UNIVERSIDAD SEÑOR DE SIPÁN

ESCUELA DE POSGRADO

TESIS

"MODELO DE MACHINE LEARNING EN LA DETECCIÓN DE SITIOS WEB PHISHING"

PARA OPTAR EL GRADO ACADÉMICO DE DOCTOR EN CIENCIAS DE LA COMPUTACIÓN Y SISTEMAS

Autor:

Mg. Villegas Cubas, Juan Elías

https://orcid.org/0000-0001-7026-9767

Asesor:

Dr. Bustamante Quintana, Pepe Humberto

https://orcid.org/0000-0001-9842-8432

Línea de Investigación:

Infraestructura, Tecnología y medio ambiente

Pimentel – Perú

2021

UNIVERSIDAD SEÑOR DE SIPÁN

ESCUELA DE POSGRADO

DOCTORADO EN CIENCIAS DE LA COMPUTACIÓN Y SISTEMAS

"MODELO DE MACHINE LEARNING EN LA DETECCIÓN DE SITIOS WEB PHISHING"

AUTOR

MG. VILLEGAS CUBAS, JUAN ELIAS

PIMENTEL – PERÚ

2021

MODELO DE MACHINE LEARNING EN LA DETECCIÓN DE SITIOS WEB PHISHING

APROBACIÓN DE LA TESIS

_____

Dr. Dios Castillo Christian Abraham

Presidente del jurado de tesis

_____ _____

Dr. Callejas Torres Juan Carlos

Secretario del jurado de tesis

Dr. Bustamante Quintana Pepe Humberto

Vocal del jurado de tesis

Dedicatoria

A mis padres Segundo Teófilo Villegas Vásquez (Q.E.P.D.) y María Dionisia Cubas Guevara, por creer siempre en mí, por su invalorable amor, sacrificio y esfuerzo que me permiten seguir adelante y por todas sus enseñanzas que sirvieron para lograr todo lo que soy ahora.

A mis hijos Juan Leonel, Ángel Gabriel y Victoria Carolina, y a mi esposa Zoila Carolina por ser las personas que me impulsan para seguir mejorando cada día; por el sacrificio y comprensión que realizan para lograr las metas paso a paso.

A mis hermanos María, Lidia, Zarela, Josué, Marcos, Elizabeth y Yanina, por el apoyo incondicional, el cariño inmenso y la confianza hacia mi persona.

Y a mis sobrinos como muestra de que nunca dejen de creer en ustedes y que son capaces de lograr cada meta que se tracen en la vida.

Mg. Juan Elías Villegas Cubas

Agradecimientos

A Dios, por darme vida y sobre todo por brindarme a la familia que tengo. A mis docentes del doctorado por sus sabias enseñanzas y sus consejos.

En especial a mi asesor metodológico Dr. Juan Carlos Callejas Torres por su motivación y enseñanzas que me ha permitido encaminar y terminar la Tesis. Y a mi asesor Dr. Pepe Humberto Bustamante Quintana por su guía en el desarrollo de la tesis.

A mis amigos y colegas que contribuyeron con sus conocimientos y aportes a mejorar mi trabajo; gracias porque su apoyo fue vital para lograr que éste culminara con éxito.

Mg. Juan Elías Villegas Cubas

Resumen

En la actualidad se evidencia el crecimiento de ataques informáticos y de forma específica los ataques phishing, la presente investigación tiene como problema fundamental el rendimiento en la detección de sitios web phishing. Las causas encontradas sugieren profundizar el proceso de la ciberseguridad y la detección de phishing, por lo que se plantea como objetivo: Aplicar un sistema de detección de phishing, sustentada en un modelo de machine learning, para el rendimiento en la detección de sitios web falsos.

Se propone un modelo de machine learning en la detección de sitios web phishing, construida en seis dimensiones: Sitio web, inteligencia de amenazas, preparación de datos, algoritmos de machine learning, entrenamiento y detección; visto holísticamente que permite la integración de todas las dimensiones; este modelo propuesto se materializa mediante un sistema de detección de phishing que se desarrolla en seis fases: recolección de datos, preparación de datos, selección de algoritmos, entrenamiento del sistema, detección de sitios phishing y evaluación del rendimiento.

Finalmente, se implementó el sistema de detección de phishing, utilizando datos de 11055 sitios web que son clasificados como sitios web legítimos y sitios web phishing, de los cuales 2211 sitios web se utilizaron para la evaluación del rendimiento del sistema y se obtiene un accuracy de 97.42% en la detección correcta de forma global de los sitios web, que es mayor en comparación con los resultados de estudios previos.

Palabras Clave.

Inteligencia de amenazas, aprendizaje automático, sitios web falsos, sistema de detección de phishing, anti-phishing.

Abstrac

At present, the growth of computer attacks and specifically phishing attacks is evidenced, the present investigation has as a fundamental problem the performance in the detection of phishing websites. The causes found suggest deepening the process of cybersecurity and the detection of phishing, for which the objective is: Apply a phishing detection system, based on a machine learning model, for the performance in the detection of fake websites.

A machine learning model is proposed in the detection of phishing websites, built in six dimensions: Website, threat intelligence, data preparation, machine learning algorithms, training and detection; seen holistically that allows the integration of all dimensions; This proposed model is materialized through a phishing detection system that is developed in six phases: data collection, data preparation, algorithm selection, system training, detection of phishing sites, and performance evaluation.

Finally, the phishing detection system was implemented, using data from 11,055 websites that are classified as legitimate websites and phishing websites, of which 2,211 websites were used for the evaluation of the performance of the system and a accuracy of 97.42% in the correct detection of websites globally, which is higher compared to the results of previous studies.

Keywords:

Threat intelligence, machine learning, fake websites, phishing detection system, anti-phishing

Índice

Lista de Figuras.

Lista de Tablas

I. INTRODUCCION.

1.1. Realidad Problemática.

En la actualidad muchas aspectos de nuestra vida cotidiana han cambiado con el uso de las Tecnologías de la Información y Comunicación (TIC); la mayoría de personas lleva siempre un Smartphone, utiliza el correo electrónico, envía mensajes de texto y realiza videoconferencias para comunicarse; se puede hacer transferencias de dinero, compra en línea a través de un dispositivo conectado, la educación ha cambiado la forma de aprender, muchas personas trabajan desde su casa a través de una conexión y reciben atención médica mediante una conexión virtual.

Las redes informáticas se han convertido en una herramienta importante para todo tipo de organizaciones, como instituciones financieras, médicas, industriales, de transporte, educativas, etc. Sin embargo, el crecimiento del uso de las redes e internet ha conllevado a mejorar la protección frente a los ciberataques.

Los ciberataques afectan a la intimidad e integridad de los datos de las personas, organizaciones y gobiernos, causan cuantiosas pérdidas económicas, según (Singh & Sharma, 2019) aproximadamente 2 billones de dólares de pérdidas en el 2019 debido a los ciberataques. El ciber crimen es tema de mucha importancia tanto que el FBI (Departamento de Justicia de los Estados Unidos) lo considera al mismo nivel que el terrorismo o la de contra inteligencia (FBI, 2019).

En la literatura se encuentra varios tipos de ciberataques, según (Bendovschi, 2015) los principales ciberataques son: del hombre en el medio, de fuerza bruta, de denegación de servicio distribuido, ransomware y phishing.

Los ciberataques han afectado a todo tipo de organizaciones para obtener información confidencial y privilegiada de sus clientes, empleados y del negocio, incluso se le asigna la muerte de una persona debido a un ciberataque ransomware (Clarin, 2020).

Los ciberataques cada día van creciendo en número, tamaño y velocidad, según (Medina & Molist, 2017) los ataques ransomware crecieron en 800%, el Phishing creció hasta el 100% en algunos meses; entre otros ataques en crecimiento son la denegación de servicio y el robo de datos personales.

Según Modi (2019) cerca de cuatro millones de ataques de denegación de servicio distribuido en seis meses, el crecimiento en número de ataques es del 39%, el tamaño en los ataques creció del 776% entre 100Gbps y 400Gbps.

Según APWG (2021) el número de ataques phishing crece continuamente, alcanzando un récord en crecimiento en el mes de octubre del 2020 con 225,304 ataques phishing, como se muestra en la figura 1.

Figura 1. Ataques phishing en el 2020. Fuente (APWG, 2021)

Entre las principales manifestaciones del problema del phishing se muestran a continuación.

- El 84% de profesionales en Estados Unidos indican que han sufrido al menos un tipo de incidente de seguridad y ponen a los ataques phishing y ransomware, como los tipos de ataques con mayor ocurrencia. (Trend Micro, 2021)

- El 31% de empresas en Latinoamérica han percibido un aumento de los ciberataques en el 2020 y la principal amenaza de la ciberseguridad son los eventos relacionados al phishing. (Marsh & Microsoft, 2020)
- El número de ataques phishing crecieron continuamente y hasta se duplicaron a lo largo del 2020, llegando en octubre a 225,304 ataques phishing. (APWG, 2021)
- Los países de Latinoamérica que fueron más afectados con los ataques phishing durante el 2020 son las empresas de Brasil (26,4%), seguidos por las empresas de Perú (22,8%) y luego por las empresas de México (12%). (Eset, 2021)
- En el Perú hasta octubre del 2021, se eleva el número de denuncias de ciberdelitos, siendo la modalidad más frecuente el phishing, que en el 2021 se duplicaron con respecto al 2020. (Divindat , 2021)
- En el Perú, en de enero a septiembre del 2021, se han generado 259 alertas integradas de Seguridad Digital, y en la mayoría de las alertas se ha reportado a los ataques phishing. (PECERT, 2021)
- Según (Trend Micro, 2021) el 50% de los profesionales en Estados Unidos consideran ineficiente las técnicas de hacer frente al phishing y al ransomware.
- El 65% de envíos de phishing ingresan a las bandejas de los usuarios. (Trend Micro, 2021).
- Además, el 65% de los usuarios hacen clic en los enlaces phishing. (Trend Micro, 2021).
- Los sistemas de detección de phishing en el 2020 neutralizaron 434 898,635 phishing, sin embargo, representa una menor cantidad de phishing detectado en comparación con el año 2019. (Kaspersky, 2021)

Las manifestaciones enumeradas en el párrafo anterior se sintetizan en el problema científico: Insuficiencias en el proceso de la ciberseguridad y la detección de phishing, **limitan el rendimiento en la detección de sitios web falsos.**

En la búsqueda de posibles causas del problema antes planteado, se observan.

- Insuficiente diagnóstico contextual en la detección de sitios web falsos, durante el proceso de ciberseguridad.
- Limitados referentes teóricas del proceso de ciberseguridad y la detección de phishing.
- Limitadas referencias prácticas sistemáticas del proceso de ciberseguridad y la detección de phishing.
- Insuficiente uso de modelos de machine learning, en el proceso de ciberseguridad y la detección de phishing.
- Insuficientes sistemas de detección de phishing para el desarrollo del proceso de ciberseguridad y la detección de sitios web phishing.

Estas valoraciones causales sugieren profundizar en el **proceso de la ciberseguridad y la detección de phishing**, objeto de la presente investigación.

El **campo de acción** que de esta investigación es la dinámica del proceso de la ciberseguridad y la detección de phishing.

Es así que existe una **brecha epistémica** en donde el estudio del proceso de la ciberseguridad, la detección de phishing y su dinámica revelan, que no ha sido lo suficientemente aplicados modelos y sistemas para la detección de phishing, desde una lógica que integre la información obtenida de las URL, la información obtenida de inteligencia de amenazas y las técnicas de machine learning en la detección de sitios web falsos.

1.2. Trabajos Previos

Los primeros trabajos para la detección de intrusos en las redes informáticas estaban relacionados con las auditorias de seguridad, que consistía en la revisión manual de las actividades de los usuarios.

Anderson (1972) documentó un sistema de clasificación que distingue entre ataques internos y los ataques externos, basándose en los accesos de un usuario tiene o no tiene para ingresar a un ordenador, este trabajo se considera como el primero que habla de la detección de intrusiones y que remplazaba a las auditorias de seguridad.

Denning (1987) propone un Sistema Experto de Detección de Intrusiones (IDES) en tiempo real con la capacidad de detectar robos, penetraciones y otras formas de abuso informático. Su propuesta se basa en el monitoreo de los registros de auditoría, que permite detectar patrones anormales en el uso del sistema.

Heberlein (1995) en la Universidad de California desarrollaron el "Network System Monitor", un Sistema de Detección de Intrusos con capacidad de monitoreo de red. El funcionamiento del NSM es la base de muchos de los sistemas de detección de intrusos de red que se utilizan hoy en día.

Rami , Fadi & Lee (2014) indican que se puede combatir el phishing con soluciones legales, educación y soluciones técnicas. Desarrollaron y evaluaron una red neuronal con 500 épocas, utilizaron un conjunto de datos con 1400 sitios web entre phishing y legítimos; y obtuvieron un accuracy de 92.48%.

Mohammad, Thabtah & McCluskey (2014) indican que para combatir el phishing se puede hacer mediante soluciones legales, con capacitaciones a los usuarios y con soluciones técnicas; y proponen un modelo para predecir ataques phishing basado en redes neuronales artificiales. Experimentan con un conjunto de datos de 19 características y obtienen un accuracy máximo de entrenamiento de 94.07% con 1000 épocas.

De la Hoz (2016) propone un enfoque PCA/FDR para la detección de ataques en redes, aplica el poder discriminante para la selección de características. Obtiene un alto rendimiento en la clasificación de los ataques usando el ajuste de las probabilidades de activación previa, usa las métricas de a precisión de la clasificación, o la sensibilidad.

Jain & Gupta (2017) precisan que la detección de phishing se realiza mediante la educación a los usuarios y mediante el uso de software; las técnicas basadas en software son el uso de listas negras, similitud visual, motores de búsqueda y aprendizaje automático. Proponen un modelo anti-phishing que extrae los datos solamente del lado cliente, utilizando las características de la url y el código fuente HTML, evalúan varios algoritmos de aprendizaje automático con un conjunto de datos de 2141 phishing y 1918 sitios web legítimos, obtienen un accuracy máximo con Random Fortest.

Yi, y otros (2018) presentan dos tipos de funciones para la detección de sitios web phishing, que una es la característica original y la otra es las características de interacción. Con un conjunto de datos pequeño los parámetros adecuados y Luego entrenan el modelo DBN y evalúan DBN obteniendo el 89.20% de accuracy.

Feng, y otros (2018) proponen un modelo para la clasificación de sitios web phishing aplicando redes neuronales, además evaluaron el rendimiento de los algoritmos como Naive Bayes, regresión logísticas, arboles de decisión, LSVM, RSVM y análisis de discriminante líneas (LDA). Utilizaron el conjunto de datos con acceso público en el repositorio UCI, obteniendo un resultado de accurracy de 97.71%.

Jain & Gupta (2018) proponen un sistema para detectar url phishing denominado PHISH-SAFE basado en aprendizaje automático y en las características de la URL, el modelo es entrenado con un conjunto de datos de más de 33000 direcciones URL, con 14 características URL seleccionadas; usaron los clasificadores SVM y Naive Bayes, obtuvieron un accuracy de mas de 90% mediante el algoritmo SVM.

Niakanlahiji, Chu, & Al-Shaer (2018) proponen PhishMon, un marco de aprendizaje automático con funciones para detectar páginas web de phishing. Se basa en un conjunto de datos de quince características que se pueden calcular de manera eficiente desde una página web sin requerir servicios de terceros, como motores de búsqueda o servidores de WHOIS. Estas funciones capturan varias características de las aplicaciones web legítimas, así como sus infraestructuras web subyacentes. A través de una evaluación de un conjunto de datos que consta de 4800 phishing distintos y 17,500 páginas web benignas distintas, demuestran que PhishMon puede distinguir el phishing invisible de las páginas web legítimas con un grado muy alto de precisión. En los experimentos, PhishMon logró una acurracy del 95,4%

Abutair, Belghith, & Al-Ahmadi (2018) proponen un sistema de detección de phishing de razonamiento basado en casos (CBR-PDS) que se basa en casos anteriores para detectar ataques de phishing y que puede adaptarse para detectar nuevos ataques de phishing. Evalúan al modelo propuesto con varios conjunto de datos y obtienen un accuracy máximo de 96.26%.

Patil y otros (2018) utilizan tres enfoques para la detección de sitios web phishing, el primero es analizando las características de la URL, el segundo es verificando la legitimidad del sitio web y el tercero basado en la apariencia visual verificando la autenticidad del sitio web. Evalúan los datos con los algoritmos regresión logística, arboles de decisión y random forest, obteniendo un accuracy máximo de 96.58% con Random Forest.

Wei, y otros (2019) diseñaron un sensor para la detección de phishing aplicando técnicas de aprendizaje profundo, utilizaron un conjunto de datos de 1’523,966 direcciones URL con sitios legítimos y sitios phishing. El modelo propuesto fue un Red Neuronal Profunda y logró tener una tasa de detección real de 86.63% de accuracy.

Kumar y otros (2019) utilizan el conjunto de datos de código abierto Kaggle, realiza la extracción de las características y trabaja con los siguientes parámetros: Longitud de la URL, dirección IP, subdominio, uso de HTTPs, trafico del sitio web, SVM, puntos, SSL y vectores de características. El modelo propuesto evalúa diferentes algoritmos como Naive Bayes, Random Forest, KN vecino, obteniendo mejores resultados de clasificación con el algoritmo SVM.

Ubing y otros (2019) utilizan la selección de características y se integra con voting y se compara con diferentes modelos de clasificación, incluidos el Random Forest y regresión logística. Usan a la estructura y componentes de la URL, y el conjunto de datos con acceso público en el repositorio UCI, obteniendo un accuracy de 95%.

Wang, Zhang, Luo, & Zhang (2019) proponen un modelo denominado PDRCNN para la detección de sitios web phishing utilizando únicamente las características de la dirección URL del sitio web. Combinar dos tipos de rede y generan un conjunto de datos de casi 500,000 URL obtenidas a través de Alexa y PhishTank. Los resultados experimentales muestran que PDRCNN logra un accuracy de detección del 95.79%.

Zabihimayvan & Doran (2019) aplican la teoría Fuzzy Rough Set (FRS) como una herramienta para seleccionar las características más efectivas de tres conjuntos de datos. Y se seleccionan tres clasificadores para entrenar y validar con un conjunto de datos de 14000 direcciones de sitios web y lograron obtener un accuracy máximo de 95% con el clasificador Random Forest.

Sountharrajan, y otros (2019) utilizan las características de las direcciones URL seleccionadas de forma dinámica que dependen del tipo de aprendizaje. Evalúan los resultados aplicando las técnicas de aprendizaje profundo DBM y SAE red neuronal profunda; siendo el modelo DNN la que disminuye la tasa de falsos positivos, y brindando mejores resultados.

Kulkarni & Brown (2019) indican que debido a que los humanos son tan susceptibles a ser engañados, es necesario contar con métodos automatizados para diferenciar un sitio web legítimo de uno falso o phishing. Desarrollan un sistema utilizando técnicas de aprendizaje automático como el Árbol de decisión, Naive Bayes, SVM, y una red neuronal para clasificar los sitios web en función de la URL. Probaron un conjunto de datos con 1353 direcciones URL del mundo real y lograron obtener un mejor rendimiento de 91.5%, con los árboles de decisión.

Abdulhamit & Kremicb (2020) comparan algoritmos de machine learning en modo simple, en modo Adaboost y modo MultiBoosting, para la detección de sitios web phishing, usa el conjunto de datos con acceso público en el repositorio UCI y la evaluación da como resultado que el Adaboost con Maquinas de Vectores de Soporte brinda los mejores resultados.

Christou y otros (2020) consideran que incluso con la formación adecuada y una alta conciencia puede resultar difícil que un usuario pueda clasificar adecuadamente una página que visita como sitio phishing. También indican que la detección tradicional con listas de bloqueo y el análisis de contenido requiere mucho tiempo y verificación humana. Desarrollan un sistema de filtrado predictivo de los sitios web phishing, con algoritmos de Maquinas de Vectores de Soporte, Random Forest, evalúan el rendimiento del sistema con un conjunto de datos propio y obtienen un rendimiento máximo de 90% con el algoritmo Maquinas de Vectores de Soporte.

Chavan (2020) proponen utilizan los algoritmos de regresión logísticas, arboles de decisión, Random Forest, KN vecinos, SVM y redes neuronales artificiales, para evaluar el rendimiento con el conjunto de datos disponible en kagle con 1782 registros y 19 características; y obtienen un rendimiento máximo de 96% de rendimiento con Random Forest.

Zamir, y otros (2020) realizan una comparación de enfoques de aprendizaje automático supervisado y modelado de apilamiento para detectar sitios web phishing. Propone características con PCA y apila mecanismos de machine learning como Maquinas de Vectores de Soporte, Naive Bayes, Random Forest, KN vecinos, Bagging y redes neuronales; usa el conjunto de datos disponible en Kaggle con 11055 sitios web y con 32 atributos; y obtiene un rendimiento máximo de 97.4%.

Shahrivari, Darabi, & Izadi (2020) indican que las formas para evitar los ataques phishing es capacitando al usuario a que estén preparados para ataques phishing futuros, se trata de un método preventivo y se capacita a los usuarios a distinguir entre los sitios web phishing y los sitios web legítimos; sin embargo, los usuarios tienden a olvidarse y a cometer errores; la otra forma es mediante un sistema de detección automatizado que advierta al usuario, y desarrollan un sistema de clasificación de phishing, utilizando un conjunto de datos con 6157 sitios web legítimos y 4898 sitios web phishing y evalúan doce algoritmos de machine learning.

Alsariera, Adeyemo, Balogun, & Alazzawi (2020) propone cuatro modelos de meta aprendizaje basados todos ellos en el algoritmo Extra Tree: AdaBoost (ABET), Bagging (BET), Rotation Forest (RoFBET) y LogitBoost (LBET), evalúan el desempeño de los mismos y obtienen un accuracy de no menor de 97%, y un rendimiento máximo de 97.404% con el modelo BET.

Opara, Wei, & Chen (2020) proponen HTMLPhish, un enfoque de clasificación de páginas web de phishing basado en datos y basado en aprendizaje profundo. Específicamente, HTMLPhish recibe el contenido del documento HTML de una página web y emplea redes neuronales convolucionales (CNN) para aprender las dependencias semánticas en el contenido textual del HTML. Realizan realizan experimentos integrales en un conjunto de datos de más de 50.000 documentos HTML y que arroja un rendimiento superior al 93%.

1.3. Teorías relacionadas al tema.

1.3.1. Ciberseguridad y detección de phishing

El proceso de seguridad informática se considera como parte del proceso de seguridad de la información, y según (ISO27000, 2017) tiene como propósito la protección de los riesgos, logrando que estos sean conocidos, asumidos, gestionados y minimizados por toda organización.

En los últimos años, está tomado mayor importancia al proceso de la ciberseguridad que según (ISACA, 2015) consiste en el proceso de proteger a los activos de información de una organización, cuando es procesada, almacenada y transmitida en dispositivos digitales y en las redes informáticas de las amenazas.

La Ciberseguridad contempla las diferentes normas, prácticas, herramientas y conceptos que se relacionan con la seguridad de la información y la seguridad TI operacional. Es decir, la ciberseguridad, está considerada como parte de la seguridad de la información y se orienta a la protección de los activos de información en formato digital que se transmiten a través de sistemas interconectados.

