62
6.5 RQ3: Differenza nelle matrici di correlazione	63
6.5.1 Confronto delle Correlazioni	64
6.5.2 Osservazioni Critiche	65
6.5.3 Conclusioni e Soluzioni	66
6.6 RQ4: Analisi di Distribuzioni note	66
6.6.1 Obiettivo e Metodologia	66
6.6.2 Risultati	67
6.6.3 Conclusioni	67
Bibliografia	69

CAPITOLO 1

Introduzione

Nell'era della digitalizzazione, la sicurezza informatica rappresenta una sfida cruciale per istituzioni, aziende e utenti finali. Uno degli attacchi più diffusi e insidiosi in questo ambito è il phishing, una tecnica di ingegneria sociale finalizzata a sottrarre informazioni sensibili sfruttando la fiducia degli utenti. Contrastare efficacemente tali minacce richiede non solo una solida infrastruttura tecnologica, ma anche l'impiego di metodologie avanzate di analisi dei dati e machine learning.

Il presente progetto si propone di analizzare e modellare un dataset relativo al phishing, utilizzando tecniche di analisi statistica e machine learning. In particolare, il dataset preso in esame, "Phishing Dataset for Machine Learning" disponibile su Kaggle, contiene informazioni strutturali e contenutistiche sugli URL, fornendo un'opportunità concreta per studiare le caratteristiche distintive di un sito fraudolento.

La scelta di concentrarsi sul problema del phishing non è casuale: in un contesto in cui le truffe online sono in costante evoluzione, migliorare la capacità di identificare tempestivamente attività fraudolente può avere un impatto significativo sulla sicurezza digitale globale. Inoltre, questo progetto rappresenta un'applicazione pratica delle competenze acquisite nel corso di Statistica e Analisi dei Dati, integrando metodi teorici con strumenti computazionali moderni.

1.1 Obiettivi del progetto

Il lavoro si articola in diverse fasi, ciascuna finalizzata all'estrazione di conoscenza utile dal dataset e alla costruzione di un modello predittivo efficace. Gli obiettivi principali includono:

- **Esplorazione del dataset:** Analisi della struttura e delle caratteristiche delle variabili, individuazione di pattern, distribuzioni e anomalie.
- **Analisi descrittiva e inferenziale:** Studio delle relazioni tra le feature e la variabile target, valutazione della significatività statistica di ciascuna caratteristica nel predire la presenza di phishing.
- **Sviluppo di un modello predittivo:** Implementazione e confronto di algoritmi di machine learning per la classificazione degli URL come legittimi o di phishing, con valutazione delle prestazioni attraverso metriche standard.

L'analisi dei dati e la costruzione del modello contribuiranno a comprendere quali variabili giocano un ruolo chiave nel rilevamento delle attività fraudolente e a sviluppare strumenti più efficienti per la prevenzione degli attacchi di phishing.

Nel prosieguo del report verranno presentate in dettaglio le fasi di:

- **Analisi del dataset:** Descrizione del dataset e valutazione delle correlazioni tra le variabili e la variabile target.
- **Distribuzioni di frequenza:** Studio delle frequenze assolute, relative e percentuali delle principali feature.
- **Statistica descrittiva univariata:** Calcolo di misure di centralità, dispersione e simmetria per le variabili numeriche e binarie.
- **Statistica descrittiva bivariata:** Analisi della covarianza e del coefficiente di correlazione tra le variabili selezionate e la variabile target.
- **Creazione del modello predittivo:** Preparazione dei dati, addestramento del modello di regressione logistica, validazione e valutazione delle sue performance su un dataset di test e su un dataset sconosciuto.

- **Generazione di un dataset sintetico:** Applicazione delle stesse analisi statistiche sul dataset generato, al fine di confrontare i risultati ottenuti e valutare la coerenza del modello.

1.2 Descrizione del Dataset

Il dataset utilizzato per questo studio è il **Phishing Dataset for Machine Learning**[1], disponibile su Kaggle¹. Questo dataset è stato progettato per supportare la ricerca nel campo della sicurezza informatica, fornendo un insieme di caratteristiche estratte dagli URL al fine di identificare siti di phishing e distinguerli da quelli legittimi. Grazie alla varietà di feature incluse, esso rappresenta una risorsa preziosa per l'analisi statistica e per lo sviluppo di modelli di apprendimento automatico in grado di rilevare minacce informatiche in modo automatico ed efficace.

1.2.1 Struttura e caratteristiche generali

Il dataset è composto da **10.000 osservazioni**, ognuna delle quali rappresenta un URL analizzato. Le informazioni sono organizzate in **50 colonne**, che descrivono le diverse caratteristiche di ciascun URL, tra cui aspetti strutturali, elementi contenutistici e informazioni sulla sicurezza. Alcune di queste variabili assumono valori numerici discreti, altre sono di tipo binario e indicano la presenza o l'assenza di una determinata caratteristica.

Le variabili presenti nel dataset coprono un'ampia gamma di informazioni. Tra queste, vi sono indicatori legati alla struttura dell'URL, come la lunghezza complessiva e il numero di sottodomini, e variabili relative al contenuto della pagina, come l'uso di iframe o la presenza di riferimenti a risorse esterne. Inoltre, il dataset include una variabile target, fondamentale per la classificazione, che identifica se l'URL è legittimo o rappresenta un tentativo di phishing.

¹<https://www.kaggle.com/datasets/amj464/phishing>

1.2.2 Tipologia dei dati e intervalli di valore

Le colonne del dataset presentano esclusivamente variabili di tipo numerico. Tuttavia, la maggior parte di esse sono di natura discreta, con alcune eccezioni che assumono valori continui su un intervallo definito.

Variabili come la lunghezza dell'URL o il numero di caratteri speciali presenti assumono valori numerici con un intervallo ben definito. Ad esempio, la lunghezza degli URL varia da un minimo di **12 caratteri** a un massimo di **253 caratteri**, mentre il numero di punti all'interno di un dominio può andare da **1** fino a **21**.

Alcune variabili invece sono binarie e assumono valori pari a **0** o **1**, indicando la presenza o l'assenza di una determinata caratteristica.

Vi sono poi alcune variabili continue, ovvero:

- **UrlLength**, che è l'unica variabile completamente continua nel dataset.
- **PctNullSelfRedirectHyperlinks** e **PctExtHyperlinks**, che assumono valori compresi tra **0** e **1**.
- **NumDots**, che nonostante possa sembrare discreta, si distribuisce su un intervallo continuo.
- **NumQueryComponents**, che rappresenta il numero di componenti nella query di un URL e assume valori continui.
- **PathLevel**, che pur rappresentando il livello del percorso nell'URL, è una variabile continua.
- **NumUnderscore**, che indica il numero di underscore presenti nell'URL e si distribuisce su un intervallo continuo.
- **NumPercent**, che rappresenta il numero di caratteri percentuali nell'URL ed è anch'essa una variabile continua.
- **NumAmpersand**;
- **NumHash**;
- **NumNumericChars**;

1.2.3 La variabile target

Un elemento cruciale del dataset è la variabile target, denominata **CLASS_LABEL**, che identifica la natura di ciascun URL. Essa assume due valori distinti:

- **0**: Indica un URL legittimo.
- **1**: Indica un URL di phishing.

Questa variabile è essenziale per la costruzione e la valutazione di modelli predittivi, poiché consente di addestrare algoritmi di machine learning a distinguere tra siti affidabili e siti malevoli. Nel dataset, gli URL classificati come phishing risultano più numerosi rispetto a quelli legittimi, rendendo il problema della classificazione particolarmente interessante e potenzialmente soggetto a sbilanciamento delle classi.

1.2.4 Variabili principali e loro significato

All'interno del dataset sono presenti numerose feature, ma alcune di esse risultano particolarmente rilevanti. Tra queste troviamo:

- **UrlLength**: La lunghezza dell'URL. URL molto lunghi sono spesso utilizzati nei tentativi di phishing per nascondere il vero dominio all'utente.
- **NumDots**: Il numero di punti all'interno dell'URL. Un elevato numero di punti può indicare la presenza di sottodomini sospetti.
- **SubdomainLevel**: Il numero di sottodomini nell'URL. I siti di phishing tendono a utilizzare sottodomini multipli per imitare domini legittimi.
- **NumDash**: Il numero di trattini (“-”) nell'URL. I siti fraudolenti spesso incorporano trattini per simulare indirizzi affidabili.
- **AtSymbol**: La presenza del carattere “@” nell'URL. Questo simbolo è frequentemente utilizzato nei tentativi di phishing per reindirizzare gli utenti verso siti malevoli.
- **IframeOrFrame**: L'uso di elementi `iframe` o `frame` all'interno della pagina. I siti di phishing spesso li sfruttano per incorporare contenuti da altre fonti, nascondendo la vera origine della pagina.

- **PctExtResourceUrlsRT:** La percentuale di risorse esterne rispetto alle risorse totali caricate nella pagina. Un'alta dipendenza da contenuti esterni può essere un segnale di phishing.
- **ExtMetaScriptLinkRT:** La presenza di metadati e script provenienti da fonti esterne. Questa caratteristica può indicare che il sito sta caricando codice da server sospetti.
- **NumQueryComponents:** Indica il numero di componenti nella parte dedicata alla query dell'URL

Di seguito viene riportata una tabella dove si mostrano tutte le correlazioni delle variabili

Feature	Correlation
PctExtNullSelfRedirectHyperlinksRT	0.5405
FrequentDomainNameMismatch	0.4640
NumDash	0.3722
SubmitInfoToEmail	0.3576
PctNullSelfRedirectHyperlinks	0.3428
InsecureForms	0.3164
NumDots	0.2941
PctExtHyperlinks	0.2597
NumSensitiveWords	0.2552
IframeOrFrame	0.2352
PathLevel	0.2295
AbnormalExtFormActionR	0.1858
UrlLengthRT	0.1695
HostnameLength	0.1692
NumDashInHostname	0.1504
NumQueryComponents	0.1474
AbnormalFormAction	0.1451
EmbeddedBrandName	0.1418

Tabella 1.1: Correlazione delle feature con CLASS_LABEL

CAPITOLO 2

Distribuzioni di frequenza

Un primo passo nell'analisi del nostro dataset consiste nella **distribuzione di frequenza**, che ci consentirà di ottenere un quadro generale sulle principali caratteristiche presenti nei dati, attraverso lo sfruttamento di tabelle e grafici.

Nel dataset si possono trovare features quantitative che possono assumere valori continui, come la lunghezza dell'URL o il numero di parametri per le query, ma anche features qualitative che assumono valori discreti come l'utilizzo del protocollo Htts. Per questo motivo, per effettuare le analisi preliminari con i grafici, è stato necessario suddividere in intervalli detti **classi** i dati continui.

Quando ci occupiamo del calcolo della frequenza, è fondamentale distinguere tra frequenza assoluta e frequenza relativa. La frequenza assoluta rappresenta il numero di occorrenze di un dato all'interno del nostro campione. Questo significa che la somma di tutte le frequenze assolute sarà sempre uguale alla dimensione del campione, fatta eccezione per i casi in cui vi siano dati mancanti. D'altra parte, la frequenza relativa viene calcolata dividendo la frequenza assoluta per la dimensione totale del campione. La somma di tutte le frequenze relative sarà sempre uguale a 1, tranne nei casi in cui vi siano dati mancanti. Per visualizzare entrambe le frequenze, utilizzeremo grafici a barre. Questi grafici forniranno una chiara rappresentazione delle distribuzioni dei dati e delle proporzioni relative, consentendo una comprensione

visiva immediata delle tendenze nel nostro dataset.

2.1 Istogrammi

2.1.1 Distribuzione delle classi

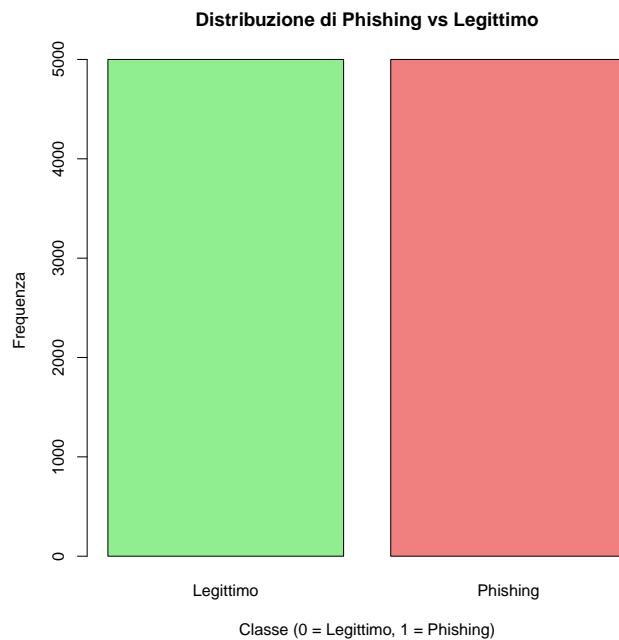


Figura 2.1: Distribuzione di frequenza assoluta delle classi Phishing e Leggittimo

- A questo punto dello studio emerge che il dataset è ben bilanciato e ritroviamo 5000 pagine web leggittime e 5000 pagine phishing

2.1.2 Distribuzione di PctExtNullSelf

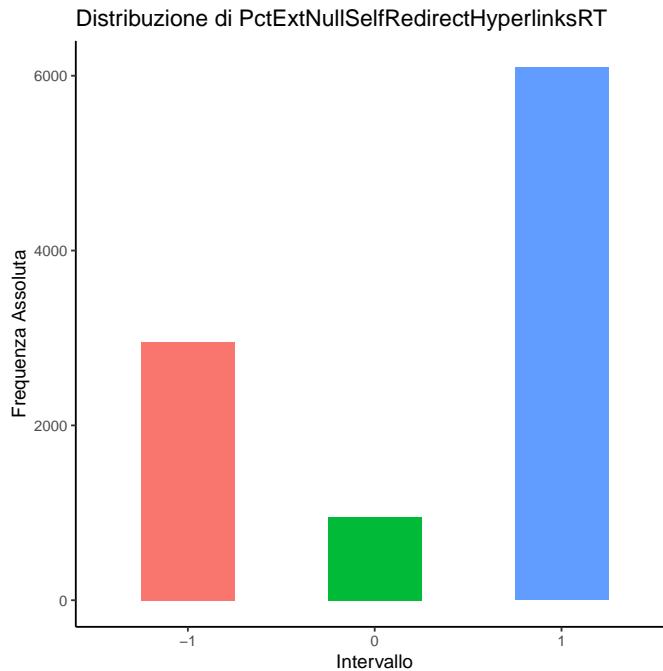


Figura 2.2: Distribuzione di frequenza assoluta di PCT

Valore	Frequenza Assoluta	Frequenza Relativa
-1	2953	0.2953
0	953	0.0953
1	6094	0.6094

Tabella 2.1: Distribuzione della variabile PctExtNullSelfRedirectHyperlinksRT

La variabile **PctExtNullSelfRedirectHyperlinksRT** mostra una distribuzione fortemente sbilanciata verso il valore 1, che rappresenta il 60,94% del totale, mentre gli altri valori hanno una presenza significativamente inferiore. Il valore -1 è presente nel 29,53% dei casi, mentre lo 0 solo nel 9,53%. Questo suggerisce che la maggior parte delle pagine analizzate presentano un'alta percentuale di link esterni nulli o autoreferenziali, caratteristica comune nei siti di phishing. Tale sbilanciamento potrebbe influenzare un modello di classificazione, portandolo a prediligere la classe più frequente e riducendo la sensibilità agli altri intervalli.

