Detección y mitigación de ataques de ingeniería social tipo Phishing utilizando minería de datos

Rosero Gomezcoello, Johanna Mishell

Departamento de Ciencias de la Computación

Carrera de Ingeniería de Sistemas e Informática

Trabajo de titulación previo, la obtención del título de Ingeniera en Sistemas e Informática

Dr. Fuertes Díaz ,Walter Marcelo

18 de diciembre del 2020

Resultados de Urkund

URKUND

Document Information

Analyzed document	Tesis_Johanna_Mishell_Rosero_Gomezcoello_ID_L00263618_Final.docx (D92413972)
Submitted	1/15/2021 10:00:00 PM
Submitted by	FUERTES DIAZ WALTER MARCELO
Submitter email	WMFUERTES@ESPE.EDU.EC
Similarity	3%
Analysis address	wmfuertes.espe@analysis.urkund.com

WALTER MARCELO FUERTES DIAZ

Firmado digitalmente por WALTER MARCELO FUERTES DIAZ
Nombre de reconocimiento (DN): c=EC, o=SECURITY DATA S.A. 1, ou=ENTIDAD DE CERTIFICACION DE INFORMACION, serialNumber=090720114320, cn=WALTER MARCELO FUERTES DIAZ
Fecha: 2021.01.16 07:05:10 -05'00'

Sources included in the report

SA	**Universidad de las Fuerzas Armadas ESPE / OnaZapata_TESIS (002).docx** Document OnaZapata_TESIS (002).docx (D47114989) Submitted by: rgguerrero@espe.edu.ec Receiver: rgguerrero.espe@analysis.urkund.com	1
SA	**Universidad de las Fuerzas Armadas ESPE / tesisV1.8._rev4docx.docx** Document tesisV1.8._rev4docx.docx (D85170985) Submitted by: aobaldeon@espe.edu.ec Receiver: aobaldeon.espe@analysis.urkund.com	1
W	URL: http://dspace.unl.edu.ec:9001/jspui/bitstream/123456789/11457/1/Jaramillo%20Zhingr ... Fetched: 12/2/2020 12:25:06 PM	1

WALTER MARCELO FUERTES DIAZ

Firmado digitalmente por WALTER MARCELO FUERTES DIAZ
Nombre de reconocimiento (DN): c=EC, o=SECURITY DATA S.A. 1, ou=ENTIDAD DE CERTIFICACION DE INFORMACION, serialNumber=090720114320, cn=WALTER MARCELO FUERTES DIAZ
Fecha: 2021.01.25 18:14:40 -05'00'

Dr. Fuertes Díaz, Walter Marcelo

C.C: 1707017701

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

CARRERA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

CERTIFICACIÓN

Certifico que el trabajo de titulación, **Detección y mitigación de ataques de ingeniería social tipo Phishing utilizando minería de datos** fue realizado por la señorita **Rosero Gomezcoello, Johanna Mishell**, el mismo ha sido revisado en su totalidad, analizado por la herramienta de verificación de similitud de contenido; por lo tanto cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la "Universidad de las Fuerzas Armadas ESPE", razón por la cual me permito acreditar y autorizar para que lo sustenten públicamente.

Sangolquí, 18 de Diciembre del 2020

WALTER MARCELO FUERTES DIAZ

Firmado digitalmente por WALTER MARCELO FUERTES DIAZ
Nombre de reconocimiento (DN): c=EC, o=SECURITY DATA S.A. 1, ou=ENTIDAD DE CERTIFICACION DE INFORMACION, serialNumber=090720114320, cn=WALTER MARCELO FUERTES DIAZ
Fecha: 2021.01.25 18:15:21 -05'00'

Dr. Fuertes Díaz, Walter Marcelo
C.C: 1707017701

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

CARRERA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

RESPONSABILIDAD DE AUTORÍA

Yo, **Rosero Gomezcoello, Johanna Mishell**, con cédula de ciudadanía n° 1727349076, declaro que el contenido, ideas y criterios del trabajo de titulación: **Detección y mitigación de ataques de ingeniería social tipo Phishing utilizando minería de datos** es de mi autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la "Universidad de las Fuerzas Armadas ESPE", respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Sangolquí, 18 de Diciembre del 2020

Rosero Gomezcoello, Johanna Mishell
C.C. 1727349076

ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

CARRERA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

AUTORIZACIÓN DE PUBLICACIÓN

Yo, **Rosero Gomezcoello, Johanna Mishell**, con cédula de ciudadanía n° 1727349076, autorizo a la "Universidad de las Fuerzas Armadas ESPE" publicar el trabajo de titulación: **Detección y mitigación de ataques de ingeniería social tipo Phishing utilizando minería de datos** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi autoría y responsabilidad.

Sangolquí, 18 de Diciembre del 2020

Rosero Gomezcoello, Johanna Mishell
C.C. 1727349076

Dedicatoria

Quiero dedicar este proyecto de titulación, primeramente, a Dios por darme la vida y ser mi guía en cada paso que he dado en mi vida, por haberme dado la fuerza y sabiduría necesaria para afrontar cualquier reto o adversidad que se me ha presentado.

A mi familia por el amor, tiempo, dedicación y esfuerzo que me han entregado durante todos estos años, por motivarme a llegar lejos y ser una buena persona, por no permitirme rendir y seguir adelante hasta cumplir este sueño que hoy se hace realidad. Gracias a su apoyo hoy estoy cumpliendo esta meta.

Especialmente a mis padres Irina y Byron quien, me impulsaron a continuar, en aquel momento que pensé que esta carrera no era para mí. Ellos me enseñaron a no rendirme y seguir adelante, gracias por motivarme y no permitir que deje de luchar mis sueños. Se los dedico como agradecimiento de todo el esfuerzo, amor que me entregan día a día.

A mi mami Diana que siempre me acompañado desde el primer día que me toco ir a la universidad y estar pendiente en cada detalle y sobre todo el amor que me has dado siempre.

Especialmente este proyecto se lo dedico a mi papito Fausto que está en el cielo y estoy segura que se sentirá muy orgulloso de mi, era quien esperaba con ansias verme graduada, pero se hoy desde el cielo está celebrando, como él sabía hacerlo. Gracias porque nunca le importó dejar aún de lado sus cosas por apoyarme, irme a retirar, y acompañarme en cada logro en mi vida. Esto va para el que me enseñó el valor del esfuerzo y la dedicación y sobre todo la responsabilidad, valores que en mi vida diaria y profesional las llevaré presente siempre.

Agradecimiento

Primeramente, quiero agradecer a Dios por la vida que me ha dado y por permitirme cumplir un objetivo más en mi vida. Gracias por que me ha dado la sabiduría y fortaleza necesaria para poder afrontar toda adversidad que se me han presentado en el transcurso de mi etapa universitaria.

Agradezco a mi familia por el amor que me dan día a día, por la comprensión y paciencia, porque muchas veces no pude compartir con ellos por mis obligaciones y responsabilidades en la universidad. Gracias por motivarme a cumplir mi sueño que hoy se hace realidad y no rendirme. Gracias por acompañarme en cada paso que dado y sobre todo hacerme sentir lo orgullosos que están de mí, eso me ha impulsado a seguir adelante. Han sido mi apoyo en mi crecimiento profesional, pero especialmente en mi crecimiento personal, enseñándome a ser fuerte contra las adversidades, y ser perseverante pese a que existan momentos difíciles. Gracias a ustedes me convertido en una persona de valores, que tiene muy en claro que, lo más importante es la felicidad propia y luchar por ello.

Quiero agradecer también a esta institución que me abrió las puertas y a cada uno de los ingenieros que compartieron sus conocimientos, y contribuyeron en mi formación profesional, gracias por sus consejos y prepararme para la vida.

Especialmente quiero agradecer a mi tutor de tesis Dr. Walter Fuertes, por ser una maravillosa persona, gracias por compartir sus conocimientos y tiempo para el desarrollo de este proyecto, por su amistad sincera y cada consejo en clase que estoy segura que lo tendré muy presente toda mi vida.

Por último, pero no menos importante quiero agradecer a cada persona que formó parte de mi vida en esta etapa universitaria y contribuyó con mi crecimiento académico y personal.

Contenido

ÍNDICE DE TABLA

ÍNDICE DE FIGURAS

Resumen

Este estudio tuvo como objetivo diseñar e implementar un modelo de precisión para detectar y mitigar ataques Phishing en correos electrónicos, utilizando técnicas de minería de datos. Como primer paso, se realizó una investigación bibliográfica sobre las técnicas, métodos y herramientas actuales de minería de datos empleados en la detección de Phishing. Luego se identificaron las características de correos infectados que hacen que el ataque de Phishing sea exitoso. Para ello, se realizó un análisis de diferentes correos con Phishing de tres importantes fuentes tales como: www.monkey.org, www.enron.org y www.PhishTank.com , los mismos que permitieron la generación de un dataset. Para el diseño e implementación del modelo, se empleó la metodología CRISP-DM. Con ello se generó el modelo de detección, en base a las características que reconocen a un correo como Phishing. Dentro del proceso de minería de datos se desarrolló un análisis predictivo de datos que consistió en la extracción de información existente y su utilización para predecir tendencias y patrones de comportamiento. Así mismo, se desarrolló un análisis descriptivo utilizando algoritmos de minería de datos siendo Random Forest la de mayor precisión. Por último, instalando la librería Twilio en Python, se implementó el despliegue de un mensaje a WhatsApp al detectar un correo con Phishing, motivo por el que se otorga mayor validez a la investigación realizada. En último lugar, se evaluó el modelo, mediante pruebas de concepto en un ambiente controlado, cuyos resultados muestran la funcionalidad del modelo, puesto que alcanzó un grado de precisión superior al 97% en la detección de correos infectados con Phishing.

KEYWORDS:

- **MINERÍA DE DATOS**
- **PHISHING**
- **ATAQUES DE INGENIERÍA SOCIAL**
- **CIBERSEGURIDAD**

Abstract

This study aimed to design and implement a precision model to detect and mitigate phishing attacks in emails, using data mining techniques. A literature review was conducted on current data mining techniques, methods, and tools used in Phishing detection as a first step. Then, the characteristics of infected emails that make a phishing attack successful were identified. To do this, an analysis was carried out of various phishing emails from three important sources: www.monkey.org, www.enron.org, and www.PhishTank.com, which allowed the generation of a dataset. The CRISP-DM methodology was used to design and implement the model. This generated the detection model, based on the characteristics that recognize an email as Phishing. A predictive data analysis was developed within the data mining process, which consisted of the extraction of existing information and its use to predict trends and behavior patterns. Likewise, a descriptive analysis using data mining algorithms was developed, being Random Forest the most accurate. Finally, by installing the Twilio library in Python, a message was displayed to WhatsApp when a phishing mail was detected, which is why the research was given more validity. Finally, the model was evaluated through proofs of concept in a controlled environment, the results of which show the functionality of the model, as it achieved a degree of accuracy greater than 97% in detecting phishing emails.

KEYWORDS:

- **DATA MAINING**
- **PHISHING**
- **SOCIAL ENGINEERING ATTACKS**
- **CYBER SECURITY**

Capítulo I

Introducción

Antecedentes

El crecimiento exponencial del Internet ha dado paso a que el mundo de la ciberdelincuencia crezca (Tabares, 2015), permitiendo a los delincuentes encontrar nuevas maneras de aprovecharse de las vulnerabilidades de los sistemas. Existe un sin número de ciberdelitos como son: la suplantación de identidad, pornografía infantil, estafas, robo de información confidencial mediante Phishing (Medero), etc. Especialmente este último es uno de los métodos más antiguos y utilizados por cibernéticos.

El mundo de la ciberseguridad ha ido creciendo y combatiendo a gran escala estos ataques, sin embargo, el Phishing sigue sobresaliendo. En el artículo realizado por Juan Manuel (Arán, 2019) se menciona, uno de los motivos por los que el ataque de Phishing sigue avanzando efectivamente, es que gran porcentaje de los usuarios en internet no comprenden el significado ni el peligro que están expuestos, según datos publicador por Google.

El Phishing es un ataque informático donde los delincuentes conocidos como phishers suplantan una entidad de confianza mediante ingeniería social para la obtención de información confidencial. En la tesis de Thasphon Chuenchujit (Chuenchujit, 2016) menciona una forma común de Phishing consiste en enviar un correo electrónico a usuarios, con el remitente afirmando ser una entidad confiable. Referente a esto cabe destacar que todos los usuarios que utilizan correos electrónicos están expuestos a ser vulnerados.

Por otro lado, en el ámbito laboral las empresas usualmente se manejan con un correo institucional para el desarrollo de su trabajo y comunicación entre los usuarios. Hecho que ha obligado a las empresas a tomar medidas de seguridad efectivas para limitar el acceso de intrusos, y detectar cualquier tipo de ataque. Pese a esto los sistemas aún no

detectan de manera autónoma patrones o características de un correo infecto con Phishing. Esto ha ocasionado gran pérdida de información en las empresas, y ha causado gran incertidumbre en las personas expertas en seguridad., Aunque las medidas de seguridad crecen, los atacantes encuentran nuevas maneras de atacar y persuadir. Se ha buscado diferentes técnicas de detección de Phishing entre ellas se mencionan en el artículo escrito por Diana Sastoque (Tabares, 2015) :detección de ataques de Phishing basado en la categorización de enlaces (Atighetchi, 2009), prevención de ataques de Phishing basada en atributos (Dunlop, 2010), desarrollo de un sistema experto basado en las características de las páginas web para detectar sitios web de Phishing (Shreeram, 2010).

Problemática

El Phishing es una de las amenazas más comunes y graves existentes durante los últimos 20 años en la red. Se efectúa generalmente a través de servicios de mensajería instantánea o de correo electrónico (Gyuris, 2018), al aprovechar las vulnerabilidades de las empresas y de los usuarios en el ciberespacio. En este tipo de ataque, los “phishers” intentan robar información confidencial del usuario al hacerse pasar por una entidad confiable (Chuenchujit, 2016) y efectuar fraudes, generalmente financieros.

De acuerdo con el Reporte de Investigaciones de Brechas de Datos (DBIR) (verizon, 2016), emitido por Verizon en 2019, el Phishing es la principal variedad de acción de amenazas en brechas de seguridad. También señala que los usuarios son más susceptibles a ser víctimas de Phishing si reciben el ataque por medio de dispositivos móviles.

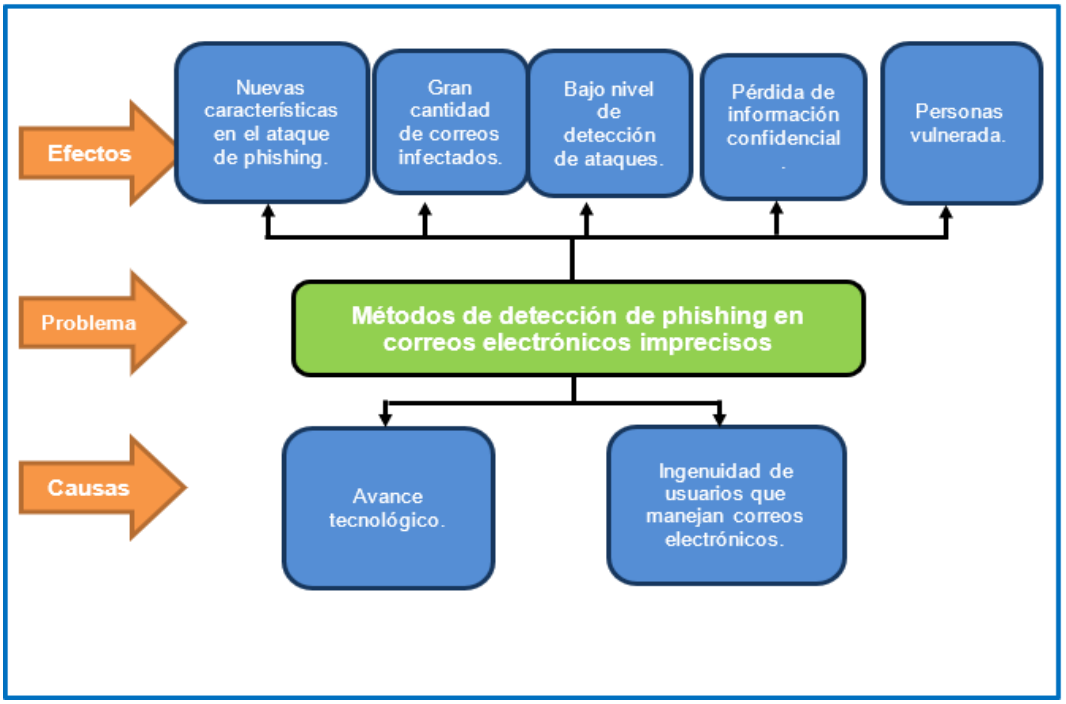
Según el reporte del Anti-Phishing Working Group (APWG) (APWG, 2019), el número total de sitios de Phishing detectados en el primer trimestre del 2019 fue de 180.768, superior a los 138.328 vistos en el cuarto trimestre de 2018. En el segundo trimestre en cambio (APWG, 2019) la cifra aumentó a 182,465. En este mismo contexto, según datos compartidos durante la presentación en BlackHat (Arán, 2019), el 68% de los correos de

Phishing bloqueados de manera diaria por Gmail contienen nuevas variantes, lo que obliga a diseñar e implementar una herramienta de protección para la ciberseguridad a fin de detectarlos y establecer contramedidas.

A continuación, en la Ilustración 1 se representa el árbol de problemas sobre el tema, en donde se especifica el problema con sus respectivas causas y efectos.

Figura 1

Árbol De Problemas



Como se observa en el árbol de problemas las principales causas de que los métodos actuales de detección de Phishing sean imprecisos se dan por el vertiginoso avance tecnológico el mundo actualmente y la ingenuidad de los usuarios al manejar los correos electrónicos, facilitando a los phishers encuentra nuevos métodos para poder realizar ataques, aprovechándose del desconocimiento de las víctimas y a su vez de las vulnerabilidades de los sistemas en este caso correos electrónicos presentan.

Así mismo los métodos actuales que son encargados de la detección de estos ataques ya no son suficientes debido a la existencia de nuevos patrones o características que los

métodos convencionales no abarcan, retrasando el tiempo de detección de correos infectados y permitiendo al usuario abrir correos sin ningún problema, provocando el aumento del número de ataques diarios actualmente.

Es por esta razón que el presente trabajo de investigación plantea un modelo de detección de ataque de ingeniería social tipo Phishing con el fin de una detección temprana de estos ataques en los correos electrónico y así poder mitigarlos correctamente.

Formulación del Problema

¿Es posible desarrollar un Modelo que detecte y mitigue ataques de ingeniería social tipo Phishing que permita disminuir el número de correos con Phishing?

Justificación

Hoy en día las empresas y los usuarios realizan grandes esfuerzos de inversión en mecanismos de seguridad de la información, tales como UTMs, SIEMs, firewalls, IDS/IPS, antimalware, antispam, antivirus, etc. Sin embargo, los delincuentes cibernéticos logran atravesar los citados mecanismos, de tal forma que el correo infectado ya está almacenado en el servidor de correo electrónico de la empresa. Esto motiva a la industria y a la academia a que se sigan buscando soluciones utilizando nuevas técnicas relacionadas con la seguridad de la información, el aprendizaje automático y la analítica de datos.

Justificación Teórica

Primera parte: caracterizar correos infectados con Phishing. Para obtener las características de los correos infectados con Phishing, se plantea extraer ejemplos de enlaces verificados contaminados de fuentes de listas negras de correos. En general, el análisis de estos sitios es un proceso que se automatizará con algoritmos de minería de datos para obtener un dataset (conjunto de datos) que se procesará hasta contar con características suficientes para la detección de Phishing en correo electrónico. Como fuente

primaria

de información, se utilizarán las bases de datos descargables de PhishTank (PhishTank, 2011), a través de su API. PhishTank es un sitio comunitario gratuito donde los usuarios envían, verifican, buscan y comparten datos de Phishing continuamente. Las bases de datos proporcionadas están disponibles en varios formatos y se actualizan cada hora para una rápida detección de Phishing.

Segunda parte: diseño e implementación, la minería de datos ha sido utilizada en soluciones para la detección de Phishing. Sin embargo, día a día se vienen implementando nuevas técnicas y algoritmos que incrementan el nivel de precisión de sus hallazgos. En este sentido, se planea utilizar para el desarrollo del modelo la metodología CRISP-DM y la herramienta Python.

Tercera parte: detección y mitigación de ataques de Phishing. Existen estándares internacionales debidamente aprobados de seguridad de la información y ciberseguridad como la NIST, ISO27000, COBIT e ITIL.

Justificación Metodológica.

Se aplicará la metodología de investigación - acción, la cual es un proceso sistemático que involucra la recopilación y el análisis de datos, así como la reflexión y discusión entre investigadores (Ali, Examining the Efficacy of Online Self-Paced Interactive Video-Recordings inNursing Skill Competency Learning: Seeking Preliminary Evidence Through an Action Research, 2009). Esta metodología incluye un “proceso de investigación y aprendizaje a través de la relación a largo plazo del investigador con un problema" (Juan Miguel Moine,Ana Silvia Haedo). En donde el proceso de la indagación y el producto de la misma proporcionan conocimiento y aportan a la aplicabilidad del proyecto.

-El manejo de datos se llevará a cabo con la metodología de investigación descriptiva, la cual utiliza la clasificación, la medición y la comparación para determinar,

describir o identificar un fenómeno (Rusydiana, 2016). El resultado de esta metodología intenta describir relaciones entre las diferentes variables (Awi, 2014).

-En lo que tiene que ver con el diseño e implementación del modelo, se empleará la metodología Cross Industry Standard Process for Data Mining (CRISP-DM) debido a que es el modelo de procesos más utilizado para proyectos de minería de datos, por su aceptación en la industria y porque cumple con la mayoría de los argumentos evaluados en (A. Guayasmín, 2018).

