FACULTAD DE HÁBITAT, INFRAESTRUCTURA Y CREATIVIDAD

INGENIERÍA EN SISTEMAS

MAESTRÍA EN SISTEMA DE INFORMACIÓN, MENCIÓN DATA SCIENCE

TEMA

Diseño y evaluación comparativa de modelos supervisados de machine learning para la detección de correos electrónicos de phishing en español

AUTOR: Ing. Jahel Liliana Nogales Carrera

DIRECTOR: Mgtr. Sebastián Felipe Tamayo Proaño

QUITO – 2026

Plan Final del Anteproyecto

1. Título del proyecto

Diseño y evaluación comparativa de modelos supervisados de machine learning para la detección de correos electrónicos de phishing en español.

2. Planteamiento del problema

El phishing por correo electrónico continúa siendo una amenaza relevante para la seguridad de la información, debido a que combina mecanismos técnicos como enlaces maliciosos, suplantación de identidad, adjuntos y redirecciones; con estrategias de ingeniería social orientadas a inducir al usuario a entregar credenciales, ejecutar acciones no autorizadas o interactuar con contenido fraudulento. Además, la detección basada exclusivamente en reglas, firmas o listas negras resulta limitada cuando los mensajes varían su redacción, ocultan enlaces, emplean dominios recientes o incorporan tácticas persuasivas difíciles de capturar mediante patrones estáticos.

En el contexto hispanohablante, la disponibilidad de corpus públicos de correos electrónicos con etiquetas y atributos técnicos o psicológicos ha sido más limitada que en inglés. Esta situación dificulta evaluar modelos reproducibles y comparables para detectar correos fraudulentos redactados en español, por lo que, resulta pertinente diseñar y evaluar modelos supervisados que integren variables asociadas a persuasión o personalización, utilizando conjuntos de datos secundarios documentados y de acceso público.

3. Pregunta de investigación

> ¿Qué modelo supervisado de machine learning, entrenado con representaciones textuales y variables técnicas de corpus públicos en español, ofrece el mejor resultado para clasificar correos electrónicos de phishing frente a correos legítimos?

4. Objetivos generales y específicos

4.1. Objetivo general

Diseñar, implementar y evaluar comparativamente modelos supervisados de machine learning para la clasificación de correos electrónicos legítimos y de phishing en español.

4.2. Objetivos específicos

- Caracterizar los conjuntos de datos seleccionados.
- Preparar el corpus mediante técnicas de preprocesamiento de datos.
- Implementar modelos de aprendizaje automática para la clasificación de correos electrónicos legítimos y de phishing.
- Evaluar el desempeño de los modelos.
- Analizar los errores de clasificación para identificar patrones de phishing que el modelo no detecta adecuadamente y discutir sus implicaciones técnicas y operativas.

5. Justificación del proyecto

El phishing mediante correo electrónico constituye una de las amenazas más relevantes para la seguridad de la información, debido a su capacidad para afectar a usuarios y organizaciones mediante técnicas de ingeniería social, suplantación de identidad y manipulación de elementos técnicos presentes en los mensajes digitales (Anti-Phishing Working Group [APWG], 2025, p. 9). De hecho, esta modalidad de ataque se ha intensificado con el uso creciente de servicios digitales y del correo electrónico como canal de comunicación institucional, comercial y personal, lo que amplía la superficie de exposición frente a mensajes fraudulentos (Carroll et al., 2022).

Además, el avance de herramientas basadas en inteligencia artificial generativa ha incrementado la posibilidad de crear mensajes más personalizados, coherentes y difíciles de identificar mediante señales tradicionales, como errores gramaticales, formatos sospechosos o redacciones poco naturales (Brightside AI, 2025). Por ello, resulta pertinente investigar mecanismos de detección más adaptativos que permitan analizar características presentes en los correos electrónicos como base para el diseño de modelos de aprendizaje automático orientados a la identificación temprana de posibles intentos de phishing (Mendoza Vega et al., 2025).

