Sistema de clasificación de sitios web de phishing costo eficiente basado en aprendizaje automático para diferentes tipos de despliegue en 2023.

Presentado por:

Javier Arturo Velandia Yepes
Ingeniero electrónico

Trabajo de grado

Facultad de ingeniería
Maestría en analítica aplicada

Directora: Luz Helena Mancera Méndez Msc.
Codirectora: Johana Maria Florez Lozano PhD

2023

Página de aceptación

Chía

27 de octubre de 2023

Dedicatoria

A todos aquellos inconformes que se atreven a retar la normalidad con actos de profundo amor y entrega a sus semejantes y no semejantes.

Agradecimientos

Con profundo agradecimiento a mi familia quién cedió su tiempo y duplicó sus esfuerzos en pro de alcanzar esta nueva meta. A mis compañeros de trabajo por ayudarme a entender las dinámicas de este flagelo y a mis directoras por apoyarme y guiarme en los momentos de mayor incertidumbre.

Tabla de contenido

Lista de figuras

Glosario

Hosting: Es el servicio que almacena y pone a disposición de los usuarios del contexto de navegación al que sirve recursos, tales como páginas web, aplicaciones, documentos para descarga e interacción, entre otros.

SOC: El centro de operaciones de seguridad o SOC por sus siglas en inglés es el lugar donde convergen agentes que trabajan en pro de la detección y gestión de amenazas y riesgos de ciberseguridad. Estos centros de operaciones reciben miles e incluso millones de alertas a diario para ser evaluadas y calificadas como potenciales amenazas o falsas alertas.

Spear Phishing: Tipo de campaña de phishing que se dirige a una persona o grupo específico y que a menudo incluye información que se sabe que es de interés para el objetivo, como eventos actuales o documentos financieros.

General Phishing: Tipo de campaña de phishing abierta con un público objetivo amplio y que no suele exhibir grandes detalles de adaptación.

Hipervínculo (*link*): Es un tipo de etiqueta HTML que permite hacer direccionamientos entre sitios web.

Free hosting: Alojamiento para sitios web de muy bajo o nulo pago que hace uso de infraestructura compartida y subdominios para hacer publicaciones de los sitios web allí alojados.

URL: Localizador Uniforme de Recursos. Una URL es la dirección de un recurso único y determinado en la Web.

Abusebox: Un buzón de correo de abuso es donde los usuarios informan sobre amenazas de correo electrónico, como ataques de phishing, que se envían para su análisis. Los analistas de SOC utilizan el buzón de abuso para clasificar, analizar y responder a los correos electrónicos sospechosos informados por los usuarios.

Arañas o rastreadores (*crawlers)*: Un rastreador web, a veces llamado araña o *spiderbot* y a menudo abreviado como rastreador, es un robot de Internet que navega sistemáticamente por *la World Wide Web* y que normalmente es operado por motores de búsqueda con el fin de indexar la web (*web spidering*).

Web Scraping: es el proceso de utilizar herramientas automatizadas (*bots)* para extraer contenido y datos de un sitio web.

Javascript: JavaScript (JS) es un lenguaje de programación multiplataforma orientado a objetos utilizado por los desarrolladores para hacer que las páginas web sean interactivas. Permite a los desarrolladores crear contenido que se actualiza dinámicamente, usar animaciones, menús emergentes, botones en los que se puede hacer clic, controlar multimedia, etc.

Registros de dominio (DNS *registers*): Los registros DNS (también conocidos como archivos de zona) son instrucciones que se encuentran en servidores DNS autorizados y brindan información sobre un dominio, incluida qué dirección IP está asociada con ese dominio y cómo manejar las solicitudes para ese dominio.

Nombre de Dominio (*domain name*): Es una cadena que identifica un ámbito de autonomía, autoridad o control administrativo. Los nombres de dominio se utilizan a menudo para identificar servicios proporcionados a través de Internet, como sitios web, servicios de correo electrónico y más.

Subdominio: Un subdominio es un prefijo agregado a un nombre de dominio para separar una sección de su sitio web.

Centro de operaciones de seguridad (*SOC*): Es el acrónimo en inglés para *Security Operation Center* y es un equipo de profesionales de seguridad informática que monitorea las 24 horas del día durante los 7 días de la semana una infraestructura determinada, para detectar eventos de ciberseguridad en tiempo real y abordarlos de la manera más rápida y efectiva posible.

Contraseña de única vez (*One time password OTP*): Es una contraseña que se expide periódicamente para cada usuario o por eventos, la ventaja de estas contraseñas es que expiran y resuelve el problema de robo de credenciales.

Extraction Transformation and Load (ETL): Acrónimo en inglés para los procesos de extracción, transformación y carga de la información que suelen ser necesario en todos los proyectos de analítica de datos.

Resumen

La detección de sitios web maliciosos se ha convertido en un problema cada vez más relevante, dado que en la actualidad gran parte de la actividad humana se lleva a cabo en línea. Los sitios web maliciosos, también conocidos como sitios web de *phishing*, pueden engañar a los usuarios para que revelen información personal y financiera confidencial, lo que puede llevar a fraudes y delitos financieros.

Este es un desafío importante para la seguridad en línea y afecta significativamente a todo el entorno de operaciones digitales, no solo por las pérdidas económicas que se generan sino por el deterioro de la confianza de los consumidores. Puesto que se trata de un problema de gran escala e involucra ingentes cantidades de datos la analítica de datos se ha posicionado como una herramienta idónea para abordar este flagelo.

En este trabajo, se aborda desde una perspectiva innovadora la identificación de sitios web de phishing, contemplando múltiples variables que consideran diversos aspectos y permite que la solución sea efectiva frente a los diferentes tipos de despliegues empleados por los ciberdelincuentes en 2023.

El objetivo de este trabajo es desarrollar un sistema automatizado basado en un modelo de aprendizaje automático que sea capaz de identificar páginas web de phishing con una precisión superior al 98%, empleando datos de relacionamiento entre el sitio web analizado consigo mismo y su entorno. Para ello se consideran diferentes variables contenidas en las direcciones URL y el código HTML de estos sitios, así como algunas relaciones entre estas variables.

En el desarrollo de este trabajo se implementó un sistema de captura de datos modular encargado de recopilar la información necesaria en forma automática para entrenar y probar el modelo de aprendizaje automático. El sistema se compone de un módulo de *scraping* que se encarga de hacer la descarga del código HTML a

partir de una tabla con direcciones URL y un módulo encargado de hacer la extracción de las características para cada uno de los sitios web listados enriqueciendo la matriz de características. Esta matriz de características es la que servirá para entrenar y probar el modelo de clasificación.

Los resultados de este estudio indican que el enfoque propuesto puede identificar sitios web maliciosos con alta precisión, con una tasa de detección superior al 95% y una tasa de falsos positivos inferior al 2%. Además, el modelo es capaz de identificar nuevas amenazas de phishing con alta precisión, lo que lo convierte en una herramienta valiosa para la detección de amenazas en tiempo real.

Palabras Clave: Phishing, *free-hosting, web-scraping*, aprendizaje de máquina, modelos de clasificación binaria, análisis multivariado.

Resumen gráfico

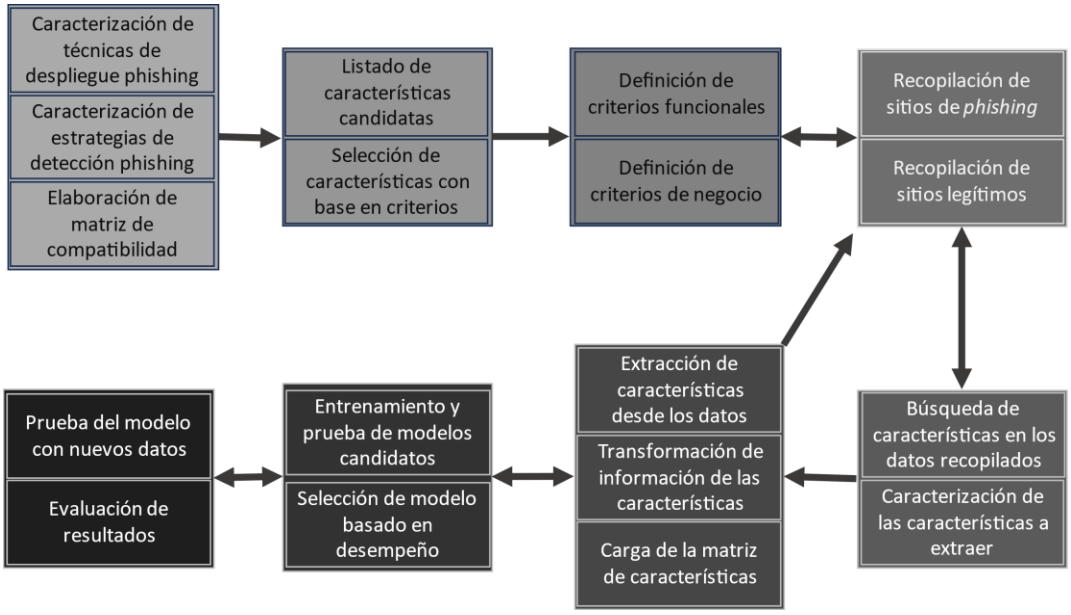
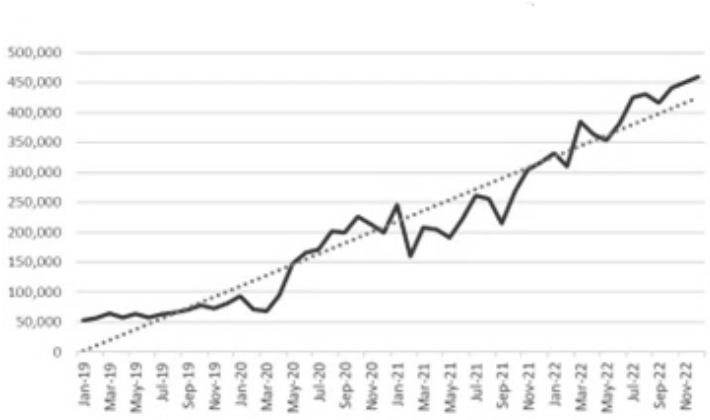


Figura 1 Resumen Gráfico,
Fuente: Elaboración propia

1. Introducción

El *phishing* es la modalidad de delito cibernético dominante y a su vez la de mayor crecimiento en los últimos años (TechDay, 2023). Solamente durante el año 2022 se enviaron más de tres mil millones de mensajes de correo electrónico diarios orientados al phishing en el mundo, lo que en conjunto con otros mensajes de correo electrónico *spam* representan el 48% del total de correos electrónicos enviados durante todo el año (The latest phishing statistics, 2023). La tasa de crecimiento anual de esta modalidad delictiva, según la cantidad de casos reportados se ubica cerca al 150% y la tendencia creciente se ha mantenido consistente en el tiempo desde el año 2019. Para el último trimestre del año 2022 se llegaron a reportar más de un millón cuatrocientos mil casos de phishing, lo que supuso la cifra más alta de casos reportados en la historia (APWG, 2023). En la siguiente figura se puede observar la cantidad de casos reportados de phishing ante la APWG[1] desde el año 2019.