Esta metodología creada por el grupo de empresas SPSS, NCR y Daimer Chrysler en el año 2000 (Juan Miguel Moine,Ana Silvia Haedo), propone seis fases: comprensión del negocio, comprensión de datos, preparación de datos, modelado, evaluación y desarrollo; necesarias para descubrir patrones en el conjunto de datos recopilado, determinar las características claves de un correo con Phishing y desarrollar

el modelo más apto para su detección y mitigación.

Justificación práctica

El proyecto está dirigido a pequeñas y medianas empresas del Ecuador que deseen aumentar su nivel de seguridad en la organización y de esta manera puedan prevenir el robo de su información a través de ataques de Phishing por correo electrónico.

Objetivo

Objetivo general

Diseñar e implementar, un modelo de precisión para detectar y mitigar Phishing en correos electrónicos utilizando técnicas actuales de minería de datos para aumentar el nivel de seguridad de los correos electrónicos.

Objetivos específicos

- Investigar las técnicas, métodos y herramientas actuales de minería de datos empleados para la detección y mitigación de Phishing.

- Identificar las características de correos infectados que hacen que el ataque de Phishing sea exitoso.
- Diseñar e implementar un modelo de detección de Phishing a través de minería de datos.
- Evaluar el modelo diseñado, mediante pruebas en un ambiente controlado, donde se pueda observar el comportamiento del modelo e identificar la cantidad de correos infectados con Phishing que son detectados.

Alcance

Este proyecto comprende el diseñar e implementar un modelo para la detectar y mitigar de Phishing en correos electrónicos. Este modelo será diseñado mediante el uso de minería de datos, permitirá realizar la detección de ataque Phishing eficaz en un ambiente controlado. Además, el modelo contará con un dataset donde se encontrarán almacenadas la mayor cantidad de características que hacen de un correo electrónico ser considerado afectado por Phishing, para posteriormente poder mitigarlos. Todo el proceso se lo realizará bajo la metodología CRISP-DM, la cual define fases para un correcto diseño del modelo bajo minería de datos.

Para delinear de forma adecuada el alcance de la investigación planteada, se proponen varias preguntas de investigación asociadas a los objetivos específicos, que se muestra en la Tabla 1:

Tabla 1

Preguntas de Investigación

Objetivo específico	Pregunta de investigación
i. Investigar las técnicas, métodos y herramientas actuales de minería de datos empleados para la detección y mitigación de Phishing.	**RQ1.** ¿Cuáles son las técnicas de minería de datos más utilizadas para detectar Phishing?
ii. Identificar las características de correos infectados que hacen que el ataque de Phishing sea exitoso.	**RQ2.** ¿Cuáles son las características principales que contiene un correo electrónico exitoso de Phishing?
iii. Diseñar e implementar un modelo de detección de Phishing a través de minería de datos.	**RQ3.** ¿Cuál sería un modelo óptimo de minería de datos para detectar Phishing?
iv. Evaluar el modelo diseñado, mediante pruebas en un ambiente controlado, donde se pueda observar el comportamiento del modelo e identificar la cantidad de correos infectados con Phishing que son detectados.	**RQ4.** ¿Cuál es el porcentaje deefectividad en la detección de Phishing en correo electrónicos, en el ambiente de pruebas?

Hipótesis

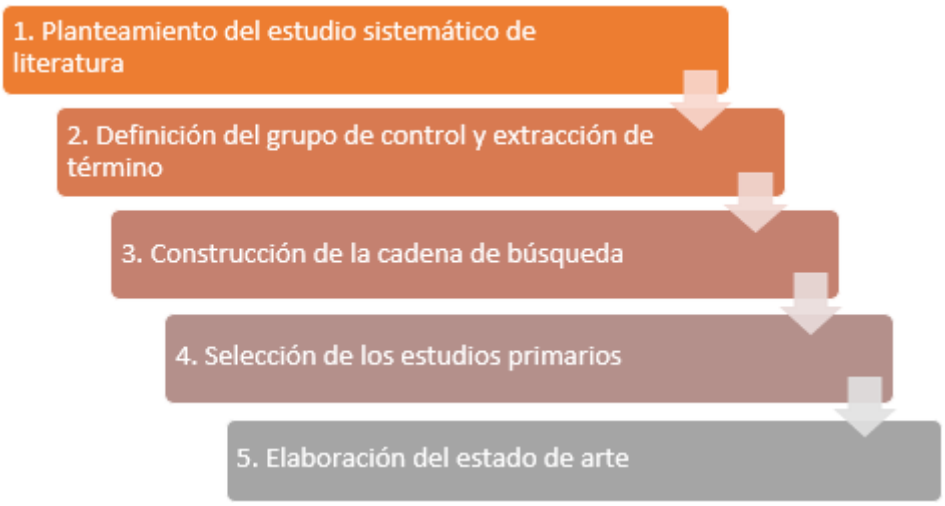
Un modelo predictivo de detección de Phishing en correos electrónicos permitirá disminuir el número de correos infectados con ataques de Phishing.

Estado del Arte

Para el análisis y revisión de la literatura enfocada a diferentes métodos de detección para Phishing en los correos electrónicos se realizó un proceso de revisión literaria preliminar con la obtención de palabras claves que faciliten la construcción de una cadena de búsqueda, realizando un mapeo sistemático que se detalla a continuación. Ver **Figura 2**

Figura 2

Proceso de búsqueda científica



Planteamiento del estudio de literatura

Para iniciar el proceso de desarrollo del mapeo sistemático de la literatura, en esta fase se realizó la descripción del problema central del proyecto de titulación, objetivos alcanzar y sus respectivas preguntas de investigación. Además, se especificó los criterios de inclusión y de exclusión para la extracción e identificación de los estudios primarios.

Definición del grupo de control y extracción de términos

En esta fase se identificar los diferentes estudios que conforman el grupo de control, basándose en los criterios de inclusión y exclusión definidos en la fase anterior. Estos estudios primarios pertenecen a paper o bases digitales importantes. Los estudios propuestos se muestran en la **Tabla 2**.

Tabla 2

Grupo de control

Título	Cita	Palabras clave
Understanding User Behaviors When Phishing Attacks Occur	(Li Y., Xiong K., Li X., 2019)	Phishing, Electronic mail, Sorting, Task analysis, Ducts, Time measurement, Indexes
Feature selection for Spam and Phishing detection	(Toolan F., Carthy J., 2010)	feature selection, spam detection, Phishing detection, unsolicited bulk email, junk filters, Ham corpora.
A Literature Survey on Social Engineering Attacks: Phishing Attack	(Surbhi Gupta., Abhishek Singhal., Akanksha Kapoor.,2016)	Phishing attack; Social engineering attack; spoofed email; Personal data;
A Methodical Overview on Phishing Detection along with an Organized Way to Construct an Anti-Phishing Framework	(Patil S., Dhage S., 2019)	Phishing websites, Machine Learning, anti-Phishing, Phishing attack, security and privacy, website features, classification, Phishing approaches
Phish-Net: Investigating phish clusters using drop email addresses	(Zawoad S., Kumar Dutta A., Sprague A., Hasan R., Britt J., Warner G)	Phishing, Clúster, Investigation, Detection

Tras un análisis de los estudios del GC, se permitió seleccionar las palabras más relevantes y con mayor frecuencia en los artículos científicos, alineados al objetivo de la búsqueda, en este caso fueron: Phishing, cybercrime, cyberattack, detection, mitigation y email.

Construcción de la cadena de búsqueda

Una vez identificado las palabras clave del grupo de control, en esta fase se crean posibles cadenas de búsqueda en la base digital seleccionada. Se construyó la siguiente cadena: **((((("All Metadata": Phishing) AND "All Metadata": CYBERCRIME) OR "All Metadata": cyberattack) AND "All Metadata": detection) OR "All Metadata": mitigation) AND "All Metadata": email)**, obteniendo 234 resultados estrechamente relacionados con la

temática, aplicando conectores al aplicar la búsqueda en la base digital de IEEE xplore.

Selección de los estudios primarios

Al realizar la búsqueda en la base digital con la cadena idónea, se obtuvo 206 artículos relacionados al tema. Se aplica filtros a los resultados bajo los siguientes criterios.

- ✓ Año 2016-2019
- ✓ Términos del índice: Informática

En base a los filtros antes mencionados, y el criterio de los investigadores, se eligieron 6 estudios primarios, los cuales constituyen la base para realizar el estudio del estado del arte, los cuales se muestran en la **Tabla 3.**

Tabla 3

Estudios seleccionados

Código	Título	Cita
EP1	Detection of Phishing attacks	Muhammet Baykara., Zahit Ziya Güre. (2018)
EP2	Employing machine learning techniques for detection and classification of Phishing emails	Moradpoor N., Clavie B., Buchanan B, (2017)
EP3	Detecting Phishing Websites through Deep Reinforcement Learning	Chatterjee M., Namin A, (2019)
EP4	Formally Reasoning about the Cost and Efficacy of Securing the Email Infrastructure	Speicher A., Steinmetz M., Künnemann R., Simeonovski M., Pellegrino G., Hoffmann J
EP5	Phishing in Depth – Modern Methods of Detection and Risk Mitigation	Backes M, (2018) Bikov T., Iliev T., Gr. Y., Mihaylo G., Stoyanov I, (2019)
EP6	Phishlimiter: A Phishing Detection and Mitigation Approach Using Software-Defined Networking	Chin T., Xiong K., Hu C, (2018)

Elaboración del estado del arte

EP1 (Muhammet Baykara., Zahit Ziya Güre. (2018)): Detection of Phishing attacks

En el presente documento los autores hacen un análisis íntegro sobre lo que es el Phishing, aclarando el panorama para conocer a fondo cómo se realiza el ataque que puede ser el envío de enlaces maliciosos o archivos adjuntos infectados por correo electrónico u otros medios con el fin de capturar credenciales de inicio de sesión o información de las cuentas de las víctimas, para la detección de esto plantean el desarrollo de un software llamado "Anti Phishing Simulator" que ayuda con la detección de Phishing sobre correoselectrónicos.

EP2 (Moradpoor N., Clavie B., Buchanan B, (2017)): Employing machine learning techniques for detection and classification of Phishing emails

El estudio se centra en los correos electrónicos y cómo detectar el Phishing sobre ellos, se plantea la posibilidad de que un correo tenga aspectos legítimos y aun así este está infectado, esto causó una pérdida de 174 millones de libras al reino Unido en el 2015, para el estudio se presenta un modelo que da resultados precisos bajo los parámetros de verdaderos positivos y falsos positivos.

EP3 (Chatterjee M., Namin A, (2019)): Detecting Phishing Websites through Deep Reinforcement Learning

La investigación se centra en la detección de sitios web maliciosos antes de que estos puedan causar daños perjudiciales a las víctimas, para esto el enfoque del documento es con el aprendizaje de refuerzo profundo utilizado para modelar y detectar direcciones maliciosas, la razón que motiva el uso de un modelo de este tipo es la naturaleza cambiante de los sitios web, por lo tanto, si se extrae las características asociadas es posible detectarlo a tiempo.

EP4 (Speicher A., Steinmetz M., Künnemann R., Simeonovski M., Pellegrino G., Hoffmann

JBackes M, (2018)): Formally Reasoning about the Cost and Efficacy of Securing the Email Infrastructure

Hacer una evaluación del costo que representa tener una estrategia para la mitigación de ataques sobre la infraestructura de correos electrónicos, la metodología propuesta en el presente documento consiste en automatizar y realizar evaluaciones formales de implementación, es decir que se incluye el impacto, la rentabilidad de diferentes estrategias de mitigación como la combinación de protocolos como IPsec, DNSSEC, DANE, SMTP STS, SMTP sobre TLS y otras técnicas de mitigación como la reubicación del servidor para mejorar la confidencialidad de los usuarios de correo electrónico en 45 combinaciones de países atacantes y defensores y nueve escenarios de costos.

EP5 (Bikov T., Iliev T., Gr. Y., Mihaylo G., Stoyanov I, (2019)): Phishing in Depth – Modern Methods of Detection and Risk Mitigation

En la actualidad se plantea que todos viven en un mundo digital con diferentes experiencias y realidades, las amenazas no son diferentes a estas realidades y una de las más comunes es sobre los correos electrónicos, que por más básicos y simples causan grandes problemas.
El documento hace énfasis en que Incluso las medidas y sistemas de seguridad modernos no son capaces de identificar y prevenir todo el contenido fraudulento que se crea y distribuye todos los días. En este documento se cubre los vectores de ataque más comunes, que incluyen las infraestructuras de correo electrónico ya masivas, las medidas de contrarrestar requeridas para minimizar el impacto sobre los entornos corporativos y qué más se debe desarrollar para mitigar los sofisticados ataques de correo electrónico modernos

EP6 (Chin T., Xiong K., Hu C, (2018)): Phishlimiter: A Phishing Detection and Mitigation Approach Using Software-Defined Networking

El presente propone una nueva técnica para la inspección profunda de paquetes (DPI) y luego la aprovechamos con una red definida por software (SDN) para identificar actividades

de Phishing a través del correo electrónico y la comunicación basada en la web. El enfoque DPI propuesto consta de dos componentes: clasificación de firma de Phishing y DPI en tiempo real. Con base en la capacidad de programación de SDN, se desarrollas el modo de almacenamiento y el modo de reenvío e inspección al tráfico directo de la red a través de una red neuronal artificial para clasificar las firmas de ataques de Phishing y diseñar el DPI en tiempo real para que Phishlimiter pueda abordar con flexibilidad

Características del estado del arte

Existen estudios que muestran métodos actuales, en los que se han basado para la detección de ataques Phishing en correos electrónicos, entre ellos la seguridad cognitiva. Sin embargo, los estudios enfocados en un modelo de detección de phishing basado en procesos de minería de datos son muy pocos. En los estudios primarios recolectados, se menciona que aun los sistemas modernos de seguridad no son capaces de identificar y prevenir todo el contenido fraudulento que viene en los correos, y las características que definen que un correo está infectado. Esto se debe a que el avance tecnológico provoca que phishers busquen nuevas maneras de realizar su ataque.

Estas dificultades conducen a proponer la generación de un modelo predictivo, que pueda detectar si un correo tiene Phishing de manera efectiva, mediante algoritmos y técnicas efectivas de minería de datos.

Capítulo II

Marco metodológico

En este capítulo se detalla la metodología más adecuada que se aplicó para el proceso de investigación y desarrollo de este proyecto. Seguidamente se describe el marco teórico, que permitió tener un amplio conocimiento del tema, para poder lograr conseguir resultados efectivos.

Metodología de investigación

Tomando en cuenta la problemática de esta investigación, se debe determinar que el Phishing es un riesgo social actualmente que requiere una solución, y emprender un cambio en cuanto a los efectos que causa. En base a esto se puede determinar que su proceso se ajustó a una investigación-acción, la mima que fue descrita por el científico Lewin en 1946.

Según (Creswell, 2015) la investigación -acción tiene una semejanza con los métodos de investigación mixtos, ya que utiliza una colección de datos que pueden ser de tipo cuantitativo o cualitativo, con la diferencia que este se centra en una problemática en específico. El autor clasifica este tipo de investigación como práctica o participativa. En la Tabla 2 se puede observar las características de cada una de estas clasificaciones.

Tabla 4

Características de la clasificación de la investigación-acción

Práctica	Participativa
Estudia todas las prácticas realizadas localmente.	Su enfoque va dirigido a estudiar temas sociales diversos como la vida de un grupo de personas.
Realiza una investigación individual o en equipo	Realiza una investigación equitativa del grupo
Su objetivo se enfoca en el desarrollo y aprendizaje de los participantes	Su objetivo se enfoca en los cambios que se pueda realizar , con el fin de lograr mejorar el desarrollo humano.

Práctica	Participativa
Para la resolución del problema esta metodología implementa un plan de acción en cuanto a la mejora o cambio La persona que lleva el mando lo hace conjuntamente quien investiga y los miembros de la comunidad.	Da libertad para desarrollar tanto al investigador como a los participantes. -

Nota. Esta tabla identifica las características de la clasificación de la investigación-acción, para poder diferenciarla adecuadamente, extracto de Creswell (2005). Educational research. Planning, conducting and evaluating quantitative and qualitative research. USA: Pearson. (Creswell, 2015)

En base a las características definidas en la Tabla 2 y tomando en cuenta el objetivo y alcance de la investigación se determina que la metodología a utilizar es la investigación-acción práctica, la misma que va a desarrollarse mediante etapas que tiene la metodología con un espiral. Según (Hernández, Fernández y Baptista, 2014) los ciclos del proceso son:

- Detección y diagnóstico del problema de investigación
- Elaboración del plan para la solución del problema
- Implementación del plan y evaluación de resultados.

Realimentación, la cual conduce a un nuevo diagnóstico y una nueva etapa en espiral.

Estos autores describen la importación de cada fase dentro el proceso de la investigación. La primera etapa es importante ya que se llega a conocer a profundidad el problema, su consecuencia, personas inmersas, eventos, situaciones. Todo esto permitirá definir correctamente cual va ser la solución buscada para resolver el problema y sobre todo elegir con exactitud el camino por el cual empezar el proceso de investigación. También permite la recolección de toda la información que permitirá ayudar con la investigación.

La segunda etapa, se basa en la elaboración del plan que se va a implementar para la solución del problema, seguida por la tercera etapa en donde ya se pone en marcha el plan a

desarrollar, realizando prueba con los datos recolectados, generando resultados. Estos resultados deben ser evaluados hasta conseguir un resultado eficaz.

Como se puede observar además que la metodología se acopla con el proceso de investigación a realizarse, cada una de las etapas están inmersas en el proceso de minería de datos que se va a detallar. Asegurando así que este proyecto tendrá un sustento eficaz en cuanto a su investigación y la aplicación de la minería adecuada, permitiendo obtener el modelo deseado que resuelva el problema inicial.

Capítulo III

Marco Teórico

A continuación, se identifican las variables que forman parte de la hipótesis planteada en el capítulo anterior.

Señalamiento de variables

Variable independiente: modelo predictivo y detección de Phishing en correos electrónicos.

Variable dependiente: cantidad de correos infectados con ataques Phishing.

Red de categorías de la investigación

Con la finalidad de definir el marco teórico se realizó una red de categorías, partiendo de las variables dependiente e independientes de la hipótesis, identificadas en el punto anterior. Estas redes de categorías se detallan en la **Figura 3 y 4 .**

Figura 3

Red de la categoría de la variable dependiente.

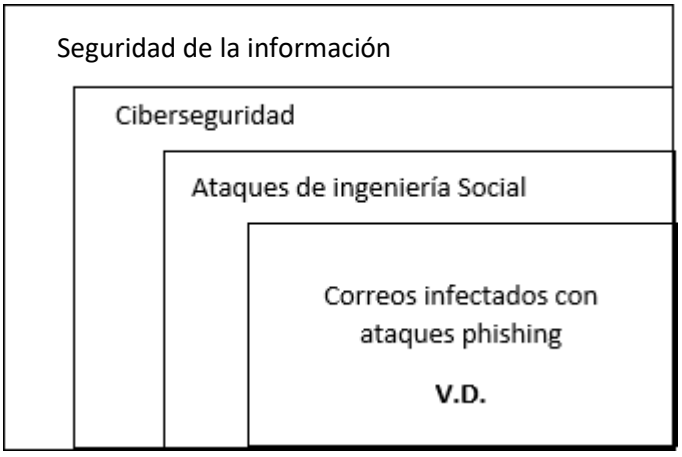
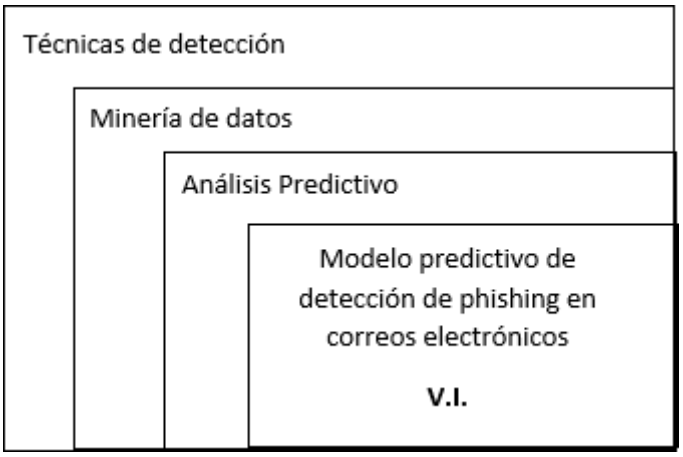


Figura 4

Red de la categoría de la variable independiente.



Seguridad de la información

El activo principal de todo sistema informático es la información y es necesario protegerlo de manera cautelosa, tomando en cuenta todos los aspectos posibles que podrían causar vulnerabilidad hacia este.

La seguridad de la información permite realizar una valoración, asegurar que se pueda identificar y gestionar riesgos que pueda existir en contra del activo de información.(Bertolín, 2008).

Para esto es necesario realizar ciertas consideraciones tales como, la información es más sensible para la protección, si esta información se perdiera valorar qué consecuencias podría traer, las amenazas que puede haber en contra de esta información y que se puede hacer en el caso de que esta información sea sustraída (Bertolín, 2008).

Es necesario mencionar que existe la norma ISO 27001 que contempla varios procedimientos y estándares para realizar el proceso de aseguramiento de la información,

Esta norma al estar en un enfoque basado en el ciclo de vida de mejora continua o Deming, que consiste en Planificar, Hacer, Controlar y Actuar (PDCA por sus siglas en ingles), que son esenciales en la construcción del sistema de seguridad de la información (SGSI) que

sugiere la norma realizar.

