UNIVERSIDAD POLITÉCNICA SALESIANA

SEDE QUITO

CARRERA DE COMPUTACIÓN

DETECCIÓN Y MITIGACIÓN DEL PHISHING EN CORREOS ELECTRÓNICOS MEDIANTE APRENDIZAJE AUTOMÁTICO Y ANÁLISIS DE DATOS PREEXISTENTES

Trabajo de titulación previo a la obtención del
Título de Ingeniero e Ingeniera en Ciencias de la Computación

AUTORES: ALEX SEBASTIAN PANCHI CANCHIGNIA
DENISSE ANAHI TUPIZA TUPIZA

TUTOR: JOSÉ LUIS AGUAYO MORALES

Quito-Ecuador

2026

CERTIFICADO DE RESPONSABILIDAD Y AUTORÍA DEL TRABAJO DE TITULACIÓN

Nosotros, Alex Sebastian Panchi Canchignia con documento de identificación N° 1754494910 y Denisse Anahí Tupiza Tupiza con documento de identificación N° 1750843862: manifestamos que:

Somos los autores y responsables del presente trabajo; y, autorizamos a que sin fines de lucro la Universidad Politécnica Salesiana pueda usar, difundir, reproducir o publicar de manera total o parcial el presente trabajo de titulación.

Quito, 25 de febrero del 2026

Atentamente,

Alex Sebastian Panchi Canchignia
1754494910

Denisse Anahí Tupiza Tupiza
1750843862

CERTIFICADO DE CESIÓN DE DERECHOS DE AUTOR DEL TRABAJO DE TITULACIÓN A LA UNIVERSIDAD POLITÉCNICA SALESIANA

Nosotros, Alex Sebastian Panchi Canchignia con documento de identificación N° 1754494910 y Denisse Anahí Tupiza Tupiza con documento de identificación N° 1750843862; expresamos nuestra voluntad y por medio del presente documento cedemos a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud de que somos autores del Proyecto Técnico: "Detección y mitigación del phishing en correos electrónicos mediante aprendizaje automático y análisis de datos preexistentes", el cual ha sido desarrollado para optar por el título de: Ingenieros en Ciencias de la Computación, en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En concordancia con lo manifestado, suscribimos este documento en el momento que hacemos la entrega del trabajo final en formato digital a la Biblioteca de la Universidad Politécnica Salesiana.

Quito, 25 de febrero del 2026

Atentamente,

Alex Sebastian Panchi Canchignia
1754494910

Denisse Anahí Tupiza Tupiza
1750843862

CERTIFICADO DE DIRECCIÓN DEL TRABAJO DE TITULACIÓN

Yo, José Luis Aguayo Morales con documento de identificación N° 1709562597, docente de la Universidad Politécnica Salesiana, declaro que bajo mi tutoría fue desarrollado el trabajo de titulación: DETECCIÓN Y MITIGACIÓN DEL PHISHING EN CORREOS ELECTRÓNICOS MEDIANTE APRENDIZAJE AUTOMÁTICO Y ANÁLISIS DE DATOS PREEXISTENTES, realizado por Alex Sebastian Panchi Canchignia con documento de identificación N° 1754494910 y por Denisse Anahí Tupiza Tupiza con documento de identificación N° 1750843862, obteniendo como resultado final el trabajo de titulación bajo la opción de Proyecto Técnico que cumple con todos los requisitos determinados por la Universidad Politécnica Salesiana.

Quito, 25 de febrero del 2026

Atentamente,

_____

Ing. José Luis Aguayo Morales, MSc.
1709562597

DEDICATORIA

Con sincero afecto, dedico este trabajo de titulación, culminación de mi esfuerzo académico, a los pilares fundamentales de mi vida:

A mi padre y madre, por su cariño y amor incondicional, el constante aparo y los inconyables sacrificios que hicieron para que este logro fuera posible.

A mis hermanos, por su compañía, consejo y por ser un ejemplo de perseverancia.

A mis sobrinos, cuya alegría renovó mi motivación y me recordó la importancia de aspirar a un futuro mejor.

Alex Sebastian Panchi Canchignia

Dedico este logro a mis personas favoritas, mis padres, quienes han sido mi mayor inspiración para esforzarme cada día y llenan mi corazón de gratitud. Por su incansable esfuerzo, por el apoyo incondicional que siempre he recibido, por demostrarme cómo mantenerme firme ante cualquier dificultad y por enseñarme a confiar en que siempre existe una forma de vencer los retos.

Por ser mis pilares fundamentales, por inculcarme principios valiosos que me han orientado en cada etapa de mi vida y han dado forma a mi existencia. Sin su fe inquebrantable, su sacrificio sin descanso y sus palabras de ánimo en las situaciones más complicadas, este trayecto no habría sido posible.

Denisse Anahi Tupiza Tupiza

AGRADECIEMIENTOS

Mi agradecimiento a mi querida familia, por el apoyo emocional de ellos, con cada segundo de su amabilidad, y por creer en mí durante todo este trayecto, especialmente en los últimos meses de mí de mi camino profesional.

Mis sinceros agradecimientos van a la UPS, mi alma mater, por una educación de calidad con valores humanos y profesionales que llevo conmigo, junto con los recursos necesarios y oportunidades que he ganado para formar mi identidad profesional.

La gratitud que extiendo a mis queridos docentes, por su apoyo, su orientación y su determinación para enseñar y transmitir el conocimiento para realizar este último trabajo de mi formación profesional.

Un agradecimiento especial a mi director de tesis, por su guía experta, sus observaciones precisas, su constante disponibilidad y su paciencia a lo largo del proceso de investigación y desarrollo.

A mis compañeros de estudio, que compartieron la misma experiencia universitaria y con quienes tuve esos debates, apoyo y tiempo para desahogarme les agradezco por su camaradería en todo este camino.

A mi compañera de tesis, por todo su esfuerzo inquebrantable, espíritu cooperativo sobre todo para la realizar este proyecto. Fue un privilegio trabajar junto a ella.

Alex Sebastian Panchi Canchignia

AGRADECIEMIENTOS

Quiero transmitir mi gratitud más profunda a Dios, quien ha sido mi fuente de sabiduría, fortaleza y orientación a lo largo de este camino académico. Gracias a Él he podido comprender, aprender y perseverar, incluso cuando esta meta parecía inalcanzable. Su bendición ha sido pilar que me ha permitido llegar a este importante logro en mi vida.

A mis amigos, tanto a los que estuvieron conmigo desde el principio de esta travesía como a aquellos que conocí a lo largo del camino. Gracias por enseñarme lo que realmente significa la amistad. Compartir con ustedes ha sido una experiencia algo mucho más enriquecedor que solo un aprendizaje académico. Juntos hemos reído, nos hemos apoyado y hemos descubierto la importancia de disfrutar cada etapa del proceso; su compañía fue, sin duda, lo que lo hizo inolvidable.

A mis hermanos, quienes son mi mayor motivación, mi alegría diaria y la razón por la cual sigo esforzándome por ser mejor cada día. Les agradezco por su apoyo incondicional, por mantenerse siempre a mi lado y por su ejemplo de superación. Han sido mi refugio en los momentos difíciles y una fuerza constante que me ha impulsado a seguir adelante.

Y, desde el fondo de mi corazón, mi agradecimiento eterno es para mis padres. Su amor incondicional, sus esfuerzos incansables y los innumerables sacrificios que han hecho han sido la base sobre la que he construido este sueño. Garcias por iluminar mis decisiones con su guía, por ser mi roca en mis tiempos difíciles y por llenarme de valor cuando lo necesitaba. Este logro no es solamente mío, también es suyo, porque sus enseñanzas, su aliento contante y su confianza en mí me han inspirado más de lo que puedo expresar.

Gracias por estar ahí en mis días más agotadores, por ser mi fortaleza y por celebrar conmigo cada pequeño paso hacia adelante. Todo lo que he alcanzado está impregnado de su amor, su paciencia infinita y su fe en mí. Con todo mi cariño, este logro les pertenece tanto como a mí.

Denisse Anahi Tupiza Tupiza

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

Institución académica es organización muy importante, si el correo electrónico institucional es vulnerado, va a riesgo la confidencialidad del correo universal y acceso a muchos recursos. Este proyecto procura un modelo de detección de correos fraudulentos por medio de técnicas de aprendizaje automático (M_L) supervisado, procesamiento del lenguaje natural sobre datos previos. Se hace una revisión del contenido textual, con pretratamiento, extracción de caracteres de entrada y el entrenamiento de modelos de clasificación binaria, la instancia establece el umbral para distinguir correos legítimos de peligrosos. El trabajo se realiza en un ambiente académico en la Carrera de Ciencias Computacionales de la UPS. Se encuentra en fase de investigación, aunque el modelo no se lleva a cabo en sistemas reales ni en tiempo real, a modo de referente se constata su efectividad mediante métricas como precisión, recall y F1-score. El presente proyecto de investigación ha constatado –en entornos educativos– la viabilidad de utilizar el aprendizaje automático y el análisis textual para detectar phishing, sentando las bases para futuras investigaciones y elevando el nivel de conciencia en materia seguridad informática en la carrera de Ciencias Computacionales.

Palabras clave: phishing, ciberseguridad, aprendizaje automático; procesamiento de lenguaje natural; clasificación de texto.

Abstract

An academic institution is a very important organization; if its institutional email is compromised, the confidentiality of the centralized email system and access to many resources are at risk. This project seeks to develop a model for detecting fraudulent emails using supervised machine learning (M_L) techniques and natural language processing applied to pre-existing data. The process involves reviewing the textual content, including preprocessing, extracting input characters, and training binary classification models. This process establishes the threshold for distinguishing legitimate emails from malicious ones. The work is being carried out in an academic setting at the Carried of Computer Science of the UPS. It is currently in the research phase, and although the model is not yet implemented in real-world systems or in real time, its effectiveness is being assessed using metrics such as accuracy, recall, and F1 score. This research project has confirmed—in educational environments—the feasibility of using machine learning and textual analysis to detect phishing, laying the groundwork for future research and raising awareness of cybersecurity within the Computer Science program.

Keywords: phishing, cybersecurity, machine learning; natural language processing; text classification

CAPÍTULO I

ANTECEDENTES Y GENERALIDADES

1.1 Introducción

En nuestros días, es prácticamente imposible imaginar el funcionamiento de cualquier organización, independientemente de si es sector público como en el privado, sin el uso intensivo del correo electrónico. Se ha consolidado como la columna vertebral de la comunicación institucional. La dependencia es alta en cuanto a lo digital y trae consigo una exposición masiva a amenazas cibernéticas, en la que destaca el phishing por su efectividad. Dado que el ataque se trata de generar vulnerabilidades en distintos tipos de cuentas humanas a través de la ingeniería social, pues tiende a tener mutaciones tan intensas que los sistemas típicos de detección en ocasiones no son efectivos.

En la computación en entornos universitarios, el email institucional se utiliza para actividades académicas, coordinación administrativa y entrada a recursos a plataformas. Es por esto por lo que, al ser objetivos directos, debido a que, en caso de comprometer una cuenta, se puede lograr accesos no autorizados, fuga de información o manipular actividades.

Las T_M_L y el N_L_P son dos opciones que hacen posible, además de complementar la seguridad que brindan los mecanismos tradicionales, es decir, hacen posible el análisis de textos a una escala grande y en la detección de patrones que filtros estáticos no llegan a detener.

Este proyecto se concentra en el análisis y clasificación de correos electrónicos por medio de modelos basados en información pública. Por lo tanto, se articula y vincula con uno de los diseños ya utilizados en el Trabajo de Titulación de Computación: aplicar T_M_L y N_L_P en los ámbitos académicos.

El uso masivo del correo electrónico institucional en la universidad, específicamente en Computación, para la realización de actividades académicas, la coordinación de procesos administrativos, el acceso a plataformas tecnológicas y la administración de laboratorios, hace de éste un recurso tentativo para el ataque de phishing. Ello debido a que, si en caso de que una cuenta

se ha hackeada, puede ocasionar accesos no autorizados, fuga de información o frenar las actividades académicas.

1.2 Problema de estudio

1.2.1 Antecedentes

El phishing ha evolucionado hacia comunicaciones cada vez más sofisticadas que simulan ser mensajes oficiales, Logrando engañar incluso a los usuarios más precavidos. Debido a esta complejidad, los filtros tradicionales basados en firmas y listas negras han perdido eficacia. Por ello, es necesario contar con modelos dinámicos que empleen aprendizaje automático (M_L) y análisis de contenido para detectar estos mensajes fraudulentos. Almomani et al. (2013) respalda el uso de estos enfoques, destacando su capacidad para descubrir patrones de fraude ocultos en datos históricos que no son evidentes con métodos clásicos.

A partir de los años noventa, los estafadores han mejorado mucho en la forma en que envían correos electrónicos y en cómo engañan a la gente. Hoy en día, hay correos electrónicos y páginas web de estafa que están completamente personalizadas. Los mensajes que envían ahora se ven muy similares a los mensajes legítimos, lo que hace que sea difícil detectarlos. El hecho de que sean muy personalizados y utilicen información real hace que parezcan más creíbles.

Los métodos convencionales no logran detectar modificaciones avanzadas ni casos de spoofing avanzado, lo cual refuerza la importancia de explorar modelos de M_L capaces de ajustarse al progreso de las amenazas.

1.2.2 Importancia

Diversas investigaciones señalan que las instituciones educativas suelen ser un objetivo frecuente de ataques de phishing. Esto se debe, en parte, al uso constante del correo electrónico y a la gran cantidad de usuarios que interactúan a diario, entre estudiantes, docentes y personal administrativo. Además, la renovación continua de miembros dentro de la comunidad universitaria dificulta mantener niveles uniformes de concienciación en seguridad (Almomani et al., 2013).

En el caso de la UPS, en los laboratorios de la carrera de Computación se agrupan algunos de los recursos tecnológicos, entornos de desarrollo, repositorios académicos y cuentas

institucionales. Si alguno de estos elementos llegara a verse comprometido, el desarrollo normal de las actividades académicos podría verse seriamente afectado.

La magnitud del problema radica en que un incidente de phishing exitoso puede traducirse en robo de credenciales, acceso sin autorización a sistemas críticos, fuga de información relevante y, en general, afectaciones a la sostenibilidad de las labores educativas y administrativas. Por estas razones, resulta importante analizar e implementar métodos de detección y mitigación mas solidos dentro de este entorno.

1.2.3 Delimitación

Esta investigación de sitúa más allá de las fronteras físicas convencionales, enfocándose en el ciberespacio como su ámbito principal. El análisis se desarrolla en el campo de las comunicaciones electrónicas, debido a que los datos provienen de diversas fuentes digitales distribuidas a nivel global.

Para la realización de este trabajo se utilizó el conjunto de datos CEAS 2008, el cual contiene correos electrónicos clasificados como legítimos y fraudulentos.

Este dataset permite entrenar a los modelos de M_L para que puedan diferenciar entre mensajes genuinos y correos maliciosos.

1.3 Justificación

Detectar correos de phishing es importante para fortalecer la ciberseguridad en las universidades, especialmente porque el correo institucional es una herramienta de uso diario para docentes y personal administrativo. Este proyecto de titulación puede aportar beneficios tanto a estudiantes como a profesores y al personal de apoyo de la carrera, ya que sus resultados podrían servir como base para futuras acciones de prevención y capacitación en temas de seguridad informática.

La viabilidad del proyecto se apoya en la existencia de los conjuntos de datos públicos que incluyen correo legítimos y maliciosos.

De esta manera, el trabajo contribuye al fortalecimiento de competencias en áreas como ciencia de datos, ciberseguridad e inteligencia artificial, las cuales forman parte del perfil profesional de la carrera.