2.1.3 FrequentDomainNameMismatch

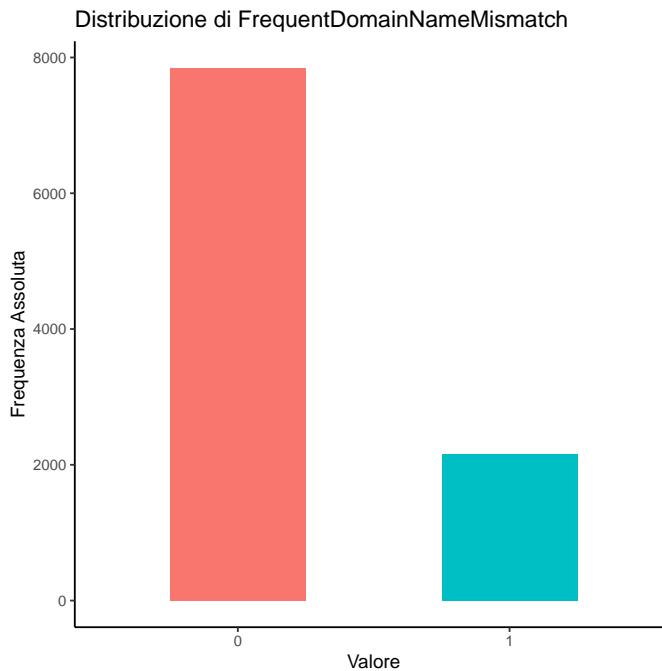


Figura 2.3: Distribuzione di frequenza assoluta di FDNM

Valore	Frequenza Assoluta	Frequenza Relativa
0	7847	0.7847
1	2153	0.2153

Tabella 2.2: Distribuzione della variabile FrequentDomainNameMismatch

La variabile **FrequentDomainNameMismatch** rappresenta la presenza di una discordanza frequente tra il nome di dominio dichiarato e il dominio effettivo utilizzato nel sito web. Questo può essere un indicatore di phishing, poiché i siti fraudolenti spesso manipolano i nomi di dominio per ingannare gli utenti.

L'analisi mostra che la distribuzione è sbilanciata, con il valore 0 che rappresenta il 78,47% del totale, mentre il valore 1 si verifica nel 21,53% dei casi. Questo suggerisce che la maggior parte dei siti analizzati non presenta una discordanza frequente nel dominio.

2.1.4 Frequenza per NumDash

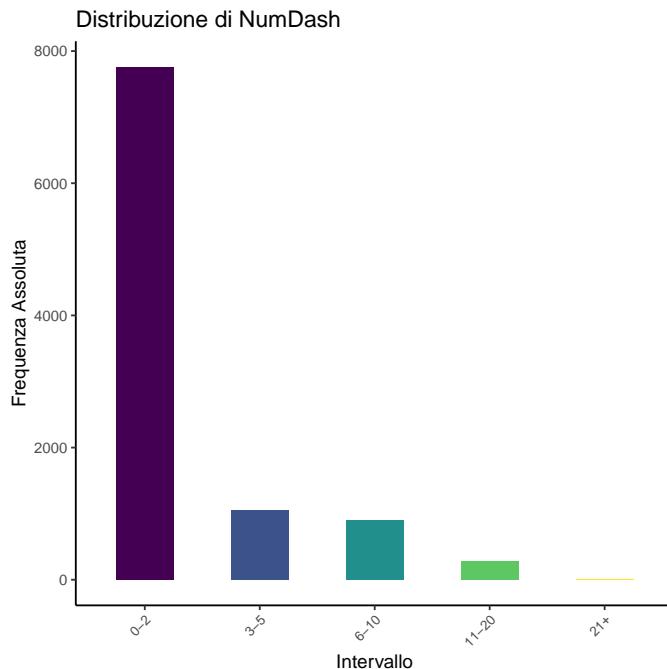


Figura 2.4: Distribuzione di frequenza assoluta di NumDash

Intervallo	Frequenza Assoluta	Frequenza Relativa
0-2	7761	0.7761
3-5	1048	0.1048
6-10	906	0.0906
11-20	277	0.0277
21+	8	0.0008

Tabella 2.3: Distribuzione della variabile NumDash

L’analisi della distribuzione mostra che la maggior parte degli URL ha tra 0 e 2 dash (77,61% del totale), mentre percentuali più basse si distribuiscono sugli altri intervalli. In particolare, gli URL con più di 10 dash sono rari (meno dello 0,1% del totale). Questo suggerisce che il numero di trattini può essere un indicatore utile per distinguere tra siti legittimi e potenzialmente dannosi.

2.1.5 Frequenza per SubmitInfo

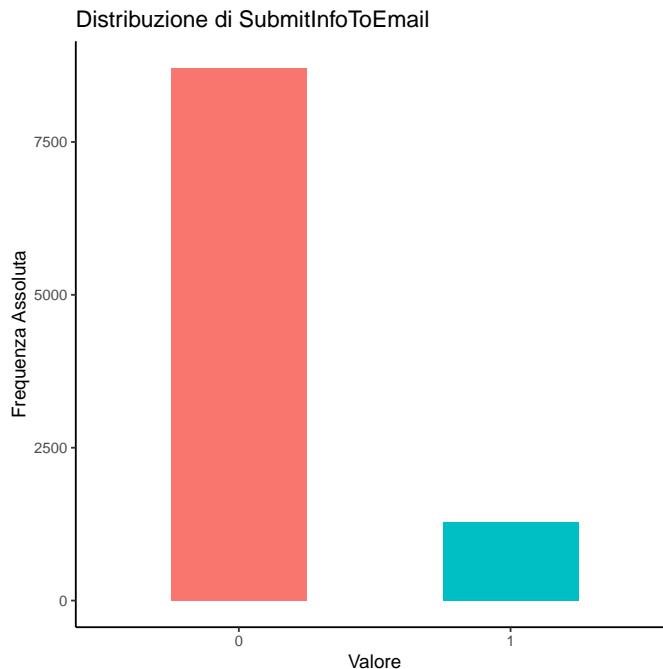


Figura 2.5: Distribuzione di frequenza assoluta di SubmitInfo

Valore	Frequenza Assoluta	Frequenza Relativa
0 (No)	8712	0.8712
1 (Sì)	1288	0.1288

Tabella 2.4: Distribuzione della variabile SubmitInfoToEmail

La variabile **SubmitInfoToEmail** indica se le informazioni di input di un modulo web vengono inviate via email anziché attraverso un’azione di backend più sicura. Questo è spesso un indicatore di phishing, poiché i siti fraudolenti utilizzano questa tecnica per raccogliere dati sensibili dagli utenti senza la necessità di un’infrastruttura server avanzata.

L’analisi mostra che la maggior parte dei siti (87,12%) non utilizza questa tecnica, mentre il 12,88% dei siti analizzati invia informazioni via email. Anche se questa percentuale è inferiore rispetto alla classe dominante, la presenza di moduli di input che inviano dati via email è comunque un segnale di potenziale rischio di phishing.

2.1.6 Frequenza per PctNullSelf

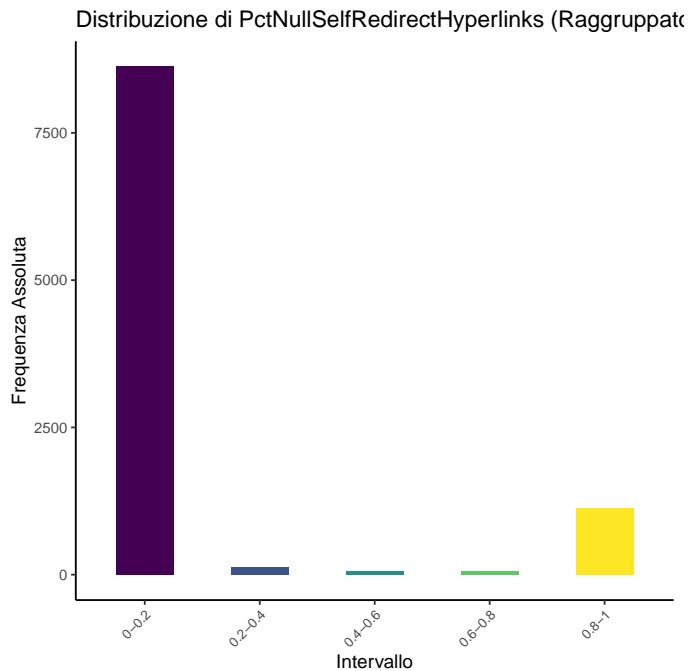


Figura 2.6: Distribuzione di frequenza assoluta di PCTNULL

Intervallo	Frequenza Assoluta	Frequenza Relativa
0-0.2	8625	0.8625
0.2-0.4	125	0.0125
0.4-0.6	62	0.0062
0.6-0.8	63	0.0063
0.8-1	1125	0.1125

Tabella 2.5: Distribuzione della variabile PctNullSelfRedirectHyperlinks

La variabile **PctNullSelfRedirectHyperlinks** rappresenta la percentuale di collegamenti ipertestuali autoreferenziali o nulli presenti in una pagina web. Questa caratteristica può essere un indicatore di pagine poco affidabili, poiché un numero elevato di tali link può indicare una struttura sospetta o una progettazione mirata al phishing.

L'analisi della distribuzione mostra che la maggior parte delle pagine analizzate (86,25%) ha una percentuale di link nulli o autoreferenziali compresa tra 0 e 0,2, mentre solo il 1,25% rientra nell'intervallo 0,2-0,4. Gli intervalli superiori hanno una presenza ancora più ridotta, con solo l'11,25% delle pagine che supera il valore di 0,8. Questo suggerisce che, sebbene la maggior parte delle pagine presenti una bassa percentuale di tali link, un piccolo sottoinsieme potrebbe indicare pagine web potenzialmente sospette.

2.2 Boxplot

In questa sezione, vengono presentati vari boxplot per visualizzare la distribuzione delle variabili chiave nel dataset, distinguendo tra siti **legittimi** e **phishing**.

2.2.1 Boxplot Multipli

Per una visione complessiva, sono stati creati boxplot multipli che confrontano contemporaneamente diverse variabili chiave.

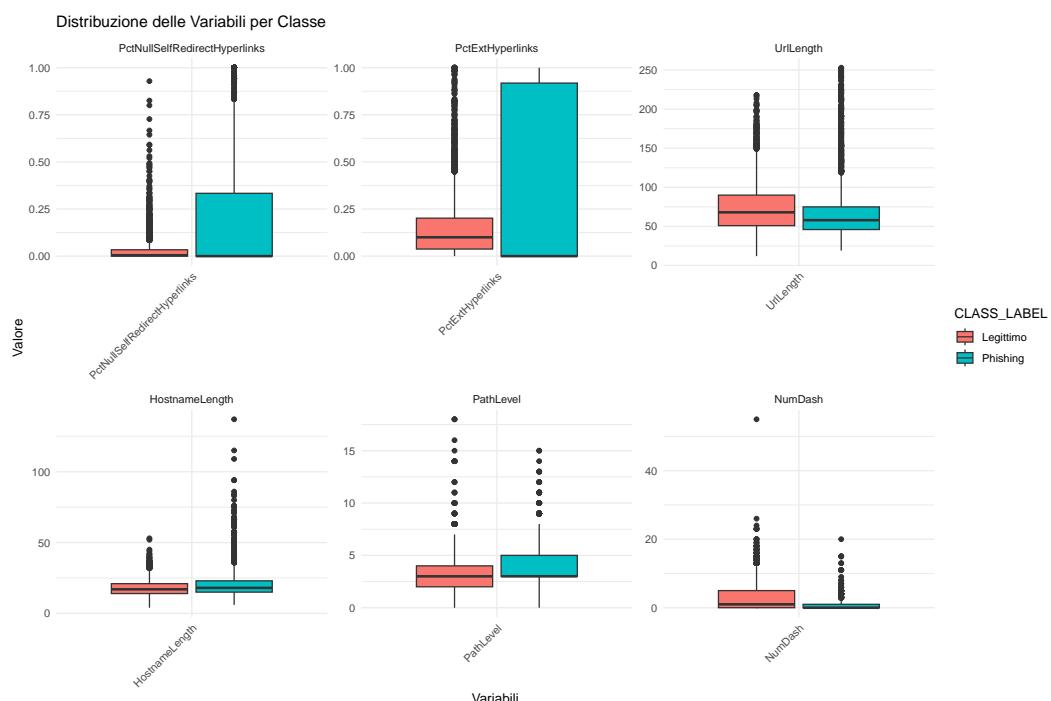


Figura 2.7: Boxplot multipli delle variabili per classe.

2.2.2 Osservazioni dai Boxplot Multipli

- **Differenze Marcate:** Variabili come `PctNullSelfRedirectHyperlinks` e `UrlLength` mostrano differenze evidenti tra le classi.
- **Distribuzioni Sovraposte:** Alcune variabili, come `PctExtHyperlinks`, mostrano distribuzioni più sovrapposte, suggerendo una minore importanza discriminante.
- **Outlier:** Sono stati identificati diversi outlier, in particolare nelle variabili `NumDash` e `UrlLength`.

2.3 Conclusioni

L’analisi tramite boxplot ha evidenziato differenze significative tra siti legittimi e phishing in diverse variabili chiave forniscono ulteriori spunti per migliorare la rilevazione di siti malevoli.

2.4 Kernel Density Plots

Il **Kernel Density Plot** è una tecnica grafica utilizzata per visualizzare la distribuzione di una variabile numerica continua. Si tratta di un metodo alternativo all’istogramma, che consente di rappresentare in modo più fluido la distribuzione dei dati senza essere influenzati dalla scelta del numero di classi (bin).

A differenza dell’istogramma, che suddivide i dati in intervalli discreti e rappresenta la frequenza di ciascun intervallo tramite barre, il Kernel Density Plot utilizza una funzione di densità per generare una curva continua che stima la distribuzione dei dati. Questa tecnica permette di identificare con maggiore precisione la presenza di picchi (modi), la forma generale della distribuzione e l’eventuale presenza di asimmetrie.

La costruzione del Kernel Density Plot si basa su due elementi fondamentali:

- **Il Kernel,** una funzione matematica utilizzata per stimare la densità locale dei dati.

- **La Bandwidth**, un parametro che controlla il grado di lisciatura della curva: un valore piccolo rende la stima più dettagliata ma soggetta a oscillazioni, mentre un valore grande produce una curva più smussata ma potrebbe nascondere dettagli importanti.

Il Kernel Density Plot è particolarmente utile per l’analisi di distribuzioni non normali, per confrontare più distribuzioni in un unico grafico e per evidenziare caratteristiche dei dati che un istogramma potrebbe non rivelare. Nelle sezioni seguenti verranno illustrate le proprietà principali di questa tecnica e il suo utilizzo nel contesto dei dati analizzati.

2.4.1 Kernel plot per NumDash

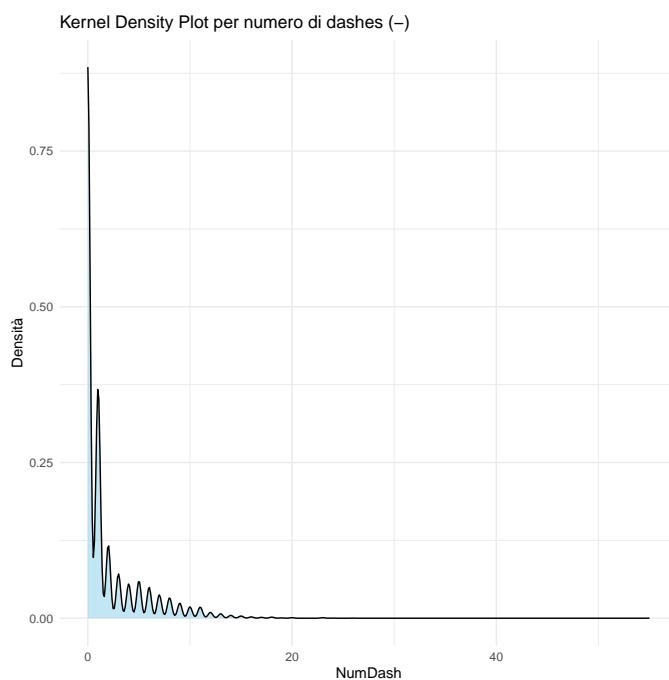


Figura 2.8: Distribuzione di numDashes

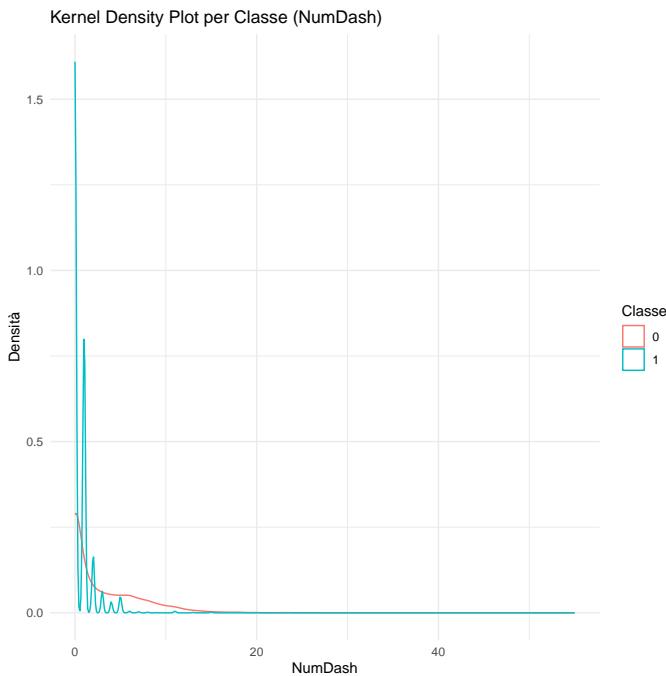


Figura 2.9: Distribuzione di numDashes rispetto alla classe

Il primo Kernel Density Plot (Figura 2.8) rappresenta la distribuzione del numero di trattini ("") negli URL senza distinzione tra siti legittimi e di phishing. Sull'asse orizzontale è riportato il numero di dashes, mentre sull'asse verticale è rappresentata la densità stimata. L'analisi della curva mostra che la distribuzione presenta un picco nella parte bassa dell'asse x , indicando che la maggior parte degli URL contiene pochi trattini. Tuttavia, la curva evidenzia anche una coda estesa verso destra, suggerendo la presenza di un numero non trascurabile di URL con un elevato numero di dashes. Questa tendenza potrebbe indicare che alcuni siti web utilizzano una quantità significativa di trattini nei loro URL, aspetto che merita un approfondimento nelle analisi successive.

