PROJECTE MINERIA DE DADES

DETECCIÓ DE PHISHING SOBRE PÀGINES WEB

MIDA

24 Octubre 2023

Eric Gonzalez Duro

ÍNDEX

1. Introducció

En un món cada vegada més interconnectat, la seguretat cibernètica s'ha convertit en una preocupació crítica. El phishing, en particular, ha sorgit com una de les formes més efectives en les quals els ciberdelinqüents poden enganyar els usuaris per a obtenir informació personal i financera sensible. Aquests atacs sofisticats i cada vegada més difícils de detectar representen una amenaça constant per a la integritat i la privacitat en línia.

En aquest projecte, s'aborda la tasca de desenvolupar un sistema de detecció de phishing basat en aprenentatge automàtic. S'utilitzarà un conjunt de dades que inclou 11,430 URL amb 87 característiques extretes. Aquestes característiques es divideixen en tres classes: 56 extretes de l'estructura i sintaxi de les URL, 24 extretes del contingut de les pàgines corresponents i 7 extretes mitjançant consultes a serveis externs. El que fa que aquest projecte sigui especialment desafiador és que el conjunt de dades està equilibrat, amb un 50% d'URL de phishing i un 50% d'URL legítimes, la qual cosa reflecteix la realitat de la ciberdelinqüència en línia.

L'objectiu principal d'aquest treball és aplicar tècniques de mineria de dades per a desenvolupar un model capaç de distingir entre URLs legítimes i URLs de phishing amb precisió. Per a aconseguir aquest objectiu, es duran a terme una sèrie de passos crucials que involucren la selecció d'un conjunt de dades no trivial, preprocessament de dades, ajust de paràmetres d'algorismes d'aprenentatge automàtic, interpretació de models i, finalment, una avaluació exhaustiva de diferents enfocaments.

Aquest projecte se centra en comprendre i aplicar el procés complet de mineria de dades, des de la selecció del conjunt de dades fins a la discussió crítica dels resultats obtinguts. La precisió dels models no és l'únic resultat rellevant; la presa de decisions fonamentades i l'avaluació crítica dels resultats exerciran un paper crucial en l'avaluació d'aquest projecte.

En el següent treball, es detallarà el procés complet, des de la selecció del conjunt de dades fins a la discussió de les conclusions finals. S'explicaran els mètodes utilitzats, els enfocaments de preprocessament i ajust de paràmetres, i es

presentaran i analitzaran els resultats obtinguts. Aquest projecte té com a objectiu no sols desenvolupar models efectius de detecció de phishing, sinó també brindar una comprensió profunda de les etapes de la mineria de dades i la lògica darrere de cada decisió presa en el procés.

2. Descripció de les dades originals

El dataset que utilitzarem per a poder realitzar aquesta pràctica, s'ha obtingut del següent link:

https://www.kaggle.com/datasets/shashwatwork/web-page-phishing-detection-dataset/

Tot i que a la descripció del dataset de Kaggle, s'esmenta que la font real es del següent link:

https://data.mendeley.com/datasets/c2gw7fy2j4/3

Per altra banda, el conjunt de dades consta de 11,431 files i 89 columnes. Aquestes columnes representen diverses característiques que s'han extret de les URL i s'utilitzen per a la tasca de detecció de phishing. Les dades inclouen una varietat de tipus de dades com números enters, nombres de coma flotant i indicadors binaris. Cada columna té un significat específic i representa una característica diferent que es pot utilitzar per avaluar la probabilitat que una URL sigui de pishing o legítima. A continuació, es proporciona una descripció general d'aquestes columnes, la qual cosa ajudarà a comprendre millor la naturalesa de les dades originals:

url → La URL completa.
length_url → Longitud de la URL.
length_hostname → Longitud del nom d'amfitrió.
ip → Indicador de si la URL conté una adreça IP en lloc d'un nom d'amfitrió.
nb_dots → Número de punts a la URL.
nb_hyphens → Número de guions a la URL.
nb_at → Número de símbols "@" a la URL.
nb_qm → Número de signes d'interrogació a la URL.
nb_and → Número de signes "&" a la URL.
nb_or → Número de signes "|" a la URL.
nb_eq → Número de signes "=" a la URL.
nb_underscore → Número de guions baixos a la URL.
nb_tilde → Número de virgulilles a la URL.
nb_percent → Número de signes de percentatge a la URL.
nb_slash → Número de barres diagonals a la URL.