1.4 Objetivos

1.4.1 Objetivo general

Desarrollar un modelo basado en técnicas de aprendizaje automático que permita identificar correos electrónicos de tipo phishing dentro de un conjunto de datos ya existente, utilizando un umbral de decisión predefinido para clasificar los mensajes.

1.4.2 Objetivos específicos

Preprocesar y transformar el conjunto de datos mediante técnicas de limpieza, normalización y extracción de características, asegurando su calidad y adecuación para el entrenamiento de modelos de aprendizaje automático.

Diseñar la arquitectura del modelo de clasificación bajo algoritmos supervisados, con parámetros y criterios de evaluación para la detección del phishing.

Implementar un modelo de clasificación binaria a través de la regresión logística o árboles de decisión para clasificar los correos en los clasificados anteriormente.

Evaluar la calidad de clasificación del modelo mediante métricas como la precisión y el recall, y hacer una interpretación de los resultados del modelo en función al umbral de decisión acordado.

Proponer un esquema de mitigación institucional para tener alertas sobre correos detectados caracterizados como sospechosos por el modelo, como propuesta para docente en entornos académicos.

1.5 Alcance

La investigación se desarrolló con fines académicos y de experimentación. No se trabajó directamente con el sistema de correo institucional, sino con dataset públicos empleados comúnmente en estudios de ciberseguridad. Esta decisión metodológica permite trabajar de forma ética y controlada, sin comprometer la privacidad de usuarios reales.

Este trabajo es presentado como académico y experimental, ya que involucra análisis de datasets públicos de ciberseguridad y no interfiere con la infraestructura de un correo real. Se orienta al contenido de los mensajes en términos de su contenido textual únicamente y excluye enlaces, metadatos y archivos adjuntos con el fin de estudiar las T_M_L en un análisis de texto. No hay ninguna integración con plataformas de correo real ni se lleva a cabo un posible procesamiento en tiempo real.

CAPÍTULO II

MARCO TEÓRICO

2.1 Ciberseguridad y uso del correo electrónico

"La ciberseguridad ... es relevante debido a la creciente exposición de individuos, empresas, gobiernos, instituciones financieras y usuarios finales a diversas amenazas en el entorno digital. Nadie está exento de ser víctima de un ciberataque" (Tucker, sf). Esto es porque no es un problema limitado a un sector específico, sino un problema global que afecta a todos los que usan internet. Dado que las amenazas no establecen distinciones entre sus objetivos, resulta imperativo que tanto los individuos como las instituciones de cualquier escala mantengan un estado de alerta y adopten medidas de seguridad. La preparación adecuada constituye la principal defensa contra los riesgos capaces de afectar la integridad de la información y la continuidad de las operaciones.

"Los atacantes, conocidos como phishers, utilizan técnicas de ingeniería social para hacerse pasar por personas o instituciones confiables. Para ello, imitan comunicaciones oficiales a través de distintos medios electrónicos, como el correo electrónico, servicios de mensajería, redes sociales, mensajes SMS e incluso llamadas telefónicas" (Tucker, s. f.). En esencia, estos ataques se basan en engañar a los usuarios y aprovechar la confianza que depositan en este tipo de mensajes.

Esta situación demuestra que las herramientas tecnológicas de seguridad, por si solas, no son suficientes. La protección también depende del criterio de los usuarios, de su nivel de atención y de su capacidad para desconfiar de mensajes sospechosos. Por ello, una estrategia de prevención realmente efectiva debe integrar tanto soluciones tecnológicas como afirmación y concienciación de las personas.

2.2 Ingeniería social y phishing

2.2.1 Ingeniería social

En las universidades, el correo institucional es una herramienta básica para enviar tareas, recibir avisos y mantener comunicación con los docentes. Cuando una cuenta se ve comprometida,

el problema no afecta únicamente a su propietario, sino que también puede impactar a grupos de estudiantes y a otros miembros de la institución.

Por otra parte, muchas de las técnicas de ingenieria social se basan en aprovechar la forma en que las personas reaccionan ante ciertas situaciones. En lugar de explorar fallos técnicos, los atacantes suelen centrarse en el comportamiento de los usuarios para obtener información sensible. Los atacantes se aprovechan de emociones que incluyen la delicadeza de confiar en los demás y la urgencia, así como la ignorancia en torno a la protección digital (Alhogail & Alsabih, 2021). Los Atacantes aprovechan los sentimientos del ser humano.

Cuando un mensaje exige el desbloqueo de una cuenta o una verificación inmediata, lo que busca es ejercer una presión psicológica que nuble nuestro juicio. Es fundamental comprender que, ante la vulnerabilidad de muchos usuarios, las instituciones legítimas jamás solicitarán contraseñas a través de un correo electrónico, ni emplearán enlaces sospechosos o técnicas de suplantación para enmascarar sus direcciones oficiales.

La falta de conciencia en materia de seguridad no solo presenta dificultades para reconocer riesgos técnicos, sino que además implica una comprensión falsa de cómo operar la seguridad regular, por lo que adopta posturas vulnerables que cualquier usuario entendido rápidamente podría detectar.

2.2.2 Definición y tipos de phishing

Un método de ingeniería social, el phishing es el más común. Actuando de esta manera, los actores sienten que enviar correos electrónicos haciéndose pasar por una entidad legítima (por ejemplo, un banco, un servicio de redes sociales o una institución pública) y recuperar información sensible, como contraseñas y credenciales bancarias (INCIBE, 2024a, 2024b), añade el Instituto Nacional de Ciberseguridad de España. Sin embargo, gradualmente comenzaron a ocurrir tantos casos de phishing, y han aparecido algunas variedades de phishing:

- **Phishing clásico/phishing masivo**: mensajes enviados a las masas, sin personalización.
- **Spear phishing**: Un ataque específico que ataca por nombre a una entidad (conocida como objetivo) o número utilizando cierta información sobre esa

entidad, proporcionando una mejora de la credibilidad del mensaje o declaración realizada (INCIBE, 2024c).

- **Whaling**: Este objetivo apunta a ejecutivos o empresarios internos en posiciones de alto privilegio dentro de la red de una organización.
- **Ataques relacionados**: (robo de identidad a través de SMS o llamadas a otros) pero utilizando la misma mentalidad fraudulenta, pero a través de otro medio (INCIBE, 2024b).

En última instancia, todos estos casos implican lo mismo: atraer a los usuarios a hacer clic en un enlace malicioso, descargar un archivo infectado o ingresar los detalles de su cuenta y contraseña en un sitio web falso que es un proveedor original de servicios (Verma y Das 2017).

Las campañas de phishing clásico y masivo son casos amplios y generales en los que la tasa de fallo para el individuo único es pequeña, pero la tasa de fallo dentro de un individuo puede superarla fácilmente; en una escala mucho mayor que la de unos pocos intentos. Con estos ataques, el contenido genérico llega a todo tipo de usuarios y probablemente esté asociado con servicios populares (redes de correo electrónico, redes sociales y servicios de logística).

Aunque estas técnicas se han refinado mucho más, difieren en varios aspectos de su lenguaje, apariencia y preguntas generales. El spear phishing es una variación cualitativa, cambiando el equilibrio entre volumen y precisión. Mientras que el tipo genérico de mensajes que los atacantes enviarían son generales (o mensajes genéricos sin nombre), tales ataques apuntan a una sola persona (o grupo de personas) a través de comunicados personalizados con datos legítimos sobre el individuo atacante. Por ejemplo, un profesor universitario podría atacar con proyectos específicos para apuntar y nombrar a colaboradores existentes, incluso replicando algunos detalles de discusiones previas con grupos afiliados a autoridades de financiamiento (Bendale, 2021).

El whaling es un tema al que los usuarios con poder y acceso a la información pueden prestar atención. Y en entornos universitarios, esos objetivos pueden ser rectores, decanos y administradores senior. Y debido a que pueden autorizar transacciones, así como proteger información institucional sensible, este acceso a la cuenta es estratégicamente valioso.

El smishing y el vishing tienen más de un mecanismo para el phishing contra otros sitios. El smishing utiliza SMS para explotar que confiamos mucho más en los textos que en los correos electrónicos, que las URL completas están geográficamente restringidas y son difíciles para nosotros con URL completas. En contraste, el vishing puede ser interactivo a través del teléfono, en el cual operadores entrenados adaptan su voz según la respuesta del anfitrión instantáneamente, lo que puede crear historias más convincentes y reaccionar a objeciones en tiempo real (Almomani, 2013).

2.2.3 Phishing en entornos universitarios

Debido al impulso de los usuarios, el hecho de que constantemente ingresan nuevos estudiantes y que se utilizan intensamente servicios online sin embargo la falta de madurez en políticas de seguridad para compararse con la industria financiera ello hace instituciones educativas particularmente susceptibles al phishing (Almomani et al., 2013).

En los laboratorios computacionales de la universidad donde se utilizan herramientas de desarrollo y accesos a servidores que pueden verse afectados por correos falsos. En el correo electrónico institucional de la UPS, es vital tener una dirección para estos recursos informáticos. Si un ataque tiene éxito probablemente llevando a un robo de credenciales, en los sistemas académicos entrarán sin autorización o manipulación de información crítica.

Por eso, es fundamental buscar métodos de detección y respuesta que estén diseñados para este tipo de entornos, usando tecnologías actuales como los son el N_L_P y M_L.

2.3 Procesamiento de Lenguaje Natural (N_L_P) aplicado a correos electrónicos

2.3.1 Conceptos básicos de N_L_P

El método del Procesamiento de Lenguaje Natural incluye la clasificación de textos, el análisis de Como señalan Kumar et al. (2022), el N_L_P incluye desde clasificar textos hasta traducir idiomas, lo que demuestra que no es una herramienta simple, sino un conjunto de funciones para que las máquinas logren 'interpretar' lo que leen. Para el proyecto, esto es clave: no busca que la computadora solo reconozca palabras sueltas, sino que sea capaz de agrupar el mensaje completo para diferenciar si un correo es meramente informativo o si disimula una intención maliciosa. Al usar estas técnicas, estamos dotando al sistema de una especie de “criterio lingüístico” que va mucho más allá de un simple filtro de spam.

En este estudio, lo que hacemos es aplicar el N_L_P directamente a los correos electrónicos para ‘entrenar el oído’ del sistema. La idea es que el modelo aprenda a diferenciar un mensaje autentico de una trampa de phishing, captando esos pequeños detalles y mañas en la redacción que a los filtros de seguridad de toda la vida se les suelen escapar. Al final del día, los atacantes siempre dejan pistas en su forma de escribir, y con estas herramientas podemos detectar esos patrones sutiles antes de que el usuario caiga en el engaño.

2.3.2 Preprocesamiento de texto

Antes de meterle texto a un modelo de aprendizaje automático, es obligatorio pasar por una etapa de procesamiento.

Se requiere eliminar_Stop_words_y algunos otros cambios de otra manera para mejorar la calidad del texto/original (Kumar et al., 2022; Bendale & Patil, 2021).

- **Limpieza del texto (cleaning):** eliminar los signos de puntuación, etiquetas HTML, espacios redundantes y eliminar las mayúsculas.
- **Tokenización del texto:** dividir en palabras.
- **Eliminar_Stop_words:** quitar las palabras inconsecuentes demasiado comunes.
- **Lematización o stemming**: transformar una palabra en su forma básica.

Estos pasos preparan el terreno hacia una lectura más limpia de la red, disponible para detectar patrones de estafas de Internet. (Almomani, 2013).

El phishing proviene de características genéricas y se hace pasar seguido por mensajes de populares servicios. La campaña es masiva pero los resultados por individuo son bajos, en cantidad compensan, Aunque ganan en sofisticación, su idioma y diseño suelen tener errores que las delatan.

El spear phishing está dirigido a individuos o grupos pequeños de personas con mensajes personalizados, utilizando información real sobre el objetivo (Kumar et al. 2022). Es propio de los individuos de alto nivel en las universidades, como rector y decano, ya que ellos son los que poseen información vital.

El smishing y el vishing son expansiones del phishing en los canales de comunicación. El smishing usa SMS, aprovechando la creencia de que los textos deben ser de confianza; el vishing

consiste en llamadas telefónicas, que pueden crear conversaciones convincentes ya que la gente disminuye sus esperanzas sobre veracidad en tiempo real a través de ellas. (Almomani, 2013).

2.3.3 Representación vectorial del texto

Las computadoras trabajan con números, por lo que el texto debe transformarse en representación vectorial. Métodos comunes incluyen:

- **Bolsa de palabras (BoW):** cuenta las apariciones de cada palabra sin considerar el orden.
- **TF-IDF (Frecuencia del término – Inverso de la frecuencia en el documento):** las palabras importantes pesan más en un documento y las comunes menos.
- **Incrustaciones de palabras:** representan palabras en un espacio vectorial donde palabras parecidas están juntas.

Un término que aparece muchas veces en un documento tendrá una alta frecuencia del término. La IDF penaliza a los términos comunes en muchos documentos. Por tanto, TF-IDF combina ambos para dar un peso alto a términos que aparecen con frecuencia en un documento específico pero rara vez en general.

Esto es útil para buscar palabras clave en correos electrónicos fraudulentos. En la detección de la suplantación de identidad, estos vectores se utilizan en algoritmos de clasificación para identificar las combinaciones típicas de correos fraudulentos.

2.3.4 Rasgos lingüísticos en correos de phishing

Junto con la representación general del texto, varios análisis cualitativos y cuantitativos también identifican propiedades lingüísticas que se encuentran frecuentemente en correos electrónicos de phishing (Alhogail & Alsabih, 2021; Sahingoz et al., 2019) por ejemplo:

- Mensajes de urgencia o que señalan amenazas (por ejemplo, “su cuenta será bloqueada”, “última advertencia”).
- Solicitudes directas de credenciales o información personal.
- Referencias a nombres o instituciones reconocidas, pero con ligeras variaciones en el nombre o dominio.

- Errores gramaticales o estilísticos que son menos comunes, pero están presentes en muchas campañas.

Estos elementos se combinan con técnicas de PLN para ayudar a que los modelos automáticos sean más precisos en la detección.

2.4 Aprendizaje automático supervisado (Machine Learning) para detección de phishing

2.4.1 Aprendizaje supervisado y clasificación binaria

El M_L supervisado es una rama del M_L donde el algoritmo aprende a partir de ejemplos ya etiquetados. En otras palabras, se le proporcionan datos de entrada junto con la salida correcta (suplantación de identidad o legítimo) y el modelo va ajustando sus parámetros para reducir los errores de clasificación (Sahingoz et al., 2019).

Una vez entrenado, el modelo puede recibir un nuevo correo y devolver una probabilidad de que sea de phishing. En el caso de la localización de agresiones de phishing, este problema se plantea como una distinción binaria, donde cada correo pertenece a una de dos clases: 1 para correo de phishing y 0 para correo legítimo. A partir de esa probabilidad y de un umbral de decisión, se define si el mensaje se considera sospechoso o no, tal como se contempla en el diseño de este trabajo.

2.4.2 Algoritmos más usados en la literatura

En la literatura se han probado distintos algoritmos supervisados para la detección de correos phishing. Algunos de los más mencionados son (Almomani et al., 2013; Khonji et al., 2013; Alhogail & Alsabih, 2021):

- **Regresión logística:** modelo lineal sencillo pero muy utilizado para clasificación binaria; funciona bien con representaciones como TF-IDF.
- **Árboles de decisión y Random Forest:** permiten capturar relaciones no lineales entre las características y suelen tener buen rendimiento con datos de texto.
- **Máquinas de soporte vectorial (SVM):** son eficientes en áreas de mucha superficie, como los que se generan al trabajar con muchas palabras.
- **Naive Bayes:** algoritmo probabilístico clásico en clasificación de textos, por su simplicidad y rapidez.

- **Redes neuronales y modelos profundos:** trabajos más recientes exploran arquitecturas más complejas y modelos basados en deep learning para mejorar la detección (Kyaw, Gutierrez & Ghobakhlou, 2024; Tamal et al., 2024).