La Organización Internacional de Normalización ha publicado el estándar ISO/IEC 27032 para la ciberseguridad (ISO, 2017) . Su objetivo es de facilitar directrices para mejorar el estado de la ciberseguridad en infraestructuras críticas.

El NIST (Instituto Nacional de Estándares y Tecnología) ha publicado un Marco de Trabajo de Ciberseguridad para infraestructuras críticas. El Framework NIST se compone de tres elementos principales: El marco Core, los niveles de implementación Tiers y los perfiles del marco (Profiles). El marco principal (Core) emplea cinco funciones fundamentales de la ciberseguridad que son: Identificar, Proteger, Detectar, Responder y Recuperar (NIST, 2018).

Para el NIST la detección de intrusos es en proceso de monitorear eventos que ocurren en sistemas de computación o redes, con la finalidad de detectar signos de posibles incidentes como pueden ser violaciones, amenazas de violación de políticas de seguridad o uso de recursos en forma abusiva (Scarfone & Mell, 2007).

Se han propuesto varias definiciones de Phishing, investigadores e instituciones de ciberseguridad siguen discutiendo referido al tema que sigue evolucionado de acuerdo con la función y al contexto.

Phishing es el proceso de engaño a los usuarios de las redes informáticas para que divulguen información confidencial para fines nefastos. Los ataques phishing se realiza por correos masivos hasta millones de direcciones de destinatarios o ataques altamente direccionados a usuarios específicos. (Ollmann, 2004).

El phishing es una actividad fraudulenta definida como la creación de una página web falsa pero muy similar a la existente con la finalidad de engañar a un usuario para que ingrese datos personales, financieros o de contraseña (Merwe, Loock, & Dabrawski, 2005).

El phishing es una forma de ingeniería social en la que un atacante también conocido como “phisher” intenta obtener de forma fraudulenta información confidencial o sensible de usuarios legítimos, imitando de forma automatizada las comunicaciones electrónicas de una organización. (Jakobsson & Myers, 2006)

El phishing es un delito que se emplea las técnicas de ingeniería social y el engaño para robar identidad personal y las credenciales de las cuentas financieras de los usuarios. Los esquemas de ingeniería social se aprovechan de las victimas desprevenidas, engañándoles, haciéndoles creer que están tratando con una parte legítima y de confianza. (APWG, 2021)

El phishing utiliza diversos vectores de ataque como ataques del hombre en el medio, registradores de claves y la falsificación completa de un sitio web. (Ollmann, 2004)

1.3.2. Machine Learning

Machine Learning, está definido según (Gori, 2018) como un tipo de Inteligencia artificial que proporciona a un sistema (hardware y/o software) la capacidad de aprender, sin ser explícitamente programadas.

El aprendizaje automático es una área de inteligencia artificial que permite que un sistema aprenda a partir de un conjunto de datos en lugar de a través de ser explícitamente programados (Hurwitz & Kirsch, 2018).

El aprendizaje automático según (Mueller & Guido, 2016) permite extraer conocimiento de los datos. Las técnicas de aprendizaje automático se han vuelto omnipresente en casi todas las actividades de nuestra vida cotidiana. Desde recomendaciones automáticas como por ejemplo qué películas elegir, qué tipo de comida consumir o qué productos adquirir, hasta reconocer a amigos en imágenes, muchas aplicaciones y dispositivos actuales utilizan algoritmos de aprendizaje automático en su funcionamiento.

Según (Mueller & Guido, 2016) son dos los enfoques para los problema de aprendizaje; el aprendizaje supervisado y el aprendizaje no supervisado; sin embargo (Hurwitz & Kirsch, 2018) indica además al aprendizaje por refuerzo y al aprendizaje profundo como enfoques del aprendizaje automático.

El aprendizaje supervisado según (Mueller & Guido, 2016), se usa datos de entradas también llamadas características y salidas deseadas; y el algoritmo encuentra una manera de producir la salida deseada dada una entrada especifica. El algoritmo puede crear o predecir una salida para datos de entrada sin la ayuda de una persona.

El aprendizaje supervisado según (Hurwitz & Kirsch, 2018) inicia con un conjunto de datos de características y una etiqueta que brinda un significado de los datos. El aprendizaje supervisado tiene como objetivo buscar y reconocer patrones en los datos analizados.

Los modelos de aprendizaje supervisado se aplican en una amplia variedad de problemas comerciales, como la detección de fraudes, detección de enfermedades, sistemas de recomendación o reconocimiento de voz.

En el aprendizaje no supervisado, el algoritmo solo recibe los datos de entrada o características y no se proporcionan las etiquetas de los datos de salida conocidos, es decir no se bridna el significado de los datos. (Mueller & Guido, 2016)

El aprendizaje no supervisado según (Hurwitz & Kirsch, 2018) se usa principalmente cuando el problema solo usa una gran cantidad de datos de solamente las características, sin etiquetar. Los algoritmos de aprendizaje no supervisados segmentan los conjuntos de datos en grupos de ejemplos (clústeres) o grupos de características. Los datos sin etiquetar crean los valores de los parámetros y la clasificación de los datos.

En el aprendizaje por reforzamiento según (Hurwitz & Kirsch, 2018) algoritmo recibe una retroalimentación por parte del usuario, para el análisis de los datos y obtener un mejor resultado; es decir, se aprende con estímulos de da un peso alto si está cerca del objetivo o un peso bajo si comete errores en el resultado.

El aprendizaje profundo o deep learning, es un tipo de aprendizaje que se basa en redes neuronales y se utiliza para problemas mas complejos como en los problemas de aprendizaje con imágenes.

Los algoritmos de aprendizaje automático supervisados se clasifican en dos tipos de problemas clasificación y regresión.

En la clasificación según (Mueller & Guido, 2016), el objetivo es predecir la salida del tipo categórico o de clase, que es una alternativa dentro de las posibilidades; si la lista de posibilidades son dos se denomina problemas de clasificación binaria, pero si la lista de posibilidad hay mas de dos, entonces se denominada clasificación multiclase.

En la regresión según (Mueller & Guido, 2016), el objetivo es predecir la salida numérica, puede se continuo o de punto flotante. La predicción de los ingresos anuales de una persona a partir de algunas características como su educación, su edad y el lugar donde vive es una tarea de regresión.

El papel de los algoritmos en aprendizaje automático es muy importante; los algoritmos se definen son un conjunto de instrucciones secuenciales que debe realizar un sistema sobre cómo interactuar, manipular y transformar datos. Un algoritmo puede ser simple como realizar una operación matemática con dos números o tan complejo como reconocer un objeto en una imagen.

Los tipos de algoritmos de aprendizaje automático según (Hurwitz & Kirsch, 2018) son los Bayesianos, agrupación (clustering), arboles de decisión, reducción de la dimensionalidad, basados en instancias, regresión lineal, regularización, basado en reglas, redes neuronales y de aprendizaje profundo.

Entre los algoritmos más comunes para los problemas de aprendizaje supervisado según (Mueller & Guido, 2016) son:

- k-Nearest Neighbors. El algoritmo k-NN se basa en analizar los datos de los vecinos más cercanos para hacer una predicción de un nuevo punto de datos. El principal factor de análisis es el número de vecinos para obtener el mejor rendimiento. Se puede implementar KNN para los problemas de clasificación y para problemas de regresión.
- Los modelos lineales. Son una clase de modelos que se utilizan ampliamente en la práctica y hacen una predicción utilizando una función lineal de las características de entrada. Los algoritmos de modelos lineales para problemas de regresión son Regresión lineal ordinario, Regresión Ridge y regresión LASSO. Y para problemas de clasificación son los algoritmos Regresión Logística y las Maquinas de vectores de soporte lineal (Lineal SVM).

- Clasificadores Naive Bayes. Los clasificadores Naive Bayes son una familia de clasificadores que son bastante similares a los modelos lineales, sin embargo, tienden a ser incluso más rápidos en el entrenamiento, pero su rendimiento generaliza y es ligeramente menor que las clasificaciones lineales.
- Arboles de decisión. Son modelos ampliamente utilizados para tareas de clasificación y regresión. Esencialmente, aprenden una jerarquía de preguntas si / si no, lo que lleva a una decisión. Por lo general, la construcción de un árbol y continuar hasta que todas las hojas sean puras conduce a modelos que son muy complejos.
- Ensamblados de árboles de decisión. Los conjuntos son métodos que combinan varios modelos de aprendizaje automático para crear modelos más potentes. Hay muchos modelos, hay dos modelos de conjunto que han demostrado ser efectivos en una amplia gama de conjuntos de datos para clasificación y regresión, los cuales usan árboles de decisión como Random Forest y árboles de decisión impulsados por gradientes.
- Kernelized Support Vector Machines. A menudo denominadas SVM, son una extensión que permite modelos más complejos que no están definidos simplemente por hiperplanos en el espacio de entrada. Hay máquinas de vectores de soporte para clasificación y regresión.
- Redes neuronales. Una familia de algoritmos conocida como "aprendizaje profundo". Los métodos más básicos son los perceptrones multicapa para clasificación y regresión, que pueden servir como punto de partida para métodos de aprendizaje profundo más complejos. Los perceptrones multicapa (MLP) también se conocen como redes neuronales de retroalimentación o, a veces, simplemente redes neuronales.

El ciclo de aprendizaje automático según (Hurwitz & Kirsch, 2018) es un proceso continuo y los pasos son los siguientes:

- Identificar los datos: Identificar las fuentes de datos relevantes es el primer paso del ciclo.
- Preparación los datos: consiste en realizar las actividades necesarias para asegúrese de que sus datos estén limpios, protegidos y gobernados. La importancia de este paso radica en que si una aplicación de aprendizaje automático aprende basado en datos con errores, la aplicación cometerá errores en las predicciones.
- Selección el algoritmo de aprendizaje automático: puede tener varios algoritmos aplicables a sus datos y en este paso consiste en seleccionar un algoritmo adecuado y que obtenga buen desempeño.
- Entrenar: Se refiere a entrenar un algoritmo con el conjunto de datos para crear el modelo.
- Evaluar: Se refiere a la evaluación de los modelos para elegir el algoritmo de brinda el mejor desempeño.
- Implementar: Consiste a implementar los algoritmos de aprendizaje automático credos.
- Predecir: después de la implementación, se pueden hacer predicciones basadas en datos de ingreso nuevos.
- Evaluar predicciones: Consiste en la evaluación de las predicciones realizadas por el modelo. La información que recopila al analizar la validez de las predicciones se retroalimenta luego en el ciclo de aprendizaje automático para tratar de colaborar a mejorar la precisión.

Después de aplicar los algoritmos de aprendizaje automático, necesitamos medir el rendimiento o desempeño de las predicciones realizadas por el modelo. Contamos con un significativo número de métricas para medir el desempeño, Por lo tanto, para cada problema de aprendizaje automático, necesitamos utiliza métricas adecuadas para la evaluación del rendimiento.

Las métricas más comunes de evaluación del desempeño de un modelo de clasificados de aprendizaje automático más comunes según (Vakili, Ghamsari, & Rezaei, 2020) y (Borja-Robalino, Monleón-Getino, & Rodellar, 2020) son la matriz de confusión, accuracy, precision, recall, f1-score, y ROC-AUC; que se describen a continuación:

- Matriz de confusión. Esta tabla de frecuencias es una de las métricas más intuitivas y descriptivas que se utilizan para encontrar la precisión y corrección de un algoritmo de aprendizaje automático. Su uso principal es en problemas de clasificación, la matriz se muestra en la tabla 1.

  Dónde:

  VP: Verdaderos positivos, es el número correcto de predicciones que la instancia positiva.

  FP: Falsos Positivos, es el número incorrecto de predicciones que la instancia es positiva.

  FN: Falsos Negativos, es el número incorrecto de predicciones que la instancia negativa.

  VN: Verdaderos Negativos, es el número correcto de predicciones que la instancia negativa

  Tabla 1: Matriz de confusión.

	Clases	**Resultado del clasificador**	
		Positivo	**Negativo**
Resultado real	Positivo	VP	FN
	Negativo	FP	VN

  Adaptado (Borja-Robalino, Monleón-Getino, & Rodellar, 2020)

- Accuracy: Es el más utilizado y quizás la primera opción para evaluar el desempeño de un algoritmo en problemas de clasificación. Se definirse como la relación entre elementos de datos clasificados con precisión y el número total de observaciones.

- Precision: muestra "qué número de elementos de datos seleccionados son relevantes". En otras palabras, de las observaciones que un algoritmo ha predicho que serán positivas, ¿cuántas de ellas son realmente positivas? La precisión se calcula dividiendo el número de verdaderos positivos dividido por la suma de verdaderos positivos y falsos positivos:
- Recall o Tasa de Verdaderos Positivos (TVP): Presenta “qué número de elementos de datos relevantes se seleccionan”. De hecho, de las observaciones que son realmente positivas, cuántas de ellas han sido predichas por el algoritmo. La sensibilidad es igual al número de verdaderos positivos dividido por la suma de verdaderos positivos y falsos negativos:
- F1-Score. Esta métrica, tiene en cuenta tanto la exactitud como la sensibilidad para calcular el rendimiento de un algoritmo; es la media armónica del accuracy y la recall.
- Curva ROC y AUC. La curva de características operativas del receptor o curva ROC de forma abreviada, es la curva que muestra la Tasa de Falsos Positivos (TFP) frente a la Tasa de Verdaderos Positivos (TVP). La curva ideal está cerca de l aparte superior izquierda, es decir que produzca un recall alta, mientras mantiene una Tasa de Falsos Positivos bajo. (Mueller & Guido, 2016)

1.3.3. Inteligencia de Amenazas

Gartner (2013) introduce el término “Threat intelligence” o Inteligencia de amenazas, que lo define como el conocimiento basado en evidencia, incluyendo su contexto, mecanismos, indicadores, implicaciones y acciones concretas, sobre la amenaza o peligro existente o emergente a los activos y que pueda ser usada para tomar decisiones informadas y acciones de respuesta por parte del afectado por la amenaza o peligro.

El tipo de información de inteligencia se clasifica en estratégica que se refiere a la inteligencia acerca de los riesgos y consecuencias asociadas a las amenazas que se usa para la toma de decisiones a alto nivel y direccionamiento de la estrategia. Puede ser táctica, que se refiere a la información de carácter técnica que normalmente contiene información específica de una dirección IP, URL, dominio, etc; y puede ser operacional que se refiere a la inteligencia que se enfoca en las técnicas, herramientas, metodologías de los adversarios.

La información de inteligencia de las amenazas y los agentes de amenazas proporcionan una comprensión suficiente para mitigar un evento dañino. Las fuentes de inteligencia pueden ser internas, se acceso libre, comerciales y organizacionales.

Actuar sobre la información de amenazas es el proceso de hacerla procesable localmente y que este conocimiento se distribuye tácitamente e informa directamente las interpretaciones futuras de la información de amenazas, lo que afecta la capacidad de las organizaciones para recibir alertas anticipadas sobre las amenazas, para contener el daño y aumentar la seguridad. Además, indican que la creación y utilización de conocimiento relevante es un proceso en el que los analistas de inteligencia de amenazas confían en gran medida y que puede ser respaldado a través de la automatización y el aumento de procesos impulsados por software. (Ahrend, Jirotka, & Jones, 2016)

La inteligencia de amenazas como la información que utiliza una organización para comprender las amenazas que tiene, se dirigirán o se dirigen actualmente a la organización; y el propósito principal de la inteligencia de amenazas es ayudar a las organizaciones a comprender los riesgos de las amenazas externas más comunes y graves. (Cascavilla, Tamburri, & Heuvel, 2021)

La inteligencia de código libre es el conocimiento obtenido a partir del procesamiento y análisis de fuentes de datos públicas, como transmisiones de televisión y radio, redes sociales y sitios web. Estas fuentes proporcionan datos en formatos de texto, video, imagen y audio. (Cascavilla, Tamburri, & Heuvel, 2021)

La inteligencia de amenazas es la disciplina cuya intención es proporcionar información organizada, analizada y refinada sobre ataques potenciales o actuales que amenazan a una organización, o gobiernos (Tounsi & Rais, 2018).

El ciclo de vida de la inteligencia de amenazas consta de seis fases y son: dirección, recopilación, procesamiento, análisis, diseminación y retroalimentación (Cascavilla, Tamburri, & Heuvel, 2021)

De lo analizado se evidencia que se utilizan modelos de la ciberseguridad, y modelos de machine learning, pero hay insuficientes referentes teóricos y prácticos relacionados con la detección de sitios web phishing, basadas en machine learning, teniendo en cuenta la información de las URL, la información de la inteligencia de amenazas y la lógica de machine learning.

1.3.4. Marco Conceptual.

Amenaza. En informática una amenaza es una posible acción o evento negativo contra un sistema o dispositivo informático, que, aprovechando debilidades o vulnerabilidades del sistema, puede generar una violación a la seguridad y causar un impacto no deseado en el dispositivo y/o en la información importante para una organización o persona. (Baca Urbina, 2016).

Ataque. Un ataque es una acción que aprovecha o explota debilidad de un sistema hardware o software para causar un impacto sobre él e incluso (Escrivá Gascó, Romero Serrano, & Ramada, 2013).

Ciberataque. Para (Larrieu-Let, 2015) un ciberataque es toda acción malintencionada que se realiza con la finalidad de comprometer los pillares de la seguridad como la confidencialidad, integridad o disponibilidad de un equipo, red, o sistema, como un sitio web.

Ciberseguridad. Para (Sánchez, 2011) la ciberseguridad es el conjunto de tecnologías, políticas, técnicas, normas, capacitación, herramientas, salvaguardas de seguridad, directrices, y cualquier buena práctica que permite proteger los activos de información de una organización y los usuarios en un entorno conectado.

Código malicioso. Llamado también malware, se refiere al código que está diseñado para infiltrarse en un sistema sin autorización, o interrumpir el funcionamiento del mismo. Los tipos de código malicioso son virus y gusanos informáticos, troyanos y ransomware.

Código fuente. Se refiere al conjunto de líneas de texto, con la lógica que un software debe seguir para ejecutar paso a paso.

Dirección IP. Una dirección IP es una secuencia numérica decimal de 32 bits organizado en cuatro grupos separado por puntos (IPv4) o una secuencia hexadecimal de 128 bits organizado en ocho grupos separados por dos puntos (IPv6) que identifica de forma única a una tarjeta de red de un dispositivo conectado a la red.

Dirección URL. URL proviene de Uniform Resource Locator (Localizador uniforme de Recursos) y es una dirección de dominio o IP única que se asocia a cada uno de los recursos disponibles en la WWW (World Wide Web). Las direcciones URL permiten que los recursos en internet sean localizados y visitados por los usuarios.

Dominio. Un dominio es un nombre que identifica a una empresa, grupo de empresas, un sistema o un activo dentro de internet. El Sistema de nombres de dominio conocido generalmente como DNS permite la traducción del nombre de dominio a una dirección IP de cada activo en la red.

Información. La información según (Escrivá Gascó, Romero Serrano, & Ramada, 2013), es el conjunto de datos que es interpretable para una empresa o persona, datos importantes que, en manos de ciberdelincuentes, pueden llevar a una empresa o persona hasta a la ruina.

Informática. Según (Real Academia Española, 2017) es el conjunto de conocimientos científicos y técnicas que permiten el procesamiento, almacenamiento y transmisión de la información por medio de dispositivos informáticos y de comunicaciones.

Ingeniería social. Técnica de ciberataque para obtener información confidencial como datos personales, financieros o contraseñas. También se considera ingeniería social a pedir información de autenticación como usuario y clave como un favor a un compañero de trabajo.

Modelo. Un modelo es considerado como una abstracción teórica de la realidad principalmente con la finalidad de reducir la complejidad del objeto en estudio, mostrando las partes importantes de un proceso, obviando o ignorando los detalles y centrándose en lo primordial del proceso o de la realidad.

Phishing. Un tipo de ataque en que el atacante se pone en contacto con la víctima fingiendo ser una empresa legítima con la que la víctima tenga alguna relación como un banco o un operador de telefonía. A través de este mensaje el atacante trata de convencer a la víctima que ingrese a un enlace que le llevará a un sitio web falso donde se le solicita información confidencial como números de tarjetas, usuarios o claves.

Seguridad. Según (Real Academia Española, 2017) el término seguridad, se define como: "cualidad de seguro"; y el término seguro es definido como un adjetivo "libre y exento de riesgo".

1.4. Formulación del Problema.

Insuficiencias en el proceso de la ciberseguridad y la detección de phishing, **limitan el rendimiento en la detección de sitios web falsas.**

1.5. Justificación e importancia del estudio.

Los ciberataques se han convertido en un problema que afecta a todas las personas y a todo tipo de organizaciones que hacen uso de las tecnologías de la información y comunicaciones. Siendo los ataques mediante phishing, los más comunes en la actualidad. Según (Marsh & Microsoft, 2020) las industrias que han percibido incremento en los ciberataques son la financiera, energía, minería y transporte.

Según (Singh & Sharma, 2019) aproximadamente 2 billones de dólares de pérdidas en el 2019 debido a los ciberataques y que va creciendo día a día. Y además según (APWG, 2021) las estafas por phishing son cada vez más costosas para las víctimas. Las solicitudes de transferencia promedio en los ataques BEC aumentó de $ 48,000 a $ 75,000 en el cuarto trimestre del 2020.

Es por eso la importancia del desarrollo de esta investigación que trata de aportar un modelo de Machine Learning para detectar los sitios web falsos, analizando la estructura de las direcciones URL y la información de inteligencia de las amenazas.

Teniendo como **aporte teórico** el modelo de machine learning que permitirá la detección de sitios web falsos, teniendo en cuenta la estructura de las URL y la inteligencia de las amenazas y las técnicas de machine learning.

Y como **aporte práctico** se considera un sistema de detección de phishing con técnicas de machine learning y la inteligencia de las amenazas.

La **novedad científica** será revelar que las técnicas de machine learning con la información de los sitios web con la URL y la información de inteligencia de amenazas pueden ser integradas para el rendimiento en la detección de sitios web phishing.

La **significación práctica**, radica en el impacto social y la relevancia al desarrollar el sistema de detección de phishing para el rendimiento en la detección de los sitios web phishing.

1.6. Hipótesis y operacionalización de las variables.

1.6.1. Hipótesis.

Si se aplica un sistema de detección de phishing, basado en un modelo de machine learning, que tenga en cuenta la información de las características de la URL, la información de la inteligencia de amenazas y las técnicas de machine learning, se contribuye al rendimiento en la detección de sitios web falsos.

1.6.2. Variables

Variable independiente:

Sistema de detección de phishing, basado en un modelo de Machine Learning.

Variable dependiente:

Rendimiento en la detección de sitios web falsos.

El rendimiento de un sistema de detección según (Martínez Puentes, 2011) es el “número de eventos que es capaz de analizar un sistema correctamente”.

Para evaluar un sistema de detección según (De la Hoz, De la Hoz, Ortíz, & Ortega, 2012) es necesario evaluar las métricas como el accuracy, precision, recall, f1-score y principalmente la matriz de confusión con los siguientes indicadores.

Verdaderos positivos (VP). Cantidad de pruebas clasificadas correctamente un sitio web falso, como sitio web falso.

Verdaderos negativos (VN). Cantidad de pruebas clasificadas correctamente un sitio web real, como sitio web real.

Falsos positivos (FP). Cantidad de pruebas clasificadas erróneamente un sitio web real, como sitio web falso.

Falsos Negativos (FN). Cantidad de pruebas clasificadas erróneamente un sitio web falso, como sitio web real.

1.7. Objetivos

1.7.1. Objetivo General

Aplicar un sistema de detección de phishing, sustentada en un modelo de machine learning, para el rendimiento en la detección de sitios web falsos.

1.7.2. Objetivos Específicos

- Caracterizar científicamente el proceso de la ciberseguridad, la detección de phishing y su dinámica.
- Diagnosticar el estado actual del proceso de ciberseguridad y la detección de phishing.
- Elaborar un modelo de machine learning, utilizando las características de la URL, información de la inteligencia de amenazas y técnicas de machine learning.
- Elaborar un sistema de detección de phishing, basado en el modelo de machine learning.
- Validar los resultados de la investigación.

II. MATERIAL Y MÉTODO

2.1 Tipo y Diseño de Investigación.

El tipo de investigación de este trabajo es mixto, porque tiene objetivos cuantitativos y objetivos cualitativos.

El diseño de contrastación de hipótesis es cuasi experimental, dado que se realizarán dos experimentos con los datos, el primer experimento determinará el rendimiento en la detección de phishing según los estudios previos y el segundo experimento se realizará con el sistema de detección propuesto que estará basado en el modelo propuesto, haciendo uso de la información de la inteligencia de amenazas, la información de las características de los sitios web y las técnicas de machine learning.

2.2 Población y muestra.

La población de la investigación se refiere al número de los sitios web que pueden ser clasificadas como sitios legítimos o sitios phishing; el número de sitios web es desconocida por tal motivo la población es infinita.

En los problemas de machine learning hay dos etapas principales que son en entrenamiento (para la construcción del sistema) y la validación (para la evaluación del rendimiento del sistema), por tanto, se usa como muestra 11055 sitios web como se muestra en la tabla 2, de los cuales se utilizan 8844 sitios web para el entrenamiento y 2211 para la evaluación del rendimiento.

Tabla 2: Muestra de sitios web utilizados en la tesis

Sitios Web	Total	Porcentaje	Clases	
			Phishing	Legítimo
Entrenamiento	8844	80%	3935	4909
Validación	2211	20%	963	1248
Total	11055		4898	6157

2.3 Técnicas e instrumentos.

Los métodos de recolección de datos, que se utilizaron en el presente trabajo de investigación son el análisis documentario y la ficha de observación.

Análisis documentario: Consiste en extraer la información de los diferentes documentos, artículos, cuaderno de incidencias, libros, revistas, publicaciones, gráficos, etc. Los cuales presentan una serie de estándares, teorías y recomendaciones, para el uso correcto de los componentes de la arquitectura de seguridad, además se pueden analizar las características del proceso de seguridad informática a través de los mecanismos de seguridad a emplear y por último analizar los métodos para dar solución al problema planteado.

Ficha de Observación. Es el registro visual de lo que ocurre en una situación real; se basa en la experimentación que consiste en la observación de la detección o clasificación de los sitios web como sitios web phishing o como sitios web legítimas. Las observaciones se darán en dos escenarios uno con los sistemas actuales convencionales y otra observación con el sistema de detección basado en el modelo de machi learning propuesto, que luego servirán para comparar y analizar los resultados con la hipótesis; los cuales se pueden observar y documentar.

Entre los **Métodos y Técnicas** se emplean las siguientes:

Para la caracterización del proceso de ciberseguridad y la detección de phishing, así como en la construcción del aporte teórico y aporte práctico se utilizarán:

- **Análisis-síntesis**. En el estudio del proceso de la ciberseguridad y la detección de phishing transitando por toda la lógica del machine learning, información de las URL, y la información de la inteligencia de amenazas.

- **Hipotético-deductivo**: Desde la formulación de la hipótesis hasta la aplicación del sistema de detección de phishing, para la evaluación de los resultados.
- **Histórico–lógico**: En el análisis de la evolución del proceso de ciberseguridad y en la detección de phishing, con el fin de verificar los avances y tendencias en este proceso.
- **Sistémico-estructural-funcional**: Los algoritmos de machine learning permiten aprender bajo el proceso de entrenamiento, pero es necesario dotarle de los datos procesados y limpios, para un aprendizaje adecuado, además permite optimizarlo con diferentes parámetros y técnicas de machine learning y análisis de datos.
- **Holístico Dialéctico:** Este es un proceso integral que vincula los estándares y buenas prácticas internacionales en gestión de la ciberseguridad, con las técnicas y algoritmos de machine learning y usando la inteligencia de amenazas. con el fin de detectar con los sitios web en phishing o sitios web legítimos, con un alto rendimiento en la detección.

Para la corroboración de la factibilidad y el valor científico-metodológico de los resultados de la investigación de la implementación del sistema predictivo de phishing, con técnicas de aprendizaje automático.

- **Estadístico:** Se utilizará principalmente Python como herramientas estadísticas para la determinación de la hipótesis y su contrastación, se hará uso de diferentes librerías que empaqueta fórmulas y métricas, y permite la visualización de los resultados.
- **Guía de Observación:** Se realizarán verificaciones de detección de sitios web phishing y sitios web legítimas, con el objetivo de medir el rendimiento del sistema predictivo; las pruebas se registrarán en las guías de observación y se registrarán de forma automatizada haciendo uso de python.

2.4 Procedimientos de análisis de datos.

La información recolectada a través de los instrumentos será validada, luego codificada, seguidamente se registrará y por último se tabulará para el análisis de los datos, haciendo uso de los programas Python y Microsoft Excel. Los resultados de este análisis serán presentados a través gráficos y tablas estadísticas, así como la interpretación de los datos tomando como base los indicadores y variables que se medirán.

El error en la clasificación errónea se representa como la proporción de instancias clasificadas incorrectamente a todas las instancias.

$$\text{Misclassification Error} = \frac{FP + FN}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Para calcular el accuracy, es la proporción de clasificación correcta de todas las instancias que se utilizan se calcula mediante la fórmula.

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Recall o Tasa de verdaderos positivos (TPR) o sensibilidad se calcula utilizando la proporción de instancias clasificadas correctamente como positivas a todas las instancias positivas, representadas por la fórmula

$$TPR = \frac{TP}{TP + FN}$$

La tasa de falsos positivos (FPR) o la especificidad se da utilizando la proporción de instancias clasificadas incorrectamente como positivas a todas las instancias negativas, y la fórmula es:

$$FPR = \frac{TN}{\text{TN} + \text{FP}}$$

La Precisión es la proporción de instancias clasificadas correctamente como positivas a todas las instancias clasificadas positivamente:

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

2.5 Criterios éticos

Los criterios éticos que se tomaron en cuenta en este trabajo están basados en los principios básicos del Informe Belmont y son:

- **El respeto a las personas.** Citando correctamente las investigaciones y libros que se han utilizado como base para el Desarrollo del aporte teórico y del aporte práctico.
- **Beneficencia.** Los beneficios obtenidos en este trabajo permiten un beneficio para toda la población que hace uso de las tecnologías de la información y que están expuestos a acceder a sitios web falsos.

2.6 Criterios de Rigor científico.

En esta tesis se aplicaron un conjunto de procedimiento para estructurar el aporte teórico y el aporte práctico en base los criterios científicos siguientes:

- **Credibilidad.** Se ha valorado las técnicas de machine learning en un sistema de detección de phishing y cuyos resultados obtenidos se muestra en el modelo de machine learning y en el sistema de detección de phishing.
- **Objetividad.** El diagnóstico del proceso de ciberseguridad y la detección de phishing que se pretende dar a conocer se basará en estudios previos, criterios técnicos e imparciales.
- **Originalidad.** Se citarán las fuentes bibliográficas, a fin de reflejar la inexistencia de plagio intelectual.
- **Veracidad.** La información utilizada en esta investigación es verdadera, y se mantiene la confidencialidad de los datos.
- **Relevancia**. El modelo de machine learning y el sistema de detección de phishing aportes de esta investigación resultan substancialmente importante por la implicancia de prevenir el acceso a sitios web falsos por los usuarios de las tecnologías de información.

III. RESULTADOS

3.1 Resultados en Tablas y Figuras

Se analizaron los resultados del rendimiento en la detección de sitios web falsos, en la Tabla 3 se muestra el resumen con los autores, el algoritmo y el accuracy como métrica del rendimiento.

Tabla 3: Rendimiento en la detección de phishing, estudios previos

Autores	Algoritmo	Accuracy
(Rami , Fadi , & Lee , 2014)	Redes Neuronales	92.48%
(Mohammad, Thabtah, & McCluskey, 2014)	Redes Neuronales	94.07%
(Yi, y otros, 2018)	Redes Neuronales (DBN)	89.20%
(Jain & Gupta, 2018)	SVM	90.00%
(Niakanlahiji, Chu, & Al-Shaer, 2018)	PhishMon	95.40%
(Abutair, Belghith, & Al-Ahmadi, 2018)	CBR-PDS	96.26%
(Patil, Thakkar, Shah, Bhat, & Godse, 2018)	Random Forest	96.58%
(Ali & Ahmed, 2019)	Red Neuronal Profunda	91.13%
(Kulkarni & Brown, 2019)	Arboles de decisión	91.50%
(Ubing , y otros, 2019)	Ensamblado	95.40%
(Wang, Zhang, Luo, & Zhang, 2019)	PDRCNN (Deep learning)	95.79%
(Zabihimayvan & Doran, 2019)	Random Forest	95.00%
(Wei, y otros, 2019)	Red Neuronal Profunda	86.63%
(Christou y otros, 2020)	SVM	90.00%
(Opara, Wei, & Chen, 2020)	Red Neuronal Convolucional	93.00%
(Zamir, y otros, 2020)	Stack: RF, NN y bagging	97.40%
(Chavan, y otros, 2020)	Arboles de decisión	96.82%
(Harinahalli & BoreGowda, 2020)	Random Forest	96.87%
(Aljofey, Jiang, Qu, Huang, & Niyigena, 2020)	Deep learning	95.02%
(Christou, y otros, 2020)	SVM	90.00%
(Anupam & Kar, 2020)	SVM	90.38%
(Lakshmi, Reddy, Santhaiah , & Reddy , 2021)	Deep learning	96.00%
(Yang, y otros, 2021)	NIOSELM	97.30%

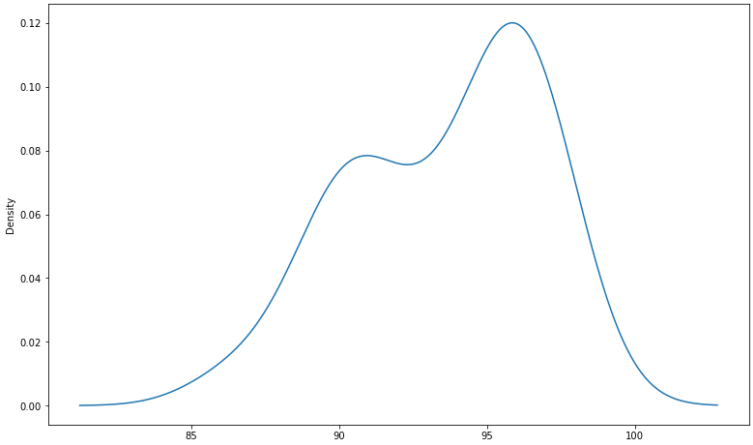
En la Tabla 4, se muestra el rendimiento promedio de 93.58% en la detección de sitios web phishing por estudios previos, y una desviación estándar de 3.17, un mínimo 86.63% y un máximo de 97.61%.

Tabla 4: Resumen del rendimiento de estudios previos

Métrica	Media	Desviación	Mínimo	Máximo
Accuracy	93.58	3.13	86.63	97.40

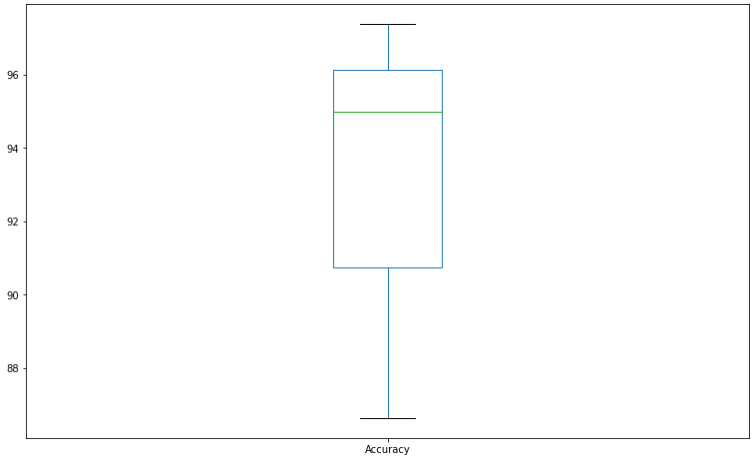
En la Figura 2 se muestra la densidad del rendimiento y se observa un sesgo en la parte derecha.

Figura 2: Densidad del rendimiento de estudios previos.



En la Figura 3 con el diagrama de caja, se observa que no existen valores atípicos o fuera de rango, en los resultados de estudios previos

Figura 3: Diagrama de caja del rendimiento de estudios previos.



Los resultados de los estudios previos muestran que el 50% del rendimiento están entre 90.76% y 96.13%. La mediana es de 95.00% y es mayor que la media (93.58%), por lo que se confirma que los datos son asimétricos y tienen un sesgo a la derecha.

Además de los resultados de rendimiento de los estudios previos, se construyó un Sistema Base, usando las técnicas de machine learning, y la información de las direcciones URL, según estudios previos.

Para la evaluación de los resultados del sistema base se utilizaron en total 2211 sitios web, divididos en 963 (44%) sitios phishing y 1248 (56%) sitios web legítimos, según la muestra determinada.

Tabla 5: Matriz de confusión del sistema base

	Clases	**Resultado del Sistema Base**		
		Phishing	**Legítimo**	**Total**
Sitio web (Clasificación real)	Phishing	884 (VP)	79 (FN)	963
	Legítima	58 (FP)	1190 (VN)	1248

En la tabla 5 se muestra la matriz de confusión y se observa que 2074 (884+1190) sitios web fueron correctamente clasificado por el sistema base y 137 (58 +79) sitios web fueron incorrectamente clasificados por el sistema base. En la Tabla 6, se muestra a detalle los resultados de los indicadores VP, FP, VN y FN.

Tabla 6: Resultados de VP, FP, VN, FN del sistema base

Métrica Evaluada	**Descripción**	**Resultado**
VP: Verdaderos Positivos	Numero de sitios web phishing, correctamente clasificados por el sistema	884
FP: Falsos Positivos	Número de sitios web legítimos, clasificados incorrectamente por el sistema	58
VN: Verdaderos Negativos	Números de sitios web legítimos, clasificados correctamente por el sistema	1190
FN: Falsos Negativos	Números de sitios web legítimos, clasificados incorrectamente por el sistema	79

En la tabla 7 se muestra los resultados de la clasificación del sistema base; donde se muestra que el 93.80% es la proporción de clasificación correcta del sistema en global, el 6.20% es el error en la clasificación errónea del sistema base en global, el 91.80% de sitios web phishing clasificadas correctamente, el 95.35% de sitios web legítimos clasificados correctamente y el 93.84% es la proporción de sitios web phishing clasificados correctamente en relación a la clasificación como sitios web phishing por el sistema.

Tabla 7: Resultados del rendimiento del sistema base

Métrica Evaluada	Fórmula	Resultado
Accuracy	Accuracy = (VP+VN) /(VP+VN+FP+FN)	93.80%
Classification Error	Classification Error = (FP+FN)/(VP+VN+FN+FP)	06.20%
Recall (TVP)	recall = (VP)/(VP+FN)	91.80%
Specificity (TVN)	Specificity = (VN)/(VN+FP)	95.35%
Precision	Precision = VP / (VP + FP)	93.84%

3.2 Discusión de resultados

Los estudios previos muestran que se utilizan técnicas de machine learning, para la detección de sitios web phishing, utilizando la información de las direcciones URL, y de anomalías en el código fuente, analizando solamente al cliente.

El accuracy es la principal métrica en la evaluación del rendimiento, y mide la proporción de clasificación correcta del sistema de todas las instancias que se utilizan en el experimento.

El accuracy obtenido como promedio de los estudios previos, es de 93.58%, y del sistema base construido tiene un accuracy de 93.80%.

Estos resultados serán utilizados como base para contrastar con la propuesta de la tesis.

3.3 Construcción del Aporte teórico

3.3.1 Fundamentación del aporte teórico

El aporte teórico es un modelo de machine learning, para la detección de sitios web falsos, que se fundamenta en las teorías de machine learning, sitios web, detección de phishing y la inteligencia de amenazas.

Para construir la estructura del modelo planteado y las dimensiones, tomamos como base principalmente en **el ciclo de Machine Learning**, que según (Hurwitz & Kirsch, 2018) es un proceso continuo y considera las siguientes etapas:

- Identificar los datos: Identificar las principales fuentes de datos relevantes, para el problema en estudio, es el primer paso del ciclo.
- Preparación los datos: consiste en realizar las actividades necesarias para asegúrese de que sus datos estén limpios, protegidos y gobernados. La importancia de este paso radica en que si una aplicación de aprendizaje automático aprende basado en datos con errores, la aplicación cometerá errores en las predicciones.
- Algoritmos de Machine Learning. puede tener varios algoritmos aplicables a sus datos y en este paso consiste en seleccionar un algoritmo adecuado y que obtenga buen desempeño.
- Entrenar: Se refiere a entrenar un algoritmo o varios algoritmos de machine learning con el conjunto de datos para crear el modelo.
- Evaluar: Se refiere a la evaluación de los modelos para elegir el algoritmo de brinda el mejor desempeño.
- Implementar: Consiste a implementar los algoritmos de aprendizaje automático credos.
- Predecir: después de la implementación, se pueden hacer predicciones basadas en datos de ingreso nuevos.

- Evaluar predicciones: Consiste en la evaluación de las predicciones realizadas por el modelo. La información que recopila al analizar la validez de las predicciones se retroalimenta luego en el ciclo de aprendizaje automático para tratar de colaborar a mejorar la precisión.

Machine Learning son técnicas computacionales que utilizan la experiencia para mejorar el rendimiento o lograr predicciones precisas. La experiencia se refiere a la información previa disponible (normalmente conocido como conjunto de datos) para sirve como entrada para el proceso de entrenamiento. (Subasi, 2020)

La clasificación de las técnicas de Machine Learning, es aprendizaje supervisado, aprendizaje no supervisado, aprendizaje semi supervisado y aprendizaje por reforzamiento (Mohammed, Khan, & Mohammed, 2017) y (Sarker, 2021). Esta tesis se orienta solucionar un problema de **aprendizaje supervisado**, porque se trabajarán con datos que a priori se conocen si son sitios web phishing y sitios web legítimos, es decir se cuenta con las etiquetas.

(Mohammed, Khan, & Mohammed, 2017) precisan que el aprendizaje supervisado, se cataloga dentro de dos grupos que son la clasificación y la regresión, mientras que (Sarker, 2021) clasifica a las tareas y algoritmos de Machine Learning en **clasificación**, regresión, agrupamiento, reducción de dimensionalidad y selección de características, aprendizaje de reglas de asociación, y aprendizaje por reforzamiento.

(Sarker, 2021) cataloga a los problemas de clasificación en binaria y multiclase. La **clasificación binaria** que se refiere a las tareas de clasificación cuando se tiene dos etiquetas de clase, como “verdadero y falso” o como “si y no”. En estas tareas una clase puede ser el estado normal, mientras que la otra clase podría ser una anomalía, en la presente tesis se considerará a un sitio web como legítimo o como un sitio web phishing.

Subasi (2020) presenta un marco de trabajo, de machine learning con aprendizaje supervisado, donde resaltan dos procesos cruciales en el marco que son el entrenamiento del modelo y la predicción del modelo. El marco además cuenta con la etapa de procesamiento y análisis de los datos, y algunos procesos complementarios que como el escalado, la extracción y la selección de características, que deben permanecer constantes en la forma en que se utilizan las mismas características para entrenar el modelo y las mismas características se extraen de muestras de datos de prueba no vistas para probar el modelo en la fase de predicción.

Sarkar, Bali, & Sharma (2018) describen el modelo CRISP-DM (Cross Industry Standard Process for Data Mining) que integra el uso de las técnicas de machine learning con la ciencia de datos y describen como procesos fundamentales a: identificación de los datos, preparación de los datos, Modelado, Evaluación e implementación.

Varios autores como (Sarker, 2021), (Sandoval, 2018) muestran una estructura modelos predictivos basados en machine learning, con dos fases principales que son: la fase de **entrenamiento** y la fase de predicción.

La **fase de entrenamiento** permite construir un modelo con datos históricos y utilizando **algoritmos de machine learning**. La fase de evaluación se analiza el rendimiento del modelo construido, con datos nuevos que el modelo no conoce.

Los tipos de **algoritmos de machine learning** para problemas de aprendizaje supervisado, en problemas de clasificación según (Mohammed, Khan, & Mohammed, 2017), (Sarker, 2021) y (Mueller & Guido, 2016) son:

- Naive Bayes (NB). Los clasificadores Naive Bayes son una familia de clasificadores que son bastante similares a los modelos lineales, sin embargo, tienden a ser incluso más rápidos en el entrenamiento, pero su rendimiento generaliza y es ligeramente menor que las clasificaciones lineales.

- Análisis discriminante lineal (LDA). Son una clase de modelos que se utilizan ampliamente en la práctica y hacen una predicción utilizando una función lineal de las características de entrada. Para problemas de clasificación son los algoritmos Regresión Logística y las Maquinas de vectores de soporte lineal (Lineal SVM).
- Regresión Logística. (RL). Es un método estadístico que se puede aplicar a problemas de clasificación
- Clasificadores de KN Vecinos (KNN). Se basa en analizar los datos de los vecinos más cercanos para hacer una predicción de un nuevo punto de datos. El principal factor de análisis es el numero de vecinos para obtener el mejor rendimiento.
- Máquinas de vectores soporte (SVM). son una extensión que permite modelos más complejos que no están definidos simplemente por hiperplanos en el espacio de entrada.
- Arboles de decisión. (DT). Son modelos ampliamente utilizados para tareas de clasificación y regresión. Esencialmente, aprenden una jerarquía de preguntas si / si no, lo que lleva a una decisión. Por lo general, la construcción de un árbol y continuar hasta que todas las hojas sean puras conduce a modelos que son muy complejos.
- Redes Neuronales (NN). Una familia de algoritmos conocida como "aprendizaje profundo". Los métodos más básicos son los perceptrones multicapa para clasificación y regresión, que pueden servir como punto de partida para métodos de aprendizaje profundo más complejos.
- Ensamblados: Los conjuntos son métodos que combinan varios modelos de aprendizaje automático para crear modelos más potentes. Hay muchos modelos, hay dos modelos de conjunto que han demostrado ser efectivos en una amplia gama de conjuntos de datos para clasificación y regresión, los cuales usan árboles de decisión como Random Forest y árboles de decisión impulsados por gradientes.