Il secondo Kernel Density Plot (Figura 2.9) confronta la distribuzione del numero di trattini negli URL distinguendo tra siti legittimi (0) e siti di phishing (1). Anche in questo caso, l'asse x rappresenta il numero di dashes, mentre l'asse y mostra la densità stimata.

L'analisi del grafico suggerisce che le due classi presentano pattern differenti nella distribuzione del numero di trattini. Se le curve mostrano picchi distinti o aree di sovrapposizione ridotte, ciò potrebbe indicare che il numero di dashes è una variabile potenzialmente utile per distinguere tra URL legittimi e malevoli. Nello

specifico, se gli URL appartenenti alla classe phishing tendono ad avere un numero più elevato di trattini rispetto agli URL legittimi, questo potrebbe suggerire una strategia utilizzata dai creatori di siti malevoli per costruire domini ingannevoli o manipolare il riconoscimento degli URL.

2.4.2 Kernel plot per UrlLength

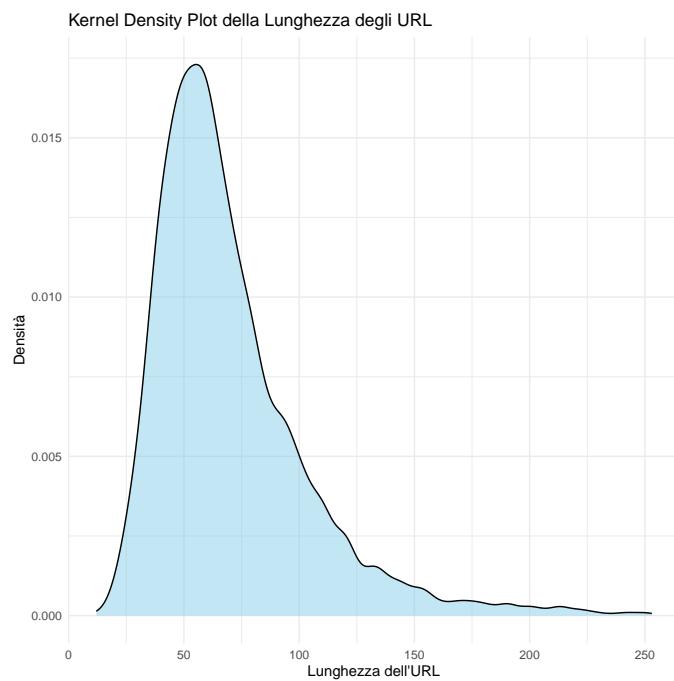


Figura 2.10: Distribuzione di UrlLength

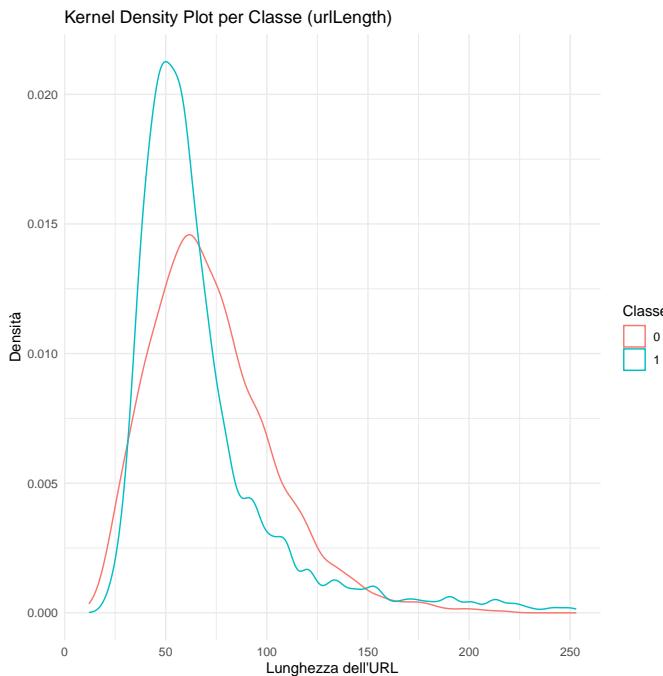


Figura 2.11: Distribuzione di UrlLength rispetto alla classe

I Kernel Density Plots presentati in questa sezione mostrano la distribuzione della lunghezza degli URL all'interno del dataset.

Il primo Kernel Density Plot (2.10) rappresenta la distribuzione della lunghezza degli URL senza distinguere tra classi. L'asse orizzontale mostra la lunghezza degli URL, mentre l'asse verticale rappresenta la densità stimata. La curva presenta una distribuzione unimodale, con un picco principale tra valori relativamente bassi e una coda estesa verso destra. Questo indica che la maggior parte degli URL nel dataset ha una lunghezza contenuta, mentre un numero inferiore di URL è significativamente più lungo. La presenza di questa coda suggerisce una distribuzione asimmetrica positiva, con alcuni URL molto lunghi che influenzano la media complessiva.

Il secondo Kernel Density Plot (2.11) suddivide la distribuzione in base alla classe dell'URL (siti legittimi vs siti di phishing). Anche in questo caso, la lunghezza dell'URL è riportata sull'asse orizzontale, mentre la densità è rappresentata sull'asse verticale. La distinzione tra classi permette di osservare eventuali differenze nella distribuzione della lunghezza tra i due gruppi.

L'analisi del grafico evidenzia che le distribuzioni della lunghezza degli URL per le due classi presentano differenze sottili, con variazioni nella concentrazione dei valori. In particolare, se le curve mostrano picchi distinti o aree di sovrapposizione,

ciò suggerisce che la lunghezza dell'URL potrebbe essere un fattore discriminante tra siti legittimi (0) e siti di phishing (1). Nello specifico, se gli URL più lunghi risultano più frequenti nella classe di phishing, ciò potrebbe indicare che i siti malevoli tendono ad avere strutture URL più complesse rispetto ai siti legittimi.

2.4.3 Kernel plot per NumQueryParams

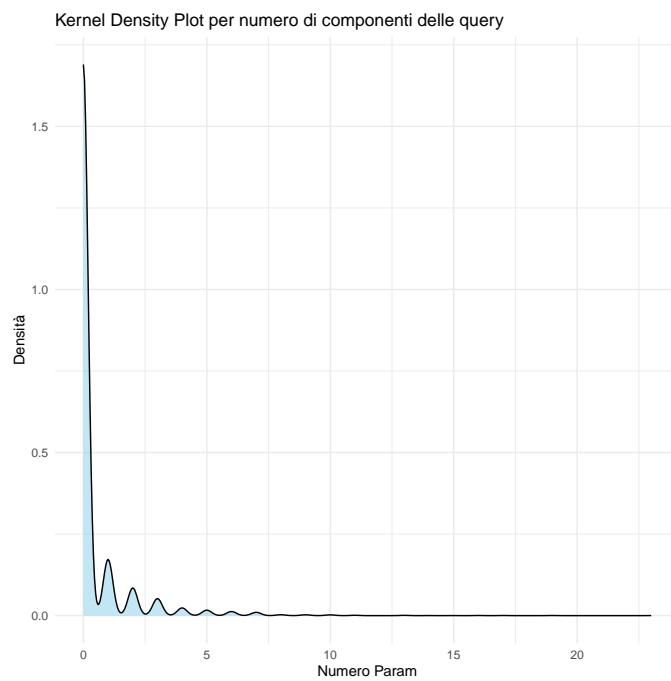


Figura 2.12: Distribuzione di NumQueryParams

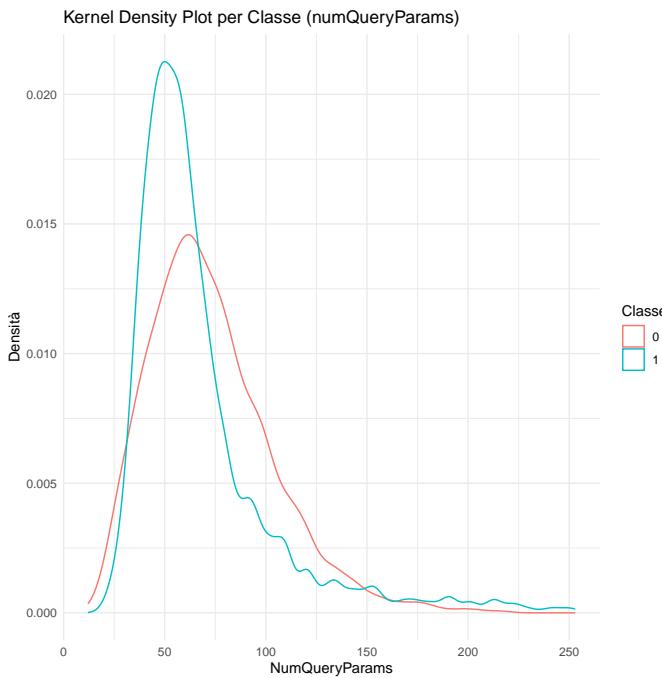


Figura 2.13: Distribuzione di NumQueryParams rispetto alla classe

Il primo Kernel Density Plot (Figura 2.12) mostra la distribuzione del numero di componenti delle query negli URL senza distinzione tra classi. Sull’asse orizzontale è riportato il numero di parametri presenti nell’URL, mentre sull’asse verticale è rappresentata la densità stimata.

L’analisi del grafico evidenzia che la distribuzione presenta un picco nella parte bassa dell’asse x , indicando che la maggior parte degli URL contiene pochi parametri nelle query. Tuttavia, si osserva una coda verso destra, suggerendo che esistono anche URL con un numero significativamente più alto di parametri. Questa tendenza potrebbe indicare che alcuni URL fanno un uso intensivo di query lunghe, un aspetto che potrebbe essere rilevante per l’analisi della sicurezza.

Il secondo Kernel Density Plot (Figura 2.13) suddivide la distribuzione in base alla classe dell’URL, distinguendo tra siti legittimi (0) e siti di phishing (1). L’asse orizzontale rappresenta il numero di parametri delle query, mentre l’asse verticale mostra la densità stimata per ciascuna classe.

Dall’analisi del grafico emerge che le due classi seguono pattern differenti nella distribuzione del numero di parametri nelle query. Se la curva relativa agli URL di phishing mostra una tendenza a includere un numero più elevato di parametri rispetto agli URL legittimi, ciò potrebbe suggerire che le query particolarmente lunghe

siano una caratteristica più frequente nei siti malevoli. Un’eventuale separazione significativa tra le due curve potrebbe indicare che questa variabile è un elemento discriminante nella classificazione degli URL.

CAPITOLO 3

Statistica Descrittiva Univariata

3.1 Funzione di Distribuzione Empirica

Quando si analizza un dataset, è fondamentale comprendere come sono distribuiti i dati. Un modo per farlo è attraverso la **funzione di distribuzione empirica (FDE)**, che permette di capire, per ogni valore osservato, quanta parte del dataset è minore o uguale a quel valore.

In parole semplici, la FDE può essere vista come un *contatore cumulativo*: dato un insieme di numeri ordinati in ordine crescente, essa indica la percentuale dei dati che è già stata "raggiunta" fino a un certo valore. Se ad esempio abbiamo un dataset con 100 valori e la funzione ci dice che per il numero 50 la FDE è pari a 0.7, significa che il **70% dei dati è minore o uguale a 50**.

Questa funzione è utile perché fornisce una rappresentazione visiva e numerica della distribuzione dei dati, aiutando a individuare eventuali concentrazioni, asimmetrie o anomalie nella distribuzione. Nei paragrafi successivi verrà illustrato il metodo di calcolo della funzione di distribuzione empirica e la sua applicazione al dataset analizzato.

Tuttavia, nel dataset analizzato, la **quasi totalità delle feature è di tipo binario**, assumendo unicamente valori 0 o 1. In questi casi, la distribuzione empirica è

già completamente descritta dalla **frequenza relativa** di ciascun valore: la FDE di una variabile binaria si riduce semplicemente alla percentuale di osservazioni che assumono il valore 0 o 1, rendendo superfluo il suo calcolo.

Di conseguenza, la funzione di distribuzione empirica verrà calcolata e analizzata unicamente per la variabile **NumDash**, che è una delle poche variabili continue del dataset. Questa scelta è motivata dal fatto che, per le variabili binarie, la FDE assume solo due valori discreti, mentre per una variabile continua può fornire informazioni dettagliate sulla distribuzione dei dati nel dataset.

3.2 Funzione di Distribuzione Empirica Continua

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

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

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

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

3.2.1 Distribuzione Empirica Continua - NumDash

Dunque, per calcolare la **funzione di distribuzione empirica continua** (FDEC), è necessario suddividere il dominio della variabile analizzata in classi di ampiezza uniforme. La suddivisione dell'intervallo consente di ottenere una rappresentazione chiara e dettagliata della distribuzione dei dati.

Nel nostro caso, la variabile analizzata assume valori compresi tra un minimo di 0 e un massimo di 55. La scelta delle classi deve quindi coprire l'intero intervallo $[0, 55]$, garantendo al contempo un buon bilanciamento tra precisione e leggibilità della distribuzione.

Un criterio comune per la suddivisione in classi consiste nell'utilizzo di circa 10 intervalli. Per determinare l'ampiezza ottimale di ciascuna classe, si utilizza la formula:

$$\text{Am piezza classe} = \frac{\text{valore massimo} - \text{valore minimo}}{\text{numero di classi desiderato}}$$

Applicando i valori del nostro dataset:

$$\frac{55 - 0}{10} = 5.5$$

Poiché l'ampiezza deve essere un valore intero per semplificare l'interpretazione, arrotondiamo il risultato a 5. In questo modo, le classi definite sono:

$[0, 5), [5, 10), [10, 15), [15, 20), [20, 25), [25, 30), [30, 35), [35, 40), [40, 45), [45, 50), [50, 55]$

Questa suddivisione permette di mantenere una buona granularità nella distribuzione senza perdere leggibilità nei dati.

Se fosse necessaria una maggiore precisione, si potrebbe optare per un'ampiezza inferiore, come 2.5, raddoppiando il numero di classi e ottenendo una suddivisione ancora più dettagliata. Tuttavia, per il nostro scopo, la scelta di classi di ampiezza 5 rappresenta un ottimo compromesso tra precisione e interpretabilità dei risultati.

3.2.2 Distribuzione Empirica Continua - NumQueryComponents

Per costruire la funzione di distribuzione empirica continua della variabile *NumQueryComponents*, è necessario suddividere l'intervallo $[0, 23]$ in classi di ampiezza uniforme. La formula per il calcolo delle classi è riportata nella precedente sezione.

