UNIVERSIDAD SEÑOR DE SIPÁN

FACULTAD DE INGENIERÍA, ARQUITECTURA Y URBANISMO

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

TESIS

IMPLEMENTACIÓN DE UN MÉTODO DE CLASIFICACIÓN DE MINERÍA DE DATOS PARA DETECTAR PÁGINAS WEB DE TIPO PHISHING

PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO DE SISTEMAS

Autor(a) (es):

Bach. Maguiña Maza, Jean Carlos

ORCID: 0000-0002-1477-9073

Bach. Soto Calderón, José Luis

ORCID: 0000-0002-1688-9727

Asesor:

Mg. Ing. Bances Saavedra, David Enrique

ORCID: 0000-0001-7164-8918

Línea de Investigación:

Infraestructura, Tecnología y Medio Ambiente

Pimentel – Perú

Año 2020

APROBACIÓN DEL JURADO

IMPLEMENTACIÓN DE UN MÉTODO DE CLASIFICACIÓN DE MINERÍA DE DATOS PARA DETECTAR PÁGINAS WEB DE TIPO PHISHING

_____

Maguiña Maza Jean Carlos

Autor

_____

Soto Calderón José Luis

Autor

_____

Mg. Ing. Bances Saavedra, David Enrique

Asesor

_____ _____

Dr. Ramos Moscol Mario Fernando

Presidente de Jurado

Mg. Mejía Cabrera Heber Iván

Secretario de Jurado

_____

Mg. Diaz Vidarte Miguel Orlando

Vocal

Dedicatorias

A todas las personas que hicieron posible la realización de este trabajo.

Agradecimientos

A nuestras familias, educadores y a nuestra institución. Un reconocimiento especial a mi amigo Jean Carlos Maguiña Maza coautor de este trabajo. ¡Este logro va para ti!

Resumen

En Latinoamérica, el uso de plataformas virtuales no tenía mayor relevancia que los canales tradicionales. Sin embargo, con la problemática mundial respecto al COVID-19, y el confinamiento, que casi la mayoría de países adoptaron, el canal virtual tuvo un incremento exponencial nunca antes visto, y con ello también la ciberdelincuencia. En la actualidad, una de las estafas online más utilizada es el Phishing, páginas idénticas que se construyen para engañar al usuario, y obtener información personal sensible, suplantarlos y robar su dinero o extorsionarlos. Por consiguiente, desde hace un buen tiempo, se vienen desarrollando herramientas para poder combatir el Phishing, mismas que parten por reconocer patrones que logren caracterizar la página web como fraudulenta. Sin embargo, así como evolucionan las técnicas anti-phishing, también evolucionan las técnicas de suplantación. Por lo que los métodos de detección pierden vigencia, y ya no detectan correctamente. Es por ello, que el presente trabajo implementa un método de detección de páginas web utilizando minería de datos, con base en un análisis teórico de la literatura y la selección de los 3 mejores métodos con una excelente precisión. Así como también, la selección de los 32 atributos más utilizados en los 10 mejores métodos de clasificación de páginas web de tipo Phishing. El resultado muestra cifras muy positivas, que además se han puesto a prueba con las 3 mejores técnicas de la actualidad que son AdaBoost, SVM y XGBoost, los mismos que han logrado una precisión de 94%, 95% y 99% respectivamente. Además, el consumo de recursos de los 3 clasificadores mencionados fue en CPU: AdaBoost 43.17%, SVM 15.5% y XGBoost 21.71%. Con respecto a la RAM: AdaBoost consumió 409MB, SVM 17.64MB y XGBoost 4MB. En tal sentido, XGBoost ha tenido un desempeño sobresaliente en su técnica, por formar grupos de datos bien definidos usando técnicas de dimensionamiento y con clasificadores simples, y además con un eficiente uso de recursos computacionales.

Palabras claves:

Plataformas virtuales, Ciberdelincuencia, Phishing, Suplantación, Patrones, Atributos, Precisión (en estadística), Dimensionamiento, Clasificadores

Abstract

In Latin America, the use of virtual platforms was not more relevant than traditional channels. However, with the global problem of COVID-19, and the confinement, which almost all countries adopted, the virtual channel had an exponential increase never seen before, and with it also the cybercrime. Currently, one of the most widely used online scams is Phishing, identical pages that are built to trick users, and obtain sensitive personal information, impersonate them and steal their money or extort them. Therefore, for some time now, tools have been developed to combat phishing, which are based on recognizing patterns that characterize a Web page as fraudulent. However, as anti-phishing techniques evolve, so do phishing techniques. Therefore, detection methods are no longer valid, but are now correctly detected. For this reason, this paper implements a method of web page detection using data mining, based on a theoretical analysis of the literature and the selection of the 3 best methods with excellent accuracy. Also, the selection of the 32 most used attributes in the 10 best methods of classification of web pages of the Phishing type. The result shows very positive figures, which have also been tested with the 3 best techniques of the moment, which are AdaBoost, SVM and XGBoost, the same ones that have achieved an accuracy of 94%, 95% and 99% respectively. In addition, the resource consumption of the 3 classifiers mentioned was in CPU: AdaBoost 43.17%, SVM 15.5% and XGBoost 21.71%. Regarding RAM: AdaBoost consumed 409MB, SVM 17.64MB and XGBoost 4MB. In this sense, XGBoost has had an outstanding performance in its technique, for forming well-defined data groups using dimensioning techniques and with simple classifiers, and also with an efficient use of computational resources.

Keywords:

Virtual platforms, Cybercrime, Phishing, Impersonation, Patterns, Attributes, Accuracy (in statistics), Dimensioning, Classifiers.

INDICE

I INTRODUCCIÓN 9

1.1 Realidad Problemática 9

1.2 Antecedentes de estudio 14

1.3 Teorías relacionadas al tema 19

1.4 Formulación del Problema 31

1.5 Justificación e importancia del estudio 31

1.6 Hipótesis 32

1.7 Objetivos 32

1.7.1 Objetivo General 33

1.7.2 Objetivos Específicos 33

II MATERIAL Y MÉTODO 33

2.1 Tipo y diseño de investigación 33

2.2 Población y muestra 33

2.3 Variables, Operacionalización 35

2.4 Técnicas e instrumentos de recolección de datos, validez y confiabilidad 36

2.5 Procedimiento de análisis de datos 37

2.6 Criterios éticos 38

2.7 Criterios de rigor científico 38

III RESULTADOS 39

3.1 Resultados en Tablas y Figuras 39

3.2 Discusión de resultados 43

3.3 Aporte práctico 45

IV CONCLUSIONES Y RECOMENDACIONES 63

REFERENCIAS 65

ANEXOS 70

Lista de tablas

Tabla 1. TOP 10 Países Objetivos del Phishing según RSA. 12
Tabla 2. Clasificación de método de selección de características. Elaboración propia. 31
Tabla 3. Lista de Métodos de clasificación de minería de datos. Elaboración propia. 34
Tabla 4. Fórmulas para los indicadores. 35
Tabla 5. Análisis de resultados de otras técnicas similares. Elaboración propia. . 43
Tabla 6. Análisis de rendimiento de los métodos de clasificación. Elaboración propia. 45
Tabla 7. Criterios de selección para la selección de los mejores métodos. Elaboración propia. 46
Tabla 8. Métodos seleccionados con los criterios aplicados. Elaboración propia. 46
Tabla 9. Muestra de un extracto de la base de datos original de 11054 registros. 49
Tabla 10. Análisis teórico de las diferencias entre los métodos de selección de características. Elaboración propia basada en las definiciones de (Masters, 2019). 50
Tabla 11. Atributos considerados en el dataset. Elaboración propia. 52
Tabla 12. Atributos considerados en el dataset. Elaboración propia. 53
Tabla 13. Atributos considerados en el dataset. Elaboración propia. 54
Tabla 14. Resultados de la separación del dataset. Elaboración propia. 55
Tabla 15. Interpretación de la Matriz de confusión. Elaboración propia. 58
Tabla 16. Matriz de Confusión - AdaBoost. Elaboración propia. 59
Tabla 17 Matriz de Confusión - SVM. Elaboración propia. 59
Tabla 18 Matriz de Confusión - XGBoost. Elaboración propia. 59
Tabla 19 Métricas de rendimiento con Adaboost. Elaboración propia. 60
Tabla 20 Métricas de rendimiento con SVM. Elaboración propia. 61
Tabla 21 Métricas de rendimiento con XGBoost. Elaboración propia. 62
Tabla 22. Ficha de registro de datos 71
Tabla 23. Ficha de Registro de Resultados 72

Lista de figuras

Figura 1. Conocimiento de Phishing a nivel Global. (Proofpoint, 2020) 11
Figura 2. Porcentaje de cumplimiento según NCSI. 12
Figura 3. División de las técnicas de Minería de datos (Ñaupa Caraza, 2016) 21
Figura 4. Estructura de Árbol de decisión. Elaboración propia. 23
Figura 5. Diagrama de Redes Bayesianas. Elaboración propia. 23
Figura 6. Clasificación de método de algoritmo SVM. Elaboración propia. 25
Figura 7. Gráfica SVM con el hiperplano para separar las clases. 25
Figura 8. Ejecución de AdaBoost. (James, Witten, Hastie, & Tibshirani, 2017). .. 27
Figura 9. Cálculo con XGBoost. (James, Witten, Hastie, & Tibshirani, 2017). 29
Figura 10. Etapas de la metodología KDD. 30
Figura 11. Grado de consumo de CPU durante la ejecución de las pruebas. Elaboración propia. 42
Figura 12. Consumo en Megabytes de RAM durante la ejecución de las pruebas. Elaboración propia. 42
Figura 13. Método de selección de características. 47
Figura 14. Cantidad registros por su clasificación. Elaboración propia. 51
Figura 15. Arreglo de base X_train y X_test. Elaboración propia. 56
Figura 16. Matriz de confusión 58

I INTRODUCCIÓN

1.1 Realidad Problemática

Desde siempre, han existido personas que hacen el bien y el mal. Esto se da en todos los contextos de la humanidad, en la política, en las ideologías sociales, en la salud y por supuesto en el ámbito de la tecnología. En la medida que evoluciona el conocimiento, también evoluciona la tecnología con una gran velocidad, este cambio requiere que los usuarios se adapten continuamente.

Los expertos en tecnologías desarrollan nuevos métodos para procesar, distribuir información entre los millones de usuarios. Muchos de los expertos motivados por el avance del conocimiento, utilizan su talento para favorecer el avance de la tecnología, mientras otros, desarrollan métodos para cometer delitos, valiéndose de la tecnología. (Real Academia Española, 2020). En el mundo de la cibernética, estos delitos, se les denomina ciberdelincuencia, que es la acción de ir en contra de los sistemas de cómputo y de la información con el propósito de tener acceso no autorizado a los dispositivos o información, así como también restringir el acceso al usuario legítimo. (Interpol, 2020).

En la actualidad, una de las estafas online más utilizada es el Phishing, en donde a los ciberdelincuentes se les denomina Phishers, se hacen pasar por las entidades legítimas para engañar al usuario, con la única intención de obtener información personal sensible. Esta información es explotable y sirven en todas sus formas para sacar provecho económico y/o sustraigan información sensible. (UNAM-CERT, 2019)

El Phishing evoluciona al avance de la tecnología y a los nuevos medios por los cuales puedan obtener información. Sus variantes sorprenden a muchos, como el Vishing, el Smishing, el Spear Phishing, el Pharming y el más popular: el Phishing web. (Gonzales, 2020)

El Phishing web es la forma preferida por los ciberdelincuentes. Esta, consiste en publicar en internet una página web fraudulenta, muy parecida a la legítima, para

capturar la información que ingresas en ella, tales como datos personales, números de identificación nacional, correos y claves, números de tarjetas de pago, dirección de casa, etc.

En el mundo entero se han identificado millones de estas páginas gracias a la denuncia de personas que cayeron en la trampa y otros gracias a sistemas de monitoreo y detección. Los clientes de las entidades financieras, que realizan transacciones por internet, son las víctimas más frecuentes, sin embargo, se puede generalizar, que el ataque estará dirigido a cualquier usuario, de una plataforma web, que realice una transacción financiera en ella, o en cualquier otra página web legítima de registro de información, para obtener un beneficio social económico. (Malwarebytes, 2020).

El objetivo principal del Phisher, es obtener dinero ilícitamente a través de la suplantación de identidad. En el Perú, según la División de Investigación de Alta Tecnología DIVINDAT de la policía, informó en un diario, que este tipo de delitos crecieron un 8% durante el 2019 respecto al año anterior, llegando a 2097 denuncias. (Gestión Perú, 2020).

Lamentablemente, uno de los problemas para recopilar información es que si bien, las denuncias han aumentado, aún son muy pocas. En el mundo en general, la tendencia es que las personas no hagan la denuncia de las páginas fraudulentas. (Agencia Peruana de Noticias, 2020)

El Phishing es una técnica de ataque que se vale de otras técnicas y varía en el tiempo constantemente para evitar ser detectada. Es por ello que distintos especialistas alrededor del mundo crean algoritmos que ayuden a detectar con mejor precisión el Phishing con el fin de prevenir el fraude electrónico. Sin embargo, se debe tener en cuenta que los Phishers desarrollan nuevos mecanismos de ataque constantemente y mejoran sus estrategias para seguir delinquiendo.

Según el reporte de Seguridad de Microsoft del 2019, para el primer mes del año, las detecciones de ataques de tipo phishing llegó a 225,000 diarios, en febrero subió a 300,000. Y dado que normalmente, el phishing funciona con campañas, solo en el día de los enamorados, se llegó en un día a las 480,000 detecciones. (Microsoft Prensa, 2019)

Cabe mencionar que la empresa Proof Point emitió un informe anual sobre el conocimiento del phishing a nivel global; dando como resultado que el 61% de las personas tienen conocimiento de lo que significa phishing, el 24% no tiene conocimiento y el 15% no conoce la terminología; por lo que esta estadística hace prever que sean mayores las víctimas de esta modalidad fraudulenta. (Proofpoint, 2020).