Para construcción de un sistema de seguridad de la información (SGSI) que es el aspecto fundamental de procedimientos, protocolos, controles, políticas y auditorias tal que estas en conjunto permitan a la organización un mejor control y mitigar los riesgos que existen sobre su activo principal que es la información, la norma plantea el seguimiento de nueve que guían a la organización en la construcción e implementación del mismo, estos pasos son los siguientes:

- Análisis y evaluación de riesgos.
- Implementación de controles.
- Definición de un plan de tratamiento de los riesgos o esquema de mejora.
- Alcance de la gestión.
- Contexto de organización.
- Partes interesadas.
- Fijación y medición de objetivos.
- Proceso documental.
- Auditorías internas y externas

Importancia de la seguridad de la información.

La creciente amenaza sobre la información según (kaspersky, 2019) es de un 60% a nivel de Latinoamérica, hace importante para los usuarios de sistemas de información el implementar medidas para proteger este activo, entre las amenazas actuales se tiene: fraudes asistidos por computador, espionaje, sabotaje, vandalismo, fenómenos naturales y la ingenuidad por parte de los usuarios (Dirección General de Modernización Administrativa, 2012).

La importancia de la seguridad de la información ha ido creciendo con el paso del

tiempo y el constante crecimiento del uso de sistemas de información por parte de organizaciones que hacen atractivo el realizar ataques a estos sistemas ya que sobre ellos se maneja cada vez más cantidad de información y hace que las consecuencias de la sustracción de esta sean cada vez más graves, por lo tanto la cantidad de ataques crece de igual manera en la que los sistemas son utilizados y la diversificación de estos es totalmente cambiante ya que los atacantes consiguen nuevas maneras de vulnerar los esquemas de seguridad.

Por esta razón fue necesario la construcción de una norma que abarque de manera evolutiva la mejora de la seguridad de la información como la norma ISO 27001 la cual se expuso con anterioridad.

Elementos de seguridad de la información.

Para poder realizar la implementación de un sistema de seguridad de la información es importante tomar en cuenta los siguientes puntos (Dirección General de Modernización Administrativa, 2012).

- Procesos: Es el conjunto de actividades que están relacionadas con el fin de transformar las entradas en una salida deseada.
- Tecnología: Es el medio que maneja la información, su elaboración y almacenamiento.
- Personas: Son aquellos que utilizan esta información con el fin de obtener provecho de la misma.

Seguridad informática o ciberseguridad.

La seguridad informática es un elemento primordial en los sistemas actuales de información, este permite prevenir y detectar el uso no autorizados de dichos sistemas, el proceso de la seguridad informática implica la protección en contra de intrusos que podrían utilizar de manera maliciosa ciertos recursos informáticos obteniendo beneficios de ellos.

Clasificación.

Al ser tratado el tema de ciberseguridad o seguridad informática es importante identificar qué tipos de seguridad existe, de acuerdo con la función que realicen dentro del sistema, como están implementados o los elementos que utiliza para realizarlos (RAMOS, 2011).

Seguridad activa y pasiva.

Esta clasificación se deriva del nivel de actuar que tiene la medida dentro del sistema.

- Activa.

La seguridad activa es aquella que a medida que va detectando las amenazas usa mecanismos para ir solucionando el problema, por ejemplo, un sistema de login que a medida que se ingresa las contraseñas este permite o deniega el acceso (RAMOS, 2011).

- Pasiva.

La seguridad pasiva es un conjunto de medidas que se ponen en acción una vez haya ocurrido un incidente, y hace que el impacto de este sea el menor posible, por ejemplo, un sistema de respaldos que guarda información cada cierto tiempo y en el momento de existir una pérdida o fallo es posible acudir a estos respaldos y normalizar el sistema con el menor impacto (RAMOS, 2011).

Seguridad Física y Lógica.

Esta clasificación viene dada a nivel de software o hardware.

- Física.

Como su nombre lo indica este nivel de seguridad viene dado a nivel de hardware con la implementación de dispositivos que protejan los sistemas de información de alguna manera física, por ejemplo, un dispositivo biométrico para el acceso a los servidores.

- Lógico.

El nivel de seguridad lógico es el más utilizado, ya que es uno de los que se intentan

de vulnerar como mayor frecuencia como lo indica las estadísticas de Kaspersky (kaspersky, 2019), por esta razón es la que se debe considerar al momento de implementar seguridad informática, la seguridad informática se encarga de controlar que el acceso al sistema informático, desde el punto de vista software, se realice correctamente y por usuarios autorizados, esto desde adentro del sistema como por fuera como una red externa (RAMOS, 2011).

Para ello debe tomar distintas medidas de seguridad. Cada vez los sistemas controlan más la seguridad del equipo informático ya sea por arte de un error, por un uso incorrecto del sistema o del usuario, o por un acceso no controlado físicamente o a través de la red, o por programas maliciosos, como los virus, espías, troyamos, gusanos o Phishing (RAMOS, 2011).

Es casi imposible establecer sistemas que sean totalmente seguros, pero existen medidas que reducen el riesgo al mínimo.

Ataques de ingeniería social

Después de haberse detallado y evidenciado los tipos de ataque que la seguridad informática busca mitigar, se hace un enfoque en los ataques de ingeniería social que se aprovecha de las vulnerabilidades de las personas que se encuentran conectadas al mundo del internet.(Blanco, 2018).

Los ataques de ingeniería social se los realiza por diferentes medios que son los siguientes:

Mediante correo electrónico, que son los ataques del tipo Phishing.

Por teléfono, ataques del tipo Phishing que conociste en realizar llamadas telefónicas con el fin de suplantar identidades y conseguir cierta información.

Por medio de redes sociales, extorsionar al usuario con la obtención de información personal de los mismos y de esta forma estos brinden información de accesos.

Mediante unidades externas, el uso de dispositivos USB que están cargados con malwares que sin darse cuenta puede infectar dentro de una organización y todo su sistema externo de seguridad se ve vulnerado.

Por mensaje de texto, suplantar identidades mediante el uso de mensajes de texto que proporcionan enlaces o llamen a un número de teléfono o respondan un mensaje.

Para evitar ser víctima de este tipo de ataques es necesario tomar ciertas conductas al momento de hacer uso de los sistemas informáticos.

- No revelar datos personales ni datos confidenciales.
- Cuidado al compartir información en redes sociales.
- Verificar los ficheros adjuntos.
- Instalar y mantener siempre actualizado el antivirus.
- Sentido común y precaución.

Correos infectados con ataques Phishing

Según datos proporcionado por la página oficial de Kaspersky se registra más de 746 mil ataques de malware diarios durante los últimos 12 meses en Latinoamérica, que da por resultado un promedio de 9 ataques de malware por segundo, los ataques del tipo Phishing son constantes en esta región principalmente en Brasil que pasa a formar parte de los 20 países más atacados por Phishing a nivel mundial (kaspersky, 2019).

Fases

Las fases del ataque Phishing se divide en tres que son las siguientes (Velasquez, 2013).

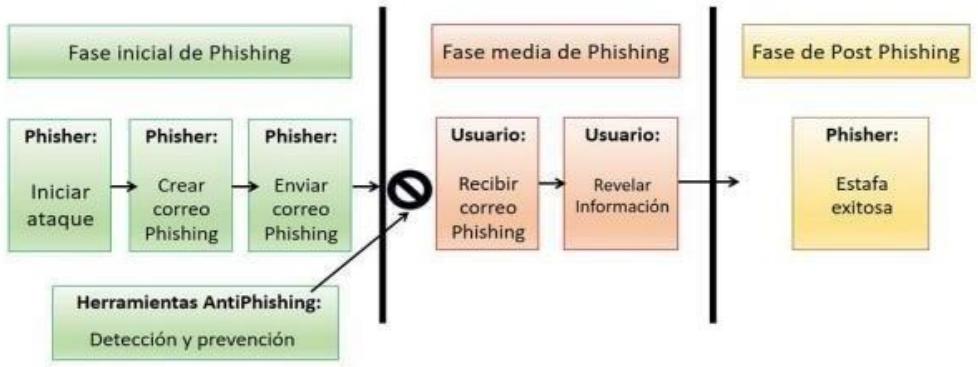
- Primera fase: los phishers alcanzan gran nivel de usuarios mediante el usode correos, foros, enlaces o chat, donde envían mensajes llamativos, como ofertas, empleos, descuentos. En el caso de que el usuario llegue a caer en la

estafa, las herramienta o persona denominadas intermediarios proceden a llenar campos donde se encuentran datos confidenciales como: número de tarjetas, cuentas y datos personales.

- Segunda fase: se realiza el ataque de Phishing, mediante el envío de correos masivos que toman el nombre de entidades bancarias o dominios públicos. Estos mensajes solicitan información o existen técnicas que realizan ataques específicos.
- Tercera fase: esta última fase ya se enfoca en el proceso donde los phishers empiezan a obtener beneficios del fraude realizado, sacando sumas grandes de dinero. Normalmente los intermediarios se encargan de realizar las transferencias a los estafadores.

Figura 5

Fases del Phishing



Nota. El gráfico indica las fases que tiene un el Phishing en un correo electrónico. Tomado de (Velasquez, 2013)

Daños causados.

El phishing puede provocar, grandes daños a los usuarios que fueron víctimas de este ataque como es la perdida de credenciales del correo electrónico o a gran escala

pérdidas económicas representativas. La suplantación de identidad está avanzando considerablemente, debido a las vulnerabilidades del usuario como es la falta de desconfianza al momento de entregar información personal a personas desconocidas, muchas veces se ha dado el caso que los phishers han logrado obtener hasta número de tarjetas de crédito l de seguridad social, solo con hablar con la víctima. Es suficiente tener esta información, para que los phishers hagan uso de estos datos y creen cuenta falsas, y así aprovecharse de lo beneficios de las cuentas obtenidas. (Velasquez, 2013).

Medidas de prevención.

- Las medidas de prevención que se debe considerar para disminuir el riesgo de ser víctimas de un ataque tipo Phishing son.
- Si recibe un correo electrónico en donde se solicita información personal o financiera, no se debe responder o responder.
- Si el correo le envía enlace, URL, hipervínculos para que los abra, no lo haga.
- No envié su información personal o confidencial a través de un correo electrónico.
- No acceder a lugar públicos con poca seguridad
- Verifique los dominios, y certificados de seguridad de un sitio web, antes de ingresar información personal.
- Actualizar el sistema operativo y antivirus de la PC.
- Revisar mensajes o notificaciones bancarias.
- Evitar la descarga de documento o archivos de fuentes desconocidas.

Técnicas de detección

Luego de describir el ataque de ingeniería social tipo Phishing en los correos electrónicos en el punto anterior, es necesario conocer cuáles son las técnicas de detección más utilizadas en la actualidad.

Se han desarrollado varias herramientas informáticas a nivel mundial, utilizadas en la detección y mitigación de Phishing. El estudio y análisis literario permite agrupar estas herramientas como soluciones en grupos diferentes (Hernández Dominguez & Сторчак, 2019) que se detallan a continuación.

Educación y legal

De acuerdo con (Hernández Dominguez & Сторчак, 2019) una manera eficaz encontrada para contrarrestar y detener un poco este tipo de ataques, fue el adaptarlo a un marco legal y así poder condenar a las personas involucradas, a través de un conjunto de leyes y normas que determinan a la suplantación como delito.

Paralelo a lo mencionado existen soluciones que han sido dedicadas a la formación de usuarios para la identificación de estos ataques, los mismo que se realizan en un ambiente de entrenamiento, con ciertas simulaciones de situaciones reales. Entre proveedores de estas soluciones se tiene:

- Sitio web APWG: es consorcio que reúne empresas que han sido afectadas por Phishing, patrocinando una serie de soluciones para ese ataque. (Anti-Phishing Working Group, 2019).
- Confíenseles el proveedor líder en soluciones de detección y defensa conta el Phishing. Ofreciendo el reconocimiento, acción y reporte de estos ataques enfocados a la ciberseguridad.

Con respecto al ámbito de la educación, se han ido desarrollando diferentes

soluciones a base de la experiencia del usuario. La necesidad de contrarrestar este delito ha provocado que se formen distintas comunidades en línea dedicadas a monitorear y estudiar información y actividades relacionadas con el Phishing. El objetivo de estas investigaciones es el de poder generar herramientas y métodos que puedan combatir con todo tipo de ataque de suplantación. Existen varias comunidades anti-Phishing, como son: APWG, *PhishTank*, *Millersmiles* y Symantec (Anti-Phishing Working Group, 2019), (PhishTank, 2011), (Bright, 2011) , (Nahorney, 2015)

Computacional utilizando métodos semiautomáticos:

Actualmente manejar correos electrónicos se ha vuelto una práctica muy común en la mayoría de usuarios en el internet, convirtiéndose en un medio para poder comunicarse además de ser una forma para poder enviar archivos e información, tanto en la vida cotidiana como en la profesional. Este hecho ha dado paso a que se desarrollen diferentes herramientas como antivirus y spam, encargados bloquear el paso o envió de correos sospechosos. Sin embargo, estos programas no detectan correctamente algunos correos, eliminando así correos genuinos, clasificándolo como fraudulento (R.M., F, & McCluskey, 2015).

Según (Hernández Dominguez & Сторчак, 2019) uno de los objetivos de las herramientas anti-Phishing es aumentar la precisión al momento de detecta phishing. Y de esa manera reducir el nivel de falsos positivos, ya que esto daría más seguridad a los usuarios, evitando realizar un trabajo manual, para verificar si los correos en spam son los correctos.

Para poder implementar esta técnica se utiliza las bases de datos donde se almacenan listas negras y blancas, las misma que tiene nombres, direcciones o dominios de correos electrónicos, reconociendo así los correos verdaderamente dañinos infectados de Phishing.

En el artículo de (S., M, P, & J., 2010) realiza un estudio enfocado en el análisis de enlaces infectados, donde se define un 47% a 83% de efectividad en la detección de phishing.

Métodos inteligentes de Machine Learning

Teniendo en cuenta que Phishing es un problema típico de clasificación, su tratamiento se lo puede focalizar en diferentes técnicas adecuadas que logran la obtención del conocimiento a partir de la caracterización del phishing. Entre estas técnicas se menciona la minería de datos y Machine Learnning. (Hernández Dominguez & Сторчак, 2019).

Según la UNESCO define que los sistemas de ML pueden realizar predicciones exactas, cuando los datos empleados tengan similitud (UNESCO, 2019). Por lo que es necesario definir correctamente las características y datos de los correos infectados con Phishing para un correcto desarrollo del modelo utilizando diferentes algoritmos enfocados en esta técnica, entre ellos se encuentra: árbol de decisiones, clasificación basado en reglas, modelos probabilísticos, etc.

Minería de Datos

La Minería de Datos (MD) es considerada uno de los métodos definidos como principal en la identificación y recolección de datos y características. Aquí se definirá y detallará dicha técnica siendo la base principal para la construcción de este proyecto de investigación.

La técnica denominada como Data Mining puede ser definida como el proceso de extracción de información y patrones de comportamiento que permanecen ocultos entre grandes cantidades de información: Es un proceso iterativo en el que a los avances que se van produciendo en cada paso se les denomina descubrimientos. (KDD).

La minería de datos es una integración de varias tecnologías como el aprendizaje automático, la estadística. Soporte en la toma de decisiones, gestión de bases de datos y procesamiento de gran cantidad de datos. Para el desarrollo de estos procesos existen diferentes técnicas que se desarrollan en diversas áreas como puede ser aquella que utilizan la inteligencia artificial. (Martínez, 2018).

Proceso de Minería de Datos

Para iniciar el proceso de minería de datos es necesario identificar todos los datos con los que se va a tratar, es necesario que se recolecte la mayor cantidad de datos necesarios. Una vez que se tiene los datos, se procede almacenarlos en una base de datos en un formato adecuado, seguidamente se selecciona los datos esenciales y se elimina los datos que no son necesarios.

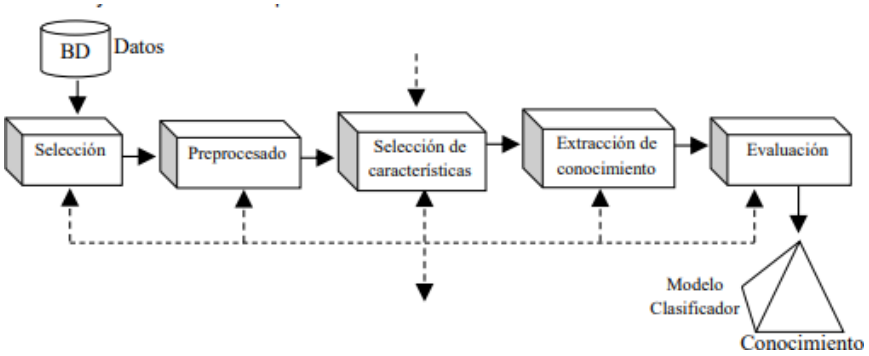
Se debe seleccionar la herramienta adecuada a utilizarse para realizar el procesamiento y análisis de datos, además de definir concretamente el objetivo que se tratando de conseguir con los datos analizarse, para así enfocarse en el desarrollo del modelo de manera adecuada.

Cuando ya se aplican las herramientas elegidas o desarrolladas por un mismo, se debe interpretar los resultados o patrones que se obtuvieron para poder seleccionar únicamente los resultados correctos. Una vez definido los patrones correctos, se identifica las acciones que deben ser tomadas y los procedimientos que se deben implementar.

Por último, se tiene la evaluación donde se monitorea el comportamiento de cada dato, del modelo construido y los resultados conseguidos. En la Figura se puede tener una visión general del proceso detallado.

Figura 6

Proceso de minería de datos



Nota: El gráfico indica cuál es el proceso de l minería de datos. Tomado de (Martínez, 2018)

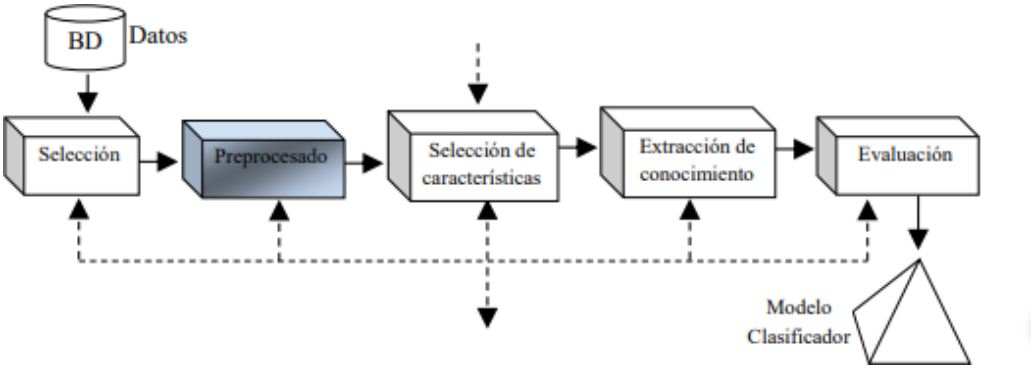
En el artículo de (Martínez, 2018) se definen los estados por los cuales pasa este proceso, a continuación, se detalla cada uno:

- Procesado de datos
- Selección de características
- Uso de un algoritmo para la extracción de la información
- Interpretación y evaluación.

Procesado de los Datos

Figura 7

Proceso Minería de Datos Preprocesados



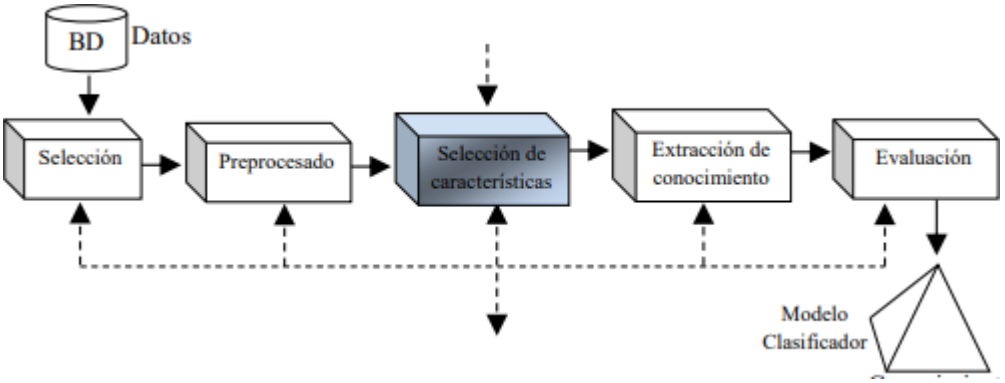
Nota: Gráfico que identifica la fase de preprocesado en el proceso de minería de datos. Tomado de (Martínez, 2018)

Cuando se obtiene datos de una fuente, no siempre todos los datos son los adecuados para procesarlos en bruto. En esta fase de preprocesamiento, se eliminan todos aquellos valores incompletos, inválidos que no proporcionan una información adecuada para el proceso respectivo. Este proceso también permite que se reduzca el número de datos, lo que facilita el proceso en cuanto a velocidad y eficacia. (Martínez, 2018)

Selección de Características

Figura 8

Proceso de Selección



Nota: Gráfico que identifica la fase de selección de características en el proceso de minería de datos. Tomado de (Martínez, 2018)

La selección de las características permite reducir el tamaño de los datos, eligiendo las variables más relevantes en el problema, y las misma que identifican si un dato tiene validez o no. Los métodos para poder seleccionar las características son:

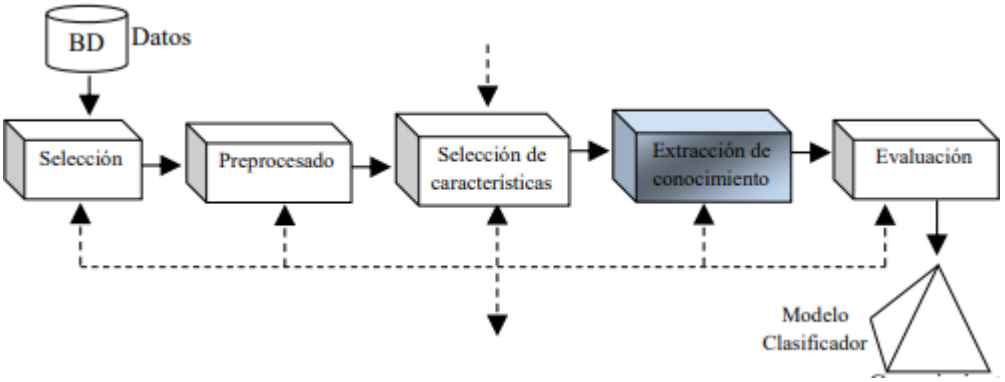
- Elección de los mejores atributos del problema.
- Variables independientes mediante test de sensibilidad, algoritmos de

distancia o heurísticos.