La elección de un algoritmo específico depende de factores como el tamaño del dataset, los recursos computacionales y el objetivo del proyecto.

2.4.3 Métricas de evaluación

Para saber si un modelo de detección de phishing funciona bien, no basta con entrenarlo; es necesario medir su desempeño con métricas adecuadas. Entre las más utilizadas se encuentran (Almomani et al., 2013; Sahingoz et al., 2019):

- **Accuracy (exactitud):** porcentaje total de correos clasificados correctamente.
- **Precisión (precision):** de todos los correos que el modelo marcó como phishing, cuántos realmente lo eran.
- **Recall o sensibilidad:** de todos los correos de phishing del conjunto de prueba, cuántos logró detectar el modelo.
- **F1-score:** media armónica entre precisión y recall, útil cuando se quiere un equilibrio entre ambos valores.
- **Matriz de confusión:** tabla que muestra cuántos verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos se obtuvieron.

El uso de un umbral de decisión permite ajustar el modelo según las necesidades del entorno.

2.5 Conjuntos de datos para la detección de phishing

Para que un modelo de aprendizaje supervisado funcione, necesitamos "enseñarle" con datos que ya estén bien etiquetados, de modo que el algoritmo aprenda a distinguir qué es un correo legítimo y qué es una trampa. Aunque históricamente se han usado repositorios como el de Enron o el SpamAssassin Public Corpus, en esta investigación hemos decidido trabajar con el CEAS 08 Dataset.

Esta elección no fue al azar. El dataset de la *Conference on Email and Anti-Spam* (CEAS) de 2008 es un referente en la comunidad científica porque contiene correos reales con sus

encabezados completos, lo que permite un análisis mucho más profundo que otros datasets que solo traen el texto plano. Al ser un estándar en la literatura de ciberseguridad, nos permite entrenar los modelos con ejemplos que, aunque tienen su tiempo, siguen siendo la base de las estructuras de fraude que vemos hoy en día. Además, usar un benchmark como el CEAS 08 garantiza que cualquier otro investigador pueda replicar mis experimentos y comparar los resultados, algo que es vital en la ciencia de datos.

Eso sí, hay que ser realistas: aunque estos datasets globales son excelentes para el laboratorio, no siempre captan al 100% las mañas del phishing en universidades latinoamericanas, donde el lenguaje y los modismos locales juegan un papel importante. Por eso, planteo este enfoque como una fase inicial de experimentación controlada. La idea es dejar sentadas las bases para que, más adelante y con los permisos éticos necesarios, se puedan integrar datos reales de nuestra propia universidad para terminar de pulir el sistema y adaptarlo a nuestra realidad local.

2.6 Mitigación académica y cultura de seguridad

Además de detectar automáticamente los correos fraudulentos, varios autores destacan que es igual de importante educar a los usuarios y aplicar medidas para mitigar el phishing como parte de una estrategia completa (Almomani et al., 2013; INCIBE, 2024a). Aunque en este proyecto no se conecta a un sistema real al correo institucional, si se plantea una propuesta de mitigación desde el ámbito académico, basada en:

- Etiquetar correos como sospechosos cuando su probabilidad de phishing supere un determinado umbral.
- Generar reportes o ejemplos que puedan utilizarse en talleres, charlas o materiales formativos sobre ciberseguridad.
- Sensibilizar a estudiantes y docentes sobre los signos de alerta más comunes en este tipo de mensajes.

De este modo, el modelo no solo se ve como una herramienta técnica, sino también como un recurso pedagógico que puede contribuir a fortalecer la cultura de seguridad de la información dentro de la UPS.

CAPÍTULO III

METODOLOGÍA

3.1 Revisión de la literatura y análisis del estado actual del conocimiento

Como primer paso en la metodología del proyecto, se realizó una revisión bibliográfica a con el fin de reconocer los enfoques técnicos utilizados en la identificación de correos electrónicos de phishing utilizando M_L supervisado y procesamiento de lenguaje natural. El meta principal fue reunir referencias técnicas que apoyaran la toma de decisiones durante el diseño del experimento.

Una revisión de la literatura técnica evidenció que, para diferenciar correos auténticos de intentos de phishing, la gran parte de las investigaciones elige modelos supervisados. Estos algoritmos son principalmente seguros cuando se dispone de un registro de correos previamente catalogados, dado que aprenden a identificar los patrones sutiles que separan un mensaje legítimo de uno dañino (Basit et al., 2020; Santosh Paradkar, 2023).

Para que el programa pueda interpretar el contenido de un correo, primero es necesario traducir el texto a un formato que pueda comprender los números. Para ello, aplicamos ciertos procedimientos de limpieza y descomposición de las palabras, conocidos como tokenización y lematización, y luego utilizamos una fórmula matemática, el conocido método TF-IDF, que asigna un valor numérico a cada palabra según su relevancia en el texto (Alhogail & Alsabih, 2021; Gupta et al., 2025).

Con esta base, el camino a seguir en el proyecto es bastante claro. Implementaremos un modelo de aprendizaje previamente entrenado para identificar patrones. Este modelo tomará los valores numéricos generados y, en lugar de emitir una respuesta categórica de "sí" o "no", calculará un porcentaje que indica la probabilidad de que el correo sea malicioso. Finalmente, si este porcentaje supera un umbral previamente establecido, se clasificará como phishing (Tamal et al., 2024).

3.2 Selección y justificación del conjunto de datos

3.2.1 Caracterización de la población del conjunto de datos

La población está compuesta por correos electrónicos digitales, tanto auténticos como fraudulentos (phishing), extraídos de conjuntos de datos públicos utilizados con fines académicos en el ámbito de la ciberseguridad. En cada correo electrónico se inspecciona de manera individual como si fuera un caso único que el modelo debe evaluar por separado. Este enfoque representa la forma tradicional de trabajar con algoritmos de M_L para la revisión de correos electrónicos y se considera un método bien establecido (Alhogail & Alsabih, 2021).

3.2.2 Proceso de selección de los datos CEAS_08

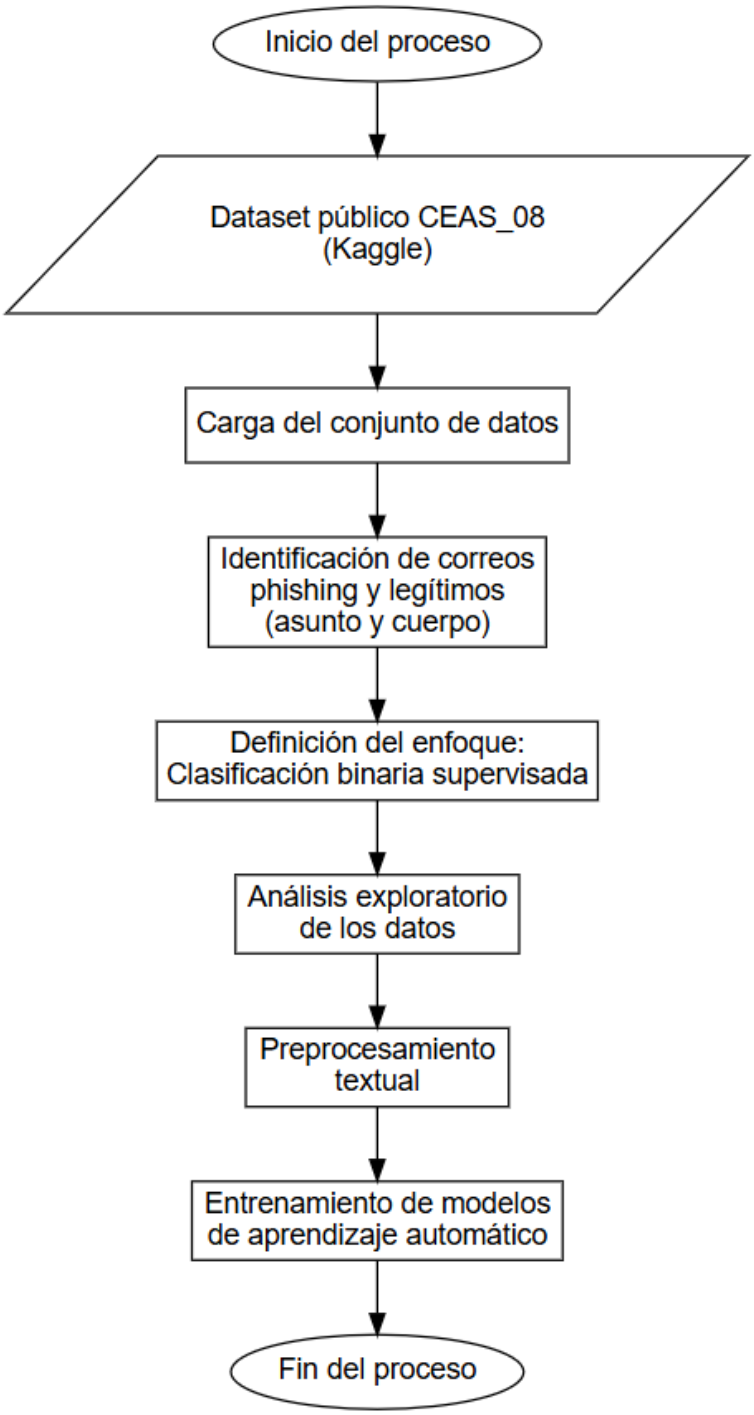
Para desarrollar el sistema de detección de phishing, utilizamos una base de datos pública proporcionada por Kaggle, llamada CEAS_08. Este conjunto incluye miles de correos electrónicos que ya han sido previamente etiquetados como "normales" o "fraudulentos", contando tanto con el asunto como con el contenido completo de cada mensaje. Gracias a su estructura, se pudo definir el problema con claridad: entrenar un modelo capaz de distinguir entre estas dos categorías, siguiendo un enfoque similar al de estudios previos (Alam, 2024).

La estructura del dataset es adecuada para utilizar técnicas de procesamiento de lenguaje natural (N_L_P) y entrenar modelos de aprendizaje automático (M_L). Este enfoque está alineado con prácticas documentadas en investigaciones recientes sobre detección de phishing mediante M_L y N_L_P, donde se resalta el uso de datasets públicos para garantizar la reproducibilidad y mantener un control experimental adecuado (Alhogail & Alsabih, 2021; Paradkar, 2023).

La Figura 1 se ilustra el flujo general del procedimiento metodológico desarrollado en este proyecto, desde la carga del dataset CEAS_08 hasta su empleo en las etapas de análisis exploratorio, preprocesamiento textual y entrenamiento del modelo. Esto pone de manifiesto la importancia crucial del conjunto de datos dentro de la estrategia metodológica planteada.

Figura 1.

Diagrama general del flujo metodológico del proyecto basado en el dataset CEAS_08



Nota. El diagrama presenta el proceso metodológico que se emplea para detectar correos electrónicos de phishing mediante el dataset CEAS_08. Elaborado por: Los Autores.

3.2.3 Distribución de clases y partición de los datos

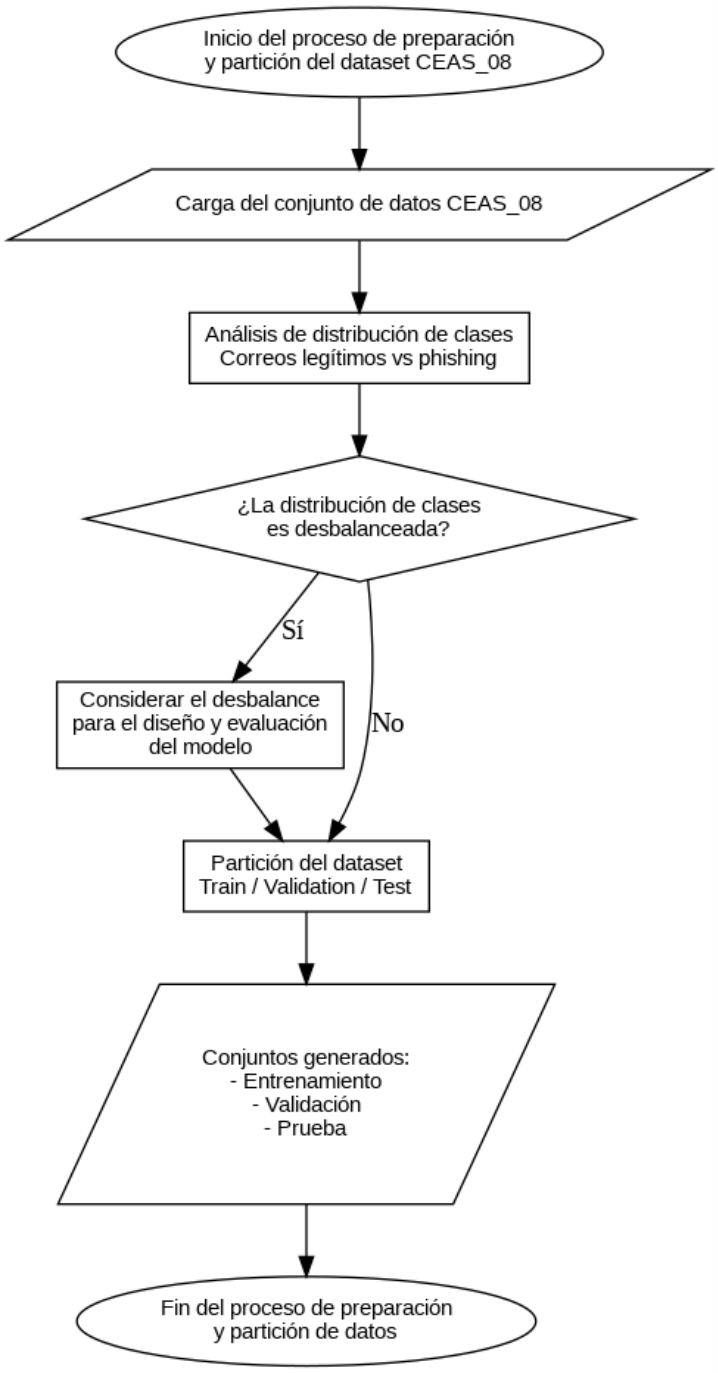
Para el entrenamiento del modelo, utilizamos los correos del conjunto de datos CEAS_08, que contiene mensajes legítimos y de phishing. En estos casos, es muy común que la cantidad de correos normales se exceda en gran medida a la de los mensajes fraudulentos. Desde el inicio, para asegurar que se pueda identificar ambos tipos de mensajes y así evitar un sesgo en el modelo, se considera en la desproporción en los datos.

Seguimos un método estándar para clasificar la información, dividiendo los datos en tres secciones distintas. El primero se usó para entrenar el modelo, la segunda para ajustar sus parámetros durante el desarrollo y la tercera se realizó únicamente para la prueba final. Este último paso es esencial porque permite la evaluación del modelo con correos completamente nuevos. Así se asegura que su aprendizaje no se limite a recordar ejemplos vistos durante su entrenamiento.

En la Figura 2 ilustra este proceso de elaboración y división de los datos antes de iniciar el entrenamiento.

Figura 2.

Diagrama de flujo del proceso de preparación y partición del conjunto de datos CEAS_08



Nota. El diagrama presenta las fases esenciales del procedimiento, que incluyen la carga del conjunto de datos, el análisis de la distribución de las clases y la división en conjuntos de entrenamiento, validación y prueba. Elaborado por: Los autores.

