In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split
```

Dataset Analysis

This dataset contains 48 features extracted from 5000 phishing webpages and 5000 legitimate webpages, which were downloaded from January to May 2015 and from May to June 2017. An improved feature extraction technique is employed by leveraging the browser automation framework (i.e., Selenium WebDriver), which is more precise and robust compared to the parsing approach based on regular expressions.

In [2]:
```
df = pd.read_csv('./dataset/Phishing_Legitimate_full 2.csv')
df.head()
```

Out[2]:

	id	NumDots	SubdomainLevel	PathLevel	UrlLength	NumDash	NumDashInHostname	AtSymbol	TildeSymbol	NumUnderscore	...	IframeOrFrame	MissingTitle	ImagesOnlyInForm	SubdomainLevelRT	UrlLengthRT	PctExtResourceUrlsRT	AbnormalExtFormActionR	ExtMetaScriptLinkRT	PctExtNullSelfR
0	1	3	1	5	72	0	0	0	0	0	...	0	0	1	1	0	1	1	-1	
1	2	3	1	3	144	0	0	0	0	2	...	0	0	0	1	-1	1	1	1	
2	3	3	1	2	58	0	0	0	0	0	...	0	0	0	1	0	-1	1	-1	
3	4	3	1	6	79	1	0	0	0	0	...	0	0	0	1	-1	1	1	1	
4	5	3	0	4	46	0	0	0	0	0	...	1	0	0	1	1	-1	0	-1	

5 rows × 50 columns

In [3]:
```
df.shape
```

Out[3]:
```
(10000, 50)
```

In [4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 50 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   id                                  10000 non-null  int64
 1   NumDots                             10000 non-null  int64
 2   SubdomainLevel                      10000 non-null  int64
 3   PathLevel                           10000 non-null  int64
 4   UrlLength                           10000 non-null  int64
 5   NumDash                             10000 non-null  int64
 6   NumDashInHostname                   10000 non-null  int64
 7   AtSymbol                            10000 non-null  int64
 8   TildeSymbol                         10000 non-null  int64
 9   NumUnderscore                       10000 non-null  int64
 10  NumPercent                          10000 non-null  int64
 11  NumQueryComponents                  10000 non-null  int64
 12  NumAmpersand                        10000 non-null  int64
 13  NumHash                             10000 non-null  int64
 14  NumNumericChars                     10000 non-null  int64
 15  NoHttps                             10000 non-null  int64
 16  RandomString                        10000 non-null  int64
 17  IpAddress                           10000 non-null  int64
 18  DomainInSubdomains                  10000 non-null  int64
 19  DomainInPaths                       10000 non-null  int64
 20  HttpsInHostname                     10000 non-null  int64
 21  HostnameLength                      10000 non-null  int64
 22  PathLength                          10000 non-null  int64
 23  QueryLength                         10000 non-null  int64
 24  DoubleSlashInPath                   10000 non-null  int64
 25  NumSensitiveWords                   10000 non-null  int64
 26  EmbeddedBrandName                   10000 non-null  int64
 27  PctExtHyperlinks                    10000 non-null  float64
 28  PctExtResourceUrls                  10000 non-null  float64
 29  ExtFavicon                          10000 non-null  int64
 30  InsecureForms                       10000 non-null  int64
 31  RelativeFormAction                  10000 non-null  int64
 32  ExtFormAction                       10000 non-null  int64
 33  AbnormalFormAction                  10000 non-null  int64
 34  PctNullSelfRedirectHyperlinks       10000 non-null  float64
 35  FrequentDomainNameMismatch          10000 non-null  int64
 36  FakeLinkInStatusBar                 10000 non-null  int64
 37  RightClickDisabled                  10000 non-null  int64
 38  PopUpWindow                         10000 non-null  int64
 39  SubmitInfoToEmail                   10000 non-null  int64
 40  IframeOrFrame                       10000 non-null  int64
 41  MissingTitle                        10000 non-null  int64
 42  ImagesOnlyInForm                    10000 non-null  int64
 43  SubdomainLevelRT                    10000 non-null  int64
 44  UrlLengthRT                         10000 non-null  int64
 45  PctExtResourceUrlsRT                10000 non-null  int64
 46  AbnormalExtFormActionR              10000 non-null  int64
 47  ExtMetaScriptLinkRT                 10000 non-null  int64
 48  PctExtNullSelfRedirectHyperlinksRT  10000 non-null  int64
 49  CLASS_LABEL                         10000 non-null  int64
dtypes: float64(3), int64(47)
memory usage: 3.8 MB
```