Extracción de conocimiento

Figura 9

Proceso de Extracción



Nota: Gráfico que identifica la fase de extracción de conocimiento en el de minería de datos. Tomado de (Martínez, 2018) *proceso*

Una vez de haber realizado el proceso de formateo y selección de características, los datos quedan listo para poder ser procesados de manera correcta. En esta fase se selecciona la técnica que permite generar un modelo que se adhiere al cumplimiento del objetivo. Cabe recalcar que para esta fase se puede hacer uso de varias técnicas obteniendo resultados favorables, en distintos modelos con procesamientos distintos, esto permite ver de mejor manera el comportamiento y así también poder seleccionar la mejor opción que permita cumplir con los objetivos de minería de datos planteados. (Martínez, 2018).

Evaluación y Validación

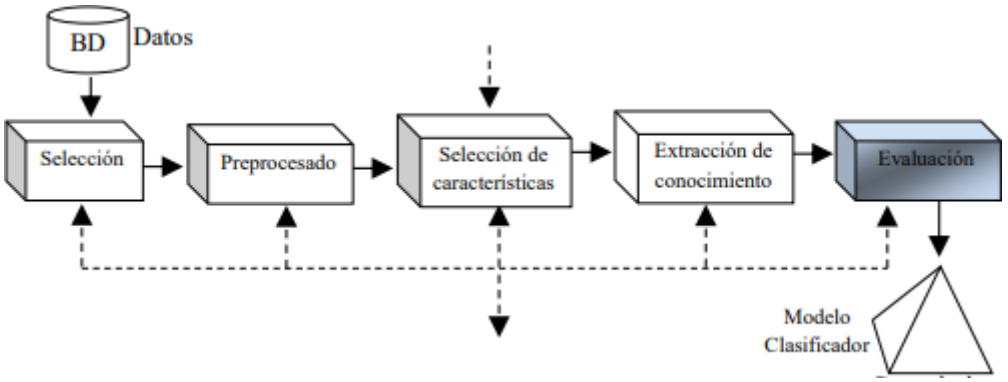
Luego de haber obtenido el modelo y el procesamiento de los datos, se debe comprobar y verificar los resultados que se obtiene, para determinar si son válidas y suficientes como para obtener resultados en cuanto al cumplimiento de objetivos. En el caso de que se hayan aplicado varios modelos, es necesario hacer una comparativa en todos los

resultados que cada modelo arroja, seleccionando aquel se ajuste mejor a la solución del problema.

Es importante mencionar que, si los resultados no son favorables, se debe retrocedes a una fase y volver a seleccionar diferentes datos o realizar un preprocesamiento diferente hasta lograr alcanzar resultados confiables y favorables.

Figura 10

Proceso de Evaluación



Nota: Gráfico que identifica la fase de extracción de conocimiento en el proceso de minería de datos. Tomado de (Martínez, 2018)

Análisis predictivo

Dentro del proceso de minería de datos se encuentra una serie de herramientas y algoritmos que permiten, la solución de problemas planteados que generalmente tiene un enfoque en dos áreas que son: Análisis predictivo y análisis descriptivo. (Romero, 2019).

El análisis predictivo es considerado un área en la minería de datos que tiene un enfoque en la extracción de información adecuada existente en los datos y su utilización para predecir tendencias de comportamiento, que puede aplicarse sobre cualquier evento desconocido, ya sea en el pasado, presente o futuro. (Timón, 2017).

Para poder realizar un análisis predictivo es indispensable tener una considerable cantidad de datos, donde se extraen características específicas que generan el problema a tratarse. La obtención de datos y procesamiento que se realizó en el proceso detallado en el punto anterior de minería de datos son fundamentales para generar un análisis predictivo.

A continuación, se detalla las tareas que abarca este análisis predictivo:

- **Clasificación o discriminación:** esta tarea es una de las más comunes en minería de datos y se basa en determinar a qué clase perteneces un dato u objeto en base a sus características. (Cortina V. G., e-archivo, 2015)
- **Clasificación suave:** Realiza la misma función de la clasificación, con la diferencia que este agrega una función que determina el nivel de confianza que tuvo la predicción. (Cortina V. G., 2015).
- **Estimación de probabilidad de clasificación:** esta tarea clasifica un objeto o dato en diferentes clases, con un nivel de probabilidad que pertenezca o no a dichas clases. (Cortina V. G., 2015).
- **Categorización:** la función de esta tarea es determinar las características de un objeto que puede corresponder a n clases. (Cortina V. G., e-archivo, 2015).
- **Regresión:** esta tarea es similar a la clasificación, ya que se le determina una clase para cada objeto con la diferencia que este se representa en base a un valor numérico. (Cortina V. G., 2015)

En el presente trabajo de investigación se realizó un modelo predictivo el cual se detallará en punto siguiente, el mismo que se hizo en base a un análisis predictivo con los dataset obtenidos con la técnica de minería de datos.

Modelo predictivo de detección de Phishing en correos electrónicos

En el desarrollo del modelo predictivo se considera una seria de métodos que ayudan a cumplir con el objetivo planteado. A continuación, se detalla los métodos cuya funcionalidad se aplica en modelos predictivos y permitirán el desarrollo de este proyecto. Cabe mencionar que para la selección de algún método es importante saber que uno puede resolver diferentes tareas, o que 1 tarea puede ser resuelta por diferentes métodos obteniendo resultados positivos. (Cortina V. G., e-archivo, 2015).

- **Técnicas bayesianas:** esta técnica utiliza el teorema de Bayes, que se puede determinar como un aprendizaje basado en parámetros, que dada una estructura de datos se puede determinar la probabilidad antes de que el hecho suceda y bajo condiciones requeridas.
- **Técnicas basadas en árboles de decisión y sistema de aprendizaje de reglas:** esta técnica clasifica un objeto en base a una serie de reglas, que pertenecen al tipo de algoritmos “divide y vencerás”.
- **Técnicas basadas en redes neuronales artificiales:** esta técnica requiere de un entrenamiento, el cual se basa en el peso de cada neurona que se interconecta entre sí, en base a la topología de la red y sus pesos en las diferentes conexiones, determina un patrón de clasificación.

Capítulo IV

Descripción del sistema

Descripción de Herramientas y librerías

Python.

Python es un lenguaje de programación orientado a objetos, de gran versatilidad y rapidez de desarrollo en todo tipo de programas para Windows, servidores de red o páginas web. (Alvarez, 2003).

Es un lenguaje de programación multiplataforma y multiparadigma, que sobresale por su código de programación legible y limpio. Python es el lenguaje indicado para trabajar con grandes volúmenes de datos, debido a que, al ser multiplataforma, permite la extracción y procesamiento de los datos, por lo que es el lenguaje elegido por las empresas de big data. (Robledano, 2019)

Este lenguaje se ha hecho muy reconocido, debido a las siguientes características:

- ✓ Contiene una gran cantidad de librerías, diferente tipo de datos y funciones implícitas en el mismo lenguaje, permitiendo realizar diversas tareas sin la necesidad de generar código desde 0.
- ✓ Lenguaje sencillo, elegante y legible que sigue reglas que hacen de su aprendizaje muy rápido y fácil
- ✓ Este lenguaje da facilidad y rapidez al momento del desarrollo de programas.
- ✓ Se pude desarrollar en diferentes plataformas como es: Unix, Windows, Os/2, Ubuntu, entre otros.
- ✓ Python es un lenguaje de programación gratuito.

Imaplib

Esta librería se implementa para comunicarse con servidores del protocolo IMAP4.Este protocolo define un conjunto de diferentes comandos enviados al servidor y las respuestas al cliente. (Schmidt, 2019)

urllib. request

Esta librería define funciones y clases que ayudan abrir diferentes enlaces, especialmente Http, utilizando autenticación básica, redirecciones, cookies, etc. (Foundation., Python, 2020)

pandas as pd

Es una herramienta para el manejo y manipulación de datos de alto nivel. Esta librería e contenida y construida con el paquete Numpy. Su estructura de datos se la denomina DataFrame, el cual re permite almacenar y manejar datos tabulados en filas y columnas de variables. (learnpython.org, 2019)

Os

Este módulo permite el acceso a las funcionalidades dependientes del sistema Operativo, además de permitir la manipulación de la estructura de directorios, para la lectura, escritura de archivos. (Uniwebsidad, 2020)

Re

Es un módulo que proporciona soporte para las expresiones regulares. Re es una secuencia especial de caracteres que ayuda a coincidir o buscar otras cadenas o conjunto de cadenas, mediante una sintaxis especializada dentro de un patrón. (Foundation., Python, 2020)

xml.dom import minidom

Es una implementación de Document Object Model. Esta librería provee un método que es el encargado de leer un archivo y analizar el contenido den un XML y poder representarlos en el programa Python. (Bassi, 2019)

From email.header import decode header

Esta librería es parte de la API de email en Python. Encargada de la codificación y decodificación de las cabeceras. Permite controlar el conjunto de caracteres que se usa cuando se codifican las cabeceras. (Foundation., Tutorial de Python 3.6.3 documentation, s.f.)

Pprint

Es una librería que permite imprimir estructuras de datos de Python de una manera que se la pueda utilizar como entradas para el intérprete utilizado. (Foundation., Python, 2020)

Mailbox

Esta librería permite la manipulación de archivos de correo electrónicos, defina una interfaz para lograr acceder a mensajes de correos electrónicos en varios formatos: Maildir, mbox, Mh, Babyl, MMDF. (Schmidt, 2019)

Tldextract

Esta librería permite la separación del dominio superior genérico como: .com, .org, .edu o el dominio geográfico .ec, .uk de subdominio de un enlace, mediante el uso de sufijos de una lista pública. (Esteve, 2018)

Twilio

Es una librería auxiliar de Twilio en Python, que permite la conexión de la API desde la aplicación desarrollada con el lenguaje Python. (Twilio Docs, 2020)

Metodología de desarrollo

Para realizar la construcción del modelo de precisión se aplicó la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), considerado como un modelo y guía eficaz, para la elaboración de un modelo basado en minería de datos."

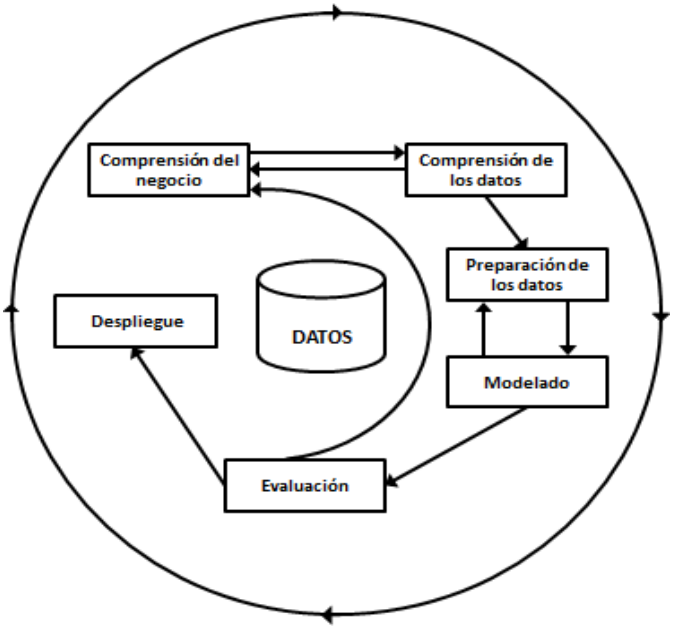
Considerado como un modelo y guía eficaz, para la elaboración de un modelo basado en minería de datos.

La metodología CRISP-DM contempla el proceso de análisis de datos como un proyecto profesional, estableciendo un contexto mucho más amplio en la elaboración de los modelos. (Román, 2016).

La metodología CRISP-DM está estructurada por 6 fases, cuya secuencia no es rígida, es decir permite el movimiento bidireccional entre cada fase, según se considere necesario. En la Figura 1 se puede visualizar las fases en las que la metodología está dividida y las posibles secuencias a seguir durante el desarrollo del modelo.

Figura 11

Fases de la metodología CRISP-DM



Nota: El gráfico representa las fases en las que se divide la metodología CRISP-DM y las secuencias que se puede seguir. Tomado de (Cortina V. G., 2015)

Primera Fase Comprensión del negocio

En este apartado de documento, se detallará el proceso de desarrollo de todas las tareas abarcadas en la primera fase de esta metodología. Con el fin de determinar los objetivos y requisitos principales del proyecto desarrollado desde la perspectiva del negocio.

Determinar los Objetivos del Negocio

Los ataques de ingeniería social tipo Phishing se han convertido en una de las amenazas más comunes y graves durante los últimos años. Los atacantes han encontrado diferentes métodos para robar información confidencial del usuario al hacerse pasar por una entidad confiable.

Actualmente estos ataques mediante correos electrónicos han aumentado, a pesar de que existen varios métodos para la detección y mitigación de los mismos. Es por esto que se busca realizar un modelo de detección de Phishing en los correos electrónicos con la ayuda de la minería de datos.

Contexto

En referencia a la situación de la problemática, en un inicio este proyecto cuenta con una serie de características que un correo electrónico infectado con Phishing tiene. Sin embargo, no existe ningún estudio en profundidad sobre el comportamiento de correos con estas características para poder identificar correctamente un ataque Phishing y así sacar conclusiones o patrones para poder definir el modelo. También se tiene 2 bases de datos importantes, la primera con información de URL infectados de Phishing con sus respectivos dominios y propietarios, obtenido de la página PhishTank que certifica que esos datos son Phishing y la segunda contiene más de 2000 correos electrónicos con Phishing, obtenida de la página Monkey.org., la misma que se ha utilizado en varios estudios mencionados en el capítulo III.

Objetivos del negocio

El objetivo general es el diseño de un modelo de precisión para detectar y mitigar Phishing en correos electrónicos, utilizando técnicas actuales de minería de datos para aumentar el nivel de seguridad de los correos electrónicos.

Este modelo puede ser muy útil para los usuarios que manejan información confidencial en la red y que están expuestos a este ataque, poniéndolos en alerta cuando reciban un correo infectado y disminuyendo el peligro de ser víctimas de Phishing.

Criterio de éxito del negocio

Se define como criterio de éxito la posibilidad de categorizar correctamente correos que tengan o no phishing. Logrando una precisión mayor del 95%, considerando este porcentaje como aceptable.

Evaluación de la situación

Se cuenta con una base de datos, la cual contiene gran información como es URL y direcciones detectadas como phishing. Estos datos fueron obtenidos de la base de datos de la página de PhishTank en un formato .CSV, cuya descarga se la realizo sin ningún problema con la ayuda de la dirección de descarga que la misma página proporciona en diferentes formatos. Además de una API que PhishTank proporciona para poder verificar si un enlace se encuentra en su base de datos y si es Phishing o no.

En base a estudios previos acerca de ataques Phishing , se pudo visualizar que las fuentes de datos utilizadas en común para el desarrollo de pruebas eran Monkey.org y una base de datos denominada Enron.mbox ,las mismas que brindan correos electrónicos con característica de Phishing y no Phishing , cabe recalcar que estos datos son actualizados constantemente por la comunidad dedicada al estudio de este tipo de ataque ,asegurando el manejo de nuevos datos y características para mejores resultados.

Esta información, ayuda a la identificación en el correo electrónico, de un conjunto

de características definidas previamente, que serán almacenadas en un dataset y mediante procesos de minería de datos, se podrá identificar si un correo está infectado o no.

Inventario de recursos

Para los recursos de software se dispone del lenguaje de programación de alto nivel Python que proporciona todas las herramientas necesarias para desarrollar el proceso de minería de datos de manera adecuada con bibliotecas dedicadas para cada proceso como son; recolección y limpieza, exploración de datos, modelado de datos y visualización de datos.

Los recursos de hardware que se dispone es una laptop con las siguientes características:

- Marca: DELL
- Modelo: Inspiron 3442
- Procesador: Intel(R) Core™ i5-421OU CPU @1,70GHz, 1.70 GHz
- Memoria RAM: 8,00 GB
- Tipo de sistema: Sistema operativo de 64 bits, procesador x64
- Sistema Operativo: Windows 10 Pro

La descarga de la base de datos PhishTank, entrega todos los registros que son verificados como Phishing, estos datos son actualizados cada hora. Se puede definir que para este proyecto se tiene datos de la base hasta la fecha de descarga. Los enlaces obtenidos van desde el año en que se recibió el primer mensaje 2012 hasta el 29 de septiembre del 2020.

Se tiene la API de PhishTank que permitirá la verificación de enlaces en tiempo real, ya que facilita la conexión a la base de datos de PhishTank.

La base de datos obtenida de Monkey.org entrega correos electrónicos con Phishing desde el año 2012 hasta el año 2020 y la base de Error desde el año 2000 al 2020.

Requisitos, supuestos y restricciones

Como requisito necesario tener la información de correos electrónicos y la identificación de todas las características esenciales que definen un correo con Phishing.

Restricciones

Cantidad de correos electrónicos enviados a diario desde una misma fuente, ya que son tomados como spam.

Riesgos y contingencias

Como riesgos en el desarrollo del modelo se puedo identificar los siguientes:

- ✓ Información desactualizada, como es de conocimiento el avance de la tecnología ha obligado a phisher buscar nuevos métodos y maneras de realizar el ataque. Esto hace que se corra el riesgo de que la información que se obtenga no sea muy relevante e impida tener un alto nivel de detección.

  Contingencia: generar en un ambiente controla enlaces con Phishing, y enviarlos mediante correos electrónico, abarcando todos los métodos, y herramientas posibles, para tener una amplia idea de cómo puede atacar y como se puede identificar el Phishing.
- ✓ Las funciones utilizadas en el código desarrollado pueden ser sujetas a cambio de versión o deprecadas, causando conflictos en versiones futuras de Python.

  Contingencia: Con el fin de evitar estos inconvenientes se trabajó con la última versión disponible de Python.

Terminología

- **Phishing:** es una técnica de ataque informático que utiliza la ingeniería social, para engañar a sus víctimas, suplantando la identidad de entidades reconocidas y así

obtener información personal y confidencial como pueden ser credenciales, cuenta de banco, etc.

- **Análisis predictivo de datos:** Análisis que predice tendencias y comportamientos futuros en base a datos históricos registrados.
- **Árbol de decisión**: Estructura en forma de árbol, donde cada uno de sus nodos son decisiones, las mismas que permiten generar reglas para clasificación de varios datos.
- **Teorema de Bayes:** se utiliza para calcular la probabilidad de un suceso, teniendo información específica de dicho suceso.
- **Clasificación:** Proceso encargado de separar un conjunto de datos en conjuntos mutuamente excluyentes.
- **Data Mining:** conjunto de técnicas que permiten explorar gran cantidad de datos almacenados en dataset con el fin de encontrar patrones repetitivos que expliquen el comportamiento de los datos.
- **Dataset:** corresponde a los contenidos de una única tabla de base de datos, donde cada columna representa una variable en particular.

Costes y beneficios

Los datos obtenidos de la página de PhishTank y Monkey.org al igual que los correos electrónicos ENROR no generan ningún coste adicional, ya que la misma página proporciona enlaces de generación y descarga de datos gratuita.

El beneficio de este proyecto dependerá de la eficacia del modelo generado, ya que se puede integrar en servidores de correo electrónico de diferentes empresas que deseen un mayor grado de protección en sus emails. En este caso se generará un beneficio de protección que representa un valor monetario.

Determinar los Objetivos de la Minería de datos

Los objetivos en términos de minería de datos son los siguientes:

- Determinar las características esenciales e información principal para la detección de Phishing.
- Generar un código que permita, la extracción de las características definidas previamente, para la creación de un dataset.
- Entrenar el modelo con el dataset generado a través de técnicas de minería de datos seleccionadas.
- Determinar la mejor técnica en base al porcentaje de fiabilidad de cada una.
- Detectar correos electrónicos con Phishing al momento de llegar al correo personal.

Criterios de éxito de minería de datos

El criterio de éxito definido en este punto es la extracción de las características definidas previamente con un correcto tratamiento de los datos, permitiendo el desarrollo de un modelo de detección de correos electrónicos infectados de Phishing con un elevado índice de confiabilidad, específicamente definiéndolo por lo menos en un 90%. El grado de fiabilidad lo definirá el modelo, dependiendo el número de casos detectados. Este tema se volverá abordar en el paso 5 de la metodología que es la evaluación donde ya se tiene el modelo desarrollado.

