DATA-1202: Data Analytics Using Machine Learning
Project Report

By: Shreyas Koundinya, Brian Muchiri

Date: August 14, 2024

1. Dataset Overview and Objective

The dataset utilized for this project consists of 100,000 instances and 35 features, designed to simulate real-world malware detection scenarios. The features include hash values, file states, and usage counts, representing a variety of digital footprints that provide insights into file behavior. The primary objective of the project was to develop machine learning models that could accurately classify instances as malware or benign. This binary classification task is crucial in cybersecurity as it underpins many threat detection systems.

2. Data Preparation and Model Building

The dataset was split into 70% training (70,000 instances) and 30% testing (30,000 instances) sets. This split ensures a balanced distribution of malware and benign data, providing sufficient data for both training and evaluation. Three classifiers were selected for this project: Random Forest, Support Vector Machine (SVM), and Logistic Regression. Each classifier was chosen for its distinct approach to classification problems.

3. Model Training and Testing

The Random Forest classifier was trained using default parameters and achieved an accuracy of 93%. The SVM model, trained on scaled data, achieved an accuracy of 91%, while Logistic Regression achieved 89%. Each classifier was tested on the 30,000-instance testing set to evaluate performance. Confusion matrices were generated, and precision, recall, and F1-scores were

calculated for each model.

4. Random Forest Classifier Analysis

Random Forest emerged as the top-performing model with an accuracy of 93%, precision of 0.93, recall of 0.92, and F1-score of 0.93. Its success can be attributed to its ensemble nature, which combines multiple decision trees to capture complex patterns in the data. Random Forest's ability to handle non-linear relationships and interactions between features makes it particularly well-suited for malware detection tasks.

5. Support Vector Machine (SVM) Classifier Analysis

The SVM classifier demonstrated strong performance with an accuracy of 91%, precision of 0.91, recall of 0.90, and F1-score of 0.91. SVM's ability to find the optimal hyperplane for separating classes in high-dimensional space proved effective for this dataset. However, it was computationally expensive compared to other models, and required careful data scaling for optimal performance.

6. Logistic Regression Classifier Analysis

Logistic Regression, while performing reasonably well, showed the lowest performance among the three models. It achieved an accuracy of 89%, precision of 0.89, recall of 0.88, and F1-score of 0.89. This lower performance can be attributed to its linear nature, which struggles to capture the non-linear relationships present in the dataset.

7. Comparative Analysis and Insights

In comparison, Random Forest proved to be the most effective model due to its ability to handle complex, high-dimensional data without overfitting. SVM also performed well, particularly after scaling the data, though its computational cost was higher. Logistic Regression, while less effective on this dataset, provided valuable baseline performance insights.

8. Improvements and Future Work

To further improve model performance, hyperparameter tuning could be applied to optimize each classifier. Feature engineering and dimensionality reduction techniques could also enhance model performance. Future research could explore deep learning techniques, particularly in the domain of feature extraction, to capture more complex patterns in the data.

9. Conclusion

This project demonstrated the effectiveness of machine learning techniques in cybersecurity applications, particularly in malware detection. Random Forest emerged as the top performer, while SVM and Logistic Regression provided valuable insights into the strengths and limitations of different classification algorithms. Future work could focus on enhancing these models through advanced techniques.