En el modelamiento de machine learning, para la detección de sitios web phishing, corresponde al **proceso de la ciberseguridad** y que según (NIST, 2018) dentro del elemento principal del marco de trabajo se cuenta con cinco funciones principales y son:

- Identificar: la NIST lo define como “desarrollar una comprensión organizacional para administrar el riesgo de seguridad cibernética para sistemas, personas, activos, datos y capacidades”
- Proteger. La NIST lo define como “desarrollar e implementar medidas de seguridad adecuadas para garantizar la entrega de servicios críticos”
- Detectar. La NIST lo define como “desarrollar e implementar actividades apropiadas para identificar la ocurrencia de un evento de seguridad cibernética”
- Responder. La NIST lo define como “desarrollar e implementar actividades apropiadas para tomar medidas con respecto a un incidente detectado de seguridad cibernética
- Recuperar. La NIST lo define como “desarrollar e implementar actividades apropiadas para mantener los planes de resiliencia y restablecer cualquier capacidad o servicio que se haya visto afectado debido a un incidente de seguridad cibernética”

Esta tesis se centra en la detección de sitios web phishing, ubicado en la función de detectar según el marco de trabajo ciberseguridad de NIST, y lo que se lo que se va a tener que analizar son los sitios web por lo que dos aspectos importantes para este trabajo es la detección y los sitios web.

La Función **Detectar** según (NIST, 2018) “permite el descubrimiento oportuno de eventos de seguridad cibernética.” Para lograr una adecuada detección es importante un monitoreo continuo de la seguridad y utilizar sistemas automatizados con alto rendimiento en la clasificación de eventos de seguridad.

Este trabajo se centra en la detección de sitios web phishing, que consiste en clasificar los **sitios web** en sitios phishing o sitios legítimos, según (Mohammad, Thabtah, & McCluskey, Phishing Websites Features, 2015) plantean que los sitios web phishing se pueden identificar según las características de la barra de direcciones y las características anormales en código HTML y JavaScript.

Otros autores (Jain & Gupta, 2018), (Niakanlahiji, Chu, & Al-Shaer, 2018), (Kumar, y otros, 2019), (Ubing , Binti Jasmi, Azween , Jhanjhi, & Supramaniam, 2019) y (Wang, Zhang, Luo, & Zhang, 2019), proponen sistemas y modelos para detectar phishing que se basan en la información proporcionada por las características de la URL, como la longitud de la URL, uso de direcciones IP, uso de HTTPS, etc.

(Patil, Thakkar, Shah, Bhat, & Godse, 2018) utilizan tres enfoques para la detección de sitios web phishing, el primero es analizando las características de la URL del sitio web, el segundo es verificando la legitimidad del sitio web y el tercero basado en la apariencia visual del sitio web.

Además, tenemos en la **inteligencia de las amenazas,** que según (Gartner, 2013) lo define como “conocimiento basado en evidencia, incluyendo su contexto, mecanismos, indicadores, implicaciones y acciones concretas, sobre la amenaza o peligro existente o emergente a los activos y que pueda ser usada para tomar decisiones informadas y acciones de respuesta por parte del afectado por la amenaza o peligro”

El tipo de información de inteligencia puede de carácter técnica que normalmente contiene información específica de una dirección IP, URL, dominio, etc; y puede ser operacional que se refiere a la inteligencia que se enfoca en las técnicas, herramientas, metodologías de los adversarios.

La información de inteligencia de las amenazas y los agentes de amenazas proporcionan una comprensión suficiente para mitigar un evento dañino. Las fuentes de inteligencia pueden ser internas, se acceso libre, comerciales y organizacionales.

La información de inteligencia de **acceso libre** es el conocimiento obtenido a partir del procesamiento y análisis de fuentes de datos públicas, como transmisiones de televisión y radio, redes sociales y sitios web (Cascavilla, Tamburri, & Heuvel, 2021). Estas fuentes proporcionan datos en formatos de texto, video, imagen y audio.

El ciclo de vida de la inteligencia de amenazas consta de seis fases y son: dirección, recopilación, procesamiento, análisis, diseminación y retroalimentación (Cascavilla, Tamburri, & Heuvel, 2021)

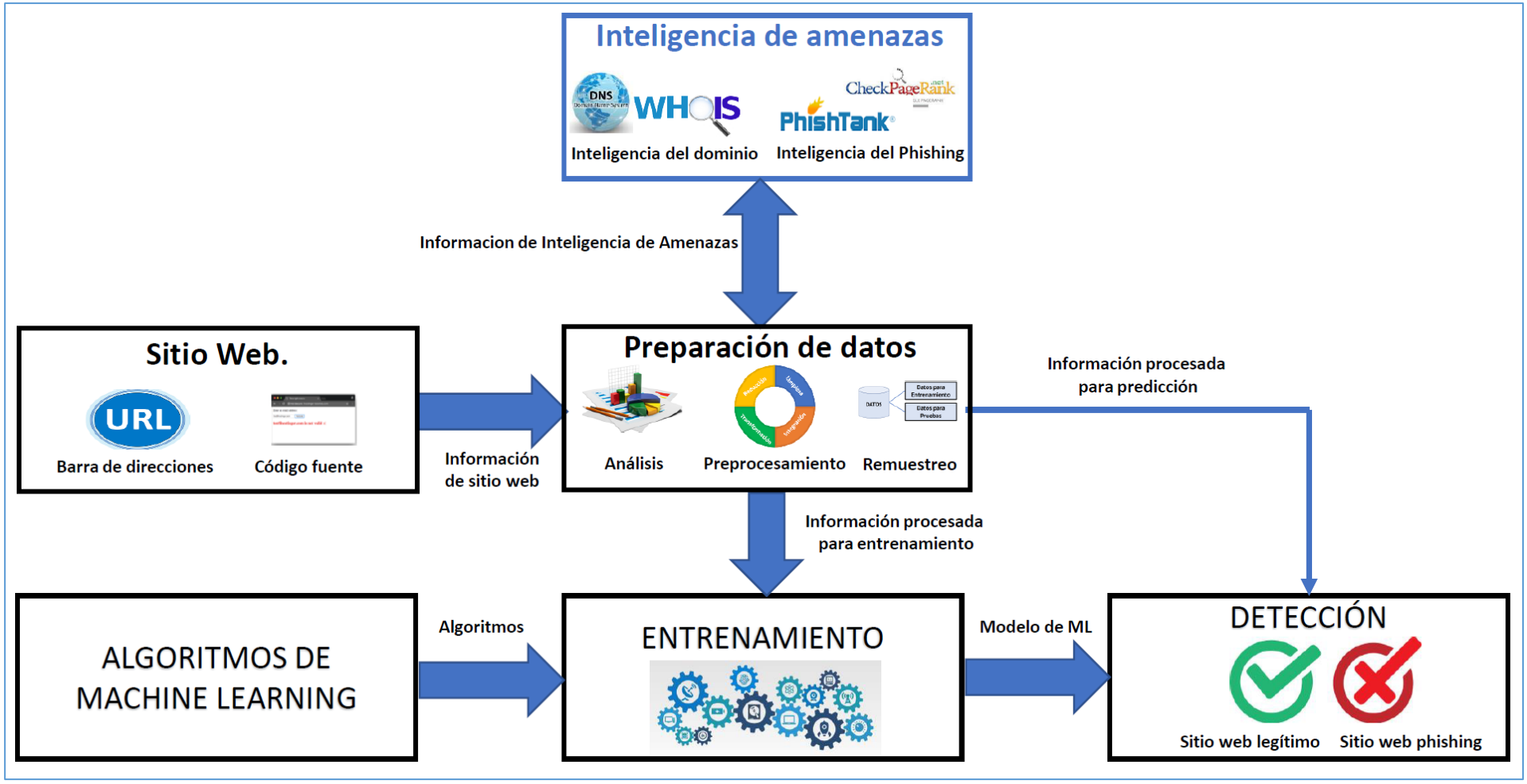
Este trabajo está basado en las teorías de machine learning, sus técnicas, algoritmos y ciclo de vida, y está integrado el ciclo de vida de la inteligencia de amenazas, recogiendo información de los sitios web, y terminando en la detección de sitios web falsos.

3.3.2 Descripción argumentativa del aporte teórico

En el presenta trabajo de investigación se desarrolló un modelo de machine learning en la detección de sitios web phishing, tomando en cuenta las técnicas de machine learning, la información de los sitios web y la información de la inteligencia de amenazas, para la detección de sitios web falsos; sabiendo que lograr la detección adecuada de los sitios web phishing implica lograr un alto rendimiento.

Se propone un modelo tomando en cuenta 6 dimensiones: Sitio web, Inteligencia de amenazas, Preparación de datos, Algoritmos de machine learning, Entrenamiento y Detección, estas dimensiones visto holísticamente que permite la integración de todas las dimensiones, como se observa en la figura 4.

Figura 4: Modelo de Machine Learning en la detección de sitios web phishing



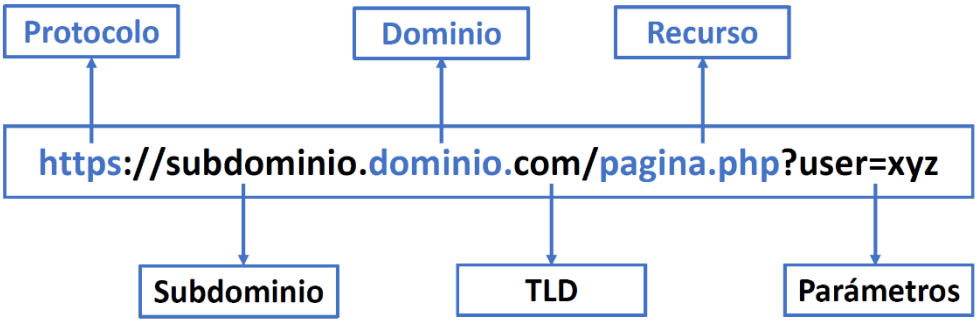
La **dimensión de sitio web**. Dado que el trabajo está orientado en la detección de sitios web phishing, se tiene que analizar los sitios web y la información que proporcionan, por ello es la primera dimensión del modelo propuesto. Un sitio web se refiere al conjunto de archivos que se muestran en forma de una página web que están alojados en un dominio de internet, y que los usuarios de internet acceden a través de una URL. Además, los sitios web están escritos en un lenguaje de programación en código HTML o dinámicamente son convertidos a HTML.

La dimensión del sitio web, en este trabajo permitirá obtener información de los sitios web de acuerdo con sus características desde dos puntos de vista que son: Estructura de la URL, Informacion de la barra de direcciones y Código del sitio web.

Figura 5: Dimensión Sitio Web

- **Barra de direcciones URL.** La estructura de la URL brinda información del sitio web, y que está basado en la información mostrada en la barra de direcciones. La estructura de una URL se organiza con el protocolo, subdominio, dominio, TLD, recurso (archivo web) y los parámetros; como se muestra en la figura 6.

Figura 6: Estructura de una URL



- **Código fuente.** Los sitios web están desarrollados en un lenguaje de programación y desde el cliente se puede acceder al código fuente, el código fuente brinda información de que tipo de datos se está recogiendo desde la página.

La **dimensión Inteligencia de Amenazas** se considera que brinda información importante y actualizada de las amenazas avanzadas para identificar posibles sitios web phishing. En esta dimensión se trabaja con información de inteligencia de fuentes abiertas (OSINT: Open Source Intelligence) y se recolectará información de inteligencia de los registros de dominio y de los registros de phishing disponible de forma pública para ser utilizados en un contexto de inteligencia.

Figura 7: Dimensión Inteligencia de amenazas

- **Los Registros de dominio.** Están referidos a la información de del nombre del dominio en los registros de servicios whois, y servicios de reputación por dominio.
- **Los registros phishing,** Son los registros con información de inteligencia relacionado al phishing como reportes de listas negras, reportes de tráfico, reportes y estadísticas de phishing.

La Inteligencia de amenazas, está referida a recopilar la información de los sitios web en análisis, pero además al procesamiento y organización adecuada de los datos, es por eso que la dimensión de inteligencia se relaciona directamente con la dimensión de Preparación de los datos.

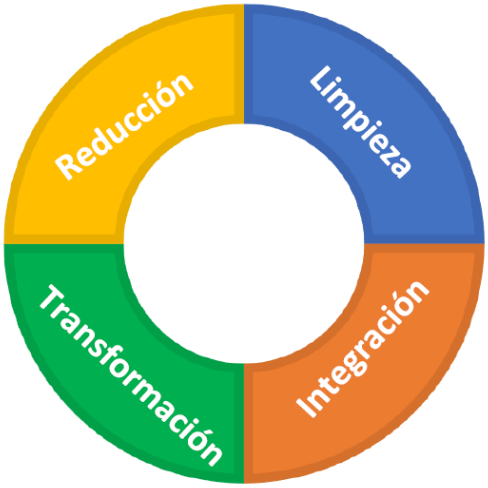
La dimensión de **Preparación de datos**. Brinda los datos de calidad para la aplicación efectiva de las técnicas de machine learning. Es el conjunto de actividades para asegúrese comprender, limpiar, transformar y controlar los datos. Las técnicas de machine learning dependen de la calidad de los datos, para un aprendizaje correcto y sin errores. Esta dimensión es una etapa previa al uso de las técnicas de machine learning, recoge los datos de las fuentes de información para realizar el análisis de los datos, el preprocesamiento, proporcionará los datos para el entrenamiento y las predicciones a través del remuestreo.

Figura 8: Dimensión Preparación de datos

- **Análisis de datos.** Es comprender los datos, realizando un análisis de los tipos de datos, una estadística descriptiva y visualización de los datos.
    - Tipos de datos. Permite describir los tipos de datos recolectados de las fuentes de información, tanto en las características y en las etiquetas de las clases.
    - Estadística descriptiva de datos. Permite entender los datos, analizar los atributos y las instancias de los datos, mediante algunas técnicas descriptivas.
    - Visualización de datos. Permite analizar de forma gráfica los datos y el comportamiento de los mismos.

- **Preprocesamiento de datos.** Consiste en dotarle calidad a los datos, a través de la limpieza, integración, transformación y reducción de los datos.

Figura 9: Preprocesamiento de datos



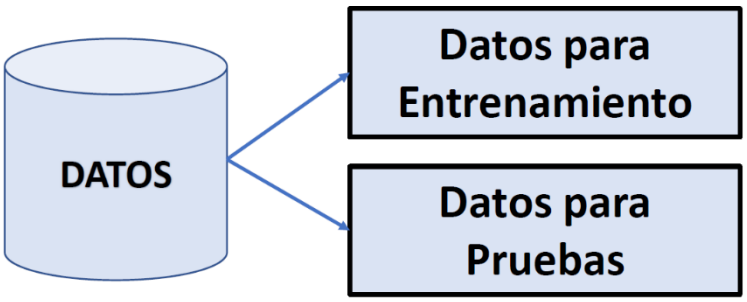
- Limpieza de datos. Verifica y corrige datos incompletos, datos fuera de rango o atípicos y la inconsistencia de los datos recolectados.
- Integración de los datos. Los datos son obtenidos de diferentes fuentes de información, por tanto, es importante integrar los mismos, creando un solo conjunto de datos homogéneo. Se resuelven los problemas de representación y codificación de los datos.
- Transformación de los datos. Consiste en operaciones de preprocesamiento adicionales que consolidan los datos para el entrenamiento más eficiente, aplicando técnicas de transformación de datos como el escalamiento, estandarización, normalización y binarización.
- Reducción de las características. Consiste en aplicar técnicas de reducción de datos o características para obtener un conjunto de datos reducido en volumen, pero con que mantenga la integridad de los datos, siendo más eficiente con respecto a los tiempos de entrenamiento y logrando obtener los mismos o similares resultados. Las estrategias de reducción de datos incluyen la reducción de la dimensionalidad, reducción de numerosidad y comprensión de datos.

- **Remuestreo de los datos.** Las técnicas de remuestreo permitirá dividir el conjunto de datos en subconjuntos para entrenamiento y un subconjunto para la predicción del modelo, además de evaluar la robustez del modelo utilizando división por porcentaje y utilizando validación cruzada.

  La división por porcentaje implica dividir los datos en un conjunto específico para el proceso de entrenamiento del modelo y otro subconjunto de datos para el proceso de evaluación del modelo. El tamaño de cada subconjunto depende del tamaño y las características del conjunto de datos en estudio y puede ser de 80% para entrenamiento y 20% para evaluación del modelo.

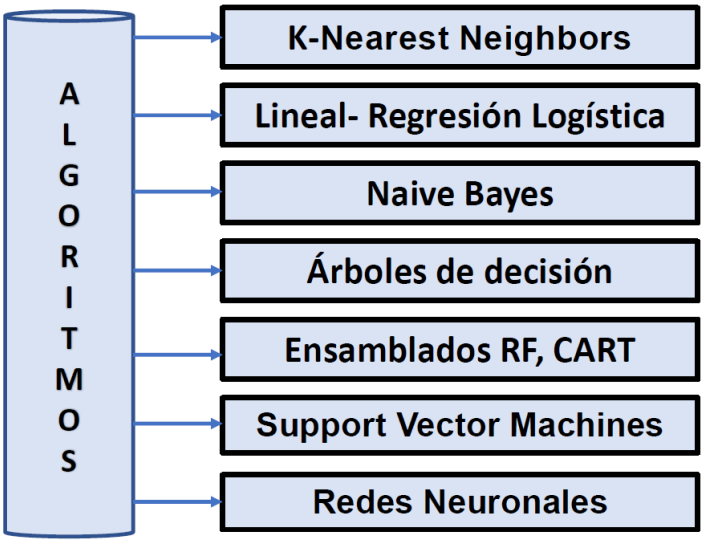
  La validación cruzada implica dividir el conjunto de datos en n-particiones llamados k-folds, cada conjunto se mantiene mientras el modelo se entrena en todas las demás particiones; este proceso se repite hasta que se determina el rendimiento de cada instancia en el conjunto de datos y se estima un promedio del rendimiento global del modelo.

Figura 10: Remuestreo de datos



La dimensión de **Algoritmos de Machine Learning**. Consiste en la selección de los algoritmos de machine learning para realizar el entrenamiento del modelo, se realiza una evaluación preliminar de los diferentes algoritmos, seleccionando los que muestren los mejores resultados Para la evaluación de los algoritmos se debe utilizar las métricas de clasificación como el accuracy, matriz de confusión y el informe de clasificación.

Figura 11: Dimensión Algoritmos de Machine Learning



Los tipos de algoritmos de machine learning se dividen en:

- **k-Nearest Neighbors.** El algoritmo k-NN se basa en analizar los datos de los vecinos más cercanos para hacer una predicción de un nuevo punto de datos.
- **Los modelos lineales.** Son una clase de modelos que se utilizan una función lineal de las características de entrada. Los algoritmos de modelos lineales son los algoritmos Regresión Logística y las Maquinas de vectores de soporte lineal (Lineal SVM).
- **Naive Bayes.** Son una familia de algoritmos en probabilidades y el teorema de Bayes.
- **Arboles de decisión.** Esencialmente, aprenden una jerarquía de preguntas si / si no, lo que lleva a una decisión.
- **Ensamblados.** Son métodos que combinan varios modelos de aprendizaje automático para crear modelos más potentes, hay dos modelos de conjunto que han demostrado ser efectivos y son Random Forest y árboles de decisión impulsados por gradientes.
- **Support Vector Machines.** El algoritmo SVM construye un hiperplano en un espacio de dimensionalidad, para lograr una separación entre una clase y otra.
- **Redes neuronales.** Son un modelo de aprendizaje profundo que consiste en un conjunto de neuronas artificiales conectadas entre si para lograr la clasificación correcta.

La dimensión de **Entrenamiento**. Consiste en la construcción del modelo de machine learning, buscando los mejores parámetros y haciendo uso del subconjunto de datos de entrenamiento.

Figura 12: Dimensión Entrenamiento

Se desarrollan los siguientes procesos:

- Entrenamiento base: Se refiere a entrenar un algoritmo o algoritmos seleccionados con el conjunto de datos para crear los modelos base, y seleccionar el (o los) de mejores resultados.
- Optimizar. Se refiere a la búsqueda de los mejores parámetros de cada algoritmo seleccionado que permita tener los mejores resultados. Se puede realizar mediante la búsqueda de cuadrículas o mediante búsquedas aleatorias.
- Entrenamiento Final, Se refiere a desarrollar el modelo final, con los mejores parámetros encontrados.

La dimensión de **Detección** es la interfaz gráfica de modelo que perimirá clasificar si un sitio web específico es phishing o es un sitio web legítimo. Además, se evalúa el rendimiento del modelo utilizando las métricas de clasificación, como en accuracy, matriz de confusión y el informe de clasificación.

Figura 13: Dimensión Detección

3.4 Aporte práctico

El aporte práctico de esta tesis consiste en el Sistema de Detección de Phishing, que contribuya al rendimiento de la detección de sitios web falsos.

3.4.1 Fundamentación del Sistema de detección de phishing.

El **sistema de detección de phishing está fundamentado** en el aporte teórico de este trabajo que es el **modelo de machine learning**, basado en la información de las URL, la información de la inteligencia de amenazas y las técnicas de machine learning.

El modelo de machine learning consta de seis dimensiones donde la dimensión de Sitios Web y la dimensión de Inteligencia de Amenazas, recolectan información tanto del sitio web y la información de inteligencia de amenazas avanzadas. Los datos recopilados de estas dos dimensiones pasan a la dimensión de Preparación de datos, para realizar el análisis, preprocesamiento y el remuestreo de los datos, donde se generan dos subconjuntos de datos, uno para el entrenamiento que se realiza en la dimensión de Entrenamiento y el otro conjunto de datos para la detección de los sitios web phishing que se realiza en la dimensión de Detección, la dimensión de entrenamiento es alimentado por la dimensión de Preparación de los datos y la dimensión de Algoritmos de machine learning. La dimensión de Detección es la salida del modelo propuesto brindando la clasificación de un sitio web en sitio phishing o en un sitio legítimo.

3.4.2 Objetivo del Sistema de detección de phishing.

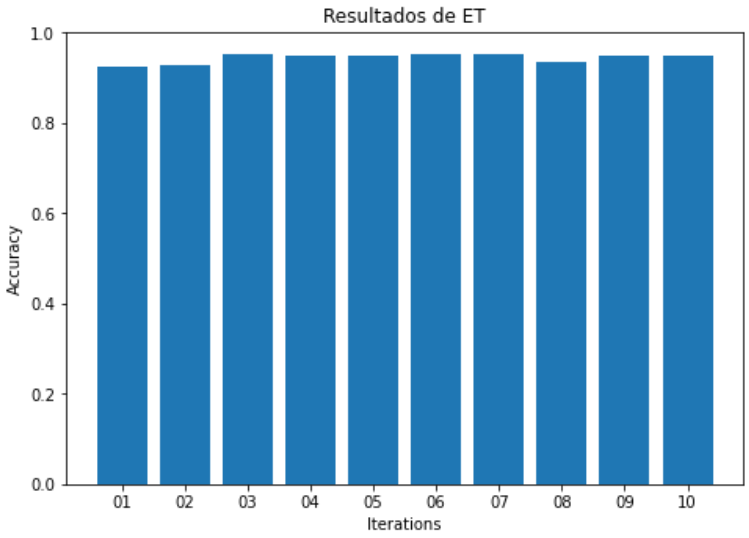
El Objetivo general del sistema de detección de phishing es Sistematizar el proceso de la ciberseguridad y la detección de phishing, mediante el diagnóstico contextual, la fundamentación teórica, su desarrollo y evaluación del rendimiento en la detección de sitios web falsos.