Applicando i valori specifici:

$$\frac{23 - 0}{10} = 2.3$$

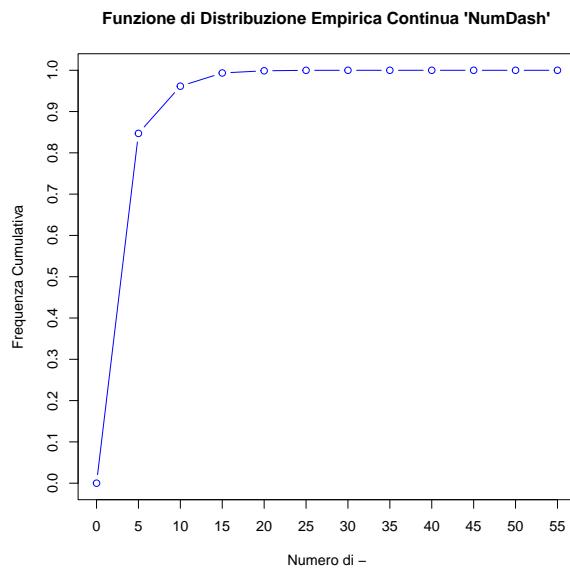


Figura 3.1: FDEC - NumDash

Dato che l’ampiezza deve essere un valore intero per garantire una suddivisione chiara e interpretabile, si arrotonda a 2, ottenendo così il seguente insieme di classi:

$$[0, 2), [2, 4), [4, 6), [6, 8), [8, 10), [10, 12), [12, 14), [14, 16), [16, 18), [18, 20), [20, 23]$$

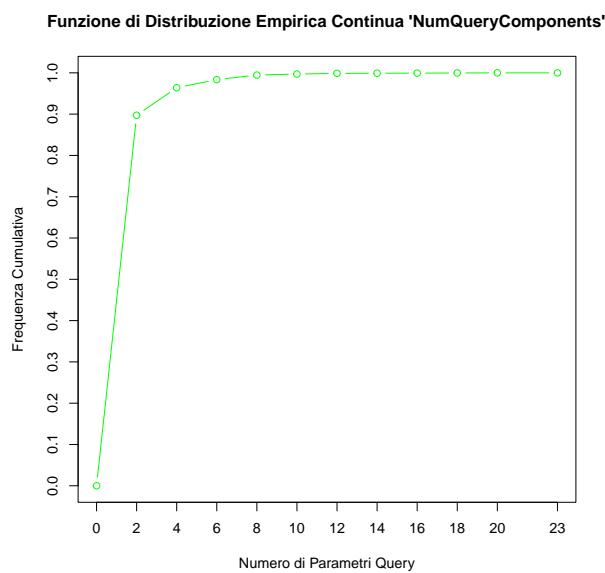


Figura 3.2: FDEC - NumQueryComponents

3.3 Indici di Sintesi

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

3.3.1 Misure di Centralità

Le misure di centralità permettono di individuare il valore attorno al quale tendono a concentrarsi i dati di un dataset. Esse forniscono un riferimento utile per comprendere la posizione centrale della distribuzione e per confrontare diversi insiemi di dati.

Media Campionaria

La **media campionaria** rappresenta il valore medio dei dati osservati ed è calcolata come la media aritmetica dei valori presenti nel dataset. Data una serie di osservazioni numeriche $X = \{x_1, x_2, \dots, x_n\}$, la media campionaria \bar{x} è definita dalla formula:

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

Per ogni osservazione x_i , è possibile calcolare la sua deviazione dalla media, chiamata scarto dalla media campionaria, definito come:

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

Un'importante proprietà della media è che la somma algebrica di tutti gli scarti dalla media è sempre pari a zero, il che riflette il fatto che la media rappresenta un punto di equilibrio nella distribuzione dei dati.

Mediana Campionaria

La **mediana campionaria** è un valore che divide il dataset in due parti di uguale numerosità, in modo che metà dei valori siano inferiori o uguali alla mediana e l'altra metà sia maggiore o uguale ad essa. Per calcolarla, è necessario ordinare i dati in ordine crescente. La sua definizione varia in base alla numerosità del campione:

- Se n è dispari, la mediana è semplicemente il valore in posizione $\frac{n+1}{2}$.
- Se n è pari, la mediana è calcolata come la media dei due valori centrali, ovvero quelli in posizione $\frac{n}{2}$ e $\frac{n}{2} + 1$.

Confrontare la media campionaria e la mediana campionaria permette di ottenere informazioni sulla forma della distribuzione:

- Se media e mediana sono uguali, la distribuzione è simmetrica.
- Se la media è maggiore della mediana, la distribuzione è asimmetrica positiva, cioè con una coda più lunga a destra.
- Se la media è minore della mediana, la distribuzione è asimmetrica negativa, con una coda più lunga a sinistra.

Moda Campionaria

La **moda campionaria** è il valore più frequente all'interno del dataset, ovvero il valore che compare con la maggiore frequenza assoluta o relativa. In alcuni casi, la moda può non essere unica:

- Se esiste un solo valore con frequenza massima, si parla di distribuzione unimodale.
- Se esistono due valori con la stessa frequenza massima, la distribuzione è bimodale.
- Se ci sono più di due valori modali, la distribuzione è detta multimodale.

A differenza della media e della mediana, la moda è particolarmente utile per variabili di tipo discreto o categoriale, poiché permette di identificare il valore più comune senza richiedere calcoli complessi.

Misure di Centralità per NumDash

La **media** è pari a 1.18, indicando che il valore medio della variabile analizzata è leggermente superiore a 1. Tuttavia, la **mediana** e la **moda campionaria** sono

Media	Mediana	Moda campionaria
1.1818	0	0

Tabella 3.1

entrambe pari a 0, suggerendo che la maggior parte delle osservazioni nel dataset assume valori uguali a zero.

Questa discrepanza tra media e mediana indica che la distribuzione è **asimmetrica e influenzata da valori estremi**. In particolare, la presenza di una media superiore alla mediana suggerisce una asimmetria positiva, con alcuni valori elevati che innalzano il valore medio.

Il fatto che la moda campionaria sia anch'essa pari a 0 conferma che il valore più frequente osservato nel dataset è 0, il che può indicare una distribuzione fortemente concentrata in un singolo valore, con una parte della distribuzione caratterizzata da pochi valori significativamente più grandi.

Questa analisi suggerisce che la variabile analizzata potrebbe presentare una distribuzione molto sbilanciata, dove la maggior parte dei dati si trova vicino allo zero, ma con la presenza di alcuni valori elevati che influenzano il valore medio. Pertanto, nelle analisi successive sarà importante considerare metodi robusti, come l'uso della mediana o delle trasformazioni dei dati, per mitigare l'effetto dei valori estremi.

Misure di Centralità per NumQueryComponents

Media	Mediana	Moda campionaria
0.4586	0	0

Tabella 3.2

La **media** della variabile è pari a 0.4586, indicando che il valore medio è inferiore a 1. Tuttavia, sia la **mediana** che la **moda campionaria** sono pari a 0, il che significa che più della metà delle osservazioni assume il valore zero e che questo è anche il valore più frequente nel dataset.

La discrepanza tra la media e la mediana suggerisce che la distribuzione è **asimmetrica e influenzata da valori più alti**, anche se in misura meno pronunciata rispetto a distribuzioni fortemente skewed. In particolare, il fatto che la media sia superiore alla mediana indica una leggera asimmetria positiva, dovuta alla presenza di alcuni valori più grandi che innalzano il valore medio.

La moda campionaria pari a zero conferma che il valore più frequente è 0, suggerendo una distribuzione fortemente concentrata in un singolo valore con una coda destra più lunga. Questo potrebbe indicare che la variabile analizzata è caratterizzata da una grande quantità di osservazioni con valori nulli e da un insieme più ristretto di valori positivi distribuiti su un range più ampio.

Questa analisi evidenzia la necessità di considerare attentamente la natura della distribuzione nei passi successivi, soprattutto nell'uso di misure statistiche e modelli predittivi. L'uso di metriche robuste, come la mediana, potrebbe fornire un quadro più rappresentativo dei dati rispetto alla media, specialmente se la presenza di valori elevati è limitata ma influente.

Misure di Centralità per UrlLength

Media	Mediana	Moda campionaria
70.2641	62	48

Tabella 3.3

La **media** della variabile è pari a 70.26, mentre la **mediana** è pari a 62. Il fatto che la media sia più alta della mediana indica una leggera asimmetria positiva della distribuzione, suggerendo che esistono alcuni valori più elevati che innalzano la media complessiva.

La **moda campionaria** è pari a 48, il che significa che il valore più frequente osservato nel dataset è inferiore sia alla mediana che alla media. Questo suggerisce che la distribuzione presenta una certa concentrazione di valori attorno a 48, ma con una parte della distribuzione estesa verso valori più alti.

Questa situazione è tipica di una distribuzione con una coda lunga verso destra, dove molti dati si concentrano in una fascia più bassa, mentre un numero più ridotto di osservazioni assume valori sensibilmente più alti, influenzando il valore medio.

Nelle analisi successive, potrebbe essere utile verificare la presenza di outlier e valutare metriche robuste, come la mediana, per rappresentare meglio la tendenza centrale della variabile, soprattutto se la distribuzione presenta una coda lunga o una concentrazione significativa attorno a valori più bassi.

3.3.2 Misure di Dispersione

Le misure di dispersione sono statistiche fondamentali per quantificare quanto i valori di un dataset si distribuiscono intorno alla media. Esse forniscono un'indicazione della variabilità dei dati, aiutando a comprendere se i valori sono strettamente concentrati intorno alla media o se presentano una maggiore dispersione.

Varianza

La **varianza campionaria** è una misura che quantifica il grado di dispersione dei dati rispetto alla loro media campionaria. Essa indica, in media, quanto i valori di un campione si discostano dalla media del campione stesso. Data una serie di osservazioni numeriche x_1, x_2, \dots, x_n , la varianza campionaria, indicata con s^2 , è calcolata come:

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

dove \bar{x} rappresenta la media campionaria, n è la numerosità del campione e x_i è il valore della i -esima osservazione. La varianza è espressa in unità al quadrato rispetto ai dati originali, rendendola meno intuitiva da interpretare.

Deviazione Standard

Per ovviare al problema dell'unità di misura della varianza, si utilizza la **deviazione standard campionaria**, che corrisponde alla radice quadrata della varianza:

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

Poiché la deviazione standard è espressa nelle stesse unità dei dati originali, risulta più facilmente interpretabile rispetto alla varianza. Un valore elevato di deviazione standard indica una grande dispersione dei dati attorno alla media, mentre un valore più basso suggerisce una maggiore concentrazione dei dati intorno alla media.

Coefficiente di Variazione

Un’ulteriore misura utile per analizzare la dispersione è il **coefficiente di variazione (CV)**, che misura la dispersione dei dati in termini relativi rispetto alla media. È definito come:

$$CV = \frac{s}{\bar{x}} \times 100$$

Il coefficiente di variazione è espresso in percentuale e permette di confrontare la variabilità di dataset con scale di misura diverse. Un valore alto indica una maggiore dispersione dei dati rispetto alla media, mentre un valore basso suggerisce una distribuzione più stabile e meno variabile. In particolare:

- Se $CV < 10\%$, la variabilità è molto bassa.
- Se $10\% \leq CV < 30\%$, la variabilità è moderata.
- Se $CV \geq 30\%$, la variabilità è elevata e i dati mostrano un’alta dispersione.

Questa misura è particolarmente utile quando si vogliono confrontare la dispersione di variabili con unità di misura diverse o quando si deve valutare la stabilità di una distribuzione rispetto alla sua media.

Misure di Dispersione per NumDash

Varianza	Std Dev	Coefficiente di Variazione
9.65	3.11	170.86%

Tabella 3.4

La **varianza** della variabile è pari a 9.65, indicando un certo grado di dispersione dei dati rispetto alla media. La **deviazione standard**, che rappresenta la radice quadrata della varianza, è pari a 3.11, suggerendo che i valori della variabile si discostano mediamente di circa 3 unità dalla media.

Il **coefficiente di variazione (CV)**, pari a 170.86%, è particolarmente elevato. Il coefficiente di variazione è una misura normalizzata della dispersione, ottenuta come rapporto tra la deviazione standard e la media, espresso in percentuale. Un valore così alto indica che la variabilità dei dati è molto elevata rispetto alla media, suggerendo una distribuzione con una forte dispersione relativa.

Questa elevata variabilità potrebbe essere il risultato della presenza di valori estremi o di una distribuzione fortemente asimmetrica. Nelle analisi successive, sarà importante valutare eventuali trasformazioni dei dati o metodi robusti per ridurre l'effetto della dispersione elevata e migliorare la rappresentazione dei dati.

Misure di Dispersione per NumQueryComponents

Varianza	Std Dev	Coefficiente di Variazione
1.81	1.34	293.24%

Tabella 3.5

La **varianza** della variabile è pari a 1.81, indicando un certo grado di dispersione dei dati rispetto alla media. La **deviazione standard**, che rappresenta la radice quadrata della varianza, è pari a 1.34, suggerendo che i valori della variabile si discostano mediamente di circa 1.34 unità dalla media.

Il **coefficiente di variazione (CV)** è pari a 293.24%, un valore estremamente elevato. Questo coefficiente, calcolato come il rapporto tra deviazione standard e media, misura la dispersione relativa dei dati rispetto alla loro media. Un valore così alto indica che la variabilità dei dati è molto elevata rispetto alla media, suggerendo una distribuzione con una forte dispersione relativa.

Un coefficiente di variazione superiore al 100% generalmente indica una distribuzione in cui la deviazione standard è maggiore della media, suggerendo la presenza

di una grande variabilità nei dati. Questo potrebbe essere dovuto alla presenza di valori estremi o di una distribuzione altamente asimmetrica.

Misure di Dispersione per UrlLength

Varianza	Std Dev	Coefficiente di Variazione
1113.55	33.37	47.49%

Tabella 3.6

La **varianza** della variabile è pari a 1113.55, indicando una significativa dispersione dei dati rispetto alla media. La **deviazione standard**, pari a 33.37, fornisce un'indicazione più intuitiva della variabilità, suggerendo che i valori della variabile si discostano mediamente di circa 33 unità dalla media.

Il **coefficiente di variazione (CV)** è pari a 47.49%, un valore moderatamente elevato. Questo coefficiente, ottenuto rapportando la deviazione standard alla media e moltiplicandolo per 100, misura la dispersione relativa dei dati. Un valore vicino al 50% suggerisce una distribuzione con una discreta variabilità, ma non estremamente dispersa.

Un CV inferiore al 100% generalmente indica che la deviazione standard è inferiore alla media, suggerendo una distribuzione più stabile rispetto ai casi in cui il CV è molto più alto. Tuttavia, la presenza di una varianza elevata potrebbe indicare che alcuni valori nel dataset si distaccano sensibilmente dalla media.

3.3.3 Misure di Simmetria

Le misure di simmetria permettono di valutare la forma di una distribuzione, indicando se essa è simmetrica o presenta un certo grado di asimmetria. Questi indicatori aiutano a comprendere se i dati sono distribuiti in modo equilibrato intorno alla media o se tendono a concentrarsi in una direzione piuttosto che nell'altra.

Skewness Campionaria

La **skewness campionaria** misura il grado di asimmetria della distribuzione di un insieme di dati numerici. Data una serie di osservazioni x_1, x_2, \dots, x_n , la skewness si calcola come:

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

dove m_2 è il secondo momento campionario (varianza) e m_3 è il terzo momento campionario, che indica quanto la distribuzione si discosta dalla simmetria.

L'interpretazione della skewness è la seguente:

- Se $\gamma_1 = 0$, la distribuzione è **simmetrica**, ossia non presenta squilibri tra i valori più bassi e quelli più alti.
- Se $\gamma_1 > 0$, la distribuzione è **asimmetrica positiva**, cioè ha una coda più lunga a destra, indicando la presenza di valori elevati che influenzano la media.
- Se $\gamma_1 < 0$, la distribuzione è **asimmetrica negativa**, con una coda più pronunciata a sinistra, suggerendo la presenza di valori più bassi che distorcono la distribuzione.

Curtosi Campionaria

La **curtosi campionaria** misura quanto la distribuzione dei dati è concentrata attorno alla media e la presenza di valori estremi. È definita come:

$$\gamma_2 = \beta_2 - 3$$

dove β_2 , noto come *indice di Pearson*, è calcolato come il rapporto tra il quarto momento campionario m_4 e il quadrato del secondo momento campionario m_2 :

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

L'interpretazione della curtosi è la seguente:

- Se $\beta_2 = 3$, la distribuzione è **normocurtica** (o *mesocurtica*), con una forma simile a quella della distribuzione normale.
- Se $\beta_2 < 3$, la distribuzione è **platicurtica**, ovvero più piatta rispetto a una normale, con code meno pronunciate e meno valori estremi.
- Se $\beta_2 > 3$, la distribuzione è **leptocurtica**, con un picco più accentuato e code più pesanti, indicando una maggiore presenza di valori estremi.