Realizar el Plan del Proyecto

Para el desarrollo del modelo en este proyecto, se contempla las siguientes etapas:

- Etapa 1: Investigación y recolección de estudios primarios relacionados con ataques de Phishing. Tiempo: **1 semana**
- Etapa 2: Análisis y recolección de datos desde la página de PhishTank empleando un enlace que la página proporciona en diferentes formatos, en este caso se lo

descargó en formato .CSV. Recolección de datos con y sin Phishing de Monkey.org y Enron obtenidos en formato. mbox. En esta fase también se incluye la depuración inicial. Tiempo estimado:**2 semanas**

- Etapa 3: Realizar consultar para la exploración y verificación la calidad de los datos, en cuanto a diferentes muestras. Tiempo estimado: **1 semana**
- Etapa 4: Preparación de los datos para su respectivo análisis, lo que implica selección, limpieza, conversión y formateo de datos que no son relevantes y así facilitar el proceso de Data Mining. Tiempo estimado: **2 semanas**
- Etapa 4: Selección de características específicas de un correo infectado con Phishing, en esta etapa también se considera la generación del código en Python para la extracción de las características seleccionadas, generando una matriz con 0 y 1, los mismo que representa la presencia o no de cada uno de las características. Tiempo estimado: **1 semana y media**
- Etapa 5: Selección y aplicación de las herramientas y técnicas adecuadas para el modelado y ejecución. Tiempo estimado: **1 semana**
- Etapa 6: Generación del código para la generación del modelo de precisión para la detección de Phishing, cuyo aprendizaje se realiza con los correos Phishing obtenidos de la base Monkey.org y correos sin Phishing de la fuente Enron. Tiempo estimado: **2 semanas**
- Etapa 6: Análisis de resultados, en un proceso de evaluación en donde se medirá el porcentaje de fallos y acierto en cuanto a la detección de Phishing. Si es necesario se repetiría la etapa 4 hasta obtener resultados exitosos. Tiempo estimado: **1 semana**
- Etapa 7: Desarrollo de conclusiones y recomendaciones en base a los resultados obtenidos y los objetivos planteados. Tiempo estimado: **1 día**

- Etapa 8: Elaboración de informe, detallando el proceso realizado y los resultados obtenidos en función a los objetivos de negocio y criterios de éxitos establecidos.

  Tiempo estimado: **1 semana**

- Etapa 9: Presentación del modelo final y de los resultados obtenidos.

Evaluación inicial de herramientas y técnicas

Para poder desarrollar el proyecto en base a la minería de datos, se seleccionó la herramienta Python , ya que proporciona herramientas eficaces para el manejo y tratamiento de grandes cantidades de datos .Además de ser un lenguaje fácil de usar ya que está basado en intérpretes y maneja casi todo tipo de datos.

Para la selección de técnicas correctas que permitan desarrollar el modelo, se realizó un exhaustivo análisis de las todas las tareas necesarias, para seleccionar cual permite cumplir con eficacia el objetivo de negocio y minería de datos planteados.
A continuación, se detalla la definición y la tarea que fue seleccionada:

Tarea: Se puede definir como un tipo de problema a ser resulto por un algoritmo de minería de datos. (Cortina V. G., APLICACIÓN DE LA METODOLOGÍA CRISP-DM A UN PROYECTO DE MINERÍA DE DATOS EN EL ENTORNO UNIVERSITARIO, Octubre 2015)

Las tareas de datamining pueden dividirse en dos grupos que son: predictivas o descriptivas.

Para el desarrollo de este modelo se seleccionó el grupo predictivo ya que en base a su objetivo que es determinar casos futuros basados en datos o variable definidas, se al proceso que se deseas implementar. Dentó de este grupo se encuentran diferentes tareas cuyo objetivo dependerá del resultado o el proceso que se quiera desarrollar. Estas tareas son:

- ✓ Clasificación o discriminación: El objeto de esta tarea es clasificar objetos nuevos a la clase correcta, mediante el análisis de sus atributos y comparación con la información conocida.
- ✓ Clasificación suave: Cumple la misma función que la tarea anterior, con la diferencia que aquí se tiene un grado de certeza de la predicción realizada.
- ✓ Estimación de probabilidad de clasificación: Su proceso es parecido a la clasificación suave. La diferencia es que los resultados obtenidos, se define como un conjunto de probabilidades de que objetos pertenezca a una u otra clase.
- ✓ Categorización: Es muy diferente a las tareas descritas anteriormente, esta tare se encarga de etiquetar a un objeto en n clases.
- ✓ Regresión: al igual que la clasificación entrega un solo valor como resultado, pero en este caso este valor será numérico.

Una vez analizada cada una de estas tareas, se seleccionó como tareas adecuadas en el desarrollo del proyecto a la clasificación suave, ya que se necesita una predicción para determinar el Phishing de un correo, y su grado de certeza que determina el resultado final.

Para poder resolver la tarea identificada se necesita de técnicas, algoritmos o métodos. Es importante señalar que la resolución de estas tareas se las puede realizar con cualquier método, y la eficacia de los resultados pueden variar según la técnica utilizada.

Python como se menciona antes da la facilidad de manejar cualquier técnica o método que maneja la minería de datos, lo que permitió elegir los siguientes algoritmos para resolver el problema: Técnica Bayesiana, Técnica basada en árbol de decisión y Random Forest.

Las técnicas bayesianas tienen como objetivo estimar una probabilidad de pertenencia de un objeto o dato a una clase o grupo, a través de la estimación de

probabilidades. Para esta técnica se utiliza el teorema de Bayes.

La técnica basada en árboles de decisión y sistema de aprendizaje de reglas se basan en el principio de “divide y vencerás”. Estos árboles se creando, al momento de formar nodos, en donde se almacenan diferentes condiciones de uno o varios atributos, dividiendo un conjunto de datos en subconjuntos , dependiendo las condiciones especificadas. Este proceso se realizará continuamente hasta poder logra el resultado correcto o cumplir un criterio definido.

La técnica de Random Forest a diferencia de los árboles de decisión, este método selecciona variables al azar en cada nodo del árbol, generando múltiples árboles, creando un bosque donde el resultado final será la clase con mayor número de votos. Proporcionando resultados mucho más efectivos.

Segunda Fase Comprensión de los datos

En esta fase de la metodología CRISP-DM se realiza la recolección inicial de todos los datos necesarios para poder establecer un primer contacto con el problema, familiarizarse con los datos a utilizar y verificar la calidad de los mismos, verificando estos brinden el apoyo y características necesarias para la generación del modelo. El análisis de estos datos también ayudará a la formulación de las primeras hipótesis del proyecto.

Recolectar los datos Iniciales

Los datos recolectados inicialmente para el desarrollo de este proyecto fueron:

- **Base de datos PhishTank**

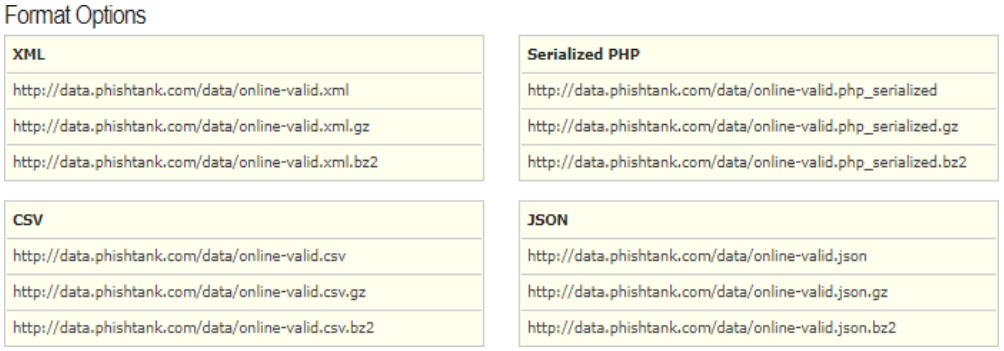
  Se selecciona la base de datos de esta fuente debido a que cuentan con grandes registros actualizados cada hora, esto permite la detección de Phishing rápida y actualizada ideal para integrar en el modelo a diseñarse. Esta es una página de compensación colaborativa de datos e información sobre Phishing en Internet. (PhishTank, 2011).

Los desarrolladores de esta página cuentan con un Api abierta para que los desarrolladores e investigadores puedan integrar en códigos anti-Phishing. Además, da la facilidad de proporcionar enlaces para la descargar de sus registros o a su vez código que permite encontrar o identificar un determinado registro con Phishing. Para la recolección de datos de este proyecto se procedió a realizar la descarga de sus registros en 2 formatos .CSV y JSON.A continuación se detalla el proceso:

- ✓ Antes de poder descargar la base de datos, fue necesario registrarse como desarrollador en la página https://PhishTank.com/.
- ✓ Luego se ingresó a las opciones de desarrollador en donde se puedo visualizar los formatos de descarga, códigos y el modelo de URL a utilizarse.

Figura 12

Formatos de descarga PhishTank

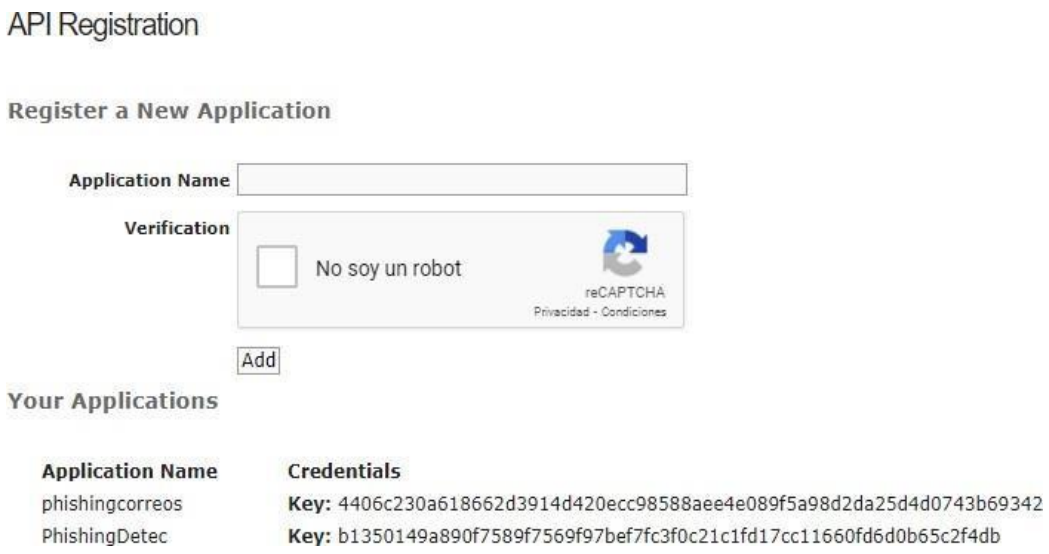


N*ota. El gráfico indica las opciones de formato que proporciona la página de PhishTank para la descarga. Tomado de* (PhishTank, 2011)

- ✓ Para poder obtener los datos de PhishTank, se solicitó una clave, la misma que fue proporcionada, al registrar el proyecto o modelo.

Figura 13

Obtención de clave para descarga de datos.



Nota. La imagen indica como se obtiene la clave, para poder realizar la descarga de los datos en la página de PhishTank. Tomado de (PhishTank, 2011)

- ✓ La clave que se obtuvo fue:

  4406c230a618662d3914d420ecc98588aee4e089f5a98d2da25d4d0743b69342

- ✓ En el buscador se ingresó la URL respectiva, con la clave generada y el tipo de formato en que se desea obtener los datos.
- ✓ La descarga de los datos iniciará.

- **Base de datos correos Monkey.org**

  La revisión bibliográfica previa, permitió encontrar esta base de datos, ya que en varios artículos científicos la utilizaron, para investigaciones de ataques Phishing obtenido resultados eficaces.

  Monkey.org es una página comunitaria, donde un grupo de personas dedicadas al estudio de ataques Phishing en correos electrónicos, suben correos infectados

diariamente, en una base de datos completa totalmente accesible para cualquier persona que necesite realizar pruebas o estudios en base a los correos almacenados en su base de datos. (Nazario, 2005).

Para la descarga de los datos se realizó los siguientes pasos:

- ✓ Se ingresó a la página de Monkey.org con el siguiente enlace para la obtención la última versión disponible https://monkey.org/~jose/Phishing/
- ✓ Se seleccionó la opción private-Phishing4.mbox que es la base de datos más actual registrada.

Figura 14

Página de descarga de archivos en Monkey.org

Index of /~jose/phishing

Name	Last modified	Size	Description
Parent Directory		-	
20051114.mbox	2014-02-10 17:00	3.9M	
README.txt	2018-07-13 17:15	844	
phishing-2015	2018-05-15 19:18	8.4M	
phishing-2016	2018-05-15 19:18	12M	
phishing-2017	2018-05-15 19:18	11M	
phishing-2018	2019-01-11 19:41	8.4M	
phishing-2019	2020-01-03 15:08	5.0M	
phishing0.mbox	2005-06-13 14:37	3.0M	
phishing1.mbox	2005-11-15 16:01	4.0M	
phishing2.mbox	2006-08-07 13:09	9.9M	
phishing3.mbox	2007-08-07 01:20	19M	
private-phishing4.mbox	2018-07-24 21:46	31M	

Nota. La imagen indica la página donde se realiza la descarga de los diferentes archivos, los mismo que se identifican con el nombre y la fecha de su última modificación. Se obtuvo de (Monkey.org, 2020)

- **Base de datos correos ENRON**

  Al igual que la base de datos descrita anteriormente, esta se obtuvo mediante el análisis y revisión de estudios primarios. Esta base de datos, tiene más de 5000

correos electrónicos regulares o sin Phishing, lo mismo que permitirán una mejor clasificación y aprendizaje del modelo.

El conjunto de datos de correo electrónico de Enror fue recopilado y preparado por el proyecto CALO (Un asistente cognitivo que aprende y organiza). (William W. Cohen, MLD, CMU, 2015).

En el enlace de la página http://www.enron-mail.com/ se realizó la descarga del archivo en formato. mbox.

- **Datos existentes de dominios principales**

  Se realizó una investigación, sobre los casos de Phishing en el Ecuador, donde se pudo determinar que el nombre de entidades como: Banco pichincha, Netflix, Spotify o YouTube son repetitivos en la suplantación de identidad, lo que resulta necesario e importante dentro de nuestro proyecto.

  Estos datos se obtuvieron mediante la lectura de sitios web donde se redactan diferentes casos de Phishing en el país. (LA HORA, 2020)

- **Caracterización del Phishing en correos electrónicos**

  Para la recopilación de las características que definen un correo electrónico como Phishing, se realizó una revisión exhaustiva de artículos, trabajos e investigaciones referentes al tema, así como también una comparación entre correos electrónicos legítimos y con Phishing. Todo este proceso permitió la obtención de la caracterización adecuada del Phishing y crear el conjunto de datos de características, la misma que será la entrada para los algoritmos de aprendizaje del modelo a desarrollarse.

A continuación, se listará los datos adquiridos de cada proceso detallado anteriormente:

PhishTank

- **Phish_id**

  Cada URL y registro de PhishTank tiene un identificador , uno que se relaciona con la URL, y permite encontrarlo con mayor facilidad.

- **URL**

  En esta columna se encuentran todas las URL tipo Phishing.

- **phish_detail_url**

  Aquí se tiene URL que direcciona directamente a una información mucho más detallada del Phishing seleccionado.

- **submission_time**

  En esta columna están las fechas y horas de cuando se realizó el Phishing.

- **verified**

  Los datos obtenidos en esta columna determinan con un “yes” o “no” si cada registro es Phishing. Estos datos se los identificó de tipo booleano.

- **verification_time**

  Estos datos definen el tiempo y fecha de verificación de Phishing

- **Online**

  Determina si los URL infectados se encuentra aún en línea.

- **Target**

  Aquí se tienen los objetivos o el medio por el cual se realizó el ataque.

Monkey.org y Enron (archivos. Mbox)

- **Return-Path**

  Dirección de procedencia del correo electrónico

- **X-Original-To**

  Dirección de destino

- **Delivered-To**

  Dirección de destino final del correo

- **Received**

  Dominio del destinatario

- **From**

  Nombre y correo del remitente

- **To**

  Nombre y correo del destinatario

- **Subject**

  Asunto del correo electrónico

- **Sender**

  Ratificación del sujeto que envía el mensaje

- **User-Agent**

  Identificación de la aplicación, del sistema operativo y el proveedor

- **X-Priority**

  Nivel de prioridad del mensaje

- **MIME-Version**

  Se utiliza para declarar la versión del estándar de formato de cuerpo del mensaje

- **Content-Type:**

  Identifica como interpretar el contenido del correo electrónico.

- **Message-Id**

  Código único que identifica el mensaje

- **Date**

  Fecha que se recibe el correo electrónico

- **Content-Length:**

  Tamaño del contenido del mensaje

- **Lines**

  Número de líneas del mensaje.

- **Status**

  Estado del correo

La cantidad exacta de la recopilación de todos estos datos se los puede visualizar en la Tabla 5.

Tabla 5

Número de datos recolectados

Total de correos	PhishTank	Monkey.org	Enror.org
23757	14857	3900	5000

Nota. Esta tabla muestra la cantidad exacta de datos recopilados con los que se trabajará para el desarrollo del modelo.

Descripción de los datos

En este punto se describe cada una de las características definidas y su relación con las cabeceras mencionadas en el punto anterior. Se logro agrupar y sintetizar de una manera efectiva todas aquellas características principales que definen que un correo electrónico es Phishing, como se mencionó en el punto anterior estudios previos ayudaron al análisis y determinación de las mismas. (Andronicus A. Akinyelu, 2014) (Sonowal, 2020)

- Asunto vacío
- Detección del carácter @ en las URL del correo electrónico y el asunto.

- Presencia de URL en el texto contenido en el cuerpo del correo.
- Detección de contenido y formularios HTML en el correo electrónico.
- Detección de direcciones IP en URL, en vez de nombre de dominio para enmascarar el nombre del sitio al que redirige el enlace.
- Presencia de código de JavaScript que puede ser incrustado en el cuerpo del correo electrónico o en un enlace.
- Número de puntos en el dominio no debe ser superior a 3
- Presencia de Http en las URL
- Presencia de Https en las URL
- Formato del correo del remitente
- Tamaño de la URL >74
- Uso de CSS en el cuerpo del correo electrónico
- Presencia del evento OnClick
- Grupos de palabras que aparecen regularmente en correos electrónicos infectados con Phishing se utilizó como características. Estas palabras se agruparon en 9 grupos distintos y cada uno con una característica diferente. Los grupos de palabras son:
    1. Actualizar, Confirmar, Enviar
    2. Usuario; Cliente; Estimado; Miembro;
    3. Suspender; Restringir; Sostener.
    4. Verificar; Cuenta; Notificación
    5. Iniciar; Sesión; Username Hacer clic aquí; Contraseña; Login
    6. Ganancia; Sorteo; Felicitaciones; Premio; Ganaste; Gratis
    7. Cuenta; Seguridad;
    8. Importante; Aviso; Verificar; Verificación
    9. Crédito; Banco; En línea, Transferir

Exploración de datos

Una vez identificado los datos recolectados, se procede con el paso de exploración, que se realiza mediante el desarrollo de pruebas estadísticas, que indicarán la validez de los datos adquiridos. Es recomendable en esta fase desarrollar tablas de frecuencias y gráficos de distribución para poder tener una mejor visión de los resultados. Este punto sirve principalmente para determinar si los datos recolectados son completos, tienen consistencia y permitirán cumplir con el objetivo planteado.

La exploración se realizó mediante búsquedas y consultas básicas de los documentos txt y CSV en donde se encuentran los datos obtenidos.

En la Figura 1 se muestra el porcentaje de datos Phishing y no Phishing obtenidos de las fuentes mencionadas PhishTank, Monkey.org, Enron.

Tabla 6

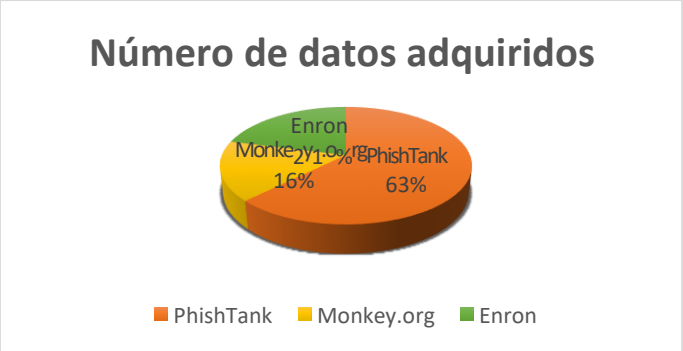
Número de datos Phishing y no Phishing recolectados.

Fuentes de datos	**Número de datos**
PhishTank	14857
Monkey.org	3900
Enron.org	5000
TOTAL	23757

Nota: Esta tabla muestra la cantidad de datos recolectados en cada uno de las fuentes utilizadas en esta investigación.

Figura 15

Porcentaje de datos recolectados



Nota. Este gráfico indica cual es el porcentaje, referente al número de datos recolectados, cuyos valores están en la Tabla 1.

En base a los datos con Phishing recolectados, se determinaron cuáles son los dominios más utilizados para la suplantación de identidad. A continuación, se puede visualizar los datos de las dos fuentes Monkey.org y PhishTank.

Tabla 7

Dominios más utilizados para la suplantación de identidad

Dominios	PhishTank	Monkey.org
Windows	6	60
eBay	388	1
Netflix	93	100
Microsoft	217	124
PayPal	609	1020
YouTube	60	54
Steam	57	0
Facebook	345	25

Dominios	PhishTank	Monkey.org
Instagram	38	9
WhatsApp	75	1
Amazon	119	24
Google	620	231
Halifax	124	0
HSBCgruop	7	0
LinkedIn	40	0
Vodafone	9	0
Apple	99	695
Yahoo!	55	993
Hotmail	12	56
Outlook	52	61
Adobe	84	13
Virus total	34	0
Run escape	1039	0
Twitter	29	17
Bank	1300	227
JPMorgan Chase and Co.	22	0
Gmail	1200	3512
office	259	138
Android	99	393
American express	10	100
Admin	0	159

Nota. Esta tabla indica los dominios más utilizados por los phisher en correos y enlaces.

