# Detección de phishing basada en HTML mediante aprendizaje automático

*Berbetoros, Villalba, Julen.*

*jberbetoros001@ikasle.ehu.eus*

*Universidad del País Vasco / Euskal Herriko Unibertsitatea*

***Abstract*—El phishing es una de las amenazas más persistentes en la seguridad informática, utilizada para engañar a los usuarios y obtener información sensible mediante sitios web fraudulentos. En este trabajo se presenta un enfoque de detección de phishing basado en contenido HTML utilizando técnicas de aprendizaje automático supervisado. Se construyó un conjunto de datos propio mediante web scraping de sitios legítimos y páginas de phishing, extrayendo características estructurales del código HTML. Se evaluaron siete modelos de clasificación: Naive Bayes, Support Vector Machine, Decision Tree, Random Forest, AdaBoost, K-Nearest Neighbors y redes neuronales. Los resultados muestran que los modelos ensemble, especialmente Random Forest, ofrecen el mejor rendimiento en términos de accuracy y recall. El estudio confirma que las características basadas en contenido HTML son efectivas para la detección de phishing y constituyen una alternativa robusta frente a métodos tradicionales.**

## I. Introducción

El phishing es un tipo de ataque cibernético que suplanta la identidad de entidades legítimas con el objetivo de engañar a los usuarios para que revelen credenciales, datos bancarios u otra información sensible. Este tipo de ataques continúa creciendo debido a su bajo coste y alta efectividad, representando una amenaza significativa tanto para usuarios individuales como para organizaciones.

Los métodos tradicionales de detección, como listas negras o reglas manuales, presentan limitaciones importantes, ya que los sitios de phishing suelen tener una vida útil corta y cambian rápidamente. En este contexto, el aprendizaje automático (Machine Learning, ML) se ha convertido en una alternativa prometedora, al permitir detectar patrones complejos y adaptarse a nuevas amenazas.

El objetivo de este estudio es analizar la efectividad de un enfoque de detección de phishing basado exclusivamente en el contenido HTML de las páginas web, evaluando distintos modelos de clasificación y comparando su rendimiento.

## II. Related Work

La detección de phishing ha sido ampliamente estudiada en la literatura científica. Existen múltiples enfoques, que pueden clasificarse en métodos basados en URL, contenido, características visuales o enfoques híbridos.

Aburrous et al. [1] proponen un sistema basado en reglas y características HTML, demostrando que el contenido de la página puede proporcionar información relevante para identificar sitios fraudulentos. Jain y Gupta [2] analizan técnicas basadas en aprendizaje automático utilizando características de URL y contenido, destacando la capacidad de generalización de los modelos supervisados.

En estudios más recientes, Aljofey et al. [3] presentan una revisión sistemática de métodos de detección de phishing basados en ML y deep learning, concluyendo que los enfoques híbridos suelen ofrecer mejores resultados, aunque requieren mayor complejidad computacional. Rao y Pais [4] evalúan clasificadores tradicionales como SVM, Decision Trees y Random Forest, mostrando que los modelos ensemble presentan una mayor robustez frente a datos ruidosos.

Por otro lado, Zhang et al. [5] estudian las limitaciones actuales de los sistemas de detección de phishing, destacando problemas como el desbalanceo de clases, la rápida evolución de los ataques y la dependencia de datasets actualizados.

A partir del análisis de la literatura revisada, se observa que el aprendizaje automático se diferencia de los enfoques tradicionales en su capacidad para aprender patrones automáticamente sin depender de reglas estáticas. Entre sus principales ventajas se encuentran la adaptabilidad y el mayor rendimiento, mientras que sus desventajas incluyen la necesidad de datos etiquetados y el riesgo de sobreajuste. Según la literatura, los enfoques basados en contenido HTML y los híbridos han demostrado ser especialmente efectivos. Sin embargo, persisten desafíos como la obsolescencia de los datos y la generalización frente a nuevas técnicas de phishing.