Por otra parte, el phishing genera consecuencias económicas, tecnológicas y organizacionales que pueden comprometer la seguridad y la continuidad operativa de las instituciones. Entre sus principales impactos se encuentran las pérdidas financieras, el robo de credenciales, la filtración de información sensible, el acceso no autorizado a sistemas internos y el deterioro de

la confianza de clientes, usuarios o miembros de la organización (Enzoic, 2025). Asimismo, estos ataques pueden convertirse en un punto de entrada para incidentes de mayor gravedad, especialmente cuando el usuario interactúa con enlaces maliciosos, entrega credenciales o descarga archivos comprometidos (Emmanuel y Oghie, 2026).

Desde esta perspectiva, técnicas como el acortamiento de URL, la ofuscación de enlaces y el uso de redirecciones dificultan la identificación del destino real del enlace y reducen la efectividad de mecanismos de detección basados únicamente en reglas estáticas, firmas o listas negras (Odgerel et al., 2025). En consecuencia, el desarrollo de modelos basados en aprendizaje automático puede aportar beneficios al área de ciberseguridad al facilitar el análisis automatizado de indicadores asociados al phishing y contribuir a una detección más oportuna de amenazas.

En conclusión, el análisis automatizado de correos electrónicos mediante modelos de aprendizaje automático representa una vía prometedora para superar algunas limitaciones de los sistemas de detección convencionales. Esto se debe a que dichos modelos pueden identificar patrones latentes en el contenido textual, los metadatos y los atributos estructurales de los mensajes, sin depender exclusivamente de reglas estáticas que requieren actualización manual constante (Lee et al., 2024). Por ello, la presente investigación propone el desarrollo y la evaluación comparativa de modelos supervisados de clasificación binaria sobre un conjunto de datos, con el propósito de contribuir a la detección temprana de correos electrónicos de phishing en entornos organizacionales.

6. Marco teórico

6.1. Phishing por correo electrónico

El phishing es una amenaza crítica de ciberseguridad basada en la ingeniería social, cuyo objetivo es manipular psicológicamente a los usuarios para que revelen información confidencial, como credenciales o datos financieros. El correo electrónico constituye su principal vector de ataque debido a su alcance masivo, bajo costo y facilidad de uso (Gangavarapu et al., 2020).

Los correos fraudulentos imitan la apariencia, el lenguaje y la identidad de organizaciones legítimas. Suelen incorporar enlaces maliciosos o archivos adjuntos diseñados para comprometer los sistemas. Las características principales para su detección se centran en identificar anomalías en:

- La cabecera del mensaje: Direcciones de remitente falsificadas o falta de autenticación de dominios.

- Las URLs: Enlaces que redirigen a sitios fraudulentos, uso de direcciones IP directas, o dominios acortados y ofuscados.
- El cuerpo del texto: Uso de lenguaje coercitivo, sentido de urgencia, solicitudes inusuales de credenciales y errores léxicos.

6.2. Métodos de detección de phishing en correos electrónicos

Los métodos de detección de phishing en correos electrónicos pueden clasificarse principalmente en dos enfoques: análisis de URLs y análisis de contenido. Ambos buscan identificar características distintivas de los mensajes maliciosos con el propósito de mejorar la precisión de los sistemas de detección. Para procesar este contenido no estructurado, se emplean técnicas de Procesamiento de Lenguaje Natural (NLP). Una herramienta matemática fundamental en este ámbito es TF-IDF (Term Frequency-Inverse Document Frequency), la cual permite vectorizar el texto. TF-IDF evalúa la relevancia de una palabra específica dentro de un correo electrónico en relación con su frecuencia en un conjunto completo de correos (corpus). Esta técnica permite a los algoritmos ponderar adecuadamente términos que son frecuentemente utilizados en ataques de phishing (por ejemplo, "urgente", "contraseña", "cuenta suspendida"), al mismo tiempo que penaliza palabras de uso común que carecen de valor discriminativo. Gracias al NLP y a la vectorización de texto, investigadores como Peng et al. (2018) han logrado precisiones superiores al 95 % en la identificación semántica de correos maliciosos.

