Data Collection and Preprocessing Phase

Date	27 September 2024
Team ID	LTVIP2024TMID24973
Project Title	Detection of Phishing Websites from URLs Using Machine learning
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	This machine learning project aims to develop a model that can accurately detect phishing websites based on their URLs. By analyzing various features of URLs, the objective is to differentiate between legitimate and malicious websites, ultimately enhancing online security and protecting users from phishing attacks.
Data Collection Plan	Data will be collected from a variety of sources, including publicly available datasets, web scraping of known phishing websites, and API services that provide real-time URL classification. This multi-source approach will ensure a comprehensive dataset for training and testing the model.

Raw Data Sources Identified	The following raw data sources have been identified for this project:

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Kaggle Dataset	Dataset containing known phishing URLs and their features.	Web page Phishing Detection Dataset	CSV	1 GB	Public