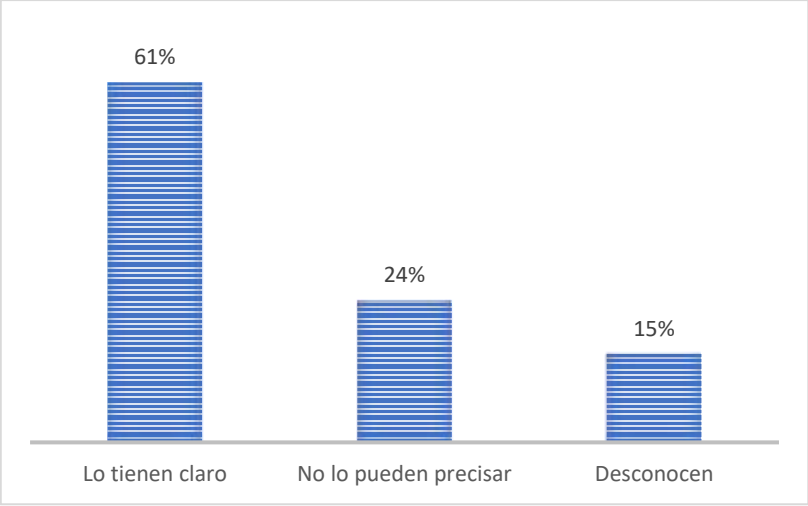


Figura 1. Conocimiento de Phishing a nivel Global. *(Proofpoint, 2020)*

Según la National CiberSecurity Index (NCSI), el Perú ocupa el puesto 70º; donde para esta institución la protección de datos personales y la identificación electrónica y servicios confiables son los puntos más fuertes de este ranking; además se puede apreciar que en los puntos de desarrollo de políticas de ciber seguridad, la información sobre el análisis de ciber amenazas, servicios de protección digital y servicios esenciales de protección, cuentan con una calificación bastante mala (0%); según la siguiente imagen. (NCSI, 2020)

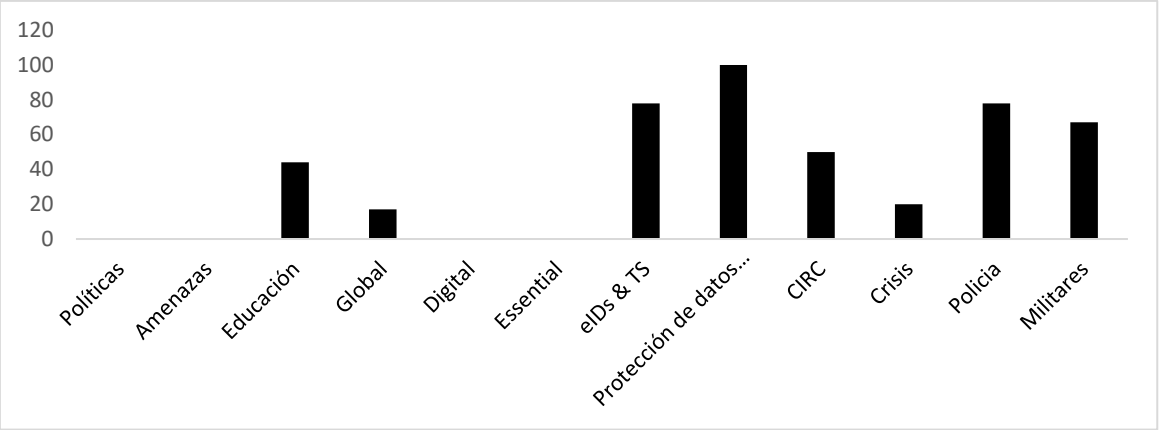


Figura 2. Porcentaje de cumplimiento según NCSI.

Según el boletín electrónico emitido en el 2018, por la empresa de seguridad RSA el Perú ocupa el 9º puesto del Top Ten de países que registran más ataques de Phishing. (RSA, 2018)

Tabla 1. TOP 10 Países Objetivos del Phishing según RSA.

País	Puesto
Canadá	1
Estados Unidos	2
Holanda	3
India	4
España	5
Brasil	6
Colombia	7
Francia	8
Perú	9
México	10

Es por ello que resulta lógico que muchas empresas, universidades, gobiernos y hasta profesionales entendidos del tema, dediquen tiempo y esfuerzo en innovar en soluciones o mejorarlas, y aunque se sabe que nada es 100% seguro, resulta alentador que los esfuerzos por combatirlos hayan crecido en el tiempo.

En la actualidad, las técnicas de detección de phishing se basan en la capacidad de identificar el ataque antes de que éste logré masificarse, e inclusive antes de que se tenga víctimas que lamentar. Es por ello que, con la gran cantidad de ataques debidamente registrados, es posible extraer sus datos y explotarlos con la ayuda de la minería de datos predictiva, que permitirá identificar los patrones de las páginas web de tipo phishing, y poder predecir oportunamente su clasificación.

En el análisis realizado, se encuentran las técnicas basadas en reconocimiento de patrones por diseño, reconocimiento de texto, análisis de URL simple y encriptado, en base a campañas, por reconocimiento de IP, de análisis por combinación de algoritmos. La mayoría de técnica de detección de phishing utilizan la minería de datos, dada su capacidad para la predicción usando algoritmos bastantes conocidos, tales como Árbol de decisión, Redes bayesianas, AdaBoost, XGBoost, Redes neuronales, Support Vector Machine (SVM), etc.

Es sabido también, que mientras más sofisticado se haga el ataque, se hace más complejo el algoritmo para la detección del mismo, por la cantidad de variables que pueden tener. Es precisamente, en esa búsqueda interminable de solución que se ha complicado el escenario, pues el uso de recursos y sistemas más complejos ha hecho que se vuelva ineficiente para unos y eficiente para otros. Siendo así, todas las técnicas conocidas resultan ser eficaces, pero no necesariamente eficientes, pues lo que se da a conocer son las pruebas y resultados de estos métodos en escenarios bien específicos, pero no se da a conocer bajo qué condiciones son más eficaces, e incluso llegan a ser eficientes a un nivel aceptable. En algunos inclusive, se evidencias resultados distintos y nada aceptables para un mismo método con distinta muestra. Así también, realizan comparaciones de los métodos con una única muestra, pero no muestran resultados de métodos mixtos.

Si bien, las técnicas utilizadas en distintas pruebas muestran los resultados obtenidos a nivel de efectividad de la técnica, no indican la cantidad de recursos que consumen ante pruebas con muestras de mayor volumen o mayor complejidad de análisis.

1.2 Antecedentes de estudio

Tengku Tengku Abd Rashid, Jamaludin bin Sallim and Yusnita binti Muhamad Noor, (2020), realizó la investigación, *A comparative Analysis on Artificial Intelligence Techniques for Web Phishing Classification*, en la Universiti Malaysia Pahang. Hay una gran variedad de algoritmos de clasificación que ayudan a proteger a los usuarios a través de herramientas, bloqueadores, plugins, a fin de controlar y/o bloquear el acceso a contenido web no requerido, inapropiado, y hasta a prevenir el acceso no autorizado que podría comprometer la seguridad de la red, sin embargo, los estudios no están mostrando la suficiente comparación de técnicas de clasificación para la detección de páginas web de tipo phishing y tampoco el análisis que se debe hacer entre ellas. En ese sentido, los autores probaron una base de datos con páginas web phishing con los siguientes 3 algoritmos de clasificación: redes neuronales, árboles de decisión y la técnica de vectores de soporte (SVM), a fin de identificar la técnica con mejor precisión, pero con el menor esfuerzo computacional para clasificar una página web. Los resultados arrojaron que la técnica de Artificial Neural Network tuvo un mejor resultado en las tres características medidas: Precisión 96%, Recall 96% y F-Measure 96%. En términos generales, los 3 clasificadores dieron buenos resultados, durante el procesamiento se seleccionaron ciertas características, y los 3 algoritmos de clasificación funcionan bien, sin embargo, si se requiere una precisión más alta, es mejor usar la técnica de detección de redes neuronales sin la selección de características, pues así el rendimiento será mejor.

Abdulhamit Subasi, Emir Kremic, (2020), realizó la investigación, *Comparison of Adaboost with Multiboosting for Phishing Website Detection*, en la Effat University, College of Engineering de Arabia Saudita. Muchas empresas de Software han implementado novedosas técnicas de detección como heurística, listas negras, reconocimiento visual y métodos de aprendizaje automático, pero con una sola no es posible eliminar el ataque por completo. Por ello, en este estudio los autores presentan un marco de detección inteligente de páginas web de tipo phishing, donde se emplean diferentes métodos de aprendizaje automático para determinar si la página es legítima o Phishing. Como resultado se tuvo que Adaboost con

Support Vector Machine (SVM) tuvo la mejor precisión logrando 97.61% que los demás clasificadores. Con este artículo los autores propusieron juntar métodos de aprendizaje automático: Adaboost y Multiboost, con los que lograron mejorar el resultado de precisión y eficiencia de los algoritmos de clasificación.

Jian Mao, Jingdong Bian, Weqian Tian, Shishi Zhu, Tao Wei, Aili Li, Zhenkai Liang, (2018), realizaron la investigación, *Detecting Phishing Websites via Aggreagation Analysis of Page Layouts*, en la School of Electronic and Information Engineering, Beihang University de China. Los autores indican que los mecanismos de aprendizaje automático enfocados a la detección de páginas web de tipo phishing no son suficientes para detectar nuevas técnicas de ataque. Por tal razón, proponen un método automático de aprendizaje por agregación para identificar mediante la similitud del diseño de la página legítima, si es una página phishing. Según los resultados obtenidos, ambas técnicas SVM y Decision Tree tuvieron un resultado mayor a 93% de exactitud y más del 95% de precisión, por lo que el enfoque de los autores, para detectar páginas web de tipo phishing, es efectivo. La técnica permite auto aprender de las similitudes de las páginas, recopilando las características del diseño, pudiendo mejorar el rendimiento de otros mecanismos antiphishing existentes.

A.A. Orunsolu, A.S. Sodiya, A.T. Akinwale, (2019), realizarón la investigación, *A predective model for phishing detection*, en Moshood Abiola Polytechmic de Nigeria. A pesar de la existencia de muchísimos mecanismos y sistemas de detección de phishing, estos siguen siendo ineficientes para detecciones en el día CERO del ataque. Por consiguiente, los autores presentan este proyecto dando una mejora basada en el aprendizaje automático, pero de manera predictiva, utilizando como base 2 algoritmos: Support Vector Machine (SVM) y Redes bayesianas. Como resultado, se obtuvo un 99.96% de precisión. La propuesta de los autores logra la construcción de un modelo de que va construyendo incrementalmente el dataset, logrando ser un sistema flexible y gestionable, y, por consiguiente, superior comparándolo con otros modelos bajo el mismo escenario.

Opara Chidimma Ugochi, (2018), realizó la investigación, *A novel web page anti-phishing approach using URL cosine similarity and IP Address comparison*, en Teeside University de United Kingdom. En la siguiente investigación, el autor indica que la mayoría de técnicas de detección de páginas web de tipo phishing, utilizan la extracción de características de la página, causando que el método sea más complejo impactando en el uso de recursos y el tiempo de análisis. En tal sentido, el autor propone una novedosa solución, utilizando la extracción de la URL, para compararla con la URL legítima. Para esta comparación, el autor utiliza un método de similitud por coseno, las URLs son elevadas con la fórmula de coseno, el resultado es comparado. Además, se comparan las direcciones IP detectadas, y con ambos resultados, resultado del coseno y la dirección IP, se concluye con una respuesta final: legítima o phishing. Los resultados muestran un 100% de efectividad en tiempo real. Por lo que es importante saber que este método es efectivo siempre que las URLs comparadas contengan el nombre de la entidad.

Erzhou Yuyang Chen, Chengcheng Ye, Xuejun Li, Feng Liu, (2019), realizaron la investigación, *OFS-NN: An Effective Phishing Websites Detection Model Based on Optimal Feature Selection and Neural Network*, en Anbui University de China. En este artículo los autores ponen en evidencia el problema que existe en los modelos de detección, ya que al momento de formar los dataset, muchas veces de consideran muchas características que por lo general son irrelevantes haciendo poco eficientes a los algoritmos disminuyendo la efectividad, es decir, no detectan adecuadamente, pero no porque el algoritmo sea malo, sino porque están mal calibrados. En ese sentido, los autores proponen un modelo que optimiza la selección de características, llamado OFS-NN utilizando Redes neuronales. En consecuencia, lo propuesto es preciso y estable para detectar los diferentes tipos de páginas web de tipo phishing. En tal sentido, el modelo optimiza la selección de características a fin de tener un óptimo resultado utilizando el algoritmo de redes neuronales, de manera que se puede entrenar eficientemente al vector para la detección de páginas web de tipo phishing.

Walled Ali, Sharaf Malebary, (2020), realizaron la investigación, *Particle Swarm Optimization-Based Feature Weighting for Improving Intelligent Phishing Website*

Detection, en King Abdulaziz University de Arabia Saudita. Los autores reconocen que hay muchas técnicas sofisticadas para la detección de páginas web de tipo phishing en la actualidad, sin embargo, ninguna es eficiente cuando se trata de identificar una página web de tipo phishing en el día CERO. La investigación propone utilizar maching learning con una técnica llamada Particle swarm optimization (PSO), que permitirá dar un mayor peso en la evaluación, a las características que, con mayor frecuencia, tipifican a una web como Phishing, de manera tal, que se pueda identificar con mayor precisión en tiempo real. Como resultado, se tuvo que utilizar la técnica PSO logra mejorar la precisión con la que el modelo clasifica entre legítima o fraudulenta, pero con menos esfuerzo del algoritmo, en reunir más características, pudiendo también mejorar las tasas de falsos positivos y viceversa. Por consiguiente, la investigación concluye que, con los pesos ideales, a las características más frecuentes en una web de tipo phishing, usando PSO, todos los algoritmos probados en esta investigación mejoraron considerablemente su nivel de precisión.

Peng Yang, Guangzhen Zhao, Peng Zeng, (2019), realizaron la investigación, *Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning*, en Southeast University de China. El presente artículo menciona que en todo modelo de detección por características para detectar páginas web de tipo phishing, es crítico el conocimiento del tema para la selección adecuada de las características, pues de ello depende la precisión con que el modelo detecte el phishing. Sin embargo, además de ello, el mismo proceso de extracción demanda mucho tiempo. Lo que proponen los autores es un método de detección de páginas web de tipo Phishing utilizando una técnica rápida de selección de características multidimensional. De las pruebas realizadas, se tuvo como resultado que el nivel de detección fue de un 98,99%, con una tasa de falsos positivos de 0,59%. Por lo que la investigación concluye que con el método de clasificación de aprendizaje profundo (Deep Learning) utilizando la selección rápida y automática de características, se logra mejorar no solo la rapidez con la que detecta, sino también, la precisión en la identificación de tasa de Falsos positivos.

Diwakar Tripatji, Bhawana Nigam, Damodar Reddy Edla, (2017), realizaron la investigación, *A novel Web Fraud Detection Technique using Association Rule Mining* en NIT Goa de la India. Los autores han identificado que el phishing aprovecha que muchos sitios web requieren de publicidad para sostenerse, y a ella se filtra publicidad engañosa, que en realidad es phishing, abriendo una y otra página a la que la víctima no desea y que finalmente es engañado. Lo que plantea el artículo, es utilizar un algoritmo de minería de reglas de asociación, que analiza el registro web del usuario, a fin de detectar una secuencia anómala, pre procesa la data extraída, la transforma e identifica el patrón, para así, determinar si es legítima, publicidad o suplantación de identidad de la página real. Los resultados muestran que el algoritmo utilizado, llamado A priori, permite detectar mediante el acceso a los registros web, la legitimidad de la página, logrando identificarla con un nivel de precisión de 89.4%, un tiempo de 2.4 segundos, consumiendo 0,35kb de memoria. El novedoso método puede dejar de ser eficiente si la base de datos de los patrones aumenta, pues aumenta el tiempo de búsqueda, y que podría mejorarse usando algún otro algoritmo de búsqueda por indexación o etiquetado.

Gunikhan Sonowal, K.S. Kuppusamy, (2017), realizaron la investigación, *PhiDMA – A Phishing detection model with multi-filter approach*, en la Pondicherry University de la India. El artículo manifiesta que hay dificultades para identificar las diferentes categorías de Phishing dada la cada vez más sofisticada técnica de suplantación de identidad. Para lo cual, se propone un modelo multicapa llamado Phishing Detection using Multi-filter Approach (PhiDMA), único modelo que propone una interfaz accesible para personas con discapacidad visual, y que considera 5 capas: características de URL, firma léxica, coincidencia de cadenas, y de comparación de accesibilidad. Como resultado se tuvo que el modelo es capaz de detectar los sitios web de tipo phishing con una precisión de 92.72%. Como conclusión, el artículo reconoce que, por la diversificación de las técnicas de phishing, una única solución es compleja y por tanto un solo modelo de detección no sería óptimo, por ello en este modelo cada capa funciona como un filtro para ir perfilando la legitimidad de la web visitada, además que es especialmente para personas con discapacidad visual.