In [5]:
```
df.describe()
```

Out[5]:

	id	NumDots	SubdomainLevel	PathLevel	UrlLength	NumDash	NumDashInHostname	AtSymbol	TildeSymbol	NumUnderscore	...	IframeOrFrame	MissingTitle	ImagesOnlyInForm	SubdomainLevelRT	UrlLengthRT	PctExtResourceUrlsRT	AbnormalExtFormActio
count	10000.00000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	...	10000.000000	10000.00000	10000.000000	10000.000000	10000.000000	10000.000000	10000.0000
mean	5000.50000	2.445100	0.586800	3.300300	70.264100	1.818000	0.138900	0.000300	0.013100	0.32320	...	0.339600	0.03220	0.030400	0.956600	0.020200	0.353300	0.7932
std	2886.89568	1.346836	0.751214	1.863241	33.369877	3.106258	0.545744	0.017319	0.113709	1.11466	...	0.473597	0.17654	0.171694	0.248037	0.820036	0.888908	0.521(
min	1.00000	1.000000	0.000000	0.000000	12.000000	0.000000	0.000000	0.000000	0.000000	0.00000	...	0.000000	0.00000	0.000000	-1.000000	-1.000000	-1.000000	-1.0000
25%	2500.75000	2.000000	0.000000	2.000000	48.000000	0.000000	0.000000	0.000000	0.000000	0.00000	...	0.000000	0.00000	0.000000	1.000000	-1.000000	-1.000000	1.0000
50%	5000.50000	2.000000	1.000000	3.000000	62.000000	0.000000	0.000000	0.000000	0.000000	0.00000	...	0.000000	0.00000	0.000000	1.000000	0.000000	1.000000	1.0000
75%	7500.25000	3.000000	1.000000	4.000000	84.000000	2.000000	0.000000	0.000000	0.000000	0.00000	...	1.000000	0.00000	0.000000	1.000000	1.000000	1.000000	1.0000
max	10000.00000	21.000000	14.000000	18.000000	253.000000	55.000000	9.000000	1.000000	1.000000	18.00000	...	1.000000	1.00000	1.000000	1.000000	1.000000	1.000000	1.0000

8 rows × 50 columns

In [6]:
```
df.drop("id", axis=1, inplace=True)
df.head(1)
```

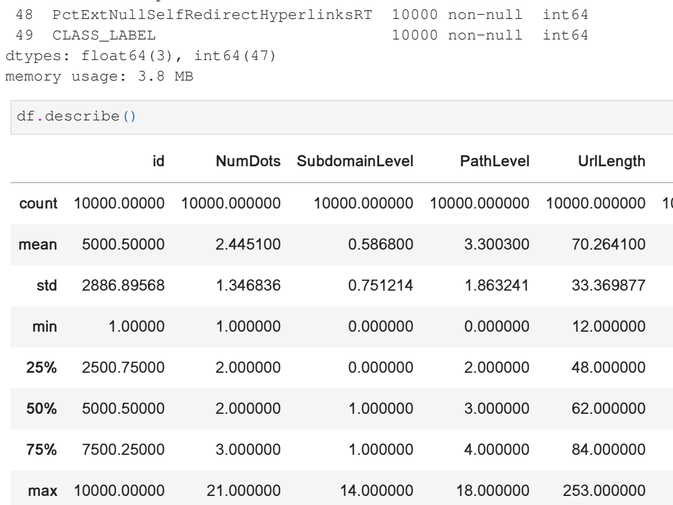
Out[6]:

	NumDots	SubdomainLevel	PathLevel	UrlLength	NumDash	NumDashInHostname	AtSymbol	TildeSymbol	NumUnderscore	NumPercent	...	IframeOrFrame	MissingTitle	ImagesOnlyInForm	SubdomainLevelRT	UrlLengthRT	PctExtResourceUrlsRT	AbnormalExtFormActionR	ExtMetaScriptLinkRT	PctE
0	3	1	5	72	0	0	0	0	0	0	...	0	0	1	1	0	1	1	-1	

1 rows × 49 columns

In [7]:
```
df['CLASS_LABEL'].hist()
```

Out[7]:
```
<Axes: >
```



Splitting data

In [8]:
```
X = df.drop("CLASS_LABEL", axis=1)
y = df["CLASS_LABEL"]
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.2, random_state=42)
X_train.shape
```

Out[8]:
```
(8000, 48)
```

Training Random Forest Algorithm

In [9]:
```
from sklearn.ensemble import RandomForestClassifier
```

In [10]:
```
rfc = RandomForestClassifier()
rfc.fit(X_train, y_train)
```

Out[10]:
```
RandomForestClassifier()
```

Model evaluation metrics

In [11]:
```
prediction = rfc.predict(X_test)
```

In [12]:
```
from sklearn.metrics import accuracy_score, precision_score, recall_score
```

In [13]:
```
accuracy_score(y_test, prediction)
```

Out[13]:
```
0.984
```

In [14]:
```
precision_score(y_test, prediction)
```

Out[14]:
```
0.9822834645669292
```

In [15]:
```
recall_score(y_test, prediction)
```

Out[15]: 0.9861660079051383

In [ ]: