Diseño de un modelo predictivo para la detección de URLs maliciosas con técnicas de Machine Learning

JHON SEBASTIAN USUGA FERRARO, jhon.usuga1@udea.edu.co
SEBASTIAN RENTERIA PALACIOS, sebastian.renteria@udea.edu.co

26 de octubre de 2025

1. Descripción del problema

1.1. Contexto y utilidad del problema

La detección de sitios *phishing* es un problema crítico, y que cada vez ha ganado más relevancia dentro del ámbito de la seguridad informática, ya que millones de usuarios son víctimas cada año de páginas falsas que suplantan sitios legítimos, resultando en pérdidas financieras, el robo de información sensible y la pérdida de la confianza en las instituaciones que manejan estos datos.

El conjunto PhiUSIIL Phishing URL Dataset fue creado con el propósito de ofrecer una base estandarizada para entrenar y evaluar modelos de detección automática de phishing. Este dataset integra características léxicas, de contenido y de red, lo que permite a los algoritmos aprender patrones que distinguen una URL real de una fraudulenta.

Automatizar este proceso mediante *Machine Learning* permite reducir el tiempo de respuesta ante amenazas, permitiendo la detección proactiva de ataques desconocidos (*zero-day attacks*) y complementar las soluciones basadas en listas negras, que suelen quedar obsoletas con rapidez frente a la evolución constantey las sofistificación de las tácticas de los atacantes.

Debido a su estructura limpia y su número elevado de registros, el dataset PhiUSIIL se ha convertido en una referencia para estudios de detección de phishing con modelos de *Machine Learning* y *Deep Learning*.

1.2. Análisis del conjunto de datos

El conjunto PhiUSIIL Phishing URL Datasetcontiene 235,795 registros, de los cuales cada uno representa una URL etiquetada como:

- 1:Sitio legítimo
- 0:Sitio de phishing

Cada registro incluye 55 atributos que describen propiedades de la URL y de la página web asociada con su respectiva etiqueta, de estos, 51 son de tipo númerico, y 4 de tipo categórico (FILENAME, URL, TLD y Title). Entre las categorías de características más relevantes se encuentran:

Características léxicas: longitud de la URL (URLLength), número de subdominios (NoOfSubDomain), presencia de caracteres sospechosos como “@”, “//”, “-” (NoOfOtherSpecialCharsInURL).

Características basadas en contenido HTML:* Número total de líneas de código (LineOfCode), la existencia de un título (HasTitle) y su relación con el dominio o la URL (DomainTitleMatchScore, URLTitleMatchScore); la presencia de etiquetas visuales o de contenido como imágenes (NoOfImage), hojas de estilo (NoOfCSS) y scripts (NoOfJS).

Características derivadas del dominio: La detección de una dirección IP en lugar de un nombre de dominio (IsDomainIP), TLD y TLDLenght, y TLDLegitimateProb que evalúa la parte final de las URL (.gov, .com, .net), presencia de HTTPS (IsHTTPS) para validar si le sitio posee un certificado.

El dataset fue recopilado a partir de múltiples fuentes y validado manualmente. el análisis exploratorio evidenció que las distribuciones tienen una alta asimetría, con muchos valores concentrados en el valor de 0 y presencia de valores extremos, por lo que para algunas técnicas se tendría que aplicar transformaciones logarítmicas o algún otro método de escalado, además no contiene valores faltantes y está balanceado entre clases con 57 % legítimas vs 43 % phishing.

Esta estructura lo hace adecuado para evaluar distintos algoritmos dentro del paradigma de aprendizaje supervisado. Los modelos basados en árboles, como *Random Forest*, *XGBoost* y *LightGBM*, pueden entrenarse directamente sobre las variables originales, mientras que modelos sensibles a la escala, como *SVM*, *Logistic Regression* y redes neuronales profundas, requieren previamente un proceso de normalización o escalado.

