Project Development Phase

Delivery of Sprint-1

Date	19 November 2022
Team ID	PNT2022TMID11399
Project Name	Web Phishing Detection
Maximum Marks	4 Marks

Importing Libraries and Dataset

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

In [3]:
```
data=pd.read_csv("dataset_website.csv")
data.head()
```

Out[3]:

	index	having_IPhaving_IP_Address	URLURL_Length	Shortining_Service	having_At_Symbol	double_slash_redirecting	Prefix_Suffix	having_Sub_Domain	SSLfinal_State	Domain_re
0	1	-1	1	1	1	-1	-1	-1	-1	
1	2	1	1	1	1	1	-1	0	1	
2	3	1	0	1	1	1	-1	-1	-1	
3	4	1	0	1	1	1	-1	-1	-1	
4	5	1	0	-1	1	1	-1	1	1	

5 rows × 32 columns

Numerical Analysis

In [4]:
```
data.size
```

Out[4]: 353760

In [5]:
```
data.shape
```

In [5]:
```
data.shape
```

Out[5]: (11055, 32)

In [6]:
```
data.info()
```

```
RangeIndex: 11055 entries, 0 to 11054
Data columns (total 32 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   index                        11055 non-null  int64
 1   having_IPhaving_IP_Address   11055 non-null  int64
 2   URLURL_Length                11055 non-null  int64
 3   Shortining_Service           11055 non-null  int64
 4   having_At_Symbol             11055 non-null  int64
 5   double_slash_redirecting     11055 non-null  int64
 6   Prefix_Suffix                11055 non-null  int64
 7   having_Sub_Domain            11055 non-null  int64
 8   SSLfinal_State               11055 non-null  int64
 9   Domain_registeration_length  11055 non-null  int64
 10  Favicon                      11055 non-null  int64
 11  port                         11055 non-null  int64
 12  HTTPS_token                  11055 non-null  int64
 13  Request_URL                  11055 non-null  int64
 14  URL_of_Anchor                11055 non-null  int64
 15  Links_in_tags                11055 non-null  int64
 16  SFH                          11055 non-null  int64
 17  Submitting_to_email          11055 non-null  int64
 18  Abnormal_URL                 11055 non-null  int64
 19  Redirect                     11055 non-null  int64
 20  on_mouseover                 11055 non-null  int64
 21  RightClick                   11055 non-null  int64
 22  popUpWidnow                  11055 non-null  int64
 23  Iframe                       11055 non-null  int64
 24  age_of_domain                11055 non-null  int64
 25  DNSRecord                    11055 non-null  int64
 26  web_traffic                  11055 non-null  int64
 27  Page_Rank                    11055 non-null  int64
 28  Google_Index                 11055 non-null  int64
```

```
 29  Links_pointing_to_page       11055 non-null  int64
 30  Statistical_report           11055 non-null  int64
 31  Result                       11055 non-null  int64
dtypes: int64(32)
memory usage: 2.7 MB
```

In [8]:
```
data.describe()
```

Out[8]:

	index	having_IPhaving_IP_Address	URLURL_Length	Shortining_Service	having_At_Symbol	double_slash_redirecting	Prefix_Suffix	having_Sub_Domain	SSLfinal_State
count	11055.000000	11055.000000	11055.000000	11055.000000	11055.000000	11055.000000	11055.000000	11055.000000	11055.000000
mean	5528.000000	0.313795	-0.633198	0.738761	0.700588	0.741474	-0.734962	0.063953	0.250927
std	3191.447947	0.949534	0.766095	0.673998	0.713598	0.671011	0.678139	0.817518	0.911892
min	1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000
25%	2764.500000	-1.000000	-1.000000	1.000000	1.000000	1.000000	-1.000000	-1.000000	-1.000000
50%	5528.000000	1.000000	-1.000000	1.000000	1.000000	1.000000	-1.000000	0.000000	1.000000
75%	8291.500000	1.000000	-1.000000	1.000000	1.000000	1.000000	-1.000000	1.000000	1.000000
max	11055.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

8 rows × 32 columns

In [9]:
```
data.isnull().any()
```

Out[9]:
```
index                        False
having_IPhaving_IP_Address   False
URLURL_Length                False
Shortining_Service           False
having_At_Symbol             False
double_slash_redirecting     False
Prefix_Suffix                False
having_Sub_Domain            False
SSLfinal_State               False
Domain_registeration_length  False
Favicon                      False
port                         False
HTTPS_token                  False
Request_URL                  False
URL_of_Anchor                False
Links_in_tags                False
SFH                          False
Submitting_to_email          False
Abnormal_URL                 False
Redirect                     False
on_mouseover                 False
RightClick                   False
popUpWidnow                  False
Iframe                       False
age_of_domain                False
DNSRecord                    False
web_traffic                  False
Page_Rank                    False
Google_Index                 False
Links_pointing_to_page       False
Statistical_report           False
Result                       False
dtype: bool
```

