

SPRINT 1

Date	29 October 2022
Team ID	PNT2022TMID15890
Project Name	Web Phishing Detection

Sprint Objective:

In Sprint 1 the team members download the given dataset from the given link. Then the dataset should be pre-processed by checking all the information in the given datasets like the features given for the URL, what the values mean for each Feature (URL's feature) in the dataset; then, they should check for the null values, if there is any null value they should have them.

Screenshots:

```
In [7]: import pandas as pd
import numpy as np

In [8]: df = pd.read_csv("../dataset_website.csv")
df.head()

Out[8]:
```

	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 x 12 columns

```
In [9]: df.info()

RangeIndex: 11055 entries, 0 to 11054
Data columns (total 12 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_registration_length          11055 non-null  int64
10  Favicon                             11055 non-null  int64
11  port                                11055 non-null  int64
12  HTTPS_token                         11055 non-null  int64
```

```

In [7]: df.isnull().any()

Out[7]:
Index                                False
having_IPhaving_IP_Address            False
URLURL_Length                        False
Shortening_Service                    False
having_At_Symbol                      False
double_slash_redirecting              False
Prefix_Suffix                        False
having_Sub_Domain                     False
SSLfinal_State                        False
Domain_registration_length            False
Favicon                              False
port                                 False
HTTPS_token                           False
Request_URL                           False
URL_of_Anchor                         False
Links_in_tags                         False
SPH                                   False
Submitting_to_email                   False
Abnormal_URL                          False
Redirect                              False
on_mouseover                          False
RightClick                            False
popupwindow                           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

```

Sprint Result:

Team members downloaded the given data. The dataset contains 11,055 data samples. Each data samples have 30 different features and these features has either of these 3 different values (-1, 0, 1). The dataset is checked for null values and there is none.