```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confu

df = pd.read_csv("dataset_phishing.csv")

# Step 2: EDA
print(df.shape)
print(df.info())
print(df.describe())
print(df['status'].value_counts())
print(df.columns.tolist())
```

```
(11430, 89)
<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  int64
dtypes: float64(13), int64(75), object(1)
memory usage: 7.8+ MB
None
          length_url   length_hostname              ip       nb_dots  \
count   11430.000000     11430.000000   11430.000000   11430.000000
mean       61.126684        21.090289       0.150569       2.480752
std        55.297318        10.777171       0.357644       1.369686
min        12.000000         4.000000       0.000000       1.000000
25%        33.000000        15.000000       0.000000       2.000000
50%        47.000000        19.000000       0.000000       2.000000
75%        71.000000        24.000000       0.000000       3.000000
max      1641.000000       214.000000       1.000000      24.000000

          nb_hyphens          nb_at         nb_qm         nb_and     nb_or  \
count   11430.000000   11430.000000  11430.000000   11430.000000   11430.0
mean        0.997550       0.022222      0.141207       0.162292       0.0
std         2.087087       0.155500      0.364456       0.821337       0.0
min         0.000000       0.000000      0.000000       0.000000       0.0
25%         0.000000       0.000000      0.000000       0.000000       0.0
50%         0.000000       0.000000      0.000000       0.000000       0.0
75%         1.000000       0.000000      0.000000       0.000000       0.0
max        43.000000       4.000000      3.000000      19.000000       0.0

               nb_eq  ...  domain_in_title   domain_with_copyright  \
count   11430.000000  ...      11430.000000            11430.000000
```

```
mean       0.293176  ...         0.775853               0.439545
std        0.998317  ...         0.417038               0.496353
min        0.000000  ...         0.000000               0.000000
25%        0.000000  ...         1.000000               0.000000
50%        0.000000  ...         1.000000               0.000000
75%        0.000000  ...         1.000000               1.000000
max       19.000000  ...         1.000000               1.000000

       whois_registered_domain   domain_registration_length     domain_age  \
count              11430.000000                11430.000000   11430.000000
mean                   0.072878                  492.532196    4062.543745
std                    0.259948                  814.769415    3107.784600
min                    0.000000                   -1.000000     -12.000000
25%                    0.000000                   84.000000     972.250000
50%                    0.000000                  242.000000    3993.000000
75%                    0.000000                  449.000000    7026.750000
max                    1.000000                29829.000000   12874.000000

         web_traffic     dns_record   google_index     page_rank         status
count   1.143000e+04   11430.000000   11430.000000  11430.000000   11430.000000
mean    8.567566e+05       0.020122       0.533946      3.185739       0.500000
std     1.995606e+06       0.140425       0.498868      2.536955       0.500022
min     0.000000e+00       0.000000       0.000000      0.000000       0.000000
25%     0.000000e+00       0.000000       0.000000      1.000000       0.000000
50%     1.651000e+03       0.000000       1.000000      3.000000       0.500000
75%     3.738455e+05       0.000000       1.000000      5.000000       1.000000
max     1.076799e+07       1.000000       1.000000     10.000000       1.000000

[8 rows x 88 columns]
0    5715
1    5715
Name: status, dtype: int64
['url', 'length_url', 'length_hostname', 'ip', 'nb_dots', 'nb_hyphens', 'nb_at', 'nb_qm',
'nb_and', 'nb_or', 'nb_eq', 'nb_underscore', 'nb_tilde', 'nb_percent', 'nb_slash', 'nb_sta
r', 'nb_colon', 'nb_comma', 'nb_semicolumn', 'nb_dollar', 'nb_space', 'nb_www', 'nb_com',
'nb_dslash', 'http_in_path', 'https_token', 'ratio_digits_url', 'ratio_digits_host', 'puny
code', 'port', 'tld_in_path', 'tld_in_subdomain', 'abnormal_subdomain', 'nb_subdomains',
'prefix_suffix', 'random_domain', 'shortening_service', 'path_extension', 'nb_redirectio
n', 'nb_external_redirection', 'length_words_raw', 'char_repeat', 'shortest_words_raw', 's
hortest_word_host', 'shortest_word_path', 'longest_words_raw', 'longest_word_host', 'longe
st_word_path', 'avg_words_raw', 'avg_word_host', 'avg_word_path', 'phish_hints', 'domain_i
n_brand', 'brand_in_subdomain', 'brand_in_path', 'suspecious_tld', 'statistical_report',
'nb_hyperlinks', 'ratio_intHyperlinks', 'ratio_extHyperlinks', 'ratio_nullHyperlinks', 'nb
_extCSS', 'ratio_intRedirection', 'ratio_extRedirection', 'ratio_intErrors', 'ratio_extErr
ors', 'login_form', 'external_favicon', 'links_in_tags', 'submit_email', 'ratio_intMedia',
'ratio_extMedia', 'sfh', 'iframe', 'popup_window', 'safe_anchor', 'onmouseover', 'right_cl
ic', 'empty_title', 'domain_in_title', 'domain_with_copyright', 'whois_registered_domain',
'domain_registration_length', 'domain_age', 'web_traffic', 'dns_record', 'google_index',
'page_rank', 'status']
```

```python
# Unstack and sort correlation pairs

corr = df.corr()
high_corr = corr[(corr > 0.8) | (corr < -0.8)]
high_corr_pairs = high_corr.unstack().dropna().sort_values(ascending=False)
print("\nTop correlated feature pairs (abs > 0.8):")
display(high_corr_pairs)
```

```
Top correlated feature pairs (abs > 0.8):
```

```
length_url             length_url               1.000000
brand_in_subdomain     brand_in_subdomain       1.000000
ratio_extRedirection   ratio_extRedirection     1.000000
nb_extCSS              nb_extCSS                1.000000
ratio_extHyperlinks    ratio_extHyperlinks      1.000000
                                                  ...
nb_and                 nb_eq                    0.906404
longest_word_host      avg_word_host            0.816313
avg_word_host          longest_word_host        0.816313
shortest_word_host     avg_word_host            0.800014
avg_word_host          shortest_word_host       0.800014
Length: 90, dtype: float64
```

```
desc_stats = df.describe()
display(desc_stats)
```

