CodeB_Internship / modell.ipynb

MontyVasita week 4 done 41e23f4 · 1 minute ago

5624 lines (5624 loc) · 1.41 MB

Monty K Vasita

Setting up libraries, logging, and pandas display — no data insights yet, just environment setup

In [1]:
```
# Import Data Manipulation Libraries
import numpy as np
import pandas as pd

# Import Data Visualization Libraries
import matplotlib.pyplot as plt
import seaborn as sns

# Import Filter Warning Libraries
import warnings
warnings.filterwarnings('ignore')

# Import Logging  Files
import logging

logging.basicConfig(
    level=logging.INFO,
    filemode='w',
    filename='app.log',
    format='%(asctime)s - %(levelname)s - %(message)s')
```

In [2]:
```
pd.set_option("display.max_columns", None)
pd.set_option("display.max_rows", 100)
```

Loading Data Set

In [3]:
```
# DataSet
url="https://raw.githubusercontent.com/MontyVasita18/CodeB_Internship/refs/he
df=pd.read_csv(url)
df.sample(frac=1) # To make the code execution faster
```

Out[3]:

	url	length_url	length_hostname
6121	https://www.tumblr.com/safe-mode?url=http%3A%2...	73	14
2935	http://microsoft-secure-online.oa.r.appspot.co...	53	40
2328	http://www.sloaneandhyde.com/imm/new2015/pvali...	55	21
8609	http://articles.extension.org/pages/26436/ways...	88	22
10199	http://108.166.202.103/pc/	26	15
...	...	...	...
3229	http://cns-international2.com/s.htm	35	22
1360	http://brighant.com/1122/?sec=Jochen%20Kuntermann	49	12
5336	https://elexusgirisim1.blogspot.com/	36	27
10478	http://docuelectronicsignatureadmin.weebly.com/	47	39
8687	http://www.hoursmap.com/s/ohio/save-a-lot-hour...	66	16

11430 rows × 89 columns

📊 Exploratory Data Analysis (EDA)

Checking shape, data types, and missing values. The dataset looks clean, with no major null-value issues. That means no need for imputation or heavy cleaning

In [4]:
```
df.shape
```

Out[4]: (11430, 89)

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11430 entries, 0 to 11429
Data columns (total 89 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   url                         11430 non-null  object 
 1   length_url                  11430 non-null  int64  
 2   length_hostname             11430 non-null  int64  
 3   ip                          11430 non-null  int64  
 4   nb_dots                     11430 non-null  int64  
 5   nb_hyphens                  11430 non-null  int64  
 6   nb_at                       11430 non-null  int64  
 7   nb_qm                       11430 non-null  int64  
 8   nb_and                      11430 non-null  int64  
 9   nb_or                       11430 non-null  int64  
 10  nb_eq                       11430 non-null  int64  
 11  nb_underscore               11430 non-null  int64  
 12  nb_tilde                    11430 non-null  int64  
 13  nb_percent                  11430 non-null  int64  
 14  nb_slash                    11430 non-null  int64  
 15  nb_star                     11430 non-null  int64  
 16  nb_colon                    11430 non-null  int64  
 17  nb_comma                    11430 non-null  int64  
 18  nb_semicolumn               11430 non-null  int64  
 19  nb_dollar                   11430 non-null  int64  
 20  nb_space                    11430 non-null  int64  
 21  nb_www                      11430 non-null  int64  
 22  nb_com                      11430 non-null  int64  
 23  nb_dslash                   11430 non-null  int64  
 24  http_in_path                11430 non-null  int64  
 25  https_token                 11430 non-null  int64  
 26  ratio_digits_url            11430 non-null  float64
 27  ratio_digits_host           11430 non-null  float64
 28  punycode                    11430 non-null  int64  
 29  port                        11430 non-null  int64  
 30  tld_in_path                 11430 non-null  int64  
 31  tld_in_subdomain            11430 non-null  int64  
 32  abnormal_subdomain          11430 non-null  int64  
 33  nb_subdomains               11430 non-null  int64  
 34  prefix_suffix               11430 non-null  int64  
 35  random_domain               11430 non-null  int64  
 36  shortening_service          11430 non-null  int64  
 37  path_extension              11430 non-null  int64  
 38  nb_redirection              11430 non-null  int64  
 39  nb_external_redirection     11430 non-null  int64  
 40  length_words_raw            11430 non-null  int64  
 41  char_repeat                 11430 non-null  int64  
 42  shortest_words_raw          11430 non-null  int64  
 43  shortest_word_host          11430 non-null  int64  
 44  shortest_word_path          11430 non-null  int64  
 45  longest_words_raw           11430 non-null  int64  
 46  longest_word_host           11430 non-null  int64  
 47  longest_word_path           11430 non-null  int64  
 48  avg_words_raw               11430 non-null  float64
 49  avg_word_host               11430 non-null  float64
 50  avg_word_path               11430 non-null  float64
 51  phish_hints                 11430 non-null  int64  
 52  domain_in_brand             11430 non-null  int64  
 53  brand_in_subdomain          11430 non-null  int64
```