[1] APWG: Acrónimo de Anti-Phishing Working Group es una coalición internacional de personal de respuesta a la ciberdelincuencia, investigadores forenses, organismos encargados de hacer cumplir la ley, empresas de tecnología, empresas de servicios financieros, investigadores universitarios, ONG y organizaciones de tratados multilaterales que operan como una organización sin fines de lucro. https://apwg.org/

Figura 2 Casos reportados de phishing 2019-2022
Fuente: Tomado de Phishing Activity Trends Report (p. 4), por APWG, 2023.

El crecimiento y popularidad del *phishing* entre las diferentes amenazas cibernéticas se explica en gran medida por su buena relación costo-eficiencia dado que, con una inversión relativamente baja, los ciberdelincuentes pueden llegar masivamente a su público objetivo (Frauenstein & Von Solms, 2014), con una alta tasa de éxito pues consiguen una materialización del delito en el 17.8% de los casos de *phishing* general y hasta de 53.2% en los casos de *phishing* dirigido (The latest phishing statistics, 2023).

Las instituciones financieras son las empresas más afectadas por esta modalidad delictiva, siendo el objetivo en el 28% de los casos reportados de phishing durante el año 2022, en la figura 3 se puede observar la afectación por industrias en el año 2022 (APWG, 2023).

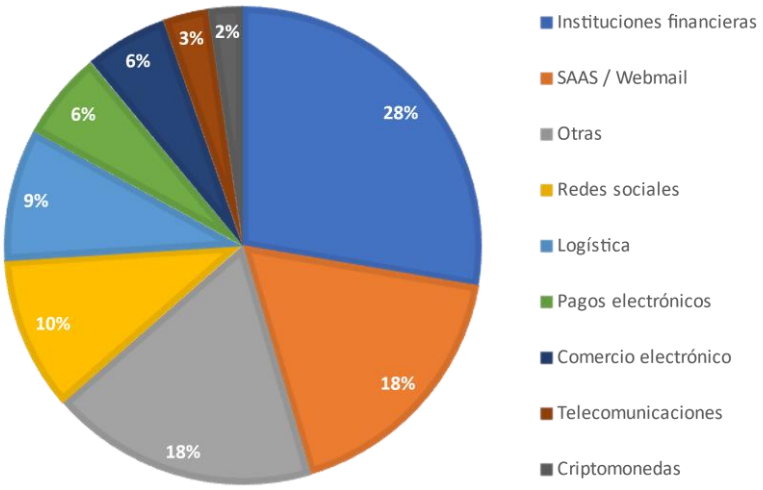


Figura 3 Industrias más afectadas por el phishing,
Fuente: Adaptado de Phishing *Activity Trends Report* (p. 5), por APWG, 2023.

Es importante notar que la principal táctica de los ciberdelincuentes para capturar su objetivo en las campañas de *phishing* es el uso de *links* engañosos que

conducen a sitios web, que simulando pertenecer a entidades de confianza instan al usuario a ingresar su información sensible (Cloudfare Inc, 2023).

El presente trabajo parte de un estudio actualizado de las técnicas vigentes para la puesta en marcha de sitios web de *phishing*, un análisis de diferentes técnicas que se han publicado en los últimos cinco años para la detección e identificación de estos sitios web fraudulentos, para finalmente proponer una alternativa que permita identificar eficientemente sitios web de phishing con bajos costos operacionales y que basándose en algoritmos de aprendizaje de máquina y tareas automatizadas, permita obtener una tasa de falsos positivos menor al 2% en la clasificación de los sitios web objetivo.

Para su realización se emplearon datos actualizados de sitios web de *phishing* confirmados en los repositorios de PhishTank[2] y Openphish[3], así como de sitios web legítimos, provenientes del proyecto CommonCrawl[4]. Todos ellos extraídos entre los meses de febrero y mayo de 2023.

Se espera que este proyecto sirva como base para la elaboración de sistemas más económicos y eficientes en la lucha y prevención contra el *phishing* en diferentes industrias, igualmente se espera que sirva como referencia para posteriores investigaciones que permitan prevenir la puesta en marcha y materialización de campañas de *phishing*.

[2] Phishtank: Sitio comunitario gratuito donde cualquiera puede enviar, verificar, rastrear y compartir datos de phishing. https://phishtank.org/

[3] OpenPhish: Plataforma autónoma totalmente automatizada para inteligencia de phishing. https://openphish.com/

[4] CommonCrawl: Repositorio abierto y gratuito de datos de rastreo web que cualquier persona puede utilizar https://commoncrawl.org/

2. Pregunta de Investigación

¿Cómo se puede desarrollar un sistema automático y costo-eficiente basado en un modelo de aprendizaje automático que permita la detección de sitios web de phishing contemplando los diferentes tipos de despliegue que se utilizan en 2023?

3. Marco Conceptual

Aun cuando el concepto y la finalidad del phishing se mantienen en términos generales desde su primera mención (Jerry Felix, 1987). El phishing ha evolucionado y adoptado nuevas tecnologías para mantener su vigencia a lo largo del tiempo (Cui et al., 2018), esto hace necesario que cada nuevo estudio que se realice para identificar, rastrear o mitigar esta modalidad delictiva, deba iniciar por un análisis actualizado de los mecanismos empleados por los ciberdelincuentes para la puesta en marcha de las campañas de *phishing*.

De igual forma es imperativo tomar como punto de partida trabajos previos en la identificación y detección de *phishing*, identificando las premisas que se mantienen vigentes y las estrategias que han demostrado éxito dentro del marco que han sido desarrolladas(Do et al., 2022).

3.1. El *phishing*

Según el departamento de seguridad nacional de los Estados Unidos el *phishing* es un tipo de ataque informático que tiene como objetivo engañar a las personas para que revelen información personal o financiera como números de teléfono, direcciones, contraseñas, números de tarjeta de crédito o información de cuentas bancarias. Para conseguirlo los atacantes a menudo se hacen pasar por empresas legítimas, organizaciones benéficas u otras entidades de confianza y utilizan correos electrónicos, mensajes de texto, llamadas telefónicas y sitios web falsos para hacer que las víctimas divulguen información confidencial. (DHS, 2017)

Esta información robada puede ser empleada para la suplantación frente a establecimientos de comercio y entidades financieras, así como fraude electrónico, entre otras actividades delictivas configurando una estrecha relación entre el

phishing, el robo de identidad y los delitos de fraude electrónico (Avast Software s.r.o., 2020)**.**

3.2. Modalidades de despliegue de sitios web de phishing

Dependiendo de la forma en que son desplegados, los sitios web de phishing pueden ser clasificados en los siguientes grupos:

3.2.1. Dominios web comprometidos. Según la ICANN[5] en esta modalidad de despliegue, los ciberdelincuentes aprovechan vulnerabilidades en los sitios de hosting, para desplegar páginas de phishing como subdominios o recursos en dominios legítimos y bien reputados, lo que ayuda al encubrimiento de la acción delictiva. (ICANN, 2022).

3.2.2. Registro de dominios maliciosos. En esta modalidad de despliegue los ciberdelincuentes registran un dominio nuevo, que servirá para direccionar los sitios web de *phishing*. Estos dominios pueden guardar semejanzas con los dominios que se pretenden suplantar para mejorar las tasas de éxito entre los afectados. (Paloalto networks, 2023).

3.2.3. Abuso de free-hosting. Existen múltiples servicios en internet para hospedar sitios web sin necesidad de adquirir un dominio propio. Dichos sitios web serán accesibles a través de un subdominio del dominio principal de la empresa de hospedaje. Estos hospedajes compartidos ofrecen múltiples ventajas a sus usuarios desde el bajo o nulo costo, hasta facilidades para gestionar los sitios web de forma automática. Esto permite crear contenidos tanto legales como ilegales de forma masiva a un bajo costo operativo, lo que ha llevado a posicionar este despliegue como uno de los más empleados para publicar páginas de phishing (Roy et al., 2022).

En la figura 4 se observa la distribución de los sitios de phishing desde el punto de vista de su despliegue para el año 2021 (Ellis, 2021)

---

[5] ICANN: Es la organización encargada de garantizar el funcionamiento estable y seguro de los sistemas de identificadores únicos de Internet. https://www.icann.org/

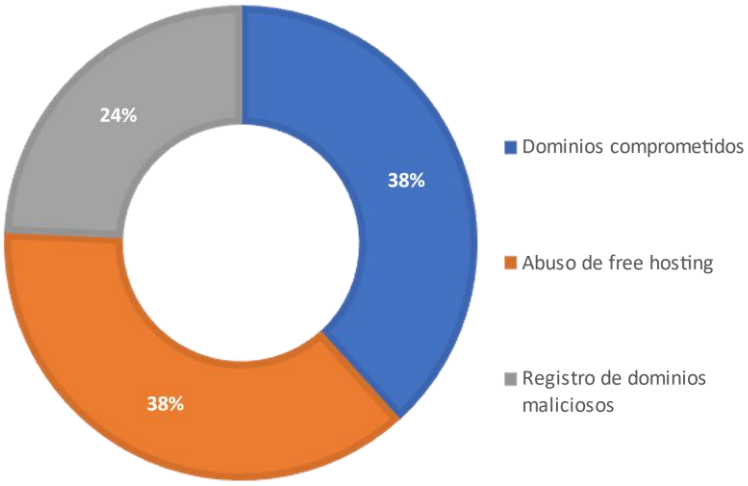


Figura 4 Despliegue de sitios web de phishing por tecnología
Fuente: Adaptado de Most Phishing Attacks Use Compromised Domains and Free Hosting por Jessica Ellis, 2021.

3.3. El proceso de detección de phishing

Tanto empresas enfocadas en la ciberseguridad, como entidades financieras y empresas de consumo implementan procesos para mitigar los efectos negativos del *phishing*, estos procesos suelen comprender tres fases bien demarcadas. Cada una de estas tres fases se enfoca en resolver un objetivo concreto y dependiendo de cada una de las entidades donde se desarrolle puede involucrar una cantidad mayor o menor de tareas, así como de un mayor o menor grado de automatización en ellas (Frauenstein & Von Solms, 2014).

En la figura 5 se ilustran las tres fases del proceso general de detección y mitigación de los efectos del *phishing*.