length_url length_hostname ip nb_dots nb_hyphens nb_at nb_qm nb_and nb_or nb_eq nb_underscore nb_tilde nb_percent nb_slash nb_star nb_colon nb_comma nb_semicolumn nb_dollar nb_space nb_www nb_com nb_dslash http_in_path https_token ratio_digits_url ratio_digits_host punycode port tld_in_path tld_in_subdomain abnormal_subdomain nb_subdomains prefix_suffix random_domain shortening_service path_extension nb_redirection nb_external_redirection length_words_raw char_repeat shortest_words_raw shortest_word_host shortest_word_path longest_words_raw longest_word_host longest_word_path avg_words_raw avg_word_host avg_word_path phish_hints domain_in_brand brand_in_subdomain brand_in_path suspecious_tld statistical_report nb_hyperlinks ratio_intHyperlinks ratio_extHyperlinks ratio_nullHyperlinks nb_extCSS ratio_intRedirection ratio_extRedirection ratio_intErrors ratio_extErrors login_form external_favicon links_in_tags submit_email ratio_intMedia ratio_extMedia sfh iframe popup_window safe_anchor onmouseover right_clic empty_title domain_in_title domain_with_copyright whois_registered_domain domain_registration_length domain_age web_traffic dns_record google_index page_rank status