nb_star → Número d'asteriscs a la URL.

nb_colon → Número de dos punts a la URL.

nb_comma → Número de comes a la URL.

nb_semicolon → Número de punts i comes a la URL.

nb_dollar → Número de signes de dòlar a la URL.

nb_space → Número d'espais a la URL.

nb_www → Indicador de si "www" està present a la URL.

nb_com → Número de vegades que apareix ".com" a la URL.

nb_dslash → Número de barres diagonals dobles a la URL.

http_in_path → Indicador de si "http" apareix a la ruta de la URL.

https_token → Indicador de si es troba el token "https" a la URL.

ratio_digits_url → Relació entre dígits i caràcters a la URL.

ratio_digits_host → Relació entre dígits i caràcters al nom d'amfitrió.

punycode → Indicador de si s'utilitza punycode a la URL.

port → Port a la URL.

tld_in_path → Indicador de si el domini de nivell superior (TLD) apareix a la ruta de la URL.

tld_in_subdomain → Indicador de si el TLD apareix en un subdomini.

abnormal_subdomain → Indicador de si el subdomini és anormal.

nb_subdomains → Número de subdominis a la URL.

prefix_suffix → Indicador de si s'utilitzen prefixos o sufixos a la URL.

random_domain → Indicador de si el domini sembla ser aleatori.

shortening_service → Indicador de si s'utilitza un servei d'acurtament d'URL.

path_extension → Extensió de la ruta a la URL.

nb_redirection → Número de redireccions a la URL.

nb_external_redirection → Número de redireccions externes a la URL.

length_words_raw → Longitud de les paraules a la URL original.

char_repeat → Indicador de repeticions de caràcters a la URL.

shortest_words_raw → Longitud de la paraula més curta a la URL original.

shortest_word_host → Longitud de la paraula més curta al nom d'amfitrió.

shortest_word_path → Longitud de la paraula més curta a la ruta.

longest_words_raw → Longitud de la paraula més llarga a la URL original.

longest_word_host → Longitud de la paraula més llarga al nom d'amfitrió.

longest_word_path → Longitud de la paraula més llarga a la ruta.

avg_words_raw → Mitjana de la longitud de les paraules a la URL original.
avg_word_host → Mitjana de la longitud de les paraules al nom d'amfitrió.
avg_word_path → Mitjana de la longitud de les paraules a la ruta.
phish_hints → Indicador de pistes de phishing a la URL.
domain_in_brand → Indicador de si el domini està en una marca registrada.
brand_in_subdomain → Indicador de si la marca està en un subdomini.
brand_in_path → Indicador de si la marca està a la ruta.
suspecious_tld → Indicador de domini de nivell superior sospitós (TLD).
statistical_report → Indicador de si es proporciona un informe estadístic.
nb_hyperlinks → Número d'hiperenllaços a la pàgina.
ratio_intHyperlinks → Relació d'hiperenllaços interns a la pàgina.
ratio_extHyperlinks → Relació d'hiperenllaços externs a la pàgina.
ratio_nullHyperlinks → Relació d'hiperenllaços nuls a la pàgina.
nb_extCSS → Número de fulls d'estil enllaçats externament.
ratio_intRedirection → Relació de redireccions internes a la pàgina.
ratio_extRedirection → Relació de redireccions externes a la pàgina.
ratio_intErrors → Relació d'errors interns a la pàgina.
ratio_extErrors → Relació d'errors externs a la pàgina.
login_form → Indicador de si es troba un formulari d'inici de sessió a la pàgina.
external_favicon → Indicador de si s'utilitza un favicon extern a la pàgina.
links_in_tags → Número d'enllaços en etiquetes HTML.
submit_email → Indicador de si s'envia correu electrònic en fer clic a un enllaç.
ratio_intMedia → Relació d'elements multimèdia interns a la pàgina.
ratio_extMedia → Relació d'elements multimèdia externs a la pàgina.
sfh → Indicador de si existeix "same origin" a la pàgina.
iframe → Indicador de si s'utilitzen iframes a la pàgina.
popup_window → Indicador de si s'utilitzen finestres emergents a la pàgina.
safe_anchor → Indicador de si els ancoratges són segurs.
onmouseover → Indicador de si s'utilitza "onmouseover" a la pàgina.
right_clic → Indicador de si es permet fer clic dret a la pàgina.
empty_title → Indicador de si el títol de la pàgina està buit.
domain_in_title → Indicador de si el domini està en el títol de la pàgina.
domain_with_copyright → Indicador de si el domini té drets d'autor.
whois_registered_domain → Indicador de si el domini està registrat.

