Data Preprocessing for Website Phishing Detection

1. Overview of Preprocessed Dataset

The dataset contains 11,430 records with 8 key features. The preprocessing steps ensure the data is clean, structured, and suitable for training deep learning models.

2. Key Features in the Dataset

- length_url: Length of the URL
- nb_dots: Number of dots in the URL
- nb_hyphens: Number of hyphens in the URL
- nb_at: Number of '@' symbols in the URL
- nb_slash: Number of slashes in the URL
- nb_www: Presence of 'www' in the URL
- nb_com: Presence of '.com' in the URL
- status: Label indicating phishing (1) or legitimate (0)

3. Preprocessing Steps

- Feature Extraction: Extracting characteristics from URLs.
- Data Cleaning: Removing duplicates and handling missing values.
- Normalization: Applying Z-score normalization for consistency.
- Feature Scaling: Ensuring numeric values are within a similar range.
- Data Augmentation: Enhancing phishing sample variations.
- Dataset Splitting: Dividing into training, validation, and test sets.

4. Example of Preprocessed Data

-0.4363	0.3791	-0.4779	-0.1429	-0.6852	1.0989	-0.3377	0
0.2871	-1.0811	-0.4779	-0.1429	0.3774	-0.8936	-0.3377	1
1.1732	1.1092	0.0012	-0.1429	0.3774	-0.8936	2.3009	1
-0.7799	-0.3510	-0.4779	-0.1429	-1.2165	-0.8936	-0.3377	0
-0.1108	-0.3510	0.4803	-0.1429	0.3774	1.0989	-0.3377	0

5. Conclusion

The preprocessed dataset has been cleaned, normalized, and structured to enhance deep learning model performance. By leveraging feature extraction, scaling, and augmentation, this dataset is optimized for phishing detection.