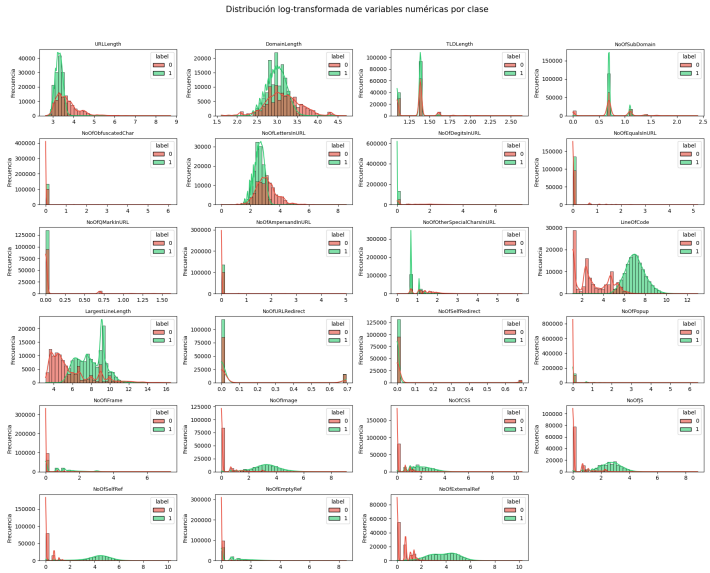

Figura 1: Distribución log-transformada de variables numéricas por clase

1.3. Paradigma de aprendizaje y justificación

El proyecto se enmarca en el aprendizaje supervisado con una tarea de clasificación binaria, cuyo objetivo es distinguir entre URLs legítimas y maliciosas (*phishing*). Dado que cada muestra cuenta con una etiqueta previamente asignada (0 o 1), los modelos pueden aprender patrones discriminativos a partir de los atributos descritos.

Las métricas empleadas para la evaluación del rendimiento incluirán:

- Accuracy
- Precision
- Recall
- F1-score

La métrica *F1-score* será especialmente importante para reducir los falsos negativos, es decir, los casos en los que un sitio de phishing se clasifica erróneamente como legítimo, con el uso de este dataset se permitirá comparar diversos modelos de aprendizaje automático, con técnicas de optimización y validación cruzada, asegurando una evaluación rigurosa, reproducible y aplicable a contextos reales de ciberseguridad.

2. Estado del arte

2.1. Vajrobol et al. [VGG24]

Dado que la base de datos PhiUSIIL contiene ejemplos previamente etiquetados como URLs legítimas y URLs maliciosas, los diferentes trabajos abordados se enmarcan bajo el paradigma de aprendizaje supervisado, en donde el modelo aprende a distinguir entre las clases a partir de las etiquetas proporcionadas. Para ello, se utiliza la técnica de regresión logística en conjunción con selección de características basada en información mutua. Los autores resaltan que la combinación de estas técnicas hacen al modelo simple y eficiente, facilitando su interpretabilidad y manteniendo un equilibrio entre precisión y simplicidad. Como metodología de validación, se particionó el conjunto de datos a un ratio de 80:20. El desempeño del modelo fue evaluado mediante las métricas de *Accuracy*, *Precision*, *Recall* y *F1-score*, y se complementó con matrices de confusión. En la conclusión, lograron una *accuracy* del 99.97 % alcanzando esos resultados sólo con 5 características: "URLSimilarityIn- dex," "LineOfCode," "NoOfExternalRef," "NoOfImage," y "NoOfSel- fRef.". Con ello, concluyeron que la técnica de selección de características Mutual Information, en conjunto con la regresión logística, puede ser un modelo muy útil, confiable, preciso y fácil de entender para tareas de identificación de URLs maliciosas.