Figura 16

Número de enlaces con Phishing que utilizan Dominios conocidos en PhishTank.

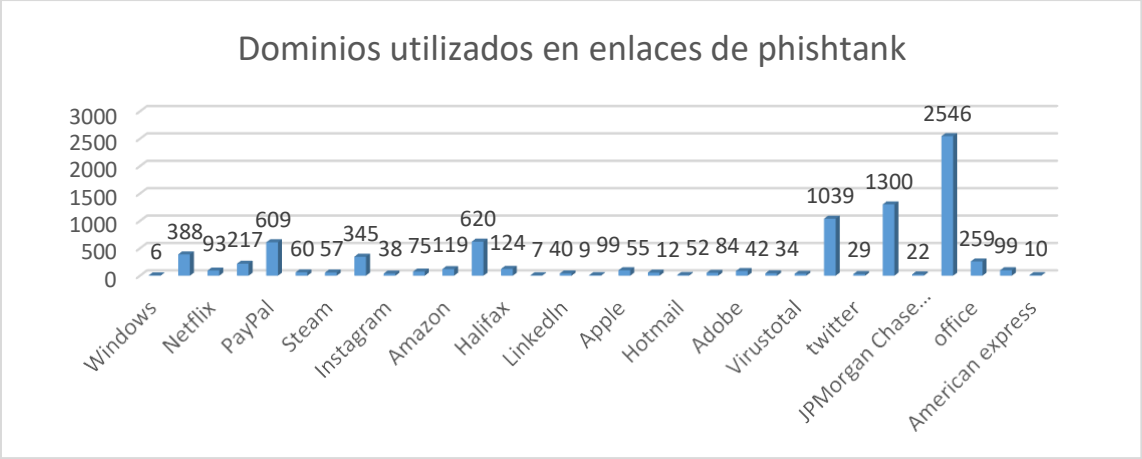


Figura 17

Número de correos con Phishing que utilizan Dominios conocidos en Monkey.org

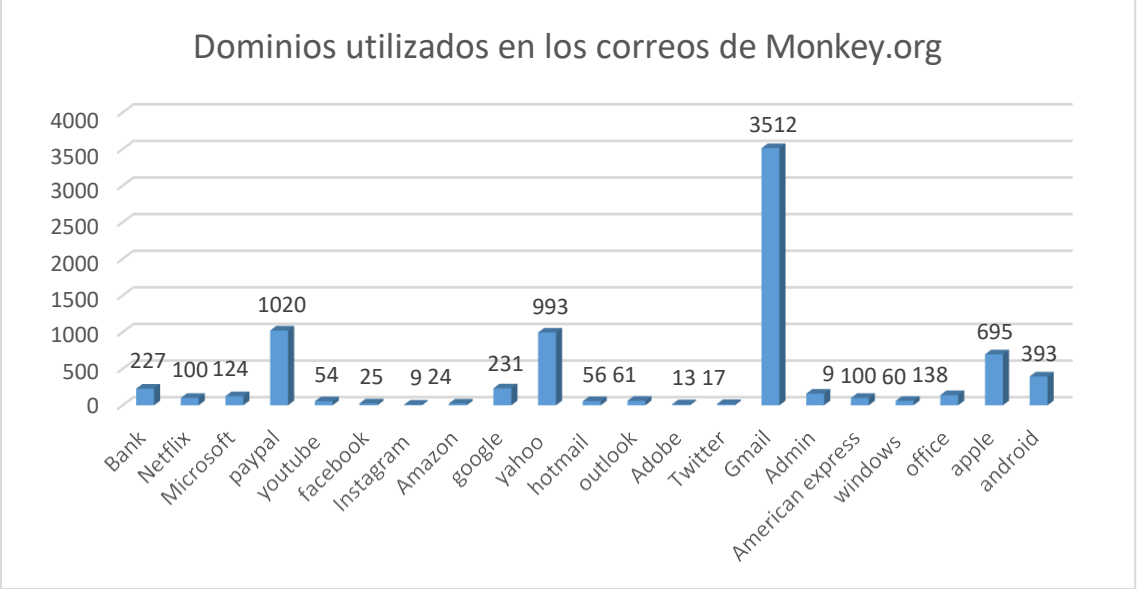
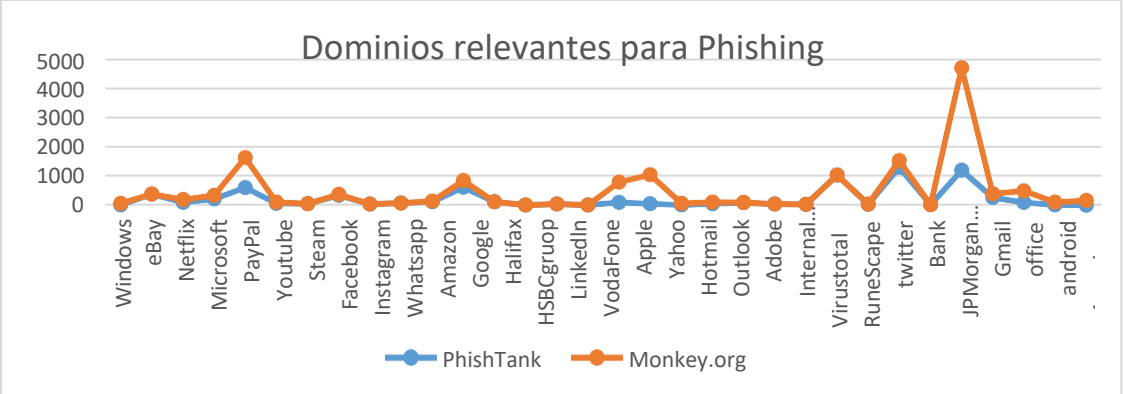


Figura 18

Comparación de datos entre las Fuentes PhishTank yMonkey.org



La exploración de los datos, ayudó a identificar y confirmar cuales son las palabras más frecuentes en correos electrónicos con Phishing, por lo que se consideran como características utilizadas en el código para el modelo de detección. A continuación, se observa las palabras identificadas y la frecuencia de uso en correos con Phishing y no Phishing, lo que permite observar con más claridad la diferencia.

Tabla 8

Número de palabras determinadas en correos electrónicos.

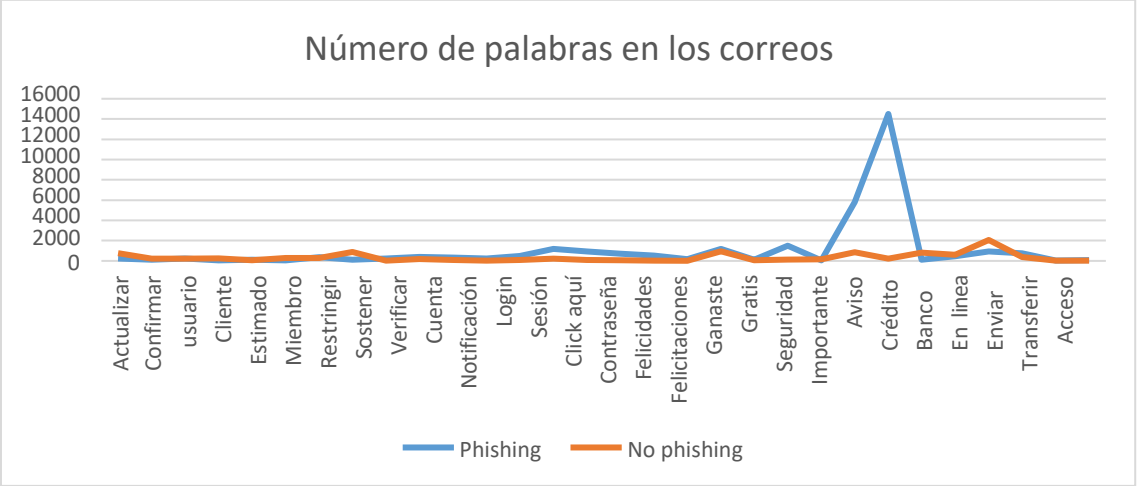
Palabra	Phishing	No Phishing
Actualizar	233	746
Confirmar	121	197
usuario	244	220
Cliente	45	245
Estimado	112	65
Miembro	44	258
Restringir	365	257
Sostener	120	871

Palabra	Phishing	No Phishing
Verificar	242	18
Cuenta	402	179
Notificación	343	80
Login	236	12
Sesión	500	81
Clic aquí	1200	194
Contraseña	930	81
Felicidades	700	38
Felicitaciones	523	22
Ganaste	189	0
Gratis	1200	934
Seguridad	118	40
Importante	1500	125
Aviso	47	135
Crédito	5800	833
Banco	14500	219
En línea	124	818
Enviar	448	593
Transferir	945	2057
Acceso	743	369
Username	38	3
suspender	68	26

Nota. En esta tabla se puede observar el número de veces que aparece, las palabras que se determinaron como características para detección de Phishing.

Figura 19

Número de palabras utilizadas en correos con Phishing y sin Phishing



Luego de haber determinado el número de veces que se repite cada palabra, se obtuvo la frecuencia relativa, para así identificar cual es la diferencia y la influencia de su uso en correos con Phishing y sin Phishing. A continuación, se puede observas una tabla en donde se detalla estas frecuencias obtenidas.

Tabla 9

Número de frecuencia de la palabra en los correos

PALABRAS	PHISHING	NO PHISHING
Palabra	5,97	19,13
Actualizar	3,10	5,05
Confirmar	6,26	5,64
usuario	1,15	6,28
Cliente	2,87	1,67
Estimado	1,13	6,62
Miembro	9,36	6,59
Restringir	3,08	22,33
Sostener	6,21	0,46
Verificar	10,31	4,59
Cuenta	8,79	2,05
Notificación	6,05	0,31
Login	12,82	2,08
Sesión	30,77	4,97
Clic aquí	23,85	2,08
Contraseña	17,95	0,97
Felicidades	13,41	0,56
Felicitaciones	4,85	0,00
Ganaste	30,77	23,95
Gratis	3,03	1,03
Seguridad	38,46	3,21

PALABRAS	PHISHING	NO PHISHING
Importante	1,21	3,46
Aviso	148,72	21,36
Crédito	371,79	5,62
Banco	3,18	20,97
En línea	11,49	15,21
Enviar	24,23	52,74
Transferir	19,05	9,46
Acceso	0,97	0,08
Username	1,74	0,67
suspender	1,74	0,67

Nota. Para obtener la frecuencia relativa se hizo en base al total de correos con Phishing y no Phishing que se obtuvo de la fuente de datos Monkey.org con 3900 correos con Phishing y Enron con 5000 correos sin Phishing.

Figura 20

Palabras determinadas en correos con Phishing y sin Phishing.

Nota. Este gráfico muestra cual es a frecuencia de uso de las palabras, donde se puede observar que en su mayoría se las encuentra en correos con Phishing.

Verificar la calidad de los datos

Después de hacer la exploración de los datos, se puede llegar a la conclusión que las fuentes de datos obtenidas son completas. Proporcionan la información necesaria para la obtención de características específicas, las misma que ayudarán en la generación del modelo con una alta precisión.

En cuanto a los correos electrónicos con Phishing, vienen de una fuente confiable donde varios estudios lo utilizan, estos datos se encuentran actualizados, lo que asegura que se puede obtener resultados nuevos en este modelo a generarse. En esta base de datos existen algunos correos que no se poseen información completa, cuya solución es eliminar mediante el proceso de minería de datos, ya que estos datos no van aportar, ni tampoco influye el no tomarlos en cuenta.

La base de datos con correos sin Phishing es de gran ayuda, al momento de realizar el aprendizaje del modelo, así se podrá dar mayor validez a la identificación de correos electrónicos con Phishing, porque proporciona mayores características para discernir entre un correo infectado y otro no.

La base de datos de PhishTank es una fuente confiable, cuya información ayudará mucho en el momento del análisis de URL y dominios en un correo.

Todos los datos obtenidos, entregan información importante y precisa, lo que asegura que no hay riesgo de ruido en el proceso, y que los resultados serán positivos y reales.

Tercera Fase Preparación de los Datos

Aquí se procede a preparar los datos para adecuarlos de manera correcta a las técnicas seleccionadas de minería de datos seleccionadas. Esta fase es una de las más

importantes y con frecuencia la que más tiempo lleva, ya que es necesario empaquetar y seleccionar correctamente los datos que van a ser procesados en la minería de datos. Se procedió a la limpieza de dos bases de datos obtenidas que son, private-Phishing4.mbox y emails-Enron. mbox, que contienen la información correspondiente a correos electrónicos infectados y no infectados con Phishing respectivamente, de estas bases de datos se seleccionó un subconjunto de datos para limpiarlos y mejorar su calidad que permite la extracción de características en el formato requerido para que las herramientas que intervienen en el modelado las puedan procesar.

Seleccionar los datos

El formato que contiene la información es del tipo mbox, formato utilizado para contener correos electrónicos en una estructura de cabeceras para lo cual cada correo electrónico cuenta con cabeceras específicas a manera de campos y los registros son representados por el número de correos que existen, algunos de los campos o cabeceras no representan mayor información por lo que fueron irrelevantes para el proceso extracción de las características deseadas, razón por la que se decidió no tomarlas en cuenta en el proceso de minería de datos.

Se seleccionó los siguientes datos para realizar el análisis:

- **header['From']**
- **header['Subject']**
- **header [Content-Type]**
- **header['Date']**
- **header['Body']**

Se considera que los datos seleccionados contienen la información necesaria para el correcto formateo y extracción de características descritas en la fase 2.

Limpiar los datos

Los archivos con los que se cuenta como fuente de datos son explícitamente contenedores de correos electrónicos infectados con Phishing y correos no infectados, razón por la que no precisa de una discriminación pues se realiza un procesamiento por separado, ya que, a pesar de ser dos archivos totalmente diferentes en sus registros, sus campos son semejantes y el proceso de extracción de características es igual en ambos casos.

Aun así se necesita realizar una limpieza d datos ya que la información contenida en los datos seleccionados si cuentan con información no deseada en su estructura o valores nulos en algunos de los campos, dichos valores nulos no requieren un proceso de estimación de valor ya que en el proceso de Data Mining la ausencia de estos valores brindaran un tipo de característica relevante en la detección, la información no deseada si requiere de un tratamiento para lo que se realizó una función dentro del lenguaje de alto nivel Python que permita eliminarlos sin alterar los datos en su naturaleza, solo en su estructura.

Los datos que se desea eliminar es:

- Salto de línea (\n)
- Tabulaciones (\t)
- Espacios en blanco (\s)
- Continuación de línea (=\s)

Esta información se la elimino a base de la siguiente función que está en la **Figura 21**

Figura 21

Función para la limpieza de datos de las fuentes obtenidas.

```
def getURLs_string(string):
result = []
cleanPayload = re.sub(r'\s+', ' ', str(string))
cleanPayload1 = re.sub(r'\n+','', cleanPayload)
cleanPayload2 = re.sub(r'\t+','', cleanPayload1)
cleanPayload3 = re.sub(r'=\s+','', cleanPayload2)
linkregex = re.compile(HREFREGEX, re.IGNORECASE)
links = linkregex.findall(cleanPayload3)
for link in links:
if isurl(link):
result.append(link)
urlregex = re.compile(URLREGEX_NOT_ALONE, re.IGNORECASE)
links = urlregex.findall(cleanPayload3)
for link in links:
if link not in result:
result.append(link)
return links
def isurl(link):
return re.compile(URLREGEX, re.IGNORECASE).search(link) is not None
```

Para este proceso se utilizó la librería re - Regular expression operation de Python, que permite utilizar funciones nativas y eliminar porciones específicas del contenido sin alterar la integridad de la información, así se logra realizar el proceso de limpieza en tres fases toda la información que se obtuvo.

En cuanto a los datos de PhishTank, solo se realizó la limpieza de aquellos que tenían datos incompletos, o en la columna no identificaba si es o no Phishing la URL.

En la **Tabla 10** se puede visualizar el número de datos finales después de la limpieza.

Tabla 10

Número de datos obtenidos

Fuentes de datos	Número de datos antes de la limpieza	Número de datos después de la limpieza
Monkey.org	3900	2000
Enron	5000	2000
PhishTank	14857	13368

Construir los datos

Selección de características

Como se describió en la primera fase de este modelo, las características seleccionadas se las pudo determinar mediante el análisis y síntesis de estudios previos detallados en el **CAPITULO I.** Estas características fueron agrupadas y sintetizadas de tal manera que se pueda determinar de manera eficaz que un correo es Phishing. Cabe recalcar que este proceso es uno de los principales para la construcción de los datos del modelo, debido a que, en base a las características seleccionadas, se define la información y partes necesarias de un correo y poder realizar un correcto proceso de minería de datos. En la **Tabla 11** se puede visualizar las características seleccionadas.

Tabla 11

Características seleccionadas

Características seleccionadas	**Código**
Asunto vacío	TJO1
Detección del carácter @ en las URL del correo electrónico y el asunto.	TJO2
Presencia de URL en el cuerpo del correo.	TJO3
Detección de contenido y formularios HTML en el correo electrónico.	TJO4
Detección de direcciones IP en URL, en vez de nombre de dominio para enmascarar el nombre del sitio al que redirige el enlace.	TJO5
Presencia de código de JavaScript que puede ser incrustado en el cuerpo del correo electrónico o en un enlace.	TJO6
Número de puntos en el dominio no debe ser superior a 3	TJO7
Presencia de Http en las URL	TJO8
Presencia de Https en las URL	TJO9
Formato del correo del remitente	TJO10
Tamaño de la URL >74	TJO11
Uso de CSS en el cuerpo del correo electrónico	TJO12
Presencia del evento OnClick	TJO13

Características seleccionadas	Código
Grupos de palabras que aparecen regularmente en correos electrónicos infectados con Phishing se utilizó como características. Estas palabras se agruparon en 9 grupos distintos y cada uno con una característica diferente. Los grupos de palabras son:	
1. Dear member, Dear customer, Dear client, Dear user	TJO14
2. Actualizar, Confirmar, Enviar	TJO15
3. Usuario; Cliente; Estimado; Miembro;	TJO16
4. Suspender; Restringir; Sostener.	TJO17
5. Verificar; Cuenta; Notificación	TJO18
6. Iniciar; Sesión; Username; Hacer clic aquí; Login	TJO19
7. Felicitaciones; Ganaste; Gratis; Felicidades	TJO20
8. Seguridad; Contraseña	TJO21
9. Importante; Aviso; Verificar; Verificación	TJO22
10. Crédito; Banco; En línea, Transferir	TJO23
Fechas vacías	TJO24

Atributos derivados

Para la construcción de datos se utilizó el lenguaje Python, que ayudó a realizar el proceso correcto de limpieza de todos los datos , para que seguidamente se pueda separar cada parte del correo electrónico contenido en los archivos mbox, para esto es crucial un dato seleccionado con anterioridad que es header['Content-Type'], ya que este contiene el

formato necesario para la extracción del cuerpo del correo electrónico como se puede ver en la **Figura 22**

Figura 22

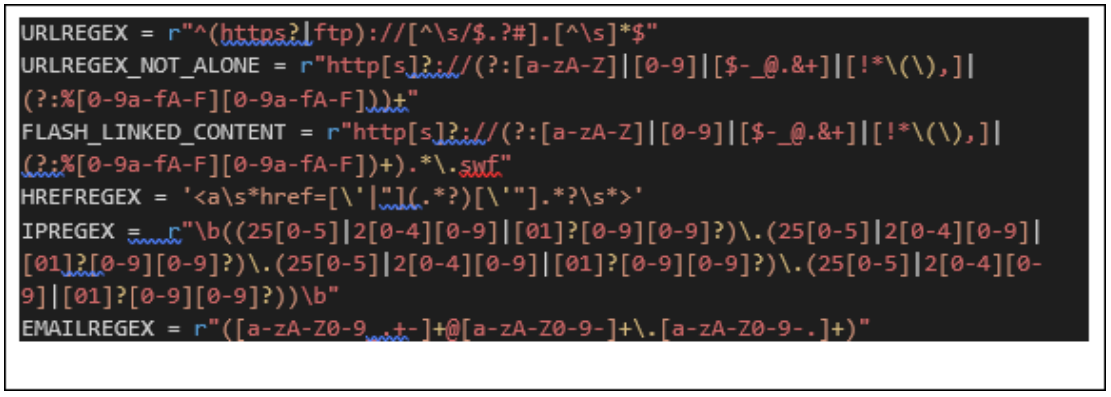
Código para la extracción del cuerpo de los correos electrónicos.

```
def get_body(tmsg):

    body = ""
    if tmsg.is_multipart():
        for part in tmsg.walk():
            ctype = part.get_content_type()
            cdispo = str(part.get('Content-Disposition'))
            # skip any text/plain (txt) attachments
            if ctype == 'text/html' and 'attachment' not in cdispo:
                body = str(part)
                break
            if ctype == 'text/plain' and 'attachment' not in cdispo:
                body = str(part.get_payload(decode=True))  # decode
                body=body.replace("\\r\\n","\n")
                break
    # not multipart - i.e. plain text, no attachments, keeping fingers
crossed
    else:
        print('no is multipart')
        body = str(tmsg.get_payload(decode=True))

    return body.upper()
```

Los siguientes datos que son necesarios extraer se los caracterizó bajo el formato de librería re - Regular expression operation, permitiendo de esta manera identificar dentro del texto del correo electrónico patrones que hacen referencia a una dirección URL y una dirección IP dentro de las URL.