1.3 Teorías relacionadas al tema

1.3.1 Definición de Phishing

El Phishing es una forma de ataque que utiliza la ingeniería social para poder cometer delitos informáticos a usuarios o empresas que realizan sobre todo operaciones financieras electrónicas. En Internet podemos encontrar infinidad de conceptos sobre phishing, dado que es una técnica de estafa muy usada por los atacantes (Phishers), siendo el medio más común de ataque el correo electrónico. Para esto, hacen llegar información engañosa a los usuarios solicitando, por lo general, ingresen sus credenciales o números de tarjetas a portales web fraudulentos. (Association of Certified Fraud Examiners, Inc., 2017).

Según ESET, la definición del término fue escuchado por primera vez en una conferencia de seguridad hecha por Jerry Felix y Chris Hauck, quienes describían el término como: “la técnica de un atacante que imitaba una entidad o servicio con buena reputación”. (ESET, 2020). La Palabra en sí es un homófono de “pescar” víctimas. La “ph” del principio hace referencia a “phreaks”, que era un grupo de hackers de los años 90. (ESET, 2020).

1.3.2 Tipos de Phishing

a. Phishing tradicional

De todas las modalidades de phishing, a la hora de analizarlo es técnicamente el más elemental; por lo general, es una copia de un sitio web con buena reputación, mayormente conocido por las personas o empresas, en donde se solicitan los datos personales y/o sensibles. Luego de realizada la captura de información, los Phishers se apropian de las credenciales y/o datos sensibles ingresados por las víctimas para cometer actos delictivos como extorsión, robo de dinero o suplantación de identidad. (APWG, 2020)

b. Spear phishing

Es una táctica usada por los Phishers a un grupo destino específico de una empresa con la finalidad de obtener información sensible y/o datos

personales. La diferencia está, en que los Phishers han hecho un estudio o seguimiento, para saber cómo abordar a su objetivo y puedan obtener información sensible. En los últimos tiempos, se han creado grupos de usuarios en distintas redes sociales como Facebook, WhatsApp, LinkedIn o Twitter, con la finalidad de recolectar información como afinidades y gustos, que luego serán aprovechadas para engañar a la víctima. (ESET, 2020).

c. Smishing

El Smishing es otra forma de fraude electrónico que tiene como objetivo principal abordar los dispositivos móviles, a través de mensajes SMS, donde generalmente envían un link que los deriva a una página web fraudulenta. (AVG Technologies, 2020).

d. Vishing

También conocido como Phishing de voz; técnica donde se utiliza la llamada telefónica para hacerse pasar por una entidad legítima y engañar a la víctima. En muchas ocasiones se configuran hasta centrales telefónicas desde donde realizan y reciben las llamadas de las víctimas. (NortonLifeLock, 2020)

e. CatPhishing

Es otra forma de ciberdelincuencia en la que se crea a una persona en línea quien de manera ficticia atrae a personas a relaciones amicales o emocionales. El Phisher por lo general creará varias cuentas falsas (fakes) en diferentes webs con fotografías de personas que no reales. El objetivo del Phisher podría ser la extorsión u obtener información confidencial. (Bernardo, 2019).

f. Rock phishing

Los rock Phishers controlan varias computadoras utilizando Bots, para enviar correos electrónicos masivos de phishing a miles de usuarios; estos generalmente contienen un mensaje de una entidad financiera que invita a las víctimas a entrar al enlace fraudulento. (Association of Certified Fraud Examiners, Inc., 2017).

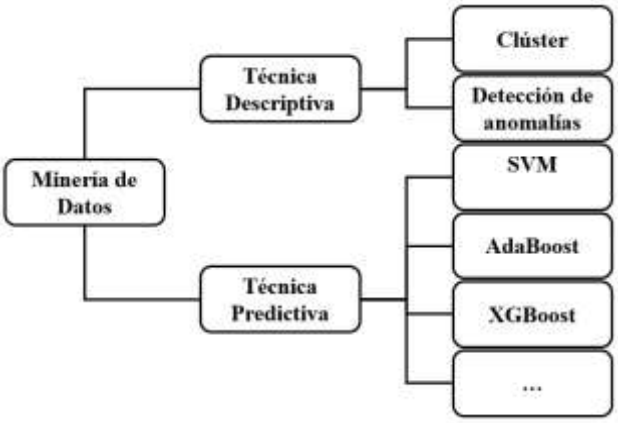
g. Pharming

Pharming es un tipo de suplantación de identidad, en la capa de red. Explota una vulnerabilidad a nivel de servidor de nombres de dominio (DNS), permitiendo al Phisher redireccione el tráfico de un sitio web legítimo a uno falso, utilizando la técnica de hombre en el medio. Por lo que, aunque el usuario ingrese la dirección correcta del sitio web, el software malicioso lo redirecciona al sitio web fraudulento, engañando al usuario que finalmente, confiado, ingresará sus credenciales y/o información sensible, que serán capturados por la página falsa y enviado a un servidor de datos o enviado por correo al atacante, quien finalmente lo utilizará para un acto ilícito. (Panda Security, S.L, 2020).

1.3.3 Minería datos

La minería de datos es un proceso del análisis, ligado a la estadística e informática, su finalidad es extraer grandes volúmenes de datos para procesarlos, analizarlos, y mostrar información relevante del resultado para la toma de decisiones. Parte del análisis, es encontrar patrones que puedan ayudar a dar una idea de lo que está ocurriendo para poder orientar a la solución. Siendo así, en el proceso de análisis se utilizan muchas técnicas de aprendizaje automático, donde se aplican diversos algoritmos, que en la actualidad vienen siendo mejorados por la comunidad científica, así como también, el uso de herramientas y científicos especialistas en interpretación de datos complejos. (ESET, 2020).

Asimismo, existen técnicas de minería de datos que nos ayudan a poder orientar la investigación según el propósito, ya que no todos los métodos están orientados a lo mismo, pues existen técnicas predictivas y descriptivas, y se deberá usar según el propósito del análisis. A continuación, la siguiente imagen podrá gráficamente ver su clasificación.



A. Técnica descriptiva

El Clustering, o también conocida como técnica de agrupamiento, tiene como finalidad que se construya una tabla de datos, en donde se los agrupa por características similares, así también la diferencia entre los grupos, mientras más distintos sean los grupos se dará una mejor clasificación. (EcuRed, 2020).

Así también, la técnica de anomalías descriptiva está enfocada en el descubrimiento y análisis de datos con comportamientos o patrones inusuales. Los cuales analizados con otra técnica no podrían tener resultados válidos. Esta técnica se basa en el agrupamiento, y tiene 2 formas de abordarla: cuando se toma como índice la distancia desigual de los datos, o cuando el índice está basado en la densidad de cada grupo. (Torres-Domínguez, Omar et al, 2018).

B. Técnica Predictiva

Es una técnica que permite representar gráficamente un flujo de decisiones que están relacionadas, y que puede ir incrementando, como las ramas de un árbol, hasta llegar a un resultado final. Todos los mapas de árboles de decisión, tiene 3 tipos de nodos: de probabilidad representado gráficamente por un círculo, de decisión representado por un rectángulo y los nodos terminales, que serían los de decisión, representados por un triángulo. Tal como se muestra en la siguiente imagen. (Lucid Software Inc., 2020).

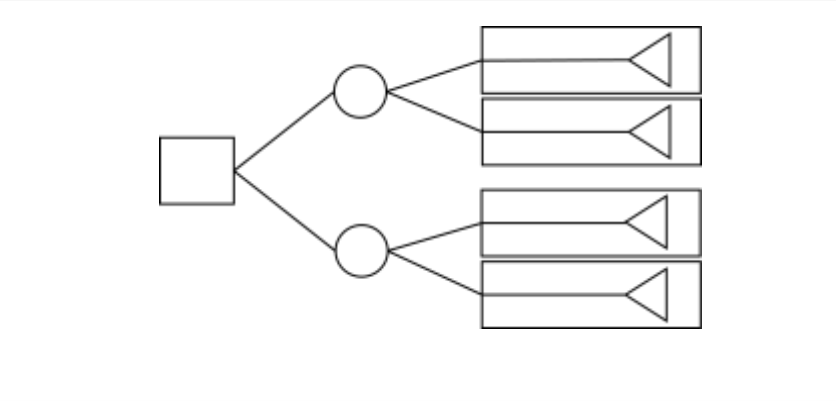


Figura 4. Estructura de Árbol de decisión. Elaboración propia.

Otra técnica conocida es la de Redes Bayesianas, también conocido como red de creencias, porque se grafica el conocimiento de un análisis de una situación incierta. Al igual que en los árboles de decisión, este modelo utiliza nodos para representar las variables del análisis, sin embargo, se tiene una variante, con respecto al árbol de decisión, que los nodos de borde son especialmente para las dependencias probabilísticas. Es conocido también como el método combinado, pues es una mezcla de teorías de probabilidad, de informática, estadística y de grafos. (IBM Knowledge Center, 2020)

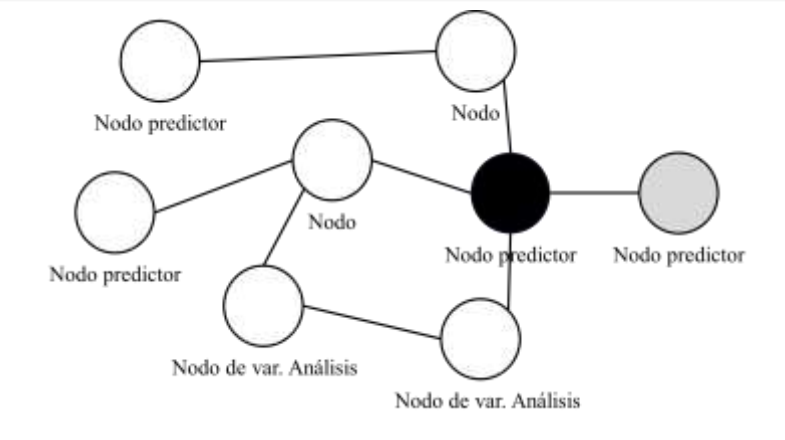


Figura 5. Diagrama de Redes Bayesianas. Elaboración propia.

Del mismo modo, se tiene la técnica Redes Neuronales, que se basa en el diseño de las redes neuronales consiste en la interconexión de nodos, basados en el modelo neuronal biológico, y éstos a su vez, están organizados en capas secuenciales que están interconectadas entre sí. Cada nodo evalúa los valores de entrada, para compararlo con el mecanismo de filtrado, y enseguida se determina su propio valor de salida.

Este método puede ser entrenado de manera supervisada y no supervisada, y su aplicabilidad, depende del objetivo, en el primer caso, usualmente se utiliza para la predicción y clasificación basado en una fuente histórica que es alimentada continua y automáticamente. Por el otro lado, el no supervisado, está más orientado en la descripción de datos. El punto bajo de esta técnica es el tiempo prolongado de aprendizaje, pero tiene un buen nivel de precisión, por lo que es frecuentemente utilizado para identificar los patrones de comportamiento en escenarios como el fraude. (Universidad Nacional del Noreste de Argentina, 2020).

Otra técnica muy conocida es llamada Support Vector Machine (SVM), ya que es una de las herramientas estándar respecto a la minería de datos. Pero tiene una característica particular, y es que el modelo permite la clasificación en 2 categorías, todo el conjunto de características que tiene un caso, se le llama vector, y este separa los casos en 2 categorías, uno a cada lado de la variable objetivo, de manera tal, que en la realidad el modelo pueda clasificar en, por ejemplo: fraude o no fraude, Es o No es, Si o No. Este modelo es normalmente utilizado para la categorización de texto, reconocimiento de caracteres, clasificación de imágenes, entre otras. (DTREG - Software For Predictive Modeling and Forecasting, 2020).

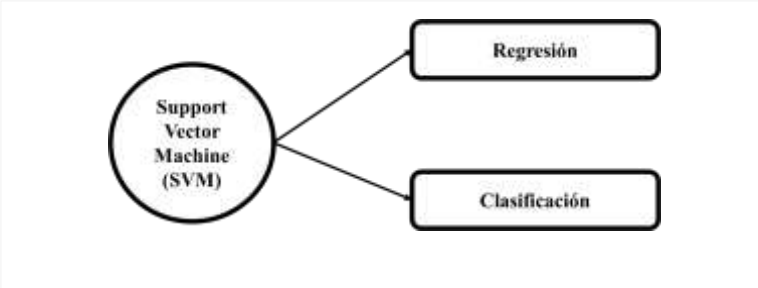


Figura 6. Clasificación de método de algoritmo SVM. Elaboración propia.

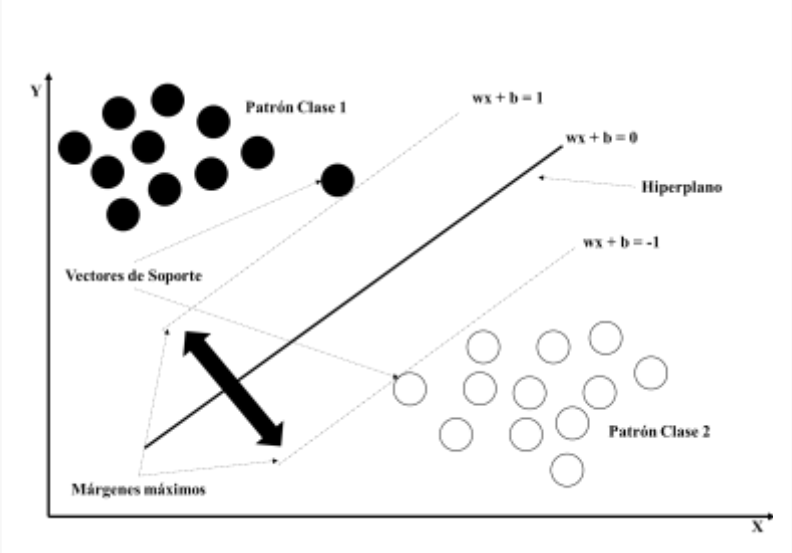


Figura 7. Gráfica SVM con el hiperplano para separar las clases.

Cada hiperplano tiene su respectiva fórmula o definición matemática que servirá para medir el margen con los otros hiperplanos. Mientras más amplio sea la separación de los hiperplanos, significará que se tendrá una mejor clasificación. Siendo así, las ecuaciones matemáticas para definir la distancia entre los hiperplanos serán:

Hiperplano Positivo: $wx + b = 1$

Hiperplano Medio: $wx + b = 0$

Hiperplano Negativo: $wx + b = -1$

Estás fórmulas serán fundamentales aplicarlas en la investigación, ya que de ellas dependerá que nuestra clasificación sea la más acertada posible. Habiendo ya entrenado al algoritmo, es decir, que el algoritmo

ya identificó los patrones en grupos de datos y los etiqueta para su correcta identificación, se pasa al algoritmo data nueva de prueba, este a su vez, genera un nuevo hiperplano, que como ya se ha mencionado, no es más que una línea para dividir en 2 el espacio, quedando cada clase en su respectivo lado, pudiendo así lograr la clasificación.

Teniendo los datos ya explotados, en donde aplicamos el método SVM con algoritmo de clasificación, se tiene que someter los resultados a mediciones científicas, para asegurar que estos cumplen eficientemente con la detección de páginas web de tipo Phishing.

Así también se tienen técnicas combinadas, que buscan lograr un mejor rendimiento a las técnicas clásicas simples ya conocidas. Una de las técnicas combinadas más conocidas es AdaBoost, que proviene de Adaptative Boosting, su finalidad es obtener un aprendizaje fortalecido partiendo de clasificadores simples, para lograr una clasificación binaria. Esta se compone de una cantidad definida por el investigador de clasificadores simples que se ejecutan en N iteraciones secuenciales, para finalmente ponderar sus salidas. Es muy parecido a Random Forest, con la diferencia que la salida de cada iteración tiene una dependencia directa con la anterior.

AdaBoost, parte inicialmente de manera similar al árbol de decisión, pero de un solo nodo, es decir tiene un atributo con 2 posibilidades y una tasa de error, estás a su vez no están enlazadas con otras 2 posibilidades, sino que se mantienen separadas, formándose un bosque de nodos simples separados.

El beneficio de esta técnica combinada, es que permite un mejor rendimiento, ya que el procesamiento de los resultados se da de manera secuencial y según el peso de la tasa de error. Es decir, dado que los nodos son independientes, se puede ordenar y priorizar los nodos simples según la cantidad de errores, mientras más

clasificadores, menor la tasa de error. (James, Witten, Hastie, & Tibshirani, 2017)

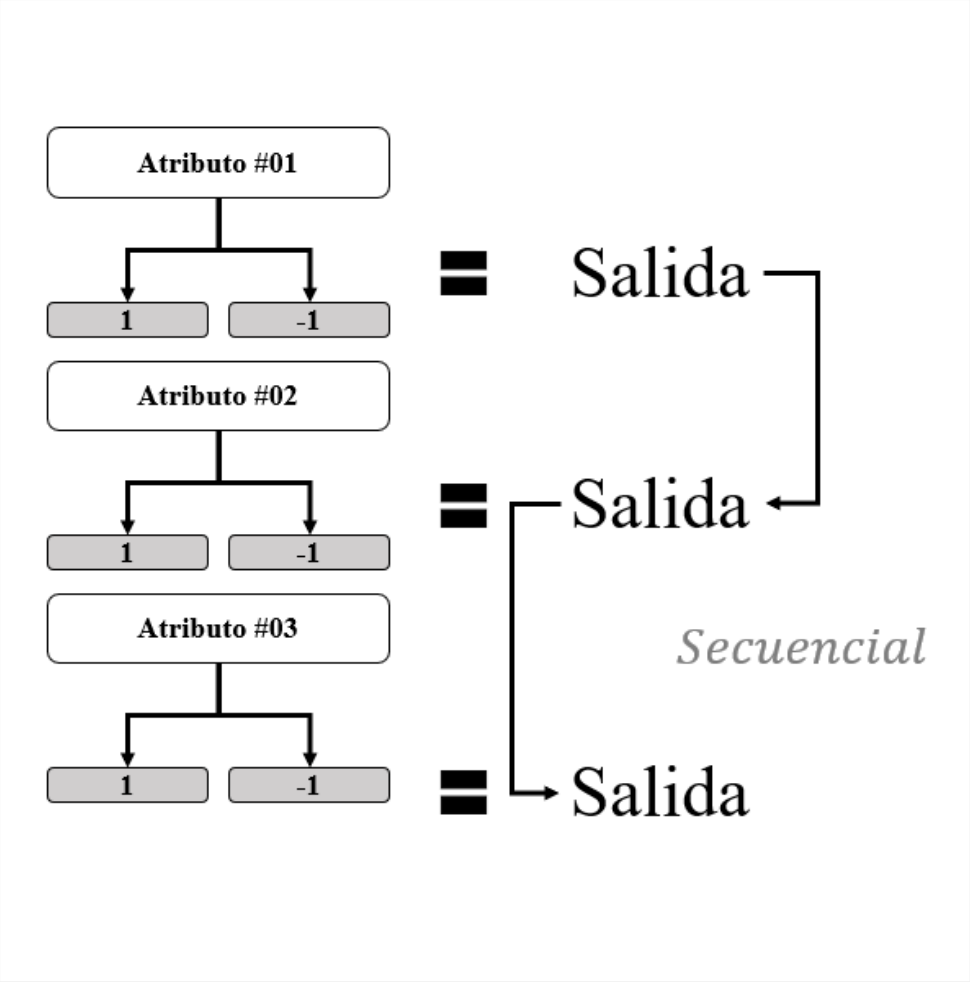


Figura 8. Ejecución de AdaBoost. *(James, Witten, Hastie, & Tibshirani, 2017)*.

Habiéndose explicado gráficamente cómo funciona la técnica de AdaBoost, ahora se explicará cómo funciona el algoritmo como tal.

En primer lugar, como se mencionó líneas arriba, AdaBoost le da un peso al conjunto de entrenamiento, y todos los grupos o instancias en que es dividido el dataset, inicialmente tienen el mismo peso, es decir 1/m, donde m es el número de instancias del entrenamiento.

Con los pesos asignados, se empieza con el primer set de entrenamiento y se seguirá de manera secuencial, obteniendo el primer resultado y la cantidad de errores que tuvo el clasificador. Siendo así, la fórmula es la siguiente:

$$For\ m = 1\ hasta\ M$$

Se extraen las instancias k_m

$$\alpha_m = \sum_{y_i \neq k_m(x_i)} w_i^{(m)}$$

Se define el peso

$$\alpha_m = \frac{1}{2} ln\left(\frac{1 - e_m}{e_m}\right)$$

donde $e_m = w_e / w$

Seguidamente, se actualizan los pesos, para la siguiente iteración. Siempre que exista una siguiente, la fórmula es:

$$w_i^{(m+1)} = w_i^{(m)} e^{-\alpha_m} = w_i^{(m)} \sqrt{\frac{e_m}{1 - e_m}}$$

De lo contrario, la fórmula es:

$$w_i^{(m+1)} = w_i^{(m)} e^{\alpha_m} = w_i^{(m)} \sqrt{\frac{1 - e_m}{e_m}}$$

Así también, se tiene la técnica de XGBoost, que viene de extra gradiente boosting, mismo que funciona igual que el AdaBoost, con la diferencia de que no es secuencial, es decir, no tiene que acabar modelo para continuar con el siguiente, sino que puede hacerlo en paralelo. (James, Witten, Hastie, & Tibshirani, 2017)

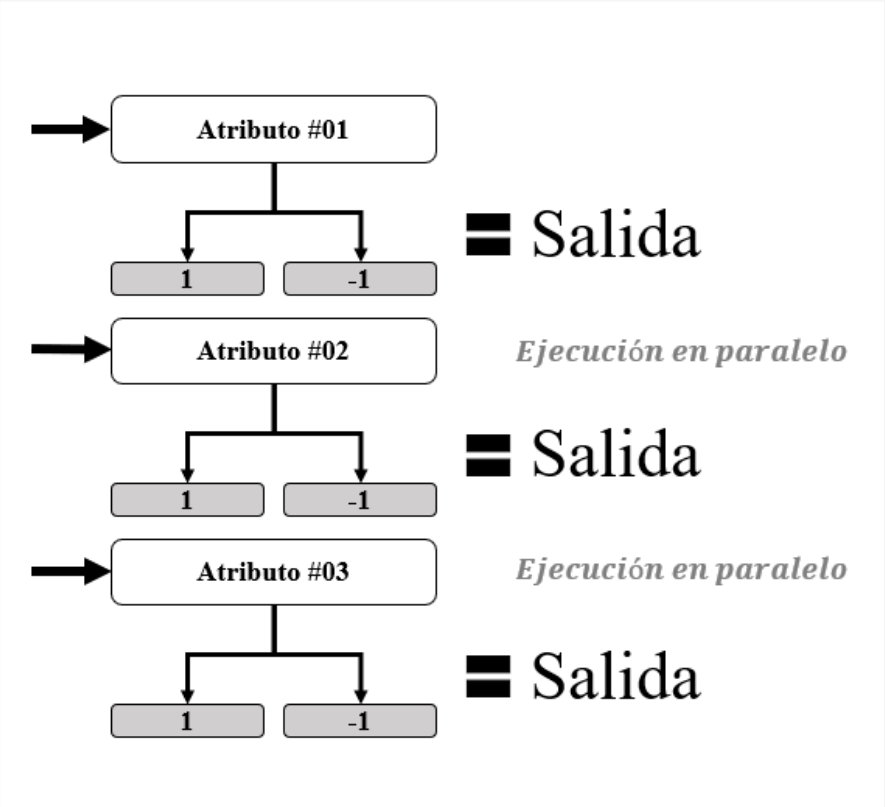


Figura 9. Cálculo con XGBoost. (James, Witten, Hastie, & Tibshirani, 2017).

Por consiguiente, el algoritmo realiza el cálculo de la siguiente manera:

$$£^{(\tau)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(\tau-1)} + \int t(x_i)\right) + \Omega\left(\int \tau\right)$$

Y la clasificación binaria optimizada para reducir pérdida de registro

$$yln(p) + (1-y)ln(1-p)$$

Donde $p = \frac{1}{(1+e^{-x})}$

1.3.4 Knowledge Discovery in Database (KDD)

En el mundo de la explotación de los datos, el proceso de minado, es una etapa del descubrimiento del conocimiento. Explotar la data,

significa descubrir información valiosa a través de los resultados de análisis, pero ese paso es solo para extraer conocimiento técnico. Es por ello, que es necesario complementarlo con una metodología muy conocida en el mundo del Big Data, pues facilita la automatización del procesamiento datos para que el enfoque esté más centrado en el análisis. (© 2018 Minerva Data Mining, 2020).

Como toda metodología, ésta tiene etapas, pero que son consecutivas e iterativas, pues permite regresar a una etapa anterior para redefinir cualquier variable, así como también interactiva, pues es necesario que el especialista participe en todas sus fases.

Figura 10. Etapas de la metodología KDD.

1.3.5 Selección de características

En este paso se determinan los atributos que caracterizan a las páginas web de tipo Phishing de las legítimas. Existen 3 métodos: de filtro, de envoltura y los métodos integrados. Sin embargo, es preciso indicar que el grupo de los métodos integrados es una combinación de los dos anteriores.

Tabla 2. Clasificación de método de selección de características. Elaboración propia.

Filtro	Envoltura	Integrados
Correlación de Pearson	Selección hacia adelante	Regresión LASSO
Análisis discriminante LDA	Selección hacia atrás	Regresión RIDGE
Análisis de varianza ANOVA	Eliminación de características recursivas	
Chi-cuadrado		

1.4 Formulación del Problema

¿Cómo detectar páginas web de tipo Phishing?

1.5 Justificación e importancia del estudio

En el ámbito de la seguridad informática, los gobiernos y entidades privadas ha mejorado los esfuerzos e inversión para fortalecer sus estrategias en este campo. Sin embargo, en la actualidad, donde el confinamiento a nivel mundial es un factor clave para superar la emergencia sanitaria a nivel mundial, la suplantación de identidad o Phishing ha tomado especial relevancia para los Phishers, pues las personas al no poder salir, realizan las actividades financieras y/o todo tipo de transacciones a través de internet, creando oportunidades de ataque para los ciberdelincuentes. Siendo así, el escenario se hace aún mejor para los Phishers, pues si le agregamos que muchos sistemas no son lo suficientemente seguros, y que la educación de las personas, en temas de seguridad informática y de la información, es bastante bajo; incrementa sus posibilidades de perpetrar su delito.

Es importante considerar que, ante un ataque masivo, no solo se ve afectado el usuario, sino también las entidades públicas y privadas, pues ellas verán

disminuir sus clientes, por el impacto reputacional que ocasiona. Por ejemplo, si el 60% de los clientes de un banco han tenido problemas de suplantación de identidad en un periodo de tiempo muy corto, y esto es difundido públicamente, no solo perderá a los clientes ya ganados, sino también a los futuros, pues preferirán apostar por un banco que garantice, en cierta medida, mayor seguridad en sus plataformas.

Por otro lado, en minería de datos, la extracción de patrones y las relaciones que se logran identificar entre ellas son evaluadas y analizadas bajo parámetros específicos, para luego convertir la data extraída en información que genere conocimiento de valor para los interesados, puesto que finalmente servirá para que el negocio pueda tomar decisiones certeras. Es por ello que, para la seguridad informática, ha tomado bastante relevancia el uso de los algoritmos de minería de datos, principalmente para identificar patrones de características típicas y de comportamiento anormal de los miles de páginas fraudulentas, para que con el uso de herramientas de apoyo complementarias se puedan clasificar como fraudulentas o legítimas. La innovación en el campo es muy buena, porque no solo hay interés de los profesionales en mejorar o innovar los métodos de detección de Phishing, sino también que hay apoyo de las entidades privadas y del estado.

Por tanto, para la aplicación de minería de datos es importante que se defina los criterios mínimos de selección de características exclusivamente para phishing y que se adopte un método para realizar una correcta clasificación.

1.6 Hipótesis

Mediante la implementación de un método de clasificación de minería de datos se detectará páginas web de tipo Phishing.

1.7 Objetivos

1.7.1 Objetivo General

Implementar un método de clasificación de minería de datos para detectar páginas web de tipo Phishing.

1.7.2 Objetivos Específicos

a) Analizar los métodos de clasificación de minería de datos usadas para detectar páginas web de tipo phishing con los mejores resultados de precisión.

b) Elegir el método de selección de características para páginas web de tipo Phishing.

c) Implementar un método basado en técnicas de clasificación de minería de datos que permita detectar páginas web de tipo phishing.

d) Evaluar el método basado en la técnica de clasificación de minería de datos para detectar páginas web de tipo Phishing.

II MATERIAL Y MÉTODO

2.1 Tipo y diseño de investigación

Será cuantitativa, pues a través de las técnicas ya descritas y el uso de herramientas computacionales, se tendrán los resultados que nos permitirán dar una conclusión. Además, el diseño será cuasiexperimental, pues el dataset que se utilizará de muestra tiene una recopilación de características de las páginas web de tipo phishing y que el método propuesto detectará como legitima o fraudulenta, el cual es el propósito del trabajo.

2.2 Población y muestra

A través de un estudio teórico del presente trabajo, la población está conformada por todos los métodos, de tipo clasificación, que usen minería de datos, para la detección de páginas web de tipo phishing, encontradas en la

revisión de artículos científicos, publicadas desde el año 2016 al 2020. Los mismos que se listan a continuación.

Tabla 3. Lista de Métodos de clasificación de minería de datos. Elaboración propia.

Item	Método de clasificación
1	Método utilizando inteligencia artificial combinado con varios algoritmos de clasificación
2	Método utilizando Adaboost y Multiboosting
3	Método de detección por características de la página web
4	Método predictivo para la detección de páginas phishing en día cero.
5	Método matemático de comparación con 2 atributos
6	Método OFS-NN para mejorar la selección de características
7	Método de optimización usando SWARM por valorización de atributos.
8	Método basado en selección de características usando DL
9	Método basado en asociación de reglas
10	Método PhiDMA basado en múltiples filtros

Para la muestra, se tomó a 3 mejores métodos de clasificación de minería de datos para la detección de páginas web de tipo phishing de la población con base a las siguientes características:

a) Tengan una precisión mayor o igual a 96%, que es el promedio de precisión toda la población.
b) Que el método no requiera que la marca de la entidad en la URL para que sea efectiva la detección del phishing.

2.3 Variables, Operacionalización

Tabla 4. Fórmulas para los indicadores.

Variables	Dimensión	Indicador	Fórmula	Técnica o Instrumento de recolección de datos	Leyenda
Independiente	Consumo de recursos	Consumo de CPU	$\mathrm{Cc} = \sum_{j}^{n} cc_j/n$	Observación directa Registros electrónicos	Cc = Consumo de CPU j = Consumo resultado n = Número de iteración
		Consumo de RAM	$\mathrm{Cm} = \sum_{j}^{n} cm_j/n$		Cm = Consumo de RAM j = Consumo resultado n = Número de iteración
Dependiente	Rendimiento	Exactitud	$\frac{\mathrm{VN} + \mathrm{VP}}{\mathrm{VN} + \mathrm{FP} + \mathrm{FN} + \mathrm{VP}}$	Registros electrónicos	VN = Verdadero Negativo VP = Verdadero Positivo FN = Falso Negativo FP = Falso Positivo
		Precisión	$\frac{\mathrm{VP}}{\mathrm{FP} + \mathrm{VP}}$		
		Sensibilidad	$\frac{VP}{FN + VP}$		
		F1-score	$(2)\frac{P * R}{P + R}$		P = Resultado de Precisión R = Recall o exhaustividad

2.4 Técnicas e instrumentos de recolección de datos, validez y confiabilidad

Se utilizará la observación directa y registros electrónicos para la recolección de datos del consumo de CPU y memoria RAM. Para lo cual, se generará una ficha de registro de datos electrónicos. Por otro lado, también se medirá los rendimientos del método, puesto que, bajo condiciones controladas, y a fin de manipular las variables según las características del Phishing, se realizará la medición de exactitud, precisión y sensibilidad, a fin de identificar la técnica de detección de páginas web de tipo Phishing más eficiente con el método propuesto.

A continuación, se describirá la aplicación de las siguientes fórmulas para las variables independientes y dependientes, que nos ayudarán con la medición antes mencionada, para que la interpretación del resultado sea imparcial.

Detalle del procesamiento de datos de las variables independientes:

Respecto al consumo de CPU:

Al iniciar el proceso de minería de datos, se realizará la medición del consumo, utilizando la ficha de registro de datos mencionada en el anexo. Para ello, se registrará el porcentaje de uso, antes y durante la ejecución del proceso de minado.

Respecto al grado de consumo de Memoria RAM:

El cálculo se basa en obtener el consumo de Megabytes de memoria utilizada antes y durante la ejecución del proceso de minado, mismo que será registrado en la ficha electrónica mostrada en el anexo.

Así también, para medir el rendimiento del método, se utilizará la matriz de Confusión, en la cual se debe considerar el significado de cada uno de los siguientes valores:

VP: Total de casos que son Phishing y que el algoritmo acertó.
FP: Total de casos que son Phishing y que el algoritmo no acertó.
VN: Total de casos que no son Phishing y que el algoritmo acertó.

FN: Total de casos que no son Phishing y que el algoritmo no acertó.

Con esta aclaración, se describe las fórmulas de medición, variables dependientes, que se utilizarán para validar que los resultados sean válidos.

Exactitud:

Será la cantidad de casos verdaderos que el modelo dará como resultado. Está fórmula científica, permitirá determinar de manera imparcial, el porcentaje real, no sesgado, del resultado obtenido, a fin de que estos resultados sean bien valorados.

$$Exactitud = \frac{\mathrm{VN} + \mathrm{VP}}{\mathrm{VN} + \mathrm{FP} + \mathrm{FN} + \mathrm{VP}}$$

Precisión:

Con esta fórmula se dará a conocer el nivel de precisión que obtuvo el algoritmo de manera imparcial. Se deberá entender como la cantidad de casos que son Phishing y que el algoritmo acertó versus la cantidad de casos que son Phishing y que el algoritmo no acertó.

$$\frac{\mathrm{VP}}{\mathrm{FP} + \mathrm{VP}}$$

Sensibilidad

Se deberá entender como la cantidad de casos que son Phishing y que el algoritmo acertó versus la cantidad de casos que no son Phishing y que el algoritmo no acertó.

$$\frac{VP}{FN + VP}$$

2.5 Procedimiento de análisis de datos

Este procedimiento se hará utilizando las técnicas de estadísticas usando Python como lenguaje de programación durante todo el análisis de los datos, que va desde el entrenamiento, creación de patrones, entrenamiento, pruebas y análisis de resultados, en conjunto con los indicadores detallados en las variables de operacionalización 2.3.

2.6 Criterios éticos

Confidencialidad: En esta investigación se mantendrá el anonimato de las fuentes de información externas utilizados para la recopilación, preparación y análisis de los DataSets, así como como; los autores y participantes.

Derecho de autor: Esta investigación se citan a los autores de las fuentes citadas cumpliendo el criterio ético; con lo cual protegemos y respetamos los derechos de propiedad intelectual.

Búsqueda del bien: Esta investigación permitirá aportar un método de clasificación que permita detectar páginas web de tipo phishing, dado que es un punto que debe de ser tomado en cuenta dentro de la seguridad informática, ya que en la actualidad podemos ser testigos que la ciberdelincuencia; que se manifiesta a través de los Phishers, sigue siendo un problema para la sociedad.

2.7 Criterios de rigor científico

Fiabilidad: Este trabajo garantiza resultados consistentes, dado que utilizará una metodología de minería de datos con algoritmos de clasificación para obtener datos fiables.

Validez: En esta investigación se usarán indicadores reales que se especifican en la tabla de Operacionalización; estos ayudarán a poder medir las variables por lo que esta validez es de tipo lógico en vista que utilizan modelos matemáticos y estadísticos.

Consistencia: La investigación que presentamos se fundamenta con pruebas fehacientes y demostrables como son los artículos científicos que hemos plasmado en el punto Trabajos previos y en la matriz de revisión.

III RESULTADOS

3.1 Resultados en Tablas y Figuras

El presente trabajo, tiene como objetivo principal la identificación de páginas web de tipo Phishing, y para ello se han utilizado técnicas que permitieron clasificar binariamente entre legítimas y Phishing.

A continuación, se muestra la siguiente tabla, con los resultados de los indicadores obtenidos en las 10 iteraciones de prueba, con los clasificadores AdaBoost, SVM y XGBoost. De la tabla, se muestra que, en Precisión, tanto SVM (95%) y XGBOOST (98%) muestran valores estables, por lo que son las 2 técnicas más adecuadas.

Tabla 5. Resultado de las pruebas por clasificador según su indicador de rendimiento.

	ADABOOST				SVM				XGBOOST			
	Precisión	Exactitud	Sensibilidad	F1-Score	Precisión	Exactitud	Sensibilidad	F1-Score	Precisión	Exactitud	Sensibilidad	F1-Score
Iteración 1	*93.69*	93.49	93.13	94.37	95.25	95.12	94.13	95.70	98.64	97.20	96.63	97.49
Iteración 2	*93.70*	94.03	93.20	94.81	95.39	93.94	92.59	94.49	98.59	97.06	97.66	97.38
Iteración 3	*93.76*	93.85	93.41	94.59	95.51	94.08	93.20	94.62	98.53	97.24	97.44	97.55
Iteración 4	*93.60*	93.80	93.36	94.58	95.41	94.57	93.42	95.16	98.78	96.79	96.35	97.26
Iteración 5	*94.02*	93.53	92.90	94.27	95.23	94.84	93.86	95.38	98.53	97.29	97.20	97.59
Iteración 6	*93.87*	93.53	93.53	94.25	95.24	94.84	93.61	95.31	98.67	95.52	95.10	95.99
Iteración 7	*93.71*	94.21	93.78	94.84	95.23	95.30	94.53	95.82	98.53	96.65	95.82	96.93
Iteración 8	*93.55*	93.94	92.92	94.57	95.25	94.12	92.62	94.73	98.55	97.29	96.94	97.56
Iteración 9	*93.58*	94.03	93.49	94.57	95.36	95.30	94.47	95.83	98.56	97.33	97.56	97.67
Iteración 10	93.93	93.58	93.39	94.29	95.24	94.71	93.66	95.28	98.60	96.83	97.14	97.22

3.1.1 Métricas de desempeño del método propuesto

La siguiente figura muestra los resultados obtenidos de cada técnica en la fase de pruebas. En la práctica para cada técnica, la exactitud representa cuan cerca se estuvo del resultado real, la precisión, muestra el grado de coincidencia de las respuestas, la sensibilidad, el grado de aciertos cuando es Phishing, F1-score es el grado ponderado entre la precisión y la sensibilidad.

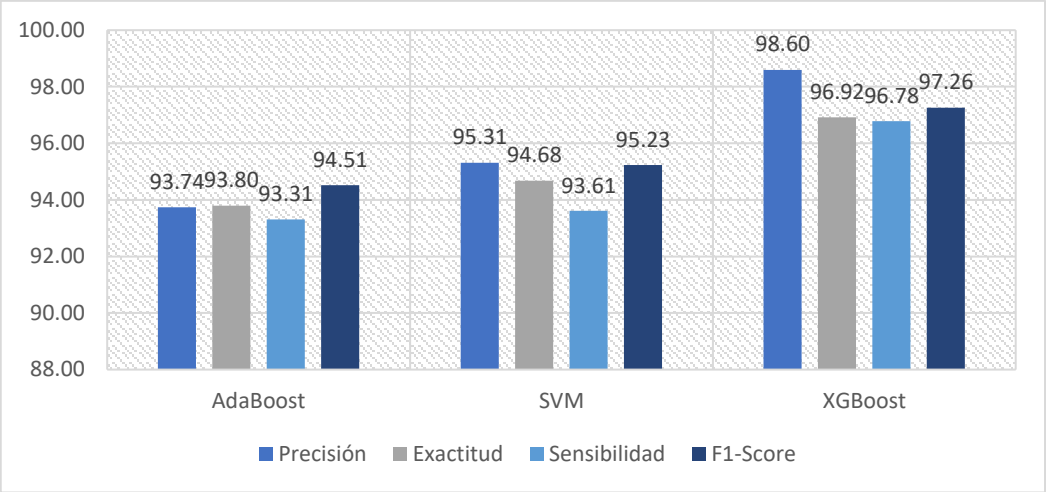


Figura 11. Ponderado de los indicadores por cada clasificador.

Asimismo, los resultados mostrados, evidencian que las 3 técnicas han tenido un alto desempeño, sin embargo, XGBoost ha sido sobresaliente de entre todas.

3.1.2 Métrica de desempeño del consumo de recursos

Los resultados mostrados, corroboran lo revisado en la literatura científica, donde las técnicas SVM y XGBoost tienen un mejor manejo de recursos de CPU logrando resultados eficientes.

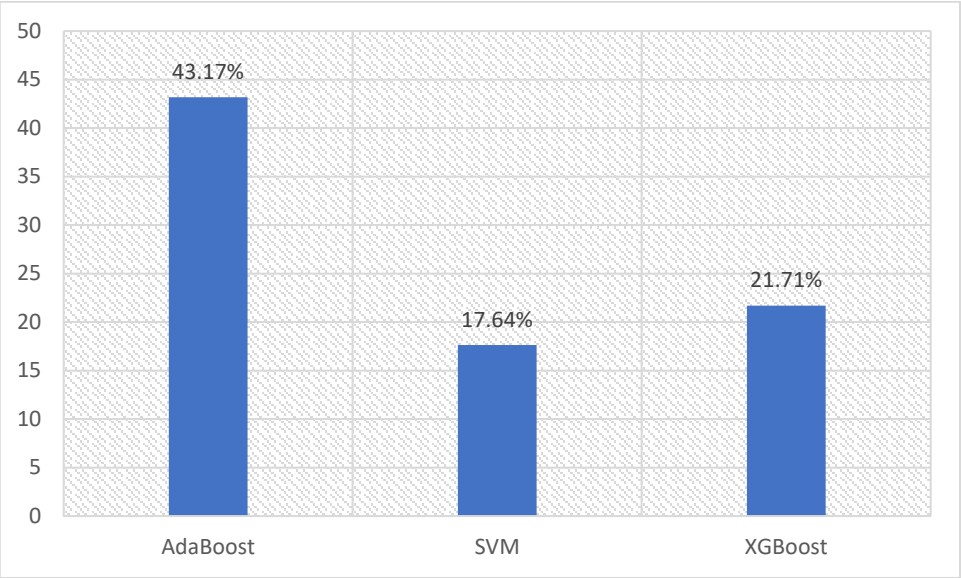


Figura 12. Grado de consumo de CPU durante la ejecución de las pruebas. Elaboración propia.

Asimismo, el mínimo consumo en Megabytes de XGBoost muestra la eficiencia de su técnica, dada la combinación de clasificadores y un destacable manejo de memoria con el kernel trick.

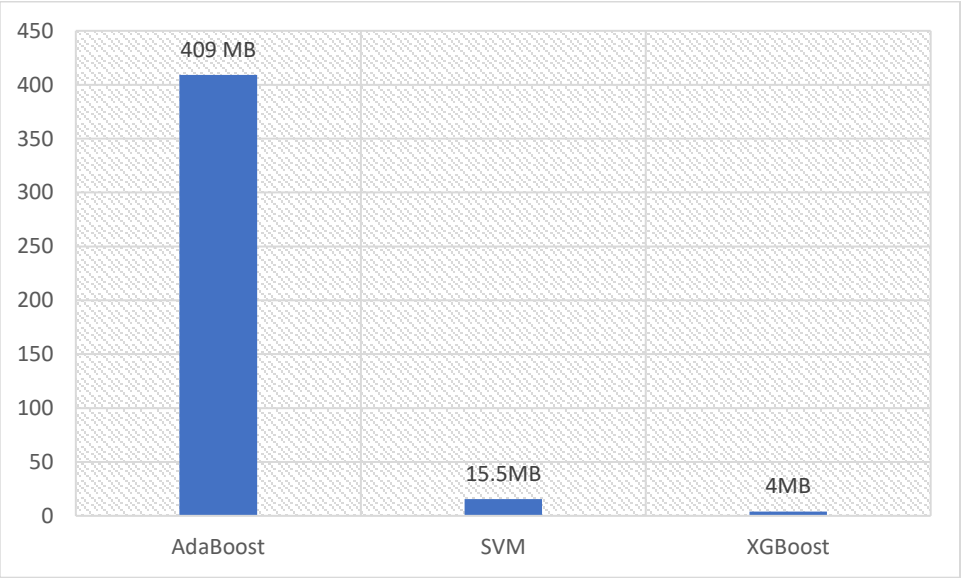


Figura 13. Consumo en Megabytes de RAM durante la ejecución de las pruebas. Elaboración propia.

3.2 Discusión de resultados

La siguiente tabla muestra los resultados de desempeño de los 3 mejores modelos con que se ha comparado el presente trabajo. El presente trabajo utiliza un dataset de 11054 registros vigentes, es decir, todos los registros son de páginas web detectadas durante el presente año a Julio 2020. De los resultados, se conoce que la investigación llamada *A predictive model for phishing detection,* que utiliza la técnica de SVM, es la que ha tenido un mejor desempeño, tanto en su precisión, exactitud y F-score. Sin embargo, es la que ha utilizado un Dataset con menor tamaño, con lo cual los resultados son mejores, dado que hay un menor esfuerzo de la técnica por la detección.