Misure di Simmetria per NumDash

Skewness	Curtosi
2.791489	14.71037

Tabella 3.7

La **skewness**, che misura il grado di asimmetria della distribuzione, ha un valore pari a **2.79**, indicando una **forte asimmetria positiva**. Questo significa che la distribuzione dei dati è sbilanciata verso destra, con una coda più lunga nella parte alta dei valori.

La **curtosi**, che descrive la concentrazione dei dati attorno alla media e la pesantezza delle code della distribuzione, assume un valore pari a **14.71**. Un valore così elevato indica che la distribuzione è **fortemente leptocurtica**, ovvero presenta un picco molto pronunciato e code più pesanti rispetto a una distribuzione normale. Questo suggerisce la presenza di **outlier significativi** e una distribuzione caratterizzata da una forte concentrazione di valori intorno alla media, con eventi estremi più frequenti.

Questi risultati evidenziano che la distribuzione della variabile analizzata si discosta significativamente da una distribuzione normale, con una tendenza a valori più elevati e la presenza di dati anomali che influenzano la sua forma.

Misure di Simmetria per NumQueryComponents

Skewness	Curtosi
5.119783	40.12744

Tabella 3.8

Il valore della **skewness** è pari a 5.12, il che indica una fortissima asimmetria positiva. Questo significa che la distribuzione è estremamente sbilanciata verso destra, con una coda molto lunga nella parte alta dei valori. In altre parole, la maggior parte dei dati è concentrata nella parte bassa della distribuzione, mentre esistono alcuni valori molto elevati che influiscono in modo significativo sulla forma della distribuzione.

La **curtosi** assume un valore di 40.13, un valore estremamente elevato rispetto al valore di riferimento 3, che caratterizza una distribuzione normale. Questo indica che la distribuzione è altamente leptocurtica, ovvero presenta un picco centrale molto accentuato e code estremamente pesanti. Una curtosi così alta suggerisce la presenza di numerosi outlier significativi, ossia valori anomali che si discostano fortemente dalla media.

Questi risultati evidenziano che la distribuzione è molto lontana da una distribuzione normale, con una concentrazione elevata dei dati intorno alla media e la presenza di valori molto grandi che influenzano fortemente la forma della distribuzione. Nelle analisi successive, sarà importante considerare l'impatto di questa forte asimmetria e la presenza di outlier, valutando eventualmente tecniche di trasformazione dei dati o metodi robusti per l'analisi.

Misure di Simmetria per UrlLength

Skewness	Curtosi
1.704753	4.180494

Tabella 3.9

Il valore della **skewness** è pari a 1.70, il che indica una moderata asimmetria positiva. Questo significa che la distribuzione è leggermente sbilanciata verso destra, con una coda più lunga nella parte alta dei valori. Tuttavia, l'asimmetria non è estrema e i dati sono relativamente concentrati intorno alla media.

La **curtosi** ha un valore di 4.18, che è leggermente superiore al valore di riferimento 3, caratteristico di una distribuzione normale. Questo suggerisce che la distribuzione è leggermente leptocurtica, ovvero presenta un picco centrale più pronunciato rispetto a una distribuzione normale, con una maggiore concentrazione dei dati vicino alla media e code leggermente più pesanti.

Questi risultati indicano che, pur essendo la distribuzione non perfettamente normale, essa non presenta una distorsione eccessiva. La leggera asimmetria e la curtosi moderata suggeriscono la presenza di alcuni valori estremi, ma senza un impatto particolarmente significativo sulle analisi. Pertanto, la distribuzione potrebbe essere adatta a modelli statistici standard, con un'eventuale valutazione della presenza di outlier per migliorare l'interpretazione dei risultati.

CAPITOLO 4

Statistica descrittiva Bivariata

La statistica descrittiva bivariata si occupa dell'analisi delle relazioni tra due variabili, utilizzando metodi sia grafici che statistici. Uno degli strumenti più efficaci per rappresentare la relazione tra due variabili quantitative è il **diagramma di dispersione** (scatterplot), che visualizza le coppie di osservazioni come punti in un piano cartesiano. Oltre ai metodi grafici, si impiegano misure statistiche come la **covarianza campionaria** e il **coefficiente di correlazione campionario** per quantificare il grado di associazione tra le variabili.

4.1 Correlazione fra le variabili

4.1.1 Covarianza Campionaria

La **covarianza campionaria** misura l'intensità e la direzione della relazione lineare tra due variabili quantitative. Sia (X, Y) una coppia di variabili e consideriamo un campione di n osservazioni $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. La covarianza campionaria è definita dalla formula:

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (4.1.1)$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{e} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad (4.1.2)$$

$C_{XY} > 0$ indica una relazione diretta tra X e Y (4.1.3)

$C_{XY} < 0$ indica una relazione inversa tra X e Y (4.1.4)

$C_{XY} = 0$ indica assenza di relazione lineare tra X e Y (4.1.5)

La divisione per $(n - 1)$ permette di ottenere una stima non distorta della covarianza, rendendola più affidabile nelle applicazioni pratiche.

4.1.2 Coefficiente di Correlazione Campionario

Il **coefficiente di correlazione campionario** è un indice adimensionale che misura l'intensità della relazione lineare tra due variabili, normalizzando la covarianza. La sua definizione matematica è:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} \quad (4.1.6)$$

Dove s_X e s_Y rappresentano le deviazioni standard campionarie delle variabili X e Y.

Le proprietà fondamentali di r_{XY} sono:

- $-1 \leq r_{XY} \leq 1$: il valore è sempre compreso tra -1 e 1;
- Se $r_{XY} = 1$, esiste una relazione lineare perfetta positiva;
- Se $0 < r_{XY} < 1$, esiste una relazione positiva, ma con una certa dispersione dei dati;
- Se $r_{XY} = 0$, non vi è alcuna correlazione lineare;
- Se $-1 < r_{XY} < 0$, esiste una relazione negativa;
- Se $r_{XY} = -1$, le variabili sono perfettamente correlate in modo inverso.

Una caratteristica importante di r_{XY} è la sua invarianza rispetto alle trasformazioni lineari delle variabili, il che significa che non dipende dalle unità di misura adottate.

4.2 Analisi della Relazione tra le Feature e la Variabile Target

In questa sezione vengono analizzate le feature **PctExtNullSelfRedirectHyperlinksRT**, **FrequentDomainNameMismatch**, **NumDash**, **SubmitInfoToEmail** e **PctNullSelfRedirectHyperlinks** rispetto alla variabile target **CLASS_LABEL**. L’analisi viene condotta attraverso il calcolo della covarianza campionaria, del coefficiente di correlazione di Pearson e la visualizzazione mediante scatterplot.

4.2.1 Covarianza e Correlazione

La tabella seguente riporta i valori di covarianza e correlazione per ciascuna delle feature analizzate rispetto alla variabile target:

Feature	Covarianza	Correlazione	p-value
PctExtNullSelfRedirectHyperlinksRT	-0.2427	-0.5405	≈ 0
FrequentDomainNameMismatch	-0.1742	-0.4639	≈ 0
NumDash	-0.1034	-0.3722	≈ 0
SubmitInfoToEmail	-0.1041	-0.3576	1.7070×10^{-299}
PctNullSelfRedirectHyperlinks	-0.0875	-0.3428	9.0132×10^{-274}

Tabella 4.1: Covarianza e correlazione tra le feature e la variabile target.

4.2.2 Interpretazione dei Risultati

PctExtNullSelfRedirectHyperlinksRT: La correlazione negativa di -0.5405 indica che, all’aumentare della percentuale di hyperlink autoreferenziali, la probabilità di phishing diminuisce. Questo suggerisce che i siti legittimi tendono ad avere più link interni, mentre i siti phishing spesso contengono meno collegamenti autoreferenziali.

FrequentDomainNameMismatch: La correlazione negativa di -0.4639 mostra che un alto livello di discordanza tra il dominio principale e i sottodomini è più

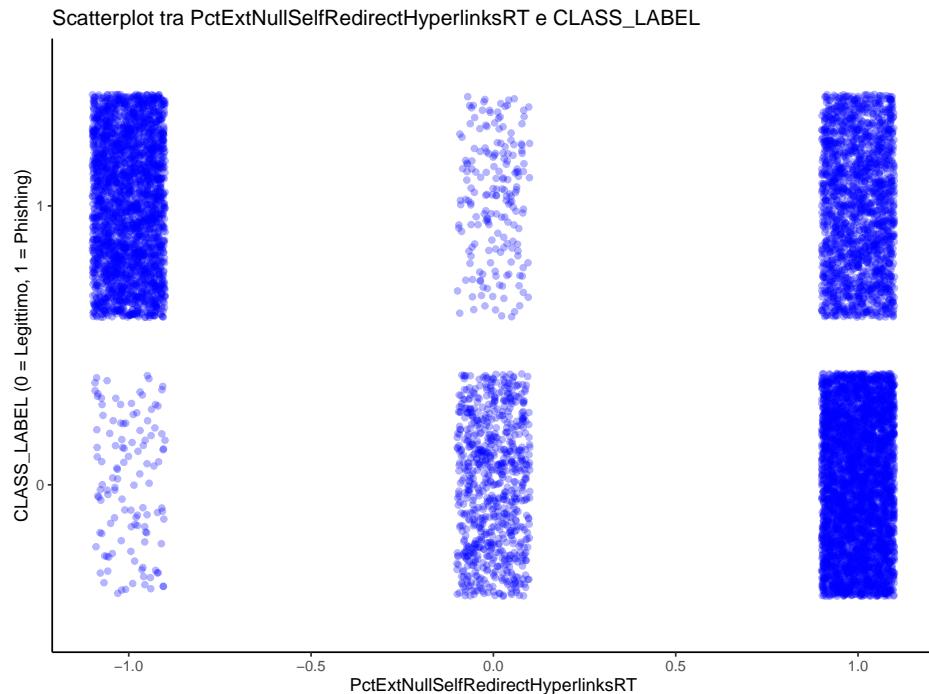


Figura 4.1: Scatterplot tra PctExtNullSelfRedirectHyperlinksRT e CLASS_LABEL

frequente nei siti legittimi che nei siti phishing. I siti fraudolenti spesso cercano di imitare domini esistenti con maggiore coerenza nei nomi di dominio.

NumDash: Con una correlazione di -0.3722, si nota che un numero elevato di trattini nell'URL non è un indicatore fortissimo di phishing, ma comunque ha un'influenza. I siti legittimi tendono ad avere URL più strutturati, mentre alcuni siti phishing possono avere più trattini per generare URL complessi.

SubmitInfoToEmail: La correlazione di -0.3576 suggerisce che i siti che inviano dati utente direttamente via email anziché a un server backend sono fortemente sospetti. Questo valore indica che i siti phishing fanno spesso uso di questa tecnica, mentre i siti legittimi evitano di inviare informazioni sensibili tramite email.

4.2.3 Visualizzazione con Scatterplot

Per comprendere meglio la distribuzione dei dati, sono stati generati gli scatterplot delle feature rispetto alla variabile target. I grafici evidenziano la relazione tra le feature e la variabile target. Si nota come alcune variabili presentino una maggiore dispersione dei dati, mentre altre mostrano un pattern più definito.