domain_registration_length → Longitud del registre del domini.
domain_age → Antiguitat del domini.
web_traffic → Trànsit web del lloc.
dns_record → Registre DNS del lloc.
google_index → Índex de Google del lloc.
page_rank → Rang de la pàgina del lloc.
status → Indica si la URL es legítima o bé es phishing

Aquestes característiques es divideixen en tres classes, que inclouen característiques relacionades amb l'estructura i sintaxi de les URL, característiques extretes del contingut de les pàgines corresponents i característiques obtingudes mitjançant consultes a serveis externs.

En l'anàlisi de les dades originals, explorarem més a fons aquestes característiques per a comprendre la seva rellevància en la detecció de phishing. A més, es considerarà la distribució de les classes de phishing i legítimes en el conjunt de dades equilibrat, la qual cosa influeix en l'avaluació dels models i la interpretació dels resulta

3. Pre-Processament de les dades

En abordar la tasca de processament del conjunt de dades "dataset_phishing.csv", el primer pas va ser la seva càrrega i exploració. Aquesta etapa inicial no només ens proporciona una visió general de les dades, sinó que també destaca qualsevol anomalia o peculiaritat que podria requerir atenció. La nostra anàlisi va revelar que el conjunt de dades tenia diverses característiques, majoritàriament numèriques, i no presentava valors faltants. La presència de dades completes simplifica molts passos posteriors, ja que no cal abordar l'imputació o la eliminació de valors faltants.

La columna "url", tot i ser informativa des d'una perspectiva referencial, no aportava valor discriminatiu per als algoritmes de machine learning. La seva naturalesa textual i única per cada fila podria introduir soroll innecessari o sobreajustament en models posteriors. Per això, es va decidir eliminar-la, simplificant així la dimensionalitat de les dades.

La selecció de característiques és un pas crític que pot influir directament en la eficàcia i eficiència dels models de machine learning. En aplicar un anàlisi de correlació, vam observar que algunes característiques tenien valors constants, cosa que implica que no varien i, per tant, no aporten informació per distingir entre URL legítimes i de phishing. La seva eliminació va ser una elecció lògica per reduir la complexitat sense comprometre la qualitat de les dades.

La normalització de les dades és fonamental, especialment quan les característiques tenen diferents unitats o escales. Algoritmes que utilitzen càlculs de distància o gradient, com ara SVM, k-NN o xarxes neuronals, poden ser altament sensibles a les escales de les dades. Al normalitzar, assegurem que cada característica té el mateix pes, optimitzant així el rendiment dels algoritmes.

Finalment, l'aplicació de PCA va ser una decisió estratègica per reduir la dimensionalitat tot mantenint la major part de la variabilitat de les dades. En un conjunt de dades amb moltes característiques, pot haver-hi una gran quantitat d'informació redundant. El PCA permet identificar i mantenir només les dimensions més informatives. El fet que 60 components capturesin el 95% de la variabilitat

indica que moltes de les característiques originals estaven correlacionades i que aquesta representació compacta podria ser igualment efectiva per a tasques de modelització.

En conclusió, cada pas del pre-processament va ser meticulosament considerat i executat amb l'objectiu d'optimitzar la qualitat i eficiència de les dades per a l'anàlisi i modelització posteriors. Aquesta preparació acurada és fonamental per assegurar que qualsevol model de machine learning que es desenvolupi posteriorment tingui una base sòlida sobre la qual operar.

