Data Collection and Preprocessing Phase

Date	29 September 2024
Team ID	LTVIP2024TMID24838
Project Title	Detection of Phishing Websites from URLs
Maximum Marks	6 Marks

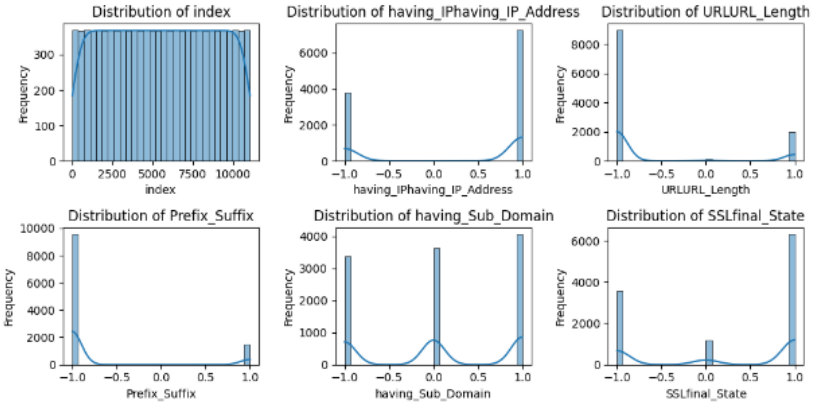
Data Exploration and Preprocessing Template

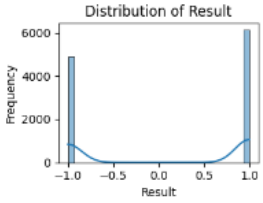
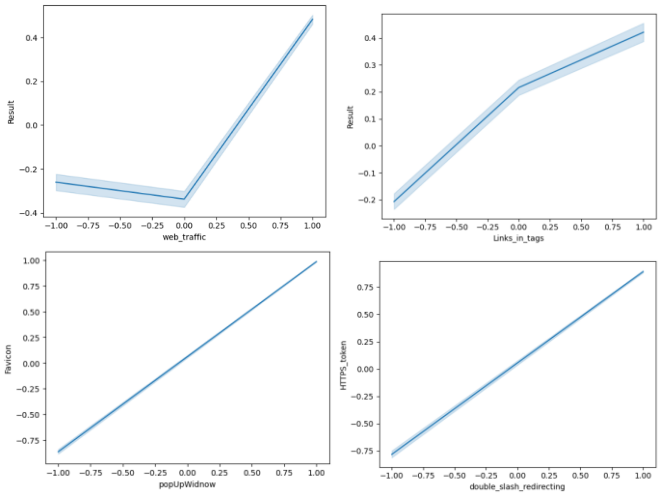
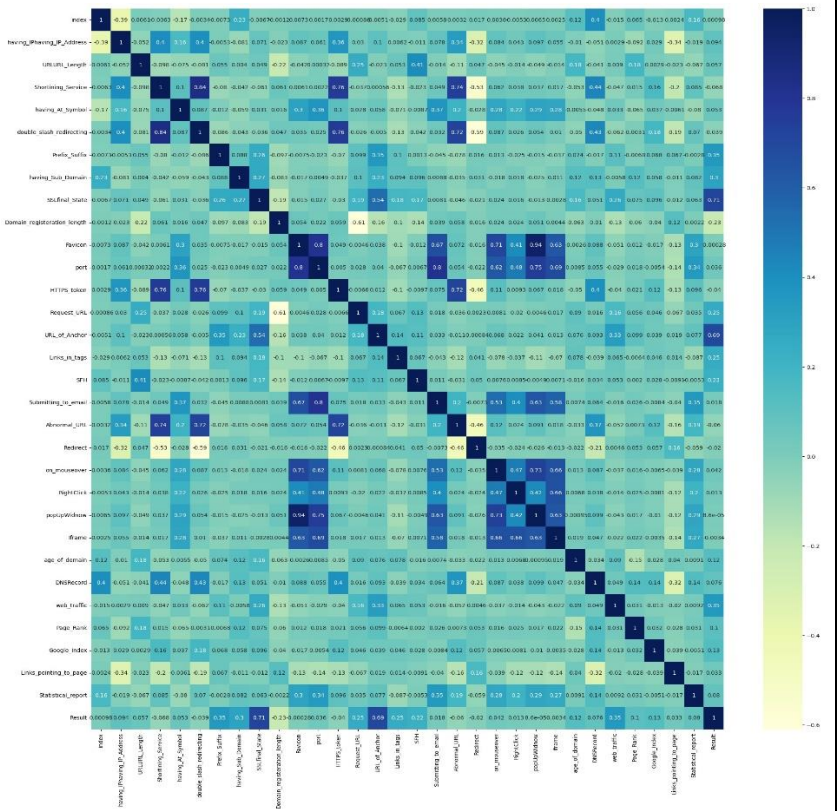
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