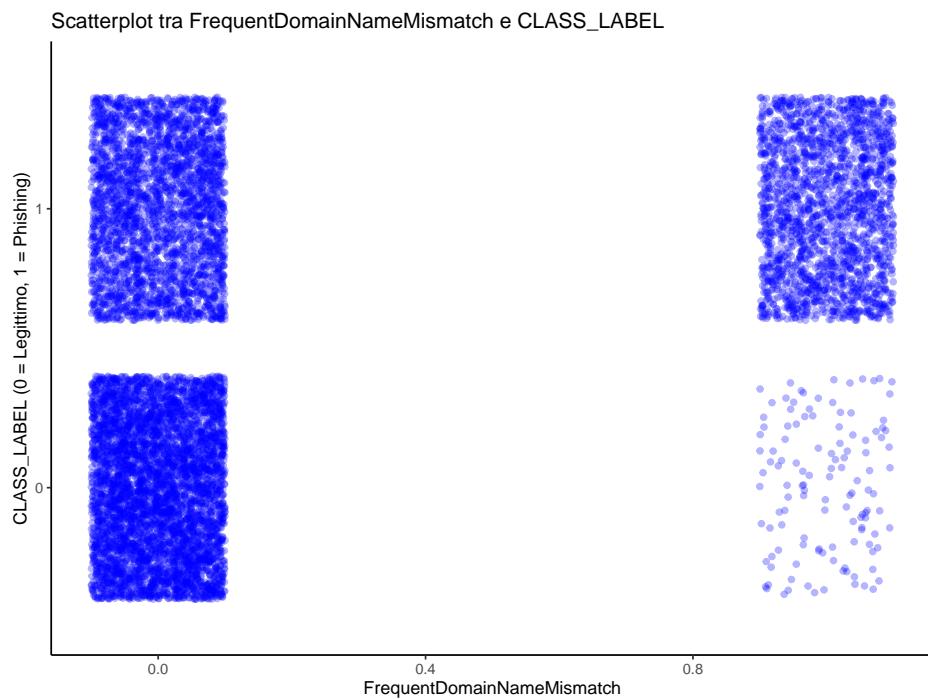


Figura 4.2: Scatterplot tra FrequentDomainNameMismatch e CLASS_LABEL

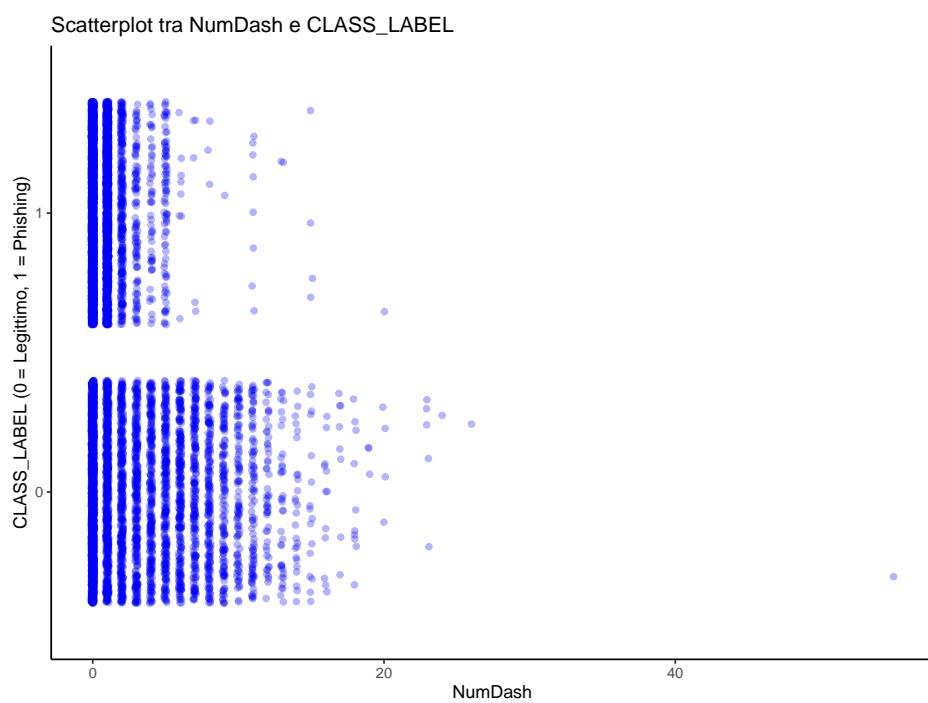


Figura 4.3: Scatterplot tra NumDash e CLASS_LABEL

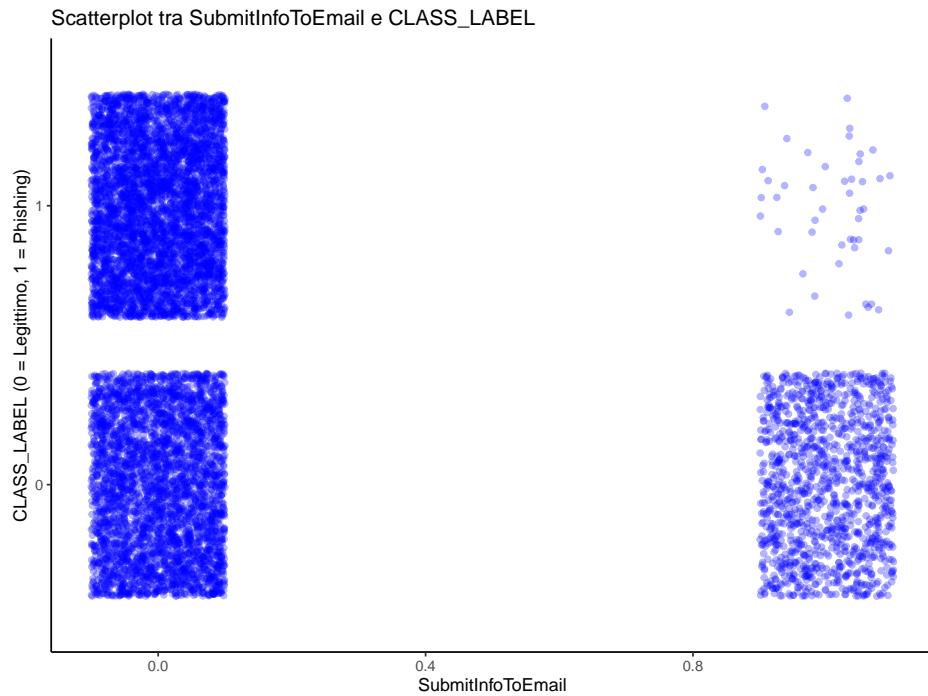


Figura 4.4: Scatterplot tra SubmitInfoToEmail e CLASS_LABEL

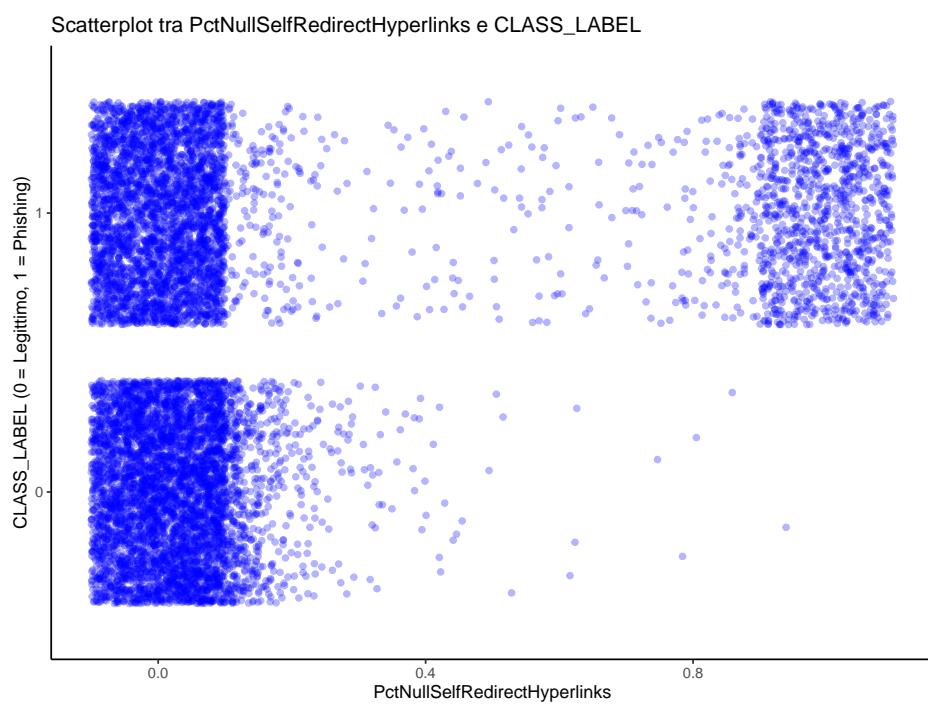


Figura 4.5: Scatterplot tra PctNullSelfRedirectHyperlinks e CLASS_LABEL

4.2.4 Analisi della Matrice di Correlazione

La matrice di correlazione riportata consente di osservare le relazioni lineari tra la variabile target **CLASS_LABEL** e le feature selezionate, nonché le interrelazioni tra le feature stesse. L'intensità delle relazioni è rappresentata cromaticamente: i valori prossimi a 1 (rosso intenso) indicano una forte correlazione positiva, mentre i valori prossimi a -1 (toni più scuri) indicano una forte correlazione negativa.

4.2.5 Interpretazione della Matrice di Correlazione

PctExtNullSelfRedirectHyperlinksRT: Presenta una correlazione negativa significativa con **CLASS_LABEL** (-0.54), indicando che i siti phishing tendono ad avere una percentuale inferiore di hyperlink autoreferenziali rispetto ai siti legittimi.

FrequentDomainNameMismatch: Ha una correlazione moderata positiva con **CLASS_LABEL** (0.46), suggerendo che la discordanza nei nomi di dominio è più comune nei siti phishing.

NumDash: La correlazione negativa di -0.37 con **CLASS_LABEL** indica che un numero elevato di trattini nell'URL è un indicatore moderato di phishing.

SubmitInfoToEmail: Correlazione di -0.36 con **CLASS_LABEL**, confermando che i siti phishing tendono a inviare più spesso informazioni tramite email rispetto ai siti legittimi.

PctNullSelfRedirectHyperlinks: Correlazione positiva di 0.34 con **CLASS_LABEL**, suggerendo che nei siti phishing è presente un numero relativamente più elevato di hyperlink nulli o autoreferenziali rispetto ai siti legittimi.

L'analisi di questa matrice conferma che alcune feature sono più utili di altre nel distinguere tra siti phishing e legittimi. La presenza di correlazioni forti o moderate tra alcune variabili suggerisce che potrebbero esserci pattern rilevanti da sfruttare in modelli di classificazione automatica.

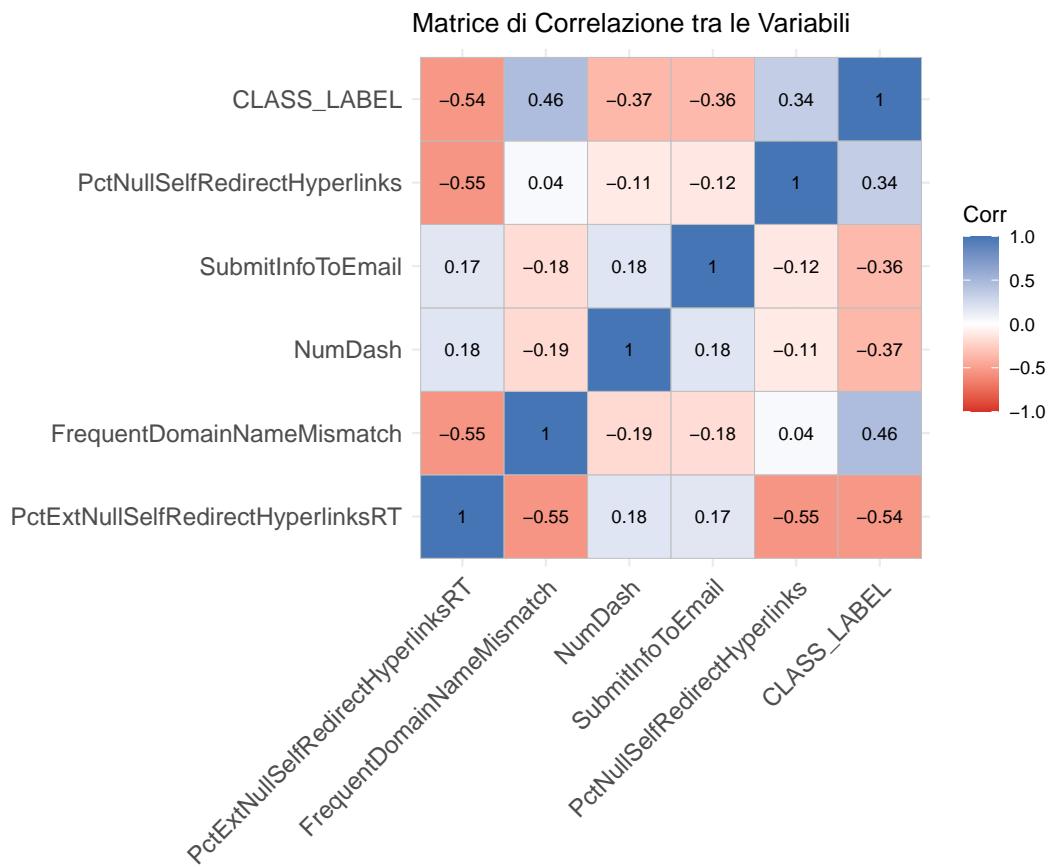


Figura 4.6: Matrice di Correlazione tra le Variabili.

CAPITOLO 5

Creazione di un Modello predittivo

5.1 Regressione Logistica

La **regressione logistica** è utilizzata per modellare la probabilità che un evento si verifichi (ad esempio, nel nostro caso, che un sito sia phishing) in funzione di una o più variabili indipendenti.

La funzione logit è definita come:

$$\text{Logit}(\pi) = \ln\left(\frac{\pi}{1 - \pi}\right) \quad (5.1.1)$$

Dove π rappresenta la probabilità che l'evento si verifichi.

L'equazione del modello logit è data da:

$$\ln\left(\frac{\pi}{1 - \pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad (5.1.2)$$

La regressione logistica è stata utilizzata per modellare la probabilità che un sito web sia phishing o legittimo in base alle feature selezionate. Dopo la fase di preprocessing, il modello ha ottenuto un'accuratezza di **87.1%**, dimostrando una buona capacità predittiva.

5.1.1 Risultati della Regressione Logistica

I coefficienti stimati dal modello, insieme ai loro errori standard e ai valori di significatività statistica, sono riportati nella tabella seguente.

Variabile	Coefficiente	Errore Std.	Z-value	p-value
Intercetta	10.5305	0.4493	23.437	< 2e-16 *
PctExtNullSelfRedirectHyperlinksRT0	-10.6685	0.4494	-23.740	< 2e-16 *
PctExtNullSelfRedirectHyperlinksRT1	-14.7679	0.6695	-22.057	< 2e-16 *
FrequentDomainNameMismatch1	3.2099	0.1789	17.938	< 2e-16 *
NumDash	-1.2671	0.0736	-17.212	< 2e-16 *
PctNullSelfRedirectHyperlinks	-3.1085	0.2339	-13.290	< 2e-16 *
PctExtHyperlinks	-4.9690	0.2435	-20.404	< 2e-16 *
PathLevel	0.8179	0.0501	16.330	< 2e-16 *
UrlLength	-0.1143	0.0377	-3.029	0.00245
HostnameLength	0.2838	0.0463	6.126	9.02e-10 *

Tabella 5.1: Coefficiente stimati per la regressione logistica.

5.1.2 Interpretazione dei Risultati

Dall’analisi dei coefficienti possiamo osservare quanto segue:

- **PctExtNullSelfRedirectHyperlinksRT0 e PctExtNullSelfRedirectHyperlinksRT1** hanno coefficienti altamente negativi, indicando una forte correlazione inversa con la probabilità che un sito sia phishing.
- **FrequentDomainNameMismatch1** ha un coefficiente positivo significativo, suggerendo che questa feature è un buon predittore per il phishing.
- **NumDash, PctNullSelfRedirectHyperlinks e PctExtHyperlinks** mostrano anch’essi coefficienti negativi, il che indica che valori elevati in queste variabili sono associati a un minor rischio di phishing.

Predetto / Reale	Classe 0 (Legittimo)	Classe 1 (Phishing)
Classe 0 (Predetto)	1076	170
Classe 1 (Predetto)	163	1176

Tabella 5.2: Matrice di confusione della regressione logistica.

- **PathLevel e HostnameLength** hanno coefficienti positivi, indicando che un numero maggiore di sottodirectory nel percorso dell'URL e una lunghezza maggiore dell'hostname sono correlati a un rischio maggiore di phishing.

5.1.3 Metriche del Modello

Le metriche principali del modello sono le seguenti:

- **Deviance Nulla:** 8401.3
- **Deviance Residua:** 3681.5
- **AIC:** 3701.5
- **Numero di iterazioni:** 7
- **Osservazioni eliminate per valori mancanti:** 936

5.1.4 Matrice di Confusione e Accuratezza

La matrice di confusione del modello mostra il numero di predizioni corrette e errate.

L'accuratezza finale del modello è pari a 87.1%, indicando una buona capacità di classificazione tra siti phishing e legittimi.

5.1.5 Analisi della Learning Curve

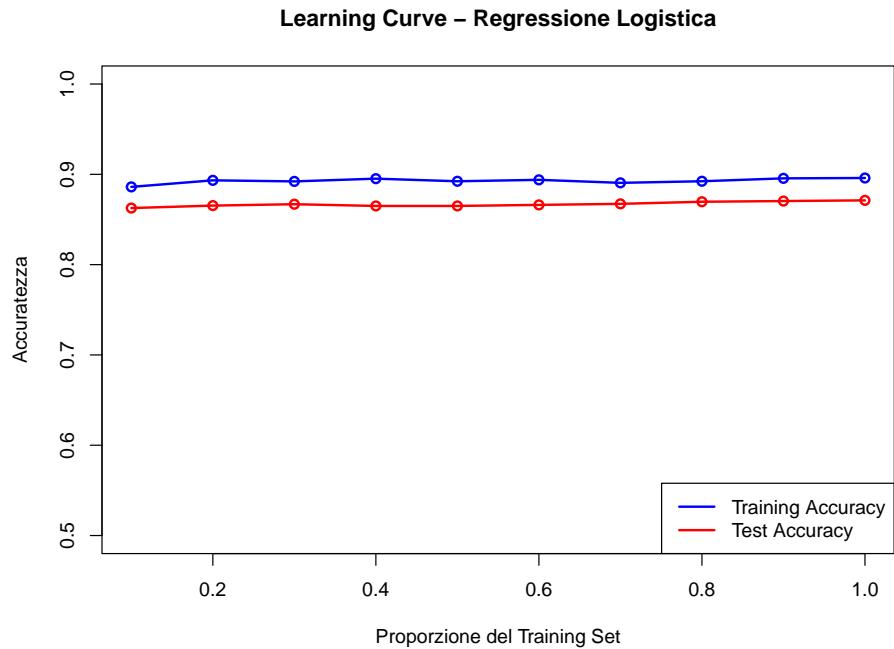


Figura 5.1: Learning Curve della regressione logistica.

La learning curve permette di osservare come varia l'accuratezza del modello all'aumentare della dimensione del dataset di training. Nel grafico seguente, la curva blu rappresenta l'accuratezza sul training set, mentre la curva rossa mostra l'accuratezza sul test set.

Dall'analisi della learning curve possiamo osservare:

- L'accuratezza sul **training set** rimane elevata per tutte le dimensioni del dataset, indicando che il modello è in grado di apprendere bene.
- L'accuratezza sul **test set** mostra una crescita stabile fino a convergere, suggerendo che il modello generalizza bene sui dati non visti.
- L'assenza di una grande differenza tra training e test accuracy suggerisce che il modello non soffre di overfitting.

5.2 Interpretazione dei Risultati con SHAP

Abbiamo utilizzato i valori SHAP per analizzare l'importanza delle variabili nel modello. Di seguito sono riportati i principali risultati.

5.2.1 Valori SHAP per una Predizione Specifica

Per una specifica osservazione del test set, i contributi delle variabili sono stati i seguenti:

- **PctNullSelfRedirectHyperlinks** = +2.52 (aumenta la probabilità di phishing)
- **HostnameLength** = -1.83 (riduce la probabilità di phishing)
- **PctExtHyperlinks** = -0.70 (riduce la probabilità di phishing)
- **UrlLength** = -0.41 (riduce leggermente la probabilità di phishing)
- **PathLevel** = +0.37 (aumenta leggermente la probabilità di phishing)

La predizione finale per questa osservazione è stata **1** (phishing), con una probabilità predetta del 100%. La media delle predizioni per l'intero dataset è di circa 45%.

5.2.2 Grafico dei Valori SHAP

Il grafico mostra come le variabili influenzano la predizione del modello. Le barre blu indicano un contributo positivo verso la classificazione come phishing, mentre le barre rosse indicano un contributo negativo.