3.3 Análisis Exploratorio de Datos (EDA)

El primer paso consistió en realizar un análisis general del conjunto de datos para comprender a detalle el material con el íbamos a trabajar. Estudiamos del cómo estaba estructurado, los tipos de correos que incluye (legítimos frente a phishing) y las características generales del texto. Este primer análisis fue clave para identificar aspectos como la presencia de patrones comunes en la redacción o posibles desbalances en la cantidad de correos de cada clase. Esta evaluación preliminar nos proporcionó una base clara para decidor como procesar los datos y que métodos de análisis aplicar

En esta etapa, es esencial destacar que solo realizamos a observar; no llevamos a cabo ningún tipo de limpieza, ajustes o filtraciones de ninguna clase. Se examinaron los datos tal como estaban como estaban, sin ningún tipo de procesamiento. El objetivo era obtener una imagen clara y auténtica de los datos originales, de manera que no se incorporara ningún sesgo desde el principio.(Alhogail & Alsabih, 2021; Tamal et al., 2024).

3.3.1 Carga y análisis estructural del conjunto de datos

El primer paso fue en cargar el conjunto de datos CEAS_08, este dataset consta de 39,154 correos electrónicos repartidos en 7 atributos, los cuales describen tanto la información general sobre el mensaje como la variable objetivo que se utiliza para la clasificación.

En esta fase se revisaron los tipos de datos y la existencia de valores nulos en cada uno de los atributos. En la Tabla 1 se ofrece un resumen estructural del conjunto de datos, que incluye el nombre de los campos, el tipo de dato y la cantidad de valores que faltan.

Tabla 1.

Estructura del conjunto de datos CEAS_08 y valores nulos

Atributo	Descripción	Tipo de dato	Valores nulos
sender	Dirección del remitente del correo	Texto	0
receiver	Dirección del destinatario	Texto	462
date	Fecha y hora de envío del correo	Texto	0
subject	Asunto del correo electrónico	Texto	28
body	Cuerpo del mensaje	Texto	0
label	Etiqueta de clasificación (legítimo/phishing)	Entero	0
urls	Presencia de URLs en el correo	Entero	0

Nota. La tabla presenta un resumen de la organización del conjunto de datos CEAS_08, señalando las categorías de datos y la existencia de valores vacíos por atributo. Elaborado por: Los Autores.

Aunque hay valores vacíos en los campos *receiver* y *subject*, esto no influye de manera relevante en el análisis inicial ni en las fases siguientes, ya que el contenido textual fundamental (*body*) y la variable objetivo (*label*) están completos en todos los registros, asegurando así la factibilidad del proceso metodológico.

Con el objetivo de mostrar cómo están organizados los registros y ayudar a entender el conjunto de datos que se usa, en el Anexo 1 se incluye un ejemplo de correos electrónicos del conjunto de datos CEAS_08.

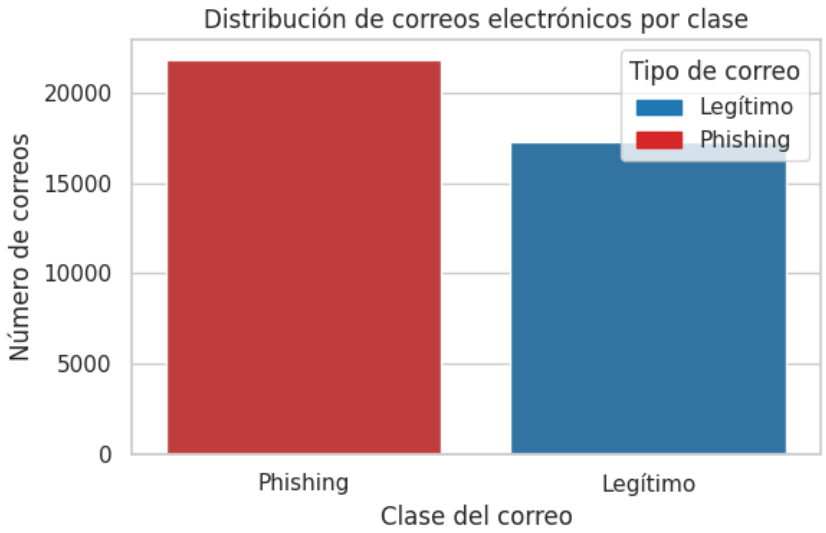
3.3.2 Estudio de la distribución de la variable objetivo

Se llevó a cabo un análisis de la distribución de la variable objetivo-asociada a la clasificación binaria de los correos electrónicos en mensajes legítimos y correos de phishing. De los datos valorados, se identificaron 21,842 correos clasificados como phishing y 17,312 como legítimos, lo que indica que hay un desbalance moderado entre ambas categorías, lo que significa que hay una mayor cantidad de ejemplos de correos fraudulentos en comparación con los normales.

La variación que observamos en la información se presenta en la Figura 3: hay una cantidad significativamente mayor de correos electrónicos que han sido etiquetados como phishing. Este desbalance no es propio de nuestro conjunto de datos, sino que también se encuentra en la mayoría de las bases utilizadas para desarrollar procesos que buscan detectar amenazas. Este punto es crucial, ya que, si se ignora, el modelo podría asimilar información de forma incorrecta y las métricas que se utilizan para evaluar su desempeño podrían dar resultados engañosos.

Figura 3.

Distribución de correos electrónicos por clase en el conjunto de datos CEAS_08



Nota. La figura representa el número de correos legítimos y de phishing presentes en los datos originales, antes de ser procesados. Elaborado por: los autores.

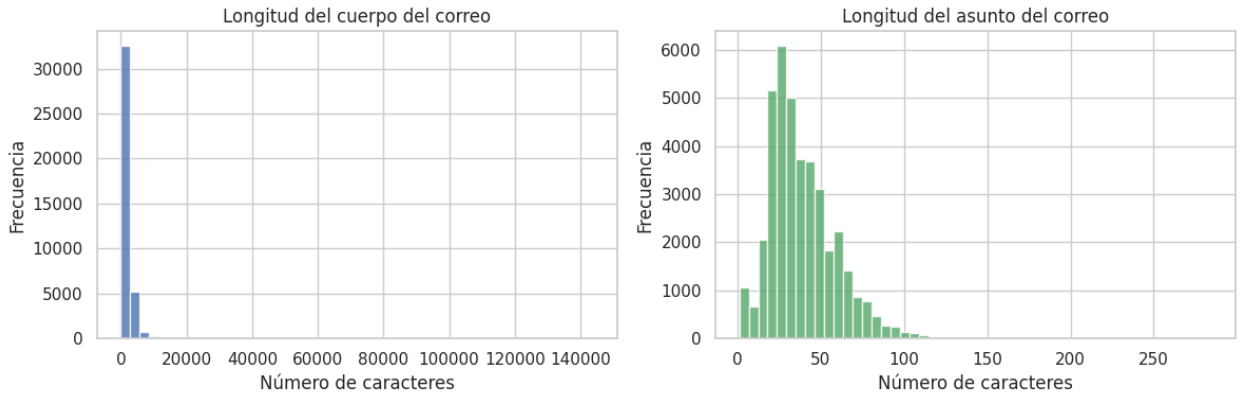
3.3.3 Análisis exploratorio de las propiedades textuales

Se exploró la amplificación textual de los correos electrónicos considerando por separado la longitud del cuerpo y del asunto de cada mensaje, medida en número de caracteres. Los resultados muestran una notable diferencia entre ambas variables: mientras que la longitud del cuerpo presenta una alta variabilidad, caracterizada por una gran dispersión y la presencia de valores atípicos, la longitud del asunto exhibe una distribución significativo más uniforme y reducida.

En la Figura 4, contiene histogramas de frecuencia para ambas dimensiones. Identificar estos rasgos en la composición textual resulta crucial desde las etapas iniciales del análisis, pues aporta evidencia preliminar para justificar la necesidad de incorporar técnicas avanzadas de N_L_P y para la representación textual en actividades de clasificación (Alhogail & Alsabih, 2021).

Figura 4.

Distribución de la longitud del cuerpo y del asunto de los correos electrónicos



Nota. La figura presenta dos histogramas que muestran cómo se distribuye la extensión, en número de caracteres, del cuerpo del correo y del asunto en el conjunto de datos analizado. Elaborado por: Los Autores.

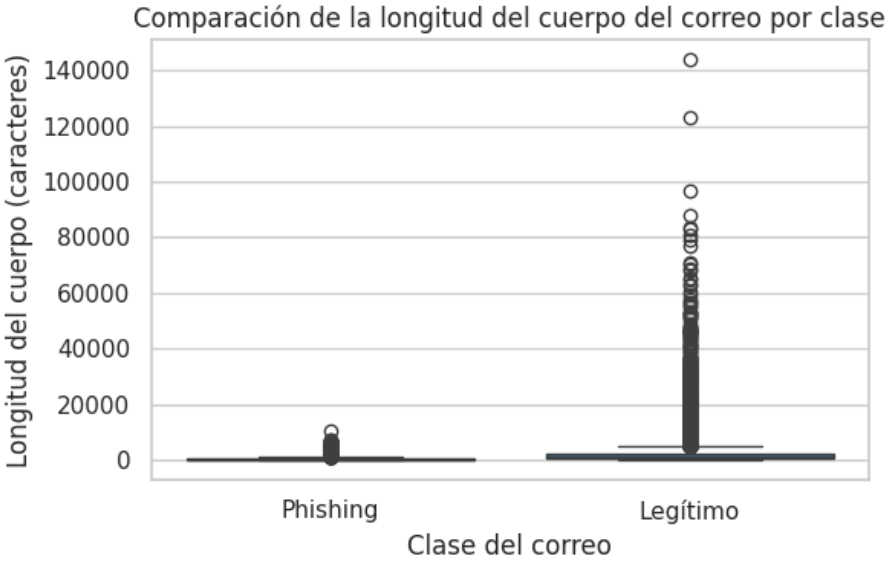
3.3.4 Análisis comparativo de la extensión del cuerpo del correo según categorías

Se analizó la longitud del cuerpo de los correos con el objetivo de comparar las características entre los mensajes legítimos y los de phishing. Para lograr esto, se utilizó una representación gráfica mediante diagramas de caja, los cuales facilitan para la identificación de la tendencia central, la dispersión y los valores atípicos en cada clase.

Sin recurrir a métodos de predicción, este método descriptivo permitió identificar diferencias notables en la distribución del tamaño entre ambas categorías. La Figura 5 presenta los resultados de esta comparación. Esto ofrece un punto de vista valioso para entender mejor los datos y orientar las decisiones relacionadas con las etapas siguientes.

Figura 5.

Comparación de la longitud del cuerpo del correo por clase



Nota. Comparación de diagramas de caja en relación con la longitud del cuerpo del correo para las dos categorías evaluadas. Elaborado por: Los Autores.

3.3.5 Alcance y delimitación del análisis exploratorio

Analizamos el contenido de los textos y la estructura del conjunto de datos CEAS_08, lo que nos permitió un análisis detallado. Es importante destacar que durante esta etapa no se realizó una limpieza ni alteramos los datos y tampoco entrenamos ningún modelo, pues la finalidad principal era solamente observar y comprender la información tal como estaba en su estado original.

Los resultados de esta revisión inicial fueron esenciales ya que establecieron la base para el resto del proyecto. A partir de ellos, logramos definir con mayor precisión cómo prepararíamos los textos y cómo diseñaríamos el proceso de modelado supervisado en las siguientes etapas.

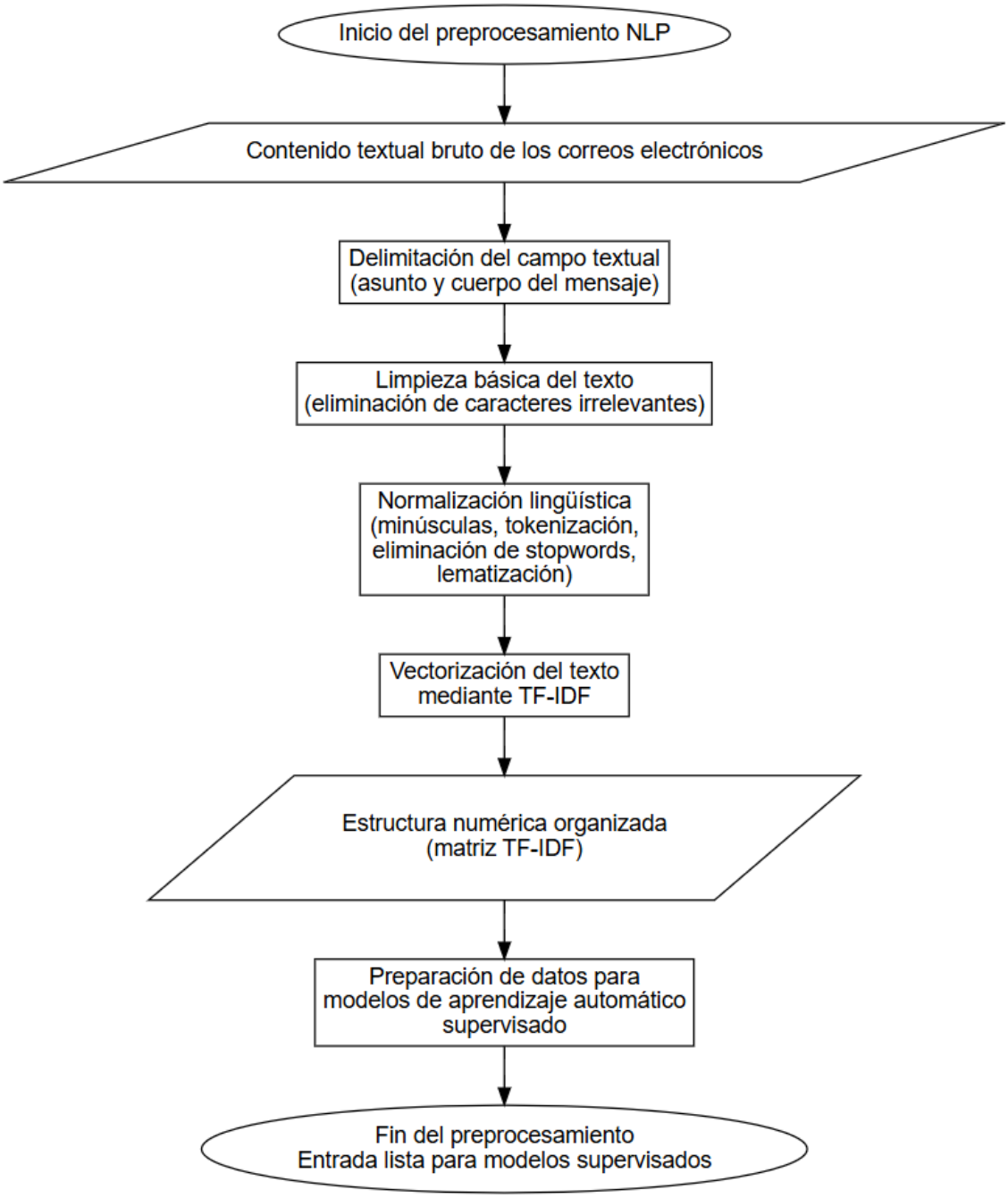
3.4 Preprocesamiento de texto a través de técnicas de Procesamiento de Lenguaje Natural (N_L_P)

Consistió en convertir el contenido textual de los correos electrónicos a un formato que sea estructurado y estándar, lo que permite su posterior procesamiento por parte de los algoritmos de M_L. Este paso es fundamental en actividades de clasificación de texto, ya que ayuda a disminuir el ruido en los datos, homogeneizar el lenguaje y gestionar de forma eficaz el tamaño del vocabulario, todos elementos importantes antes de la formación de cualquier modelo.

Con el fin de asegurar la consistencia, la repetibilidad y la calidad técnica de los resultados, se creó y aplicó un flujo de trabajo específico. La Figura 6 representa este proceso, que va desde la selección del campo de texto que se va a procesar hasta el almacenamiento final de las representaciones generadas, aplicado de manera metódica de los datos CEAS_08.

Figura 6.

