Information Security
Assignment 2

Group Members:
Rafay Kiyani
Sherwin Desouza
Section: AI-K

Data Pre-processing and feature Selection

To create the data, I used the following features:

`url_length` – The total length of the URL. Longer URLs may indicate phishing or malicious sites.

`domain_length` – The length of the domain name (excluding the full URL path). Shorter or excessively long domains may be suspicious.

`num_digits` – The number of digits in the URL. Phishing URLs often contain many numbers.

`num_special_chars` – The count of special characters (e.g., `@`, `-`, `_`, `?`, `&`). Excessive use may indicate obfuscation.

`num_subdomains` – The number of subdomains in the URL (e.g., `sub.example.com` has one subdomain). Multiple subdomains can be suspicious.

`https` – A binary flag (1 = HTTPS, 0 = HTTP). Secure websites typically use HTTPS.

`is_shortened` – A binary flag (1 = URL is shortened, 0 = normal URL). Shortened URLs can hide malicious links.

`url_depth` – The number of slashes (`/`) in the URL path, indicating how deep the URL goes. Complex structures may be suspicious.

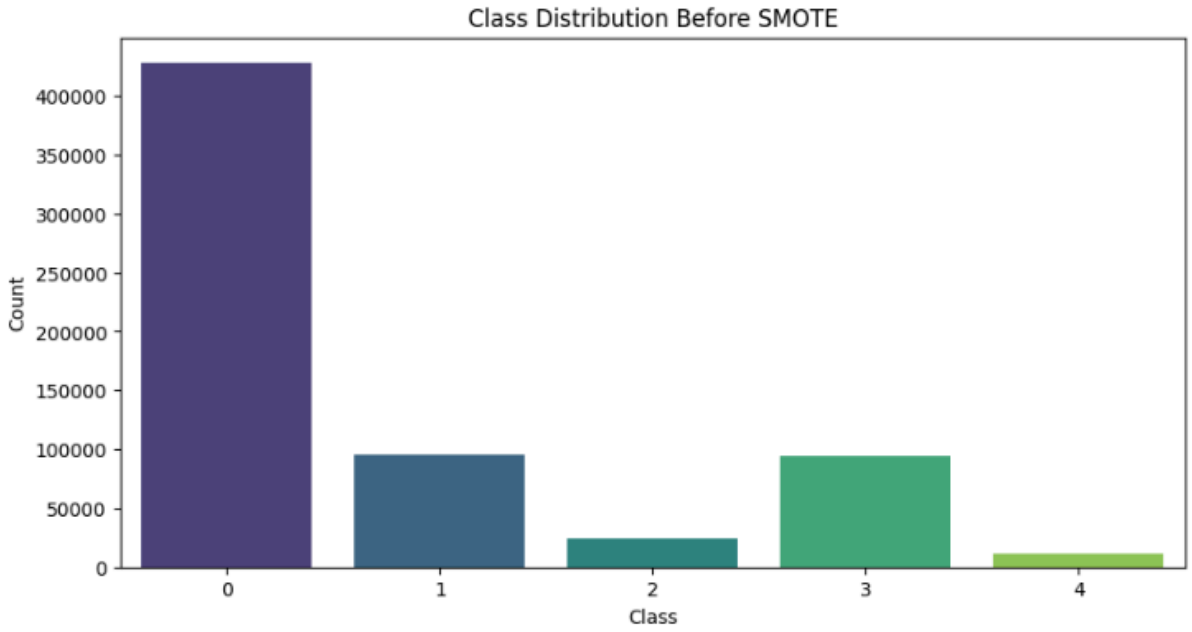
`num_params` – The number of parameters in the URL (e.g., in `?id=123&ref=abc`, there are 2 parameters). Many parameters may indicate tracking or phishing.