```
 54  brand_in_path               11430 non-null  int64
 55  suspecious_tld              11430 non-null  int64
 56  statistical_report          11430 non-null  int64
 57  nb_hyperlinks               11430 non-null  int64
 58  ratio_intHyperlinks         11430 non-null  float64
 59  ratio_extHyperlinks         11430 non-null  float64
 60  ratio_nullHyperlinks        11430 non-null  int64
 61  nb_extCSS                   11430 non-null  int64
 62  ratio_intRedirection        11430 non-null  int64
 63  ratio_extRedirection        11430 non-null  float64
 64  ratio_intErrors             11430 non-null  int64
 65  ratio_extErrors             11430 non-null  float64
 66  login_form                  11430 non-null  int64
 67  external_favicon            11430 non-null  int64
 68  links_in_tags               11430 non-null  float64
 69  submit_email                11430 non-null  int64
 70  ratio_intMedia              11430 non-null  float64
 71  ratio_extMedia              11430 non-null  float64
 72  sfh                         11430 non-null  int64
 73  iframe                      11430 non-null  int64
 74  popup_window                11430 non-null  int64
 75  safe_anchor                 11430 non-null  float64
 76  onmouseover                 11430 non-null  int64
 77  right_clic                  11430 non-null  int64
 78  empty_title                 11430 non-null  int64
 79  domain_in_title             11430 non-null  int64
 80  domain_with_copyright       11430 non-null  int64
 81  whois_registered_domain     11430 non-null  int64
 82  domain_registration_length  11430 non-null  int64
 83  domain_age                  11430 non-null  int64
 84  web_traffic                 11430 non-null  int64
 85  dns_record                  11430 non-null  int64
 86  google_index                11430 non-null  int64
 87  page_rank                   11430 non-null  int64
 88  status                      11430 non-null  object
dtypes: float64(13), int64(74), object(2)
memory usage: 7.8+ MB
```

In [6]:

```
# Checking Null Value in DataSet
df.isnull().sum()/len(df)*100
```

Out[6]:

```
url                          0.0
length_url                   0.0
length_hostname              0.0
ip                           0.0
nb_dots                      0.0
nb_hyphens                   0.0
nb_at                        0.0
nb_qm                        0.0
nb_and                       0.0
nb_or                        0.0
nb_eq                        0.0
nb_underscore                0.0
nb_tilde                     0.0
nb_percent                   0.0
nb_slash                     0.0
nb_star                      0.0
nb_colon                     0.0
nb_comma                     0.0
nb_semicolumn                0.0
nb_dollar                    0.0
nb_space                     0.0
nb_www                       0.0
nb_com                       0.0
nb_dslash                    0.0
http_in_path                 0.0
https_token                  0.0
ratio_digits_url             0.0
ratio_digits_host            0.0
punycode                     0.0
port                         0.0
tld_in_path                  0.0
tld_in_subdomain             0.0
abnormal_subdomain           0.0
```

```
nb_subdomains                0.0
prefix_suffix                0.0
random_domain                0.0
shortening_service           0.0
path_extension               0.0
nb_redirection               0.0
nb_external_redirection      0.0
length_words_raw             0.0
char_repeat                  0.0
shortest_words_raw           0.0
shortest_word_host           0.0
shortest_word_path           0.0
longest_words_raw            0.0
longest_word_host            0.0
longest_word_path            0.0
avg_words_raw                0.0
avg_word_host                0.0
avg_word_path                0.0
phish_hints                  0.0
domain_in_brand              0.0
brand_in_subdomain           0.0
brand_in_path                0.0
suspecious_tld               0.0
statistical_report           0.0
nb_hyperlinks                0.0
ratio_intHyperlinks          0.0
ratio_extHyperlinks          0.0
ratio_nullHyperlinks         0.0
nb_extCSS                    0.0
ratio_intRedirection         0.0
ratio_extRedirection         0.0
ratio_intErrors              0.0
ratio_extErrors              0.0
login_form                   0.0
external_favicon             0.0
links_in_tags                0.0
submit_email                 0.0
ratio_intMedia               0.0
ratio_extMedia               0.0
sfh                          0.0
iframe                       0.0
popup_window                 0.0
safe_anchor                  0.0
onmouseover                  0.0
right_clic                   0.0
empty_title                  0.0
domain_in_title              0.0
domain_with_copyright        0.0
whois_registered_domain      0.0
domain_registration_length   0.0
domain_age                   0.0
web_traffic                  0.0
dns_record                   0.0
google_index                 0.0
page_rank                    0.0
status                       0.0
dtype: float64
```

Summary statistics show how features are spread. You can spot skewed distributions and outliers (e.g., long URLs, high digit ratio).

In [7]:
```
df.describe()
```

Out[7]:

	length_url	length_hostname	ip	nb_dots	nb_hyphens	
count	11430.000000	11430.000000	11430.000000	11430.000000	11430.000000	114
mean	61.126684	21.090289	0.150569	2.480752	0.997550	
std	55.297318	10.777171	0.357644	1.369686	2.087087	
min	12.000000	4.000000	0.000000	1.000000	0.000000	
25%	33.000000	15.000000	0.000000	2.000000	0.000000	

50%	47.000000	19.000000	0.000000	2.000000	0.000000
75%	71.000000	24.000000	0.000000	3.000000	1.000000
max	1641.000000	214.000000	1.000000	24.000000	43.000000

Separating numerical and categorical columns. Then, for each numeric feature, you analyze spread, skewness, and outliers — very helpful for choosing scaling techniques or detecting which features might need transformation.

In [8]:
```
# Spliting data into Numerical Data and Catagorical Data
numerical_data=df.select_dtypes(exclude='object')
numerical_data

categorical_data=df.select_dtypes(include='object')
```

In [9]:
```
from collections import OrderedDict
stats=[]

for col in df.columns:
    if df[col].dtype !='object':
        numerical_stats=OrderedDict({
            'Feature': col,
            'Minimum': df[col].min(),
            'Maximum': df[col].max(),
            'Mean': df[col].mean(),
            'Mode': df[col].mode()[0] if not df[col].mode().empty else None,
            '25%': df[col].quantile(0.25),
            '75%': df[col].quantile(0.75),
            'IQR': df[col].quantile(0.75) - df[col].quantile(0.25),
            'Standard Deviation': df[col].std(),
            'Skewness': df[col].skew(),
            'Kurtosis': df[col].kurt()