5.2.3 Conclusioni sulla Regressione Logistica

La regressione logistica si è dimostrata un modello efficace per la classificazione di siti phishing, con un'accuratezza finale del 87.1%. L'analisi della learning curve conferma che il modello non soffre di overfitting, suggerendo che può essere ulteriormente ottimizzato senza rischio di peggiorare la generalizzazione. Possibili miglioramenti futuri includono l'uso di modelli più complessi come le Random

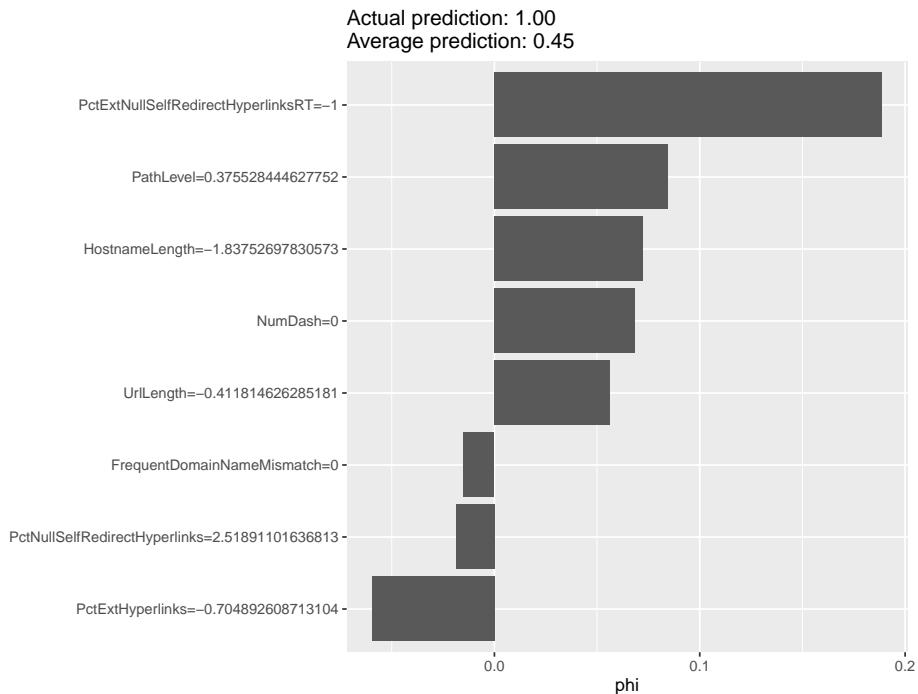


Figura 5.2: Grafico dei valori SHAP per le variabili selezionate.

Forest o i Gradient Boosting Models, che potrebbero migliorare ulteriormente la capacità predittiva.

CAPITOLO 6

Generazione e Analisi di un Dataset Sintetico

L'utilizzo di modelli di linguaggio di grandi dimensioni (LLM) nella generazione di dati sintetici è una tecnica innovativa che permette di simulare dataset realistici, preservando le proprietà statistiche fondamentali dei dati originali. In questo studio, abbiamo utilizzato un modello LLM per generare un dataset sintetico a partire dalle caratteristiche osservate nel dataset reale, con l'obiettivo di valutare la qualità della generazione attraverso un'analisi comparativa.

6.1 Metodologia di Generazione

Per la generazione del dataset sintetico, abbiamo sviluppato un **agent** in Python utilizzando la libreria LangGraph. Il modello utilizzato per la generazione è **GPT-4o**, selezionato in quanto rappresenta lo stato dell'arte nei modelli di linguaggio di grandi dimensioni ed è attualmente uno dei più performanti disponibili. L'agent sfrutta le API di OpenAI per la generazione dei dati.

La progettazione dell'architettura dell'agent si basa sul **Whitepaper di Google** [3] relativo agli agenti intelligenti, dal quale abbiamo preso ispirazione per definire il flusso di generazione dei dati sintetici.

Struttura a Grafo dell'Agent

Utilizzando LangGraph, il nostro agent viene rappresentato come un **grafo direzionale**, in cui ogni nodo rappresenta una fase specifica del processo di generazione. La struttura del grafo è illustrata in Figura 6.1.

Nel nostro agent, i nodi del grafo sono i seguenti:

1. **Nodo di generazione:** esegue la chiamata al LLM (GPT-4o) per generare i dati sintetici, basandosi sulle caratteristiche statistiche del dataset originale.
2. **Nodo di check sintattico e critica:** applica una serie di controlli programmatici sui dati generati per rilevare eventuali problemi (es. valori fuori range, incongruenze statistiche).

Dal **Nodo 2**, se vengono rilevati problemi nei dati, il flusso ritorna al **Nodo 1**, avviando una nuova generazione. Questo processo iterativo introduce un meccanismo simile al **self-refinement**[2], permettendo al sistema di correggere automaticamente eventuali anomalie nei dati generati, garantendo una maggiore fedeltà al dataset originale.

Componenti Principali dell'Agent

Nel whitepaper [3] di Google viene proposto un modello architettonale per Agents, composto da Orchestratore, Modello e Tools.

Il nostro agent si compone delle seguenti componenti chiave:

- **Orchestratore:** LangGraph, utilizzato per la gestione e il controllo del flusso di generazione dei dati.
- **Modello: GPT-4o**, impiegato per generare dati coerenti con il dataset originale.
- **Tool:** abbiamo implementato un tool specifico che verifica che i dati generati rispettino le proprietà statistiche del dataset originale. In particolare, il tool controlla che i valori generati rientrino negli intervalli minimi e massimi osservati nelle feature originali, evitando così la creazione di dati non realistici.

L'agent segue un processo iterativo di generazione e validazione, in cui ogni batch di dati sintetici passa per una funzione in cui si controlla che i valori siano

ammissibili (e.g. valori che rientrino in un range specifico). Se un valore generato risulta fuori dai limiti stabiliti, viene scartato o modificato per rientrare nel range corretto, assicurando che la distribuzione dei dati rimanga il più possibile fedele a quella del dataset originale.

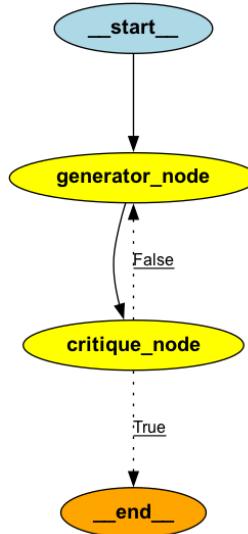


Figura 6.1: Struttura sotto forma di grafo dell'agent di Data Generation

Prompt Engineering

Per la generazione del dataset sintetico, abbiamo effettuato più iterazioni di generazione fino a ottenere un dataset di **10.000 righe**, la stessa dimensione del dataset originale. Durante questo processo, abbiamo sperimentato con due diverse versioni del prompt per ottimizzare la qualità dei dati generati.

Il primo prompt utilizzato era il seguente:

You are a data generator.

Your task is to generate 1 sample based on the information I will provide.

The dataset you must generate is about phishing and legit websites.

The columns are:

`id, NumDots, SubdomainLevel, PathLevel, UrlLength, ...`

`PctExtNullSelfRedirectHyperlinksRT, CLASS_LABEL`

Examples are:

```
2, 3, 1, 3, ..., 1;
6434, 3, 1, 1, ..., 0;
```

You must respond with data only, without any introductory or explanatory text.

The structure must be a dictionary-like format because I will use `ast.literal_eval` to parse the data you generate.

Come si può osservare, abbiamo adottato un approccio **2-shot**, fornendo al modello un esempio per la classe positiva (1, phishing) e uno per la classe negativa (0, legittimo). Questo aveva lo scopo di guidare il modello nella generazione di dati coerenti con la distribuzione osservata nel dataset originale.

Tuttavia, durante le prime generazioni, abbiamo riscontrato un problema significativo: **il modello tendeva a generare dati che non rispettavano i limiti di valori delle variabili presenti nel dataset originale**. In particolare, alcune feature assumevano valori fuori dal range osservato nei dati reali, compromettendo la qualità della simulazione.

Per risolvere questa problematica, abbiamo modificato il prompt aggiungendo informazioni esplicite riguardo:

- **I limiti di valori per ogni feature**, laddove questi erano ben definiti e sensati.
- **Il tipo di ogni colonna**, specificando se la variabile fosse numerica continua, numerica discreta o binaria.

Il prompt finale risulta essere il precedente più l'aggiunta di queste informazioni, è il seguente:

```
"id": "type": int, "min": 0, "max": None,
"NumDots": "type": int, "min": 0, "max": None,
"SubdomainLevel": "type": int, "min": 0, "max": None,
"PathLevel": "type": int, "min": 0, "max": None,
"UrlLength": "type": int, "min": 0, "max": None,
...
"ExtMetaScriptLinkRT": "type": int, "min": -1, "max": 1,
"CLASS_LABEL": "type": int, "min": 0, "max": 1,
```

Dopo questa modifica, i dati generati hanno iniziato ad avere la struttura desiderata, con valori che rispettavano i limiti del dataset originale e una distribuzione più realistica.

Finding 1. Una corretta ingegnerizzazione del prompt, con l'aggiunta di informazioni relative al contesto, può migliorare significativamente la qualità dei dati sintetici prodotti tramite modelli LLM.

6.2 Obiettivi della Generazione Sintetica

L'analisi si propone di verificare se il dataset sintetico conserva le seguenti proprietà statistiche:

- **Coerenza delle misure di centralità:** la media, la mediana e la moda delle variabili sintetiche devono essere compatibili con quelle osservate nel dataset reale.
- **Stabilità della dispersione:** la varianza e la deviazione standard devono riflettere una distribuzione simile a quella dei dati originali.
- **Mantenimento delle relazioni tra le variabili:** la correlazione tra le feature deve essere preservata, in modo da garantire che la struttura del dataset non sia alterata.
- **Presenza di pattern statistici:** distribuzioni, densità e asimmetrie devono essere rispettate, evitando distorsioni che potrebbero compromettere l'affidabilità dei dati sintetici.

Da tali obiettivi emergono le seguenti domande di ricerca (Research Questions):

Q RQ₁. In che misura la media, la mediana e la moda delle variabili sintetiche si discostano da quelle osservate nei dati reali?

Q RQ₂. La varianza e la deviazione standard delle variabili sintetiche sono statisticamente indistinguibili da quelle dei dati originali?

Q RQ₃. In che misura la matrice di correlazione delle variabili sintetiche differisce da quella dei dati originali?

Q RQ4. Le distribuzioni delle variabili sintetiche appartengono a distribuzioni note?

6.3 RQ1: Misure di centralità del dataset sintetico

Per rispondere a tale domanda vengono prese in considerazione le variabili UrlLength, NumDash, NumQueryComponents.

6.3.1 Scostamento delle Misure di Centralità

Per valutare la coerenza delle misure di centralità tra il dataset originale e il dataset sintetico, sono state analizzate la media, la mediana e la moda delle seguenti variabili: **UrlLength**, **NumDash** e **NumQueryComponents**. Di seguito, vengono riportati i risultati del confronto.

Tabella 6.1: Confronto delle Misure di Centralità per UrlLength

Misura	Dataset Originale	Dataset Sintetico	Differenza (%)
Media	70.26	251.91	+258.5%
Mediana	62	248	+300%
Moda	48	174	+262.5%

Come mostrato nella Tabella 6.1, la media, la mediana e la moda del dataset sintetico risultano notevolmente superiori rispetto al dataset originale, suggerendo una generazione di URL molto più lunghi rispetto ai dati reali.

Tabella 6.2: Confronto delle Misure di Centralità per NumDash

Misura	Dataset Originale	Dataset Sintetico	Differenza (%)
Media	1.818	4.500	+147.5%
Mediana	0	4	+∞%
Moda	0	1	+∞%

Dalla Tabella 6.2 emerge che gli URL generati sinteticamente contengono un numero significativamente maggiore di trattini ("."), con una media più che raddoppiata e una mediana che passa da 0 a 4. Questo potrebbe indicare una **distorsione strutturale** nel dataset sintetico.

Tabella 6.3: Confronto delle Misure di Centralità per NumQueryComponents

Misura	Dataset Originale	Dataset Sintetico	Differenza (%)
Media	0.4586	4.5517	+892%
Mediana	0	5	+∞%
Moda	0	5	+∞%

Dalla Tabella 6.3 si nota che gli URL sintetici hanno molte più componenti nella query rispetto ai dati reali, suggerendo un'alterazione nei pattern di generazione.

6.3.2 Conclusioni

L'analisi evidenzia scostamenti significativi tra il dataset sintetico e quello originale:

- **URL più lunghi:** le lunghezze degli URL nel dataset sintetico sono in media **3.6 volte superiori** a quelle reali.
- **Eccesso di trattini:** il numero di trattini è aumentato drasticamente nei dati sintetici, alterando la struttura degli URL.
- **Aumento esponenziale delle query components:** la maggior parte degli URL sintetici contiene numerosi parametri nella query string, mentre nel dataset reale erano quasi assenti.

In sintesi, il dataset sintetico presenta differenze sostanziali nelle misure di centralità rispetto ai dati originali, suggerendo la necessità di una revisione nel processo di generazione per garantire una maggiore fedeltà strutturale ai dati reali.

6.4 RQ2: Metriche di Dispersione

In questa sezione viene analizzata la differenza tra i dati reali e quelli sintetici riguardo le misure di dispersione. L'obiettivo è valutare se i dati generati artificialmente mantengano le stesse caratteristiche numeriche dei dati originali.

6.4.1 Confronto Numerico

La tabella seguente mostra la deviazione standard delle feature selezionate nei dati reali e sintetici.

Feature	Deviazione Standard Reale	Deviazione Standard Sintetica
PctExtNullSelfRedirectHyperlinksRT	0.8978	0.7105
FrequentDomainNameMismatch	0.4110	0.1430
NumDash	3.1063	2.8911
PctNullSelfRedirectHyperlinks	0.3124	0.2898
PctExtHyperlinks	0.3424	0.2887
PathLevel	1.8632	1.4360
UrlLength	33.3699	137.9866
HostnameLength	8.1165	25.9916
EmbeddedBrandName	0.2320	0.1359

Tabella 6.4: Confronto della dispersione tra dati reali e sintetici (Deviazione Standard).

Dalla tabella emerge che le feature sintetiche presentano una dispersione diversa rispetto ai dati reali. Ad esempio, la lunghezza dell'URL (*UrlLength*) ha una deviazione standard significativamente più alta nei dati sintetici rispetto ai dati reali, indicando una maggiore variabilità. Allo stesso modo, *HostnameLength* mostra una differenza marcata tra reale e sintetico, mentre altre feature, come *NumDash* e *PctNullSelfRedirectHyperlinks*, mantengono valori simili.

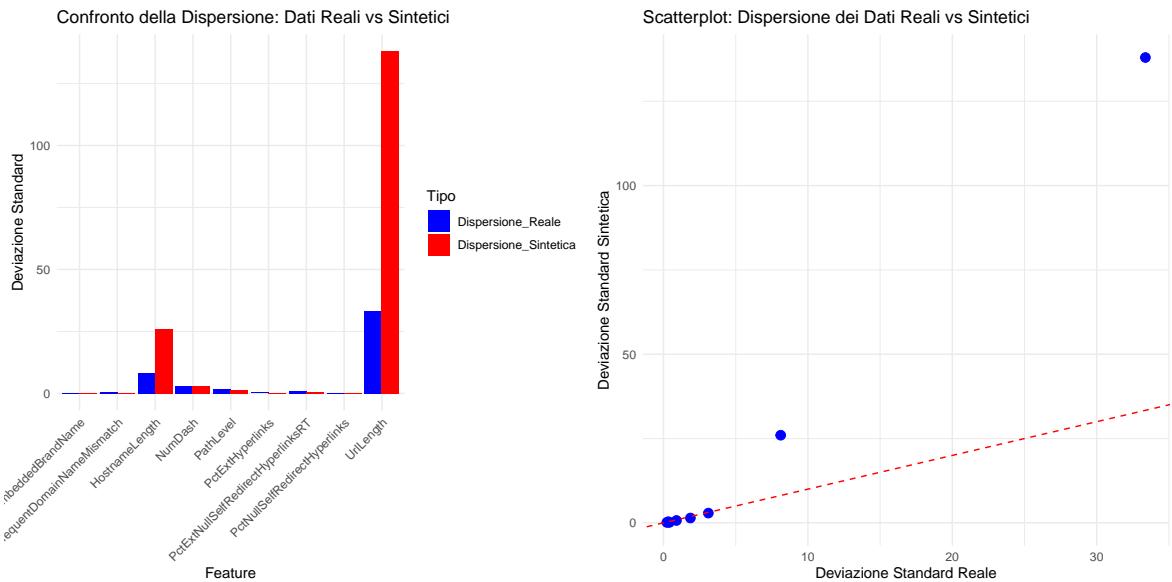


Figura 6.2: Confronto della dispersione dei dati reali e sintetici.