In [10]:
```
data.isnull().sum()
```

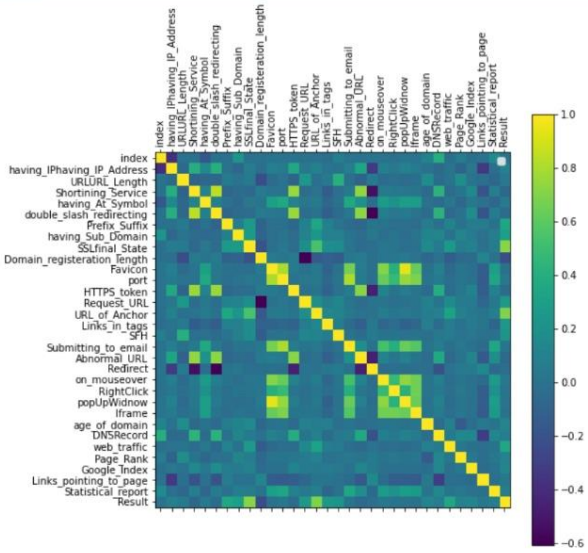
Out[10]:
```
index                        0
having_IPhaving_IP_Address   0
URLURL_Length                0
Shortining_Service           0
having_At_Symbol             0
double_slash_redirecting     0
Prefix_Suffix                0
having_Sub_Domain            0
SSLfinal_State               0
Domain_registeration_length  0
Favicon                      0
port                         0
HTTPS_token                  0
Request_URL                  0
URL_of_Anchor                0
```

```
Links_in_tags                0
SFH                          0
Submitting_to_email          0
Abnormal_URL                 0
Redirect                     0
on_mouseover                 0
RightClick                   0
popUpWidnow                  0
Iframe                       0
age_of_domain                0
DNSRecord                    0
web_traffic                  0
Page_Rank                    0
Google_Index                 0
Links_pointing_to_page       0
Statistical_report           0
Result                       0
dtype: int64
```

In [11]:

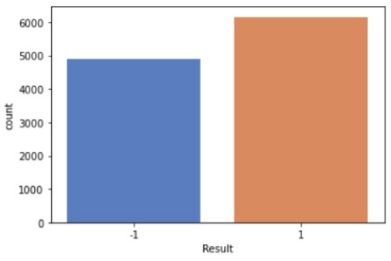
```
def plot_corr(df,size=8):
    corr=df.corr()
    fig,ax=plt.subplots(figsize=(size,size))
    ax.legend()
    cax=ax.matshow(corr)
    fig.colorbar(cax)
    plt.xticks(range(len(corr.columns)), corr.columns, rotation='vertical')
    plt.yticks(range(len(corr.columns)), corr.columns)
plot_corr(data)
```

```
No artists with labels found to put in legend.  Note that artists whose label start with an underscore are ignored when legend() is called with no argu
ment.
```



In [12]:

```
with sns.color_palette('muted'):
    sns.countplot(x=data['Result'])
```



Splitting the Data

In [13]:

```
x=data.iloc[:,1:31].values
y=data.iloc[:,-1].values
```

In [14]:

```
x
```

Out[14]:

```
array([[-1,  1,  1, ...,  1,  1, -1],
       [ 1,  1,  1, ...,  1,  1,  1],
       [ 1,  0,  1, ...,  1,  0, -1],
       ...,
       [ 1, -1,  1, ...,  1,  0,  1],
       [-1, -1,  1, ...,  1,  1,  1],
       [-1, -1,  1, ..., -1,  1, -1]], dtype=int64)
```

In [15]:
```
y
```

Out[15]:
```
array([-1, -1, -1, ..., -1, -1, -1], dtype=int64)
```

Train, Test and Split

In [16]:
```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=0)
```

In [17]:
```
x_train.shape
```

Out[17]:
```
(8844, 30)
```

In [18]:
```
y_train.shape
```

Out[18]:
```
(8844,)
```

In [19]:
```
x_test.shape
```

Out[19]:
```
(2211, 30)
```

In [20]:
```
y_test.shape
```

Out[20]:
```
(2211,)
```