Anàlisi Exploratori Inicial:

Abans de qualsevol pre-processament, és essencial entendre la naturalesa del conjunt de dades. Per fer-ho, hem visualitzat les primeres files i hem explorat les estadístiques descriptives.

```python
data.head()
data.describe([x*0.1 for x in range(10)])
```

Un dels passos clau de la fase d'anàlisi exploratòria és identificar possibles problemes, com ara valors nuls o dades duplicades.

```python
data.isnull().sum()
data.duplicated().sum()
```

Transformació de Característiques:

La columna 'status', que indica si una URL és "legitimate" o "phishing", es va transformar en una representació numèrica per facilitar l'anàlisi quantitatiu.

```python
data['status_numeric'] = data['status'].apply(lambda x: 1 if x == 'phishing' else 0)
```

Aquesta transformació ens va permetre calcular la correlació de cada característica amb la variable objectiu, ajudant-nos a comprendre quines característiques podrien ser més rellevants per a la detecció de phishing.

Preprocessament de Dades:

Per assegurar-nos que les dades estiguin en un format adequat per a l'anàlisi, hem codificat les variables categòriques i hem normalitzat les dades.

La normalització és especialment important quan es treballa amb algoritmes que són sensibles a l'escala de les característiques, com ara la regressió logística o les màquines de vectors suport.

```python
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
data_normalized = scaler.fit_transform(data.select_dtypes(include=[np.number]))
```

Reducció de Dimensions amb PCA:

L'Anàlisi de Components Principals (PCA) és una tècnica que permet reduir la dimensionalitat del conjunt de dades mentre es conserva la major part de la informació.

```python
from sklearn.decomposition import PCA
pca = PCA()
data_pca = pca.fit_transform(data_normalized)
```

Després d'aplicar PCA, vam observar la variància explicada per cada component i vam determinar que 29 components principals eren suficients per capturar almenys el 95% de la variabilitat total de les dades.

```python
variancia_acumulada = pca.explained_variance_ratio_.cumsum()
components_95_variancia = sum(variancia_acumulada < 0.95) + 1
```

Aquesta reducció de dimensions no només simplifica l'anàlisi, sinó que també pot ajudar a millorar la velocitat i l'eficàcia d'alguns algoritmes de mineria de dades.

4. Criteri d'avaluació dels models de mineria de dades

4. 1 Procediment per al desglossament de dades

Tenint en compte que tenim un conjunt de dades equilibrat amb un 50% d'URL de phishing i un 50% d'URL legítims, és essencial assegurar-se que aquest equilibri es manté quan es divideixen les dades en conjunts de formació i validació.

```python
X = data_preprocessada.drop('status', axis=1)
y = data_preprocessada['status']
```

Per a poder assegurar que els conjunts de dades d'entrenament i validació tenen aproximadament el mateix percentatge de mostres de cada classe objectiu que el conjunt de dades complet, aplicarem un mostreig estratificat

Utilitzarem un desglossament 80-20, on el 80% de les dades s'utilitza per a la formació i el 20% restant s'utilitza per a la validació. Això assegura que hi ha una quantitat suficient de dades per entrenar models robusts mentre encara hi ha un conjunt representatiu per a la validació.

```javascript
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
stratify=y, random_state=42)
```

4.2 Mètode d'avaluació del model

Validació creuada: Donada la mida del conjunt de dades (11.430 URLs), és factible utilitzar k-fold validació creuada per avaluar el rendiment del model. La validació creuada ens proporciona una millor estimació del rendiment d'un model en dades no vistes en comparació amb una divisió simple de train-test.

K-Fold Cross-Validation: Pel que fa a k-fold Cross-validation, el conjunt d'entrenament es divideix en "k" conjunts més petits. Per a cada un dels plecs "k", un model s'entrena utilitzant $k - 1$ dels plecs i validat en el plec restant. Aquest procés es repeteix "k" vegades, amb cada plec utilitzat exactament una vegada com a dades de validació. Per a obtenir una única puntuació, es realitza una mitjana amb els resultats donats.

En el nostre cas utilitzarem 10 plecs, ja que es un número recomanat i bastant comú. Per altra banda, com el conjunt de dades està bastant equilibrat, realitzarem una validació creuada k-fold estratificada per assegurar que cada plec sigui un bon representant del conjunt de dades general.