        })
        stats.append(numerical_stats)
report=pd.DataFrame(stats)
report
```

Out[9]:

	Feature	Minimum	Maximum	Mean	Mode	2
0	length_url	12.0	1.641000e+03	61.126684	26.0	33.000
1	length_hostname	4.0	2.140000e+02	21.090289	16.0	15.000
2	ip	0.0	1.000000e+00	0.150569	0.0	0.000
3	nb_dots	1.0	2.400000e+01	2.480752	2.0	2.000
4	nb_hyphens	0.0	4.300000e+01	0.997550	0.0	0.000
5	nb_at	0.0	4.000000e+00	0.022222	0.0	0.000
6	nb_qm	0.0	3.000000e+00	0.141207	0.0	0.000
7	nb_and	0.0	1.900000e+01	0.162292	0.0	0.000
8	nb_or	0.0	0.000000e+00	0.000000	0.0	0.000
9	nb_eq	0.0	1.900000e+01	0.293176	0.0	0.000
10	nb_underscore	0.0	1.800000e+01	0.322660	0.0	0.000
11	nb_tilde	0.0	1.000000e+00	0.006649	0.0	0.000
12	nb_percent	0.0	9.600000e+01	0.123097	0.0	0.000
13	nb_slash	2.0	3.300000e+01	4.289589	3.0	3.000
14	nb_star	0.0	1.000000e+00	0.000700	0.0	0.000

15	nb_colon	1.0	7.000000e+00	1.027909	1.0	1.000
16	nb_comma	0.0	4.000000e+00	0.004024	0.0	0.000
17	nb_semicolumn	0.0	2.000000e+01	0.062292	0.0	0.000
18	nb_dollar	0.0	6.000000e+00	0.001925	0.0	0.000
19	nb_space	0.0	1.800000e+01	0.034821	0.0	0.000
20	nb_www	0.0	2.000000e+00	0.448469	0.0	0.000
21	nb_com	0.0	6.000000e+00	0.127997	0.0	0.000
22	nb_dslash	0.0	1.000000e+00	0.006562	0.0	0.000
23	http_in_path	0.0	4.000000e+00	0.016710	0.0	0.000
24	https_token	0.0	1.000000e+00	0.610936	1.0	0.000
25	ratio_digits_url	0.0	7.238806e-01	0.053137	0.0	0.000
26	ratio_digits_host	0.0	8.000000e-01	0.025024	0.0	0.000
27	punycode	0.0	1.000000e+00	0.000350	0.0	0.000
28	port	0.0	1.000000e+00	0.002362	0.0	0.000
29	tld_in_path	0.0	1.000000e+00	0.065617	0.0	0.000
30	tld_in_subdomain	0.0	1.000000e+00	0.050131	0.0	0.000
31	abnormal_subdomain	0.0	1.000000e+00	0.021610	0.0	0.000
32	nb_subdomains	1.0	3.000000e+00	2.231671	2.0	2.000
33	prefix_suffix	0.0	1.000000e+00	0.202450	0.0	0.000
34	random_domain	0.0	1.000000e+00	0.083290	0.0	0.000
35	shortening_service	0.0	1.000000e+00	0.123447	0.0	0.000
36	path_extension	0.0	1.000000e+00	0.000175	0.0	0.000
37	nb_redirection	0.0	6.000000e+00	0.498250	0.0	0.000
38	nb_external_redirection	0.0	1.000000e+00	0.003150	0.0	0.000
39	length_words_raw	1.0	1.060000e+02	6.232808	2.0	2.000
40	char_repeat	0.0	1.460000e+02	2.927472	3.0	1.000
41	shortest_words_raw	1.0	3.100000e+01	3.127297	3.0	2.000
42	shortest_word_host	1.0	3.900000e+01	5.019773	3.0	3.000
43	shortest_word_path	0.0	4.000000e+01	2.398950	0.0	0.000
44	longest_words_raw	2.0	8.290000e+02	15.393876	9.0	9.000
45	longest_word_host	1.0	6.200000e+01	10.467979	9.0	7.000
46	longest_word_path	0.0	8.290000e+02	10.561505	0.0	0.000
47	avg_words_raw	2.0	1.282500e+02	7.258882	6.0	5.250
48	avg_word_host	1.0	3.900000e+01	7.678075	5.0	5.250
49	avg_word_path	0.0	2.500000e+02	5.092425	0.0	0.000
50	phish_hints	0.0	1.000000e+01	0.327734	0.0	0.000
51	domain_in_brand	0.0	1.000000e+00	0.104199	0.0	0.000
52	brand_in_subdomain	0.0	1.000000e+00	0.004112	0.0	0.000
53	brand_in_path	0.0	1.000000e+00	0.004899	0.0	0.000
54	suspecious_tld	0.0	1.000000e+00	0.017935	0.0	0.000
55	statistical_report	0.0	2.000000e+00	0.059755	0.0	0.000

56	nb_hyperlinks	0.0	4.659000e+03	87.189764	0.0	9.000
57	ratio_intHyperlinks	0.0	1.000000e+00	0.602457	0.0	0.224
58	ratio_extHyperlinks	0.0	1.000000e+00	0.276720	0.0	0.000
59	ratio_nullHyperlinks	0.0	0.000000e+00	0.000000	0.0	0.000
60	nb_extCSS	0.0	1.240000e+02	0.784864	0.0	0.000
61	ratio_intRedirection	0.0	0.000000e+00	0.000000	0.0	0.000
62	ratio_extRedirection	0.0	2.000000e+00	0.158926	0.0	0.000
63	ratio_intErrors	0.0	0.000000e+00	0.000000	0.0	0.000
64	ratio_extErrors	0.0	1.000000e+00	0.062469	0.0	0.000
65	login_form	0.0	1.000000e+00	0.063605	0.0	0.000
66	external_favicon	0.0	1.000000e+00	0.442170	0.0	0.000
67	links_in_tags	0.0	1.000000e+02	51.978211	0.0	0.000
68	submit_email	0.0	0.000000e+00	0.000000	0.0	0.000
69	ratio_intMedia	0.0	1.000000e+02	42.870444	0.0	0.000
70	ratio_extMedia	0.0	1.000000e+02	23.236293	0.0	0.000
71	sfh	0.0	0.000000e+00	0.000000	0.0	0.000
72	iframe	0.0	1.000000e+00	0.001312	0.0	0.000
73	popup_window	0.0	1.000000e+00	0.006037	0.0	0.000
74	safe_anchor	0.0	1.000000e+02	37.063922	0.0	0.000
75	onmouseover	0.0	1.000000e+00	0.001137	0.0	0.000
76	right_clic	0.0	1.000000e+00	0.001400	0.0	0.000
77	empty_title	0.0	1.000000e+00	0.124759	0.0	0.000
78	domain_in_title	0.0	1.000000e+00	0.775853	1.0	1.000
79	domain_with_copyright	0.0	1.000000e+00	0.439545	0.0	0.000
80	whois_registered_domain	0.0	1.000000e+00	0.072878	0.0	0.000
81	domain_registration_length	-1.0	2.982900e+04	492.532196	0.0	84.000
82	domain_age	-12.0	1.287400e+04	4062.543745	-1.0	972.250
83	web_traffic	0.0	1.076799e+07	856756.643307	0.0	0.000
84	dns_record	0.0	1.000000e+00	0.020122	0.0	0.000
85	google_index	0.0	1.000000e+00	0.533946	1.0	0.000
86	page_rank	0.0	1.000000e+01	3.185739	0.0	1.000

Several features showed significant skewness, suggesting non-normal distributions.

Wide ranges and high standard deviations in some columns (e.g., web_traffic, length_url) indicate the presence of outliers.

Features with high kurtosis are likely to have heavy tails or sharp peaks.

Checking frequency counts for categorical columns — this helps you see whether categories are balanced or dominated by one class (like the target label status).

In [10]:
```
# Frequency Distribution
for col in df.columns:
    if df[col].dtype=='object':
```

