

Data Collection and Preprocessing Phase

Date	06-06-2024
Team ID	740055
Project Title	DETECTION OF PHISHING WEBSITE FROM URLS
Maximum Marks	6 Marks

Data Exploration and Preprocessing Report

Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

Section	Description
---------	-------------

Data Overview	<u>Dimension:</u> 614 rows \times 13 columns <u>Descriptive statistics:</u> -
Univariate Analysis	

--	--

Bivariate Analysis	-
Multivariate Analysis	-
Outliers and Anomalies	-
Data Preprocessing Code Screenshots	

Loading Data

```
[32]: #Import Dataset
ds= pd.read_csv("dataset_phishing.csv")
ds.head()
```

```
[33]:
```

	url	length	url	length	hostname	ip	nb_dots	nb_hyphens	nb_at	nb_qm	nb_and	nb_or	...	domain_in_title	domain_with
0	http://www.crestonwood.com/router.php	37		19	0	3	0	0	0	0	0	0	...	0	
1	http://shadetreeechnology.com/V4/validation/a...	77		23	1	1	0	0	0	0	0	0	...	1	
2	https://support-appleid.com.secureupdate.dial...	126		50	1	4	1	0	1	2	0	...	1		
3	http://ngjpt.ac.in	18		11	0	2	0	0	0	0	0	...	1		
4	http://www.racing.com/tracks/gateway-motorspo...	55		15	0	2	2	0	0	0	0	...	0		

5 rows x 89 columns

Handling Missing Data

```
[35]: #Splitting data as independent and dependent
#removing index column in independent dataset
x=ds.iloc[:,1:31].values
y=ds.iloc[:, -1].values
print(x,y)
```

```
[[ 37. 19. 0. ... 0. 0. 0.]
 [ 77. 23. 1. ... 0. 0. 0.]
 [126. 50. 1. ... 0. 0. 0.]
 ...
 [185. 16. 1. ... 0. 0. 0.]
 [ 38. 30. 0. ... 0. 0. 0.]
 [477. 14. 1. ... 0. 0. 1.]] [['legitimate' 'phishing' 'phishing' ... 'legitimate' 'legitimate'
 'phishing']]
```

Data Transformation	-
Feature Engineering	Attached the codes in final submission.
Save Processed Data	-