A més, al fer servir StratifiedKFold, ens assegurem que cada divisió del conjunt de dades manté la mateixa proporció de classes que el conjunt de dades original, cosa que és important per a conjunts de dades desequilibrats (no es el nostre cas) o quan la proporció de classes és crítica per a la predicció.

```python
skf = StratifiedKFold(n_splits=10, shuffle=True, random_state=42)
```

Finalment, entrenarem un model de Random Forest. Hem escollit aquest model per les següents raons:

- És robust davant overfitting, especialment quan es tracta de conjunts de dades amb un gran nombre de característiques. En el nostre cas, hem de tenir 87 característiques compte.

- Pot suportar bé les característiques categòriques i numeriques, tot i que en aquest cas ja hem transformat totes les característiques a numèriques.

- És fàcil de comprendre i interpretar, proporcionant la importància de les característiques que ens pot ajudar a entendre quins factors són més rellevants en la predicció de phishing.

- Funciona bé amb dades no balancejades, cosa que no és un problema en el nostre cos però és bo tenir en compte per altres contextos.

- És un model bastant flexible, que pot funcionar bé sense necessitat d'un ajust intensiu d'hiperparàmetres.

```python
rf_classifier = RandomForestClassifier(random_state=42)
```

4.3 Mètriques d'avaluació

Atès que el conjunt de dades és equilibrat, la precisió pot ser una mètrica fiable. No obstant això, en la detecció de phishing, tant els falsos positius (URLs legítims classificats com a phishing) com els falsos negatius (URLs de phishing classificats com a legítims) poden ser problemàtics. Per tant, considerarem mètriques adicionals

Accuracy: És la proporció d'instàncies classificades correctament sobre el total d'instàncies.

$$Accuracy = \frac{Nu\acute{}mero\ d'insta\grave{}ncies\ correctament\ classificades}{N\acute{u}mero\ total\ d'inst\grave{a}ncies}$$

Com que el dataset està equilibrat (50% de URLs de phishing i 50% de URLs legítimes), l'accuracy pot ser una mètrica fiable. No obstant això, en detecció de phishing, és important no confiar únicament en l'accuracy, ja que un model podria simplement classificar tot com a "legítim" i encara aconseguir un 50% d'accuracy.

Recall (Sensibilitat): Mesura quantes de les instàncies positives reals han estat identificades correctament.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

El recall és crític en la detecció de phishing. Un valor baix de recall indicaria que moltes URLs de phishing s'estan classificant erròniament com a legítimes.

F1-Score: És la mitjana harmònica de precisión i recall, i proporciona un equilibri entre ambdues mètriques.

$$F1 = \frac{Precision \times Recall}{Precision + Recall}$$

Donat que tant la precisión com el recall són importants en la detecció de phishing, l'F1-score pot ser una mètrica útil per avaluar un model. Un F1-score elevat indica que el model té un bon equilibri entre precisión i recall.

```python
precision = cross_val_score(rf_classifier, X_train, y_train, cv=skf,
scoring='precision')
recall = cross_val_score(rf_classifier, X_train, y_train, cv=skf,
scoring='recall')
f1_score = cross_val_score(rf_classifier, X_train, y_train, cv=skf,
scoring='f1')
```

Un cop executat el codi anterior, hem obtingut els següents resultats:

- Precisió mitjana de 96.69%: Això significa que quan el model prediu que una URL és de phishing, és correcte aproximadament el 96.69% de les vegades. En termes de phishing, això indica que un percentatge molt alt d'URLs marcades com a malicioses realment són malicioses, el que és positiu ja que redueix el risc d'URLs malicioses no detectades que poden causar danys.

- Recuperació mitjana de 96.35%: Aquesta és la proporció d'URLs de phishing reals que el model ha estat capaç de detectar. Un valor tan alt, ens suggereix que el model és capaç de reconèixer la majoria de les URL de phishing, deixant passar molt poques.