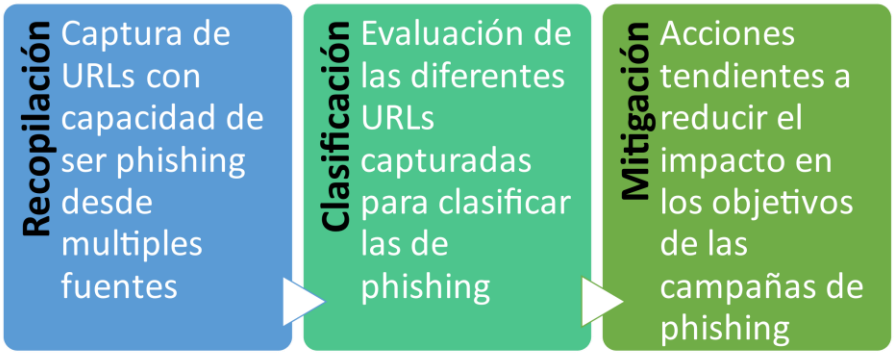


Figura 5 Fases en la detección y mitigación de campañas de Phishing
Fuente: Elaboración propia

Es importante resaltar que estas fases pueden ser automatizadas o manuales, caso en el cuál los tiempos necesarios para llegar a una clasificación correcta de un incidente pueden tomarse 24 horas desde el momento que se recibió el mensaje de phishing por parte de cualquier usuario. (Avanan, 2023). Además del tiempo es igualmente importante considerar el factor económico de los procesos de detección de phishing. Un reciente estudio realizado por una compañía de ciberseguridad muestra que una aproximación manual puede implicarle a una empresa de 10.000 empleados mantener un grupo de 90 empleados en el SOC para la gestión de incidentes con un costo de $8.1 millones de dólares. Empleando aproximaciones automatizadas organizaciones de este tamaño pueden ahorrar hasta $7.85 millones de dólares al tiempo que se estima que el riesgo por amenazas se reduce en un 51% (Fortra, 2019).

3.3.1. Recopilación. El objetivo fundamental en la etapa de recopilación es capturar la mayor cantidad de información que pueda servir para identificar posibles vectores de phishing, principalmente se busca obtener una lista de URLs con potencial de apuntar a sitios web de *phishing*.

Entre mayor sea la superficie de búsqueda mejor serán los resultados del proceso, por esa razón esta fase involucra diversas entradas como: listas de nuevos dominios registrados ante las entidades registradoras, buzones de correo destinados para el reporte (*abuse box*), logs de los sitios web de clientes, arañas o indexadores (*crawlers*), búsquedas en *deep* y *dark web*, analizadores de publicidad,

entre otras. El resultado es una extensa lista de URLs que tienen potencial de ser empleadas en campañas de *phishing*.

3.3.2. Clasificación. El análisis de las diferentes URLs recopiladas en la primera etapa, tiene como objetivo lograr su clasificación entre sitios web de *phishing* y sitios web legítimos. Nuevamente los pasos varían de una empresa a otra, pero en general las técnicas de clasificación pueden dividirse en técnicas basadas en contenido y técnicas no basadas en contenido (Chanti & Chithralekha, 2020). Estas técnicas se pueden usar en manera conjunta en flujos de trabajo donde la salida de una de ellas es la entrada de la siguiente en el proceso, por ejemplo, las listas blancas permiten descartar sitios web que evidentemente no se deben considerar como *phishing* para reducir la lista que se enviará a la siguiente etapa, mejorando la eficiencia general del proceso. (Azeez et al., 2021)

En la figura 6 se presenta un recuento de las diferentes técnicas de clasificación de sitios web de phishing, haciendo la distinción entre técnicas basadas en contenido y técnicas no basadas en contenido.

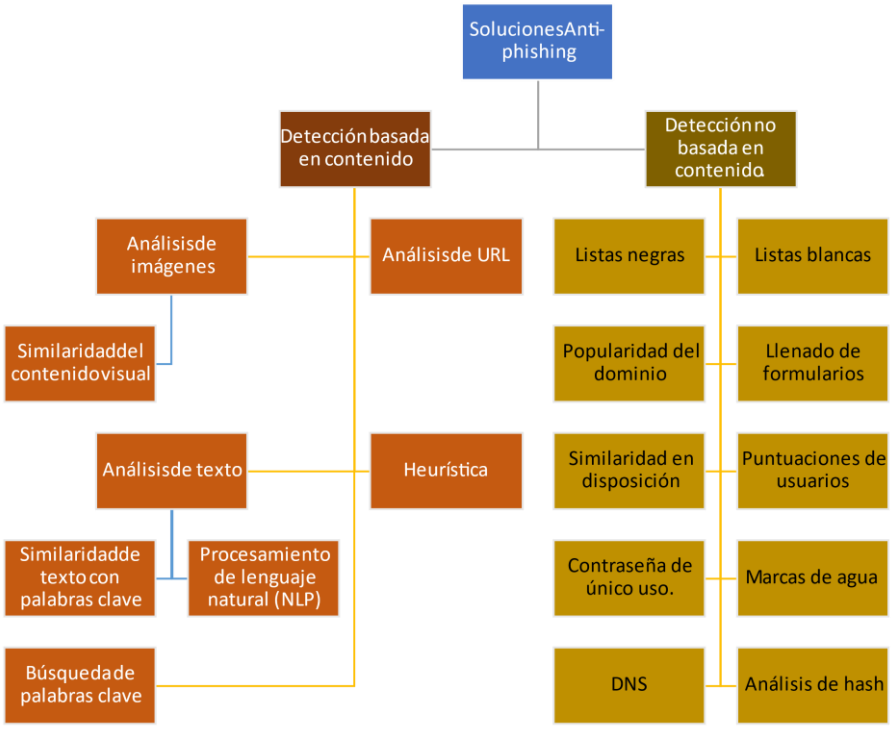


Figura 6 Técnicas de clasificación de sitios de phishing,
Fuente: Adaptado de (Chanti & Chithralekha, 2020)

3.3.2.1. Técnicas basadas en contenido. Las técnicas basadas en contenido del sitio web a analizar buscan patrones que se repiten en los sitios web de phishing como: gramática, ortografía, campos de contraseñas, vínculos, imágenes URLs, categorías de páginas, información de Whois, patrones en el código HTML y contenido dinámico como JavaScript. (Chanti & Chithralekha, 2020)

3.3.2.1.1. Análisis de imágenes. En las páginas de phishing se suelen emplear piezas características de una marca como su imagen corporativa y logos para aumentar la percepción de confianza de los usuarios objetivo, por esta

razón una búsqueda de uso sin autorización de estas piezas gráficas permite dar con la ubicación de páginas de *phishing*.(Jain & Gupta, 2017)

3.3.2.1.2. Análisis de texto. Las aproximaciones basadas en análisis de texto buscan encontrar patrones repetibles en los sitios de phishing, bien sea con la aplicación de heurística o con algoritmos de distancia de similitud comparando palabras clave previamente definidas.(Sumathi & Sujatha, 2019)

3.3.2.1.3. Análisis de URL. Tradicionalmente los análisis de URL parten de la premisa que los sitios web de phishing exhiben patrones identificables en sus direcciones URL como similitud con dominios legítimos, uso de caracteres especiales y uso de diferentes codificaciones. Normalmente se aplica heurística o algoritmos de distancia de similitud junto con la búsqueda de palabras clave. (Uddin et al., 2022)

3.3.2.2. Técnicas no basadas en contenido.

Las técnicas no basadas en contenido buscan identificar sitios web de phishing gracias a información externa al propio sitio web.(Chanti & Chithralekha, 2020)

3.3.2.2.1. Listas negras. Las listas negras son repositorios donde se almacenan direcciones IP, dominios o direcciones HTML que han sido reportadas y confirmadas como sitios de phishing.(Azeez et al., 2021)

3.3.2.2.2. Listas blancas. El objetivo de las listas blancas es reducir la cantidad de análisis de los sistemas de identificación de phishing, en dichas listas se almacena las direcciones HTML que pertenecen a las compañías que están siendo monitoreadas en búsqueda de phishing.

El mantenimiento de listas tanto negras como blancas es efectivo en términos de ahorro de reprocesos, pero es un proceso demandante y administrativamente costoso. (Azeez et al., 2021)

3.3.2.2.3. Popularidad del dominio. Esta técnica parte de la premisa que los dominios que tienen una mayor popularidad en términos de

búsqueda y listado en los buscadores tienden a presentar un menor riesgo.(Jeeva & Rajsingh, 2016)

3.3.2.2.4. Puntuaciones de usuarios De forma análoga a la popularidad del dominio, para esta técnica se emplean sitios de rankings donde los usuarios califican sitios web dando la opción de calificar en forma negativa aquellos que tienen contenido fraudulento.(Liu & Fu, 2020)

3.3.2.2.5. Llenado de formularios. Esta técnica si bien requiere información del código HTML del sitio web analizado no precisa mayores análisis sobre el contenido en sí mismo, se enfoca en la presencia de formularios para diligenciar, pues los sitios de phishing usualmente exhiben formularios para recopilar información de los usuarios afectados.(Tang & Mahmoud, 2021)

3.3.2.2.6. Similitud en disposición. Análoga a la anterior en esta técnica a partir de la disposición del contenido en los sitios web se puede encontrar patrones repetidos en algunos sitios de phishing que tienen un origen común.(Zhang et al., 2013)

3.3.2.2.7. Marcas de agua. Las empresas introducen marcas en sus páginas legítimas y establecen con sus usuarios en el momento de dada de alta en el sitio la forma en que se preguntarán estos elementos de manera que el usuario podrá considerar fraudulento un sitio sin estas marcas o cuándo no siguen el protocolo establecido.(Singh et al., 2011)

3.3.2.2.8. Contraseña de único uso. Algunos sitios web transaccionales piden a sus usuarios introducir un OTP antes de introducir cualquier información, de esta forma se educa a los usuarios que cualquier sitio que solicite su información y no solicite ese OTP podrá ser tomado como fraude.(Ulqinaku et al., 2019)

3.3.2.2.9. Registros *DNS.* En esta técnica se busca conocer la antigüedad de los registros DNS para un dominio analizado, a menor antigüedad existe un mayor riesgo de estar frente a un posible caso de phishing pues da cuenta de un sitio web de reciente publicación. (Alsabah et al., 2022)

3.3.3. Mitigación.

Una vez se han clasificado los sitios web de phishing el siguiente paso es mitigar los efectos de la campaña, esto se logra pidiendo dar de baja al sitio web malicioso ante la empresa de *hosting* y/o la entidad registradora del dominio, reportando ante entidades encargadas como Openphish o PhishTank, alertando a la empresa afectada sobre la campaña en curso para prevenir a los usuarios finales, entre otras actividades.

3.4. Estado del arte

En los últimos cinco años se han realizado múltiples investigaciones que, empleando mecanismos automatizados, principalmente aprendizaje de máquina hacen más eficiente y menos costosa la detección de phishing al reducir las tareas manuales que deben desempeñar los agentes de un *SOC* para clasificar los incidentes (Tang & Mahmoud, 2021).

Para efectos de entendimiento del estado del arte se realiza una clasificación basada en las características evaluadas y se toman 22 investigaciones en esta área de los últimos cinco años. La tabla de clasificación que se presenta a continuación toma como referencia las técnicas presentadas en la figura 6 y resalta para cada una de las investigaciones la técnica o técnica en la que se basaron los autores para hacer sus respectivos desarrollos.

Tabla 1. Clasificación de estudios relacionados con la identificación de *phishing* en los últimos cinco años.