3.4.3 Diagnóstico contextual

Se construyó un Sistema Base en la Phishing basado en estudios previos, utilizando la información de las direcciones URL, sin considerar la Inteligencia de Amenazas.

En la fase de entrenamiento los mejores resultados se obtuvieron con el algoritmo Extra Tree (ET), se utilizó una validación cruzada con 10 iteraciones para su evaluación; el promedio del accuracy es fe 94.46%, con una desviación estándar de 1.03; y los resultados por iteración se muestra en la figura 14.

Figura 14: Resultados de entrenamiento del Sistema Base.

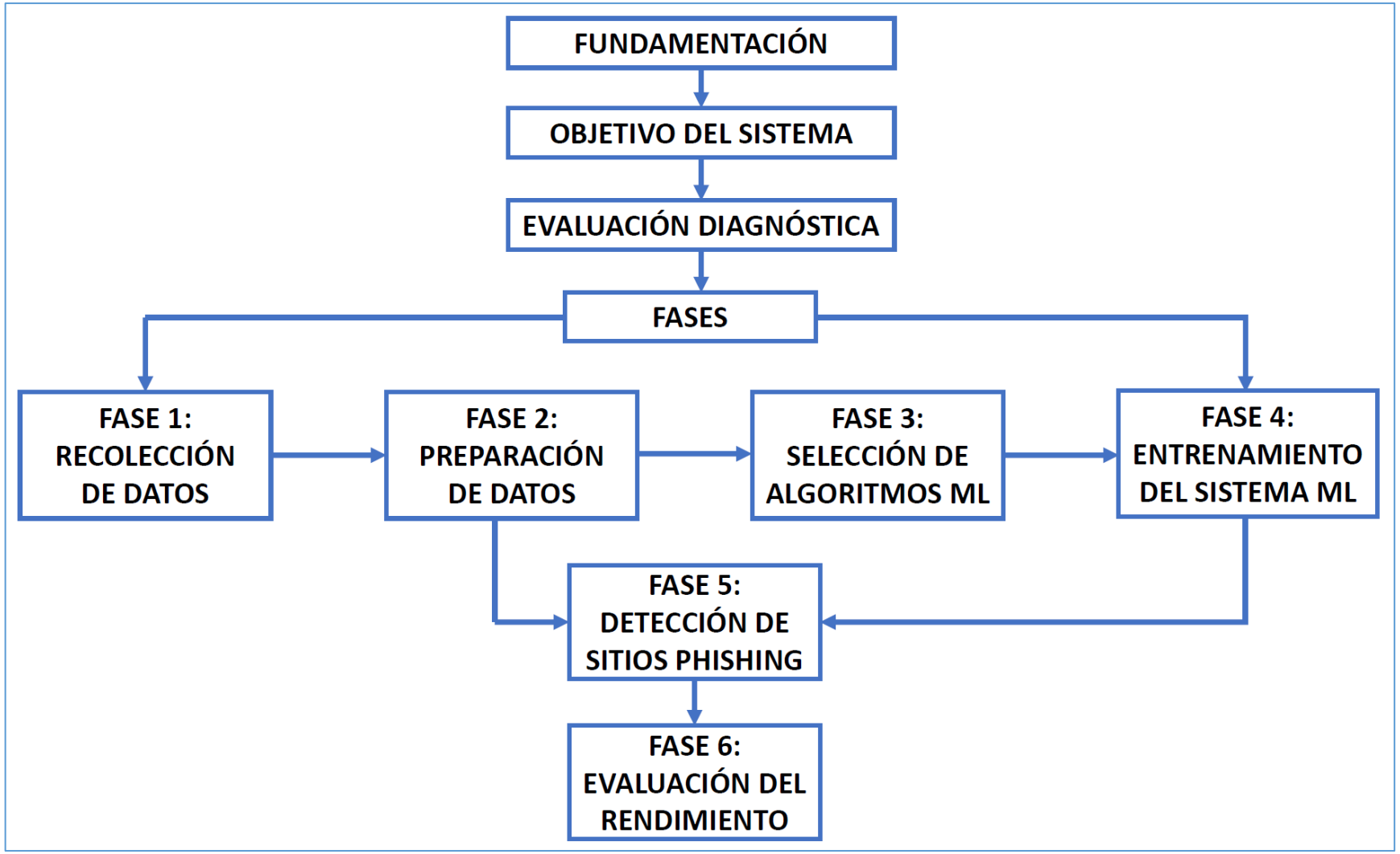


En la fase de clasificación o detección el sistema base, obtiene un accuracy de 93.80%, y todas las métricas se encuentran detalladas en las Tablas 5, 6 y 7, del punto 3.1 resultados en tablas y figuras.

3.4.4 Fases del sistema de detección de phishing

El aporte práctico está sustentado en el modelo de machine learning, integrando las características de los sitios web y la inteligencia de amenazas, y la estructura se basa en el modelo sistémico dividiendo al sistema en subsistemas o fases, en la figura 15 se muestra el Sistema de Detección de Phishing y las fases del mismo que se describen a continuación:

Figura 15: Sistema de Detección de Phishing



Fase 1. Recolección de Datos.

El objetivo de esta primera fase es la obtención de los datos para entrenamiento y para la detección.

Los datos para el entrenamiento deben tener las características del sitio web y la etiqueta si corresponde a un sitio phishing o a un sitio legítimo.

Las características de los sitios web son recolectados desde las dos dimensiones del modelo propuesto: Sitios Web (barra de direcciones y en código fuente) y de la Inteligencia de Amenazas. Las características serán catalogadas como confiable, sospechoso y riesgoso.

Características de los sitios web, basados en la barra de direcciones.

1. **IP_Address.** Se refiere al uso de una dirección IP dentro de la URL. Si no se muestra la IP en la URL es un sitio confiable, de lo contrario es un sitio no confiable.
2. **URL_longitud.** Es la longitud de la dirección URL, usualmente los atacantes usan direcciones extensas para ocultar la parte fraudulenta en la barra de direcciones. Si el tamaño es corto (menor a 54) es un sitio confiable, si el tamaño es extenso (mayor a 75) es un sitio no confiable, de lo contrario es un sitio sospechoso.
3. **URL_corto**. Es el uso del servicio de acortamiento de la URL, se usa normalmente para el redireccionamiento de un sitio web mostrando un nombre corto del sitio, pero enlace a un sitio con nombre extenso. Si no se usa el servicio de acortamiento es un sitio confiable, de lo contrario no lo es.
4. **URL_arroba.** Es el uso del símbolo “@” en la URL. Se usa normalmente para llevar al navegador a ignorar todo lo que está antes del símbolo “@” y la dirección real es la que está después del símbolo “@”. Si no se encuentra el @ en la URL es un sitio confiable, de lo contrario no lo es.

5. **URL_slash.** Es el uso del “//” dentro de la URL, se usa normalmente para redirigir a otro sitio web. Los dominios inician con http:// o https://, por lo tanto si no existe “//” después del séptimo carácter de la URL es un sitio confiable, de lo contrario no lo es.
6. **URL_linea.** El símbolo “-“ rara vez se usa en URL legítimas. No es confiable el uso de prefijos o sufijos separados por (-) al nombre de dominio. Si no se muestra una “-“ en la URL entonces es un sitio confiable, de lo contrario no lo es.
7. **URL_puntos.** El número de puntos dentro de la URL, por ejemplo, con la URL https://www.uss.edu.pe/uss/. Un nombre de dominio incluye los dominios de nivel superior de código de país (ccTLD), que en nuestro ejemplo es "pe". La parte "edu" es una abreviatura de "educativa", el "edu.pe" combinado se denomina dominio de segundo nivel (SLD) y "uss" es el nombre real del dominio. Si se cuenta con 3 o menos puntos en la URL es un sitio confiable, si cuenta con 4 puntos el sitio web es sospechoso y si cuenta con mas de 4 puntos en sitio web es no confiable.
8. **HTTPS_SSL**. El uso de un certificado SSL es muy importante para verificar la legitimidad de un sitio web. Existente autoridades de certificación, entre los mas importantes se incluyen: GeoTrust, GoDaddy, Network Solutions, Thawte, Comodo, Doster y VeriSign. Si el sitio web tiene un certificado SSL entonces es confiable, de lo contrario no lo es.
9. **Domain_registro.** El tiempo de registro de un dominio es importante, los dominios legítimos normalmente tienen varios años de registro. Si el tiempo de registro del dominio es mayor a 1 año es un sitio confiable, de lo contrario no lo es.
10. **Favicon.** Es una imagen gráfica (icono) asociada a una página web específica. Muchos agentes de usuario existentes, como navegadores gráficos y lectores de noticias, muestran favicon como un recordatorio visual de la identidad del sitio web en la barra de direcciones. Si el favicon se carga desde el mismo dominio web entones es confiable, de lo contrario no lo es.

11. **Puerto.** Se verifica los puertos abiertos en el dominio en específico, si los puertos comunes están abiertos los atacantes pueden ejecutar cualquier servicio que deseen, y como resultado la información del usuario se ve amenazada. Si los puertos comunes no están abiertos es confiable, de lo contrario no lo es.
12. **Domain_https**. Se utiliza el https como parte del dominio para aparentar una página web confiable. Si no se usa https como parte del dominio es un sitio confiable, de lo contrario no lo es.

Características de los sitios web, basados en el código fuente.

13. **Request_URL.** Examina si objetos externos como imágenes o videos contenidos en una página web, cargan desde un dominio diferente. Normalmente las páginas web legítimas cargan del mismo dominio la página web y la mayoría de los objetos incrsutados como imágenes y videos. Si las solicitudes son pocas (<22%) el sitio es confiables; si las solicitudes son muchas (>61%) el sitio no es confiable, de lo contrario es sospechoso.
14. **URL_ancla.** Un ancla es un elemento definido por la etiqueta <a>. Se examinan 4 tipos de anclas. <a href="#">, <a href="#content">, <a href="#skip"> y <a href="JavaScript ::void(0)">. Si hay pocas anclas (<31%) el sitio es confiable, si hay muchas anclas (<67%) el sitio no es confiable, de lo contrario el sitio es sospechoso.
15. **Tags.** Es común que los sitios web legítimos usen tags (etiquetas) <Meta> para ofrecer metadatos sobre el documento HTML; Etiquetas <Script> para crear un script del lado del cliente; y etiquetas <Link> para recuperar otros recursos web. Si hay pocas etiquetas (<17%) el sitio es confiable, si las etiquetas son muchas (>81%) el sitio no es confiable, de lo contrario el sitio es sospechoso.
16. **SFH**. Los SFH (Server Form Handler) que contienen una cadena vacía o "about: blank" se consideran no confiables. Además, si el nombre de dominio en SFH no coincide del nombre de dominio de la página es sospechosa, de lo contrario es sitio web confiable.

17. **Submit_email.** El formulario web permite a un usuario enviar su información personal que se dirige a un servidor para su procesamiento. Con ese fin, se puede utilizar un lenguaje de script del lado del servidor como la función "mail ()" en PHP. Otra función del lado del cliente que podría usarse para este propósito es la función “mailto:” Si se encuentra un mail() o mailto el sitio es mo confinable, de lo contrario es confinable.
18. **Abnormal_URL**. Si el hostname está incluido en el dominio entonces es confinable, de lo contrario es un sitio no confaible.
19. **Forwarding.** Es el número de veces que se ha redirigido un sitio web. Si los sitios web legítimos han sido redirigidos una vez como máximo es confiable, si han sido redirigido más de 4 veces es un sitio no confinable, de lo contrario es sospechoso.
20. **Barra_estado**. Se puede usar JavaScript para mostrar una URL falsa en la barra de estado a los usuarios. Para extraer esta función, debemos extraer el código fuente de la página web, en particular el evento "onMouseOver", y verificar si realiza algún cambio en la barra de estado. Si no hay un cambio de estado entonces es un sitio confinable, de lo contrario no lo es.
21. **Click_derecho.** Se puede usar JavaScript para deshabilitar la función de clic derecho, de modo que los usuarios no puedan ver y guardar el código fuente de la página web. Se busca el evento "event.button == 2" en el código fuente de la página web y se verifica si el clic derecho está deshabilitado, se es así, es un sitio no confiable, de lo contrario es confiable.
22. **Pop-up.** No es común que un sitio web legítimo solicite datos personales o confidenciales a través de una ventana emergente. Si no contiene una ventana emergente (popoup) entones es confinable, de lo contrario no lo es.
23. I**Frame**. Es una etiqueta HTML que se utiliza para mostrar una página web adicional en una que se muestra actualmente. Se puede hacer uso de la etiqueta "iframe" y hacerla invisible, es decir, sin bordes de marco. Si no se encuentra la función de IFrame entonces es confinable, de lo contrario no lo es.

Características de los sitios web, basados en la inteligencia de amenazas.

24. **Domain_edad.** La mayoría de sitios web legítimos tienen un tiempo de vida largo. Se extrae de la base de datos whois y si el tiempo de vida es largo (>6 meses) es confiable, de lo contrario no lo es.
25. **Registro_DNS.** La mayoría de sitios web falsos la base de datos WHOIS no reconoce la identidad o no tienen registros para el hostname. Si se encuentran registros DNS entones es confiable, de lo contrario no lo es.
26. **Trafico**. La popularidad del sitio web está determinando por el número de visitantes. Los sitios web con corta duración normalmente no se registran en sitios de popularidad, por tanto lo sitios confiables es muy probable que estén registrados y sean reconocidos por la base de datos Alexa. Si no reconoce al sitio web entonces es no confiable, si lo reconoce y lo ubica entre los 100000 sitios principales es confiable, de lo contrario es sospechoso.
27. **PageRank**. PageRank tiene como objetivo medir la importancia de una página web. Cuanto mayor sea el valor de PageRank, más importante será la página web. Si el sitio web tienen un pagerank mayor a 0.2 entonces es confiable, de lo contrario no lo es.
28. **Google_index.** Esta función examina si un sitio web está en el índice de Google o no. Si el sitio web está en el Google index, entonces es confiable, de lo contrario no lo es.
29. **Links**. El número de enlaces que apuntan a la página web puede ayudarnos a identificar a un sitio web legítimo. Si el número de enlaces es mayor a 2 entonces es confiable; si no tiene enlaces no es confiable, de lo contrario es sospechoso.
30. **Estadísticas**. Varios sitios especializados como phishTank y StopBadware publican reportes de informes estadísticos de sitios web phishing. Buscamos en esos sitios web las principales direcciones ip y direcciones de dominio phishing, y si el sitio no está en los reportes entonces es un sitio confiable, de lo contrario no lo es.

Fase 2. Preparación los datos.

El objetivo de esta fase es realizar un análisis de los datos, preprocesamiento y división de los datos, para el entrenamiento y para la detección.

El análisis de los datos consiste en visualizar los tipos de datos, hacer resúmenes de los datos, filtrar los datos, agrupar los datos por clases y visualizar si hay datos perdidos o fuera de rango. Además de realizar visualizaciones de correspondencia.

El procesamiento de los datos consiste en hacer una limpieza a los datos y una transformación a los mismos aplicando técnicas de escalamiento, normalización y binarizacion

Se realiza la reducción de características y finalmente en esta fase los datos son divididos en un subconjunto de datos para el entrenamiento y un subconjunto de datos para evaluación en la detección de phishing por el sistema.

Fase 3. Selección de algoritmos de Machine Learning.

El objetivo de esta fase es seleccionar los algoritmos para la fase de entrenamiento.

Se realiza una evaluación de todos los algoritmos de machine learning para la clasificación, con los datos de entrenamiento, se realiza una evaluación de los resultados con una validación cruzada.

Se elige los algoritmos que obtienen los mejores resultados de las métricas de clasificación, especialmente el accuracy.

Fase 4. Entrenamiento del Sistema

El objetivo de esta fase es entrenar un algoritmo que brinde los mejores resultados y guardarlo como el modelo para realizar las predicciones.

Se realiza la optimización de los algoritmos elegidos en la fase anterior, con los datos de entrenamiento. Con los mejores parámetros se realiza el afinamiento del modelo y el modelo obtenido se guarda para el uso en la siguiente fase.

Fase 5. Detección de Phishing

El objetivo de esta fase es verificar el funcionamiento del modelo almacenado, en la clasificación de un sitio web como phishing o como sitio web legitimo.

Con el subconjunto de datos de validación (datos nuevos para el sistema), se verifica la capacidad predictiva del sistema, ingresando los datos originales y comparando con las predicciones del sistema, para la evaluación del rendimiento.

Fase 6. Evaluación del rendimiento

El objetivo de esta fase es evaluar el rendimiento del modelo generado, en la clasificación de un sitio web como phishing o como legitima.

Se evalúa el modelo de machine learning en la detección de sitios web falsos, utilizando las siguientes métricas.

- Verdaderos positivos (VP). Cantidad de pruebas clasificadas correctamente un sitio web falso, como sitio web falso.
- Verdaderos negativos (VN). Cantidad de pruebas clasificadas correctamente un sitio web real, como sitio web real.
- Falsos positivos (FP). Cantidad de pruebas clasificadas erróneamente un sitio web real, como sitio web falso.
- Falsos Negativos (FN). Cantidad de pruebas clasificadas erróneamente un sitio web falso, como sitio web real.
- El error en la clasificación se representa como la proporción de instancias clasificadas incorrectamente a todas las instancias.
- Accuracy, es la proporción de clasificación correcta de todas las instancias que se utilizan se calcula mediante la fórmula.
- Tasa de verdaderos positivos (TPR) es proporción de instancias clasificadas correctamente como positivas.
- La tasa de falsos positivos (FPR), utilizando la proporción de instancias clasificadas incorrectamente como positivas.
- La Precisión. es la proporción de instancias clasificadas correctamente como positivas a todas las instancias clasificadas positivamente:

3.5 Implementación del Sistema de Detección de Phishing

La implementación del Sistema de Detección de Phishing se lleva a cabo en las fases descritas anteriormente. Se implementa en el lenguaje de programación Python, en un cuaderno de Jupyter Notebook en Anaconda. Se hace uso de las librerías open source.

A continuación, se describe el proceso de implementación.

Fase 1. Recolección de datos.

La recolección de datos, basadas en la información de la barra de direcciones y el código fuente son:

- IP_Address.
- URL_longitud.
- URL_corto.
- URL_arroba.
- URL_slash.
- URL_linea.
- URL_puntos.
- HTTPS_SSL.
- Domain_registro.
- Favicon.
- Puerto.
- Domain_https.
- Request_URL.
- URL_ancla.
- Tags.
- SFH.
- Submit_email.
- Abnormal_URL.
- Forwarding.
- Barra_estado.
- Click_derecho.
- Pop-up.
- IFrame.

Se implementa un módulo automatizado en Python, para recoger la información de las características de un sitio web y el resultado se muestra en la figura 16.
La información recogida está codificada, con los valores 1 (sitios confiables), 0 (sitios sospechosos), y 1 (sitios no confiables).

Figura 16: Recolección de datos del sitio web

```
recolectar_datos_sitioweb("https://www.aulauss.edu.pe/")
```

```
[1, 1, 1, 1, 1, 1, 1, 1, -1, 1, 1, 1, -1, -1, -1, 1, 1, -1,
-1, -1, -1, -1, 1]
```

La recolección de datos, basadas en la inteligencia de amenazas son:

- Domain_edad.
- Registro_DNS.
- Trafico.
- PageRank.
- Google_index.
- Links.
- Estadísticas.

Se implementa un módulo automatizado en Python, para recoger la información de la inteligencia de amenazas relacionadas a un sitio web y resultado se muestra en la figura 17.
La información recogida está codificada, con los valores 1 (sitios confiables), 0 (sitios sospechosos), y 1 (sitios no confiables).

Figura 17: Recolección de datos de inteligencia de amenazas

```
recolectar_datos_inteligencia("https://www.aulauss.edu.pe/")
```

```
[1, -1, 0, -1, 1, -1, 1]
```

Utilizando estos módulos automatizados, se recogen los datos de 11055 sitios web.

Fase 02. Preparación de los datos.

Luego de recoger los datos, se integran y se agregan las etiquetas por cada registro realizado, y se realizan 3 actividades.

1. Análisis de los datos.

Cargamos los datos y sus características. En la figura 18, se visualiza la salida de las dimensiones del conjunto de datos, con 11055 registros y 31 características. (23 características de sitios web, 7 características de inteligencia de amenazas y la etiqueta del sitio web).

Figura 18: Visualización de los datos totales.

```
filename = 'DATACOMPLETA.csv'
data = pd.read_csv(filename, names=None)
```

```
data.shape
```

```
(11055, 31)
```

Observamos los tipos de datos, por cada característica del conjunto de datos y de la etiqueta. En la figura 19 se muestra que todos los datos incluyendo las características y la etiqueta, son enteros.

Figura 19: Tipos de los datos.

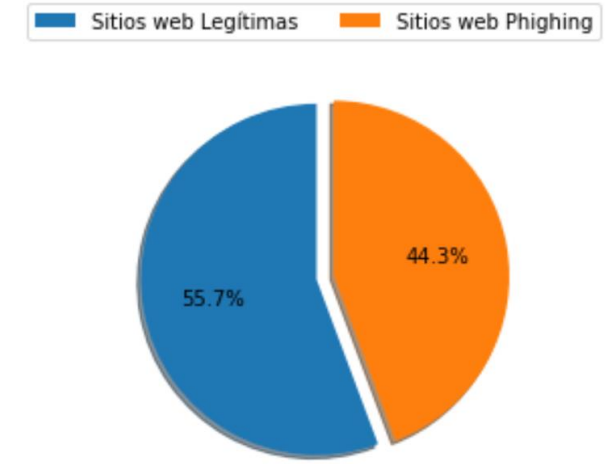
```
data.dtypes
```

```
IP_Address          int64      Submit_email        int64
URL_longitud        int64      Abnormal_URL        int64
URL_corto           int64      Fordwarding         int64
URL_arroba          int64      Barra_estado        int64
URL_slash           int64      Click_derecho       int64
URL_linea           int64      PopUp               int64
URL_puntos          int64      Iframe              int64
SSL                 int64      Domain_edad         int64
Domain_registro     int64      Registro_DNS        int64
Favicon             int64      Trafico_web         int64
Puerto              int64      Page_Rank           int64
Domain_https        int64      Google_Index        int64
Request_URL         int64      Links               int64
URL_ancla           int64      Estadisticas        int64
Tags                int64      Result              int64
SFH                 int64      dtype: object
```

Implementamos un módulo automatizado para la estadística descriptiva por característica y visualizar el comportamiento de cada característica, en la figura 20, se muestra la cantidad de datos por clase (sitios web legítimos vs sitios web phishing) en el conjunto de datos y en la figura 212, la cantidad de sitios web que tienen una dirección IP en la URL y los sitios web sin direcciones IP en la URL.

Figura 20: Datos de sitios web legítimos vs sitios web phishing.

```
show_caracteristica('Result',etiquetas)
```




```
 1    6157
-1    4898
Name: Result, dtype: int64
```

Figura 21: Datos de sitios web con y sin direcciones IP en la URL.

