Data Collection and Preprocessing Phase

Date	28 September 2024
Team ID	LTVIP2024TMID24838
Project Title	Detection of Phishing Websites from URLs
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Template

Elevate your data strategy with the Data Collection plan and the Raw Data Sources report, ensuring meticulous data curation and integrity for informed decision-making in every analysis and decision-making endeavor.

Data Collection Plan Template

Section	Description
Project Overview	The Phishing URL Detection dataset aid in the development and evaluation of models to distinguishing between phishing and legitimate URLs. Phishing attacks involve fraudulent attempts to obtain sensitive information by masquerading as a trustworthy entity in electronic communications. This project detects a url given by the user to determine its legitimacy by using a machine leaning model.
Data Collection Plan	• Searching for datasets containing multiple urls of both phishing and legitimate. • Dataset containing the characteristics which determine the chances of a url being a phishing url.

Raw Data Sources Identified	▪ The collected dataset is obtained from kaggle ▪ This dataset provides a comprehensive collection of URLs labeled as either phishing or legitimate, along with various features extracted from these URLs to facilitate machine learning and data analysis tasks.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Kaggle Dataset	• The dataset contains 11055 rows X 32 columns. • The rows contain no of urls in which 6157 are safe and 4898 are phishing websites. • The column's represent the characteristics of the urls	https://www.kaggle.com/datasets/adityachaudhary1306/phishing-url-classifier-dataset-cleaned	CSV	835 kb	Public