6.3. Aprendizaje supervisado

El aprendizaje supervisado es una rama del aprendizaje automático en la que los modelos son entrenados utilizando conjuntos de datos etiquetados, es decir, datos en los que la variable de salida o clase objetivo es conocida previamente. El proceso se desarrolla en dos etapas principales: la fase de entrenamiento, en la cual el algoritmo aprende la relación existente entre las variables de entrada y las etiquetas asociadas, y la fase de prueba, donde se evalúa la capacidad del modelo para generalizar el conocimiento adquirido y realizar predicciones sobre datos no vistos anteriormente.

En los problemas de clasificación, las etiquetas representan categorías o clases específicas que permiten al algoritmo identificar patrones y construir una función capaz de asignar nuevas observaciones a la clase más adecuada. La elección del algoritmo influye drásticamente en el

rendimiento, dependiendo de la alta dimensionalidad característica del texto procesado. En la Tabla 1, se comparan los clasificadores más pertinentes:

Tabla 1. ***Algoritmos de clasificación supervisada***

Algoritmo	Fundamento	Detección	Limitaciones
Naive Bayes	Teorema de Bayes	Rápido y eficiente	Asume independencia irreal
SVM	Hiperplano óptimo	Alta precisión	Costoso y requiere ajuste
Decision Tree	Reglas jerárquicas	Interpretable y versátil	Propenso a sobreajuste
KNN	Distancia espacial	Implementación simple	Lento con muchos datos

Nota: Adaptación de conceptos fundamentales de aprendizaje automático para la clasificación de texto.

6.4. Evaluación de Modelos de Clasificación

La validación de modelos de clasificación exige el uso de múltiples métricas de desempeño, dado que un único indicador no captura la complejidad total del sistema. El análisis debe ser multidimensional para evitar sesgos y garantizar una evaluación rigurosa del comportamiento algorítmico, a continuación se listan dichos indicadores:

- **Accuracy:** Proporción global de aciertos, puede resultar engañosa en conjuntos de datos desbalanceados.
- **Precision:** Capacidad del modelo para identificar correctamente los casos positivos y minimizar los falsos positivos.
- **Recall:** Capacidad de detectar todos los casos positivos reales, minimiza falsos negativos.
- **F1-Score:** Media armónica entre Precisión y Recall, ofrece un equilibrio balanceado en escenarios complejos.

- **AUC-ROC:** Representa la capacidad de discriminación del modelo ante distintos umbrales de decisión. Es la medida más robusta frente al desbalance de clases.

6.5. Metodología CRISP-DM

La metodología CRISP-DM se fundamenta como un modelo de procesos estructurado destinado a guiar proyectos de minería de datos e inteligencia analítica (Mariscal et al., 2010) mediante la ejecución secuencial seis fases iterativas:

1. **Comprensión del Negocio:** Definición del problema del phishing como un reto algorítmico.
2. **Comprensión de los Datos:** Exploración del corpus de correos electrónicos.
3. **Preparación de los Datos:** Limpieza y vectorización del texto mediante NLP y TF-IDF.
4. **Modelado:** Entrenamiento de los algoritmos de clasificación.
5. **Evaluación:** Validación rigurosa del rendimiento utilizando métricas como F1-score y AUC-ROC.

Despliegue: Implementación del modelo en el entorno operativo para el filtrado continuo.

7. Metodología de investigación

7.1. Enfoque, alcance y diseño

La investigación tendrá un enfoque cuantitativo, alcance aplicado y diseño experimental-comparativo sobre datos secundarios. Es cuantitativa porque mide el desempeño de modelos mediante métricas numéricas; es aplicada porque aborda un problema práctico de ciberseguridad; y es comparativa porque evalúa varios algoritmos bajo un procedimiento común. El estudio no realizará inferencia causal; por tanto, las conclusiones se limitarán al desempeño predictivo observado en los corpus analizados.