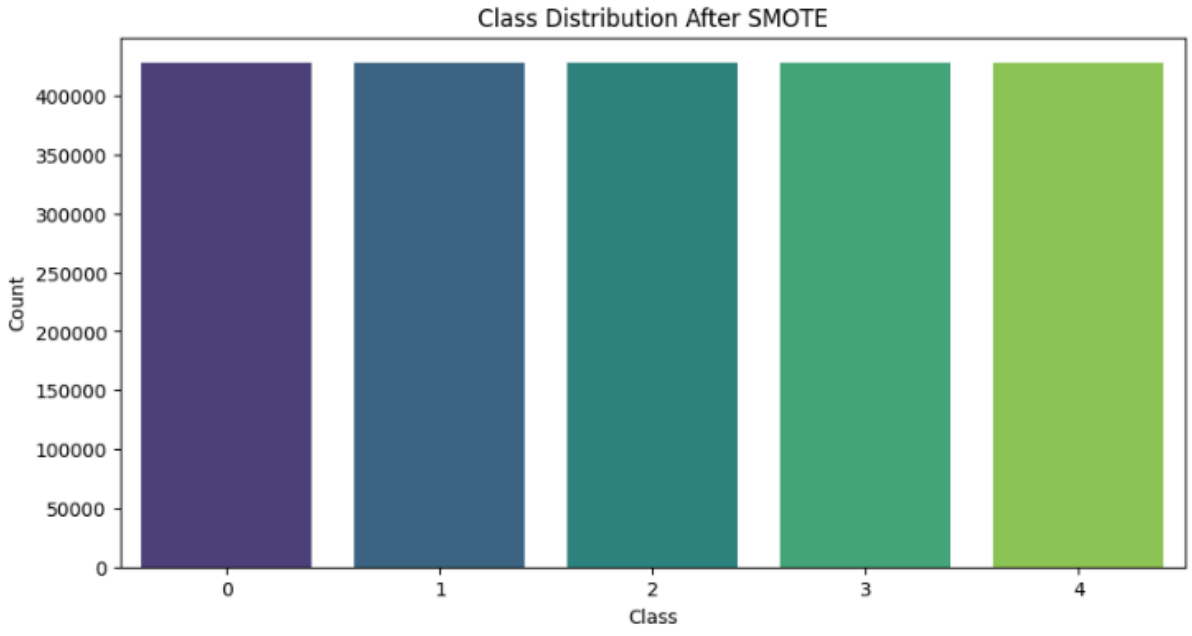
`suspicious_keywords` – A flag indicating whether the URL contains words like “login,” “verify,” “secure,” which are common in phishing sites.

`tld` – The top-level domain (e.g., `.com`, `.net`, `.xyz`). Some TLDs are associated with higher rates of malicious activity.

Data Balancing

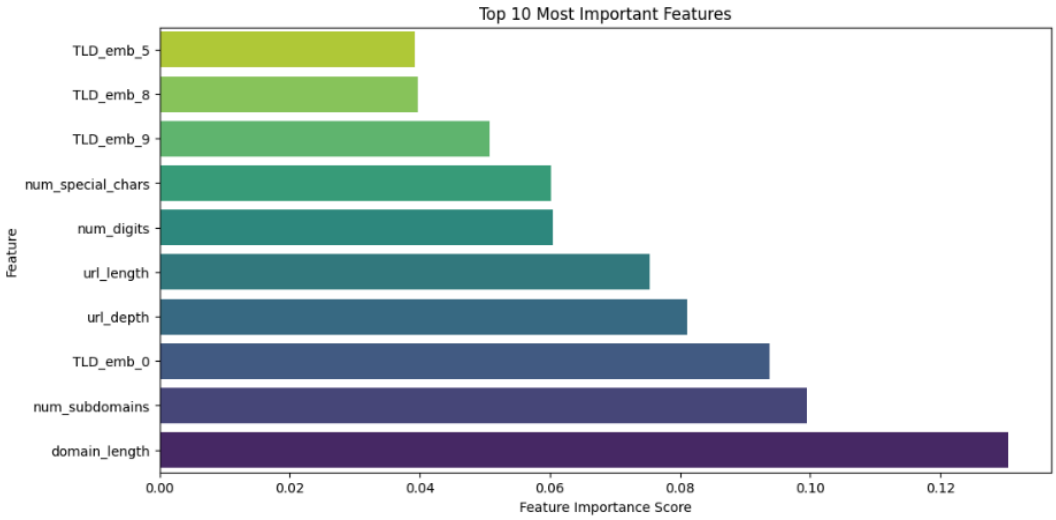
We were provided with five different classes that were highly imbalanced so we applied the SMOTE technique which synthetically generated data points.





Dimensionality Reduction

The dataset contained 20+ features which were too much for the models. So we decided to eliminate features and keep the 10 most important features.



Model Comparisons

XGBClassifier: 0.92
SVM: 0.91
Random Forest: 0.94
Custom Built Neural Network: 0.90
FTtransformer: 0.89

Comparative Analysis:

All of the models almost had the same testing accuracies. The reason traditional ML Models perform better is the reduced number of features. These models have optimized hyperparameters. If we had trained our custom neural models for a longer period of time, maybe the results would differ.