```
print(f"Frequency Distribution Of {col}\n")
print(df[col].value_counts)
```

```
Frequency Distribution Of url

<bound method IndexOpsMixin.value_counts of 0                    http://www.cre
stonwood.com/router.php
1        http://shadetreetechnology.com/V4/validation/a...
2        https://support-appleld.com.secureupdate.duila...
3                                      http://rgipt.ac.in
4        http://www.iracing.com/tracks/gateway-motorspo...
                              ...
11425        http://www.fontspace.com/category/blackletter
11426    http://www.budgetbots.com/server.php/Server%20...
11427    https://www.facebook.com/Interactive-Televisio...
11428              http://www.mypublicdomainpictures.com/
11429    http://174.139.46.123/ap/signin?openid.pape.ma...
Name: url, Length: 11430, dtype: object>
Frequency Distribution Of status

<bound method IndexOpsMixin.value_counts of 0        legitimate
1          phishing
2          phishing
3        legitimate
4        legitimate
            ...
11425    legitimate
11426      phishing
11427    legitimate
11428    legitimate
11429      phishing
Name: status, Length: 11430, dtype: object>
```

In [11]:
```
df['status'].value_counts()
```

Out[11]:
```
status
legitimate    5715
phishing      5715
Name: count, dtype: int64
```

The target column status is well-balanced, which is ideal for binary classification models and ensures fair learning across both classes.

In [12]:
```
df['status'].mode()
```

Out[12]:
```
0    legitimate
1      phishing
Name: status, dtype: object
```

Label encoding turns 'legitimate' and 'phishing' into 0 and 1 — readying the target for machine learning models.