Título del trabajo	Datos en las direcciones URL	Datos de terceros como registro de dominios	Datos semánticos en el contenido HTML	Componentes gráficos en el contenido HTML	Características de los hipervínculos
Phishing URL detection using machine learning methods(Ahammad et al., 2022)					

Título del trabajo	Datos en las direcciones URL	Datos de terceros como registro de dominios	Datos semánticos en el contenido HTML	Componentes gráficos en el contenido HTML	Características de los hipervínculos
Hybrid phishing detection using joint visual and textual identity (Tan et al., 2023)					
PhiDMA – A phishing detection model with multi-filter approach **(Sonowal & Kuppusamy, 2020)**					
HinPhish: An Effective Phishing Detection Approach Based on Heterogeneous Information Networks (Guo et al., 2021)					
Intelligent phishing url detection using association rule mining **(Jeeva & Rajsingh, 2016)**					
Identification of phishing websites through hyperlink analysis and rule extraction (Wang et al., 2020)					
An effective detection approach for phishing websites using URL and HTML features (Aljofey et al., 2022)					
A Comparative Analysis of Machine learning-Based Website Phishing Detection Using URL Information (Uddin et al., 2022)					
URL Phishing Detection using Machine learning Techniques based on URLs Lexical Analysis (Abutaha et al., 2021)					
Detection of phishing websites using an efficient feature-based machine learning framework **(Rao & Pais, 2019)**					
Cyber Threat Intelligence-Based Malicious URL Detection Model Using Ensemble Learning (Ghaleb et al., 2022)					

Título del trabajo	Datos en las direcciones URL	Datos de terceros como registro de dominios	Datos semánticos en el contenido HTML	Componentes gráficos en el contenido HTML	Características de los hipervínculos
Detection of Phishing Websites by Using Machine learning-Based URL Analysis (Korkmaz et al., 2020)	■				
Intelligent Deep Machine learning Cyber Phishing URL Detection Based on BERT Features Extraction (Elsadig et al., 2022)	■				
Detection of Phishing Websites by Investigating Their URLs using LSTM Algorithm **(Alanzi & Uliyan, 2022)**	■	■			
A Deep Learning-Based Framework for Phishing Website Detection **(Tang & Mahmoud, 2022)**	■		■	■	
Malicious URL detection based on machine learning (Xuan et al., 2020)	■				
Modeling Hybrid Feature-Based Phishing Websites Detection Using Machine learning Techniques (Das Guptta et al., 2022)	■				■
Hybrid phishing detection using joint visual and textual identity (Tan et al., 2023)			■	■	
A machine learning based approach for phishing detection using hyperlinks information **(Jain & Gupta, 2019)**					■
Effective Phishing Detection using Machine learning Approach (Yang, 2019)	■	■			
SPWalk: Similar Property Oriented Feature Learning for Phishing Detection **(Liu & Fu, 2020)**					■
Detecting Phishing URLs using Machine learning Lexical Feature-based Analysis (Alshira'H, 2020)	■				

Como se puede observar en el cuadro comparativo solo 6 de los 22 trabajos revisados emplean la revisión de las características de los hipervínculos en contraste con los 16 trabajos que optan por una aproximación basada en el análisis de las direcciones URL que la convierten el método en la más empleado para detección de páginas de phishing. Un hallazgo importante es que solo dos de los trabajos tomaron como aproximación la caracterización de los sitios web como parte de una red, ambos trabajos expusieron muy buenos resultados y requerimientos bajos en cuanto a descarga de características que requerían para la realización de sus modelos de aprendizaje de máquina, lo que resulta conveniente para la selección de características del presente trabajo, igualmente se evidenció que ninguno de los trabajos tomaba como base una caracterización de todos los elementos propios del estándar para hipervínculos y que podrían aportar pistas en sitios web maliciosos.

4. Objetivos

4.1. Objetivo General

Desarrollar un sistema de identificación automático basado en un modelo de aprendizaje de máquina que permita la identificación precisa de sitios web de phishing.

4.2. Objetivos Específicos

Desarrollar una revisión bibliográfica actualizada de los mecanismos de despliegue de páginas de phishing, identificando las características principales de cada uno de estos mecanismos y el impacto de cada una de las modalidades.

Realizar una revisión literaria en el campo de la detección de *phishing* de los diferentes enfoques que se han propuesto en los últimos cinco años para identificar la pertinencia de cada uno frente a los mecanismos actuales de despliegue de sitios web de phishing.

Recopilar y preparar un conjunto de datos representativo y actualizado que contenga ejemplos de sitios web legítimos y sitios web de phishing, así como la información adicional necesaria para extraer las características relevantes para el análisis.

Diseñar y desarrollar un modelo de aprendizaje de máquina específico para la detección de sitios web de *phishing*, explorando técnicas de extracción de características eficientes y algoritmos de clasificación.

Evaluar el rendimiento del modelo propuesto utilizando métricas de calidad como precisión, tasa de falsos positivos y tasa de falsos negativos, frente al desempeño esperado.

5. Metodología

Este trabajo adopta como marco metodológico la Metodología Fundacional para Ciencia de Datos (Jhon B. Rollins, 2015) publicada por IBM. La cuál guarda similitudes con las más conocidas KDD (Feyyad, 1996) y CRISP-DM (Chapman, et al., 1999),(Schröer et al., 2021), pero agrega unas prácticas que resultan bastante convenientes al desarrollar un problema como el que se desarrolla en este trabajo (Foroughi & Luksch, 2018). En la siguiente figura se pueden ver todas las etapas propuestas en la mencionada metodología.

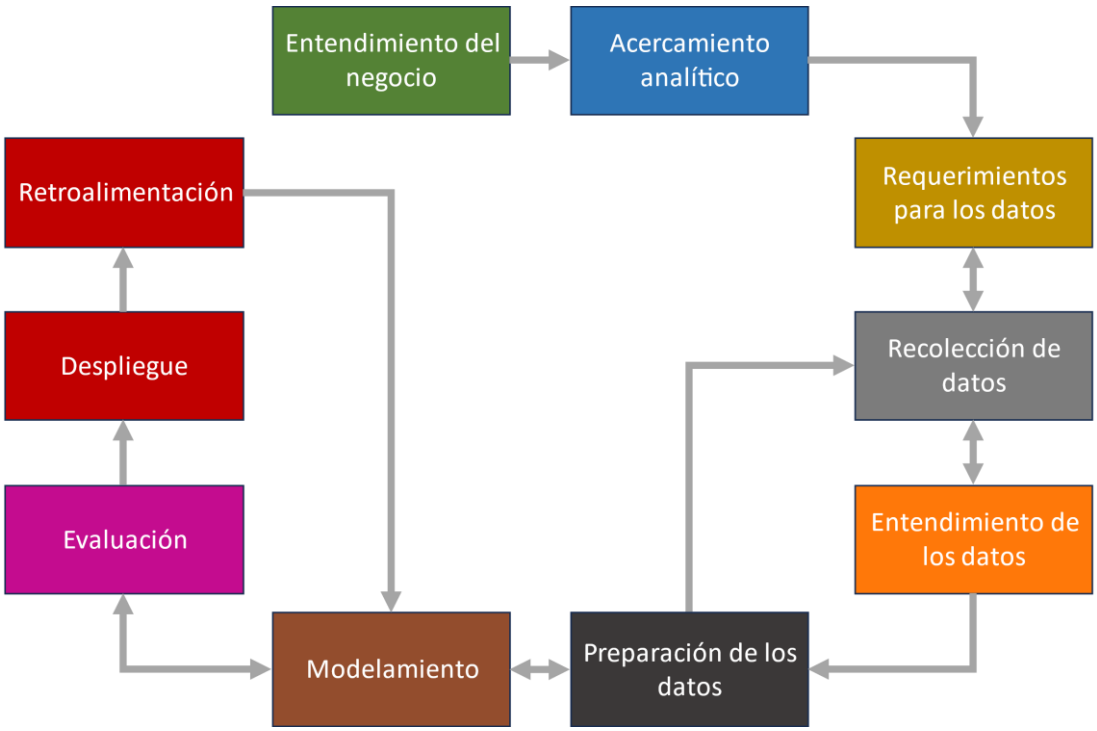


Figura 7 Metodología Fundacional,
Fuente: adaptado de (Jhon B. Rollins, 2015)

La relación y orden de las 10 etapas de la metodología (Entendimiento del negocio, acercamiento analítico, requerimientos de datos, recopilación de datos, entendimiento de datos, preparación de datos, modelamiento, evaluación, despliegue y retroalimentación) se puede observar en la figura 7.

Las etapas de despliegue y retroalimentación se dejan para una investigación futura o para una eventual puesta en operación a nivel comercial de la solución planteada en este trabajo.

A continuación, se presenta el desarrollo realizado en cada una de las ocho etapas para este trabajo:

5.1. Entendimiento del negocio

Para el entendimiento del problema de negocio a resolver, se sostuvieron reuniones explicativas por parte de los líderes del SOC de una reputada empresa de prevención de fraude y se analizaron tanto los datos recopilados, como las acciones tomadas por los agentes frente a cada nueva alerta, en el periodo de tiempo de octubre de 2022 a enero de 2023.

Con el fin de caracterizar las diferentes formas de despliegue de sitios de *phishing* se revisó bibliografía publicada en los tres últimos años por organizaciones rectoras de internet como ICANN y centros de ciberseguridad de amplia trayectoria como Fortra’s Phishlabs, dando como resultado la caracterización presentada en el capítulo 3.2 donde se exponen los datos recopilados, así como las restricciones y retos que pueden presentarse al momento de desarrollar una solución orientada a la identificación de este tipo de sitios maliciosos.

Finalmente se realizó la caracterización inicial de 20 casos confirmados de phishing que fueron capturados en el periodo de tiempo antes mencionado.

Es importante mencionar las siguientes restricciones de negocio que buscan alinear la solución propuesta a un escenario real de implementación costo-eficiente.

1. El sistema debe emplear un número reducido de características del sitio web a analizar y que dichas características no precisen de complejos procesos de extracción y elevado costo de almacenamiento.

2. Se optará por el algoritmo de aprendizaje de máquina más sencillo en términos computacionales con el fin de reducir el consumo de tiempo y recursos para analizar cada página.
3. Debe ser compatible con todas las modalidades de despliegue de phishing identificadas en el análisis del negocio.

5.2. Acercamiento analítico

A partir de los objetivos y restricciones identificados en la primera etapa y con el estudio de los datos preliminares, se determinó que el problema a resolver desde el punto de vista de aprendizaje de máquina era de tipo clasificación binaria (Harrington, 2012), donde se espera que el sistema sea capaz de clasificar una URL de entrada como *phishing* o segura. De esta forma el desarrollo del trabajo se orientará hacia algoritmos de aprendizaje de máquina supervisado. Para esta decisión se tuvo en cuenta, la disponibilidad de datos marcados, la posibilidad de determinar las características relevantes para la clasificación y la eficiencia computacional que exhiben estos algoritmos frente a otras alternativas como *Deep Learning* (Harrington, 2012).