12.000000 4.000000 0.000000 1.000000 0.000000 0.000000 0.000000 0.000000
0.0 0.000000 0.000000 0.000000 0.000000 2.000000 0.000000 1.000000
0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
0.000000 1.000000 0.000000 0.000000 0.000000 0.000000 0.000000
0.000000 1.000000 0.000000 1.000000 1.000000 0.000000
2.000000 1.000000 0.000000 2.000000 1.000000 0.000000 0.000000
0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
0.000000 0.0 0.000000 0.0 0.000000 0.0 0.000000
0.000000 0.000000 0.000000 0.0 0.000000 0.000000 0.0 0.000000
0.000000 0.000000 0.000000 0.00000 0.000000 0.000000 0.000000 0.000000
-1.000000 -12.000000 0.000000e+00 0.000000 0.000000 0.000000 0.000000
0.125

33.000000 15.000000 0.000000 2.000000 0.000000 0.000000 0.000000 0.000000
0.0 0.000000 0.000000 0.000000 0.000000 3.000000 0.000000 1.000000
0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
0.000000 2.000000 0.000000 0.000000 0.000000 0.000000 0.000000
0.000000 2.000000 1.000000 2.000000 3.000000 0.000000
9.000000 7.000000 0.000000 5.250000 5.250000 0.000000 0.000000
0.000000 0.000000 0.000000 0.000000 0.000000 9.000000 0.224991
0.000000 0.0 0.000000 0.0 0.000000 0.0 0.000000
0.000000 0.000000 0.000000 0.0 0.000000 0.000000 0.0 0.000000
0.000000 0.000000 0.000000 0.00000 0.000000 1.000000 0.000000 0.000000
84.000000 972.250000 0.000000e+00 0.000000 0.000000 1.000000 0.000000

0.125

47.000000 19.000000 0.000000 2.000000 0.000000 0.000000 0.000000 0.000000
0.0 0.000000 0.000000 0.000000 0.000000 4.000000 0.000000 1.000000
0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
1.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
0.000000 2.000000 0.000000 0.000000 0.000000 0.000000 0.000000
0.000000 5.000000 3.000000 3.000000 3.000000 2.000000
11.000000 10.000000 7.000000 6.500000 7.000000 4.857143 0.000000
0.000000 0.000000 0.000000 0.000000 0.000000 34.000000 0.743442
0.131148 0.0 0.000000 0.0 0.000000 0.0 0.000000
0.000000 0.000000 60.000000 0.0 11.111111 0.000000 0.0 0.000000
0.000000 23.294574 0.000000 0.00000 0.000000 1.000000 0.000000 0.000000
242.000000 3993.000000 1.651000e+03 0.000000 1.000000 3.000000 0.500000
0.125

55.297318 10.777171 0.357644 1.369686 2.087087 0.155500 0.364456 0.821337
0.0 0.998317 1.093336 0.081274 1.466450 1.882251 0.026448 0.240325
0.103240 0.598190 0.077111 0.375576 0.501912 0.379008 0.080742 0.169358
0.487559 0.089363 0.093422 0.018705 0.048547 0.247622 0.218225
0.145412 0.637069 0.401843 0.276332 0.328964 0.013227 0.691907
0.056035 5.572355 4.768936 2.211571 3.941580 2.997809
22.083644 4.932015 23.077883 4.145827 3.578435 7.147050 0.842600
0.305533 0.063996 0.069827 0.132722 0.331266 166.758254 0.376474
0.319958 0.0 2.758802 0.0 0.266437 0.0 0.156209
0.244058 0.496666 41.523144 0.0 46.249897 38.386577 0.0 0.036204
0.077465 39.073385 0.033707 0.03739 0.330460 0.417038 0.496353 0.259948
814.769415 3107.784600 1.995606e+06 0.140425 0.498868 2.536955 0.500022
0.125

