INTISARI

Phishing adalah ancaman terhadap keamanan dunia maya yang sering terjadi, di mana para pelaku mencoba menipu pengguna untuk menyerahkan informasi pribadi seperti kata sandi, nomor kartu kredit, dan berbagai data sensitif lainnya. Dengan perkembangan teknologi, teknik phishing semakin canggih dan sulit terdeteksi oleh metode tradisional. Oleh karena itu, sangat penting untuk merancang teknik yang mampu mendeteksi situs phishing dengan akurasi yang tinggi. Penelitian ini bertujuan untuk mengoptimasi kinerja deteksi phishing dengan mengintegrasikan analisis korelasi variabel untuk seleksi fitur dan penggunaan teknik imbalance learning guna mengatasi ketidakseimbangan data. Tahapan penelitian mencakup Data Collection, Data Preprocessing, Data Exploration meliputi analisis korelasi, pembersihan Fitur yang berkorelasi rendah, dan visualisasi data. Pada tahap Model Building and Training, dilakukan pembagian fitur dan label, Training data, dan penerapan teknik penyeimbangan data, diakhiri dengan Model Evaluation. Algoritma yang diuji mencakup Logistic Regression, Naive Bayes, K-Nearest Neighbors, Support Vector Machine, Multi-Layer Perceptron, Decision Tree, Random Forest, Gradient Boosting, CatBoost. Hasil penelitian ini menunjukkan Algoritma KNN memberikan kinerja paling baik dengan akurasi mencapai 91,25%. serta hasil optimal pada metrik Precision, Recall, dan F1-Score, masing-masing sebesar 0,906943, 0,927858, dan 0,922141, serta Hamming Loss terendah sebesar 0,0875. Sebaliknya, SVM menunjukkan hasil terendah dibandingkan algoritma lainnya, sehingga kurang cocok digunakan untuk mendeteksi URL Phishing

Kata kunci: Deteksi Phishing, Optimasi Kinerja Model, Machine learning

ABSTRACT

Phishing is a common cybersecurity threat where attackers attempt to deceive users into providing personal information such as passwords, credit card numbers, and other sensitive data. As technology advances, phishing techniques have become more sophisticated and harder to detect using traditional methods. Therefore, it is crucial to develop techniques that can accurately identify phishing websites. This study aims to optimize phishing detection performance by integrating variable correlation analysis for feature selection and using imbalance learning techniques to address data imbalance issues. The research stages include Data Collection, Data Preprocessing, Data Exploration, which involves correlation analysis, the removal of low-correlation features, and data visualization. During the Model Building and Training phase, feature and label splitting, training data, and data balancing techniques were applied, followed by Model Evaluation. The tested algorithms include Logistic Regression, Naive Bayes, K-Nearest Neighbors, Support Vector Machine, Multi-Layer Perceptron, Decision Tree, Random Forest, Gradient Boosting, and CatBoost. The results indicate that the KNN algorithm achieved the best performance with an accuracy of 91.25%, along with optimal results in Precision, Recall, and F1-Score, with respective values of 0.906943, 0.927858, and 0.922141, and the lowest Hamming Loss of 0.0875. On the other hand, SVM showed the lowest performance compared to the other algorithms, making it less suitable for detecting phishing URL.

Keywords: Phishing Detection, Model Performance Optimization, Machine Learning