Por otro lado, se tiene la investigación llamada *Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning,* que utiliza una base de datos con paginas muy antiguas, mayor a los 5 años, y muchas de ellas ya no están en línea, es decir, no son vigentes. Así también, se tiene que todas las investigaciones, tienen al menos los 31 atributos y una etiqueta de clasificación que considera el presente trabajo para el dataset. Esto es para que, al realizar la comparación, se pueda realizar bajo las mismas condiciones.

Tabla 6. Análisis de resultados de otras técnicas similares. Elaboración propia.

Artículo	Técnica	Precisión	Exactitud	F1-Score	Dataset	Observación
Comparison of Adaboost with MultiBoosting for Phishing Website Detection	AdaBoost	96.42	97.61	97.6	11,054	Base de datos dentro de los 5 años
A predictive model for phishing detection	SVM	99.96	99.96	99.96	5,041	Base de datos dentro de los 5 años

Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning	XGBoost	98.41	98.57	99	2,010,779	Base de datos contiene páginas detectadas hace más de 5 años y que muchas ya no están activas

3.3 Aporte práctico

Se realizó la revisión de la literatura científica publicada en los últimos 5 años, donde se encontraron métodos para clasificar páginas web de tipo Phishing, con la mejor precisión como resultado, mismos que se encuentran en la siguiente tabla. Para la construcción se buscó en los repositorios IEEEXPLORE, IOP Science, ScienceDirect, con las cadenas de búsqueda phishing, data mining, algorithms, classification methods.

Tabla 7. Análisis de rendimiento de los métodos de clasificación. Elaboración propia.

Método de clasificación	Técnica	Precisión	Autor(es)
De IA con algoritmos de clasificación	Redes Neuronales	96%	(Rashid, Sallim, & Noor, 2020)
Adaboost y Multiboosting	Adaboost	97%	(Abdulhamit & Emir, 2019)Abdulhamit Subasi, Emir Kremic
De detección por características de página web	Árbol de decisión	95%	(Mao, y otros, 2017)
Detección predictiva de páginas phishing en día cero	SVM	99%	(Orunsolu, Sodiya, & Akinwale, 2019)
Matemático de comparación con 2 atributos	Similitud de coseno	100%	Opara Chidimma Ugochi
OFS-NN para mejorar la selección de características	Redes Neuronales	96%	(Chen, Ye, Li, & Liu, 2019)
De optimización usando SWARM por valor de atributos.	Redes neuronales	95%	(Ali & Malebary, 2020)
Basado en selección de características usando DL	XGBoost	98%	(Yang, Zhao, & Zeng, 2019)

Basado en asociación de reglas	Alg. Asociación	89%	(Tripatji, Nigam, & Edla, 2017)
PhiDMA basado en multiples filtros	PhiDMA	92%	(Sonowal & Kuppusamy, 2017)

Asimismo, la selección de los métodos se realizó siguiendo los criterios de selección que se muestran a continuación.

Tabla 8. Criterios de selección para la selección de los mejores métodos. Elaboración propia.

N°	Criterios	Descripción
1	Use un método de clasificación	Esto es debido a que es necesario determinar o clasificar si es o no una página phishing, y no utilice métodos de aproximación como los de regresión.
2	Precisión > 96%	Se selecciona como los mejores métodos de clasificación, lo que estén por encima del promedio. Siendo 96% el promedio de todos los métodos listados en la tabla 5.
3	Sin la marca en la URL	Toda URL debe ser analizada, aunque ésta no tenga la marca de la entidad atacada en la URL.

En tal sentido, se listan todos los métodos que cumplieron con todos los criterios de selección definidos.

Tabla 9. Métodos seleccionados con los criterios aplicados. Elaboración propia.

Item	Método de clasificación	Técnica	Precisión	Autor(es)
2	Adaboost y Multiboosting	Adaboost	97%	(Abdulhamit & Emir, 2019)Abdulhamit

				Subasi, Emir Kremic
4	Detección predictiva de páginas phishing en día cero	SVM	99%	(Orunsolu, Sodiya, & Akinwale, 2019)
8	Basado en selección de características usando DL	XGBoost	98%	(Yang, Zhao, & Zeng, 2019)

A continuación, se muestra gráficamente el proceso del método que propone esta investigación, para poder determinar si una página web es de tipo Phishing o legítima.

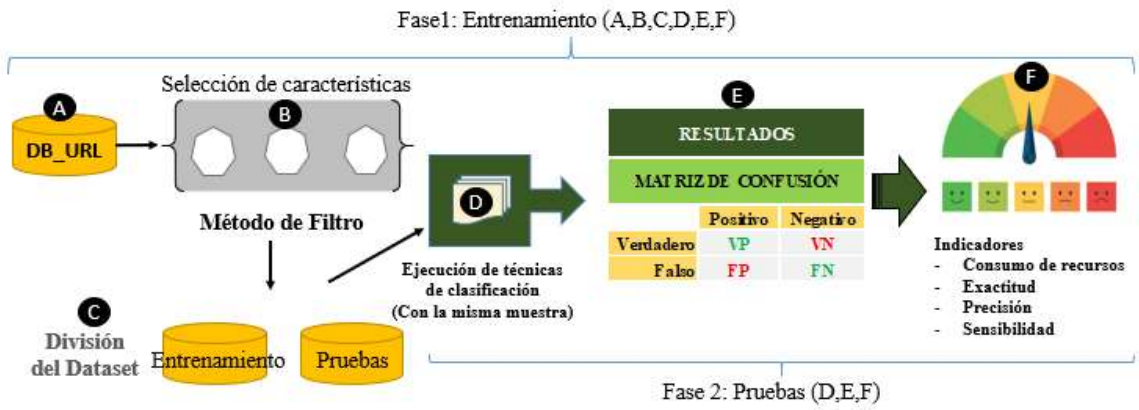


Figura 14. Método de selección de características.

A. Base de datos de URLs

En este proceso se seleccionó una base de datos con 11054 registros de páginas web, 56% legítimas y 44% Phishing. Dicha base de datos se mantiene actualizada mensualmente a nivel mundial por el reconocido sitio web anti-phishing PhishTank, por lo que los registros de las bases de datos tienen una antigüedad no mayor de 8 meses, es decir son del presente año.

La siguiente tabla, describe un extracto del dataset original que contiene 34 atributos y una etiqueta de clasificación. Los registros visibles en esta tabla son: 4 de tipo phishing y 4 legítimas.

Tabla 10. Muestra de un extracto de la base de datos original de 11054 registros.

Indice	url	Online	…	Google Index	Enlace_A_Pagina	Reporte_Estadisticas	Clasificacion
0	https://weddib.sites.com/@services.html	yes	…	1	1	1	-1
1	https://alexlaldinagency.com/rdt/redirect.html	yes	…	1	0	-1	-1
2	https://empresas-intercbank-pe.com/login/	yes	…	1	-1	1	-1
3	http://sigin-pemblokiran-fb20.gq/login.php	yes	…	1	1	1	-1
.	.	.	…	.	.	.	.
.	.	.	…	.	.	.	.
.	.	.	…	.	.	.	.
11050	https://bcpzonasegurabeta.viabcp.com/#/iniciar-sesion	yes	…	1	-1	1	1
11051	https://bancainternetempresas.scotiabank.com.pe/main/loginPage	yes	…	1	0	1	1
11052	https://zonasegura1.bn.com.pe/BNWeb/Inicio	yes	…	1	1	1	1
11053	https://mi.scotiabank.com.pe/login	yes	…	-1	1	-1	1

B. Método de selección de características:

Se eligió el método de selección de características llamado de Filtro, ya que está acorde con los objetivos planteados en este trabajo. La decisión está basada en un análisis teórico de la literatura, para lo que se elaboró la siguiente tabla, que permite identificar que el método de filtro consume menos recursos.

Tabla 11. Análisis teórico de las diferencias entre los métodos de selección de características. Elaboración propia basada en las definiciones de *(Masters, 2019)*.

Métodos de Filtro	Método de Envoltura
Es manual, no hay un proceso automatizado para la selección de características.	Se utiliza un algoritmo automatizado para selección basado en el auto aprendizaje.
Es rápida y evita el consumo de recursos, pero con mayor riesgo al error si no se tiene conocimiento de la data.	Consume mayores recursos, pero el riesgo es menor, siempre es necesario una revisión adicional, lo que se traduce en más tiempo.
Los subconjuntos creados podrían no ser las más idóneas si no se ha seleccionado las características relevantes.	Crea patrones con las características más relevantes, pues su selección es exhaustiva en base a pruebas del autoaprendizaje.

Con base en un análisis de la literatura científica, se identificó que existen 30 atributos similares y su respectiva etiqueta de clasificación, utilizados en los 3 mejores métodos de clasificación de páginas web de tipo Phishing, que son AdaBoost, SVM y XGBoost. Por lo que, a fin de hacer una evaluación bajo las mismas condiciones, de la base de datos original, se creó una nueva

base de datos, a la que llamamos "DataSet", con los 30 atributos identificados y una etiqueta de clasificación.

A continuación, se muestra la distribución del nuevo DataSet, donde se clasifica entre páginas legítimas y phishing.

```
print('Distribucion de la Base de datos:')
print(db_phishing.groupby('Clasificacion').size())
```

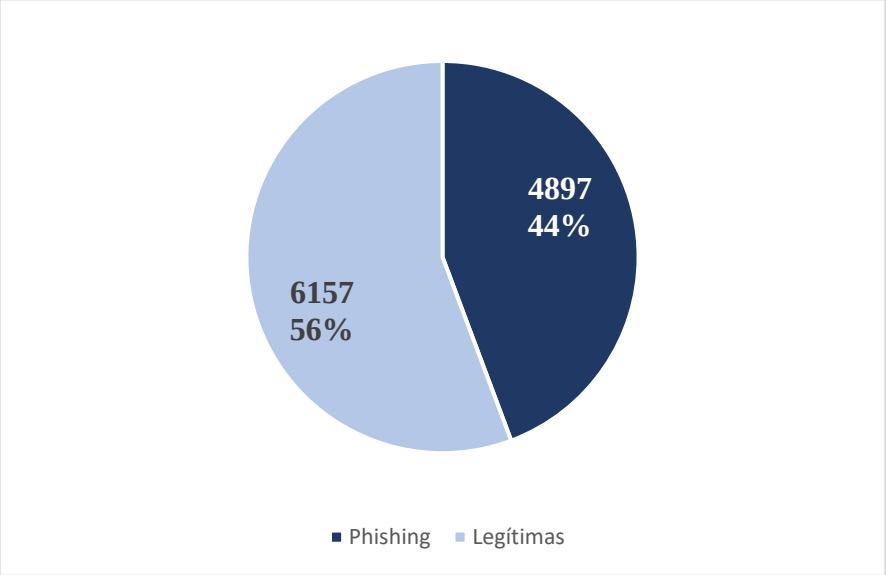


Figura 15. Cantidad registros por su clasificación. Elaboración propia.

A continuación, se describe cada uno de los atributos del dataset construido.

Tabla 12. Atributos considerados en el dataset. Elaboración propia.

Nº	ATRIBUTO	DESCRIPCIÓN	REGLA
1	**TieneIP:**	Algunas direcciones de internet maliciosas utilizan direcciones IP en vez de un nombre.	Si IP existe en URL = Phishing
2	**URL_Larga:**	Se identificó que una URL normalmente tiene una longitud máxima de 40 caracteres.	URL > 40 caracteres = Phishing URL < 40 caracteres = Legítima URL = 40 caracteres = Posible
3	**URL_Corta:**	**Por ejemplo:** https://www.uss.edu.pe/uss/TransparenciaDoc/Actas **Su URL acortada sería:** https://bit.ly/3lmPalD	8 > Cantidad de caracteres después del último “/” de URL corta > 4 = es legítima
4	**Simbolo@:**	El uso del símbolo "@" en la URL lleva al navegador a ignorar todo lo que precede al símbolo "@" y la dirección real a menudo está después del símbolo "@".	Si URL tiene @ = Phishing
5	**Redirecciona//:**	La existencia de "//" dentro de la URL significa que el usuario será redirigido a otro sitio web. Ejemplo: "http://www.legitimate.com//http://www.phishing.com".	Si URL tiene // desde 8ª posición = Phishing
6	**PrefijoSufijo:**	Rara vez las páginas legítimas usan “-“ dentro de su URL. Por ejemplo, http://www.Confirme-paypal.com/.	Si URL tiene “-“ = Phishing
7	**SubDomino:**	Este atributo es necesario para analizar si existe un subdominio.	Se omiten todos los caracteres y se deja solo el subdominio.

Tabla 13. Atributos considerados en el dataset. Elaboración propia.

Nº	ATRIBUTO	DESCRIPCIÓN	REGLA
8	**HTTPS:**	Los certificados dan la seguridad de que la página web que están visitando es segura. Los Phishers ponen certificados temporales a las páginas phishing.	Vigencia del Certificado < 6 meses = Phishing Entidad certificadora válida
9	**Registro_Dominio:**	Según la literatura científica, los dominios fiables tienen una duración no menor a 6 meses.	Registro de dominio < 7 meses = Phishing.
10	**Favicon**	Es una imagen en la URL para recordar al usuario la identidad de la página web.	Carga del Favicon ≠ interno = Phishing
11	**PuertoNoEstandar:**	Es una buena práctica que solo estén abiertos los puertos que sean necesarios.	Si otros puertos además del 80 y 443 están abiertos = Phishing
12	**HTTPSDominioURL:**	Los phishers pueden añadir la palabra "HTTPS" como parte de dominio para engañar a los usuarios. Por ejemplo, http://https-www-paypal-it-webapps-mpp-home.soft-hair.com/.	Si el dominio contiene la palabra https = Phishing
13	**RespuestaURL:**	Se examina si los objetos de una página web cargan de un dominio distinto	Objetos web cargan dominio externo = Phishing
14	**EtiquetaAnchorURL:**	Uso de etiquetas “<a> Etiqueta 	Si el domino es diferente a la etiqueta = Phishing
15	**EnlaceenScript:**	Uso de etiquetas para metadatos	Si no usa metadato = Phishing
16	**ServerFormManejo:**	Server Form Handler que contienen vacíos "about:blank" son dudosos porque se debe tomar una acción	Si existe = Phishing
17	**InfoEmail:**	Los formularios web permiten enviar su información que es dirigida a un servidor para su procesamiento.	Si domino de correo es distinto = Phishing

Tabla 14. Atributos considerados en el dataset. Elaboración propia.