Flujo del proceso de preprocesamiento del texto



Nota. Representación de las etapas secuenciales involucradas en el preprocesamiento del contenido textual de los correos electrónicos. Elaborado por: Los Autores.

3.4.1 Delimitación del ámbito textual de análisis

Para llevar a cabo el desarrollo del proceso de N_L_P se eligió el cuerpo del correo electrónico (body) como principal campo textual, dado a que contiene la mayor parte de la información semántica del mensaje. Además, se mantuvo la etiqueta (label) como la variable objetivo para garantizar que haya una relación entre los textos y clasificación. De este modo, el conjunto que se preparó para el preprocesamiento se compuso de 39,154 registros y dos columnas.

3.4.2 Proceso de limpieza fundamental del texto

En esta fase se llevó a cabo una limpieza preliminar con el objetivo de eliminar elementos que no añaden valor al análisis del texto. Las acciones realizadas incluyeron la transformación del texto a minúsculas, eliminar las etiquetas HTML, quitar URLs, deshacerse de caracteres que no son alfabéticos y ajustar los espacios en blanco. Estas medidas ayudan a disminuir diferencias sin importar en el vocabulario y a preparar el texto para una normalización lingüística más consistente.

Ejemplos del efecto de este proceso se presentan en el Anexo 2, donde se muestra el texto original y su versión tras la limpieza aplicada.

3.4.3 Normalización lingüística

Luego de limpiar el texto, se llevó a cabo un proceso de normalización lingüística que abarcó la tokenización, la eliminación de palabras vacías (*stopwords*) y la lematización. Para esto, se utilizó la librería spaCy con un modelo de lenguaje en inglés, desactivado los componentes innecesarios para esta etapa. La lematización ayudó a reducir las palabras a su forma orginal, reduciendo la dispersión del vocabulario y aumentando la coherencia semántica del texto procesado.

3.4.4 Vectorización mediante TF-IDF

Una vez que se ajustó el texto, se realizó su conversión a una representación numérica utilizando la técnica TF-IDF (*Frecuencia de Término–Frecuencia Inversa de Documento*). Para esto, se aplicó el *TfidfVectorizer* de la librería scikit-learn, que se configuró para incluir unigramas y bigramas, estableciendo un máximo de 5,000 características y un límite mínimo de frecuencia documental de cinco apariciones.

Como resultado de este procedimiento fue una matriz TF-IDF con dimensiones de 39,154 × 5,000, correspondiente al número de documentos por términos.

3.4.5 Revisión técnica del preprocesamiento

Antes de avanzar hacia la fase de modelado, se realizó una validación técnica del preprocesamiento con el fin de garantizar que la representación generada fuera coherente. En esta etapa se verificó la correspondencia entre la matriz de características y las etiquetas, así como el tamaño del vocabulario y la presencia de documentos que no contaban con una representación válida después del proceso de normalización.

La Tabla 2 se presenta una resume de los principales indicadores técnicos logrados durante esta validación.

Tabla 2.

Resumen técnico del proceso de preprocesamiento del texto

Indicador	Valor
Documentos totales	39,154
Documentos con representación válida	38,417
Documentos vacios	737
Tamaño del vocabulario TF-IDF	5,000

Nota. Los números están vinculados a la validación técnica del procesamiento previo antes al entrenamiento los modelos de M_L. Elaborado por: Los autores.

3.4.6 Conservación de elementos del preprocesamiento

Para asegurar que el experimento pueda ser reproducido y que se puede seguir el proceso, se guardaron los elementos generados durante el preprocesamiento. Esto incluye el conjunto de datos que fue preprocesado, la matriz TF-IDF, las etiquetas, el vectorizador que fue entrenado y un archivo que resumen los indicadores técnicos.

3.5 Entrenamiento y evaluación base de modelos supervisados

En esta etapa se estableció el método que se seguirá para el entrenamiento y la evaluación inicial de modelos de aprendizaje automático supervisado dedicados a la detección de correos electrónicos de phishing. El meta principal de esta fase fue crear un ambiente experimental

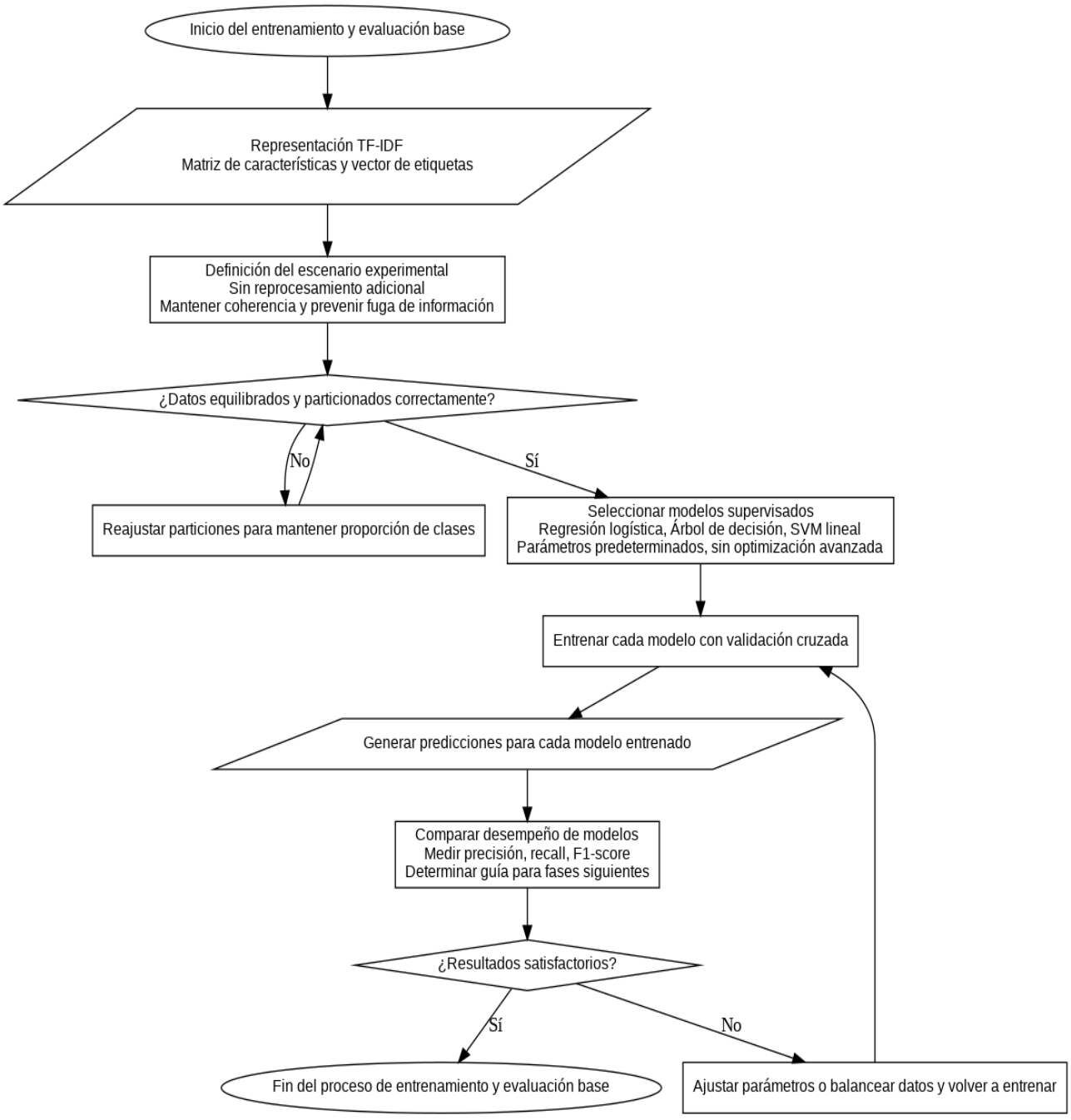
controlado, uniforme y reproducible que facilite la comprobación del rendimiento base de diferentes clasificadores condiciones similares, empleando una representación textual común y un método de validación coherente.

Para mantener la comparabilidad entre los modelos que se estaban evaluando, no se llevaron a cabo procesos de optimización de hiperparámetros ni ajustes de umbrales de decisión en este momento.

La Figura 7 presenta el flujo general seguido durante el entrenamiento y la evaluación de los modelos supervisados.

Figura 7.

Flujo metodológico del entrenamiento y evaluación base de modelos supervisados



Nota. Muestra la totalidad del flujo metodológico del estudio, que abarca el entrenamiento, la validación cruzada, la predicción y el análisis comparativo de los modelos. Elaborado por: Los Autores.

3.5.1 Preparación del escenario experimental

Se elaboro a partir de los resultados conseguidos directamente de la fase de preparación del texto. Utilizamos el vector de etiquetas y la matriz TF – IDF que se generó en el Paso 4, sin que se hiciera ningún reprocesamiento, modificación o ajuste extra a las características textuales previamente extraídas.

Este método aseguró un equilibrio en el enfoque a lo largo de todas las fases del proyecto y previno la introducción de cambios no controlados en los datos iniciales. La matriz de características empleada estuvo formada por 39,154 instancias y 5,000 atributos, asegurando una adecuada coincidencia dimensional entre los datos iniciales y sus etiquetas, lo cual es típico en tareas de clasificación de texto con representaciones de alta dimensionalidad.

3.5.2 Esquema de validación cruzada

Se utilizó un método de validación cruzada estratificada con cinco particiones (5-fold) para determinar la capacidad de los modelos supervisados, este procedimiento incluye una selección aleatoria de los datos para garantizar que cada parte represente correctamente del conjunto completo.

La utilización de validación cruzada estratificada ayuda a mantener la proporción original de las clases en cada división, lo cual resulta especialmente es muy importante en conjuntos de datos que presenta un desbalance moderado entre categorías, algo común en las tareas de detección de phishing. Este enfoque ayuda a obtener estimaciones más estables y fiables sobre el rendimiento de los modelos, al disminuir la dependencia de una sola división del conjunto de datos (James et al., 2023).

3.5.3 Elección y configuración de modelos supervisados

Se eligieron tres modelos supervisados tradicionales que son comúnmente utilizados en tareas de clasificación de textos y cuenta como respaldado en la literatura académica:

- **Regresión logística**, escogida por su capacidad de ser interpretada fácilmente y su eficiencia en entorno con alta dimensionalidad.
- **Árbol de decisión**, seleccionado por su habilidad para representar relaciones no lineales entre las diferentes características.

- **Máquina de vectores de soporte con núcleo lineal (SVM lineal)**, utilizada debidoa su rendimiento favorable en problemas de clasificación de textos con representaciones dispersas como TF-IDF.

En todas las situaciones se aplicó un ajuste de clases (*class_weight = "balanced"*) con el propósito de reducir el impacto del desbalance entre correos legítimos y de phishing, siendo esta una práctica recomendada en contextos de clasificación con categorías desiguales (Wainer, 2024)

3.5.4 Entrenamiento supervisado y generación de predicciones fuera de fold

Para el entrenamiento de los modelos y la creación de pronósticos se realizaron utilizando validación cruzada estratificada, logrando predicciones fuera de cada segmento (*out-of-fold predictions*). Este método facilita la medición del rendimiento de los clasificadores en datos que no se han utilizado durante la fase de entrenamiento, minimizando la posibilidad de sobreajuste y ofrece una valoración más imparcial del rendimiento inicial de cada modelo.

3.5.5 Almacenamiento y trazabilidad del proceso experimental

Con el fin de garantizar que el procedimiento de experimentación, se almacenaron los elementos generados en esta etapa, incluyendo las predicciones de cada modelo y los valores de las métricas calculadas. Esta acción permite repetir los experimentos, ayuda a comprobar los resultados y así garantiza la consistencia entre las etapas del método y el análisis que se realice después.

3.6 Diseño del mecanismo de mitigación académica basado en umbral

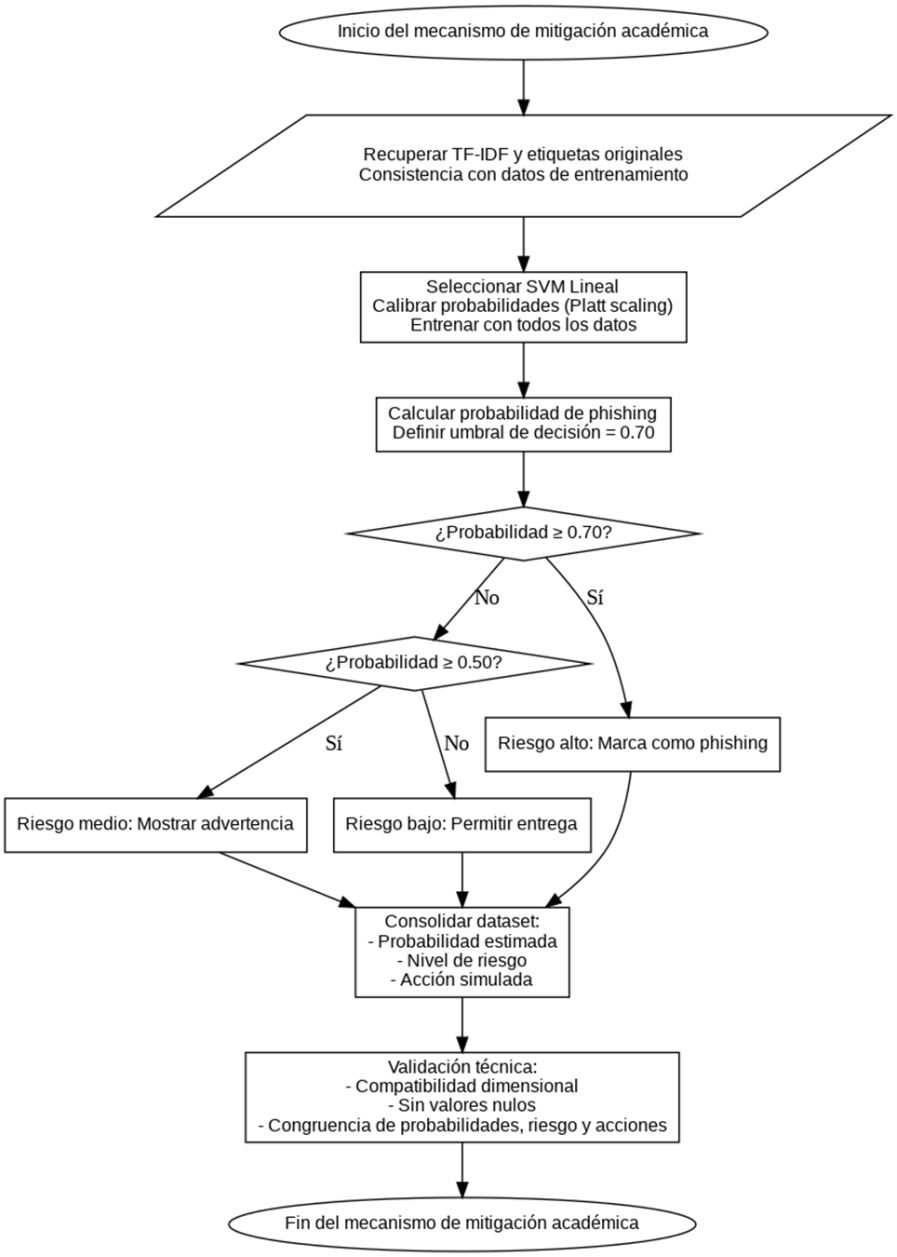
En esta fase se desarrolló un método de mitigación con un enfoque académico, dirigido a manejar el riesgo de correos electrónicos que podrían ser de phishing a partir de salidas probabilísticas. A diferencia de un método de clasificación binaria tradicional, el enfoque sugerido convierte la confianza del modelo en niveles de riesgo y en acciones simuladas (por ejemplo, informar al usuario o identificar un mensaje como phishing), lo que permite analizar cómo integrarse un clasificador dentro de un sistema básico de decisiones automáticas en ciberseguridad (Basit et al., 2020; Alhogail & Alsabih, 2021).

