Data Collection and Preprocessing Phase

Date	26 September 2024
Team ID	LTVIP2024TMID24973
Project Title	Detection of Phishing Websites from URLs Using Machine learning
Maximum Marks	2 Marks

Data Quality Report Template

The Data Quality Report Template will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies.

Data Source	Data Quality Issue	Severity	Resolution Plan
Dataset	Missing Values	Moderate	Implement imputation for numeric fields (e.g., URL length) using median. For categorical fields (e.g., presence of HTTPS), use the most frequent category. Consider dropping rows or columns with excessive missing values (>20%).
	Inconsistent Data Formats	High	Standardize URL formats by removing unnecessary characters and ensuring consistent casing (e.g., all lowercase). Use regex to validate URL formats.

	Duplicate Records	Moderate	Identify and remove exact duplicate URLs. For near-duplicates, use fuzzy matching techniques to group and consolidate similar entries.
	Incorrect Labels (Phishing vs. Legitimate)	High	Manually verify a sample of URLs to check label accuracy. Consider cross-referencing with reliable sources or databases of known phishing sites. Retrain the model with corrected labels.
	Outliers in Numeric Features	Moderate	Use Z-scores or IQR to detect outliers in features like URL length or number of parameters. Analyze whether they are valid entries or errors, and decide to remove or adjust them accordingly.
	Imbalanced Classes	High	Apply techniques such as SMOTE to generate synthetic examples for the minority class (phishing URLs) or use class-weight adjustments in model training to address imbalance.