In [13]:
```
# Encoding Target column
df['status']=df['status'].replace({'legitimate':0,'phishing':1})
df['status']
```

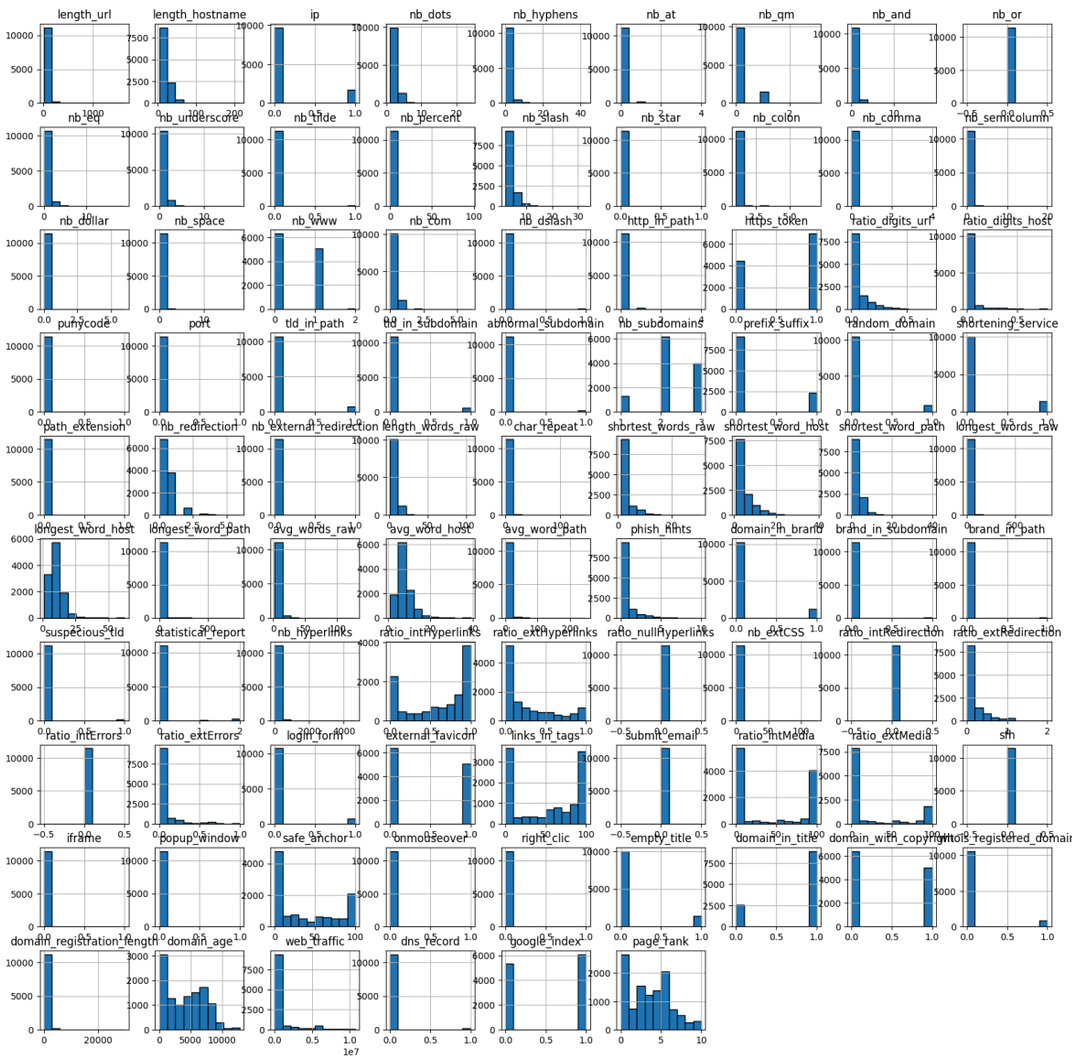
Out[13]:
```
0        0
1        1
2        1
3        0
4        0
        ..
11425    0
11426    1
11427    0
11428    0
11429    1
Name: status, Length: 11430, dtype: int64
```

The target variable status was originally categorical, labeled as “phishing” and “legitimate.” It was converted into a binary format (1 and 0) for model compatibility.

Histogram

Histograms Reveal skewed features and possible outliers. Some features like web_traffic or length_url may need scaling or normalization.

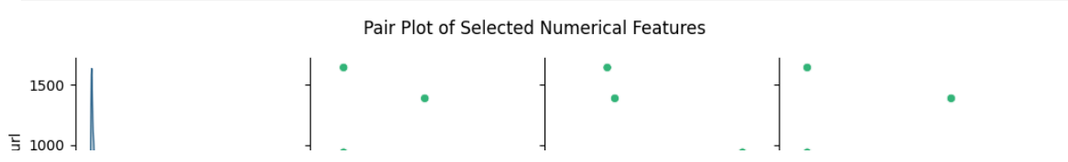
In [14]:
```
# Plotting Histogram
numerical_data.hist(figsize=(20,20),bins=10,edgecolor='black')
plt.title('Histogram example',y=1.02)
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.show()
```

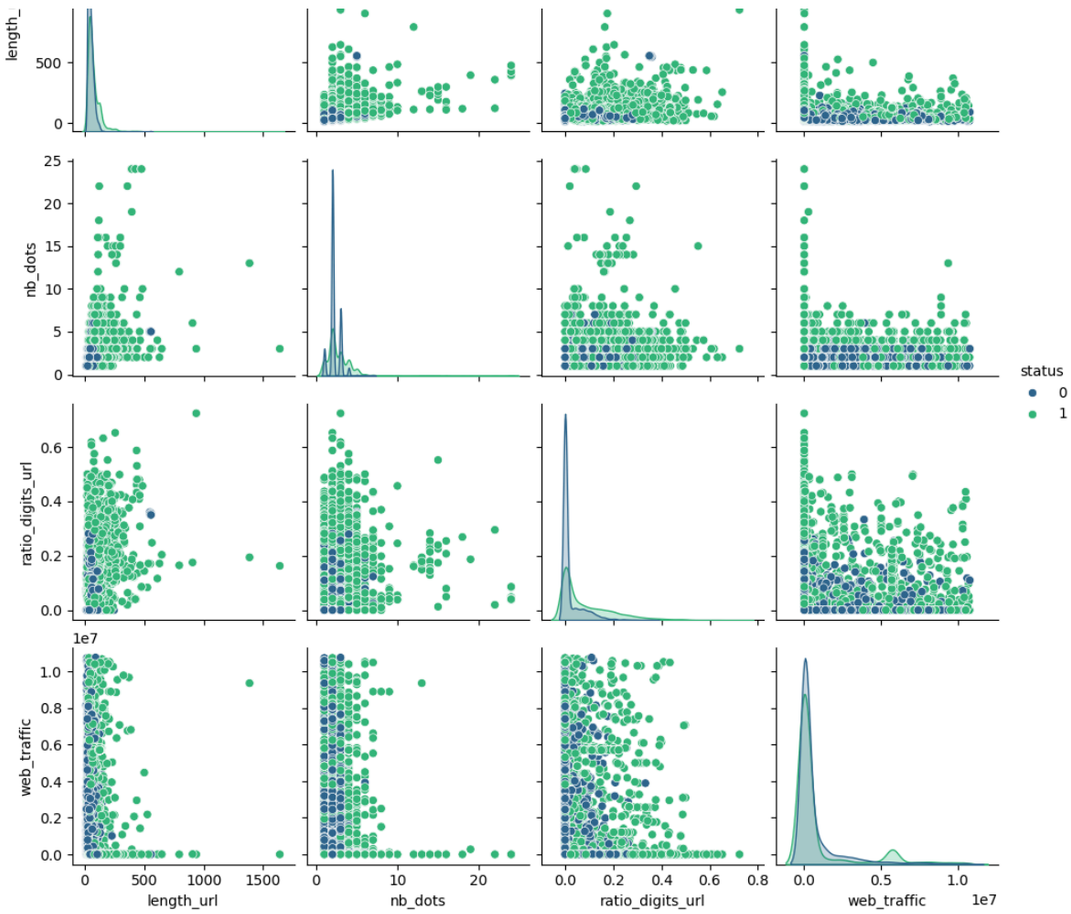


Many features are right-skewed, indicating potential preprocessing needs (e.g., log transformation). Distribution plots also highlighted high concentration of values in specific ranges for features like ratio_digits_url.

Pair Plot

In [15]:
```
selected_features = ['length_url', 'nb_dots', 'ratio_digits_url', 'web_traffi
# Plot pair plot
sns.pairplot(df[selected_features], hue='status', palette='viridis')
# Optional: Add title
plt.suptitle("Pair Plot of Selected Numerical Features", y=1.02)
plt.show()
```





The pairplot shows some visual separation between phishing and legitimate classes in selected features — especially in ratio_digits_url and web_traffic. That means these features might be strong indicators for classification.