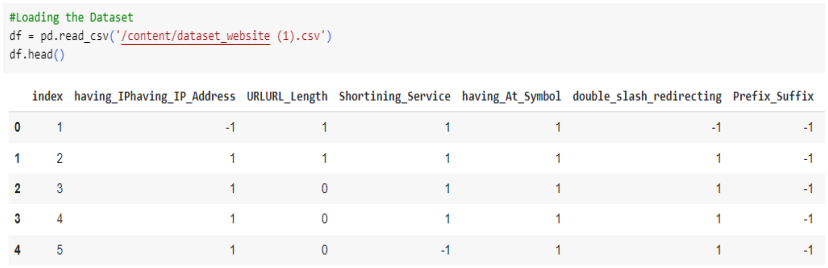
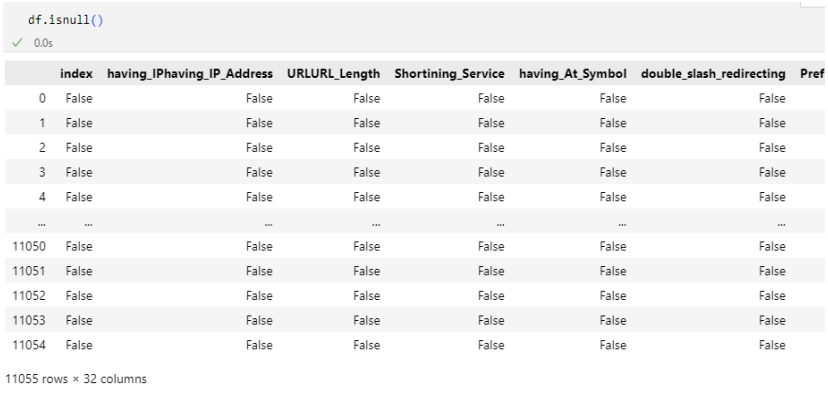
Section	Description
Data Overview	Dimensions: 11055 rows X 32 columns

	index	having_IPhaving_IP_Address	URLURL_Length	Shortining_Service	having_At_Symbol	double_slash_redirecting
count	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
std	3191.447947	0.949534	0.766095	0.673998	0.713598	0.671011
min	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
50%	5528.000000	1.000000	-1.000000	1.000000	1.000000	1.000000
75%	8291.500000	1.000000	-1.000000	1.000000	1.000000	1.000000
max	11055.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Univariate Analysis



	
Bivariate Analysis	
Multivariate Analysis	

Outliers and Anomalies	Identification and treatment of outliers.
Data Preprocessing Code Screenshots	
Loading Data	 index having_IPhaving_IP_Address URLURL_Length Shortining_Service having_At_Symbol double_slash_redirecting Prefix_Suffix 0 1 -1 1 1 1 -1 -1 1 2 1 1 1 1 1 -1 2 3 1 0 1 1 1 -1 3 4 1 0 1 1 1 -1 4 5 1 0 -1 1 1 -1
Handling Missing Data	 index having_IPhaving_IP_Address URLURL_Length Shortining_Service having_At_Symbol double_slash_redirecting Pref 0 False False False False False False 1 False False False False False False 2 False False False False False False 3 False False False False False False 4 False False False False False False ... ... 11050 False False False False False False 11051 False False False False False False 11052 False False False False False False 11053 False False False False False False 11054 False False False False False False 11055 rows × 32 columns No NULL values found in the dataset.
Data Transformation	No need for data transformation for the used dataset.
Feature Engineering	Included in the final code.
Save Processed Data	-