Figura 23

Código para identificar URL y las IP dentro del texto de los correos electrónicos



```
URLREGEX = r"^(https?|ftp)://[^\s/$.?#].[^\s]*$"
URLREGEX_NOT_ALONE = r"http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[!*\(\),]|
(?:%[0-9a-fA-F][0-9a-fA-F]))+"
FLASH_LINKED_CONTENT = r"http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[!*\(\),]|
(?:%[0-9a-fA-F][0-9a-fA-F])+).*\.swf"
HREFREGEX = '<a\s*href=[\'|"](.*?)[\'"].*?\s*>'
IPREGEX = r"\b((25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.(25[0-5]|2[0-4][0-9]|
[01]?[0-9][0-9]?)\.(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.(25[0-5]|2[0-4][0-
9]|[01]?[0-9][0-9]?))\b"
EMAILREGEX = r"([a-zA-Z0-9_+-]+@[a-zA-Z0-9-]+\.[a-zA-Z0-9-.]+)"
```

Atributos generados

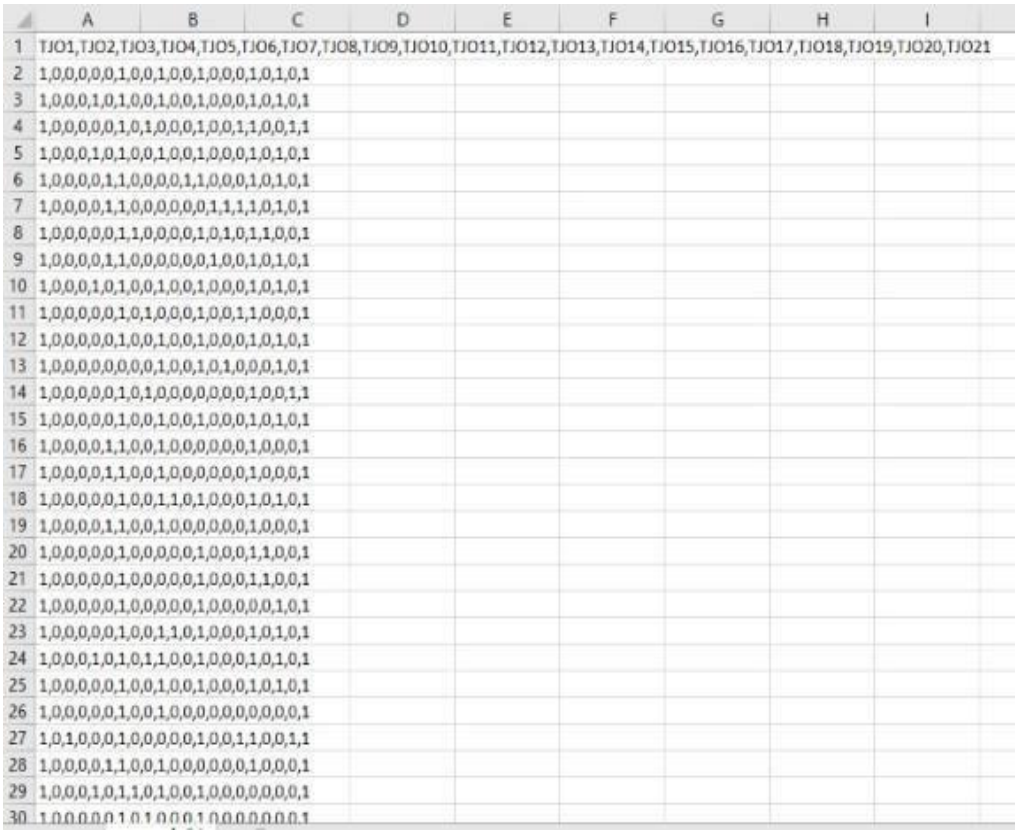
No fue necesaria la generación de atributos para el proceso requerido de minería de datos ya que toda la información relevante en la extracción de las características estas presente en todos los campos elegidos y la ausencia o presencia de alguno de ellos representa por igual alguna característica.

Integrar datos

Las características extraídas de cada uno de los correos electrónicos se las consolida dentro de una matriz en la que cada fila representa el registro y cada columna un campo característico que se extrajo, esta matriz al final del proceso de análisis y determinación es guardada dentro de un archivo CSV para su posterior análisis en los algoritmos de predicción utilizados en Data Mining, adicional de las columnas de características se coloca una extra donde se observa el nombre del archivo del cual se extrajo las características.

Figura 24

Matriz de características .CSV



	A	B	C	D	E	F	G	H	I
1	TJO1,TJO2,TJO3,TJO4,TJO5,TJO6,TJO7,TJO8,TJO9,TJO10,TJO11,TJO12,TJO13,TJO14,TJO15,TJO16,TJO17,TJO18,TJO19,TJO20,TJO21								
2	1,0,0,0,0,0,1,0,0,1,0,0,0,1,0,1,0,1								
3	1,0,0,0,1,0,1,0,0,1,0,0,0,1,0,1,0,1								
4	1,0,0,0,0,0,1,0,1,0,0,0,1,0,0,1,1,0								
5	1,0,0,0,1,0,1,0,0,1,0,0,0,1,0,1,0,1								
6	1,0,0,0,0,1,1,0,0,0,0,1,1,0,0,0,1,0								
7	1,0,0,0,0,1,1,0,0,0,0,0,0,1,1,1,1,0								
8	1,0,0,0,0,0,1,1,0,0,0,0,1,0,1,0,1,1								
9	1,0,0,0,0,1,1,0,0,0,0,0,0,1,0,0,1,0								
10	1,0,0,0,1,0,1,0,0,1,0,0,0,1,0,1,0,1								
11	1,0,0,0,0,0,1,0,1,0,0,0,1,0,0,1,1,0								
12	1,0,0,0,0,0,1,0,0,1,0,0,0,1,0,1,0,1								
13	1,0,0,0,0,0,0,0,0,1,0,0,1,0,1,0,0,1								
14	1,0,0,0,0,0,1,0,1,0,0,0,0,0,0,0,1,0								
15	1,0,0,0,0,0,1,0,0,1,0,0,0,1,0,1,0,1								
16	1,0,0,0,0,1,1,0,0,1,0,0,0,0,0,0,1,0								
17	1,0,0,0,0,1,1,0,0,1,0,0,0,0,0,0,1,0								
18	1,0,0,0,0,0,1,0,0,1,1,0,1,0,0,0,1,0								
19	1,0,0,0,0,1,1,0,0,1,0,0,0,0,0,0,1,0								
20	1,0,0,0,0,0,1,0,0,0,0,0,1,0,0,1,1,0								
21	1,0,0,0,0,0,1,0,0,0,0,0,1,0,0,1,1,0								
22	1,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,1								
23	1,0,0,0,0,0,1,0,0,1,1,0,1,0,0,0,1,0								
24	1,0,0,0,1,0,1,0,1,1,0,0,1,0,0,0,1,0								
25	1,0,0,0,0,0,1,0,0,1,0,0,0,1,0,1,0,1								
26	1,0,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,1								
27	1,0,1,0,0,0,1,0,0,0,0,0,1,0,0,1,1,0								
28	1,0,0,0,0,1,1,0,0,1,0,0,0,0,0,1,0,0								
29	1,0,0,0,1,0,1,1,0,1,0,0,1,0,0,0,0,0								
30	1,0,0,0,0,0,1,0,1,0,0,0,1,0,0,0,0,0								

Formateo de los datos

Una vez se tiene una matriz consolidada con las características que determinen o no la presencia de Phishing en un correo electrónico es necesario identificarlas de alguna manera, para ello se utiliza la columna donde se encuentra el nombre del archivo de origen y se procede con el reemplazo de esta información con el siguiente formato, los registros que contienen el nombre del archivo private-Phishing4.mbox cuyo contenido hace referencia a correos infectados se lo reemplazara con el número 1 y a los registros que contienen el nombre del archivo emails-enron.mbox que hace referencia a los correos electrónicos sin

presencia de Phishing se los reemplazara con el número 0 y de esta manera se tiene la información requerida para los modelos de predicción.

Cuarta Fase Modelado

En la presente fase se procederá a elegir cuál de los modelos de minería de datos es la indicada para el cumplimiento del objetivo planteado en la fase 1, para esto se realizó un plan de pruebas en base a métricas que definirán el porcentaje de precisión de cada modelo, por último, se evaluará si este modelo cumple con los criterios de éxito establecidos.

Escoger la técnica de modelado

Para la selección de la técnica adecuada se basó en aquellas cuya función contribuyan a cumplir los objetivos de minería de datos planteados en la fase 1, además de que se encuentren disponibles para el lenguaje Python en específico la librería “sklearn”, la que brinda funciones relacionadas con la de minería de datos. Debido a que la limpieza y construcción de los datos se realizó en este lenguaje se facilita la construcción del modelo.

De las técnicas definidas en la fase 1 y disponibles en la librería se encuentra que la técnica de Naive Bayes, Arboles de decisión y Random Forest son las que se ajustan a resolver la problemática de predicción de si un correo contiene o no Phishing en base a sus múltiples características. Estas técnicas no se ajustan a métodos lineales de predicción también disponibles, por lo que se descartaron y se seleccionaron las ya mencionadas.

Generar el plan de prueba

El plan de pruebas inicia con la matriz de datos que contiene la información de 4789 correos electrónicos que contienen o no características de Phishing, esta será separada bajo la teoría de Pareto en una proporción de 80% de los datos para el entrenamiento y un 20% utilizado para las pruebas, para esto se utiliza la función “train_test_split” que ayuda a separar en dicho porcentaje y de manera aleatoria los datos que contienen o no Phishing,

cabe mencionar que para el entrenamiento es necesario que la información sea de mayor tamaño ya que abarca la mayor cantidad de características relacionadas a lo que el objetivo deseas alcanzar.

Para cada técnica se probará su nivel de eficacia en basa a las 4 métricas siguientes: Puntaje de clasificación de precisión (Accuracy score), La precisión calculada (Precision score), La recuperación calculada (Recall score) y el Puntaje F1 (F1 score), en base a estas medidas se puede precisar el éxito obtenido al predecir si un correo contiene o no Phishing. (IArtificial.net, 2020)

Para el cálculo de estas métricas mencionadas se necesita de los siguientes valores:

Verdadero positivo (VP) es el número de casos que la prueba declaro como positiva y que son verdaderamente positivas.

Falsos positivos (FP) es el número de casos que la prueba declara como positivas y en realidad son negativas.

Verdadero negativo (VN) número de casos que la prueba declara como negativo y que en realidad es negativa.

Falso Negativo (FN) el número de casos que la prueba declara como negativo y que en realidad son positivas.

- Puntaje de clasificación de precisión (Accuracy score) esta métrica representa directamente la exactitud que tuvo el modelo al predecir correctamente en contraste al total de las predicciones. La exactitud se la calcula de la siguiente manera. (scikit learn, 2020)

$$Accuracy\ score = \frac{VP + VN}{VP + VN + FP + FN}$$

- La precisión calculada (Precision score) esta métrica representa la relación que existe entre los casos calculados como verdaderos positivos y el total de casos positivos encontrados en la prueba. (scikit learn, 2020)

$$Precision\ score = \frac{VP}{VP + VN}$$

- La recuperación calculada (Recall score) esta métrica representa qué relación existe entre los casos que fueron observados como positivos verdaderamente con la suma de estos mismos y los que fueron positivos, pero fueron observados como negativos. (scikit learn, 2020)

$$Recall\ score = \frac{VP}{VP + FN}$$

- El Puntaje F1 (F1 score) el valor F1 es la ponderación y relación entre Precision Score y Recall Score danto como resultado el contemplar tanto valores falsos positivos como falsos negativos que se obtuvieron en la observación, este valor se utiliza en casos donde difieran la distribución de casos y es realmente útil en el presente proyecto ya que la división de datos es randómico. (scikit learn, 2020)

$$F1\ score = \frac{PRECISION\ SCORE * RECALL\ SCORE}{PRECISION\ SCORE + RECALL\ SCORE} * 2$$

Construir el modelo

En la construcción del modelo se procede con la selección de los datos que formarán parte de la entrada para las técnicas seleccionadas para cumplir el objetivo propuesto en la fase 1, los datos se dividirán como fue mencionado en 80% para entrenamiento y 20% de

datos para pruebas y evaluación del rendimiento de estas.

Ya que el objetivo a cumplir es único y las entradas para las tres técnicas son semejantes únicamente se necesita de un ajuste de parámetros para que estén listos y sean procesados por estas técnicas, el ajuste y la separación de los datos se la realiza de la siguiente manera.

Se separa los datos en características en una sola matriz llamada “X” o variable independiente y la columna que identifica si el correo posee o no Phishing será colocado en un vector llamado “Y” o variable dependiente, una vez separadas las variables se obtiene lo siguiente.

- X_train: Variable que contiene el 80% de los datos que caracterizan un correo que contiene o no Phishing.
- Y_train: Variable que contiene el 80% de los datos que marcan si un correo es portador o no de Phishing
- X_test: Variable que contiene el 20% de los datos que caracterizan un correo que contiene o no Phishing.
- Y_test: Variable que contiene el 20% de los datos que marcan si un correo es portador o no de Phishing

Una vez que los datos están ajustados de tal manera que están listos para ser procesados por las diferentes técnicas de Data Mining elegidas se procede a realizar dicho análisis de la siguiente manera.

Naive Bayes

En la técnica de Naive Bayes se emplea las variables X_train y Y_train como entradas para que esta las utilice de entrenamiento como lo muestra en la **Figura 25.**

Figura 25

Aplicación de la técnica de Naive Bayes

```
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=1)
print('Number of rows in the total set: {}'.format(pima.shape[0]))
print('Number of rows in the training set: {}'.format(X_train.shape[0]))
print('Number of rows in the test set: {}'.format(X_test.shape[0]))

naive_bayes = MultinomialNB()

naive_bayes.fit(X_train, y_train)
```

Una vez entrenada la red se procede a realizar las pruebas con la variable X_test como entrada como lo muestra la **Figura 26.**

Figura 26

Código para realizar pruebas de la técnica Naive Bayes

```
y_pred = naive_bayes.predict(X_test)
```

Cuando se obtiene los datos de predicción por parte de la técnica Naive Bayes se la almacena en la variable y_pred a manera de vector, este contiene la información de la predicción realizada que contrastada con la variable Y_test se podrá evaluar si pertenece al grupo de VP, VN, FP o FN y de esta manera se obtienen los resultados como lo muestra la Figura 27.

Figura 27

Resultados de los parámetros de prueba Naive Bayes.

```
print('Accuracy score: ', format(accuracy_score(y_test, y_pred)))
print('Precision score: ', format(precision_score(y_test, y_pred)))
print('Recall score: ', format(recall_score(y_test, y_pred)))
print('F1 score: ', format(f1_score(y_test, y_pred)))
```

Arboles de decisión

En la técnica de Naive Bayes se emplea las variables X_train y Y_train como entradas para que esta las utilice de entrenamiento como lo muestra la **Figura 27**.

Figura 28

Código para la aplicación de árboles de decisión

```
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=1)
print('Number of rows in the total set: {}'.format(pima.shape[0]))
print('Number of rows in the training set: {}'.format(X_train.shape[0]))
print('Number of rows in the test set: {}'.format(X_test.shape[0]))

naive_bayes = MultinomialNB()

naive_bayes.fit(X_train, y_train)
```

Una vez entrenada la red se procede a realizar las pruebas con la variable X_test como entrada como lo muestra la **Figura 29**.

Figura 29

Código para realizar pruebas de la técnica Árboles de decisión

```
y_pred = naive_bayes.predict(X_test)
```

Cuando se obtiene los datos de predicción por parte de la técnica Naive Bayes se la almacena en la variable y_pred a manera de vector, este contiene la información de la predicción realizada que contrastada con la variable Y_test se podrá evaluar si pertenece al grupo de VP, VN, FP o FN y de esta manera se obtienen los resultados como lo muestra la **Figura 30**.

Figura 30

Resultados de los parámetros de prueba de Árboles de decisión

```
print('Accuracy score: ', format(accuracy_score(y_test, y_pred)))
print('Precision score: ', format(precision_score(y_test, y_pred)))
print('Recall score: ', format(recall_score(y_test, y_pred)))
print('F1 score: ', format(f1_score(y_test, y_pred)))
```

Random Forest

En la técnica de Naive Bayes se emplea las variables X_train y Y_train como entradas para que esta las utilice de entrenamiento como lo muestra la **Figura 31.**

Figura 31

Código para la aplicación de Random Forest

```
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=1)
print('Number of rows in the total set: {}'.format(pima.shape[0]))
print('Number of rows in the training set: {}'.format(X_train.shape[0]))
print('Number of rows in the test set: {}'.format(X_test.shape[0]))

naive_bayes = MultinomialNB()

naive_bayes.fit(X_train, y_train)
```

Una vez entrenada la red se procede a realizar las pruebas con la variable X_test como entrada como lo muestra la Figura 30.

Figura 32

Código para realizar pruebas de la técnica Random Forest

```
y_pred = naive_bayes.predict(X_test)
```

Cuando se obtiene los datos de predicción por parte de la técnica Naive Bayes se la almacena en la variable y_pred a manera de vector, este contiene la información de la predicción realizada que contrastada con la variable Y_test se podrá evaluar si pertenece al grupo de VP, VN, FP o FN y de esta manera se obtienen los resultados como lo muestra la **Figura 33**.

Figura 33

Resultados de los parámetros de prueba de Random Forest

```
print('Accuracy score: ', format(accuracy_score(y_test, y_pred)))
print('Precision score: ', format(precision_score(y_test, y_pred)))
print('Recall score: ', format(recall_score(y_test, y_pred)))
print('F1 score: ', format(f1_score(y_test, y_pred)))
```

Descripción del modelo

En esta sección, se detalla los resultados de cada métrica depués de la ejecución de cada modelo detallado, estos resultados permitirán seleccionar el modelo más óptimo para su implementación.

- **Modelo 1: Árboles de decisión**

  La ejecución de este modelo ha devuelto los siguientes resultados:

  Puntaje de clasificación de precisión (Accuracy score) tiene un valor de 97,15%.

  Precisión calculada (Precision score) tiene un valor de 96,57%.

  Recuperación calculada (Recall score) tiene un valor de 96,96%.

  Puntuación F1 (F1 score) tiene un calor de 96,77%.

- **Modelo 2: Naive Bayes**

  La ejecución de este modelo ha devuelto los siguientes resultados:

Puntaje de clasificación de precisión (Accuracy score) tiene un valor de 91,91%.

Precisión calculada (Precision score) tiene un valor de 90.4%.

Recuperación calculada (Recall score) tiene un valor de 91,31%.

Puntuación F1 (F1 score) tiene un calor de 90,85%.

- **Modelo 3: Random Foreste**

  La ejecución de este modelo ha devuelto los siguientes resultados:

  Puntaje de clasificación de precisión (Accuracy score) tiene un valor de 97.6%.

  Precisión calculada (Precision score) tiene un valor de 96,8%.

  Recuperación calculada (Recall score) tiene un valor de 97,77%.

  Puntuación F1 (F1 score) tiene un calor de 97,28%.

Evaluar el modelo

Para evaluar el modelo se utiliza las métricas planteadas en el plan de pruebas del presente documento, estas métricas son Accuracy score, Precision score, Recall score y F1 score, estos valores son proporcionados por el lenguaje de alto nivel Python y en específico la librería sklearn, que al realizar la comparación entre los valores reales y los valores predichos, estima los porcentajes en base a las fórmulas planteadas en el plan de pruebas para cada una de las métricas que se utilizara para la evaluación de los modelo planteado. Estos valores se encuentran reflejados en el punto 4.3 del presente documento.

El primero modelo evaluado será el basado en el algoritmo de Arboles de decisión, con un valor de predicción presentado de la siguiente manera, se obtiene en la primera

métrica Accuracy score un 97,15% de precisión, esta métrica como se mencionó indica el nivel predicciones correcta que se realizó en base al modelo y por esta razón genera indicios que el dataset generado tiene un nivel de confianza alto, a su vez el Precision score arroja un 96,75% de precisión al seleccionar correos con características de phishing, se observa que el porcentaje es inferior ya que este hace alusión únicamente a los mensajes que fueron predichos de manera correcta, la métrica Recall score con un 96,96% indica que el modelo puede identificar en un alto porcentaje los criterios expuestos de lo que es o no es phishing, por último el F1 score que conjuga los criterios de precisión y recuperación arroja un 96,77% indicando que el modelo en general tiene este rendimiento y se tomara en cuenta para la selección del modelo más adecuado que se implementara.

El segundo modelo a evaluar es el basado en el algoritmo de Naive Bayes con valores en las métricas que se presentaran de la siguiente manera, a primera vista los valores se notan en decremento teniendo que el Accuracy score es del 91,91% indicando que el modelo no cumple en primera instancia con el objetivo planteado ya que si se compara los valores predichos correctamente por el modelo es relativamente bajo al valor que se desea obtener por consiguiente los valores de Precision score y Recall score también se ven afectados ya que presentan un 90,4% y 91,31% respectivamente y esto confirma que los valores predichos en el modelo son relativamente bajos, si se analiza los resultados presentados se observa que si bien es cierto que el modelo predice en alto porcentaje, su nivel de fallos también es alto y esto se puede visualizar mejor en el F1 score que presenta un valor de 90,98% que es mucho más bajo que el modelo de Arboles de decisión y que del planteado como objetivo a alcanzar.

El tercer y último modelo basado en el algoritmo de Random Forest que será evaluado es el que presenta mayor porcentaje de precisión en cada una de las métricas con un Accuracy score de 97,6% comparado al 97,15% de árboles de decisión es 0,45% más

efectivo al predecir el total de correos con Phishing, casos similares se presentan en Precision score y Recall score con valores de 96,8% y 97,77% que comparados con los 97,57% y 96,96, presentan una mínima diferencia, que en primer lugar afianza los resultados del modelo de Arboles de decisión y mejora en cierto porcentaje dichos resultados, lo que ayuda a la generación de un modelo confiable para la fase de implementación.

