Data Collection and Preprocessing Phase

Date	10 june 2024
Team ID	739879
Project Title	Detection of phishing websites from URLs
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Report:

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:

Section	Description
Project Overview	The machine learning project aims to detect phishing websites based on a combination of URL analysis. Using a dataset with features such as having ip address, url length, sub domain, domain etc.. the objective is to build a model that accurately classifies URL status (legitimate or phishing), facilitating efficient and informed decision-making in the lending process.
Data Collection Plan	● Search for datasets related to detecting phishing websites ● Prioritize datasets with diverse demographic information.
Raw Data Sources Identified	The raw data sources for this project include datasets obtained from Kaggle & UCI, the popular platforms for data science competitions and repositories. The provided sample data represents a subset of the collected information, encompassing variables such as having_ip_address, url length, domain, subdomain, ssl etc..

Raw Data Sources Report:

Source Name	Description	Location/URL	Format	Size	Access Permissions
Kaggle Dataset	The dataset comprises URL features (having ip address, url length, subdomain, domain,ssl, shortening etc status is outcome.	https://www.kaggle.com/datasets/shashwatwork/web-page-phishing-detection-dataset	CSV	3.49 MB	Public