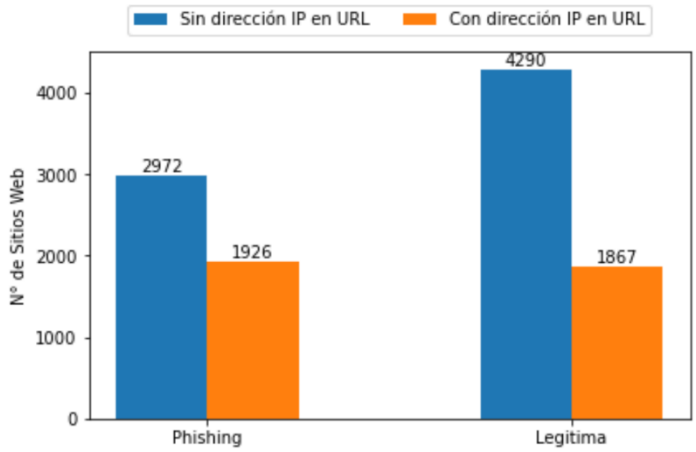
```
show_caracterisitca('IP_Address',etiquetas)
```


```
 1    7262
-1    3793
Name: IP_Address, dtype: int64
```

Implementamos un módulo automatizado para la estadística descriptiva de cada característica agrupada por clase, en la figura 22, se muestra la cantidad de sitios web de acuerdo con la característica de la dirección IP por clase, y en la figura 23, se muestra la longitud de los sitios web, agrupados por clase.

Figura 22: Datos de sitios web con direcciones IP en la URL, por clase.

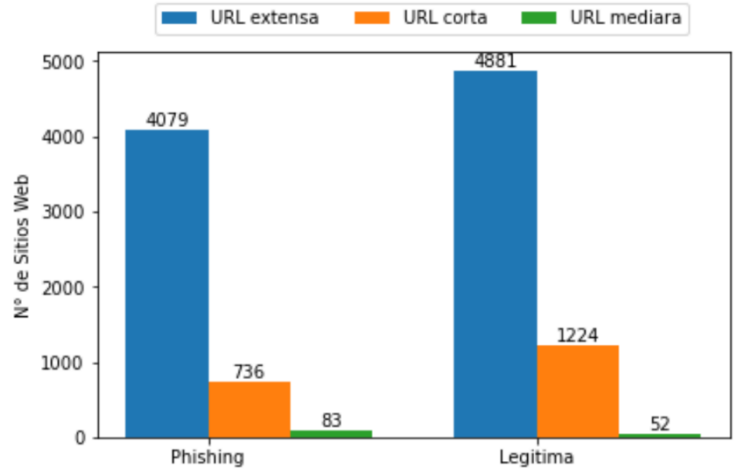
```
show_caracteristica_url('IP_Address', A, B)
```



	Sin dirección IP en URL	Con dirección IP en URL
Phishing	2972	1926
Legitima	4290	1867

Figura 23: Datos del tamaño de los sitios web, por clase.

```
show_caracteristica_url3(caracteristica, A, B, C)
```



	URL extensa	URL corta	URL mediara
Phishing	4079	736	83
Legitima	4881	1224	52

Visualizamos la correlación de cada característica con respecto a la clase, en la figura 24 se muestra una correlación muy alta (0.714) de SSL con respecto a la clase, seguido por la característica de URL_ancla (0.692); además se observa que hay varias características con baja correlación con la clase, con valores por debajo de 0.10.

Figura 24: Correlación de las características con la clase.

```
cor_target = abs(correlaciones['Result'])
print(cor_target)
```

```
IP_Address          0.094160
URL_longitud        0.057430
URL_corto           0.067966
URL_arroba          0.052948
URL_slash           0.038608
URL_linea           0.348606
URL_puntos          0.298323
SSL                 0.714741
Domain_registro     0.225789
Favicon             0.000280
Puerto              0.036419
Domain_https        0.039854
Request_URL         0.253372
URL_ancla           0.692935
Tags                0.248229
SFH                 0.221419
Submit_email        0.018249
Abnormal_URL        0.060488
Fordwarding         0.020113
Barra_estado        0.041838
Click_derecho       0.012653
PopUp               0.000086
Iframe              0.003394
Domain_edad         0.121496
Registro_DNS        0.075718
Trafico_web         0.346103
Page_Rank           0.104645
Google_Index        0.128950
Links               0.032574
Estadisticas        0.079857
Result              1.000000
Name: Result, dtype: float64
```

En la figura 25 se observa que todas las características tienen 11055 registros, indicando que no hay datos perdidos o nulos.

Figura 25: Cantidad de datos por característica.

```
data.count()
```

```
IP_Address          11055
URL_longitud        11055
URL_corto           11055
URL_arroba          11055
URL_slash           11055
URL_linea           11055
URL_puntos          11055
SSL                 11055
Domain_registro     11055
Favicon             11055
Puerto              11055
Domain_https        11055
Request_URL         11055
URL_ancla           11055
Tags                11055
SFH                 11055
Submit_email        11055
Abnormal_URL        11055
Fordwarding         11055
Barra_estado        11055
Click_derecho       11055
PopUp               11055
Iframe              11055
Domain_edad         11055
Registro_DNS        11055
Trafico_web         11055
Page_Rank           11055
Google_Index        11055
Links               11055
Estadisticas        11055
Result              11055
dtype: int64
```

Figura 26: Densidad de las características de los sitios web

2. Preprocesamiento de datos.

El conjunto de datos no cuenta con datos perdidos o NaN, como se visualiza en la figura 25; tampoco se observa datos fuera de rango o datos atípicos. Por tanto, no es necesario realizar limpieza de los mismos.

Además, los datos, están integrados y están codificados desde la recolección de los mismos, y corresponden a una distribución gaussiana, por tanto, no es necesario la integración, normalización, estandarización tampoco la binarización.

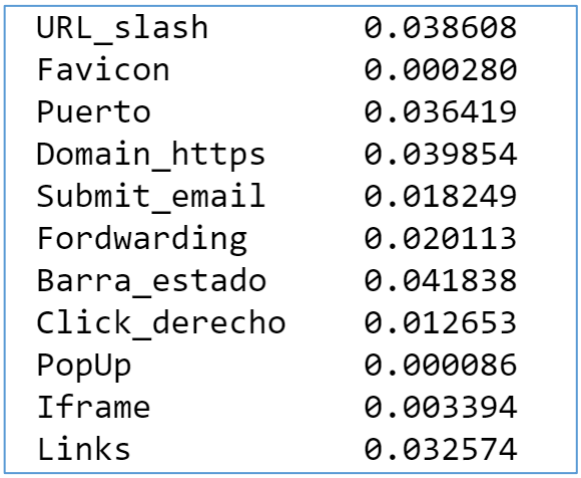
Se observó en la figura 24 que hay muchas características con baja importancia de correlación con respecto a la salida, por tanto, se procede a realizar reducción de características.

Aplicamos la técnica de Random Forest para la selección de importancia de las características y en la figura 27, se muestra que las características que representan mayor importancia en la obtención de mejores resultados están en la parte superior de la figura y resaltan URL_puntos y Request_URL; y en la parte inferior se muestran la característica con menor grado de importancia y resaltan la Barra_estado y el PopUp.

Figura 27: Importancia de las características de los sitios web

Generamos tres conjuntos de datos, para el modelado, y la evaluación de resultados. El primer conjunto de datos es el original, sin ningún cambio; el segundo conjunto de datos solo con las características que tengan una correlación mayor a 0.05 con la clase. En la figura 28 se muestran las características no relevantes, y que se eliminarán del conjunto de datos original, para crear el segundo conjunto de datos.

Figura 28: Características no relevantes de los sitios web



URL_slash	0.038608
Favicon	0.000280
Puerto	0.036419
Domain_https	0.039854
Submit_email	0.018249
Fordwarding	0.020113
Barra_estado	0.041838
Click_derecho	0.012653
PopUp	0.000086
Iframe	0.003394
Links	0.032574

El tercer conjunto de datos de generan con 5 características aplicando la transformación PCA (Análisis de Principales características). En la figura 29 se observa el tercer conjunto de datos con solo 5 características aplicando PCA.

Figura 29: Reducción de características con PCA

	PC1	PC2	PC3	PC4	PC5
0	1.476993	0.816895	2.618606	-2.603463	-0.403801
1	-0.335021	1.248301	0.669381	-1.227114	0.049686
2	1.029841	0.956666	1.460193	-1.239660	-1.615831
3	-0.645561	-1.143645	1.993959	-0.263956	0.239800
4	1.119945	0.984423	0.142774	-0.651215	0.348038

3. Remuestreo de los datos.

Dividimos los conjuntos de datos en dos porciones y de forma aleatoria; 80% para los datos de entrenamiento y 20% para los datos de validación del modelo final.

De los 11055 sitios web preprocesados 8844 sitios web serán utilizados para el entrenamiento y construcción del modelo de detección y 2211 sitios web se utilizan para la validación del modelo. Cada subconjunto de datos tiene sitios web con phishing y sitios web legítimos.

Fase 03. Selección de algoritmos.

Implementamos el código en Python, y evaluamos el rendimiento de los algoritmos de clasificación, en la figura 30 se muestra el código con el listado de algoritmos a evaluar.

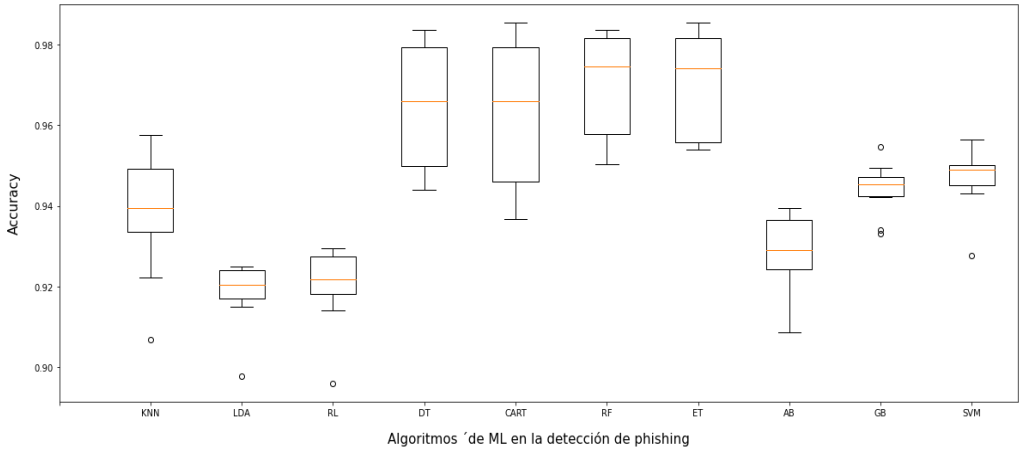
Figura 30: Código con algoritmos a evaluar el rendimiento

```
models=[]
models.append(("KNN", KNeighborsClassifier()))
models.append(("LDA", LinearDiscriminantAnalysis()))
models.append(("RL", LogisticRegression()))
models.append(("DT", BaggingClassifier(base_estimator=DecisionTreeClassifier())))
models.append(("CART", DecisionTreeClassifier()))
models.append(("RF", RandomForestClassifier()))
models.append(("ET", ExtraTreesClassifier()))
models.append(("AB", AdaBoostClassifier()))
models.append(("GB", GradientBoostingClassifier()))
models.append(("SVM", SVC()))
```

Realizamos la evaluación de los algoritmos con sus parámetros por defecto y utilizando el conjunto de datos de entrenamiento.

En la figura 31 observamos que los algoritmos DT, CART, RF y ET, tienen los mejores resultados de accuracy, así mismo son algoritmos que se necesitan analizar para el problema de la detección de phishing. Por lo tanto, los algoritmos seleccionados para la elaboración del modelo de detección de phishing son DT, CART, RF y ET. Además los datos originales son los que brindan mejores resultados y se utilizarán para las etapas posteriores.

Figura 31: Evaluación de los algoritmos a seleccionar.



Fase 04. Entrenamiento.

Implementamos el código en Python, y para el entrenamiento base, con los algoritmos seleccionados en la fase anterior, en la figura 32 se muestran los 4 algoritmos que se utilizan para el entrenamiento del modelo de detección de phishing.

Figura 32: Código con los Algoritmos para el entrenamiento

```
models=[]
models.append(("DT", BaggingClassifier(base_estimator=DecisionTreeClassifier())))
models.append(("CART", DecisionTreeClassifier()))
models.append(("RF", RandomForestClassifier()))
models.append(("ET", ExtraTreesClassifier()))
```

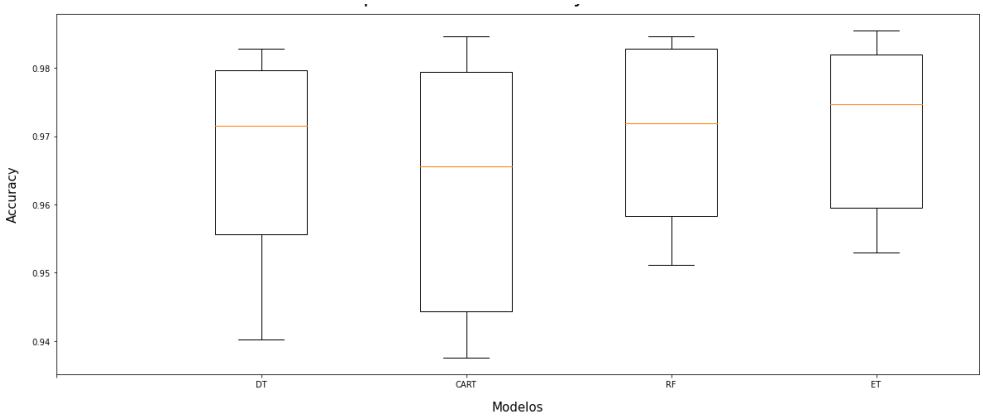
Luego del entrenamiento en la figura 33, se muestra los resultados de los 4 algoritmos en evaluación, donde se muestra que el algoritmo Extra Tree, tiene el mayor accuracy con 97.05% que los otros algoritmos, seguido por Random Forest con el 97.00% de accuracy. Además, se puede observar que los algoritmos evaluados como DT y CART no pueden descartase de la evaluación porque los resultados obtenidos están dentro del rango de RF y de ET.

Figura 33: Accuracy base en la detección de phishing

	DT	CART	RF	ET
Accuracy_Base	96.752056	96.172960	97.005368	97.050592
Desv_Stand	1.347015	1.770959	1.303008	1.274880

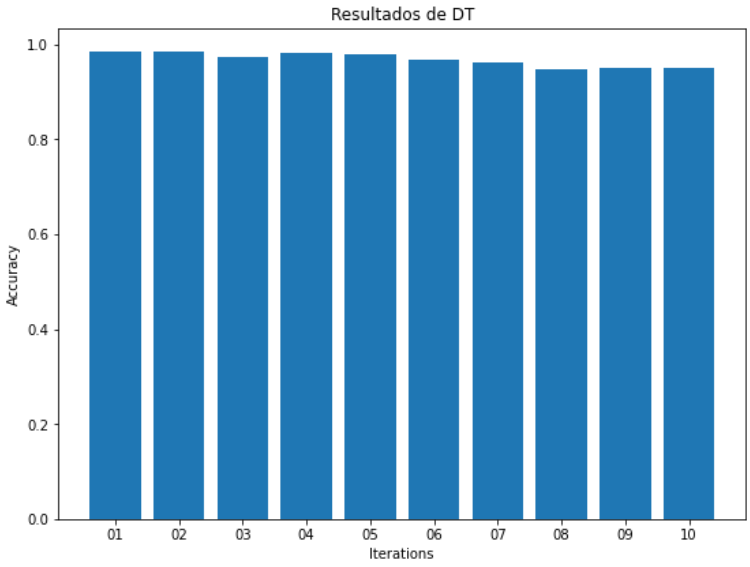
En la figura 34, mes muestran los resultados base, en forma gráfica, donde se concluye que tanto el DT, RF y ET tienen resultados muy similares, y por tanto debemos realizar la búsqueda de los mejores parámetros de cada algoritmo y evaluar sus resultados.

Figura 34: Comparación del Accuracy base en la detección de phishing



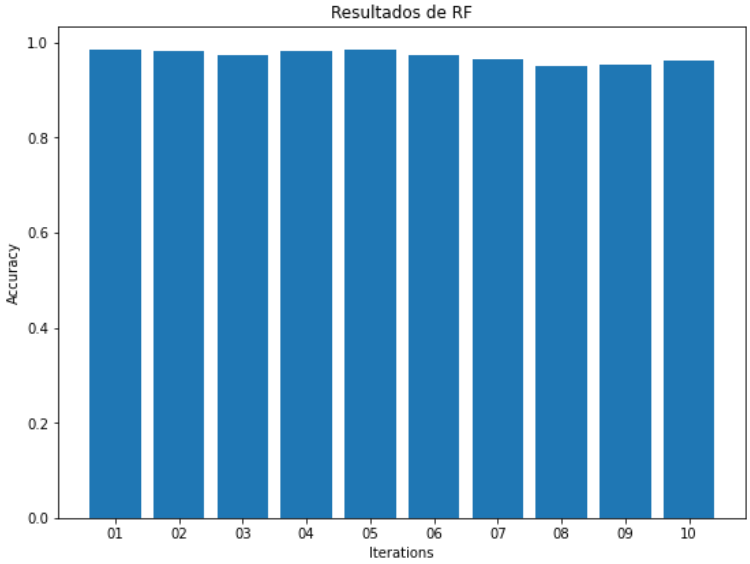
Haciendo uso de Grid Search, buscamos los mejores parámetros para los algoritmos DT, RF y ET. Con los mejores parámetros se entrena cada modelo con el conjunto de datos de entrenamiento, y para validar los resultados utilizamos validación cruzada con 10 interacciones. En la figura 35 se muestran los resultados del entrenamiento del algoritmo DT con las 10 interacciones.

Figura 35: Accuracy en el entrenamiento de RT con validación cruzada



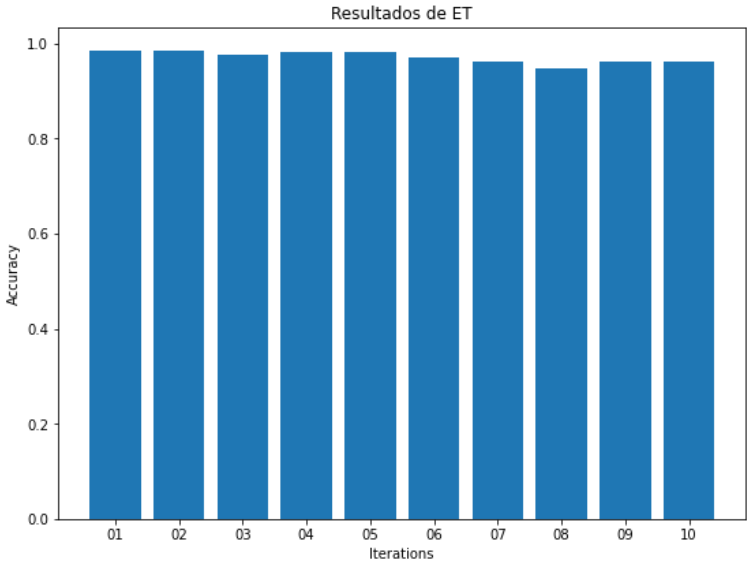
En la figura 36 se muestran los resultados del entrenamiento del algoritmo RF con las 10 interacciones.

Figura 36: Accuracy en el entrenamiento de RF con validación cruzada



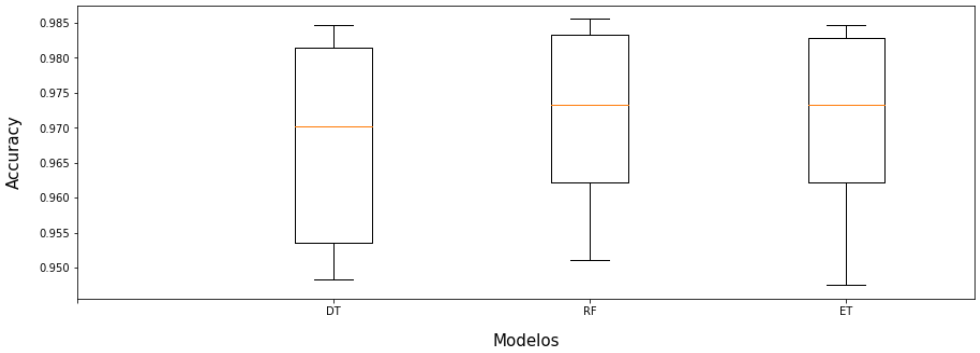
En la figura 37 se muestran los resultados del entrenamiento del algoritmo ET con las 10 interacciones.

Figura 37: Accuracy en el entrenamiento de ET con validación cruzada



Los resultados de los tres algoritmos optimizados se muestran en forma gráfica en la figura 38, donde se observa que los algoritmos RF y ET, tienen los mejores y muy similares resultados.

Figura 38: Resultados del Accuracy de algoritmos optimizados



El promedio del accuracy y la desviación estándar de los resultados por cada algoritmo se resumen en la figura 39, donde se muestra que tanto el algoritmo Random Forest y Extra Tree, tienen los mismos resultados en cuanto al promedio de 97.14%. Sin embargo, los datos de DT no se pueden descartar debido a que se encuentra en los rangos de RF y ET.

Figura 39: Accuracy promedio optimizado en la detección de phishing

	DT	RF	ET
Accuracy	96.797280	97.141090	97.141090
Desv_Stand	1.370918	1.226336	1.201401

En la figura 40, se muestra el código de afinamiento, generación y almacenamiento de los 3 modelos.

Figura 40: Código de afinamiento y almacenamiento de los modelos

```
model_DT.fit(X_train, Y_train)
filename_DT = 'Modelos/ModeloDT.sav'
jbl.dump(model_DT, filename_DT)
```

```
['Modelos/ModeloDT.sav']
```

```
model_RF.fit(X_train, Y_train)
filename_RF = 'Modelos/ModeloRF.sav'
jbl.dump(model_RF, filename_RF)
```

```
['Modelos/ModeloRF.sav']
```

```
model_ET.fit(X_train, Y_train)
filename_ET = 'Modelos/ModeloETsav'
jbl.dump(model_ET, filename_ET)
```

```
['Modelos/ModeloETsav']
```

Fase 05. Detección de Phishing.

La implementación de esta fase se realiza en dos formas.

1. Detección de phishing con la URL, que permite verificar en línea si un sitio web es legítimo o es sitio web phishing. En este caso el código realiza dos funciones, la función de recolectar datos del sitio web y luego realiza la predicción del sitio a que clase corresponde.

   En la figura 41 se visualiza que la función se prueba con un sitio web, y el sistema da como resultado la predicción, que se muestra en la parte inferior, en el ejemplo “El sitio web es Legítimo”.

   Figura 41: Prueba de detección de un sitio web en línea

```
predecir_url("https://www.aulauss.edu.pe/")
```

```
El sitio web es LEGÍTIMO
```

2. La detección de sitios web, con las características, esta funcionalidad se ha implementado con la finalidad de comprobar el funcionamiento con los sitios web phishing que ya no están en línea, pero que si se cuenta con los datos de las características.