Para determinar el conjunto de características a emplear, se partió de la tabla 1 que presenta la clasificación de trabajos recientes y una comparación de las premisas para la detección de phishing expuesta en cada uno de los trabajos citados con la información recopilada sobre las modalidades de despliegue de sitios web de phishing presentada en el capítulo 3.2.

A continuación, se presenta un resumen de la información obtenida de los diferentes trabajos citados frente a la efectividad de las técnicas y las limitaciones de cada una de ellas.

5.2.1. Detección no basada en contenido.

Los trabajos que se enfocaron en técnicas de detección no basadas en contenido exhibieron buenos resultados frente a sitios de phishing que presentaban características conocidas previamente y además mostraron que al ser complementados con técnicas basadas en contenido mejoraban significativamente la eficiencia del sistema.

5.2.2. URL o dirección web.

Los trabajos que tomaron únicamente características contenidas en las direcciones URL de los sitios sospechosos, presentaron un desempeño muy bueno tanto en tiempos de análisis como en precisión en los casos de registro de dominios maliciosos, no así en los casos de dominios web comprometidos o abuso de free-hosting.

5.2.3. Contenido HTML.

Varios trabajos tomaron diferentes características del código HTML y se pudo observar que las aproximaciones de solo texto eran más eficientes en cuanto a procesamiento que aquellas que involucraron imágenes, aunque la precisión aumentaba en este último escenario, obteniendo resultados superiores al 90%.

5.2.4. Hipervínculos.

Finalmente, un enfoque basado en hipervínculos mostró alta eficiencia en la tarea de detectar phishing sin requerir parametrizaciones previas enfocadas a proteger a una empresa en particular. El concepto desarrollado es comprender los sitios web como entidades que tienen relacionamiento entre sí y no como elementos aislados. Este enfoque es igualmente eficiente con los tipos de despliegue revisados en este trabajo y tiene como ventaja adicional que no precisa técnicas muy costosas computacionalmente como son las redes neuronales.

Adicional a esta revisión bibliográfica, se hizo un análisis de las muestras recolectadas en la primera fase del proyecto para encontrar patrones en los hipervínculos que pudieran resultar de utilidad, de esta forma se constató que un número significativo de hipervínculos en los sitios de *phishing* recolectados, exhibían algún tipo de malformación que sacaba partido de lo descrito en el estándar de la W3C[6] relacionado con HTML, para intentar oscurecer el destino real al que dirigían.

[6] W3C: El World Wide Web Consortium (W3C) desarrolla estándares y directrices para ayudar a todos a construir una web basada en los principios de accesibilidad, internacionalización, privacidad y seguridad. https://www.w3.org/

De esta forma se eligió para este trabajo características que permitieran ver el relacionamiento del sitio web con su entorno y además identificar patrones sospechosos de ocultamiento en los hipervínculos contenidos en el código HTML.

5.3. Requerimientos de datos

Dado que se eligió un modelo de aprendizaje de máquina supervisado es preciso contar con un conjunto de datos marcados y además para lograr un óptimo entrenamiento resultará benéfico contar con un conjunto de datos balanceado evitando sobre entrenamiento para cualquiera de las dos categorías. Siguiendo el enfoque elegido se determina que el set de datos debe contar con al menos los siguientes elementos: URL y código HTML del sitio a analizar, lo cual respeta las restricciones de negocio en cuanto a requerir la menor cantidad de fuentes para evitar incrementos innecesarios en el costo de la solución.

5.4. Recolección de datos

Al revisar los datos existentes en diferentes repositorios públicos se evidenció que no cumplían los requerimientos previamente expuestos, por lo que se eligió hacer un levantamiento de información por un periodo de tres meses, con el fin de conseguir un conjunto de datos significativo y con la calidad requerida para el proyecto.

Para la recolección de información se eligieron los repositorios abiertos Phishtank y OpenPhish los cuales se pueden consultar en tiempo real y presentan una gran cantidad de sitios web sospechosos y confirmados de phishing. Las principales dificultades al adquirir la información de este tipo de sitios fueron:

Restricciones para el *scraping*: en muchas ocasiones los ciberdelincuentes ponen restricciones en sus sitios de phishing para limitar la descarga del código HTML desde determinadas locaciones o determinados dispositivos.

Restricciones temporales: puesto que se eligió descargar los sitios de *phishing* a partir de la lista de sitios reportados en las dos mayores listas independientes para reporte de phishing, se contaba con una ventana de tiempo reducida pues dichos sitios ya se encontraban en un proceso de mitigación activo y al término de unas horas muchos ya habían sido desactivados, imponiendo así una ventana útil de tiempo para la descarga desde el momento de la aparición del reporte en las listas.

5.4.1. Proceso de recolección de los datos.

5.4.1.1. Identificación de URLs

La primera parte de la recolección de datos consistió en la identificación de URLs válidas tanto para sitios legítimos como para sitios de *phishing*.

En el caso de sitios web legítimos se extrajo una muestra aleatoria del proyecto CommonCrawl, tomando las primeras 100 entradas del archivo de índices más reciente disponible cada semana durante el periodo de tiempo de recolección de muestras.

Para el caso de los sitios web de phishing se tomaron a diario las primeras 10 entradas de casos confirmados de phishing en las listas públicas de Phishtank y OpenPhish, repitiendo la operación varias veces incluso en un mismo día en diferentes horas.

5.4.1.2. Descarga del contenido HTML de los sitios web identificados.

Para descargar el contenido HTML asociado a las URLs obtenidas en el paso anterior, se implementó un componente en Python que a partir de un archivo de tipo separado por comas (CSV) generaba un *dataframe* en el que además de la dirección URL recibida, se escribía el código de respuesta obtenido al intentar abrir el sitio web y el código HTML en los casos que no se recibía algún tipo de excepción durante la descarga.

5.4.1.3. Filtrado de los datos obtenidos y consolidación parcial de las muestras válidas.

El paso final en la recolección de datos consistió en filtrar datos inválidos, descartando todas las muestras con un código de respuesta diferente a 200 y todas

las entradas que no tuvieran código HTML El dataframe resultante se exportaba a un archivo separado por comas para las muestras de *phishing* y otro para las muestras legítimas, que era actualizado con cada nuevo grupo de URLs identificadas para cada categoría.

5.4.1.4. Consolidación final de resultados y balanceo de la muestra.

Al final del periodo de tiempo de recolección de información se consolidó un set de datos con 2000 muestras de datos válidas para la categoría phishing y de 2800 muestras para la categoría de sitios legítimos. A este set de datos es aplicó un filtro adicional limpiando todas las entradas con código HTML inferior a 100 caracteres, con el objetivo de limpiar todas aquellas muestras donde se hubiera podido recibir un HTML malformado o con error. El resultado final fue un set de datos de 1676 muestras de phishing y 2270 muestras de sitios legítimos, número que fue reducido a 1676 aplicando un *splitting* en Python para obtener una muestra balanceada de cara a los pasos siguientes del proyecto.

5.5. Entendimiento de los datos

A partir del estudio de las características que se han empleado en trabajos previos y junto con el análisis de propiedades definidas en el estándar de la W3C (consultar apéndice 1) de los hipervínculos de los sitios web obtenidos en la primera fase del proyecto se eligieron las siguientes características:

Cantidad de links: Esta característica muestra que tanto relacionamiento tiene un sitio web con otros y consigo mismo. Entre mayor sea este número mayor referenciación puede tener el sitio web, lo que puede no ser deseable para sitios de *phishing* pues entre menos opciones tengan de ser ubicados más tiempo pueden pasar inadvertidamente.

Links con *title*: Esta característica toma la cuenta de cuántos hipervínculos están empleando la propiedad *title*. Según el análisis de la muestra recolectada en los sitios legítimos se tiende a hacer más uso de las propiedades del estándar frente a uso que hacen los sitios fraudulentos.

Links que no dirigen a ningún lugar: Una de las posibilidades de un link es no dirigir a ningún lugar, para ello se emplea el valor *blank* en el destino. En el análisis

de las muestras se evidenció que los sitios de phishing tendían a usar mucho más este tipo de opciones.

Links relativos: La cantidad de enlaces a otros sitios que se encuentran alojados en el mismo hosting se presentó significativamente más para los sitios legítimos de la muestra analizada.

Links absolutos: La cantidad de enlaces a sitios externos resultó también un buen indicador para los sitios de phishing porque tienden a emplear más recursos o referenciar a más contenidos fuera del hosting propio.

Links a correo: Los links a correo resultaron menos frecuentes en los sitios de phishing analizados, probablemente se quiere evitar presentar contactos que podrían alertar de la falsedad del sitio de *phishing*.

Links con el mismo dominio del sitio: Esta característica permite saber cuantas relaciones tiene el sitio analizado con páginas del mismo dominio, es normal que sitios legítimos tengan una alta tasa de estas relaciones, dado que los sitios web suelen ser lugares para compartir importantes cantidades de información a los usuarios.

Links a una sección del mismo documento: Los *links* que dirigen a una sección diferente dentro de una misma página web suelen denotar organización y amplio contenido, lo que no es usual en los sitios web de phishing que suelen ser bastante más simples y enfocados en la captura de información como se pudo evidenciar con la muestra analizada.

Tabla 2. Relación del nombre, descripción y el número de la característica empleado en la presentación de resultados.

Número de la característica	**Etiqueta empleada en la codificación**	**Descripción**
1	links_number	Cantidad de links
2	links_with_text	Links con *title*
3	links_target_blank	Links que no dirigen a ningún lugar
4	relative_links	Links relativos

5	absolute_links	Links absolutos
6	mailto_links	Links a correo
7	links_with_domain	Links con el mismo dominio del sitio
8	links_to_section	Links a una sección del mismo documento

5.6. Preparación de los datos:

En la etapa de preparación de datos se aplicó un proceso de ETL para cargar una tabla con las características elegidas para cada uno de los sitios web listados en el conjunto de phishing y en el conjunto de sitios legítimos. Esta tabla de características es la que se requiere para entrenar y probar el modelo de clasificación. A continuación, se presentan los pasos desarrollados para conseguir la mencionada tabla.

5.6.1. Extracción

El primer paso consistió en extraer desde los archivos tipo CSV que se generaron en la etapa de descarga de información tanto la URL y el HTML de cada uno de los sitios web agregando una característica para cada pareja que indicara si el sitio web se trataba de *phishing* o no.

A continuación, se creó un módulo encargado de extraer las características de cada una de las parejas URL y HTML con apoyo de la librería *beatifulsoap* que facilitó la tarea de interpretación del código HTML, extrayendo las URL contenidas en él y posteriormente fue útil para hacer los conteos de las características de cada una de esas URL.

Con este paso se construyó un set de datos balanceado y marcado con contenido de sitios de phishing y sitios web legítimos por igual.

5.6.2. Transformación

Al set de datos previamente construido se le realizó un proceso de estandarización, para lo cual se empleó una funcionalidad de la librería *sklearn standard scaler* que ayuda al cumplimiento de requisitos de algunos clasificadores

como SVM que asume que todas las características están centradas en 0 y tienen una varianza semejante entre ellas, de esta forma se evita que por la diferencia de ordenes de magnitud, que puede existir entre características, se pierda significancia de algunas de ellas.