✅ Insights & Recommendations

Key Findings:

Several numerical features display non-normal distributions and contain outliers, which could affect model performance if not addressed.

Features like ratio_digits_url and web_traffic show clear separation between classes and can act as strong indicators for phishing detection.

The target column status is well-balanced, which is ideal for binary classification models and ensures fair learning across both classes.

Recommended Actions:

Normalize or transform skewed numerical features (e.g., using log or power transforms) to reduce the effect of extreme values.

Scale features using standardization (e.g., MinMaxScaler or StandardScaler) to ensure uniform treatment by algorithms.

Use feature selection techniques (e.g., correlation thresholding, mutual information, or tree-based feature importance) to focus on the most predictive variables.

Check for multicollinearity using correlation matrices or VIF to avoid redundant features

Checking duplicates

Checking duplicates

In [16]:
```
duplicates=df.duplicated()
```

In [17]:
```
duplicates.value_counts()
```

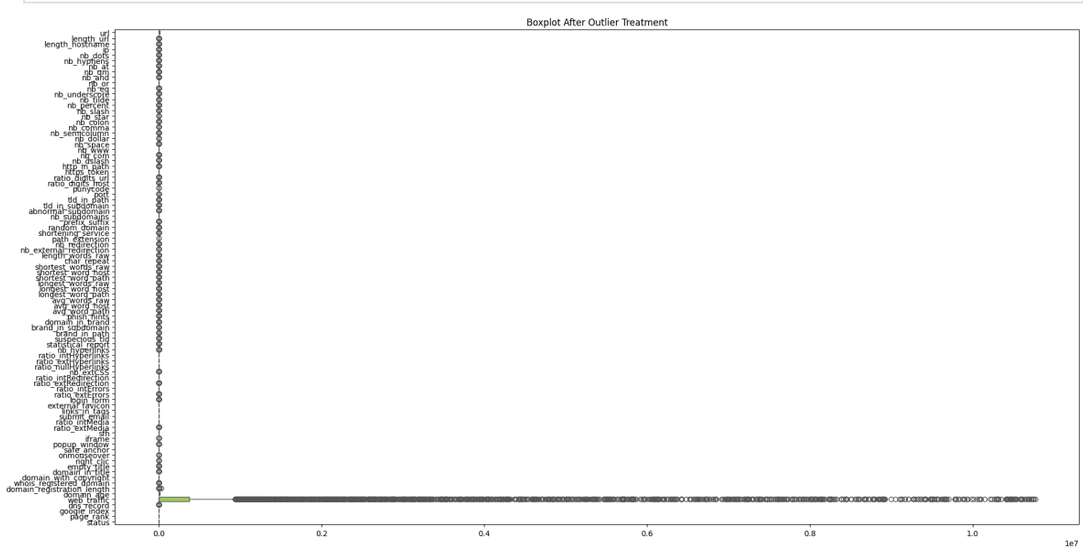
Out[17]:
```
False    11430
Name: count, dtype: int64
```

Label Encoding was applied to the url column to convert categorical values into numeric form. One-Hot Encoding was avoided because it would have significantly increased the number of columns due to the high number of unique URLs. Label Encoding keeps the dataset compact and efficient without adding unnecessary dimensions.

In [18]:
```
# Label Encoding Url Column

