Data Collection and Preprocessing Phase

Date	06-06-2024
Team ID	740031
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 Phishing Website Detection from URLs project aims to develop a robust system that utilizes machine learning techniques to identify and flag potential phishing threats. The system will extract features from URLs, classify them using machine learning algorithms, and detect phishing websites in real-time. With a focus on continuous improvement, the system will be updated regularly to stay ahead of evolving phishing tactics. The project's outcome is expected to be a highly accurate phishing detection system, providing enhanced online security and protection for users.
Data Collection Plan	First, large datasets of legitimate and phishing URLs are collected for training and testing machine learning models. Web scraping techniques are then used to extract relevant features from websites, such as HTML content, CSS, and JavaScript code. Additionally, DNS records, including IP addresses, domain names, and MX records, are collected to analyze website authenticity.
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.

	Here it is identified data sources are like webpages, urls, Dns records, whois data, network traffic, userfeedback.