El método se labora reutilizando la misma representación TF-IDF creada en la etapa de preprocesamiento, evitando el reprocesado del texto o cambios en las características. Esto asegura la consistencia en la metodológica del experimento y garantiza que el comportamiento observado

en la mitigación dependa del modelo y del criterio de decisión, y no de modificaciones en la entrada. La Figura 8 presenta un resumen del flujo general del método implementado.

Figura 8.

Diagrama de flujo del mecanismo de mitigación académica basado en umbral



Nota. Diagrama del procedimiento de mitigación de riesgos, que contiene el ajuste del modelo, cálculo de probabilidades, aplicación de umbrales y asignación de niveles de riesgo con acciones simuladas. Elaborado por: los autores.

3.6.1 Recuperación de los datos preprocesados

El proceso de mitigación dio inicio con la incorporación de los elementos generados durante el preprocesamiento: la matriz TF-IDF y el vector de etiquetas. Esta decisión permite mantener la misma base de experimentación utilizada anteriormente, asegurando coherencia y continuidad en todo el procedimiento. La forma de representación empleada se relaciona con un espacio vectorial de gran dimensionalidad, que es típico en problemas de clasificación de texto fundamentados en TF-IDF (Sastre, 2024).

3.6.2 Entrenamiento del modelo final seleccionado

Se utilizó un clasificador SVM con tiene un núcleo lineal, el cual fue seleccionado tras el análisis comparativo realizado en la etapa anterior. En esta fase, el objetivo no era volver a comparar modelos, sino contar con un clasificador consistente que sirviera como el elemento central del plan de mitigación. Por esta razón, se entrenó el modelo utilizando el conjunto completo de instancias disponibles y se mantuvo el ajuste de pesos de clase (*class_weight="balanced")* con el fin de disminuir el impacto del desbalance entre categorías (Alhogail Alsabih, 2021; Wainer, 2024).

3.6.3 Calibración de probabilidades

Dado que los modelos SVM lineales no producen probabilidades de forma directa, se añadió un paso de calibración probabilística a través de un ajuste sigmoidal, que suele relacionarse con el método de Platt scaling. Esta calibración ayuda a convertir los resultados del clasificador en estimaciones de probabilidad comprensibles, lo cual es crucial cuando el sistema necesita funcionar con umbrales de decisión y no solo con etiquetas discretas (Platt, Niculescu-Mizil Caruana, 2005).

3.6.4 Determinación de la probabilidad de phishing

Después de entrenar y ajustar el modelo, se estableció una probabilidad continua entre [0,1] para cada correo electrónico, vinculada a la identificación de phishing. Estas probabilidades son la clave del sistema, ya que facilitan la creación de una respuesta escalonada según el nivel de riesgo, en lugar de tomar decisiones estrictas y rápidas.

3.6.5 Definición de umbrales de decisión

Con el propósito poner en práctica el sistema de mitigación, se establecieron límites claros explícitos de decisión que facilitan la clasificación de los correos electrónicos según su nivel de riesgo. En esta investigación se determinaron los criterios siguientes:

Tabla 3.

Umbrales establecidos para la clasificación del nivel de riesgo

Probabilidad estimada de phishing	Nivel de riesgo	Acción simulada
≥ 0,70	Alto	Marcar como phishing
≥ 0,50 y < 0,70	Medio	Mostrar advertencia
< 0,50	Bajo	Permitir entrega

Nota. Los límites se establecidos con objetivos académicos para imitar un sistema de mitigación progresiva basado en el riesgo. Elaborado por: Los Autores.

3.6.6 Asignación de niveles de riesgo y acciones simuladas

Según los límites establecidos, se clasificó cada correo en una categoría de riesgo (alto, medio o bajo). Después, se asoció a cada categoría una acción simulada que se alinea con el propósito educativo del proyecto. Estas acciones son solo una muestra del mecanismo y no significan un bloqueo real, ajuste de bandejas de entrada ni intervención en sistemas operativos.

3.6.7 Creación del conjunto de datos para mitigación

Gracias al mecanismo sugerido, se creó un conjunto de datos de mitigación que reúne, para cada mensaje de correo electrónico, la probabilidad calculada de phishing, el nivel de riesgo corresponde y la acción simulada asociada.

Tabla 4.

Esquema del conjunto de datos para mitigación académica

Variable	Descripción
probabilidad_phishing	Probabilidad estimada de pertenencia a la clase phishing
nivel_riesgo	Categoría de riesgo asignada (alto, medio, bajo)
accion_simulada	Acción de mitigación definida según el nivel de riesgo

Nota. Esta tabla explica cómo está organizado el conjunto de datos de mitigación creado en la etapa metodológica. Elaborado por: Los Autores.

3.6.8 Comprobación técnica y almacenamiento de elementos

Por último, se realizó una comprobación técnica con el propósito de verificar la consistencia dimensional entre el total de instancias y los registros del conjunto de mitigación y la falta de valores nulos en las variables generadas. Estas comprobaciones contribuyen en disminuir los riegos de inconsistencias antes de seguir con la etapa de evaluación utilizando las métricas como ROC/AUC.

Con el fin de garantizar tanto la trazabilidad como la reproducibilidad del proceso, se almacenaron los principales elementos: el modelo final calibrado y el conjunto de datos de mitigación, ambos en formatos (pkl y .csv). Este mantenimiento permite que sea reproducible y conectar directamente la técnica aplicada con el estudio realizado en las próximas secciones.

3.7 Análisis del umbral de decisión y evaluación mediante curva ROC

En este momento se definió el método para evaluar el comportamiento del modelo final al cambiar su umbral de decisión. Para esto, se utilizaron herramientas de análisis, como la curva ROC y el cálculo del área bajo la curva (AUC).

Este enfoque permite medir la capacidad del clasificador para diferenciar entre correos legítimos y aquellos de phishing, sin depender de una configuración fija. Este análisis resulta

especialmente importante en contextos de detección de phishing donde existe un balance entre detección temprana y minimización de falsos positivos (Nahm, 2022; Li, 2024).

La explicación numérica de los valores de AUC y otros indicadores pertinentes se encontrará en el Capítulo 4 – Resultados.

3.7.1 Recuperación de los artefactos del modelo final

El análisis ROC se dio inicio mediante la recuperación de los elementos producidos en etapas anteriores del proyecto, asegurando que fueran consistentes y con el diseño experimental establecido. En específico, se manejaron los siguientes artefactos:

- El modelo SVM lineal calibrado, entrenado durante el Paso 6.
- El archivo que contiene las probabilidades estimadas de phishing para cada correo electrónico.
- Las etiquetas auténticas del conjunto de datos, generadas la fase de preprocesamiento.

Antes de realizar el cálculo de la curva ROC, se revisó que las dimensiones del vector de probabilidades coincidieran con las del vector de etiquetas verdaderas, para evitar desajustes que pudieran afectar en la validez del proyecto.

Tabla 5.

Artefactos utilizados en el análisis ROC

Artefacto	Origen	Propósito
modelo_final_svm_lineal.pkl	Paso 6	Recuperar el clasificador calibrado definido
resultado_mitigacion_academica.csv	Paso 6	Extraer probabilidad_phishing por instancia
labels.pkl	Paso 4	Conseguir las etiquetas verdaderas para el cálculo ROC

Nota. Esta tabla presenta un resumen de los insumos necesarios para realizar el análisis de umbral y la generación de la curva ROC. Elaborado por los autores.

3.7.2 Cálculo de la curva ROC

Una vez que se obtuvieron los instrumentos necesarios, se llevó a cabo el cálculo de la curva ROC basándose en las probabilidades que el modelo final había estimado y en las etiquetas verdaderas del conjunto de datos. Este proceso permitió analizar cómo se comportaba el clasificador con diferentes valores de umbral de decisión.

La curva ROC se elaboró mostrando la relación entre:

- **FPR (False Positive Rate):** porcentaje de correos legítimos clasificados incorrectamente como phishing.
- **TPR (True Positive Rate**): porcentaje de correos de phishing que fueron detectados correctamente.

El cálculo se llevó a cabo utilizando las herramientas que proporciona por la biblioteca *scikit-learn*, para utilizar las funciones (roc_curve y auc), que son ampliamente utilizadas para evaluar modelos de clasificación binaria (*Sklearn.metrics.roc_curve — Scikit-Learn 0.23.0 Documentation*, n.d.).

3.7.3 Cálculo del Área Bajo la Curva (AUC)

Como un resumen del rendimiento total del clasificador, se determinó el Área Bajo la Curva ROC (AUC). Este parámetro mide la habilidad del modelo para dar una probabilidad más alta a los casos positivos en relación con los negativos, sin considerar el umbral que se elija (Chugh, 2024).

El valor del AUC sirvió como un indicador adicional para descubrir la capacidad de distinción del modelo final, sin hacer en este momento comparaciones o juicos sobre su importancia, los cuales se abordarán en el capítulo de resultados.

3.7.4 Variables generadas para el análisis ROC

A lo largo del proceso de valoración se produjeron las siguientes variables, que forman la base técnica del análisis ROC que se llevó a cabo:

Tabla 6.

Variables generadas para el análisis ROC

Variable	Descripción
fpr	Tasa de falsos positivos para cada umbral
tpr	Tasa de verdaderos positivos para cada umbral
thresholds	Umbrales evaluados por el clasificador
roc_auc	Área bajo la curva ROC

Nota. Estas variables son el resultado del procedimiento metodológico establecido para analizar del modelo de clasificación. Elaborado por: Los Autores.

3.7.5 Verificación técnica del análisis ROC

Se realizó una evaluación técnica para garantizar la coherencia interna de los datos que se utilizaron en el análisis. Esta revisión incluyó la verificación de que el total de probabilidades generadas por el modelo coincidiera con la cantidad de etiquetas reales presentes. Este procedimiento es esencial para garantizar que las dimensiones de los datos sean íntegras y prevenir cualquier conflicto que tenga el potencial de afectar la validez del cálculo del AUC y la curva ROC.

3.7.6 Almacenamiento y trazabilidad de resultados ROC

Para asegurar que el análisis sea reproducible y que su revisión o consulta en el futuro sea fácil, los resultados se almacenaron de manera ordenada en archivos permanentes:

- Un archivo .csv que contiene los valores de fpr, tpr y su umbral correspondiente.
- Un archivo .txt que guarda el valor de AUC.

Este tipo de almacenamiento permite volver a utilizar el análisis para hacer comparaciones futuras o para justificar de manera formal la selección de los umbrales utilizados dentro de la estrategia de mitigación.

CAPÍTULO IV

RESULTADOS

4.1 Caracterización inicial del conjunto de datos CEAS_08

4.1.1 Composición y estructura del conjunto de dato

El conjunto de datos CEAS_08 está formado por 39,154 entradas y 7 características, que abarcan información del remitente, del destinatario, la fecha, el tema, el contenido del correo electrónico, una etiqueta de clasificación binaria y un indicador de la presencia de URLs. La variable objetivo se encuentra completamente especificada para cada uno de los registros, lo que simplifica su implementación en tareas de clasificación supervisada sin requerir modificaciones adicionales.

En el análisis estructural, se identificaron valores nulos en los campos de receiver y subject. Los restantes atributos no presentan registros ausentes, lo que facilita la continuación del análisis sin problemas significativos.

Este fenómeno puede observarse en la llegada de los registros de los conjuntos de las bases de datos reales de correos y no pone en peligro el análisis exploratorio del fenómeno, puesto que los protocolos actuales de preprocesamiento permiten gestionar adecuadamente estas mínimas faltas de datos.

La Tabla 7 presenta una muestra de registros del conjunto de datos CEAS_08, utilizada para verificar la coherencia entre el contenido textual de los correos electrónicos.

Tabla 7.

Muestra de registros del dataset CEAS_08

	sender	receiver	date	subject	body	label	urls
0	Young Esposito <Young@iworld.de>	user4@gvc.ceas-challenge.cc	Tue, 05 Aug 2008 16:31:02 -0700	Never agree to be a loser	Buck up, your troubles caused by small dimensi...	1	1
1	Mok <ipline's1983@icable.ph>	user2.2@gvc.ceas-challenge.cc	Tue, 05 Aug 2008 18:31:03 -0500	Befriend Jenna Jameson	\nUpgrade your sex and pleasures with these te...	1	1
2	Daily Top 10 <Karmandeep-opengevl@universalnet...	user2.9@gvc.ceas-challenge.cc	Tue, 05 Aug 2008 20:28:00 -1200	CNN.com Daily Top 10	>+=+...	1	1
3	Michael Parker <ivqrnai@pobox.com>	SpamAssassin Dev <xrh@spamassassin.apache.org>	Tue, 05 Aug 2008 17:31:20 -0600	Re: svn commit: r619753 - in /spamassassin/tru...	Would anyone object to removing .so from this ...	0	1
4	Gretchen Suggs <externalsep1@loanofficertool.com>	user2.2@gvc.ceas-challenge.cc	Tue, 05 Aug 2008 19:31:21 -0400	SpecialPricesPharmMoreinfo	\nWelcomeFastShippingCustomerSupport\nhttp://7...	1	0

Nota. Los registros que permiten verificar la correspondencia entre el contenido del correo electrónico y su etiqueta de clasificación. Elaborado por: Los Autores.

4.2 Análisis de la variable objetivo

El examen de la variable objetivo reveló una distribución asimétrica entre las diferentes clases del conjunto de datos. De la cantidad total de registros analizados, se identificaron 21,842 correos electrónicos como phishing, mientras que 17,312 fueron clasificados como legítimos, lo que indica una discrepancia moderada, con una mayoría de mensajes maliciosos.

Este patrón no es sorprendente en el ámbito de la ciberseguridad, donde los ataques tienden a ocurrir con alta frecuencia y se concentran en ciertos intervalos de tiempo.

4.2.1 Distribución de clases

La distribución que se ha evidenciado se encuentra en consonancia con las idiosincrasias del conjunto de datos CEAS_08 e influye de un modo u otro respecto a las etapas posteriores de análisis. En este sentido, este desbalance se ha de tener en cuenta a la hora de optar por la métrica de evaluación y a la hora de explicar cómo se interpreta el rendimiento de los modelos de clasificación, evitando caída en errores que nos lleven a conclusiones inadecuadas.

En la Figura 3 refleja gráficamente cómo se relacionan los emails de phishing con los legítimos, siendo una clara referencia visual de la disminución de la proporción entre ellos y que puede contribuir a entender el desbalance detectado en la fase de la exploración del conjunto de datos.

4.3 Análisis de características textuales del contenido de los correos electrónicos (EDA)

4.3.1 Distribución de la longitud del cuerpo y del asunto

El análisis de las características textuales se centró en estudiar la longitud del asunto y del cuerpo del correo electrónico, teniendo en cuenta para ambos aspectos el número de caracteres como dimensión descriptiva. Los resultados preliminares ya ponen de manifiesto diferencias notorias para estas variables, tanto en cuanto a la escala como al modo en el que se distribuyen los valores, lo que sugiere comportamientos textuales distintos.

En lo que respecta al cuerpo del correo, se demuestra que existen diferencias importantes en la longitud de los mensajes, una extensión media de 1,571 caracteres, una longitud mediana de 570 caracteres, lo que permite observar claramente una distribución visiblemente asimétrica, distorsionada por los outliers que resultan significativos. No en vano, se producen casos de correos de hasta los 140,000 caracteres, lo que pone de manifiesto que existe un grupo de mensajes

mínimos son válidos junto a otro que incluye mensajes de longitud extrema. Este patrón que observamos también es bastante típico entre los datos reales de correos, casi una reverberación de la heterogeneidad del contenido textual.

La longitud del asunto muestra por su parte un patrón más estable.