- Puntuació F1 mitjana de 96.52%: La puntuació F1 és la mitjana harmònica de la precisió i la recuperació, i es considera una mesura més robusta quan es tracta de conjunts de dades desequilibrats o quan es vol equilibrar la importància de la precisió i la recuperació. En aquest cas, la puntuació F1 també és alta, indicant que el model té un bon equilibri entre precisió i recuperació.

5. Execució de diferents models de Machine Learning

6. Comparacions i conclusions

OUT OF CONTEXT

This CSV document contains phishing detection data with 1000 rows and 89 columns. The columns represent different features extracted from URLs and their corresponding pages. Here are the descriptions of the headers:

1. 'url': This column contains the URL of the website. It is a string.
2. 'length_url': This column represents the length of the URL. It is an integer.
3. 'length_hostname': This column represents the length of the hostname in the URL. It is an integer.
4. 'ip': This column indicates whether the URL contains an IP address. It is a string.
5. 'nb_dots': This column represents the number of dots in the URL. It is an integer.
6. 'nb_hyphens': This column represents the number of hyphens in the URL. It is an integer.
7. 'nb_at': This column represents the number of '@' symbols in the URL. It is an integer.
8. 'nb_qm': This column represents the number of question marks in the URL. It is an integer.

9. 'nb_and': This column represents the number of ampersands in the URL. It is an integer.

10. 'nb_or': This column represents the number of 'or' keywords in the URL. It is an integer.

11. 'nb_eq': This column represents the number of equal signs in the URL. It is an integer.

12. 'nb_underscore': This column represents the number of underscores in the URL. It is an integer.

13. 'nb_tilde': This column represents the number of tildes in the URL. It is an integer.

14. 'nb_percent': This column represents the number of percent signs in the URL. It is an integer.

15. 'nb_slash': This column represents the number of slashes in the URL. It is an integer.

16. 'nb_star': This column represents the number of asterisks in the URL. It is an integer.

17. 'nb_colon': This column represents the number of colons in the URL. It is an integer.

18. 'nb_comma': This column represents the number of commas in the URL. It is an integer.

19. 'nb_semicolumn': This column represents the number of semicolons in the URL. It is an integer.

20. 'nb_dollar': This column represents the number of dollar signs in the URL. It is an integer.

21. 'nb_space': This column represents the number of spaces in the URL. It is an integer.

22. 'nb_www': This column represents the number of 'www' subdomains in the URL. It is an integer.

23. 'nb_com': This column represents the number of '.com' domains in the URL. It is an integer.

24. 'nb_dslash': This column represents the number of double slashes in the URL. It is an integer.

25. 'http_in_path': This column indicates whether 'http' is present in the URL path. It is a string.

26. 'https_token': This column indicates whether 'https' is present in the URL. It is a string.

27. 'ratio_digits_url': This column represents the ratio of digits in the URL. It is a floating-point number.

28. 'ratio_digits_host': This column represents the ratio of digits in the hostname. It is a floating-point number.

29. 'punycode': This column indicates whether the URL contains punycode. It is a string.

30. 'port': This column represents the port number in the URL. It is an integer.

31. 'tld_in_path': This column indicates whether the top-level domain is present in the URL path. It is a string.

32. 'tld_in_subdomain': This column indicates whether the top-level domain is present in the subdomain. It is a string.

33. 'abnormal_subdomain': This column indicates whether the subdomain is considered abnormal. It is a string.

34. 'nb_subdomains': This column represents the number of subdomains in the URL. It is an integer.

35. 'prefix_suffix': This column indicates whether the URL contains a prefix or suffix. It is a string.

36. 'random_domain': This column indicates whether the domain is randomly generated. It is a string.

37. 'shortening_service': This column indicates whether the URL uses a URL shortening service. It is a string.

38. 'path_extension': This column represents the extension of the URL path. It is a string.

39. 'nb_redirection': This column represents the number of redirections in the URL. It is an integer.

40. 'nb_external_redirection': This column represents the number of external redirections in the URL. It is an integer.

41. 'length_words_raw': This column represents the number of words in the raw URL. It is an integer.

42. 'char_repeat': This column represents the number of repeated characters in the URL. It is an integer.

43. 'shortest_words_raw': This column represents the length of the shortest word in the raw URL. It is an integer.