61.126684 21.090289 0.150569 2.480752 0.997550 0.022222 0.141207 0.162292
0.0 0.293176 0.322660 0.006649 0.123097 4.289589 0.000700 1.027909
0.004024 0.062292 0.001925 0.034821 0.448469 0.127997 0.006562 0.016710
0.610936 0.053137 0.025024 0.000350 0.002362 0.065617 0.050131
0.021610 2.231671 0.202450 0.083290 0.123447 0.000175 0.498250
0.003150 6.232808 2.927472 3.127297 5.019773 2.398950
15.393876 10.467979 10.561505 7.258882 7.678075 5.092425 0.327734
0.104199 0.004112 0.004899 0.017935 0.059755 87.189764 0.602457
0.276720 0.0 0.784864 0.0 0.158926 0.0 0.062469
0.063605 0.442170 51.978211 0.0 42.870444 23.236293 0.0 0.001312
0.006037 37.063922 0.001137 0.00140 0.124759 0.775853 0.439545 0.072878
492.532196 4062.543745 8.567566e+05 0.020122 0.533946 3.185739 0.500000
0.125

71.000000 24.000000 0.000000 3.000000 1.000000 0.000000 0.000000 0.000000
0.0 0.000000 0.000000 0.000000 0.000000 5.000000 0.000000 1.000000
0.000000 0.000000 0.000000 0.000000 1.000000 0.000000 0.000000 0.000000
1.000000 0.079365 0.000000 0.000000 0.000000 0.000000 0.000000
0.000000 3.000000 0.000000 0.000000 0.000000 0.000000 1.000000
0.000000 8.000000 4.000000 3.000000 6.000000 3.000000
16.000000 13.000000 11.000000 8.000000 9.000000 6.714286 0.000000
0.000000 0.000000 0.000000 0.000000 0.000000 101.000000 0.944767
0.474840 0.0 1.000000 0.0 0.230769 0.0 0.034483
0.000000 1.000000 98.061004 0.0 100.000000 33.333333 0.0 0.000000
0.000000 75.000000 0.000000 0.00000 0.000000 1.000000 1.000000 0.000000
449.000000 7026.750000 3.738455e+05 0.000000 1.000000 5.000000 1.000000
0.125

1641.000000 214.000000 1.000000 24.000000 43.000000 4.000000 3.000000
19.000000 0.0 19.000000 18.000000 1.000000 96.000000 33.000000 1.000000
7.000000 4.000000 20.000000 6.000000 18.000000 2.000000 6.000000 1.000000
4.000000 1.000000 0.723881 0.800000 1.000000 1.000000 1.000000 1.000000
1.000000 3.000000 1.000000 1.000000 1.000000 1.000000 6.000000
1.000000 106.000000 146.000000 31.000000 39.000000 40.000000
829.000000 62.000000 829.000000 128.250000 39.000000 250.000000 10.000000
1.000000 1.000000 1.000000 1.000000 2.000000 4659.000000 1.000000
1.000000 0.0 124.000000 0.0 2.000000 0.0 1.000000
1.000000 1.000000 100.000000 0.0 100.000000 100.000000 0.0 1.000000
1.000000 100.000000 1.000000 1.00000 1.000000 1.000000 1.000000
1.000000 29829.000000 12874.000000 1.076799e+07 1.000000 1.000000
10.000000 1.000000 0.125

11430.000000 11430.000000 11430.000000 11430.000000 11430.000000 11430.000000
11430.000000 11430.000000 11430.0 11430.000000 11430.000000 11430.000000 11430.000000
11430.000000 11430.000000 11430.000000 11430.000000 11430.000000 11430.000000 11430.000000
11430.000000 11430.000000 11430.000000 11430.000000 11430.000000 11430.000000
11430.000000 11430.000000 11430.000000 11430.000000 11430.000000 11430.000000
11430.000000 11430.000000 11430.000000 11430.000000 11430.000000 11430.000000
11430.000000 11430.000000 11430.000000 11430.000000 11430.000000
11430.000000 11430.000000 11430.000000 11430.000000 11430.000000 11430.000000
11430.000000 11430.000000 11430.000000 11430.000000 11430.000000 11430.000000
11430.000000 11430.000000 11430.000000 11430.000000 11430.0 11430.000000
11430.0 11430.000000 11430.0 11430.000000 11430.000000 11430.000000
11430.000000 11430.0 11430.000000 11430.000000 11430.0 11430.000000 11430.000000
11430.000000 11430.000000 11430.00000 11430.000000 11430.000000 11430.000000
11430.000000 11430.000000 11430.000000 1.143000e+04 11430.000000
11430.000000 11430.000000 11430.000000 0.125

