Team Analytix

VrikshaNeeti 2024

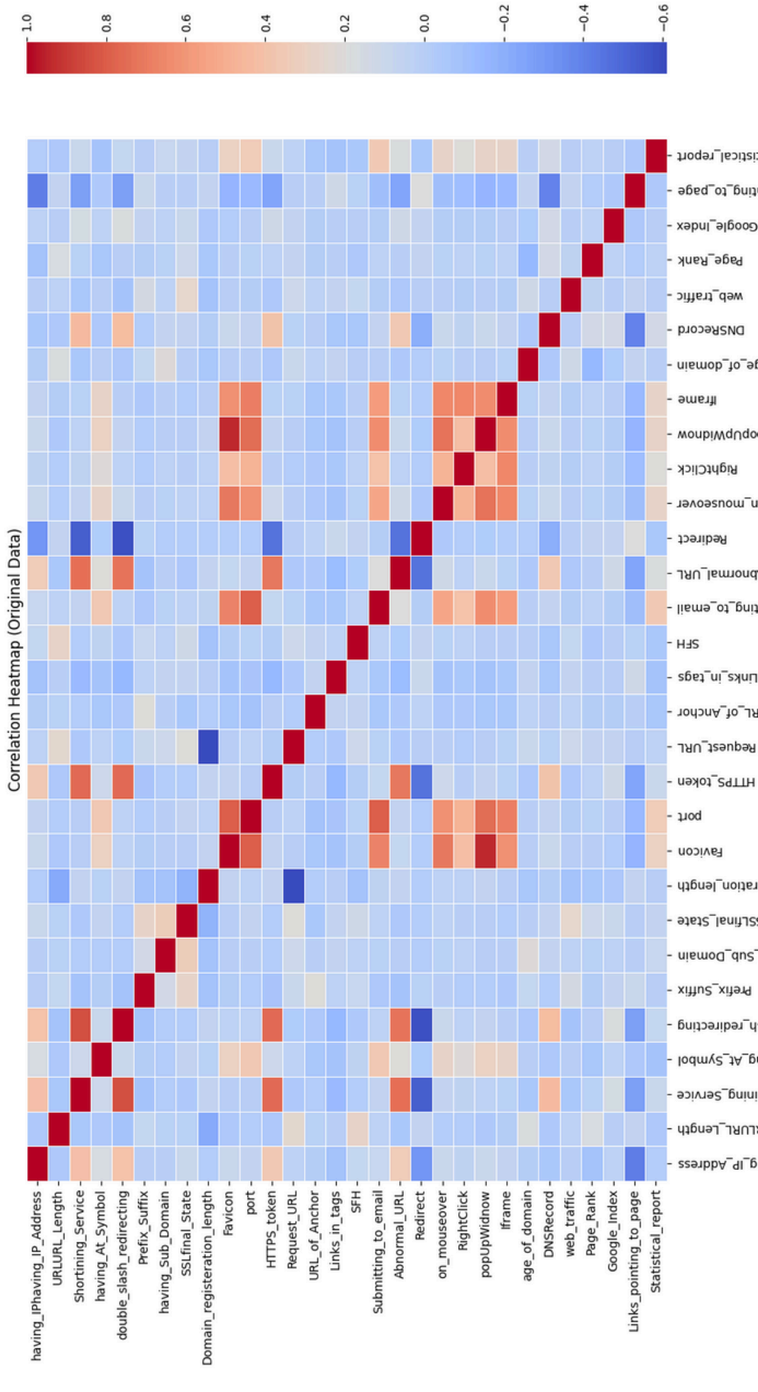
TISS MUMBAI

Problem Statement and Dataset Overview

- Problem Statement: ShopSecure faced a downfall due to security lapses exposing users to phishing scams, malware, and data breaches.
- Objective: Develop a machine learning-based phishing detection system to classify URLs as legitimate or malicious.
- Dataset Overview: Total Samples: 11,055 URLs Features: 32 (Encoded as -1: Suspicious, 0: Phishing, 1: Legitimate) Features highlight potential markers of phishing.

Task 1: Exploratory Data Analysis (EDA):

1. Plotted histograms to analyze feature distributions and identify patterns.
2. Generated a correlation heatmap to detect highly correlated features and redundancy.
3. Removed one feature from each highly correlated pair to prevent overfitting.
4. Reclassified the target variable (Result) into a binary format:
    - 0 (Malicious): Combined suspicious and phishing.
    - 1 (Legitimate): Safe URLs.



Binary Classification Models and Insights

- Task 2: Binary Classification Models
  - Models Trained: Decision Tree, Random Forest.
  - Data Split: 80% Training, 20% Testing.
- Key Insights:
  - Decision Tree:
    - AUC: ~0.98.
    - High interpretability; prone to overfitting without regularization.
  - Random Forest:
    - AUC: ~0.99.
    - Most accurate model; handles non-linear data effectively.
- K-Fold Cross-Validation:
  - Confirms models are statistically robust and generalize well.
  - Random Forest accuracy: 98.73% (low variability, high reliability).
- Conclusions:
  - Random Forest outperforms other models and is the recommended choice for deployment.
  - Phishing detection accuracy demonstrates high potential for real-world use.