6.4.2 Grafico di confronto

L’istogramma mostra la deviazione standard delle feature nei dati reali e sintetici. Se i dati sintetici riproducessero perfettamente la dispersione dei dati reali, i valori dovrebbero essere molto simili tra loro. Tuttavia, si osservano alcune differenze significative, in particolare per le feature *UrlLength* e *HostnameLength*, il che suggerisce che la variabilità nei dati sintetici non rispecchia esattamente quella dei dati reali.

Il grafico di dispersione mostra che, se i dati sintetici mantenessero la stessa distribuzione dei reali, i punti sarebbero allineati sulla diagonale rossa tratteggiata. Si nota invece una forte dispersione in alcune feature, indicando che la generazione sintetica ha alterato la variabilità dei dati in modo non uniforme.

6.5 RQ3: Differenza nelle matrici di correlazione

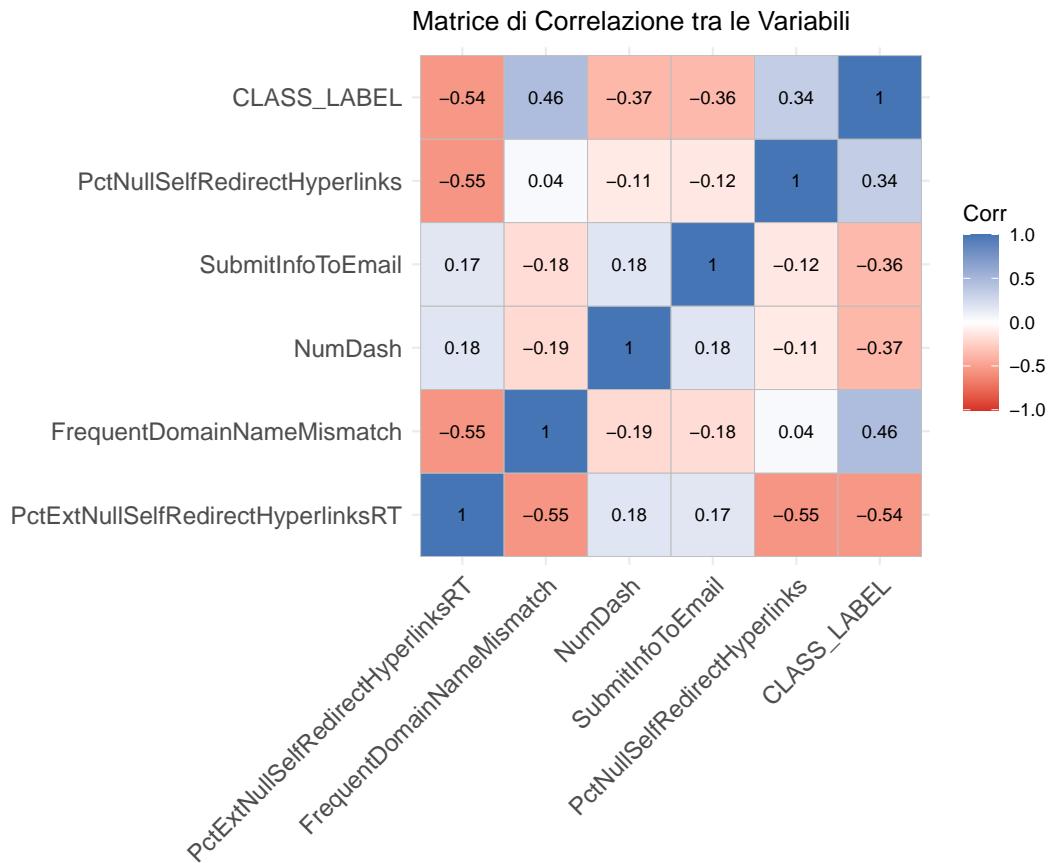


Figura 6.3: Matrice di correlazione dataset originale

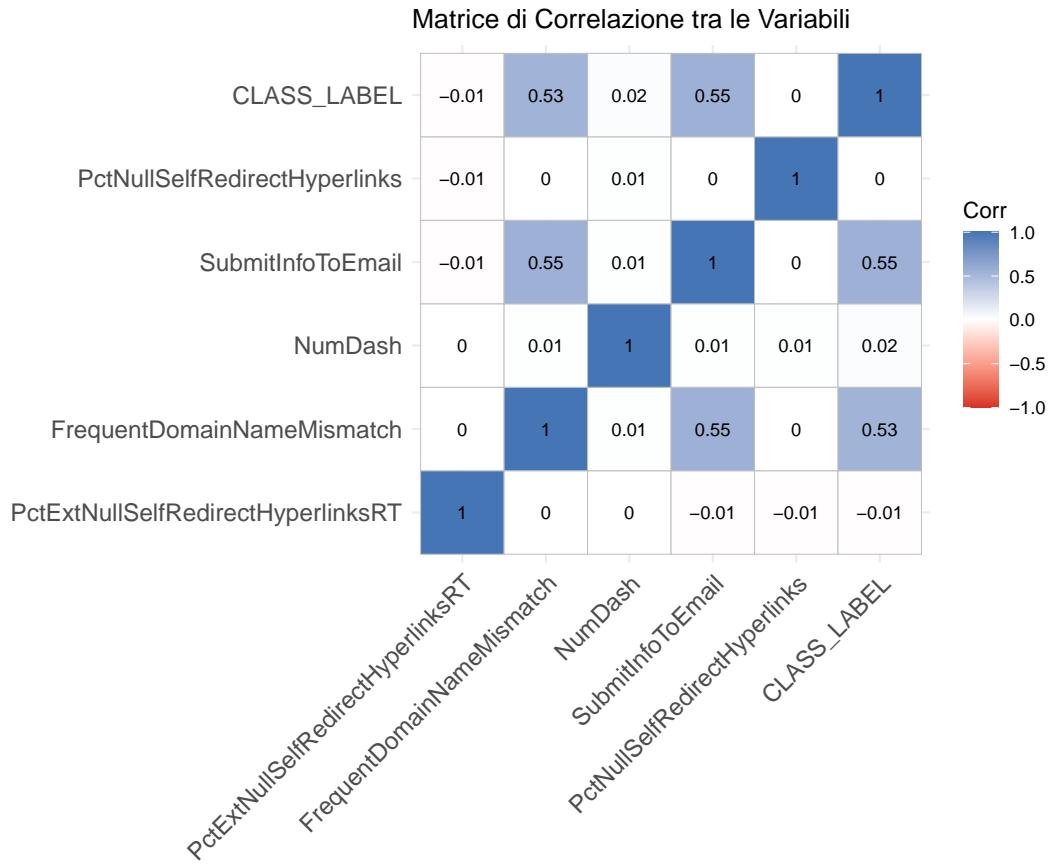


Figura 6.4: Matrice di correlazione dataset sintetico

Le due matrici di correlazione, una derivante dal dataset **originale** e l'altra dal dataset **sintetico**, mostrano differenze significative nelle relazioni tra le variabili. L'obiettivo di questa analisi è valutare in che misura il dataset sintetico ha mantenuto le relazioni statistiche presenti nei dati reali.

6.5.1 Confronto delle Correlazioni

Di seguito viene presentato un confronto tra le correlazioni delle coppie di variabili più rilevanti nei due dataset.

Tabella 6.5: Confronto delle Correlazioni tra Dataset Originale e Sintetico

Coppia di Variabili	Correlazione Dataset Originale	Correlazione Dataset Sintetico	Differenza
PctExtNullSelfRedirectHyperlinksRT - FrequentDomainNameMismatch	-0.55	0	+0.55
PctExtNullSelfRedirectHyperlinksRT - NumDash	0.18	0.01	-0.17
PctExtNullSelfRedirectHyperlinksRT - SubmitInfoToEmail	0.17	-0.01	-0.18
PctExtNullSelfRedirectHyperlinksRT - PctNullSelfRedirectHyperlinks	-0.55	-0.01	+0.54
PctExtNullSelfRedirectHyperlinksRT - CLASS_LABEL	-0.54	-0.01	+0.53
FrequentDomainNameMismatch - NumDash	-0.19	0.01	+0.20
FrequentDomainNameMismatch - SubmitInfoToEmail	-0.18	0.55	+0.73
FrequentDomainNameMismatch - PctNullSelfRedirectHyperlinks	0.04	0	-0.04
FrequentDomainNameMismatch - CLASS_LABEL	0.46	0.53	+0.07
NumDash - SubmitInfoToEmail	0.18	0.01	-0.17
NumDash - CLASS_LABEL	-0.37	0.02	+0.39
SubmitInfoToEmail - CLASS_LABEL	-0.36	0.55	+0.91
PctNullSelfRedirectHyperlinks - CLASS_LABEL	0.34	0	-0.34

6.5.2 Osservazioni Critiche

L’analisi della tabella 6.5 evidenzia diversi problemi nel dataset sintetico:

- **Gravi Perdite di Correlazione:** Alcune variabili che avevano una correlazione negativa forte nel dataset originale (-0.55) ora mostrano correlazioni quasi nulle nel dataset sintetico.
Esempio: La correlazione tra PctExtNullSelfRedirectHyperlinksRT e CLASS_LABEL è passata da -0.54 a -0.01, annullando completamente l’informazione originale.
- **Cambi di Segno nelle Correlazioni:** Alcune correlazioni sono state invertite nel dataset sintetico, il che può influenzare negativamente l’interpretazione dei dati.
Esempio: SubmitInfoToEmail e FrequentDomainNameMismatch passano da -0.18 a +0.55, un’inversione completa.
- **Aumento Drastico di Alcune Correlazioni:** Alcune relazioni hanno subito un aumento eccessivo della correlazione, indicando un possibile pattern artificiale nella generazione dei dati sintetici. **Esempio:** La correlazione tra SubmitInfoToEmail e CLASS_LABEL è passata da -0.36 a +0.55, suggerendo una possibile distorsione nei dati.

- **Correlazioni Residuali Nulle o Trascurabili:** Alcune variabili che mostravano correlazioni significative nel dataset originale ora risultano vicine a zero nel dataset sintetico. Esempio: `PctNullSelfRedirectHyperlinks` aveva una correlazione di 0.34 con `CLASS_LABEL`, ora è 0.

6.5.3 Conclusioni e Soluzioni

L'analisi evidenzia che il dataset sintetico **non ha mantenuto fedelmente le correlazioni chiave** del dataset originale. Questo potrebbe avere un impatto significativo su qualsiasi modello predittivo che utilizza il dataset sintetico, poiché alcune variabili chiave hanno perso le loro relazioni statistiche originali.

6.6 RQ4: Analisi di Distribuzioni note

6.6.1 Obiettivo e Metodologia

L'obiettivo principale di questa analisi consiste nel valutare se la variabile continua `UrlLength` possa essere considerata distribuita uniformemente. Per raggiungere tale scopo, sono stati seguiti i passaggi descritti di seguito:

1. **Suddivisione in classi:** La variabile `UrlLength` è stata suddivisa in quattro classi basate sui principali quantili della distribuzione (Minimo, 1° Quartile, Mediana, 3° Quartile e Massimo). Questa ripartizione consente di individuare eventuali variazioni significative della frequenza dei dati all'interno di intervalli chiave. I valori statistici sono i seguenti:

Minimo: 20.0

1° Quartile (Q1): 132.0

Mediana: 248.0

Media: 251.9

3° Quartile (Q3): 370.0

Massimo: 499.0

Di conseguenza, le quattro classi risultanti sono:

- 20–132
- 133–248
- 249–370
- 371–499

2. **Visualizzazione grafica:** È stata realizzata la rappresentazione grafica della distribuzione (Figura 6.5), dalla quale emerge una forma visivamente simile a una distribuzione uniforme, caratterizzata da frequenze pressoché omogenee nelle quattro classi.

3. Ipotesi di test:

- **Ipotesi nulla (H_0):** la variabile `UrlLength` segue una distribuzione uniforme.
- **Ipotesi alternativa (H_1):** la variabile `UrlLength` non segue una distribuzione uniforme.

4. **Test del Chi-quadrato:** Per verificare la validità dell'ipotesi nulla, è stato applicato il test del Chi-quadrato sulle probabilità attese, confrontando le frequenze osservate in ciascuna classe con quelle teoricamente previste da una distribuzione uniforme.

6.6.2 Risultati

Di seguito sono riportati i risultati del test del Chi-quadrato:

$$X^2 = 0.10535, \text{ df} = 3, p\text{-value} = 0.9912$$

Il valore del *p*-value (0.9912) è nettamente superiore al livello di significatività convenzionale ($\alpha = 0.05$). Pertanto, non vi sono evidenze sufficienti per rifiutare l'ipotesi nulla.

6.6.3 Conclusioni

Sulla base del test statistico e dell'analisi visiva (Figura 6.5), la variabile `UrlLength` mostra un andamento coerente con la distribuzione uniforme. L'elevato *p*-value

evidenzia l'assenza di differenze significative tra frequenze osservate e attese, sostenendo la conclusione che `UrlLength` possa essere considerata uniformemente distribuita.

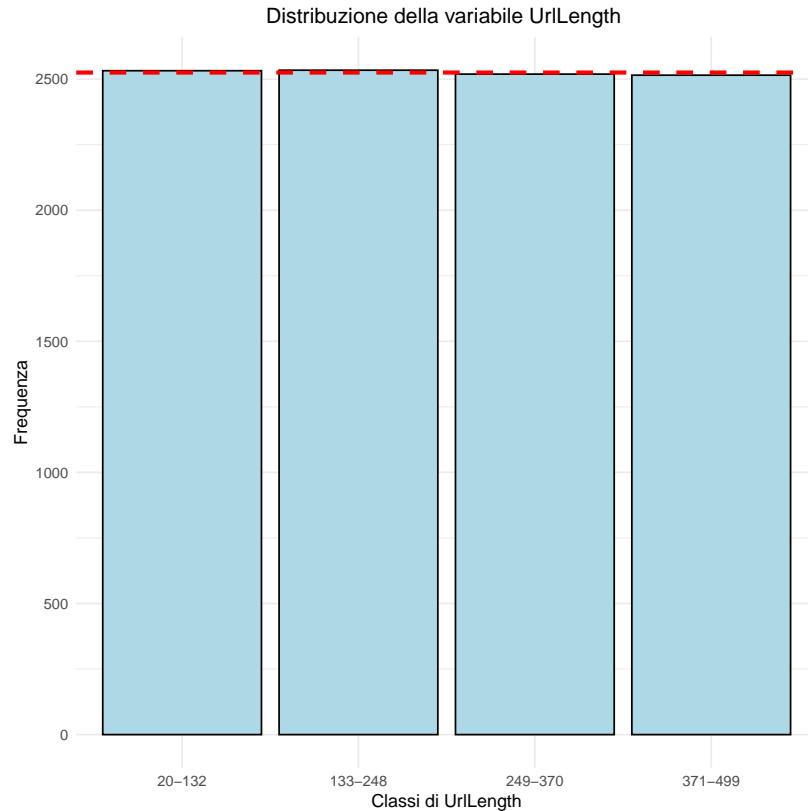


Figura 6.5: Distribuzione della variabile `UrlLength` suddivisa in classi basate sui quantili.

Bibliografia

- [1] Sultan Ahmad, Alimul Haque, Hikmat AM Abdeljaber, MU Bokhari, Jabeen Nazeer, and BK Mishra. Phishing website detection: A dataset-centric approach for enhanced security. 2024. (Citato a pagina 3)
- [2] Kumar Shridhar, Koustuv Sinha, Andrew Cohen, Tianlu Wang, Ping Yu, Ram Pasunuru, Mrinmaya Sachan, Jason Weston, and Asli Celikyilmaz. The art of llm refinement: Ask, refine, and trust. *arXiv preprint arXiv:2311.07961*, 2023. (Citato a pagina 55)
- [3] Julia Wiesinger, Patrick Marlow, and Vladimir Vuskovic. Agents. 2024. (Citato alle pagine 54 e 55)