   En la figura 42 se visualiza que la función se prueba con datos de un sitio web phishing y con datos de un sitio web legítimo; los resultados de las predicciones se realizan correctamente y se imprime en la parte inferior de cada detección.

   Figura 42: Prueba de detección de un sitio web con datos

```
from termcolor import colored
def predecir_data(dataPredecir):
    dataP= np.append([dataPredecir], [[]], axis=1)
    prediccion = loaded_model_DT.predict(dataP[:,1:30])
    if prediccion[0] == 1:
        salida = "El sitio web es LEGÍTIMO"
        color = "green"
    else:
        salida = "El sitio web es PHISHING"
        color = "red"
    print(colored(salida, color))
```

```
data=[1,1,1,1,1,-1,0,1,-1,1,1,-1,1,0,-1,-1,1,1,0,1,1,1,1,-1,-1,0,-1,1,1,1]
predecir_data(data)
```

```
El sitio web es PHISHING
```

```
data=[1,0,-1,1,1,-1,1,1,-1,1,1,1,1,0,0,-1,1,1,0,-1,1,-1,1,-1,-1,0,-1,1,1,1]
predecir_data(data)
```

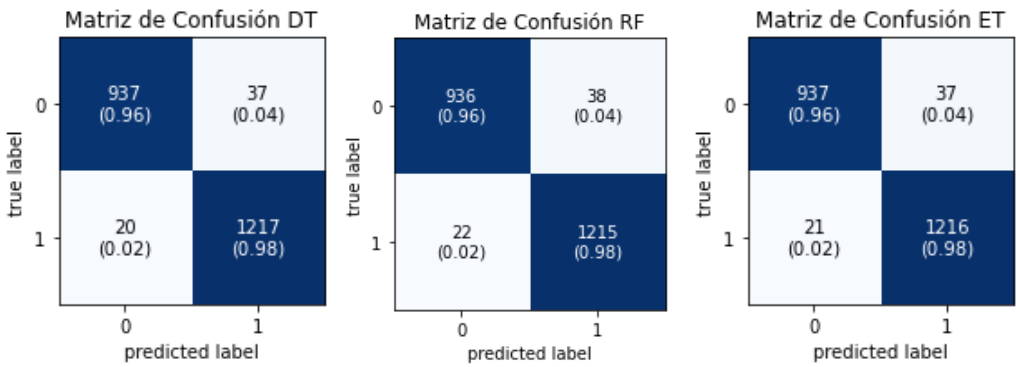
```
El sitio web es LEGÍTIMO
```

Fase 06. Evaluación del rendimiento.

Para evaluar el rendimiento del sistema en la detección de phishing, se hacen pruebas de detección, con el conjunto de datos de validación y con los modelos generados DT, RF y ET.

En la figura 43, se muestra la matriz de confusión de cada modelo, donde se muestra que el modelo DT es la que proporciona los mejores resultados en la detección de sitios web phishing y sitios web legítimos, igualando con ET en los Verdaderos Positivos (VP) y superando a RF y ET en los Verdaderos Negativos (VN); además se observa que tiene menores errores de Falsos Positivos (FN) y de Falsos Positivos (FP).

Figura 43: Matriz de confusión de los modelos evaluados



En la matriz de confusión del modelo final DT, se observa que 937 (96%) sitios web phishing detectados correctamente como sitios web phishing (VP); 1217 (98%) sitios web legítimos detectados correctamente como sitios web legítimos (VN), además se muestran 37 (4%) de sitios web phishing incorrectamente clasificados como sitios legítimos (FN) y 20 (2%) sitios web legítimos clasificados incorrectamente sitios web phishing.

En la figura 44, se muestra el reporte de clasificación del modelo DT, con los resultados de las métricas de accuracy al 97.42%, recall (TVP) del 96.20%, Sensibilidad (TVN) de 98.38%, precisión de 97.91%.

Figura 44: Reporte de clasificación del modelo final

	precision	recall	f1-score	support
-1	0.9791	0.9620	0.9705	974
1	0.9705	0.9838	0.9771	1237
accuracy			0.9742	2211
macro avg	0.9748	0.9729	0.9738	2211
weighted avg	0.9743	0.9742	0.9742	2211

En la tabla 8, se muestra la matriz de confusión del sistema construido con el algoritmo Decision Tree, y las pruebas realizadas con 2211 sitios web, donde de muestran 937 verdaderos positivos, 1217 verdaderos negativos, 37 falsos negativos y 20 falsos positivos.

Tabla 8: Matriz de confusión, del sistema propuesto

	Clases	**Resultado del Sistema Propuesto**		
		Phishing	**Legítimo**	**Total**
Sitio web (Clasificación real)	Phishing	937 (VP)	37 (FN)	974
	Legítima	20 (FP)	1217 (VN)	1237

En la tabla 9 se muestra los resultados de la clasificación del sistema propuesto; donde se muestra que el 97.42% es la proporción de clasificación correcta del sistema en global, el 2.58% es el error en la clasificación errónea del sistema en global, el 96.20% de sitios web phishing clasificadas correctamente, el 98.358% de sitios web legítimos clasificados correctamente y el 97.91% es la proporción de sitios web phishing clasificados correctamente en relación a la clasificación como sitios web phishing por el sistema.

Tabla 9: Resultados del rendimiento del sistema propuesto

Métrica Evaluada	**Fórmula**	**Resultado**
Accuracy	Accuracy = (VP+VN) /(VP+VN+FP+FN)	97.42%
Classification Error	Classification Error = (FP+FN)/(VP+VN+FN+FP)	02.58%
Recall (TVP)	recall = (VP)/(VP+FN)	96.20%
Specificity (TVN)	Specificity = (VN)/(VN+FP)	98.38%
Precision	Precision = VP / (VP + FP)	97.91%

3.6 Valoración y corroboración de los resultados

Los resultados del diagnóstico contextual, basado en estudios previos se determinó un accuracy de 93.80% como métrica de rendimiento en la detección de phishing, y los resultados de la evaluación del rendimiento con la implementación del sistema de detección de phishing que es el aporte práctico de esta tesis se obtiene un accuracy de 97.42%, notándose una diferencia superior en el rendimiento del sistema propuesto.

Para la evaluación de los resultados tanto del sistema base y del sistema propuesto se utilizaron en total 2211 sitios web, divididos en sitios phishing y sitios web legítimos.

En la tabla 10 se muestra la comparación de los resultados del rendimiento del sistema base de estudios previos y el sistema de detección de phishing, aporte práctico de esta tesis.

Tabla 10: Comparación de rendimiento en la detección de phishing

Métrica Evaluada	Sistema Base	Sistema Propuesto	Diferencia
Accuracy	93.80%	97.42%	03.62%
Classification Error	06.20%	02.58%	03.62%
Recall (TVP)	91.80%	96.20%	04.40%
Specificity (TVN)	95.35%	98.38%	03.03%
Precision	93.84%	97.91%	04.70%

Se observa una diferencia de 3.62% en el accuracy, que indica una mayor proporción de clasificación correcta global del sistema propuesto; un incremento notorio en la TVP de 4.40% con el sistema propuesto, que indica una mejora en la clasificación correcta de sitios web phishing; además de una mejora en la TVN de 3.03% que indica una mejora en la clasificación correcta de sitios web legítimos y una diferencia de 4.70% en precisión de sitios web phishing clasificados correctamente.

Además, en la etapa de entrenamiento se realizó validación cruzada con 10 interacciones y los resultados se comparan a continuación.

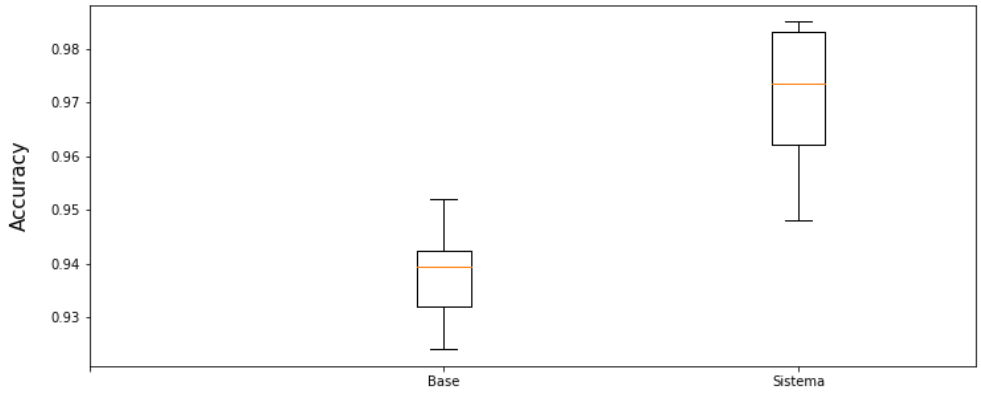
En la tabla 11, se muestra la media del accuracy, la desviación estándar, el mínimo y el máximo del accuracy en la etapa de entrenamiento, donde en la construcción del sistema tiene mayor rendimiento como media, en loa valores mínimos y en los valores máximos, por lo que se puede apreciar que el sistema propuesto tiene mejores resultados que el sistema base.

Tabla 11: Comparación del accuracy en la etapa de entrenamiento

Métrica	Media	Desviación	Mínimo	Máximo
Base	93.77	0.93	92.40	95.20
Sistema	97.16	1.27	94.80	98.50

En la figura 47, se observa el diagrama de cajas, donde se muestra que los resultados del sistema en la etapa de entrenamiento están por encima de los resultados base, y que están fuera de alcance por este.

Figura 47: Diagrama de cajas, de resultados de entrenamiento



Se realizó la prueba de correlación de Pearson entre los resultados del accuracy de entrenamiento, y el resultado es de -0.277, indicando que no hay correlación entre los dos resultados.

Tanto en la etapa de entrenamiento del sistema y en la etapa de detección, el sistema propuesto tiene mejor rendimiento en la detección de sitios web phishing que los estudios previos.

IV. CONCLUSIONES

- Se realizó la caracterización del proceso de ciberseguridad y la detección de phishing, demostrando que hay insuficiencias teóricas y prácticas que usen la información de los sitios web, la información de la inteligencia de amenazas y las técnicas de machine learning, para la detección de sitios web falsos.
- En el diagnóstico del rendimiento de la detección de sitios web phishing basado en estudios previos, se determinó que el rendimiento promedio de los sistemas de detección de phishing es de 93.80%, considerando que es la proporción de detección correcta en global; además de 91.20% de sitios web phishing clasificadas correctamente y 95.35% de sitios web legítimos clasificados correctamente.
- Se elaboró un modelo de machine learning, para la detección de sitios web phishing, utilizando las características de los sitios web en la barra de direcciones y en el código fuente, la inteligencia de amenazas y se integró con la lógica, las técnicas y los algoritmos de machine learning. El modelo está compuesto por seis dimensiones que son sitio web, inteligencia de amenazas, preparación de datos, algoritmos de machine learning, entrenamiento y detección de phishing.
- Se elaboró un Sistema de Detección de Phishing, basado en el modelo de machine learning propuesto, demostrándose su aporte y el logro de objetivos planteados.
- Se aplicó el sistema de detección de phishing y se evaluó con 2211 sitios web, demostrando el rendimiento en de 97.42% de la detección correcta del sistema propuesto en forma global, así mismo el 96.20% de sitios web phishing clasificadas correctamente, el 98.358% de sitios web legítimos clasificados correctamente
- Se corroboró estadísticamente los resultados de la evaluación de la propuesta en donde se compararon los indicadores en la detección de sitios web falsos, lográndose demostrar cambios positivos, con una mejora de 3.62%, en el rendimiento global, un incremento de 4.40% en la clasificación correcta de los sitios phishing y una mejora de 3.03% en la clasificación correcta de sitios web legítimos.

V. RECOMENDACIONES

- Se recomienda recolectar mayor número de sitios web phishing y sitios web legítimos, para evaluar el rendimiento del sistema de detección de phishing propuesto.
- Se recomienda aplicar la propuesta de investigación a otros tipos de ataques informáticos, lo que permitirá mejorar el rendimiento en la detección de ataques.
- Se recomienda adaptar el modelo propuesto, y aplicar a otros campos de investigación, debido a los buenos resultados.

VI. REFERENCIAS

Abdulhamit , S., & Kremicb, E. (2020). Comparison of Adaboost with MultiBoosting for Phishing Website Detection. *Procedia Computer Science*, 272–278.

Abutair, H. Y., Belghith, A., & Al-Ahmadi, S. A. (2018). CBR-PDS: a case-based reasoning phishing detection system. *Journal of Ambient Intelligence and Humanized Computing*, 2593-2606. doi: https://doi.org/10.1007/s12652-018-0736-0.

Ahrend, J. M., Jirotka, M., & Jones, K. (2016). *On the collaborative practices of cyber threat intelligence analysts to develop and utilize tacit Threat and Defence Knowledge.* pp. 1-10, doi: 10.1109/CyberSA.2016.7503279: International Conference On Cyber Situational Awareness, Data Analytics And Assessment (CyberSA).

Alexa the Web Information Company. (1996). *Alexa the Web Information Company*. Recuperado el 10 de November de 2011, de http://www.alexa.com/

Ali, W., & Ahmed, A. A. (2019). Hybrid intelligent phishing website prediction using deep neural networks with genetic algorithm-based feature selection and weighting. *IET Information Security*, Pages 659-669; doi: https://doi.org/10.1049/iet-ifs.2019.0006.

Aljofey, A., Jiang, Q., Qu, Q., Huang, M., & Niyigena, J.-P. (2020). An Effective Phishing Detection Model Based on Character Level Convolutional Neural Network from URL. *Electronics* , doi: https://doi.org/10.3390/electronics9091514.

Alsariera, Y. A., Adeyemo, V. E., Balogun, A. O., & Alazzawi, A. K. (2020). AI Meta-Learners and Extra-Trees Algorithm for the Detection of Phishing Websites. *IEEE Access, vol. 8*, pp. 142532-142542, doi: 10.1109/ACCESS.2020.3013699.

Anderson, J. P. (1972). *Computer Security Technology Planning Study.* Washington: Deputy for command and Management Systems.

Anupam, S., & Kar, A. K. (2020). Phishing website detection using support vector machines and nature-inspired optimization algorithms. *Telecommunication Systems*, pages 17–32; doi https://doi.org/10.1007/s11235-020-00739-w.

APWG. (2021). *Phishing Activity Trends Report, 4th Quarter 2020.* Anti-Phishing Working Group, Inc.

Baca Urbina, G. (2016). *Introducción a la seguridad informática.* Distrito Federal, MÉXICO: Grupo Editorial Patria.

Bendovschi, A. (2015). Cyber-Attacks – Trends, Patterns and Security Countermeasures. *Procedia Economics and Finance* , 24–31.

Borja-Robalino, R., Monleón-Getino, A., & Rodellar, J. (2020). Estandarización de métricas de rendimiento para clasificadores Machine y Deep Learning. *Revista Ibérica de Sistemas e Tecnologias de Información*, 184-196.

Cascavilla, G., Tamburri, D. A., & Heuvel, W.-J. D. (2021). Cybercrime threat intelligence: A systematic multi-vocal literature review. *Computers & Security*, https://doi.org/10.1016/j.cose.2021.102258.

Chavan, S., Inamdar, A., Dorle, A., Kulkarni, S., & Wu, X.-W. (2020). Phishing Detection: Malicious and Benign Websites Classification Using Machine Learning Techniques. *Algorithms for Intelligent Systems. Series Editors: Bansal, Jagdish Chand, Deep, Kusum, Nagar, Atulya K*, 437-445.

Christou, O., Pitropakis, N., Papadopoulos, P., McKeown , S., & Buchanan, W. J. (2020). *Phishing URL Detection Through Top-Level Domain Analysis: A Descriptive Approach.* Edinburgh: School of Computing, Edinburgh Napier University.

Clarin. (18 de Setiembre de 2020). *Clarin*. Obtenido de https://www.clarin.com

De la Hoz, E. M. (2016). *Mapas auto-organizativos probabilísticos y análisis en componentes de conexiones para la deteccion de anomalías en redes de computadoras.* Granada: Universidad de Granada.

De la Hoz, E., De la Hoz, E. M., Ortíz, A., & Ortega, J. (2012). *Modelo de detección de intrusiones en sistemas de red, realizando selección de caracterísitcas con FDR y entrenamiento y clasificación con SOM.* Barranquilla: Corporación Universidad de la Costa.

Denning, D. E. (1987). *An Intrusion Detection Model.* IEEE.

Divindat . (2021). *División de Investigación de Delitos de Alta Tecnología de la Policía.* Lima: El Peruano.

Escrivá Gascó, G., Romero Serrano, R. M., & Ramada, D. J. (2013). *Seguridad Informática.* Madrid: Macmillan Iberia, S.A.

FBI. (11 de Julio de 2019). *Federal Bureau of Investigation - U.S. Department of Justice*. Obtenido de https://www.fbi.gov/investigate

Feng, F., Zhou, Q., Shen, Z., Yang, X., Han, L., & Wang, J. (2018). The application of a novel neural network in the detection of phishing websites. *Springer-Verlag GmbH Germany*.

Gartner. (2013). *Threat intelligence: What is it, and how can it protect you fron today advanced cyber-attacks.* Stamford: Gartner, Inc.

Gori, M. (2018). Machine Learning: A Constraint-Based Approach. *Morgan Kaufman*, https://doi.org/10.1016/C2015-0-00237-4.

Harinahalli, L., & BoreGowda, G. (2020). Phishing website detection based on effective machine learning approach. *Journal of Cyber Security Technology*, Pages 1-14, doi: https://doi.org/10.1080/23742917.2020.1813396.

Heberlein, T. (1995). *Network Security Monitor.* California: UIniversidad de California.

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, M. (2014). *Metodología de la Investigación* (6ta Edición ed.). México D.F.: Mc Graw Hill Education.

Hurwitz, J., & Kirsch, D. (2018). *Machine Learning For Dummies, IBM Limited Edition.* United States of America: John Wiley & Sons, Inc.

ISACA. (2015). *Cybersecurity Fundamentals. Study Guide.* Rolling Meadown, USA: ISACA.

ISO. (1 de Agosto de 2017). *International Organization for Standardization*. Obtenido de https://www.iso.org/search/x/query/27032

ISO27000. (13 de Julio de 2017). *Sistema de Gestión de la Seguridad de la Información*. Obtenido de http://www.iso27000.es/

Jain, A. K., & Gupta, B. B. (2017). Towards detection of phishing websites on client-side using machine learning based approach. *Telecommun Syst 68*, 687–700 doi: https://doi.org/10.1007/s11235-017-0414-0.

Jain, A. K., & Gupta, B. B. (2018). PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning. *Advances in Intelligent Systems and Computing, vol 729. Springer, Singapore*, Pag 467-474. https://doi.org/10.1007/978-981-10-8536-9_44.

Jakobsson, M., & Myers, S. (2006). Phishing and Countermeasures: Understanding the Increasing Problem of Electronic Identity Theft. . *Wiley-Interscience*.

Kulkarni, A. D., & Brown, L. L. (2019). Phishing Websites Detection using Machine Learning. *Computer Science Faculty Publications and Presentations*, Paper 20.

Kumar, R., Gunasekaran, Nivetha, R., Sangeetha , P. K., Shanthini, G., & Vignesh., A. S. (2019). Url Phishing data analysis and detecting phishing attacks using Machine Learning in NLP. *International Journal of Engineering Applied Sciences and Technology*, 23-31.

Lakshmi, L., Reddy, M., Santhaiah , C., & Reddy , U. J. (2021). Smart Phishing Detection in Web Pages using Supervised Deep Learning Classification and Optimization Technique ADAM. *Wireless Personal Communications*, pages 3549–3564.

Larrieu-Let, E. (2015). *Ciberseguridad.* Montevideo, Uruguay: ISACA, Montevideo Chapter .

Marsh & Microsoft. (2020). *Estado del Riesgo Cibernético en Latinoamérica en tiempos de COVID-19.* Nueva York: Marsh LLC.

Martínez Puentes, J. (2011). *Sistema Inteligente de Detección de Intrusiones.* Madrid: Universidad Complutense de Madrid.

Medina, M., & Molist, M. (2017). *Ciberseguridad. Tendencias 2017.* Valencia, España.: Universidad Internacional de Valencia.

Merwe, A., Loock, M., & Dabrawski, M. (2005). Characteristics and responsibilities involved in a Phishing attack. *WISICT '05: Proceedings of the 4th international symposium on Information and communication technologies*, 249–254.

Modi, H. (2019). *Cybercrime's Innovation Machine.* California: NETSCOUT Threat Intelligence.

Mohammad, R. M., Thabtah, F., & McCluskey, L. (2014). Predicting phishing websites based on self-structuring neural network. *Neural Comput & Applic*, 443–458 doi: https://doi.org/10.1007/s00521-013-1490-z.

Mohammad, R. M., Thabtah, F., & McCluskey, L. (2015). *Phishing Websites Features.* Huddersfield: School of Computing and Engineering, University of Huddersfield.

Mohammed, M., Khan, M. B., & Mohammed, E. B. (2017). *Machine Learning: Algorithms and Applications.* London: Taylor & Francis Group.

Mueller , A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python.* Sebastopol: O'Reilly Media, Inc.

Niakanlahiji, A., Chu, B.-T., & Al-Shaer, E. (2018). PhishMon: A Machine Lea Ollmann ing Framework for Detecting Phishing Webpages. *IEEE International Conference on Intelligence and Security Informatics (ISI)*, pp. 220-225, doi: 10.1109/ISI.2018.8587410.

NIST. (2018). *Framework for Improving Critical Infrastructure Cybersecurity.* Gaithersburg, Maryland: National Institute of Standards and Technology.

Ollmann, G. (2004). The Phishing Guide. *NGSSoftware Insight Security Research*, 1-42.

Opara, C., Wei, B., & Chen, Y. (2020). HTMLPhish: Enabling PhishingWeb Page Detection by Applying Deep Learning Techniques on HTML Analysis. *IJCNN*, DOI:10.1109/IJCNN48605.2020.9207707.

Patil, V., Thakkar, P., Shah, C., Bhat, T., & Godse, S. P. (2018). Detection and Prevention of Phishing Websites Using Machine Learning Approach. *Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697412.

PECERT. (2021). *Alerta Integrada de Seguridad Digital.* Lima: Centro Nacional de Seguridad Digital .

Rami , M. M., Fadi , T., & Lee , M. (2014). Predicting phishing websites based on self-structuring neural network. *Neural Computing and Applications* , 443–458.

Real Academia Española. (5 de Julio de 2017). *Diccionario de la lengua española*. Obtenido de http://dle.rae.es

Sánchez, M. (6 de Julio de 2011). *Infraestructuras Críticas y Ciberseguridad*. Obtenido de https://manuelsanchez.com/2011/07/06/infraestructuras-criticas-y-ciberseguridad/

Sandoval, L. J. (2018). Algoritmos de Aprendizaje Automático para análisis y prediccion de datos. *Revista Tecnologica*, Pag 36-40.

Sarkar, D., Bali, R., & Sharma, T. (2018). *Practical Machine Learning with python.* New York: Spring Street. doi:https://doi.org/10.1007/978-1-4842-3207-1.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *Computer Science*, doi: https://doi.org/10.1007/s42979-021-00592-x.