from sklearn.preprocessing import LabelEncoder
LE=LabelEncoder()
df['url']=LE.fit_transform(df['url'])


df['url'].value_counts()
```

Out[18]:
```
url
1065     2
4501     1
10779    1
1315     1
9201     1
        ..
6539     1
819      1
9629     1
5956     1
62       1
Name: count, Length: 11429, dtype: int64
```

In [19]:
```
# Checking Outliers Using Boxplot
# Set figure size
plt.figure(figsize=(20, 10))

# Create boxplot for all numerical columns
sns.boxplot(data=df, orient='h', palette='Set2')

# Set title
plt.title('Boxplot After Outlier Treatment')
plt.tight_layout()
plt.show()
```



In [33]:
```
# Checking Correlation
df.corr()['status']
```

Out[33]:
```
url                          -2.909714e-01
length_url                    2.485805e-01
length_hostname               2.383224e-01
ip                            3.216978e-01
nb_dots                       2.070288e-01
nb_hyphens                   -1.001075e-01
nb_at                         1.429146e-01
nb_qm                         2.943191e-01
nb_and                        1.705464e-01
nb_or                                  NaN
nb_eq                         2.333863e-01
nb_underscore                 3.809134e-02
nb_tilde                      3.014233e-02
nb_percent                    2.810129e-02
nb_slash                      2.422700e-01
nb_star                       2.646512e-02
nb_colon                      9.283531e-02
nb_comma                      1.186465e-02
nb_semicolumn                 1.035541e-01
nb_dollar                     2.496206e-02
nb_space                     -4.193222e-03
nb_www                       -4.434677e-01
nb_com                        1.562835e-01
nb_dslash                     7.260234e-02
http_in_path                  7.077624e-02
https_token                   1.146691e-01
ratio_digits_url              3.563946e-01
ratio_digits_host             2.243349e-01
punycode                      1.871039e-02
port                          9.011116e-03
tld_in_path                   7.914651e-02
tld_in_subdomain              2.088842e-01
abnormal_subdomain            1.281598e-01
nb_subdomains                 1.128907e-01
prefix_suffix                 2.146807e-01
random_domain                 1.963062e-02
shortening_service            1.061200e-01
path_extension                5.592660e-17
nb_redirection               -2.440520e-02
nb_external_redirection       5.620994e-02
length_words_raw              1.920105e-01
char_repeat                   1.473217e-02
shortest_words_raw           -3.936361e-02
shortest_word_host            2.230840e-01
shortest_word_path            7.436495e-02
longest_words_raw             2.001466e-01
longest_word_host             1.245156e-01
longest_word_path             2.127091e-01
avg_words_raw                 1.675637e-01
avg_word_host                 1.935017e-01
avg_word_path                 1.972561e-01
phish_hints                   3.353927e-01
domain_in_brand              -9.822216e-02
brand_in_subdomain            6.425702e-02
brand_in_path                 6.515575e-02
suspecious_tld                1.100896e-01
statistical_report            1.439435e-01
nb_hyperlinks                -3.426283e-01
ratio_intHyperlinks          -2.439821e-01
ratio_extHyperlinks           8.335725e-02
ratio_nullHyperlinks                   NaN
nb_extCSS                    -8.356663e-02
ratio_intRedirection                   NaN
ratio_extRedirection         -1.508267e-01
ratio_intErrors                        NaN
ratio_extErrors              -3.470251e-02
login_form                   -1.900010e-02
external_favicon             -1.465654e-01
links_in_tags                -1.844011e-01
submit_email                           NaN
ratio_intMedia               -1.933331e-01
ratio_extMedia               -1.404059e-01
```

```
sfh                            NaN
iframe                   -1.208332e-02
popup_window             -5.760197e-02
safe_anchor              -1.733973e-01
onmouseover              -7.787061e-03
right_clic                4.680056e-03
empty_title               2.070428e-01
domain_in_title           3.428070e-01
domain_with_copyright    -1.730985e-01
whois_registered_domain   6.697907e-02
domain_registration_length -1.617188e-01
domain_age               -3.318891e-01
web_traffic               6.038772e-02
dns_record                1.221190e-01
google_index              7.311708e-01
page_rank                -5.111371e-01
status                    1.000000e+00
Name: status, dtype: float64
```

A Ranked list of features based on Variance Inflation Factor (VIF)

In [31]:

```
from statsmodels.stats.outliers_influence import variance_inflation_factor
# Checking VIF:
def calculate_vif(dataset):
    vif = pd.DataFrame()
    vif['features'] = dataset.columns
    vif['VIF_Values'] = [variance_inflation_factor(dataset.values,i) for i in
    vif['VIF_Values'] = round(vif['VIF_Values'], 2)
    vif = vif.sort_values(by = 'VIF_Values', ascending=False)
    return (vif)