En esta variable, el promedio se sitúa en 38 caracteres y el valor máximo registrado es de 285 caracteres, lo que sugiere una estructura textual más uniforme y coherente. A causa de las diferencias notables en la escala y en la dispersión entre ambas variables, las representaciones gráficas fueron elaboradas de forma separada. La Figura 4 ilustra las distribuciones correspondientes a la longitud del cuerpo del correo y del asunto, permitiendo una comparación visual de estas variaciones estructurales.

4.3.2 Comparación de la longitud del cuerpo del correo en función de la clase

El estudio comparativo de la longitud del contenido textual de los correos electrónicos, clasificado según su tipo, reveló diferencias significativas en la manera en la que se presenta la información entre los correos auténticos y aquellos identificados como phishing. En los correos legítimos, en la Figura 5 se observa una mayor diversidad en la extensión de los mensajes, así como una mayor presencia de valores atípicos, lo que indica la combinación de comunicaciones concisas con otras mucho más largas y detalladas.

4.4 Resultados del preprocesamiento del texto

4.4.1 Limpieza básica del texto

En la Tabla 8 se muestra una comparación entre el texto original y el texto limpio luego de aplicar el N_L_P.

Tabla 8.

Comparación entre texto original y texto limpio

body	clean_body
Buck up, your troubles caused by small dimensi...	buck up your troubles caused by small dimensio...
Upgrade your sex and pleasures with these te...	upgrade your sex and pleasures with these tech...
>+=+...	the daily top from cnn com top videos and stor...
Would anyone object to removing .so from this ...	would anyone object to removing so from this l...
WelcomeFastShippingCustomerSupport http://7...	welcomefastshippingcustomersupport

Nota. El proceso de limpieza básica reduce el ruido superficial del texto sin modificar su significado semántico. Elaborado por: Los Autores.

4.4.2 Normalización lingüística del texto

La Tabla 9 presenta ejemplos del resultado del proceso de limpieza y normalización lingüística.

Tabla 9.

Ejemplo del proceso de limpieza y normalización del texto

clean_body	processed_body
buck up your troubles caused by small dimensio...	buck trouble cause small dimension
upgrade your sex and pleasures with these tech...	upgrade sex pleasure technique
welcomefastshippingcustomersupport	welcomefastshippingcustomersupport

Nota. Se muestra la reducción del ruido textual y la normalización lingüística previa a la vectorización. Elaborado por: Los Autores.

4.5 Resultados del entrenamiento y evaluación de modelos supervisados

Se compararon con tres técnicas: Regresión Logística, Árbol de Decisión y Máquina de Vectores de Soporte (SVM) lineal, utilizando como entrada la representación TF-IDF que fue creada durante la fase de preprocesamiento de texto.

Se verificó el análisis a través de la validación cruzada estratificada en cinco particiones. Este método posibilitó la obtención de una estimación robusta del rendimiento de cada modelo, reduciendo el sesgo que podría ocurrir al emplear una sola división de los datos.

La tabla 10 muestra las métricas comparativas (*precision*, *recall* y *F1-score)* de cada uno de los clasificadores que se han evaluado.

Tabla 10.

Comparación de métricas de desempeño de los modelos supervisados

Modelo	Precisión	Recall	F1-score
Regresión Logística	0.9891	0.9922	0.9907
Árbol de Decisión	0.9654	0.9888	0.9770
SVM Lineal	0.9923	0.9951	0.9937

Nota. Las métricas fueron calculadas utilizando la validación cruzada estratificada de cinco particiones, tomando como entrada la representación TF-IDF del texto de los correos electrónicos. Elaborado por: Los Autores.

Los hallazgos muestran un rendimiento alto en los tres modelos evaluados. No obstante, el modelo SVM lineal se distingue por mostrar los valores más elevados de forma constante en todas las métricas, nos indica una mayor cabida de generalización y una mejor separación entre emails genuinos y phishing.

4.6 Análisis de las matrices de confusión

Con la finalidad de explorar con mayor detalle el desempeño de los modelos más allá de las métricas generales, se examinaron las matrices de confusión asociadas para cada clasificador.

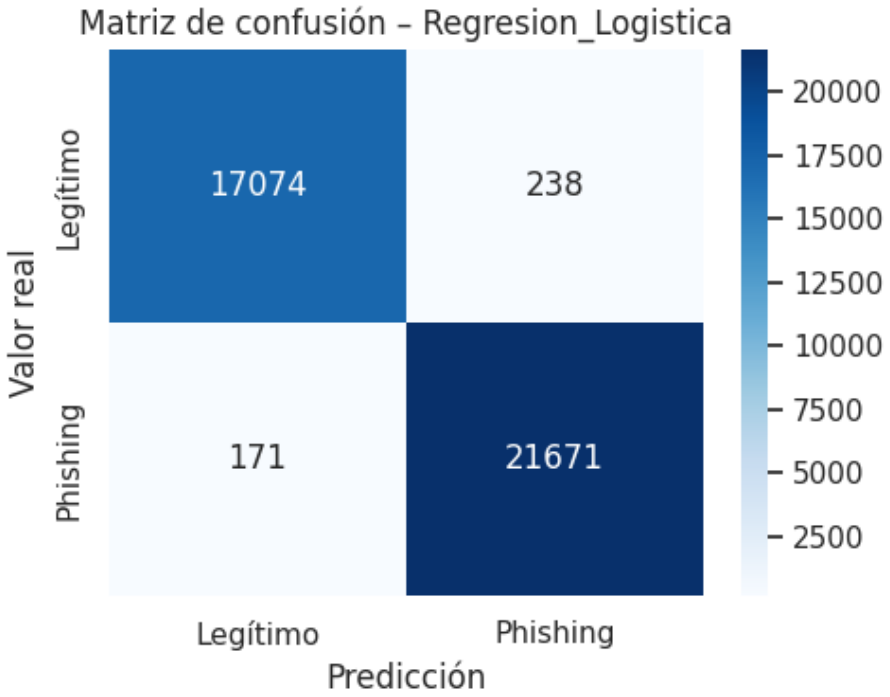
Este análisis posibilita detectar cómo se distribuye los aciertos y los errores en la clasificación, además de evaluar cómo influye los falsos positivos y falsos negativos en cada uno de los modelos.

4.6.1 Regresión Logística

Se muestra una gran cantidad para clasificar correctamente ambas categorías, con pocos errores. Sin embargo, se detectan ciertos falsos positivos y falsos negativos, lo que es normal en situaciones reales de identificación de amenazas, donde hay una superposición semántica entre mensajes legítimos y fraudulentos.

Figura 9.

Matriz de confusión del modelo de Regresión Logística



Nota. La figura muestra un balance apropiado entre la identificación de correos electrónicos de phishing y la reducción de alarmas erróneas. Elaborado por: Los Autores.

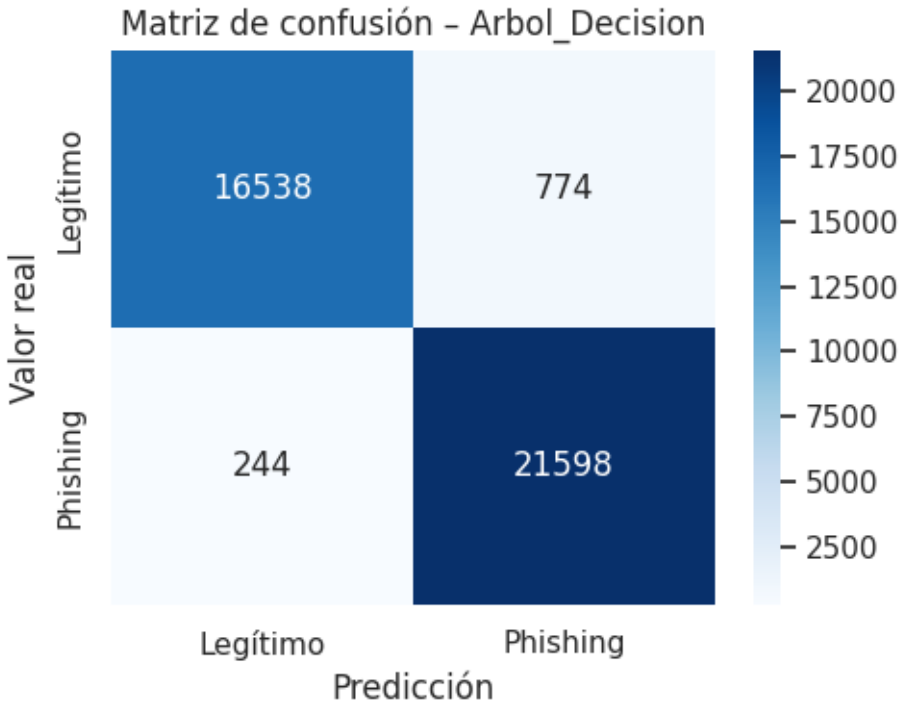
4.6.2 Árbol de Decisión

Si bien el rendimiento general es aceptable, se observa un mayor número de fallos en comparación con los otros modelos, especialmente en la clasificación de correos legítimos. Esta

situación indica una mayor vulnerabilidad al ruido del conjunto de datos, algo común en modelos que utilizan en reglas jerárquicas.

Figura 10.

Matriz de confusión del modelo Árbol de Decisión



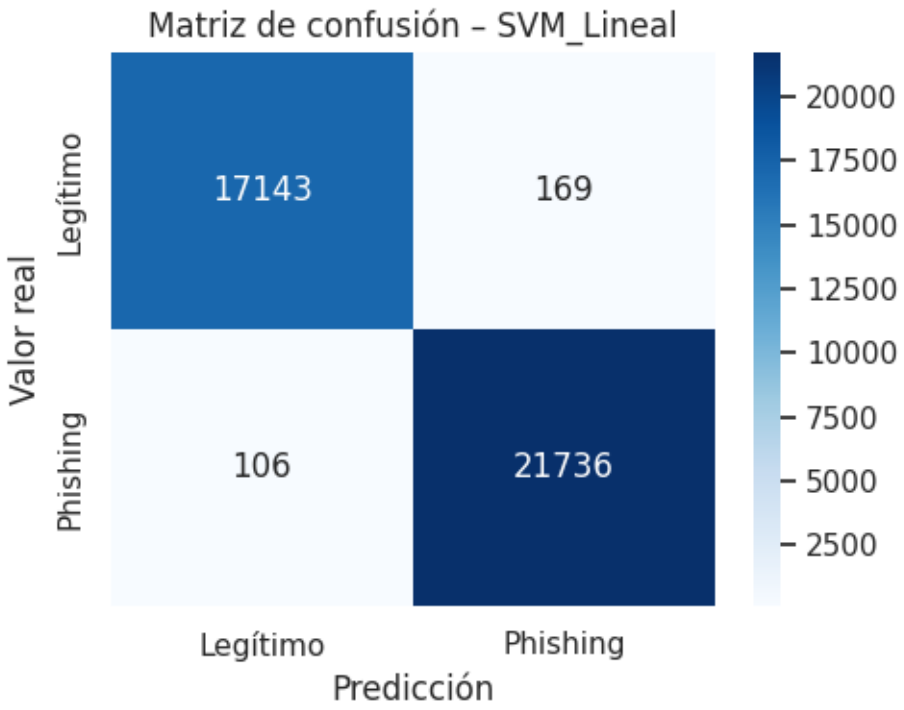
Nota. El modelo se centra en la detección de phishing, lo que conlleva un aumento en la tasa de falsos positivos sobre correos electrónicos legítimos. Elaborado por: Los Autores.

4.6.3 SVM Lineal

Este clasificador ofrece el balance óptimo entre precisión y sensibilidad, con una disminución notable en la cantidad de falsos positivos y de falsos negativos. Este hallazgo apoya su elección como modelo definitivo, mostrando una gran estabilidad y capacidad de generalización sobre los datos analizado.

Figura 11.

Matriz de confusión del modelo SVM Lineal



Nota. El modelo SVM Lineal presenta una distribución de errores más equilibrada y estable entre ambas clases. Elaborado por: Los Autores.

4.7 Resultados del modelo final y del mecanismo de mitigación académica

Luego de realizar un análisis comparativo, se optó el modelo SVM lineal como el clasificador por su mejor rendimiento. Sobre este modelo, se estableció un mecanismo de mitigación académica basado en umbrales de probabilidad, lo que posibilita clasificar los correos electrónicos de acuerdo con su nivel de riesgo.

La Tabla 11 presenta el porcentaje de los correos electrónicos clasificados en los niveles de riesgo Alto, Medio y Bajo, utilizando un umbral de decisión de 0.70 para la categoría de phishing.

Tabla 11.

Distribución porcentual de correos electrónicos por nivel de riesgo

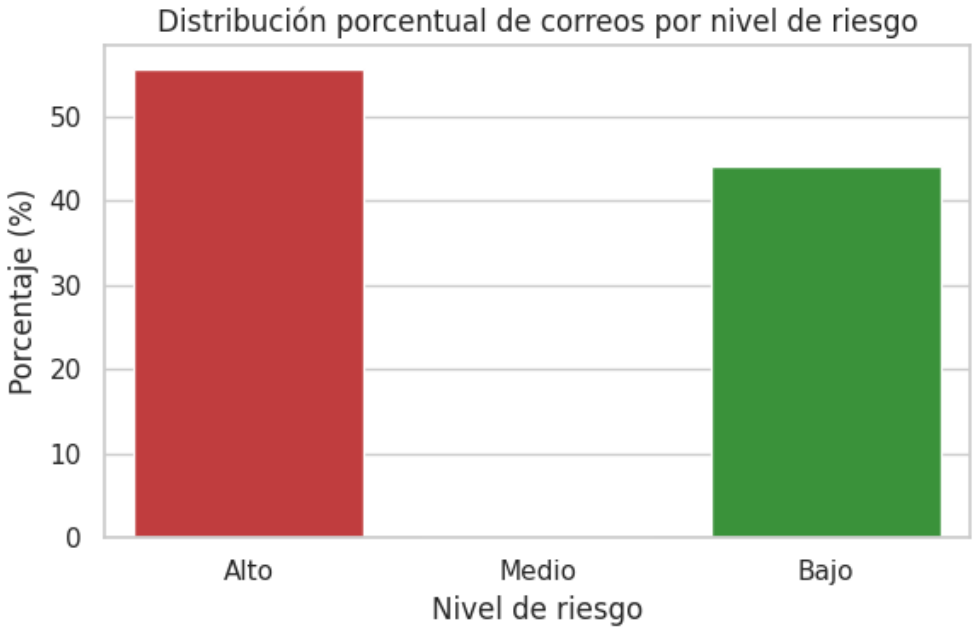
Nivel de riesgo	Porcentaje de correos (%)
Alto	55.70
Medio	0.16
Bajo	44.14

Nota: La categorización según el nivel de riesgo se derivó de las probabilidades calculadas mediante el modelo SVM lineal calibrado. Desarrollado por los autores.

La Figura 12 muestra la distribución de los correos electrónicos, destacando que más del 50% de ellos se consideran de alto riesgo. Este resultado está en línea las características de los datos y con la estrategia cautelosa del mecanismo de mitigación, que busca dar prioridad a la detección temprana de amenazas posibles.

Figura 12.

Distribución porcentual de correos electrónicos según nivel de riesgo



Nota. La figura representa la proporción de correos electrónicos clasificados en los niveles de riesgo alto, medio y bajo, de acuerdo con el mecanismo de mitigación académica sugerido. Elaborado por: Los Autores.

4.7.1 Ejemplos del resultado del mecanismo de mitigación

En la Tabla 12 se presentan ejemplos ilustrativos del resultado del mecanismo de mitigación académica.

Tabla 12.