7.2. Fuentes de datos

Se utilizarán dos conjuntos de datos secundarios de acceso público alojados en Mendeley Data:

- SpaPhish: corpus en español con 1.395 correos electrónicos, 47 variables, etiqueta binaria, 664 correos legítimos y 731 correos de phishing. Incluye asunto, cuerpo, fecha, atributos técnicos y anotaciones psicológicas asociadas a principios de persuasión.
- SpearPhishMX: corpus en español con 3.006 correos electrónicos etiquetados para clasificación binaria, donde la clase positiva corresponde a spear-phishing y la clase negativa a correos legítimos. Incluye asunto, cuerpo sanitizado, variables relacionadas con URLs y señales de personalización dirigida.

Antes de combinar los conjuntos de datos, se realizará una armonización de campos y etiquetas. Además, se considerarán tres escenarios experimentales: evaluación en SpaPhish, evaluación en SpearPhishMX y evaluación sobre corpus combinado. De ser viable, se añadirá validación cruzada entre datasets para estimar la capacidad de generalización ante diferencias de dominio.

7.3. Variables del estudio

La variable objetivo será binaria: phishing o spear-phishing como clase positiva (1) y correo legítimo como clase negativa (0).

No se utilizarán atributos que puedan introducir fuga de información, tales como identificadores derivados directamente de la etiqueta o variables generadas a partir de decisiones de anotación no disponibles en un escenario real de predicción, salvo que se justifique su uso como experimento separado de explicabilidad.

7.4. Consideraciones éticas y de seguridad

El estudio tendrá finalidad defensiva y académica. No se reactivarán enlaces, no se ejecutarán adjuntos y se utilizarán únicamente representaciones sanitizadas cuando el dataset lo indique. La investigación deberá respetar las licencias de los datos, citar los repositorios originales y evitar la exposición de información personal. Además, los resultados no se utilizarán para generar mensajes de phishing ni para optimizar tácticas ofensivas.

8. Cronograma de actividades

Fase	Actividad	Abril	Mayo				Junio				Julio				Agosto		Responsable
		S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	
Planificación	Definición del problema, objetivos y alcance																Tesista
Planificación	Revisión bibliográfica y estado del arte																Tesista
Marco Teórico	Elaboración del marco teórico y antecedentes																Tesista
Comprensión del negocio	Comprensión del problema, identificación de información requerida,																

Fase	Actividad	Abril	Mayo				Junio				Julio				Agosto		Responsable
		S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	
	planteamiento de enfoque y alcance																
Comprensión de datos	Identificación y selección de datasets																Tesista
Preparación de datos	Limpieza, transformación e ingeniería de características																Tesista
Modelado	Implementación de algoritmos de machine learning																Tesista
Modelado	Entrenamiento y ajuste de modelos																Tesista
Evaluación Englobar con los resultados	Evaluación mediante Accuracy, Precision, Recall y F1-Score																Tesista

Fase	Actividad	Abril	Mayo				Junio				Julio				Agosto		Responsable
		S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	
Resultados	Análisis e interpretación de resultados																Tesista
Cierre	Elaboración de conclusiones y recomendaciones																Tesista
Cierre	Preparación de presentación y defensa																Tesista

9. Recursos

Categoría	Recurso	Uso previsto
Humanos	Tesista	Diseño metodológico, programación, análisis y redacción.
Humanos	Director de tesis	Supervisión técnica, metodológica y académica.
Técnicos	Python, Jupyter Notebook, Pandas, NumPy	Procesamiento, análisis exploratorio y experimentación.
Técnicos	Scikit-learn	Pipelines, vectorización, entrenamiento y evaluación de modelos.
Técnicos	Matplotlib y Seaborn	Visualización de resultados y métricas.
Técnicos	Zotero o gestor bibliográfico	Gestión de referencias en formato APA 7.
Financieros	Equipo informático e Internet	Acceso a datos, bibliografía y ejecución de experimentos.