2.2. Sruthi K. & Manohar Naik S. [KN25]

Este estudio también emplea un enfoque supervisado, por las mismas razones descritas anteriormente, dado que se usó el mismo conjunto de datos. En este caso, se utiliza una red neuronal siamesa con subredes LSTM. Esta red neuronal procesa pares de URLs que, a través de cada subred LSTM, generan una representación que se compara posteriormente mediante una medida de distancia euclidiana, con el fin de determinar la similitud entre ellas. De este modo, el modelo aprende a reducir la distancia entre pares similares y a aumentarla entre pares disimilares. Se empleó la metodología de validación cruzada *5-fold cross-validation*, la cual fue complementada con estudios de ablación y sensibilidad para medir la estabilidad del modelo. El desempeño del modelo fue evaluado mediante las métricas de *Accuracy*, *Precision*, *Recall* y *F1-score*. En los conjuntos estándar, se obtuvo una *accuracy* del 99.68 %, y además, el modelo demostró tener un desempeño similar para muestras generadas usando GPT 4.0, lo que sugiere que podría ser una herramienta efectiva para mitigar los ciberataques de phishing basados en IA.

2.3. Manguli et al. [MKPR25]

Para este trabajo, se utilizaron los conjuntos de datos PhishTank para la obtención de URLs maliciosas y de Alexa para la obtención de URLs legítimas, lo que, por ende, lo clasifica como un problema de aprendizaje supervisado. La metodología empleada es una Graph Neural Network, específicamente el algoritmo GraphSAGE, que tiene la capacidad de generalizar a nuevos nodos o grafos no vistos durante el entrenamiento. Se empleó la metodología de validación cruzada 5-fold cross-validation. Asimismo, se evaluó el desempeño mediante las métricas de *Accuracy*, *Precision*, *Recall*, *F1-score* y *AUC-ROC*. El modelo GraPhish obtuvo una *accuracy* del 98.96 %, por lo que resultó ser una gran alternativa cuando se busca reducir la dependencia en el contenido del mensaje.

2.4. Fatma Hendaoui & Saloua Hendaoui [HH24]

Propusieron un sistema denominado *SENTINEY*, que combina enfoques de aprendizaje supervisado y no supervisado. Para el aprendizaje no supervisado se implementaron técnicas como *Isolation Forest*, *One-Class SVM*, *Elliptic Envelope*, *K-Means*, *DBSCAN*, *Agglomerative Clustering*, *Birch* y *GMM*; mientras que para el aprendizaje supervisado se emplearon *MLP*, *Random Forest* y *Gradient Boosting*. Dado el diseño del sistema basado en *SMPC*, no se realizó una validación estadística formal, sino pruebas en entornos controlados y distribuidos. Las métricas utilizadas fueron *Accuracy*, *Precision*, *Recall*, *F1-score*, *Average Detection Time* y *Silhouette Score*. El sistema alcanzó una *accuracy* del 99.4 % y un tiempo promedio de detección de 0.89 segundos por correo, concluyendo que el módulo supervisado fue más eficaz en la detección de correos de *phishing* conocidos, mientras que el módulo no supervisado resultó más eficiente para identificar *zero-day attacks*.

Referencias

[HH24] Fatma Hendaoui and Saloua Hendaoui. Sentiney: Securing encrypted multi-party computation for enhanced data privacy and phishing detection. *Expert Systems with Applications*, 256:124896, 2024.

[KN25] Sruthi K and Manohar Naik S. A novel framework for effective phishing url detection using an lstm-based siamese network. *Knowledge-Based Systems*, 329:114271, 2025.

[MKPR25] Kartik Manguli, Cheemaladinne Kondaiah, Alwyn Roshan Pais, and Routhu Srinivasa Rao. Graphish: A graph-based approach for phishing detection from encrypted tls traffic. *Journal of Information Security and Applications*, 94:104216, 2025.

[VGG24] Vajratiya Vajrobol, Brij B. Gupta, and Akshat Gaurav. Mutual information based logistic regression for phishing url detection. *Cyber Security and Applications*, 2:100044, 2024.