A continuación, se realizó y analizó la matriz de correlación de las variables estandarizadas encontrando que existía una fuerte correlación entre algunas de ellas como se puede observar en la siguiente figura.

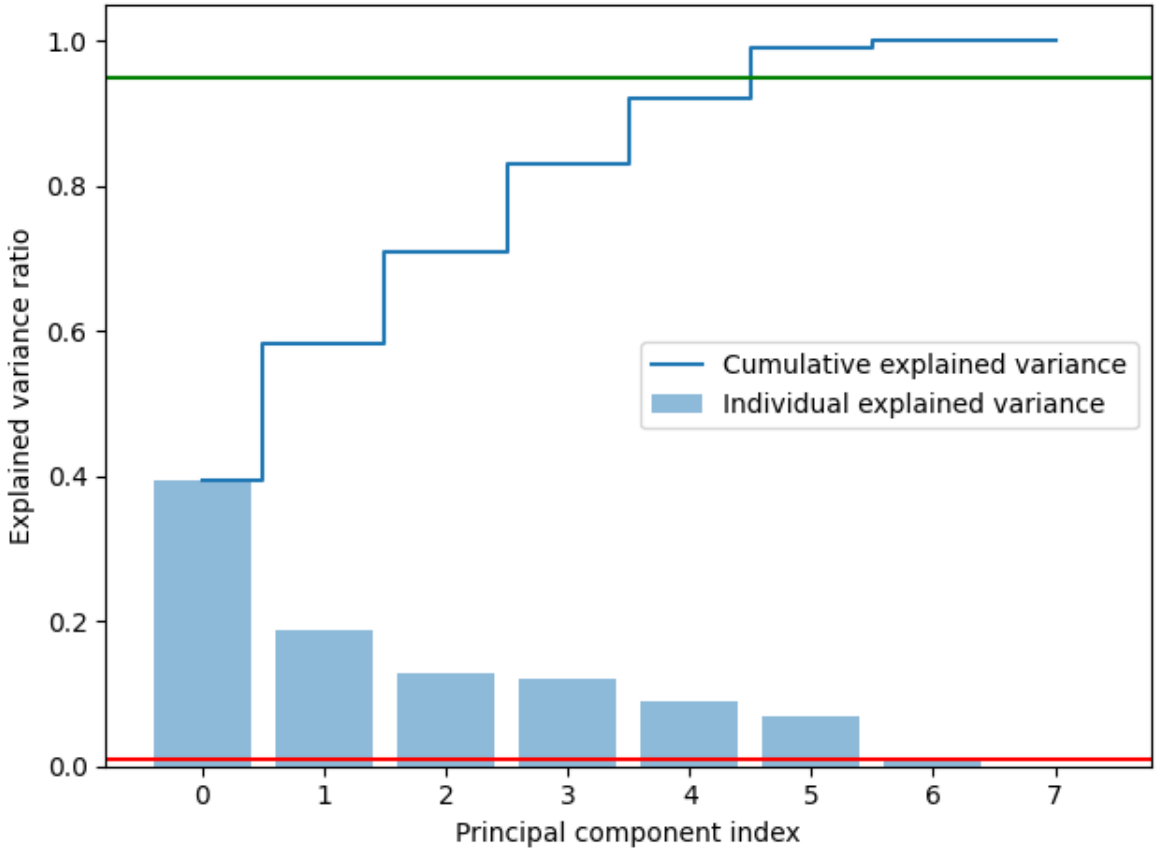
Figura 8 Matriz de correlación características seleccionadas
Fuente: elaboración propia

Esta correlación entre características no es deseada y afecta el proceso de entrenamiento de un modelo de aprendizaje automático, por tanto, se hizo necesario aplicar el método de componentes principales que contribuyó a dos efectos principales: el primero fue la eliminación de la fuerza de correlación entre

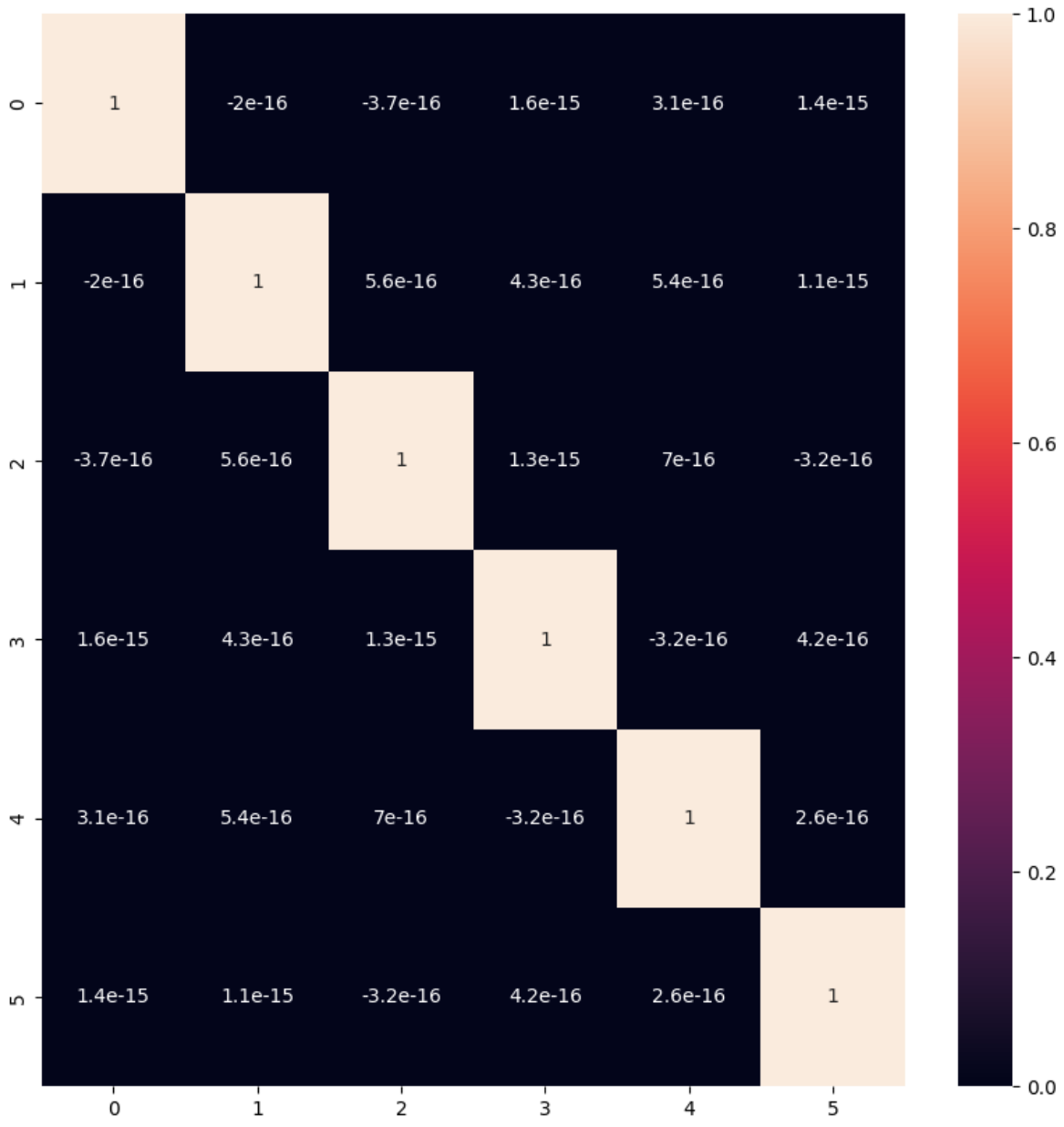
características y adicionalmente ayudó a reducir el número de características que se tendrían en cuenta.

Para establecer la cantidad apropiada de características principales que se requerían para explicar la varianza en las muestras se calculó la varianza acumulada con los componentes y se estableció el umbral de 0.95 para determinar la cantidad de componentes necesarios.

En la siguiente figura se ilustra la varianza explicada por cada una de las características y se observa la acumulación de explicación de la varianza, con este cálculo se determinó que el número apropiado de componentes principales para trabajar en los pasos sucesivos fue seis.



Al aplicar la matriz de correlación con esta muestra transformada por la aplicación de la PCA, se confirmó que los componentes que se van a emplear no tienen correlación fuerte entre ellos, como se puede ver en la siguiente figura.



5.6.3. Carga

Finalmente, los datos fueron cargados para alimentar los procesos encargados del modelamiento. El conjunto de datos contenía un total de con 3352 entradas distribuidas balanceadamente en dos categorías: phishing y legítimo con seis variables para cada una de las entradas.

5.7. Modelamiento

Antes de realizar el modelamiento se hizo una partición en el conjunto de datos para conservar una muestra estratificada con el 30% de los datos, esta partición se empleará al final del proceso para hacer las pruebas y tomar las mediciones de desempeño del modelo resultante. De esta forma se consiguieron 1173 muestras para cada una de las categorías.

Para seleccionar el algoritmo a emplear se tuvieron en cuenta los modelos clásicos de clasificación binaria referenciados en la bibliografía: *Logistic Regression, Support Vector Machine, Decision tree, Random Forest,Naive Bayes*

Para la evaluación se codificó una función que permitiera extraer las calificaciones *accuracy*, *precisión* y *recall* obtenidas para cada modelo. Para obtener resultados fidedignos se programó la función para hacer uso del k-fold con 5 dobleces y se tomó el promedio de cada una de las calificaciones como el valor a comparar.

En la siguiente gráfica se observan los resultados gráficos de la evaluación y en la tabla que le sigue se presentan los valores numéricos obtenidos.