44. 'shortest_word_host': This column represents the length of the shortest word in the hostname. It is an integer.

45. 'shortest_word_path': This column represents the length of the shortest word in the URL path. It is an integer.

46. 'longest_words_raw': This column represents the length of the longest word in the raw URL. It is an integer.

47. 'longest_word_host': This column represents the length of the longest word in the hostname. It is an integer.

48. 'longest_word_path': This column represents the length of the longest word in the URL path. It is an integer.

49. 'avg_words_raw': This column represents the average number of words in the raw URL. It is a floating-point number.

50. 'avg_word_host': This column represents the average length of words in the hostname. It is a floating-point number.

51. 'avg_word_path': This column represents the average length of words in the URL path. It is a floating-point number.

52. 'phish_hints': This column indicates whether the URL contains phishing hints. It is a string.

53. 'domain_in_brand': This column indicates whether the domain is present in the brand. It is a string.

54. 'brand_in_subdomain': This column indicates whether the brand is present in the subdomain. It is a string.

55. 'brand_in_path': This column indicates whether the brand is present in the URL path. It is a string.

56. 'suspecious_tld': This column indicates whether the top-level domain is suspicious. It is a string.

57. 'statistical_report': This column indicates whether the URL contains a statistical report. It is a string.

58. 'nb_hyperlinks': This column represents the number of hyperlinks in the URL. It is an integer.

59. 'ratio_intHyperlinks': This column represents the ratio of internal hyperlinks in the URL. It is a floating-point number.

60. 'ratio_extHyperlinks': This column represents the ratio of external hyperlinks in the URL. It is a floating-point number.

61. 'ratio_nullHyperlinks': This column represents the ratio of null hyperlinks in the URL. It is a floating-point number.

62. 'nb_extCSS': This column represents the number of external CSS files in the URL. It is an integer.

63. 'ratio_intRedirection': This column represents the ratio of internal redirections in the URL. It is a floating-point number.

64. 'ratio_extRedirection': This column represents the ratio of external redirections in the URL. It is a floating-point number.

65. 'ratio_intErrors': This column represents the ratio of internal errors in the URL. It is a floating-point number.

66. 'ratio_extErrors': This column represents the ratio of external errors in the URL. It is a floating-point number.

67. 'login_form': This column indicates whether the URL contains a login form. It is a string.

68. 'external_favicon': This column indicates whether the URL uses an external favicon. It is a string.

69. 'links_in_tags': This column indicates whether there are links in HTML tags. It is a string.

70. 'submit_email': This column indicates whether the URL contains an email submission form. It is a string.

71. 'ratio_intMedia': This column represents the ratio of internal media files in the URL. It is a floating-point number.

72. 'ratio_extMedia': This column represents the ratio of external media files in the URL. It is a floating-point number.

73. 'sfh': This column indicates whether the URL uses a server form handler. It is a string.

74. 'iframe': This column indicates whether the URL contains an iframe. It is a string.

75. 'popup_window': This column indicates whether the URL contains a popup window. It is a string.

76. 'safe_anchor': This column indicates whether the URL uses a safe anchor. It is a string.

77. 'onmouseover': This column indicates whether the URL uses the onmouseover event. It is a string.

78. 'right_clic': This column indicates whether the URL uses the right-click event. It is a string.

79. 'empty_title': This column indicates whether the URL has an empty title. It is a string.

80. 'domain_in_title': This column indicates whether the domain is present in the title. It is a string.

81. 'domain_with_copyright': This column indicates whether the domain contains a copyright symbol. It is a string.

82. 'whois_registered_domain': This column indicates whether the domain is registered. It is a string.

83. 'domain_registration_length': This column represents the length of domain registration. It is an integer.

84. 'domain_age': This column represents the age of the domain. It is an integer.

85. 'web_traffic': This column represents the web traffic of the domain. It is a string.

86. 'dns_record': This column indicates whether the domain has a DNS record. It is a string.

87. 'google_index': This column indicates whether the domain is indexed by Google. It is a string.

88. 'page_rank': This column represents the page rank of the domain. It is a string.

89. 'status': This column indicates whether the URL is legitimate or phishing. It is a string.

The data types used in this CSV file are strings, integers, and floats.