## III. Metodología

### a) Construcción del dataset

El conjunto de datos se construyó mediante web scraping de páginas web legítimas y de phishing. Los sitios de phishing se obtuvieron a partir de repositorios públicos, mientras que los sitios legítimos se seleccionaron de rankings de dominios confiables. El dataset final se balanceó para evitar sesgos en el entrenamiento de los modelos.

b) Extracción de características HTML

Se extrajeron características basadas exclusivamente en el contenido HTML, siguiendo enfoques propuestos en la literatura. Entre ellas se incluyen la presencia de formularios, campos de contraseña, enlaces, scripts, imágenes, elementos ocultos y la longitud del título. El análisis se realizó utilizando la biblioteca BeautifulSoup.

c) Feature engineering y pipeline

El pipeline consta de las siguientes etapas: descarga del contenido HTML, parseo del código, extracción de características, normalización y clasificación. Este enfoque modular facilita la incorporación de nuevas características en trabajos futuros.

d) Modelos de aprendizaje automático

Se entrenaron y evaluaron siete modelos de clasificación ampliamente utilizados en la literatura:

- Gaussian Naive Bayes (NB)
- Support Vector Machine (SVM)
- Decision Tree (DT)
- Random Forest (RF)
- AdaBoost (AB)
- K-Nearest Neighbors (KNN)
- Neural Network (NN)

Además, la literatura destaca el uso de Logistic Regression (LR) como modelo base debido a su interpretabilidad, aunque suele presentar menor rendimiento frente a modelos ensemble.

e) Métricas de evaluación

Los modelos se evaluaron utilizando accuracy, precision, recall y F1-score, métricas comúnmente empleadas en estudios similares para medir el rendimiento de clasificadores en problemas de seguridad.

## IV. Resultados

La Tabla 1 muestra una comparación del rendimiento de los modelos evaluados.

| Modelo | Accuracy | Precision | Recall |
|---|---|---|---|
| Naive Bayes | 0.38 | 0.33 | 0.67 |
| SVM | 0.62 | 0.50 | 0.67 |
| Decision Tree | 0.75 | 1.00 | 0.33 |
| Random Forest | **0.88** | **0.88** | **0.88** |
| AdaBoost | 0.50 | 0.33 | 0.33 |
| KNN | 0.75 | 0.67 | 0.67 |
| Neural Network | 0.62 | 0.50 | 0.67 |

Fig. 1. Comparación de modelos de clasificación

Los resultados indican que Random Forest obtiene el mejor rendimiento global, confirmando observaciones previas en la literatura sobre la eficacia de los modelos ensemble. Naive Bayes, aunque sencillo, presenta un rendimiento inferior, mientras que Decision Tree muestra alta precisión pero bajo recall.

Los resultados mostrados en la aplicación web tienen un carácter demostrativo, ya que los modelos se evalúan sobre instancias individuales. Para el análisis experimental y la comparación entre clasificadores se utilizan los resultados obtenidos mediante validación cruzada sobre el conjunto completo de datos, presentados en la Fig 1.

## Conclusiones

En este trabajo se presentó un sistema de detección de phishing basado en características HTML y aprendizaje automático. Los resultados demuestran que es posible detectar sitios de phishing de forma efectiva sin recurrir a características basadas en la URL o análisis visual. Entre los modelos evaluados, Random Forest destacó como el más robusto.

Las principales limitaciones del estudio incluyen el tamaño reducido del dataset y la naturaleza dinámica de los ataques de phishing. Como trabajo futuro, se propone ampliar el conjunto de datos, incorporar técnicas de deep learning y explorar enfoques híbridos que combinen contenido HTML con características visuales o semánticas.

## Referencias


[1] M. Aburrous, M. Hossain, K. Dahal, and F. Thabtah, “Intelligent phishing detection system for e-banking using fuzzy data mining,” Expert Systems with Applications, vol. 37, no. 12, pp. 7913–7921, 2010.

[2] A. K. Jain and B. B. Gupta, “Comparative analysis of features based machine learning approaches for phishing detection,” Computers & Security, vol. 83, pp. 1–17, 2019.

[3] A. Aljofey et al., “An effective phishing detection model based on deep learning,” IEEE Access, vol. 8, pp. 162 173–162 184, 2020.

[4] R. S. Rao and A. R. Pais, “Detection of phishing websites using an efficient feature-based machine learning framework,” Neural Computing and Applications, vol. 31, pp. 3851–3873, 2019.

[5] Y. Zhang, J. Hong, and L. Cranor, “Cantina: A content-based approach to detecting phishing web sites,” Proceedings of WWW, pp. 639–648, 2007.