Nº	ATRIBUTO	DESCRIPCIÓN	REGLA
18	**AnormalURL:**	La identidad suele ser parte de la URL	Si dominio es distinto a URL = Phishing
19	**WebReenvios:**	Cantidad en las que una página web ha sido reenviada.	Redirección >3 = Phishing
20	**Barra_Estado:**	URL falsa en la barra de estado	Si en evento “onMouseOver cambia = Phishing
21	**DeshabilitarBotonDerecho:**	Acción que desactiva el clic derecho	Si evento.button = 2 es Phishing
22	**UsandoPopupWindow:**	Envío de información en ventana emergente	Si existe ventana emergente con formulario = Phishing
23	**IframeRedireccion**	Etiqueta HTML para mostrar una página dentro de otra.	Si usa FrameBorder = Phishing
24	**EdadDominio:**	Registro de antigüedad en la base de datos de WHOIS.	La antigüedad del dominio < 6 meses = Phishing
25	**DNSRegistro:**	Registro de dominios	Si no existe o está vacío = Phishing
26	**Trafico_Sitio:**	Popularidad del sitio web	Si dominio no está en BD Alexa = Phishing
27	**RankingPagina:**	PageRank mide la importancia de una página de 0 a 1.	Si valor es < 0.2 = Phishing
28	**GoogleIndex:**	Usualmente, las páginas web legítimas se encuentran en la base de datos de Google Index.	Página web no está en GoogleIndex = Phishing
29	**Enlace_A_Pagina:**	Cantidad de referencias o citas que se han hecho al sitio web	Si no tiene referencias = Phishing
30	**Reporte_Estadisticas:**	Entidades como PhishTank y StopBadware publican informes sobre sitios web.	Si la url está en PhishTank o StopBadware = Phishing
31	**Clasificacion:**	Etiqueta que clasifica a los atributos	1 = Legítimo -1 = Phishing

C. División del Dataset

Para iniciar con el proceso mencionado, y dado que la base ya incluye la etiqueta de clasificación de cada página, separamos la etiqueta “Clasificacion”, para así separar los datos del resultado. Siendo así, se crearon 2 arreglos, uno llamado “X” con 30 atributos y 11054 registros, y el otro arreglo “Y” que tiene la etiqueta “Clasificación” con 11054 registros.

```
X = var_np.array(db_phishing.drop(['Clasificacion'], 1))
y = var_np.array(db_phishing['Clasificacion'])

print(X)
print(X.shape)
print(y)
print(y.shape)
```

Tabla 15. Resultados de la separación del dataset. Elaboración propia.

Arreglo	Nro Atributos	Nro. Registros	Observación
X	30	11054	Contiene todos los atributos sin resultados
Y	1	11054	Solo contiene la etiqueta Clasificación con los resultados

A continuación, se dividió el arreglo “X” en dos, para utilizar una parte de los datos en el entrenamiento y la otra en la fase de pruebas. Siguiendo la regla de Pareto:

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)
print(X_train)
print(X_train.shape)
print(X_test)
print(X_test.shape)
```

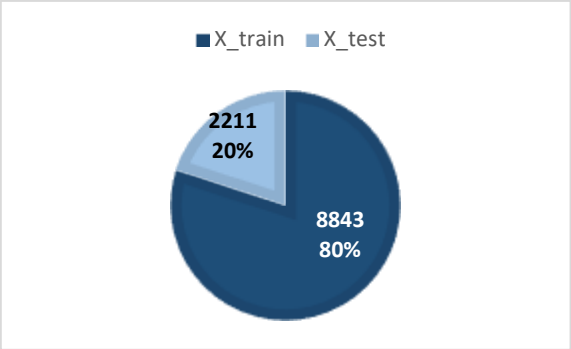


Figura 16. Arreglo de base X_train y X_test. Elaboración propia.

D. Minería de datos:

Las siguientes instrucciones, se utilizaron para hacer el cálculo de la precisión con las técnicas AdaBoost, SVM y XGBoost; y las bases de datos de entrenamiento, así como también la predicción final con las bases de datos de prueba.

Para ello, se creó una variable llamada "algoritmo", en la que se llama, según sea el caso, a la técnica AdaBoostClassifier() si el cálculo será para AdaBoost, SVC() si es para SVM, y finalmente XGBClassifier() si es para XGBoost.

```
#Técnica con AdaBoost
clasificador = AdaBoostClassifier()

#Técnica con SVM
clasificador = SVC()

#Técnica con XGBoost
clasificador = XGBClassifier()
```

Seguidamente, a la variable "clasificador" se le agregó la función "fit" para que ajuste los datos de las variables X_entrenamiento e y_entrenamiento. Como se indicó anteriormente, ambos contienen 8843 registros, el primero con todos los datos sin clasificar, y el segundo con solo las respuestas de la etiqueta de clasificación.

```
clasificador.fit(X_entrenamiento, y_entrenamiento)
```

A continuación, a la variable "clasificador" se le agregó la función "predict" para que realice la predicción con el arreglo X_prueba, y se guarda el resultado en una nueva variable llamada Y_prediccion.

```
Y_prediccion = clasificador.predict(X_test)
```

Finalmente, como resultado se muestra la precisión obtenida por cada técnica.

```
print('\n Precisión Adaboost Clasificación: {
0:.2f}'.format(clasificador.score(X_train, y_train)*100))
```

```
Precisión AdaBoost Clasificación: 93.88
```

```
print('\n Precisión con SVM: {
0:.2f}'.format(clasificador.score(X_entrenamiento,
y_entrenamiento)*100))
```

```
Precisión SVM: 95.59
```

```
print('\n Precisión XGBoost: {
0:.2f}'.format(clasificador.score(X_entrenamiento,
y_entrenamiento)*100))
```

```
Precisión XGBoost: 98.46
```

E. Matriz de confusión Permitió la visualización del desempeño del método empleado utilizando varios algoritmos durante la etapa del aprendizaje supervisado, a fin de identificar y comparar el desempeño según el algoritmo utilizado. Asimismo, en él se observa una matriz con los valores de predicción vs los valores reales, mismo que permitió ver la cantidad de errores que se tuvieron como parte del resultado.

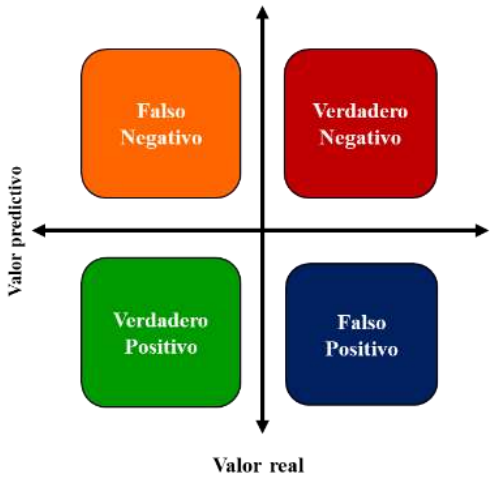


Figura 17. Matriz de confusión

Tabla 16. Interpretación de la Matriz de confusión. Elaboración propia.

Valor	Significado
Verdadero positivo:	Todos los que fueron predichos como Phishing y lo son.
Verdadero Negativo:	Todos los que fueron predichos como No Phishing, y no lo son.
Falso Positivo:	Todos los que fueron predichos como Phishing, pero no lo son.
Falso Negativos:	Todos los que fueron predichos como No Phishing, pero lo son.

```
def plot_confusion_matrix(y_prueba, prediccion_y):
 C = confusion_matrix(y_prueba, prediccion_y)
 A =(((C.T)/(C.sum(axis=1))).T)
 B =(C/C.sum(axis=0))
 plt.figure(figsize=(20,4))
 labels = ['V=1','F=-1']
 cmap=sns.light_palette("navy")
 plt.subplot(1, 3, 1)
 sns.heatmap(C, annot=True, cmap=cmap, fmt=".3f",
xticklabels=labels, yticklabels=labels)
 plt.xlabel('Valor Predecido')
 plt.ylabel('Valor Real')
 plt.title("Matriz de Confusión")
 plt.show()

#Técnica con AdaBoost
clasificador = AdaBoostClassifier()
clasificador.fit(X_train, y_train)
```

```
Y_prediccion = clasificador.predict(X_prueba)
```

Tabla 17. Matriz de Confusión - AdaBoost. Elaboración propia.

Predicción VS Realidad - AdaBoost		
Resultado	**Cantidad**	**Matriz de Confusión**
	892	VP
Fueron predichos como No Phishing, y no lo son	1182	VN
Fueron predichos como Phishing, pero no lo son	53	FP
Fueron predichos como No Phishing, pero lo son	84	FN
Total, de registros del Dataset de Prueba:	**2211 (20%)**	

```
#Técnica con SVM
clasificador = SVC()
clasificador.fit(X_entrenamiento, y_entrenamiento)
Y_prediccion = clasificador.predict(X_prueba)
```

Tabla 18 Matriz de Confusión - SVM. Elaboración propia.

Predicción VS Realidad - Máquina de Vectores de Soporte		
Resultado	**Cantidad**	**Matriz de Confusión**
Fueron predichos como Phishing y lo son	901	VP
Fueron predichos como No Phishing, y no lo son	1202	VN
Fueron predichos como Phishing, pero no lo son	33	FP
Fueron predichos como No Phishing, pero lo son	75	FN
Total, de registros del Dataset de Prueba:	**2211 (20%)**	

```
#Técnica con XGBoost
clasificador = XGBClassifier()
clasificador.fit(X_entrenamiento, y_entrenamiento)
Y_prediccion = clasificador.predict(X_prueba)
```

Tabla 19 Matriz de Confusión - XGBoost. Elaboración propia.

Predicción VS Realidad - XGBoost

Resultado	Cantidad	Matriz de Confusión
Fueron predichos como Phishing y lo son	931	VP
Fueron predichos como No Phishing, y no lo son	1212	VN
Fueron predichos como Phishing, pero no lo son	23	FP
Fueron predichos como No Phishing, pero lo son	45	FN
Total, de registros del Dataset de Prueba:	**2211 (20%)**	

F. Indicadores:

Finalmente, y en cumplimiento con los objetivos, se muestran a continuación los resultados obtenidos tras aplicar el método en las tres mejores técnicas de clasificación de páginas web de tipo Phishing.

 # Indicadores para Adaboost:

```
#Técnica con AdaBoost
clasificador = AdaBoostClassifier()
clasificador.fit(X_entrenamiento, y_entrenamiento)
Y_prediccion =
clasificador.predict(X_prueba)accuracy_score(Y_prediccion,y_prueba)
recall_score(Y_prediccion,y_prueba)
f1_score(Y_prediccion,y_prueba)

print('Métricas con la técnica '+ cabecera +' \n')
print('\n    Exactitud:
{0:.2f}'.format(accuracy_score(Y_prediccion,y_prueba)*100))
print('\n    Precisión:
{0:.2f}'.format(clasificador.score(X_entrenamiento,
y_entrenamiento)*100))
print('\n Sensibilidad:
{0:.2f}'.format(recall_score(Y_prediccion,y_prueba)*100))
print('\n     F1-Score:
{0:.2f}'.format(f1_score(Y_prediccion,y_prueba)*100))
print('------------------------------------------------------------')
```

Tabla 20 *Métricas de rendimiento con Adaboost. Elaboración propia.*

Métricas con la técnica AdaBoost	
Exactitud:	93.80
Precisión:	93.72
Sensibilidad:	93.36
F1-Score	94.52

```
#Técnica con SVM
clasificador = SVC()
clasificador.fit(X_entrenamiento, y_entrenamiento)
Y_prediccion = clasificador.predict(X_prueba)

cabecera=' SVM'
accuracy_score(Y_prediccion,y_prueba)
recall_score(Y_prediccion,y_prueba)
f1_score(Y_prediccion,y_prueba)

print('Métricas con la técnica '+ cabecera +' \n')
print('\n    Exactitud:
{0:.2f}'.format(accuracy_score(Y_prediccion,y_prueba)*100))
print('\n    Precisión:
{0:.2f}'.format(clasificador.score(X_entrenamiento,
y_entrenamiento)*100))
print('\n Sensibilidad:
{0:.2f}'.format(recall_score(Y_prediccion,y_prueba)*100))
print('\n     F1-Score:
{0:.2f}'.format(f1_score(Y_prediccion,y_prueba)*100))
print('-------------------------------------------------------------')
```

Tabla 21 *Métricas de rendimiento con SVM. Elaboración propia*.

Métricas con la técnica SVM	
Exactitud:	95.12
Precisión:	95.25
Sensibilidad:	94.13
F1-Score	95.70

```
#Técnica con XGBoost
clasificador = XGBClassifier()
clasificador.fit(X_entrenamiento, y_entrenamiento)
Y_prediccion = clasificador.predict(X_prueba)

cabecera=' XGBoost'
plot_confusion_matrix(y_prueba, Y_prediccion)

accuracy_score(Y_prediccion,y_prueba)
recall_score(Y_prediccion,y_prueba)
f1_score(Y_prediccion,y_prueba)

print('Métricas con la técnica '+ cabecera +' \n')
print('\n    Exactitud:
{0:.2f}'.format(accuracy_score(Y_prediccion,y_prueba)*100))
print('\n    Precisión:
{0:.2f}'.format(clasificador.score(X_entrenamiento,
y_entrenamiento)*100))
print('\n Sensibilidad:
{0:.2f}'.format(recall_score(Y_prediccion,y_prueba)*100))
print('\n     F1-Score:
{0:.2f}'.format(f1_score(Y_prediccion,y_prueba)*100))
```

