Data Collection and Preprocessing Phase

Date	10 june 2024
Team ID	739879
Project Title	Detection of phishing websites 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	Dimension: 11055rows X32 columns Descriptive statistics: `ds.describe()` 8 rows × 32 columns
Univariate Analysis	

For the Data Overview descriptive statistics table:

	index	having_IPhaving_IP_Address	URLURL_Length	Shortining_Service	having_At_Symbol	double_slash_redirecting	Prefix_Suffix	having_Sub_Domain
count	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
std	3191.447947	0.949534	0.766095	0.673998	0.713598	0.671011	0.678139	0.817518
min	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
50%	5528.000000	1.000000	-1.000000	1.000000	1.000000	1.000000	-1.000000	0.000000
75%	8291.500000	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

	—
Bivariate Analysis	-
Multivariate Analysis	-

Outliers and Anomalies	-
Data Preprocessing Code Screenshots	
Loading Data	`]:` `ds=pd.read_csv("dataset_website - Copy.csv")`
Handling Missing Data	-
Data Transformation	-
Feature Engineering	-
Save Processed Data	-