dtype: float64

```
# checking the outliers of the dataset

numeric_df = df.select_dtypes(include=['int64', 'float64'])
corr_matrix = numeric_df.corr()
corr_matrix = numeric_df.corr()
outlier = {}

for column in numeric_df.columns:
    Q1 = numeric_df[column].quantile(0.25)
    Q3 = numeric_df[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    outliers = numeric_df[(numeric_df[column] < lower_bound) | (numeric_df[column] > upper
    outliercount = outliers.shape[0]

    if outliercount > 0:
        outlier[column] = outliercount

# Display features with outliers
print("Features with outliers:")
```

```
for feature, count in outlier.items():
    print(f"{feature}:    {count} outliers")
```

```
Features with outliers:
length_url:    620 outliers
length_hostname:    775 outliers
ip:    1721 outliers
nb_dots:    567 outliers
nb_hyphens:    1371 outliers
nb_at:    245 outliers
nb_qm:    1555 outliers
nb_and:    761 outliers
nb_eq:    1564 outliers
nb_underscore:    1695 outliers
nb_tilde:    76 outliers
nb_percent:    355 outliers
nb_slash:    401 outliers
nb_star:    8 outliers
nb_colon:    197 outliers
nb_comma:    24 outliers
nb_semicolumn:    248 outliers
nb_dollar:    11 outliers
nb_space:    210 outliers
nb_com:    1327 outliers
nb_dslash:    75 outliers
http_in_path:    150 outliers
ratio_digits_url:    933 outliers
ratio_digits_host:    1503 outliers
punycode:    4 outliers
port:    27 outliers
tld_in_path:    750 outliers
tld_in_subdomain:    573 outliers
abnormal_subdomain:    247 outliers
prefix_suffix:    2314 outliers
random_domain:    952 outliers
shortening_service:    1411 outliers
path_extension:    2 outliers
nb_redirection:    166 outliers
nb_external_redirection:    36 outliers
length_words_raw:    264 outliers
char_repeat:    310 outliers
shortest_words_raw:    1435 outliers
shortest_word_host:    1093 outliers
shortest_word_path:    428 outliers
longest_words_raw:    1035 outliers
longest_word_host:    220 outliers
longest_word_path:    929 outliers
avg_words_raw:    725 outliers
avg_word_host:    568 outliers
avg_word_path:    282 outliers
phish_hints:    2041 outliers
domain_in_brand:    1191 outliers
brand_in_subdomain:    47 outliers
brand_in_path:    56 outliers
suspecious_tld:    205 outliers
statistical_report:    377 outliers
nb_hyperlinks:    953 outliers
nb_extCSS:    1019 outliers
ratio_extRedirection:    999 outliers
ratio_extErrors:    2149 outliers
login_form:    727 outliers
ratio_extMedia:    2012 outliers
iframe:    15 outliers
```

```
popup_window:    69 outliers
onmouseover:    13 outliers
right_clic:    16 outliers
empty_title:    1426 outliers
domain_in_title:    2562 outliers
whois_registered_domain:    833 outliers
domain_registration_length:    1529 outliers
web_traffic:    2138 outliers
dns_record:    230 outliers
```

```
# 5: Preparing the x & y cloumn
X = df.drop(['url', 'status'], axis=1)
y = df['status']

# 6: Spliting the dataset (80:20)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# 7: Feature scaling
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

```
print("Train size:", X_train.shape[0])
print("Test size:", X_test.shape[0])
print("Train distribution:", y_train.value_counts(normalize=True))
print("Test distribution:", y_test.value_counts(normalize=True))
```

```
Train size: 9144
Test size: 2286
Train distribution: 1    0.501531
0    0.498469
Name: status, dtype: float64
Test distribution: 0    0.506124
1    0.493876
Name: status, dtype: float64
```