```
print('-------------------------------------------------------')
```

Tabla 22 *Métricas de rendimiento con XGBoost. Elaboración propia.*

Métricas con la técnica XGBoost	
Exactitud:	96.92
Precisión:	98.71
Sensibilidad:	96.42
F1-Score	97.27

IV CONCLUSIONES Y RECOMENDACIONES

4.1 CONCLUSIONES

Tras el análisis teórico de la literatura científica, de las distintas técnicas de clasificación de minería de datos, las que mejor se adaptaron a la detección de páginas web de tipo Phishing, son las de tipo binarias, mismas que fueron AdaBoost, SVM y XGBoost, ya que, entre estas tres técnicas, las investigaciones mostraron una precisión alrededor del 94% en la clasificación de páginas web de tipo Phishing.

Luego de un análisis y evaluación de la literatura científica, en la aplicación de los métodos de selección de características para clasificación, se identificó a los 3 mejores llamados de filtro, de envoltura e integrado. De ello, se eligió el método de filtro, pues no genera ningún tiempo extra de aprendizaje, comparado con los otros dos mencionados, que si demandan tiempo de aprendizaje del algoritmo y además consumo de recursos.

En la implementación, se utilizó la reconocida librería Scikit-learn 0.23.2 para el uso de los clasificadores, y Python 3.8 como lenguaje de programación para la implementación del método. En ese escenario, la técnica XGBoost tuvo un 98.71% de precisión, por encima de SVM y AdaBoost, dado que es una técnica que combina clasificadores internamente, apoyándose en árbol de decisión y SVM para mejorar su desempeño. Ello indica que la técnica combinada tiene un mejor resultado.

Tras la evaluación de AdaBoost, SVM y XGBoost, luego de someter a 10 iteraciones de entrenamiento y prueba, XGBoost tuvo los mejores resultados, ya que, al ser una técnica combinada, utiliza los árboles de decisión simples de un solo nodo, y además el kernel trick base (Radial Basis Function - RBF) de SVM para las múltiples dimensiones, a fin de reducir el tiempo de análisis y el consumo de recursos.

4.2 Recomendaciones

En muy recomendable tener un amplio entendimiento de la data que se utiliza para crear el dataset, dado que cualquier característica irrelevante puesta en el dataset, podría variar la capacidad del algoritmo para predecir, es decir clasificar correctamente.

También es importante hacer un análisis teórico de la literatura científica para identificar los métodos de clasificación idóneos, que estén acorde con el propósito del trabajo, y que puedan ser demostrables en la práctica. Por ejemplo, en la actualidad, ya existen clasificadores muy específicos y cada vez más complejos, para poder analizar inclusive los datos erróneos que un clasificador simple genera.

Para lograr los resultados mostrados en el presente trabajo, es necesario considerar los 32 atributos mencionados en el dataset, de lo contrario los resultados variaran. Asimismo, es necesario considerar implementar los métodos de clasificación mencionados en el presente trabajo con las librerías de scikitlearn bajo Python.

Así también, se recomienda segmentar secuencialmente el volumen de los dataset mayores al millón de registros mejora el desempeño de cualquier clasificador, es decir, no es óptimo procesar todo el enorme dataset, cuando, con ayuda de herramientas de programación, se puede crear una secuencia de lotes.

REFERENCIAS

Association of Certified Fraud Examiners, Inc. (2017). COMPUTER AND INTERNET FRAUD. En I. Association of Certified Fraud Examiners, *2017 Fraud Examiners Manual, International Edition* (pág. 1.1305). USA: © 2020 Association of Certified Fraud Examiners, Inc.

Abdulhamit, S., & Emir, K. (2019). Comparison of Adaboost with MultiBoosting for Phishing Website Detecction. *Procedia Computer Science*, 7.

Agencia Peruana de Noticias. (30 de 06 de 2020). *Estos son los delitos informáticos más frecuentes en el Perú, según la Policía*. Obtenido de Estos son los delitos informáticos más frecuentes en el Perú, según la Policía: https://andina.pe/agencia/noticia-estos-son-los-delitos-informaticos-mas-frecuentes-el-peru-segun-policia-781320.aspx

Ali, W., & Malebary, S. (2020). PSO-Based Feature Weighting for Improving Intelligent Phishing Website Detection. *IEEE Access*, 15.

APWG. (2020). *Phishing Activity Trends Report - 1st Quarter 2020.* San Francisco: the Anti-Phishing Working Group (APWG).

Asociación de Bancos del Perú ASBANC. (01 de 01 de 2019). *MEMORIA 2019 - ASBANC.* Lima: ASOCIACIÓN DE BANCOS DEL PERÚ - Asbanc. Obtenido de Gerencia de Operaciones: https://www.asbanc.com.pe/Publicaciones/MEMORIA-ASBANC-2019.pdf

AVG Technologies. (23 de 01 de 2020). *Qué es el smishing y cómo evitarlo*. Obtenido de Qué es el smishing y cómo evitarlo: https://www.avg.com/es/signal/what-is-smishing

Bernardo, J. (08 de 03 de 2019). *Catphishing*. Obtenido de Catphishing: https://www.uoguelph.ca/ccs/catphishing

Chen, E. Y., Ye, C., Li, X., & Liu, F. (2019). OFS-NN: An Effective Phishing Websites Detection Model Based on Optimal Feature Selection and Neural Network. *IEEE Access vol. 7*, 14.

DTREG - Software For Predictive Modeling and Forecasting. (05 de 07 de 2020). *SVM - Support Vector Machines*. Obtenido de SVM - Support Vector Machines: https://web.archive.org/web/20100225141251/http://www.dtreg.com/svm.htm

EcuRed. (05 de 07 de 2020). *Clustering*. Obtenido de Clustering: https://www.ecured.cu/Clustering

EcuRed contributors. (08 de 07 de 2020). *EcuRed*. Obtenido de Weka: https://www.ecured.cu/Weka

ESET. (04 de 07 de 2020). *Phishing*. Obtenido de Phishing: https://www.eset.com/es/caracteristicas/phishing/#

Garcia, V. (15 de 05 de 2020). *Check Point alerta del aumento de campañas de phishing*. Obtenido de Check Point alerta del aumento de campañas de phishing: https://revistabyte.es/ciberseguridad/check-point/

Gestión Perú. (18 de 01 de 2020). PNP explica cómo denunciar la suplantación de identidad en redes sociales. Lima, Lima, Perú.

Gonzales, P. (01 de 07 de 2020). *Ingeniería social: evolución del phishing avanzado*. Obtenido de Ingeniería social: evolución del phishing avanzado: https://www.yolandacorral.com/ingenieria-social-evolucion-phishing-avanzado/

IBM Knowledge Center. (05 de 07 de 2020). *Nodo Red bayesiana*. Obtenido de Nodo Red bayesiana: https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_mainhelp_client_ddita/clementine/bayesian_networks_node_general.html

Interpol. (12 de 06 de 2020). *https://www.interpol.int/es*. Obtenido de Ciberdelincuencia: https://www.interpol.int/es/Delitos/Ciberdelincuencia

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). An Introduction to Statistical Learning. En G. James, D. Witten, T. Hastie, & R. Tibshirani, *An*

Introduction to Statistical Learning (pág. 426). Springer-Verlag New York: Springer Science+Business Media, LLC, part of Springer Nature.

Lucid Software Inc. (05 de 07 de 2020). *Qué es un diagrama de árbol de decisión*. Obtenido de Qué es un diagrama de árbol de decisión: https://www.lucidchart.com/pages/es/que-es-un-diagrama-de-arbol-de-decision#section_0

Malwarebytes. (06 de 07 de 2020). *Suplantación de identidad (phishing)*. Obtenido de Suplantación de identidad (phishing): https://es.malwarebytes.com/phishing/

Mao, J., Bian, J., Tian, W., Zhu, S., Wei, T., Li, A., & Liang, Z. (2017). Detecting Phishing Websites via Aggregation Analysis of Page Layouts. *Procedia Computer Sciene*, 7.

Masters, T. (2019). *Extracting and Selecting Features for Data Mining: Algorithms in C++ and CUDA C.* Berkeley: Apress.

Microsoft Prensa. (19 de 03 de 2019). *Microsoft detectó en 2018 hasta 225.000 intentos diarios de phishing*. Obtenido de Microsoft detectó en 2018 hasta 225.000 intentos diarios de phishing: https://news.microsoft.com/es-es/2019/03/19/microsoft-detecto-en-2018-hasta-225-000-intentos-diarios-de-phishing/

Minerva Data Mining. (06 de 07 de 2020). *KDD: ¿Qué es el Knowledge Discovery in Databases o KDD?* Obtenido de KDD: ¿Qué es el Knowledge Discovery in Databases o KDD?: https://mnrva.io/kdd-platform.html

NCSI. (08 de 07 de 2020). *NCSI*. Obtenido de NCSI: https://ncsi.ega.ee/country/pe/

NortonLifeLock, J. v. (07 de 07 de 2020). *What is vishing? Tips for spotting and avoiding voice scams*. Obtenido de What is vishing? Tips for spotting and avoiding voice scams: https://us.norton.com/internetsecurity-online-scams-vishing.html

Ñaupa Caraza, C. M. (01 de 01 de 2016). *Minería de datos aplicada a Fraude Electrónico en entidades Bancarias.* Lima: Repositorio UNMSM.

Orunsolu, A., Sodiya, A., & Akinwale, A. (2019). A predictive model for phishing detection. *Journal of King Saud University – Computer and Information Sciences*, 16.

Panda Security, S.L. (07 de 07 de 2020). *Pharming*. Obtenido de Pharming: https://www.pandasecurity.com/es/security-info/pharming/

Presidencia del Consejo de Ministros PCM. (07 de 07 de 2020). *Equipo de respuesta ante incidentes de seguridad digital del Perú*. Obtenido de Equipo de respuesta ante incidentes de seguridad digital del Perú: https://www.gob.pe/7739-presidencia-del-consejo-de-ministros-equipo-de-respuesta-ante-incidentes-de-seguridad-digital-del-peru

Proofpoint. (01 de 01 de 2020). *proofpoint.* California: Proofpoint Inc. Obtenido de proofpoint: https://www.proofpoint.com/sites/default/files/gtd-pfpt-us-tr-state-of-the-phish-2020.pdf

Rashid, T. T., Sallim, J. b., & Noor, Y. b. (2020). A comparative Analysis on Artificial Intelligence Techniques for Web Phishing Classification. *IOP Conference Series: Materials Science and Engineering*, 14.

Real Academia Española. (12 de 06 de 2020). *Diccionario del español jurídico*. Obtenido de https://dej.rae.es/: https://dej.rae.es/lema/delito

RSA. (01 de 05 de 2018). *RSA QUARTERLY FRAUD REPORT, Volume 1, Issue 2 Q2.* USA: RSA BUSINESS DRIVEN SECURITY.

Sonowal, G., & Kuppusamy, K. (2017). PhiDMA – A phishing detection model with multi-filter approach. *Journal of King Saud University – Computer and Information Sciences*, 14.

Torres-Domínguez, Omar et al. (2018). Detección de anomalías en grandes volúmenes de datos. *Scielo*, 1-14.

Tripatji, D., Nigam, B., & Edla, D. R. (2017). A Novel Web Fraud Detection Technique using Association Rule. *Procedia Computer Science*, 8.

UNAM-CERT. (18 de 01 de 2019). *¿Qué es el phishing?* Obtenido de ¿Qué es el phishing?: https://www.seguridad.unam.mx/que-es-el-phishing-2

Universidad Nacional del Noreste de Argentina. (05 de 07 de 2020). *MINERÍA DE DATOS – REDES.* Buenos Aires: Universidad Nacional del Noreste de Argentina. Obtenido de MINERÍA DE DATOS – REDES: http://exa.unne.edu.ar/depar/areas/informatica/dad/BDII/Presentaciones_Proyector/Mineria_de_Datos_Redes_Neuronales.pdf

Yang, P., Zhao, G., & Zeng, P. (2019). Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning. *IEEE Access*, 14.

ANEXOS

Tabla 23. *Ficha* de registro de datos

Definición:	Dada las características de las técnicas de clasificación predictivas, y a la selección de variables a considerar durante la ejecución de pruebas de detección de páginas web, los consumos de recursos pueden incrementarse tanto que pueden afectar las condiciones aceptables para tener un resultado oportuno, sin errores y preciso.

Este registro tiene la finalidad de recolectar información que permita medir el consumo de recursos de CPU y memoria RAM, durante la ejecución del ejercicio de detección de páginas web de tipo Phishing. Con ello se podrá revelar la eficiencia del método, así como dimensionar las capacidades de los recursos para ejecutar el algoritmo sin errores, interrupciones y de manera oportuna.

1. IDENTIFICACIÓN DEL ANALISTA

Nombres y apellidos:

Celular:

Correo:

2. LISTA DE MÉTODOS DE MEDICIÓN

METODO	CPU	RAM

3. OBSERVACIONES DE MÉTODOS DE MEDICIÓN

METODO	CPU	RAM

Tabla 24. Ficha de Registro de Resultados.

TABLA DE REGISTRO DE RESULTADOS

Fecha:	Analista:	
Lugar:		
REGISTRO DE DATOS PARA EL ANALISIS		
Objetivo:		
Tipo de muestra	Ruta de Muestra	Volumen de Muestra
Herramienta:	Computador:	
REGISTRO DE DATOS DURANTE EL ANÁLISIS		
Hora Inicio:	Hora Fin:	
Datos procesados:	Datos con error:	
OBSERVACIONES:		
REGISTRO DE METRICAS DE VARIABLES DEPENDIENTES		
EXACTITUD:		
PRECISIÓN:		
SENSIBILIDAD:		
OBSERVACIONES		

FACULTAD DE INGENIERÍA, ARQUITECTURA Y URBANISMO
RESOLUCIÓN N°1318-2020/FIAU-USS

Pimentel, 17 de julio de 2020

VISTO:
El Acta de reunión N°1606-2020, de fecha 16 de junio de 2020 del Comité de investigación de la Escuela profesional de INGENIERIA DE SISTEMAS, para la ejecución : **"IMPLEMENTACIÓN DE UN MÉTODO DE CLASIFICACIÓN DE MINERÍA DE DATOS PARA DETECTAR PÁGINAS WEB DE TIPO PHISHING"**, presentado por el(los) tesista(s) **MAGUIÑA MAZA JEAN CARLOS y SOTO CALDERON JOSE LUIS,** del Programa de estudios **INGENIERIA DE SISTEMAS**, y;

CONSIDERANDO:

Que, de conformidad con la Ley Universitaria N° 30220 en su artículo 48º que a letra dice: "La investigación constituye una función esencial y obligatoria de la universidad, que la fomenta y realiza, respondiendo a través de la producción de conocimiento y desarrollo de tecnologías a las necesidades de la sociedad, con especial énfasis en la realidad nacional. Los docentes, estudiantes y graduados participan en la actividad investigadora en su propia institución o en redes de investigación nacional o internacional, creadas por las instituciones universitarias públicas o privadas.";

Que, de conformidad con el Reglamento de grados y títulos en su artículo 21° señala: "Los temas de trabajo de investigación, trabajo académico y tesis son aprobados por el Comité de Investigación y derivados a la facultad o Escuela de Posgrado, según corresponda, para la emisión de la resolución respectiva. El periodo de vigencia de los mismos será de dos años, a partir de su aprobación. En caso un tema perdiera vigencia, el Comité de Investigación evaluará la ampliación de la misma.

Que, de conformidad con el Reglamento de grados y títulos en su artículo 24º señala: La tesis es un estudio que debe denotar rigurosidad metodológica, originalidad, relevancia social, utilidad teórica y/o práctica en el ámbito de la escuela profesional. Para el grado de doctor se requiere una tesis de máxima rigurosidad académica y de carácter original. Es individual para la obtención de un grado; es individual o en pares para obtener un título profesional. Asimismo, en su artículo 25° señala: "El tema debe responder a alguna de las líneas de investigación institucionales de la USS S.A.C.".

Que, en el Acta de reunión N°1606-2020 de fecha 16 de junio de 2020, del Comité de investigación de la Escuela profesional de INGENIERIA DE SISTEMAS, se indica entre los acuerdos la aprobación del Proyecto de tesis denominado "IMPLEMENTACIÓN DE UN MÉTODO DE CLASIFICACIÓN DE MINERÍA DE DATOS PARA DETECTAR PÁGINAS WEB DE TIPO PHISHING" de la línea de investigación de INFRAESTRUCTURA, TECNOLOGÍA Y MEDIO AMBIENTE, a cargo de MAGUIÑA MAZA JEAN CARLOS y SOTO CALDERON JOSE LUIS en condición de estudiante, del Programa de estudios INGENIERIA DE SISTEMAS.

Estando a lo expuesto, y en uso de las atribuciones conferidas y de conformidad con las normas y reglamentos vigentes;

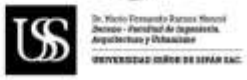
SE RESUELVE:

ARTÍCULO 1°: APROBAR, el Proyecto de denominado **"IMPLEMENTACIÓN DE UN MÉTODO DE CLASIFICACIÓN DE MINERÍA DE DATOS PARA DETECTAR PÁGINAS WEB DE TIPO PHISHING",** perteneciente a la línea de investigación de **INFRAESTRUCTURA, TECNOLOGÍA Y MEDIO AMBIENTE**, a cargo de **MAGUIÑA MAZA JEAN CARLOS y SOTO CALDERON JOSE LUIS**, del Programa de estudios **INGENIERIA DE SISTEMAS.**

ARTÍCULO 2°: ESTABLECER, que la inscripción del Título de Proyecto de tesis se realice a partir de emitida la presente resolución y tendrá una vigencia de dos (02) años.

ARTÍCULO 3°: DEJAR SIN EFECTO, toda Resolución emitida por la Facultad que se oponga a la presente Resolución.

REGÍSTRESE, COMUNÍQUESE Y ARCHÍVESE



Cc: Interesado, Archivo