calculate_vif(df.drop('status',axis = 1))
```

Out[31]:

	features	VIF_Values
49	avg_word_host	278.79
45	longest_words_raw	150.19
40	length_words_raw	144.10
47	longest_word_path	130.30
46	longest_word_host	127.15
48	avg_words_raw	92.81
43	shortest_word_host	51.16
14	nb_slash	45.65
4	nb_dots	34.09
33	nb_subdomains	33.03
16	nb_colon	29.59
0	url	28.20
1	length_url	25.48
50	avg_word_path	25.29
58	ratio_intHyperlinks	21.28
2	length_hostname	19.04
10	nb_eq	14.34
25	https_token	14.33
8	nb_and	12.27

42	shortest_words_raw	11.80
5	nb_hyphens	11.15
68	links_in_tags	8.07
87	page_rank	7.48
59	ratio_extHyperlinks	7.34
21	nb_www	6.31
79	domain_in_title	5.99
13	nb_percent	5.14
83	domain_age	5.08
26	ratio_digits_url	4.94
7	nb_qm	4.17
86	google_index	4.16
11	nb_underscore	4.02
3	ip	4.01
44	shortest_word_path	3.86
70	ratio_intMedia	3.73
27	ratio_digits_host	3.73
67	external_favicon	3.29
78	empty_title	3.07
75	safe_anchor	3.00
31	tld_in_subdomain	2.74
24	http_in_path	2.59
71	ratio_extMedia	2.52
22	nb_com	2.38
41	char_repeat	2.29
80	domain_with_copyright	2.21
32	abnormal_subdomain	2.17
52	domain_in_brand	2.05
51	phish_hints	1.97
38	nb_redirection	1.95
30	tld_in_path	1.87
18	nb_semicolumn	1.86
34	prefix_suffix	1.78
57	nb_hyperlinks	1.71
85	dns_record	1.69
82	domain_registration_length	1.66
63	ratio_extRedirection	1.66
39	nb_external_redirection	1.64
36	shortening_service	1.57
23	nb_dslash	1.52
56	statistical_report	1.51

84	web_traffic	1.47
54	brand_in_path	1.37
65	ratio_extErrors	1.34
61	nb_extCSS	1.33
81	whois_registered_domain	1.32
6	nb_at	1.30
35	random_domain	1.20
66	login_form	1.16
20	nb_space	1.15
29	port	1.14
53	brand_in_subdomain	1.14
55	suspecious_tld	1.10
12	nb_tilde	1.06
19	nb_dollar	1.05
17	nb_comma	1.04
76	onmouseover	1.04
15	nb_star	1.03
74	popup_window	1.02
28	punycode	1.02
77	right_clic	1.01
73	iframe	1.01
37	path_extension	1.00
9	nb_or	NaN
60	ratio_nullHyperlinks	NaN
62	ratio_intRedirection	NaN
64	ratio_intErrors	NaN
69	submit_email	NaN
72	sfh	NaN

In [20]:
```
# Spliting Data into Independent And target Column
X=df.drop(columns='status')
y=df['status']
```

In [21]:
```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(X,y,train_size=0.70,random_sta
```

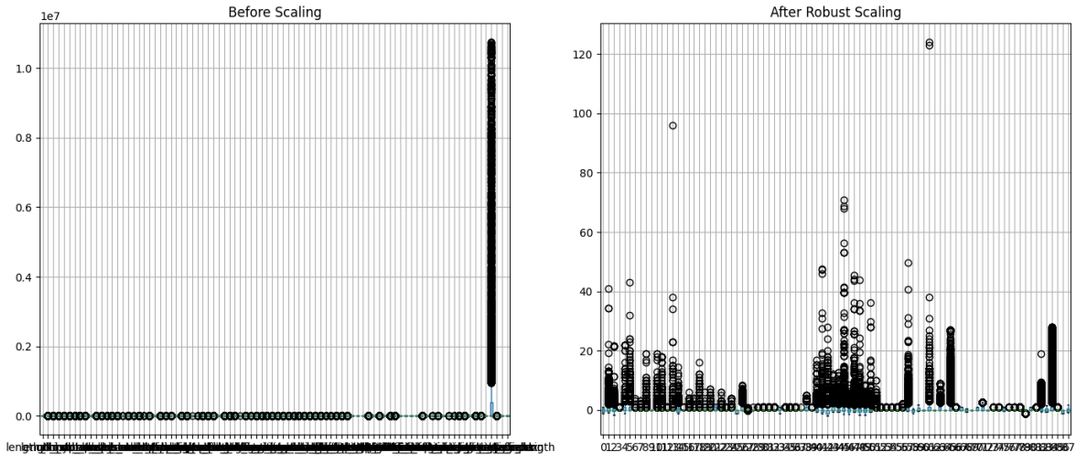
In [22]:
```
X_train_original = X_train.copy()
```

Scaling Technique:- Robust Scaler

Robust Scaler was used to handle outliers effectively, as boxplots showed many extreme values in the numerical features. It scales data based on the median and IQR, making it less sensitive to outliers compared to StandardScaler or MinMaxScaler.

In [23]:
```
from sklearn.preprocessing import MinMaxScaler,StandardScaler,RobustScaler
scaler=RobustScaler()
X_train=scaler.fit_transform(X_train)
X_test=scaler.transform(X_test)
```

In [24]:
```
import pandas as pd
import matplotlib.pyplot as plt
X_train_scaled=X_train.copy()
# If X_train is a NumPy array, convert it to a DataFrame
X_train_df = pd.DataFrame(X_train_original)
X_train_scaled_df = pd.DataFrame(X_train_scaled)

# Plot before and after scaling side by side
plt.figure(figsize=(14, 6))

plt.subplot(1, 2, 1)
X_train_df.boxplot()
plt.title("Before Scaling")

plt.subplot(1, 2, 2)
X_train_scaled_df.boxplot()
plt.title("After Robust Scaling")

plt.tight_layout()
plt.show()
```



In [ ]:

In [ ]: