# **Malware e Segurança Digital: Um Estudo a partir do Dataset TUNADROMD**

## **1. Introdução**

Nas últimas décadas, o ecossistema digital expandiu-se de forma exponencial, conectando dispositivos, sistemas críticos e dados sensíveis em escala global. Esse crescimento trouxe, entretanto, uma escalada nos ataques cibernéticos, especialmente por meio de softwares maliciosos (malware).

Segundo relatórios internacionais de cibersegurança, variantes de malware evoluem continuamente, utilizando técnicas cada vez mais sofisticadas para evadir sistemas de defesa, explorar vulnerabilidades e comprometer a integridade de redes corporativas e pessoais.

Malwares não representam apenas um risco financeiro, mas também uma ameaça à segurança nacional, à privacidade de indivíduos e à continuidade operacional de organizações estratégicas. A crescente diversidade dessas ameaças — incluindo ransomware, trojans, worms e spywares — reforça a necessidade de estudos sistemáticos que permitam entender padrões, classificar comportamentos e desenvolver métodos eficazes de detecção.

Nesse contexto, conjuntos de dados específicos para análise de malware têm papel central na pesquisa científica e no desenvolvimento de sistemas de defesa. Eles permitem a avaliação comparativa de algoritmos, o teste de robustez de métodos de detecção e a validação de novas abordagens baseadas em inteligência artificial e aprendizado de máquina.

O presente trabalho insere-se nesse cenário, apoiando-se no dataset **TUNADROMD**, disponibilizado pela *University of California, Irvine (UCI Machine Learning Repository)*, que reúne características de binários de malware para pesquisa e validação de técnicas de detecção (UCI, 2024).

## **2. Escopo e Objetivos**

O objetivo central deste estudo é investigar e validar a efetividade de métodos de detecção de malware utilizando o dataset TUNADROMD. Aplicando modelos de Machine learning supervisionados.

### **Escopo da pesquisa**

* **Análise exploratória** do dataset, compreendendo suas variáveis e distribuição das instâncias.
* **Definição de técnicas de detecção** baseadas em aprendizado de máquina, buscando identificar padrões ocultos que possam discriminar comportamentos maliciosos.
* **Discussão da importância de datasets dedicados** na evolução da pesquisa em segurança digital, destacando sua relevância para a replicabilidade de experimentos e avanço científico.

Este artigo contribui, portanto, tanto para a avaliação de métodos de detecção quanto para a consolidação de um corpo de conhecimento voltado à mitigação dos riscos associados a malwares no cenário digital contemporâneo.

## **3. Análise do Dataset**

### **3.1 Distribuição das Classes**

* **Label 1.0 (malware):** 3565 amostras (79,86%)
* **Label 0.0 (não-malware):** 899 amostras (20,24%)

Essa desproporção representava risco de **viés do modelo**, pois ele poderia tender a prever apenas a classe majoritária (1.0).

### **3.2 Estratégia de Balanceamento**

Foi adotado o **Oversampling da classe minoritária (0.0)** para atingir equilíbrio **50/50**.

* Classe 1.0: 3565 instâncias
* Classe 0.0: 3565 instâncias

## **4. Experimentos e Resultados**

Os dados foram divididos em **80% treino e 20% teste**, resultando em:

* Treino: 3571 instâncias
* Teste: 893 instâncias

Foram comparados três modelos principais: **Regressão Logística, Random Forest e Gradient Boosting**.

### **4.1 Regressão Logística**

* **Acurácia:** 0.9843
* **Precisão:** 0.9943
* **Recall:** 0.9860
* **F1-Score:** 0.9902

**Matriz de Confusão**

[170 4]

[ 10 709]

### **4.2 Random Forest**

* **Acurácia:** 0.9944
* **Precisão:** 0.9958
* **Recall:** 0.9972
* **F1-Score:** 0.9965

**Matriz de Confusão**

[171 3]

[ 2 717]

### **4.3 Gradient Boosting**

* **Acurácia:** 0.9866
* **Precisão:** 0.9876
* **Recall:** 0.9958
* **F1-Score:** 0.9917
* **MSE:** 0.0134
* **ROC-AUC:** 0.9990

**Matriz de Confusão**

[165 9]

[ 3 716]

## **5. Comparação dos Modelos**

| **Modelo** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **MSE** |
| --- | --- | --- | --- | --- | --- |
| Regressão Logística | 0.9944 | 0.9958 | 0.9972 | 0.9965 | 0.0056 |
| Random Forest | 0.9944 | 0.9958 | 0.9972 | 0.9965 | 0.0056 |
| Gradient Boosting | 0.9866 | 0.9876 | 0.9958 | 0.9917 | 0.0134 |

### **Melhores modelos por métrica**

* **Accuracy:** Regressão Logística & Random Forest (0.9944)
* **Precision:** Regressão Logística & Random Forest (0.9958)
* **Recall:** Regressão Logística & Random Forest (0.9972)
* **F1-Score:** Regressão Logística & Random Forest (0.9965)

## **6. Comparação Adicional: Regressão Logística com PCA**

### **6.1 Resultados**

**Regressão Logística COM PCA**

* **Acurácia:** 0.9937
* **Precisão:** 0.9986
* **Recall:** 0.9889
* **F1-Score:** 0.9937
* **MSE:** 0.0063
* **ROC-AUC:** 0.9982

### **6.2 Tabela Comparativa (Sem PCA vs Com PCA)**

| **Métrica** | **Sem PCA (originais)** | **Com PCA** |
| --- | --- | --- |
| Acurácia | 0.9944 | 0.9937 |
| Precisão | 0.9958 | 0.9986 |
| Recall | 0.9972 | 0.9889 |
| F1-Score | 0.9965 | 0.9937 |
| MSE | 0.0056 | 0.0063 |
| ROC-AUC | — | 0.9982 |

### **6.3 Redução de Dimensionalidade**

* **Dimensões originais:** 241
* **Dimensões após PCA:** 66
* **Redução:** 72,6%
* **Variância explicada:** 95,01%

### **6.4 Conclusões sobre PCA**

* O PCA manteve **alta performance** mesmo com número reduzido de variáveis.
* Houve **redução significativa** de dimensionalidade (72,6%), preservando 95% da variância.
* O processo de treinamento tornou-se mais rápido e computacionalmente eficiente.

## **7. Conclusões Gerais**

Os experimentos confirmam que **Regressão Logística e Random Forest** obtiveram as melhores métricas gerais, superando o Gradient Boosting.

Além disso, a análise com **PCA** demonstrou que é possível **reduzir o número de features de 241 para 66** sem perda significativa de desempenho, o que representa um ganho prático relevante em termos de eficiência computacional para sistemas de detecção em larga escala.

Assim, os achados principais são:

* **Modelos mais indicados:** Regressão Logística e Random Forest.
* **PCA como recurso adicional:** eficiente para otimização, sem comprometer a robustez do modelo.

## **8. Referências**

* UCI Machine Learning Repository. *TUNADROMD Dataset*. Disponível em: https://archive.ics.uci.edu/dataset/813/tunadromd. Acesso em: 2024.
* Relatórios de Cibersegurança (diversas fontes internacionais, 2023-2024).