BINUS UNIVERSITY GRADUATE PROGRAM Master Program

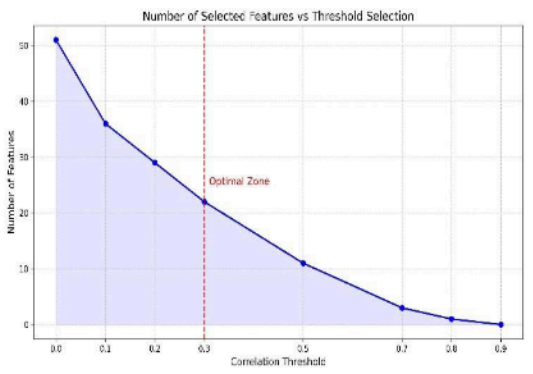
Master of Computer Science

Evaluation Feature Selection in Machine Learning Models for Malicious URL Detection

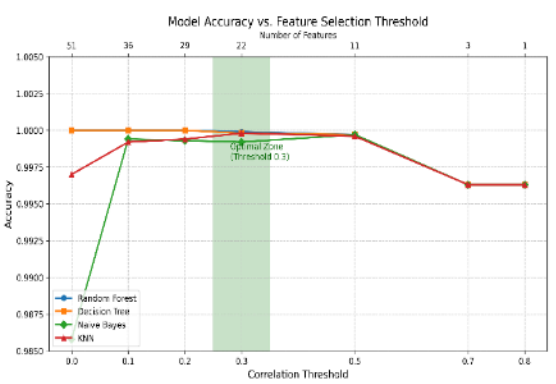
Background

Phishing attacks using malicious URLs are rapidly increasing and often bypass traditional blacklist filtering methods. Machine Learning (ML) offers better detection capabilities, yet the presence of high-dimensional and redundant features leads to significant computational overhead. To address this challenge, efficient feature selection is required to reduce redundancy while maintaining detection accuracy. Pearson correlation-based feature selection provides a practical approach to streamline features and optimize model performance. This study focuses on applying such techniques to a large phishing URL dataset to support scalable and real-time cybersecurity solutions.

Result & Discussion

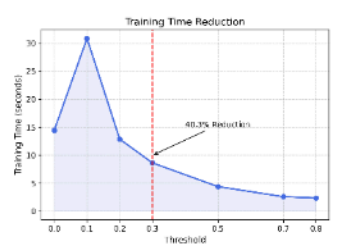


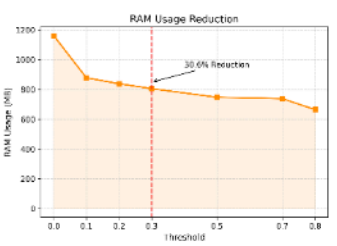
The image shows the relationship between the correlation threshold and the number of selected features. At a threshold of 0.0, all 51 features are retained, while at 0.3 this number is reduced to 22, and by 0.7 fewer than five remain. This demonstrates the strong impact of threshold selection, with 0.3 identified as the optimal balance between reducing dimensionality and maintaining model accuracy.



The image illustrates the effect of feature selection on model accuracy across Random Forest, Decision Tree, Naïve Bayes, and KNN. Accuracy remains stable when the correlation threshold is set at 0.3, showing that fewer features can be used without sacrificing performance.

This confirms the threshold of 0.3 as the optimal zone for balancing accuracy and efficiency.





Feature selection in Random Forest not only influenced model accuracy but also significantly enhanced computational efficiency. At the optimal threshold of 0.3, training time was reduced by about **40%** and RAM usage by about **30%**, demonstrating that feature selection significantly improves computational efficiency while maintaining high accuracy.

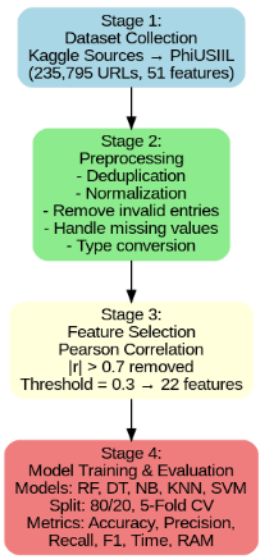
Conclusions

Feature selection with Pearson correlation improves efficiency without reducing accuracy, reducing features from 51 to 22. Future work will integrate the model into real-time security systems and explore deep learning for advanced phishing detection.

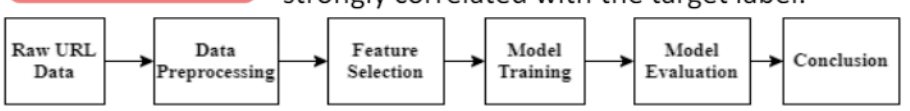
Previous Works

Extensive studies have addressed phishing URL detection using machine learning. Vajrobol et al. applied Mutual Information feature selection with logistic regression on the PhiUSIIL dataset, obtaining 99.97% accuracy using only five critical features. Mohanty et al. introduced the Multivariate Filter-Based Feature Selection Technique (MFBFST), combining correlation-based feature selection with t-tests, reporting accuracies of 97–99.25% on UCI and Kaggle datasets, though at the expense of high computational costs. Chinnasamy et al. experimented with Random Forest, SVM, and Genetic Algorithms for feature reduction, where the Genetic Algorithm achieved 94.73% accuracy. Sangra et al. utilized Pearson correlation for lexical feature selection on 10,000 URLs, with Random Forest reaching 95.3% accuracy and improved efficiency. More recently, Wazirali et al. employed RFE-SVM integrated with SDN, achieving 99.05% accuracy but with increased memory usage.

Methodology



The research methodology is structured into four main stages. In the first stage, phishing URL datasets were collected from multiple Kaggle sources, with the PhiUSIIL dataset (235,795 URLs and 51 features) selected for its balanced labels and rich feature set. The preprocessing stage involved deduplication, normalization, removal of invalid entries, conversion into structured numerical features, and handling of missing values, ensuring clean and consistent data for modeling. In the third stage, feature selection using Pearson correlation was applied in two steps: eliminating highly correlated features ($|r| > 0.7$) and selecting features strongly correlated with the target label.



An optimal threshold of 0.3 was chosen, reducing the feature set from 51 to 22 while maintaining accuracy and efficiency. Finally, in the model training and evaluation stage, five classifiers Random Forest, Decision Tree, Naïve Bayes, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) were trained and tested using an 80/20 stratified split and 5-fold cross-validation. Their performance was assessed using accuracy, precision, recall, F1-score, training time, and RAM usage, providing a comprehensive evaluation of both predictive capability and computational efficiency..

Selected References

An A. Prasad and S. Chandra, "PhiUSIIL: A diverse security profile empowered phishing URL detection framework based on similarity index and incremental learning," Comput Secur, vol. 136, Jan. 2024, doi: 10.1016/j.cose.2023.103545

Student

Hanung Febrianto
2502482973

Thesis Advisor

Dr. Aditya Kurniawan, S.Kom., MMSI. - D3448