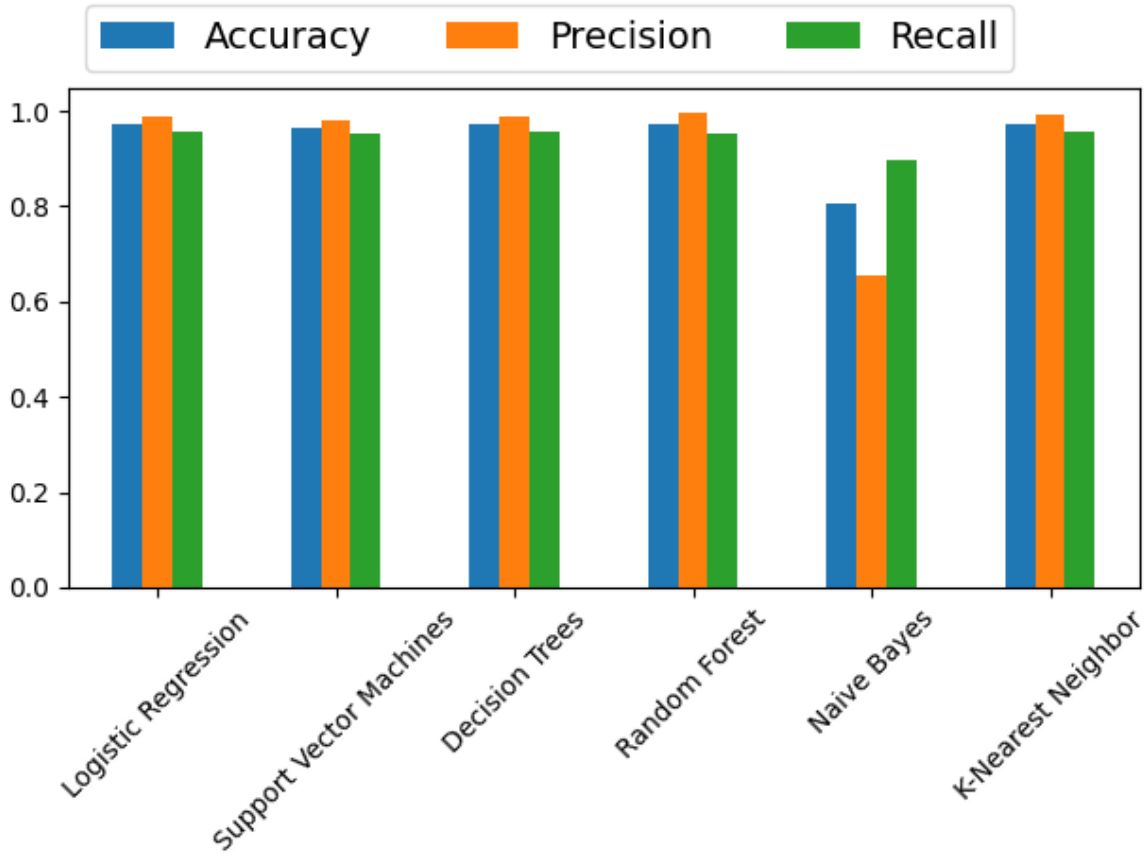


Figura 9 Comparativa de la evaluación de los diferentes modelos elegidos
Fuente: elaboración propia

Visto de forma tabular los resultados obtenidos se presentan en la siguiente tabla.

Algorithm	**Accuracy**	**Precision**	**Recall**
Logistic Regression	0.974960	0.996162	0.955624
Support Vector Machines	0.969100	0.995736	0.945517
Decision Trees	0.966329	0.978054	0.955664
Random Forest	0.974001	0.996592	0.953537
Naive Bayes	0.538945	0.290780	0.730261
K-Nearest Neighbor	0.971550	0.988283	0.956296

Dado que la diferencia en los resultados resulta mínima entre 3 diferentes algoritmos, se eligió el primero de ellos en orden del listado realizado, de esta forma se empleará la regresión logística.

5.8. Evaluación

Para hacer una adecuada evaluación del modelo se emplea la muestra que fue separada antes de cualquier tratamiento de los datos de entrada y que no ha tenido ningún contacto con el trabajo de modelado.

Este nuevo set de datos reducido es transformado con la aplicación de la estandarización y la PCA en la misma forma que se hizo con el conjunto de datos de entrenamiento para garantizar que el modelo funcione correctamente.

Tras contrastar la predicción del sistema y el valor real marcado en los datos se obtiene la matriz de confusión que se presenta en la siguiente figura.

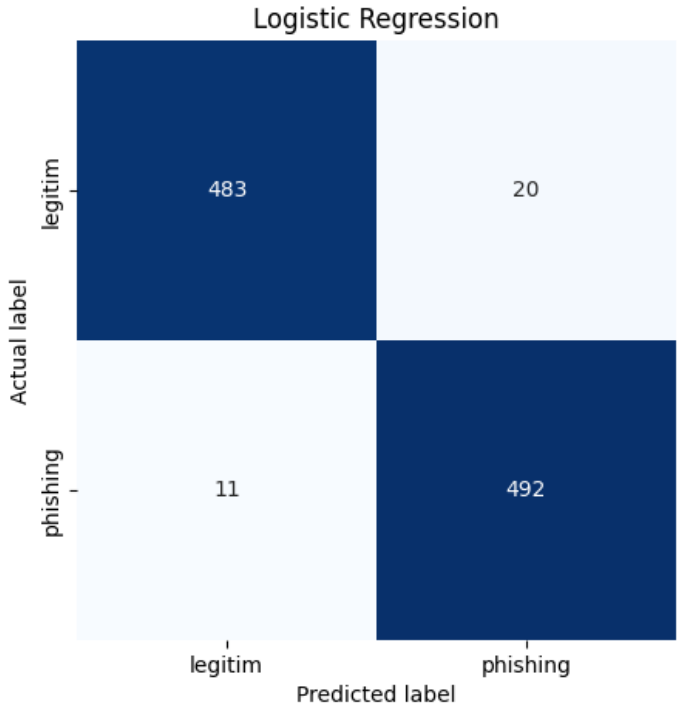


Figura 10 Matriz de confusión para la evaluación del modelo
Fuente: Elaboración propia

De esta matriz de confusión se observa que el modelo obtiene un desempeño de 11 falsos negativos (1.09%) y 20 falsos positivos (1.98%) en una muestra de 1006 entradas. Se observa que la tasa de falsos positivos es casi el doble que la de falsos negativos, aunque se mantiene bajo el umbral deseado del trabajo. En métricas de desempeño los resultados son los siguientes:

El accuracy es: 0.9738636363636364
La precisión es: 0.9940909090909091
El valor de recall es: 0.9554760748591556

El tiempo de procesamiento para la evaluación de 1006 datos es de 0.3 segundos lo que resulta bastante más bajo que el valor de 24 horas expuesto para la revisión manual y en términos de consumo de recursos computacionales es inferior a 1 segundo lo que permite procesar cargas de hasta 3000 evaluaciones por segundo constituyéndose en una alternativa eficiente para altos volúmenes de información.

6. Conclusiones y trabajo futuro

Se confirmó que la aproximación elegida respecto al análisis de las etiquetas de los hipervínculos resultó ser apropiada para abordar la clasificación de *phishing*, aportando así un enfoque novedoso en este ámbito para futuros desarrollos.

Como resultado del presente trabajo se pudo conseguir un sistema eficiente que basándose en el algoritmo regresión logística fue capaz de detectar con una precisión de 0.994 sitios web de phishing sin necesidad de parametrización previa con información de los sitios a proteger.

El sistema desarrollado presenta una alta eficiencia operacional puesto que no requiere la intervención manual de agentes para adelantar ninguna tarea previa ni posterior en la clasificación, lo que es una gran mejora frente a prácticas actuales de industria donde las parametrizaciones iniciales para proteger un sitio web particular toman varias horas hombre al igual que los procesos de confirmación de sitios de phishing.

De igual forma el sistema desarrollado muestra una alta efectividad frente a diversos tipos de despliegue de sitios de phishing dado que el paradigma empleado apunta a características propias de estos sitios que siguen siendo identificables sin importar el mecanismo de despliegue empleado.

La aproximación elegida cumple buenos estándares de desempeño y eficiencia computacional apropiados para asegurar su viabilidad económica en un eventual despliegue que se propone abordar como una investigación futura. Es importante este punto dado que como se expuso el problema del *phishing* presenta cifras de crecimiento bastante altas y es necesario que las mitigaciones resulten fácilmente implementables y replicables.

Se pudo confirmar que un acercamiento multivariable al problema genera una muy buena tasa de respuesta, dado que las técnicas de despliegue de sitios web

de phishing son variadas y para diferentes casos resulta más conveniente la aplicación de un determinado set de características.

Gracias a su alta eficiencia y precisión, el sistema de clasificación propuesto en este documento puede resultar muy valioso en un flujo de trabajo de análisis de amenazas, dado que los procesos de caracterización de amenazas pueden ser activados después de una clasificación positiva confiable, reduciendo costos y tiempos de procesamiento.

Dentro del alcance de este documento no se contemplaron los efectos de la inclusión de contenidos dinámicos en las páginas web de phishing, esta adición de contenidos dinámicos puede ayudar al ocultamiento de ciertas características empleadas por el modelo aquí desarrollado, así que se plantea como una investigación futura que puede complementar el sistema aquí descrito.

7. Apéndice 1 Descripción del hipervínculo por el estándar de W3C

7.1. Enlaces HTML - Hipervínculos

Los enlaces HTML son hipervínculos, que permiten hacer *click* y saltar a otro documento o sección de documento. Cuando se pasa el puntero del ratón sobre un hipervínculo la flecha del puntero cambia y presenta una pequeña mano.

Un enlace no tiene que ser texto. ¡Un enlace puede ser una imagen o cualquier otro elemento HTML!

7.1.1. Enlaces HTML - Sintaxis

La etiqueta HTML `<a>` define un hipervínculo. Tiene la siguiente sintaxis:

```
< a href="url">link text</a>
```

El atributo más importante del elemento `<a>` es el atributo `href`, que indica el destino del enlace.

El *texto del enlace* es la parte que será visible para el lector.

Al hacer clic en el texto del enlace, se enviará al lector a la dirección URL especificada.

7.1.2. Enlaces HTML: el atributo de destino

De forma predeterminada, la página vinculada se mostrará en la ventana actual del navegador. Para cambiar esto, debe especificar otro destino para el vínculo.

El atributo `de destino` especifica dónde abrir el documento vinculado y puede tener uno de los siguientes valores:

`_self` - Predeterminado. Abre el documento en la misma ventana/pestaña en la que se hizo clic

`_blank` - Abre el documento en una nueva ventana o pestaña

`_parent` - Abre el documento en el marco principal

`_top` - Abre el documento en el cuerpo completo de la ventana**.**

7.1.3. URLs absolutas vs. URLs relativas

Un enlace local (un enlace a una página dentro del mismo sitio web) se especifica con una **URL relativa** (sin el segmento ("https://www"), mientras que una URL absoluta si lo incluye

7.1.4. Enlace a una dirección de correo electrónico

El uso del atributo `mailto:` dentro del atributo `href` crea un enlace que abre el programa de correo electrónico del usuario (para permitirle enviar un nuevo correo electrónico):

7.1.5. Títulos de los enlaces

El atributo `title` especifica información adicional sobre un elemento. La información se muestra con mayor frecuencia como texto de información sobre herramientas cuando el puntero del ratón se mueve sobre el elemento.

8. Bibliografía

APWG. (2023). *Phishing Activity Trends Report 4th Quarter 2022.* APWG.