Scarfone, K., & Mell, P. (2007). *Guide to Intrusion Detection and Prevention Systems (IDPS). NIST Special Publication 800-94 .* Gaithersburg: National Institute of Standards and Technology.

Shahrivari, V., Darabi, M. M., & Izadi, M. (2020). *Phishing Detection Using Machine Learning Techniques.* New York: Cornell University.

Singh, R., & Sharma, T. (2019). Present Status of Distributed Denial of service (DDoS) Attacks in Internet World. *International Journal of Mathematical, Engineering and Management Sciences* , 1008-1017.

Sountharrajan, S., Nivashini, M., Shandilya, S. K., Suganya, E., Bazila , A., & Karthiga, M. (2019). Dynamic Recognition of Phishing URLs Using Deep Learning Techniques. *EAI/Springer Innovations in Communication and Computing*, 27-53.

Subasi, A. (2020). *Practical Machine Learning for data analysis using Phyton.* London: Elsevier Inc.

Tounsi, W., & Rais, H. (2018). A survey on technical threat intelligence in the age of sophisticated cyber attacks. *Computers & Security*, Pages 212-233, doi: https://doi.org/10.1016/j.cose.2017.09.001.

Trend Micro. (2021). How to Reduce the Risk of Phishing and Ransomware. *Osterman Research*, 1-29.

Ubing , A. A., Binti Jasmi, S. K., Azween , A., Jhanjhi, N. Z., & Supramaniam, M. (2019). Phishing Website Detection: An Improved Accuracy through Feature Selection and Ensemble Learning. *International Journal of Advanced Computer Science and Applications*.

Vakili, M., Ghamsari, M., & Rezaei, M. (2020). Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification. *Cornell University*.

Wang, W., Zhang, F., Luo, X., & Zhang, S. (2019). PDRCNN: Precise Phishing Detection with Recurrent Convolutional Neural Networks. *Security and Communication Networks*, Paper 15.

Wei, B., Hamad, R., Yang, L., He, X., Wang, H., Gao, B., & Woo, W. (2019). *A Deep-Learning-Driven Light-Weight Phishing Detection Sensor.* Newcastle: Sensors Northumbria University.

Yang, L., Zhang, J., Wang, X., Li, Z., Li, Z., & He, Y. (2021). An improved ELM-based and data preprocessing integrated approach for phishing detection considering comprehensive features. *Expert Systems with Applications*, doi, https://doi.org/10.1016/j.eswa.2020.113863.

Yi, P., Guan, Y., Zou, F., Yao, Y., WeiWang, & Zhu, T. (2018). Web Phishing Detection Using a Deep Learning Framework. *Wireless Communications and Mobile Computing*, 9 paginas doi: https://doi.org/10.1155/2018/4678746.

Zabihimayvan, M., & Doran, D. (2019). Fuzzy Rough Set Feature Selection to Enhance Phishing Attack Detection. *IEEE International Conference on Fuzzy Systems*, pag. 1-6 doi: 10.1109/FUZZ-IEEE.2019.8858884.

Zamir, A., Khan, H. U., Iqbal, T., Yousaf, N., Aslam, F., Anjum, A., & Hamdani, M. (2020). Phishing web site detection using diverse machine learning algorithms. *Electron*, 65-80.

Anexos

Anexo 01: Matriz de consistencia

Anexo 02: Operacionalización de las variables.

Anexo 03: Instrumentos

Anexo 04: Validación de instrumentos por juicio de expertos

Anexo 05: Consentimiento Informado

Anexo 06: Aprobación del Informe de Tesis

UNIVERSIDAD SEÑOR DE SIPÁN

ANEXO N° 1 MATRIZ DE CONSISTENCIA

Manifestaciones del problema	• El 84% de profesionales en Estados Unidos indican que han sufrido al menos un tipo de incidente de seguridad y ponen a los ataques phishing y ransomware, como los tipos de ataques con mayor ocurrencia. (Trend Micro, 2021) • El 31% de empresas en Latinoamérica han percibido un aumento de los ciberataques en el 2020 y la principal amenaza de la ciberseguridad son los eventos relacionados al phishing. (Marsh & Microsoft, 2020) • El número de ataques phishing crecieron continuamente y hasta se duplicaron a lo largo del 2020, llegando en octubre a 225,304 ataques phishing. (APWG, 2021) • Los países de Latinoamérica que fueron más afectados con los ataques phishing durante el 2020 son las empresas de Brasil (26,4%), de Perú (22,8%) y México (12%). (Eset, 2021) • En el Perú se elevan el número de denuncias de ciberdelitos, siendo la modalidad más frecuente el phishing, que en el 2021 se duplicaron con respecto al 2020. (Divindat , 2021) • En el Perú, en de enero a septiembre del 2021, se han generado 259 alertas integradas de Seguridad Digital, y en la mayoría de las alertas se ha reportado a los ataques phishing. (PECERT, 2021) • Según (Trend Micro, 2021) el 50% de los profesionales en Estados Unidos consideran ineficiente las técnicas de hacer frente al phishing y al ransomware. • El 65% de envíos de phishing ingresan a las bandejas de los usuarios. (Trend Micro, 2021). • Además, el 65% de los usuarios hacen clic en los enlaces phishing. (Trend Micro, 2021).

	• Los sistemas de detección de phishing en el 2020 neutralizaron 434 898,635 phishing, sin embargo, representa una menor cantidad de phishing detectado que en el 2019. (Kaspersky, 2021)
Problema	Insuficiencias en el proceso de la ciberseguridad y la detección de phishing, **limitan el rendimiento en la detección de sitios web falsas.**
Causas que originan el Problema	• Insuficiente diagnóstico contextual en la detección de sitios web falsos, durante el proceso de ciberseguridad. • Limitados referentes teóricas detallados del proceso de ciberseguridad y la detección de phishing. • Limitadas referencias prácticas sistemáticas del proceso de ciberseguridad y la detección de phishing. • Insuficiente uso de modelos de aprendizaje automático, en el proceso de ciberseguridad y la detección de phishing. • Insuficientes sistemas de detección de phishing para el desarrollo del proceso de ciberseguridad y la detección de sitios web phishing.
Objeto de la Investigación	Proceso de la ciberseguridad y la detección de phishing
Objetivo General de la Investigación	Aplicar un sistema de detección de phishing, sustentada en un modelo de machine learning, para el rendimiento en la detección de sitios web falsos.
Objetivos específicos	• Caracterizar científicamente el proceso de la ciberseguridad, la detección de phishing y su dinámica. • Diagnosticar el estado actual del proceso de ciberseguridad y la detección de phishing. • Elaborar un modelo de machine learning, utilizando las características de la URL, información de la inteligencia de amenazas y técnicas de machine learning. • Elaborar un sistema de detección de phishing, basado en el modelo de machine learning. • Validar los resultados de la investigación.

Campo de la investigación	Dinámica del proceso de la ciberseguridad y la detección de phishing
Título de la Investigación	Modelo de machine learning en la detección de sitios web phishing.
Hipótesis	Si se aplica un sistema de detección de phishing, basado en un modelo de machine learning, que tenga en cuenta la información de las URL, la información de la inteligencia de amenazas y las técnicas de machine learning, se contribuye al rendimiento en la detección de sitios web falsos.
Variables	Variable independiente. • Sistema de detección de phishing, basado en un modelo de Machine Learning. Variable dependiente. • Rendimiento en la detección de sitios web falsos.

UNIVERSIDAD SEÑOR DE SIPÁN

ANEXO N° 2

OPERACIONALIZACIÓN DE VARIABLES

VARIABLE INDEPENDIENTE

VARIABLES	DIMENSIONES	DESCRIIPCIÓN
V. INDEPENDIENTE **Sistema de detección de phishing, basado en un modelo de Machine Learning.**	**Introducción-Fundamentación**.	Se establece el contexto y ubicación de la problemática a resolver. Ideas y puntos de partida que fundamentan la estrategia. Se indica la teoría en que se fundamenta el aporte propuesto.
	II. Diagnóstico-	Indica el estado real del objeto y evidencia el problema en torno al cual gira y se desarrolla la estrategia, protocolo, o programa, según el aporte practico a desarrollar.
	Planteamiento del objetivo general.	Se desarrolla el objetivo general del aporte práctico. Se debe tener en cuenta que no es el de la investigación.
	Planeación estratégica	- Se definen metas u objetivos a corto y mediano plazo que permiten la transformación del objeto desde su estado real hasta el estado deseado. Planificación por etapas de las acciones, recursos, medios y métodos que corresponden a estos objetivos. Se debe tener en cuenta las dimensiones de la operacionalización de la variable dependiente.
	Instrumentación	Explicar cómo se aplicará, bajo qué condiciones, durante qué tiempo, responsables, participantes.
	Evaluación	Definición de los logros obstáculos que se han ido venciendo, valoración de la aproximación lograda al estado deseado

UNIVERSIDAD SEÑOR DE SIPÁN

OPERACIONALIZACIÓN DE VARIABLES

VARIABLE DEPENDIENTE

VARIABLES	DIMENSIONES	INDICADORES	TÉCNICAS O INSTRUMENTOS.	FUENTES DE VERIFICACIÓN
V. DEPENDIENTE Rendimiento en la detección de sitios web falsos. Clasificar correctamente los sitios web en sitios web legítimas o en sitios web phishing.	**Rendimiento**	• Verdaderos positivos (VP). Cantidad de pruebas clasificadas correctamente un sitio web falso, como sitio web falso. • Verdaderos negativos (VN). Cantidad de pruebas clasificadas correctamente un sitio web real, como sitio web real. • Falsos positivos (FP). Cantidad de pruebas clasificadas erróneamente un sitio web real, como sitio web falso. • Falsos Negativos (FN). Cantidad de pruebas clasificadas erróneamente un sitio web falso, como sitio web real. • Acurracy (Exactitud). Cantidad de pruebas positivas que fueron correctamente clasificadas. Se representa por la proporción de los VP y FN, entre el total de casos examinados (VP + VN + FP + FN) • Recall (Sensibilidad). Tasa de Verdaderos Positivos, es la capacidad de detectar correctamente los sitios web falsos, Se representa por la proporción de VP entre (VP + FN). • Especificity (Especificidad). Tasa de Verdaderos negativos, es la capacidad de detectar los sitios web reales, se representa por la proporción de VN entre (VN + FP). • Precision (Precisión). Porcentaje de casos positivos detectados. Se representa por la proporción de VP entre todos los resultados positivos (VP + FP).	• Observación. • Análisis documentario.	• Herramientas de Software

UNIVERSIDAD SEÑOR DE SIPÁN

ANEXO N° 3:

FICHA DE OBSERVACION 1

Indicaciones.

1. La presente ficha tiene como objetivo registrar las características de los sitios web legítimos y sitios web phishing.
2. Por fines de confidencialidad y protección de los datos personales, en ningún caso se debe consignar la dirección del sitio web.
3. Se registra las características del sitio web, basado en la información de la barra de direcciones, Código Fuente y de inteligencia de amenazas.

DIMENSIONES	CARACTERISTICAS DEL SITIO WEB	CATEGORIA		
Sitio Web	**¿La clase del sitio web es phishing?**	SI ( )		NO ( )
Contenido URL, en la barra de direcciones	1. ¿Se muestra una dirección IP en la barra de direcciones?	SI ( )		NO ( )
	2. ¿El tamaño de la URL, por sus caracteres?. Corto (< 54), medio (entre 54 y 75) y largo (> 75)	Corto ( )	Medio ( )	Largo ( )
	3. ¿La dirección URL, está utilizando el servicio de TinyURL (acortamiento)?	SI ( )		NO ( )
	4. ¿Se usa el símbolo "@", dentro de la URL?	SI ( )		NO ( )
	5. ¿Se usa el "//", después de la posición 7 dentro de la URL?	SI ( )		NO ( )
	6. ¿Se usa el símbolo "-" como parte del dominio o subdominio de la URL?	SI ( )		NO ( )
	7. ¿Cuál es la cantidad de puntos dentro de la URL?. Bajo (<=3), medio (4) y alto (>4).	Bajo ( )	Medio ( )	Alto ( )
	8. ¿Usa certificado SSL el sitio web? .	SI ( )		NO ( )
	9. ¿El registro del dominio es de mas de un año de antigüedad?	SI ( )		NO ( )
	10. ¿El icono de la URL, carga de una dirección diferente a la URL?	SI ( )		NO ( )
	11. ¿Hay puertos abiertos en el sitio web?	SI ( )		NO ( )
	12. ¿Se observa el uso de https, después de las "//"?	SI ( )		NO ( )

Basado en Anomalías y Análisis del código fuente	13. ¿Porcentaje de contenido que carga de otros dominios? Bajo (<22%), Medio (entre 22 y 61) o Alto (> de 61%).	Bajo ( )	Medio ( )	Alto ( )
	14. ¿Porcentaje de referencias a otros dominios que utiliza la dirección url? Bajo (<31%), Medio (entre 31 y 67%) o Alto (>67%).	Bajo ( )	Medio ( )	Alto ( )
	15. ¿Porcentaje de tags a otros dominios que utiliza la dirección url? Bajo (<17%), Medio (entre 17 y 81%) o Alto (>81%).	Bajo ( )	Medio ( )	Alto ( )
	16. ¿Los SFH que contienen una cadena vacía o “about: ¿¿blank”, a otro dominio o no?	SI ( )	NO ( )	
	17. ¿Se usa "mail()\" o\"mailto:\"?	SI ( )	NO ( )	
	18. ¿El hostname está incluido en la URL?	SI ( )	NO ( )	
	19. ¿Número de veces que se ha redirigido un sitio web?. Bajo (<=1), medio (entre 2 y 4) y alto (>4).	Bajo ( )	Medio ( )	Alto ( )
	20. ¿hay cambios en la barra de estado, con la función de "Mouse Over"?	SI ( )	NO ( )	
	21. ¿Está desactivad el clik - derecho?	SI ( )	NO ( )	
	22. ¿El sitio web usa ventana emergente?	SI ( )	NO ( )	
	23. ¿El sitio web, hace uso de Iframe?	SI ( )	NO ( )	
Informacion de Inteligencia de Amenazas	1. ¿La antigüedad del registro del dominio en WHOIS es mas de 6 meses?	SI ( )	NO ( )	
	2. ¿El dominio cuenta con un registro DNS en WHOIS?	SI ( )	NO ( )	
	3. ¿Cuál es la popularidad del sitio web en la base de datos Alexa, por el número de visita?. Bajo (No está en Alexa), Medio (<= 100000) o Alto (> a 100000).	Bajo ( )	Medio ( )	Alto ( )
	4. ¿La importancia de la web según PageRank es mayor a 0.2?	SI ( )	NO ( )	
	5. ¿El sitio web está en Google Index?	SI ( )	NO ( )	
	6. ¿Cuál es el número de enlaces que apuntan al sitio web?. Bajo (0), medio (entre 1 y 2) y alto (>2).	Bajo ( )	Medio ( )	Alto ( )
	7. ¿El dominio del sitio web, está registrado en los reportes de los top paginas phishing?	SI ( )	NO ( )	

UNIVERSIDAD SEÑOR DE SIPÁN

ANEXO N° 4:

VALIDACION DE INSTRUMENTOS POR JUICIO DE EXPERTOS

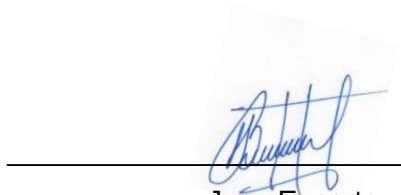
1. NOMBRE DEL JUEZ		BUSTAMANTE QUINTANA, PEPE HUMBERTO
2.	PROFESIÓN	INGENIERO DE SISTEMAS
	ESPECIALIDAD	GESTION DE TECNOLOGÍAS DE LA INFORMACION
	GRADO ACADÉMICO	DOCTOR
	EXPERIENCIA PROFESIONAL (AÑOS)	17 AÑOS
	CARGO	SECRETARIO ACADEMICO DE LA ESCUALA DE POSGRADO
Título de la Investigación: **Modelo de Machine Learning en la detección de sitios web phishing.**		
3. DATOS DEL TESISTA		
3.1	NOMBRES Y APELLIDOS	VILLEGAS, CUBAS, JUAN ELIAS
3.2	PROGRAMA DE POSTGRADO	DOCTORADO EN CIENCIAS DE LA COMPUTACION Y SISTEMAS.
4. INSTRUMENTO EVALUADO		1. Entrevista ( ) 2. Cuestionario ( ) 3. Ficha de Observación (X)
5. OBJETIVOS DEL INSTRUMENTO		<u>GENERAL:</u> • Registrar las características de los sitios web legítimos y sitios web phishing. <u>ESPECÍFICOS</u> • Procesar los datos de las características de los sitios web phishing y sitios web legitimas. • Recoger información de los sitios web con fines académicos, excluyendo datos confidenciales.

A continuación, se le presentan los indicadores en forma de preguntas o propuestas para que Ud. los evalúe marcando con un aspa (x) en “A” si está de ACUERDO o en “D” si está en DESACUERDO, SI ESTÁ EN DESACUERDO POR FAVOR ESPECIFIQUE SUS SUGERENCIAS		
N	DETALLE DE LOS ITEMS DEL INSTRUMENTO	
01	Pregunta del instrumento ¿Se muestra una dirección IP en la barra de direcciones? Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:
02	Pregunta del instrumento ¿El tamaño de la URL, por sus caracteres? Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:
03	Pregunta del instrumento ¿La dirección URL, está utilizando el servicio de TinyURL (acortamiento)? Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:
04	Pregunta del instrumento ¿Se usa el símbolo "@", dentro de la URL? Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:
05	Pregunta del instrumento ¿Se usa el "//", después de la posición 7 dentro de la URL? Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:
06	Pregunta del instrumento ¿Se usa el símbolo "-" como parte del dominio o subdominio de la URL? Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:
07	Pregunta del instrumento ¿Cuál es la cantidad de puntos dentro de la URL? Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:
08	Pregunta del instrumento ¿Usa certificado SSL el sitio web? . Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:

09	Pregunta del instrumento ¿El registro del dominio es de más de un año de antigüedad? Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:
10	Pregunta del instrumento ¿El icono de la URL, carga de una dirección diferente a la URL? Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:
11	Pregunta del instrumento ¿Hay puertos abiertos en el sitio web? Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:
12	Pregunta del instrumento ¿Se observa el uso de https, después de las "//"? Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:
13	Pregunta del instrumento ¿Porcentaje de contenido que carga de otros dominios? Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:
14	Pregunta del instrumento ¿Porcentaje de referencias a otros dominios que utiliza la dirección url? Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:
15	Pregunta del instrumento ¿Porcentaje de tags a otros dominios que utiliza la dirección url? Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:
16	Pregunta del instrumento ¿Los SFH que contienen una cadena vacía o “about: ¿¿blank”, a otro dominio o no?? Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:
17	Pregunta del instrumento ¿Se usa "mail()\" o\"mailto:\"? Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:

18	Pregunta del instrumento ¿El hostname está incluido en la URL? Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:
19	Pregunta del instrumento ¿Número de veces que se ha redirigido un sitio web?. Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:
20	Pregunta del instrumento ¿hay cambios en la barra de estado, con la función de "Mouse Over"? Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:
21	Pregunta del instrumento ¿Está desactivad el clik - derecho? Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:
22	Pregunta del instrumento ¿El sitio web usa ventana emergente? Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:
23	Pregunta del instrumento ¿El sitio web, hace uso de Iframe? Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:
24	Pregunta del instrumento ¿La antigüedad del registro del dominio en WHOIS es mas de 6 meses? Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:
25	Pregunta del instrumento ¿El dominio cuenta con un registro DNS en WHOIS? Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:
26	Pregunta del instrumento ¿Cuál es la popularidad del sitio web en la base de datos Alexa, por el número de visita?. Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:

27	Pregunta del instrumento ¿La importancia de la web según PageRank es mayor a 0.2? Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:
28	Pregunta del instrumento ¿El sitio web está en Google Index? Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:
29	Pregunta del instrumento ¿Cuál es el número de enlaces que apuntan al sitio web?. Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:
30	Pregunta del instrumento ¿El dominio del sitio web, está registrado en los reportes de los top paginas phishing? Escala de medición: Nominal	A(X) D ( ) SUGERENCIAS:
	PROMEDIO OBTENIDO:	A(X) D ( )
6	COMENTARIOS GENERALES **El instrumento propuesto reúne los requerimiento suficientes y necesarios para ser considerado válido.**	
7	OBSERVACIONES	



Juez Experto

Colegiatura CIP 210445

UNIVERSIDAD SEÑOR DE SIPÁN

ANEXO N° 5

CONSENTIMIENTO INFORMADO

Institución:

MUNICIPALIDAD PROVINCIAL DE CHICLAYO

GERENCIA DE TECNOLOGIAS DE LA INFORMACION Y ESTADISTICA

Investigador:

VILLEGAS CUBAS, JUAN ELIAS

Título:

MODELO DE MACHINE LEARNING EN LA DETECCIÓN DE SITIOS WEB PHISHING.

Yo, *ALBERTO ENRIQUE SAMILLAN AYALA* identificado con DNI N° 18134651, DECLARO:

Haber sido informado (a) de forma clara, precisa y suficiente sobre los fines y objetivos que busca la presente investigación **"Modelo de Machine Learning en la detección de sitios web phishing"**, así como en qué consiste mi participación.

Estos datos que yo otorgue serán tratados y custodiados con respeto a la intimidad, manteniendo el anonimato de la información y la protección de datos desde los principios éticos de la investigación científica. Sobre estos datos se asisten los derechos de acceso, rectificación o cancelación que podré ejercitar mediante solicitud ante el investigador responsable. Al término de la investigación, seré informado de los resultados que se obtengan.

Por lo expuesto otorgo MI CONSENTIMIENTO para que se realice la recolección de datos que permita contribuir con los objetivos de la investigación:

Objetivo general de la investigación:

Aplicar un sistema de detección de phishing, sustentada en un modelo de machine learning, para el rendimiento en la detección de sitios web falsos.

Objetivos específicos:

- Caracterizar científicamente el proceso de la ciberseguridad, la detección automática de phishing y su dinámica.
- Diagnosticar el estado actual del proceso de ciberseguridad y la detección automática de phishing.
- Elaborar un modelo de machine learning, utilizando las características de la URL, información de la inteligencia de amenazas y técnicas de machine learning.
- Elaborar un sistema de detección de phishing, basado en el modelo de machine learning.
- Validar los resultados de la investigación.
- Ejemplificar la aplicación del sistema en la predicción de sitios web phishing.

Chiclayo, 2 de AGOSTO del 2021

Dr. Ing, ALBERTO ENRIQUE SAMILLAN AYALA
Gerente de Tecnologías de la Información y Estadística
***DNI:* 18134651**

UNIVERSIDAD SEÑOR DE SIPÁN

ANEXO N° 6

APROBACIÓN DEL INFORME DE TESIS

El Docente:
DR. JUAN CARLOS CALLEJAS TORRES.

De la Asignatura:

SEMINARIO DE INVESTIGACIÓN VI: INFORME DE TESIS

APRUEBA:

*El Informe de Tesis: "***MODELO DE MACHINE LEARNING EN LA DETECCIÓN DE SITIOS WEB PHISHING"**

Presentado por:

Mg. JUAN ELIAS, VILLEGAS CUBAS.

Chiclayo, 10 de noviembre del 2021.

_____
Dr. JUAN CARLOS CALLEJAS TORRES