Ejemplos del resultado del mecanismo de mitigación académica

probabilidad_phishing	nivel_riesgo	accion_simulada
0.998519662986842	Alto	Marcar como phishing
0.9984045484894912	Alto	Marcar como phishing
0.31657655126100803	Bajo	Permitir entrega
0.04570017071885457	Bajo	Permitir entrega
0.9998111987051288	Alto	Marcar como phishing

Nota. Ejemplos ilustrativos del funcionamiento del mecanismo de mitigación académica. Elaborado por: Los Autores.

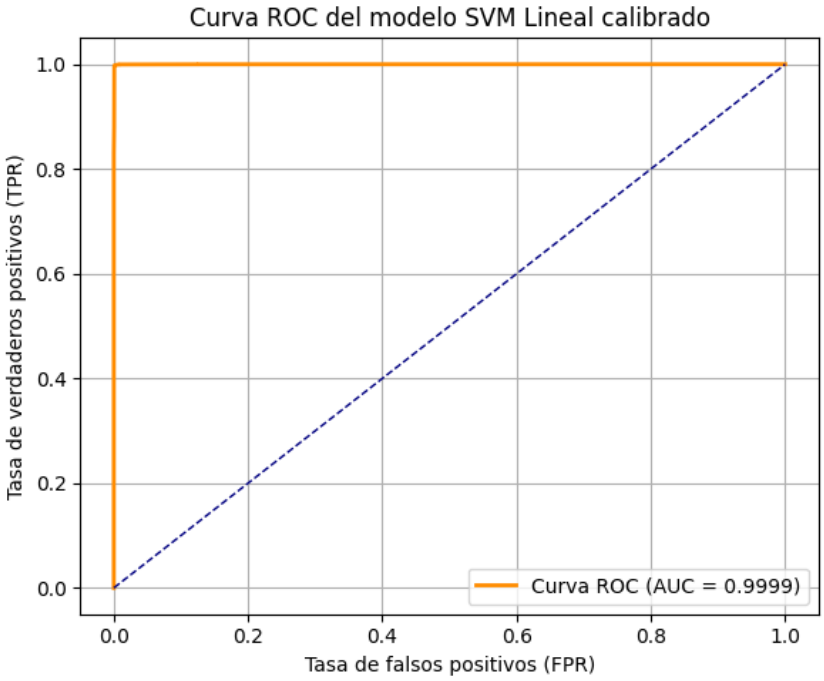
4.8 Evaluación del desempeño mediante la curva ROC

Para evaluar la capacidad discriminativa total del modelo final, se llevó a cabo un análisis la curva ROC (Receiver Operating Characteristic). Esta herramienta permite observar cómo se comporta el clasificador al enfrentar a diferentes umbrales de decisión, sin estar limitado a un solo punto de corte.

La Figura 13 presenta la curva ROC del modelo SVM lineal calibrado, logrando un área bajo la curva (AUC) de 0.9999, cifra que refleja una capacidad de distinción casi perfecta entre correos legítimos y correos phishing.

Figura 13.

Curva ROC del modelo SVM lineal calibrado



Nota. La curva ROC ilustra cómo se comporta el modelo en diferentes niveles de corte. Su proximidad al parte superior izquierdo indica un adecuado balance entre sensibilidad y tasa de falsos positivos. Elaborado por: Los Autores.

En la tabla 13 los valores determinados de FPR, TPR y los umbrales utilizados para crear la curva ROC.

Tabla 13.

Valores de FPR, TPR y umbrales utilizados

fpr	tpr	threshold
0.0	0.0	inf
0.0	0.000412	0.999999
0.0	0.001145	0.999998
0.0	0.002976	0.999994
0.003617	0.999993	0.999993

Nota. Valores utilizados para el análisis del desempeño del modelo a distintos puntos de corte. Elaborado por: Los Autores.

4.9 Validación operativa del sistema implementado

4.9.1 Interfaz web del sistema

La Figura 14 ilustra la interfaz web creada para la validación práctica del modelo. En esta pantalla se aprecia cómo el usuario puede ingresar un mensaje, mientras que el sistema muestra el resultado generado, incluyendo la probabilidad estimada de que el mensaje pertenezca a la clase phishing, así como el nivel de riesgo asociado.

Adicionalmente, se incluyen métricas del modelo recopiladas durante la fase de validación, lo que ayuda a contextualizar el rendimiento del clasificador en su entorno de ejecución. Esta evidencia demuestra que el modelo entrenado se integró de manera exitosa en una aplicación funcional, permitiendo su utilización en un entorno simulado para el análisis de correos electrónicos.

Figura 14.

Interfaz web del sistema de análisis de correos electrónicos

Nota. Captura de pantalla de la aplicación web desarrollada, en la que se muestra el texto introducido por el usuario, junto con la probabilidad estimada de phishing y el nivel de riesgo determinado por el modelo SVM lineal calibrado. Elaborado por: Los Autores.

4.9.2 Flujo operativo del sistema

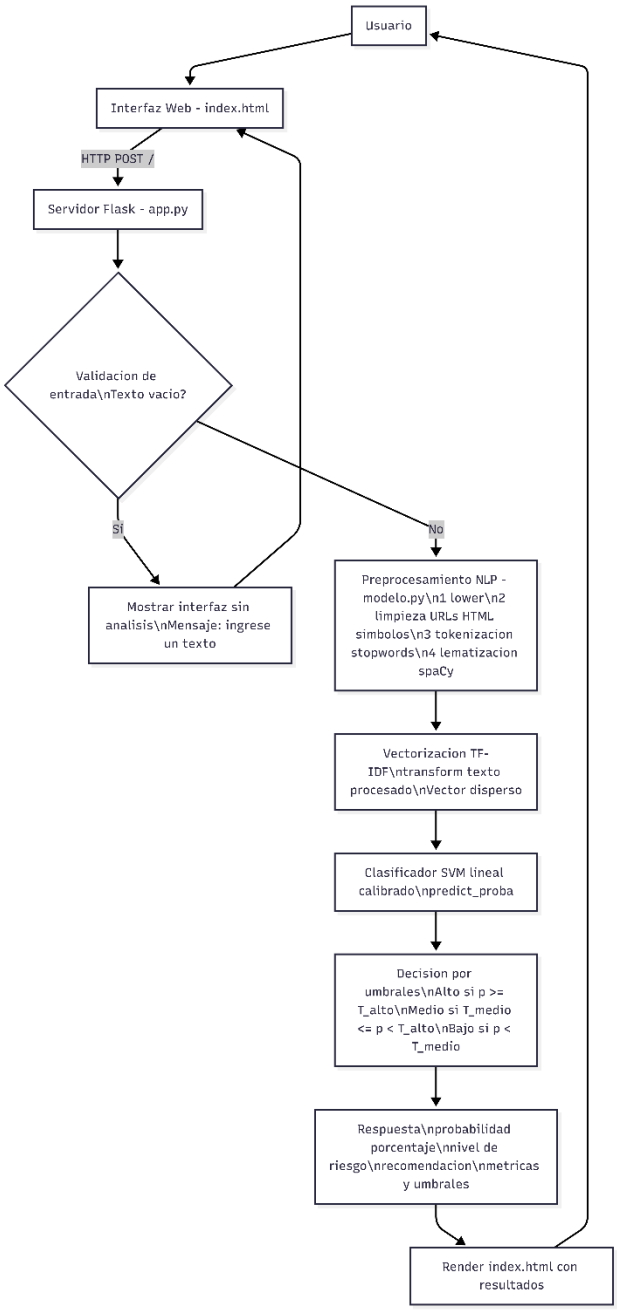
La Figura 15 ilustra el flujo operativo del sistema desarrollado. El proceso comienza con el ingreso de un texto por parte del usuario a través de la interfaz web. Luego, el servidor se encarga de procesar esta entrada mediante el módulo de preprocesamiento de lenguaje natural, que ejecuta tareas como limpieza, tokenización y lematización.

Tras la transformación del texto, se procede a vectorizarlo utilizando la técnica TF-IDF y se envía al clasificador SVM lineal calibrado, encargado de generar una probabilidad relacionada con la clasificación de la clase phishing. Finalmente, el sistema determina el nivel de riesgo según los umbrales establecidos y muestra el resultado directamente en la interfaz.

Este flujo destaca la integración total entre el modelo de aprendizaje automático y la aplicación web diseñada, demostrando la funcionalidad del sistema en un entorno operativo real.

Figura 15.

Flujo operativo del sistema de detección de phishing



Nota. Diagrama que ilustra el flujo de procesamiento desde la entrada proporcionada por el usuario hasta la creación del nivel de riesgo y la presentación de la respuesta en la interfaz web. Elaborado por: Los Autores.

CONCLUSIONES

El en evolución del proyecto se demostró que mezclar el N_L_P con modelos de M_L es una técnica muy segura para identificar correos electrónicos de phishing. Utilizando el conjunto de datos CEAS_08, se logró reconocer patrones tanto en el texto como en la estructura de los mensajes, lo que permite diferenciar entre correos genuinos y engañosos.

Durante el entrenamiento y pruebas, los tres modelos evaluados —Regresión Logística, Árbol de Decisión y SVM lineal— revelaron un desempeño positivo. No obstante, el modelo SVM lineal fue el que demostró el comportamiento más consistente entre los tres evaluados. Sus valores de *precisión, recall y F1-score* se mantuvieron por encima de los demás modelos, lo que sugiere una mejor capacidad para distinguir entre correos legítimos y fraudulentos.

Además, se incorporó un mecanismo de mitigación con base en umbrales probabilísticos, el cual permitió clasificar los correos electrónicos según su nivel de riesgo de manera ordenada. La división en niveles Alto, Medio y Bajo facilito la simulación de posibles acciones de respuesta ante mensajes sospechosos.

Con este trabajo se logró diseñar y comprobar el funcionamiento de un modelo para detectar phishing utilizando datos públicos. Esto no solo evito el uso de infraestructura reales, sino que también hace posible que el procedimiento pueda replicarse en estudios posteriores.

RECOMENDACIONES

Para futuras investigaciones, sería conveniente utilizar dataset más actuales y diversos. Así se podría analizar si el modelo mantiene su rendimiento ante nuevas técnicas de phishing y cambios en el lenguaje de los mensajes maliciosos, lo que ayudaría a validar su aplicación en contextos reales.

Como mejora futura, se podría probar el uso de embeddings contextuales y modelos de aprendizaje profundo para representar el texto. Estas técnicas suelen ofrecer una compresión más detallado del significado de os mensajes, lo que podría complementar o mejorar los resultados obtenidos con TF-IDF.

Como trabajo futuro, se podría estudiar el efecto de modificar el umbral de decisión del sistema. Al variar este valor, se pueden evaluar los cambios en los falsos positivos y falsos negativos, lo que ayudaría a adaptar el modelo a diferentes niveles de tolerancia al riesgo.

Como paso adicional, sería útil evaluar el modelo en escenarios simulados que se asemejen a sistemas reales de correos, sin intervenir directamente en infraestructuras institucionales. Así se podría observar cómo funcionan en condiciones más cercanas a la realidad, manteniendo un marco de trabajo ético y académico.

REFERENCIAS BIBLIOGRÁFICAS

Alam, N. A. (2024). *Phishing Email Dataset*. Kaggle.com. https://www.kaggle.com/datasets/naserabdullahalam/phishing-email-dataset?select=CEAS_08.csv

Alhogail, A., & Alsabih, A. (2021). Applying Machine Learning and Natural Language Processing to Detect Phishing Email. *Computers & Security*, *110*, 102414. https://doi.org/10.1016/j.cose.2021.102414

Alhuzali, A., Alloqmani, A., Aljabri, M., & Alharbi, F. (2025). In-Depth Analysis of Phishing Email Detection: Evaluating the Performance of Machine Learning and Deep Learning Models Across Multiple Datasets. *Applied Sciences*, *15*(6), 3396–3396. https://doi.org/10.3390/app15063396

Almomani, A., Gupta, B., Atawneh, S., Meulenberg, A., & Almomani, E. (2013). Un estudio sobre técnicas de filtrado de correo electrónico de phishing. IEEE Communications Surveys & Tutorials, 15(4), 2070–2090. https://ieeexplore.ieee.org/document/6489877

Basit, A., Zafar, M., Liu, X., Javed, A. R., Jalil, Z., & Kifayat, K. (2020). A comprehensive survey of AI-enabled phishing attacks detection techniques. *Telecommunication Systems*, *76*(1). https://doi.org/10.1007/s11235-020-00733-2

Bendale, R., & Patil, D. (2021). Detección de correos electrónicos de phishing mediante procesamiento del lenguaje natural y M_L. Revista Internacional de Investigación Científica en Ciencias de la Computación, Ingeniería y Tecnologías de la Información, 7(2), 234–240. https://doi.org/10.32628/CSEIT217248

Chugh, V. (2024, October). *AUC y Curva ROC en Aprendizaje Automático*. Datacamp.com; DataCamp. https://www.datacamp.com/es/tutorial/auc

Gupta, A., Kumar Mishra, A. K. M., & Arora, K. A. (2025). *Detecting Phishing Emails Using Natural Language Processing.* IEEE Xplore. https://ieeexplore-ieee-org.ecups.idm.oclc.org/stamp/stamp.jsp?tp=&arnumber=10941056

Instituto Nacional de Ciberseguridad de España (INCIBE). (2024a). ¿Qué es el phishing? Ciudadanía | INCIBE. https://www.incibe.es/ciudadania/blog/que-es-el-phishing

Instituto Nacional de Ciberseguridad de España (INCIBE). (2024b). Phishing. Aprende Ciberseguridad | INCIBE. https://www.incibe.es/aprendeciberseguridad/phishing

Instituto Nacional de Ciberseguridad de España (INCIBE). (2024c). Phishing de lanza. Aprende Ciberseguridad | INCIBE. https://www.incibe.es/aprendeciberseguridad/spear-phishing

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning*. Springer Nature.

John Tucker, director del programa de maestría en Telecomunicaciones de la Universidad Nacional de Loja (UNL). (s.f.). La ciberseguridad, uno de los problemas de la vida en internet. Universidad Nacional de Loja. https://www.unl.edu.ec/noticia/la-ciberseguridad-uno-de-los-problemas-de-la-vida-en-internet

Jurafsky, D., & Martin, J. (2024). *Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models Third Edition draft Summary of Contents*. Stanford University. https://web.stanford.edu/~jurafsky/slp3/ed3bookaug20_2024.pdf

Kyaw, P. H., Gutierrez, J., & Ghobakhlou, A. (2024). Una revisión sistemática de técnicas de aprendizaje profundo para la detección de correos electrónicos de phishing. *Electronics*, 13(19), 3823. https://doi.org/10.3390/electronics13193823

Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Detección de phishing basada en aprendizaje automático a partir de URL. *Expert Systems with Applications*, 117, 345–357. https://doi.org/10.1016/j.eswa.2018.09.029

Santosh Paradkar, N. (2023). *Phishing Email's Detection Using Machine Learning and Deep Learning*. IEEE Xplore. https://ieeexplore-ieee-org.ecups.idm.oclc.org/stamp/stamp.jsp?tp=&arnumber=10200493

Sastre, A. (2024, November 27). *Scikit Learn NLP: Clasificación de Texto*. CertiDevs. https://certidevs.com/tutorial-scikit-learn-nlp-clasificacion-texto

sklearn.metrics.roc_curve — scikit-learn 0.23.0 documentation. (n.d.). Scikit-Learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html

Tamal, M. A., Islam, M. K., Bhuiyan, T., Sattar, A., & Nayem Uddin Prince. (2024). Unveiling suspicious phishing attacks: enhancing detection with an optimal feature vectorization algorithm and supervised machine learning. *Frontiers in Computer Science*, *6*. https://doi.org/10.3389/fcomp.2024.1428013

Wainer, J. (2024). *An empirical evaluation of imbalanced data strategies from a practitioner's point of view*. Expert Systems with Applications. https://arxiv.org/pdf/1810.07168