Avanan. (15 de Marzo de 2023). *How is your SOC handling phishing*. Obtenido de A Chech Point Company Website: https://www.avanan.com/blog/how-is-your-soc-handling-phishing

Avast Software s.r.o. (5 de Febrero de 2020). *Avast Academy: Phishing*. Obtenido de ¿Qué es el phishing? Avast Academy Website: https://www.avast.com/es-es/c-phishing

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Colin, S., & Rüdiger, W. (1999). *CRISP-DM.* Obtenido de The Modelling Agency: https://www.the-modeling-agency.com/crisp-dm.pdf

Cloudfare Inc. (2023). *2023 phishing report*. Obtenido de The Cloudfare Blog: https://blog.cloudflare.com/2023-phishing-report/

DHS. (2017, 10 2). *Media library: US Department of Homeland Security*. Retrieved from Department of Homeland Security Website: https://www.dhs.gov/medialibrary/assets/videos/21694

Ellis, J. (24 de Marzo de 2021). *The phishlabs blog: Fortra Company*. Obtenido de Most Phishing Attacks Use Compromised Domains and Free Hosting: https://www.phishlabs.com/blog/most-phishing-attacks-use-compromised-domains-and-free-hosting/

Feyyad. (1996). Data mining and knowledge discovery: making sense out of data. *IEEE Expert, vol. 11, no. 5*, 20-25.

Foroughi, F., & Luksch, P. (2018). Data Science Methodology for Cybersecurity Projects. *5th International Conference on Artificial Intelligence and Applications* (págs. 1-14). Rostock, Germany: https://arxiv.org/abs/1803.04219.

Fortra. (4 de Junio de 2019). *Agari email Security*. Obtenido de soc costs employee reported phishing: https://www.agari.com/blog/soc-costs-employee-reported-phishing

Harrington, P. (2012). *Machine Learning in Action.* Shelter Island: Manning Publications Co.

ICANN. (3 de 3 de 2022). *Compromsed Domain*. Obtenido de ICANN Wiki: https://icannwiki.org/Compromised_Domain

Jerry Felix, C. H. (1987). System Security: A Hacker's Perspective. *Interex Proceedings* (pág. 1:6). Interex Proceedings.

Jhon B. Rollins, P. (2015, June). *https://www.ibm.com/downloads/cas.* Retrieved from https://www.ibm.com/: https://www.ibm.com/downloads/cas/WKK9DX51

Paloalto networks. (2023). *Unit 42 2023 Ransonware and extortion report*. Obtenido de A Paloalto Networks Web Site: https://unit42.paloaltonetworks.com/newly-registered-domains-malicious-abuse-by-bad-actors/

Roy, S. &. (2022). A Large-Scale Analysis of Phishing Websites Hosted on Free Web Hosting Domains.

TechDay. (2023). *Story: Phishing remains most dominant fastest growing internet crime*. Obtenido de Security Brief: https://securitybrief.com.au/story/phishing-remains-most-dominant-fastest-growing-internet-crime

The latest phishing statistics. (2023). Obtenido de an AAG Web Site: https://aag-it.com/the-latest-phishing-statistics/

TransUnion. (2022). *Consumer Pulse, Colombia T4 2022.* TransUnion.

Abutaha, M., Ababneh, M., Mahmoud, K., & Baddar, S. A. H. (2021). URL Phishing Detection using Machine Learning Techniques based on URLs Lexical Analysis. *2021 12th International Conference on Information and Communication Systems, ICICS 2021*. https://doi.org/10.1109/ICICS52457.2021.9464539

Ahammad, S. H., Kale, S. D., Upadhye, G. D., Pande, S. D., Babu, E. V., Dhumane, A. V., & Bahadur, M. D. K. J. (2022). Phishing URL detection using machine

learning methods. *Advances in Engineering Software*, *173*. https://doi.org/10.1016/j.advengsoft.2022.103288

Alanzi, B. M., & Uliyan, D. M. (2022). Detection of Phishing Websites by Investigating Their URLs using LSTM Algorithm. *INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND NETWORK SECURITY*, *22*(5).

Aljofey, A., Jiang, Q., Rasool, A., Chen, H., Liu, W., Qu, Q., & Wang, Y. (2022). An effective detection approach for phishing websites using URL and HTML features. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-10841-5

Alsabah, M., Nabeel, M., Boshmaf, Y., & Choo, E. (2022). Content-Agnostic Detection of Phishing Domains using Certificate Transparency and Passive DNS. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3545948.3545958

Alshira'H, M. (2020). Detecting Phishing URLs using Machine Learning Lexical Feature-based Analysis. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(4). https://doi.org/10.30534/ijatcse/2020/242942020

Azeez, N. A., Misra, S., Margaret, I. A., Fernandez-Sanz, L., & Abdulhamid, S. M. (2021). Adopting automated whitelist approach for detecting phishing attacks. *Computers and Security*, *108*. https://doi.org/10.1016/j.cose.2021.102328

Chanti, S., & Chithralekha, T. (2020). Classification of Anti-phishing Solutions. *SN Computer Science*, *1*(1). https://doi.org/10.1007/s42979-019-0011-2

Cui, Q., Jourdan, G. V., Bochmann, G. V., Onut, I. V., & Flood, J. (2018). Phishing attacks modifications and evolutions. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11098 LNCS*. https://doi.org/10.1007/978-3-319-99073-6_12

Das Guptta, S., Shahriar, K. T., Alqahtani, H., Alsalman, D., & Sarker, I. H. (2022). Modeling Hybrid Feature-Based Phishing Websites Detection Using Machine Learning Techniques. *Annals of Data Science*. https://doi.org/10.1007/s40745-022-00379-8

Do, N. Q., Selamat, A., Krejcar, O., Herrera-Viedma, E., & Fujita, H. (2022). Deep Learning for Phishing Detection: Taxonomy, Current Challenges and Future

Directions. En *IEEE Access* (Vol. 10). https://doi.org/10.1109/ACCESS.2022.3151903

Elsadig, M., Ibrahim, A. O., Basheer, S., Alohali, M. A., Alshunaifi, S., Alqahtani, H., Alharbi, N., & Nagmeldin, W. (2022). Intelligent Deep Machine Learning Cyber Phishing URL Detection Based on BERT Features Extraction. *Electronics (Switzerland)*, *11*(22). https://doi.org/10.3390/electronics11223647

Frauenstein, E. D., & Von Solms, R. (2014). Combatting phishing: A holistic human approach. *2014 Information Security for South Africa - Proceedings of the ISSA 2014 Conference*. https://doi.org/10.1109/ISSA.2014.6950508

Ghaleb, F. A., Alsaedi, M., Saeed, F., Ahmad, J., & Alasli, M. (2022). Cyber Threat Intelligence-Based Malicious URL Detection Model Using Ensemble Learning. *Sensors*, *22*(9). https://doi.org/10.3390/s22093373

Guo, B., Zhang, Y., Xu, C., Shi, F., Li, Y., & Zhang, M. (2021). Hinphish: An effective phishing detection approach based on heterogeneous information networks. *Applied Sciences (Switzerland)*, *11*(20). https://doi.org/10.3390/app11209733

Jain, A. K., & Gupta, B. B. (2017). Phishing detection: Analysis of visual similarity based approaches. En *Security and Communication Networks* (Vol. 2017). https://doi.org/10.1155/2017/5421046

Jain, A. K., & Gupta, B. B. (2019). A machine learning based approach for phishing detection using hyperlinks information. *Journal of Ambient Intelligence and Humanized Computing*, *10*(5). https://doi.org/10.1007/s12652-018-0798-z

Jeeva, S. C., & Rajsingh, E. B. (2016). Intelligent phishing url detection using association rule mining. *Human-centric Computing and Information Sciences*, *6*(1). https://doi.org/10.1186/s13673-016-0064-3

Korkmaz, M., Sahingoz, O. K., & DIri, B. (2020). Detection of Phishing Websites by Using Machine Learning-Based URL Analysis. *2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020*. https://doi.org/10.1109/ICCCNT49239.2020.9225561

Liu, X., & Fu, J. (2020). SPWalk: Similar Property Oriented Feature Learning for Phishing Detection. *IEEE Access*, *8*. https://doi.org/10.1109/ACCESS.2020.2992381

Rao, R. S., & Pais, A. R. (2019). Detection of phishing websites using an efficient feature-based machine learning framework. *Neural Computing and Applications*, *31*(8). https://doi.org/10.1007/s00521-017-3305-0

Roy, S. S., Karanjit, U., & Nilizadeh, S. (2022). A Large-Scale Analysis of Phishing Websites Hosted on Free Web Hosting Domains. *arXiv preprint arXiv:2212.02563*.

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*. https://doi.org/10.1016/j.procs.2021.01.199

Singh, A. P., Kumar, V., Sengar, S. S., & Wairiya, M. (2011). Detection and prevention of phishing attack using dynamic watermarking. *Communications in Computer and Information Science*, *147 CCIS*. https://doi.org/10.1007/978-3-642-20573-6_21

Sonowal, G., & Kuppusamy, K. S. (2020). PhiDMA – A phishing detection model with multi-filter approach. *Journal of King Saud University - Computer and Information Sciences*, *32*(1). https://doi.org/10.1016/j.jksuci.2017.07.005

Sumathi, K., & Sujatha, V. (2019). Deep learning based-phishing attack detection. *International Journal of Recent Technology and Engineering*, *8*(3). https://doi.org/10.35940/ijrte.C6527.098319

Tan, C. C. L., Chiew, K. L., Yong, K. S. C., Sebastian, Y., Than, J. C. M., & Tiong, W. K. (2023). Hybrid phishing detection using joint visual and textual identity. *Expert Systems with Applications*, *220*. https://doi.org/10.1016/j.eswa.2023.119723

Tang, L., & Mahmoud, Q. H. (2021). A Survey of Machine Learning-Based Solutions for Phishing Website Detection. En *Machine Learning and Knowledge Extraction* (Vol. 3, Número 3). https://doi.org/10.3390/make3030034

Tang, L., & Mahmoud, Q. H. (2022). A Deep Learning-Based Framework for Phishing Website Detection. *IEEE Access*, *10*. https://doi.org/10.1109/ACCESS.2021.3137636

Uddin, M. M., Arfatul Islam, K., Mamun, M., Tiwari, V. K., & Park, J. (2022). A Comparative Analysis of Machine Learning-Based Website Phishing Detection

Using URL Information. *2022 5th International Conference on Pattern Recognition and Artificial Intelligence, PRAI 2022*. https://doi.org/10.1109/PRAI55851.2022.9904055

Ulqinaku, E., Lain, D., & Capkun, S. (2019). 2FA-PP: 2nd factor phishing prevention. *WiSec 2019 - Proceedings of the 2019 Conference on Security and Privacy in Wireless and Mobile Networks*. https://doi.org/10.1145/3317549.3323404

Wang, C., Hu, Z., Chiong, R., Bao, Y., & Wu, J. (2020). Identification of phishing websites through hyperlink analysis and rule extraction. *Electronic Library*, *38*(5–6). https://doi.org/10.1108/EL-01-2020-0016

Xuan, C. Do, Nguyen, H. D., & Nikolaevich, T. V. (2020). Malicious URL detection based on machine learning. *International Journal of Advanced Computer Science and Applications*, *11*(1). https://doi.org/10.14569/ijacsa.2020.0110119

Yang, Y. (2019). Effective Phishing Detection using Machine Learning Approach. *Case Western Reserve University*.

Zhang, W., Lu, H., Xu, B., & Yang, H. (2013). Web phishing detection based on page spatial layout similarity. *Informatica (Slovenia)*, *37*(3).