A continuación, se presenta la tabla que contiene en resumen cada uno de los valores descritos para cada una de las métricas y su modelo respectivo.

Tabla 12

Porcentajes de métricas definidas

	Accuracy score	**Precision score**	**Recall score**	**F1 score**
Modelo Arboles de decisión	97,15%	96,57%	96,96%	96,77%
Modelo Naive Bayes	91,91%	90,4%	91,31%	90,85%
Modelo Random Forest	97,6%	96,7%	97,77%	97,28%

Quinta fase Evaluación

Esta fase se encuentra desarrollada en el capítulo 5 de este documento.

Sexta fase implantación

El objetivo de esta última fase de la metodología CRISP-DM es explicar correctamente el funcionamiento del proyecto construido y como se pone en marcha, además de mostrar los resultados obtenidos, que determinan la validez y cumplimiento de los objetivos descritos en la primera fase del documento.

Planear la Implantación

Para poder realizar la implantación correcta de este proyecto en un ambiente real, no se necesita de recursos extras, mas solo es necesario conectarlo al Gmail que se quiere

mitigar y número de teléfono donde se desea recibir los mensajes de WhatsApp para la notificación en el caso de que un correo recibido, tenga contenido Phishing.

Los pasos a seguir para el funcionamiento del modelo son los mismo que se detallan y se ha seguido en este documento, desde la primera fase del modelo que es la comprensión hasta la implantación. Al momento de aplicar este modelo en un ambiente no controlado, existe la posibilidad que algunas fases lleven más tiempo desarrollarlas, debido al ruido que puede existir al manejar gran cantidad de datos.

A continuación, se visualiza los diagramas de flujos de cada proceso y subproceso que permiten tener una visión más clara del funcionamiento de este proyecto.

En primera instancia se tiene 3 subprocesos muy importantes y esenciales que son: Lectura de correos no Phishing, Lectura de correos Phishing, lectura de correos. Permitiendo realizar la generación del datase con la extracción de características.

- **Primer subproceso: Lectura de correos Phishing**

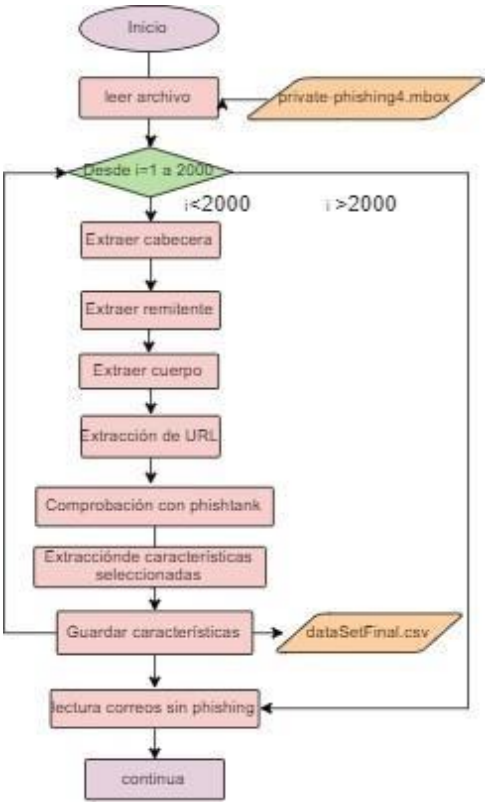
  En la Figura 34 se observa este subproceso, el algoritmo lee el archivo private-phshing4. mbox donde se encuentran correos sin Phishing, seguidamente se entra a un ciclo repetitivito For que finaliza el momento que se haya analizado los 2000 correos contenido en el archivo. Dentro de este ciclo For se realiza varias funciones para el tratamiento de cada correo y extracción de características. Se empieza con la obtención de información relevante como es la extracción de la cabecera, remitente, cuerpo del correo y URL, siendo esta última la primera en analizarse, mediante la verificación de su existencia en la base de datos del repositorio de PhishTank, a través de una API de servicio restfull, donde se envía cada una de las URL contenidas en el correo, devolviendo como resultado un archivo xml , el cual minería de dato determina si la URL existe dentro de la base de datos de PhishTank y si está definida como Phishing.

Seguido el algoritmo analiza los datos contenidos en cada correo electrónico, para posteriormente proceder con la extracción de características, cabe mencionar que cada característica será representada por un valor binario, siendo 1 detección de Phishing y 0 no Phishing, todos estos datos se guardan en el archivo datasetTesisFinal.csv por cada correo analizado.

Una vez finalizado este ciclo se continúa con el siguiente subproceso que es similar a este.

Figura 34

Lectura de archivo Phishing



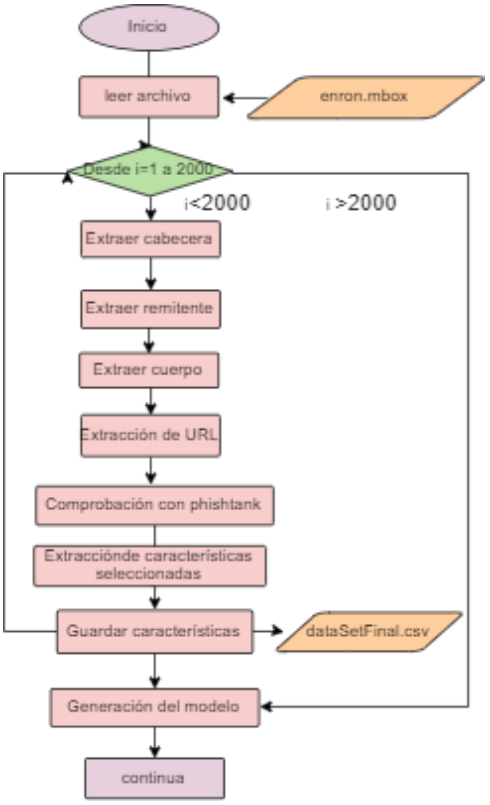
- **Segundo subproceso: Lectura de correos no Phishing**

  Los pasos de este subproceso como se puede ver en la figura 35 son exactamente igual, al descrito antes con la única diferencia que aquí se lee el archivo enron.mbox que contiene 2000 correos sin Phishing. Es importante mencionar que el dataset que se genera, crea 25 columnas de las cuales las primeras 24 representan a las características definidas, y la columna 25 identifica que correos son Phishing y cuáles no, representadas con un “1” o “0” respectivamente.

  Más adelante se explica la importancia de este punto mencionado.

Figura 35

Lectura de archivo Phishing



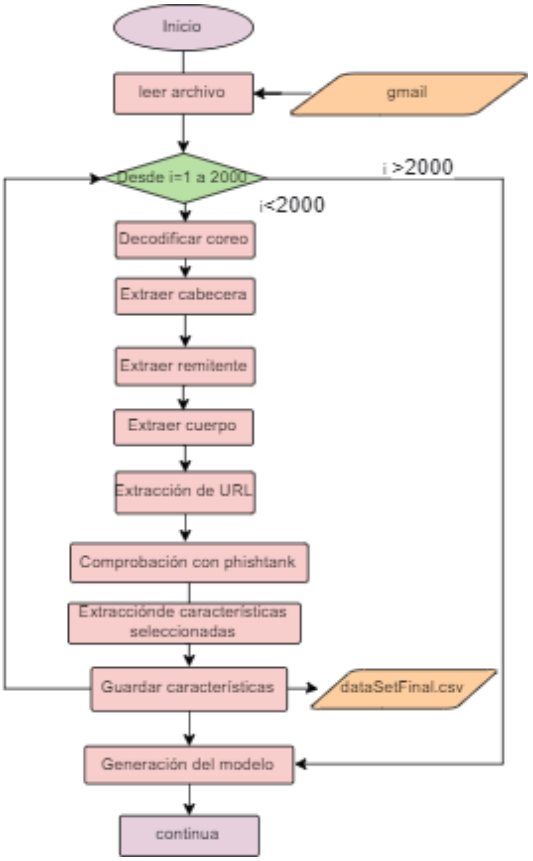
- **Lectura de correos electrónicos**

  La única diferencia a los subprocesos descritos, es la lectura de correos desde la conexión a Gmail y no la lectura de un archivo. Antes de realizar la extracción de información es necesario realizar la decodificación de los correos para así poder tener un mismo formato que permita la extracción e identificación de todas las características.

  Al finalizar este subproceso se proceda ya a la generación del modelo predictivo.

Figura 36

Lectura de correos de Gmail



Luego de haber realizado estos subprocesos, se tiene el archivo dataSetTesisFinal.csv, el mismo que permite realizar la generación de modelo y realizar las pruebas de una manera efectiva.

- **Generación del modelo predictivo**

  Esta etapa del proyecto empieza con la lectura del dataset generado, seguido por la definición de las variables “x”, “y” que permiten el entrenamiento y predicción del modelo. Se asigna a “x” las 24 características de Phishing y a “y” la variable objetivo, que define si la caracterización pertenece a un correo con Phishing o no Phishing.

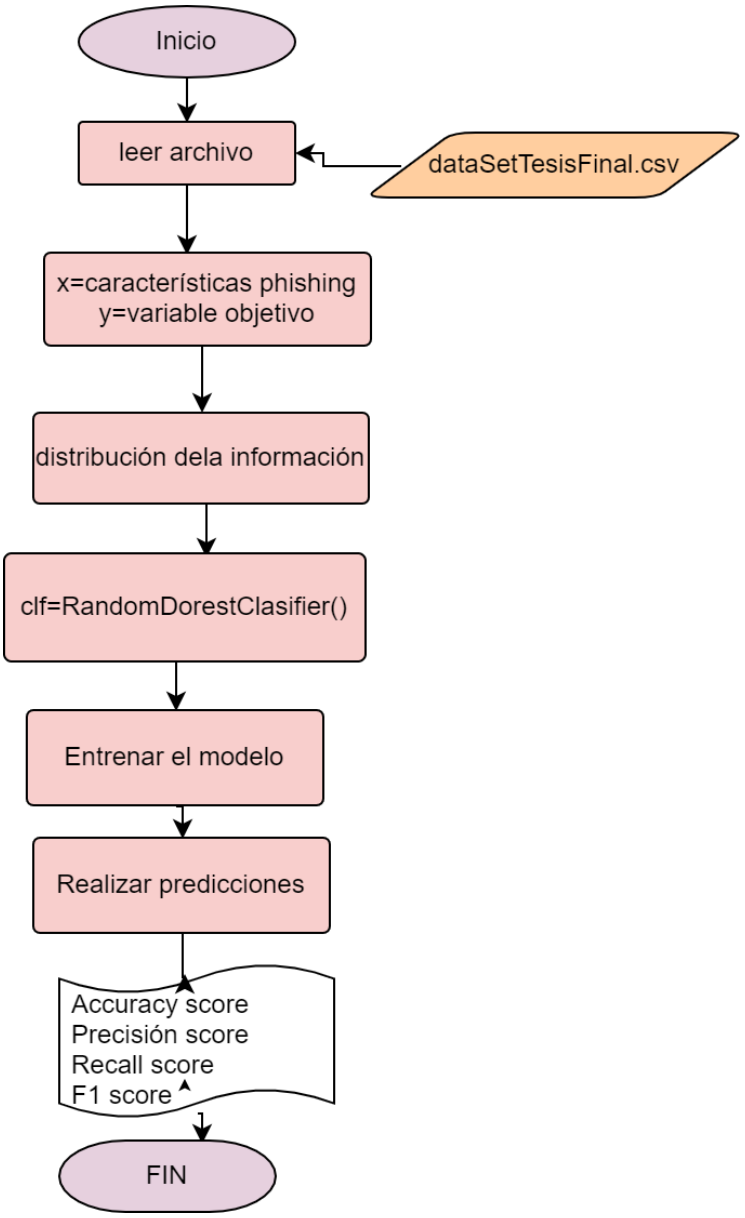
  Una vez declarada estas variables, se divide el porcentaje de datos para entrenamiento y pruebas, los datos seleccionados se harán de manera randómica. Esta división se le realiza en base al principio de Pareto siento el 80% de datos para entrenamiento y el 20% para prueba, esto se lo realiza mediante la declaración de la variable random_state=1.

  Se procede a realizar el entrenamiento del modelo con el algoritmo de RamdonForest, este algoritmo se encarga de realizar el proceso respectivo para que el modelo aprenda de las características y variable objetivo declaradas anteriormente, seguido se realiza las predicciones correspondientes. Este proceso permite obtener las métricas “Accuracy score”, “Precisión score”, “Recall score”, “F1 score”, que son las encargadas de dar validez de la precisión y fiabilidad del modelo respecto a la detección de Phishing en los correos electrónicos analizados.

  Ya generado el modelo, se procede aplicarlo en el ambiente controlado de Gmail, para el análisis de los correos electrónicos.

Figura 37

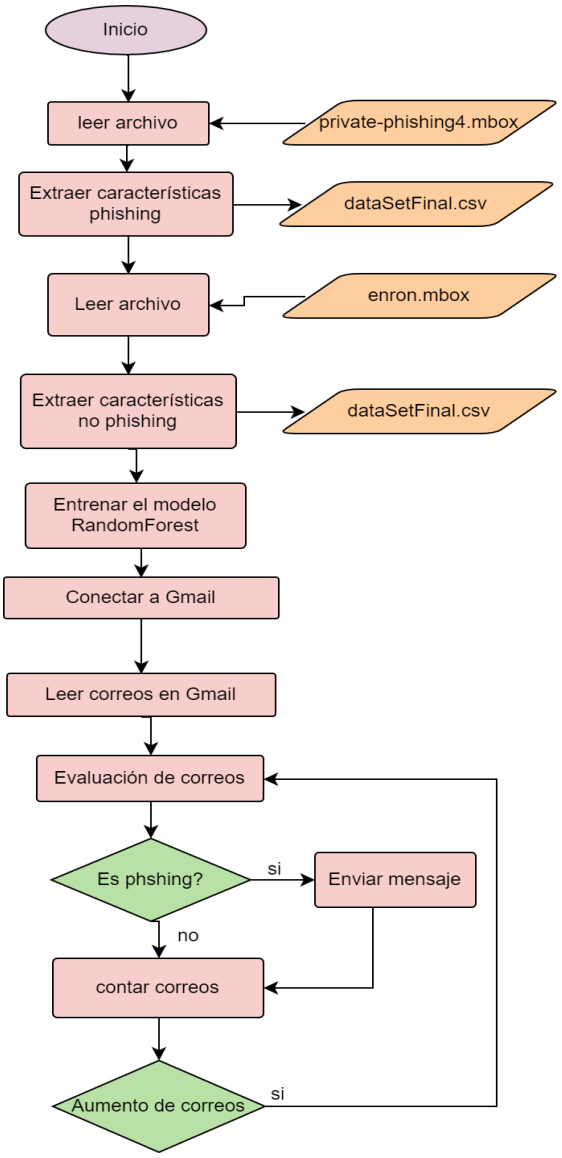
Subproceso generación del modelo de predicción



Luego de haber entendido como es el funcionamiento de cada uno de los subprocesos del modelo, se puede tener una mejor visión del proceso general del proyecto y su funcionamiento. En la Figura 38 se puede visualizar de manera general cual es el flujo del algoritmo generado cumpliendo con los objetivos definidos.

Figura 38

Proceso del modelo de predicción



La Figura 38, representa el flujo de todo el proceso que realiza el modelo. Una vez realizado cada subproceso detallado, el modelo se pone en funcionamiento bajo un ambiente controlado, para probar la efectividad del mismo. Estas pruebas se las realiza mediante la conexión de Gmail.

Para que el proceso de detención se pueda visualizar sin necesidad del código, se realizó la conexión de Python con WhatsApp, mediante la librería Twilio, que es una API encargada de generar un token, Id y número telefónico que son los que permiten vincular el código con WhatsApp y enviar mensajes. En la figura 39 se puede visualizar el código de Python de **Figura 39**

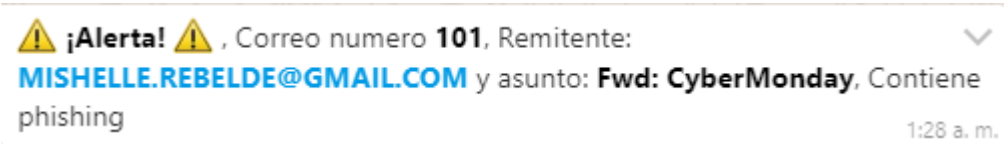
Código de conexión de Python con WhatsApp

```
def sendWhatsapp():
    account_sid = 'AC56a74523b27c5543ab5ae0f32e8745b0'
     auth_token = 'e5c46d69f0ef5ea7e9a961a1ac914103'
    client = Client(account_sid, auth_token)

    message = client.messages.create(
                                from_='whatsapp:+14155238886
',
                                body='Correo con Phishing',

                                to='whatsapp:+593963216876'
                            )

    print(message.sid)
```

El algoritmo verifica el número de correos que existen en el Gmail, en el caso de existir aumento de los mismo, el modelo hace el análisis respectivo, para verificar si es Phishing o no. En el caso que el modelo detecte que el correo es Phishing, se envía un mensaje de WhatsApp, con el asunto, remitente y una nota poniendo en alerta que ese correo contiene características de Phishing, permitiendo así realizar una mitigación al no dejar que el usuario se exponga a correos fraudulentos. El resultado de este proceso se observa en la Figura 40. Cabe mencionar que se lo realiza cada que transcurra 5 minutos, con el uso de hilos.

Figura 40

Notificación de correo con phishing



Planear la Monitorización y Mantenimiento

El mantenimiento de la implementación del modelo es una fase muy importante del mismo, debido al constante cambio o aumento de características que determinen que un correo es Phishing. Como se ha mencionado, el avance de la tecnología hace que phishers busquen continuamente, nuevas vulnerabilidades en los correos. Es por este motivo que cada determinado tiempo, el código necesita ser actualizado, para generar un dataset con mucha más precisión.

Como plan de monitorización y mantenimiento se puede establecer los siguientes procesos:

- Actualización anual de características seleccionadas que determina si un correo es Phishing o no.
- Actualización de archivos de las fuentes Enron y Monkey.
- Actualización de modelos que pueden mejorar la precisión de detección.

CAPÍTULO V

PRUEBAS Y ANÁLISIS DE RESULTADOS

El análisis de resultados, se lo realizó mediante un proceso estadístico, que muestra valores y porcentajes reales, obtenidos de cada prueba que se consideró necesaria para verificar el funcionamiento adecuado del modelo. Cabe mencionar que todas las pruebas, se realizaron en un ambiente controlando, donde se manejaron correos con phishing generados y archivos obtenidos de fuentes confiables. También se muestra la comparación del modelo final con otros algoritmos de minería de datos como árboles de decisión y Naive Bayes.

RESULTADOS

Para realizar el análisis respectivo, se la divide en 3 fases, la primera fase es el entrenamiento del modelo, seguida por la fase de predicciones, y por último la tercera hace referencia a los resultados obtenidos al aplicar el modelo en un ambiente controlado para la detección de phishing en los correos electrónicos en la plataforma de Gmail.

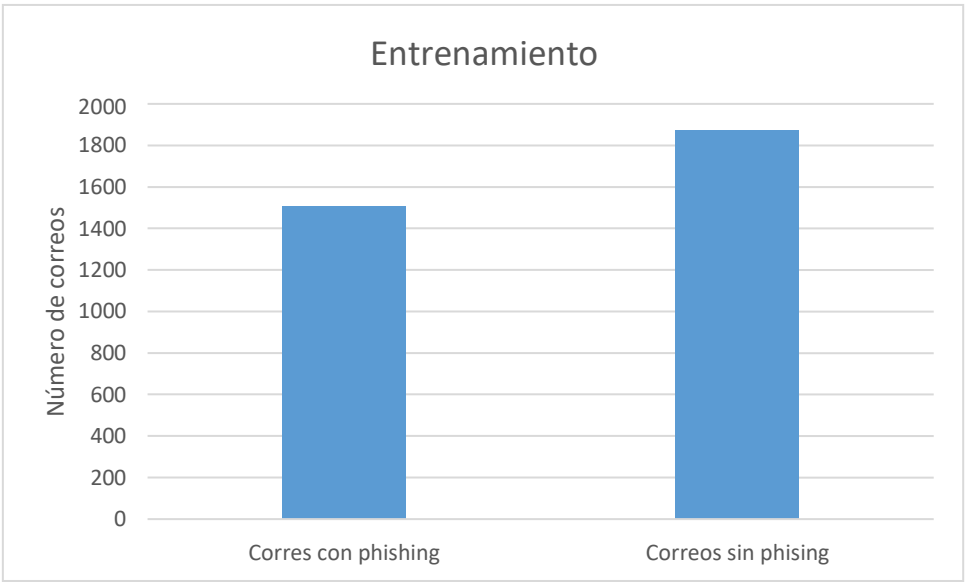
A continuación, se detallan los resultados y su análisis del modelo.

Primera fase: Entrenamiento del modelo

En esta fase se trabaja con un total 3375 correos para entrenar al modelo , de los cuales 1505 están infectados y 1870 son correos que no contiene datos con phishing. Es importante mencionar que estos datos representan el 80% del total de los datos obtenidos en el dataset. A continuación, se puede visualizar este resultado.

Figura 41

Número de correos electrónicos con phishing y sin phishing



Antes de continuar con la fase de predicción es importante mencionar que, para todo el proceso de generación del modelo se utilizó la definición de Pareto para la distribución de los datos, determinando el 80% para aprendizaje y el 20% para pruebas. En la Figura 41 se puede visualizar dicha distribución.

Figura 42

Distribución de datos mediante la definición de Pareto

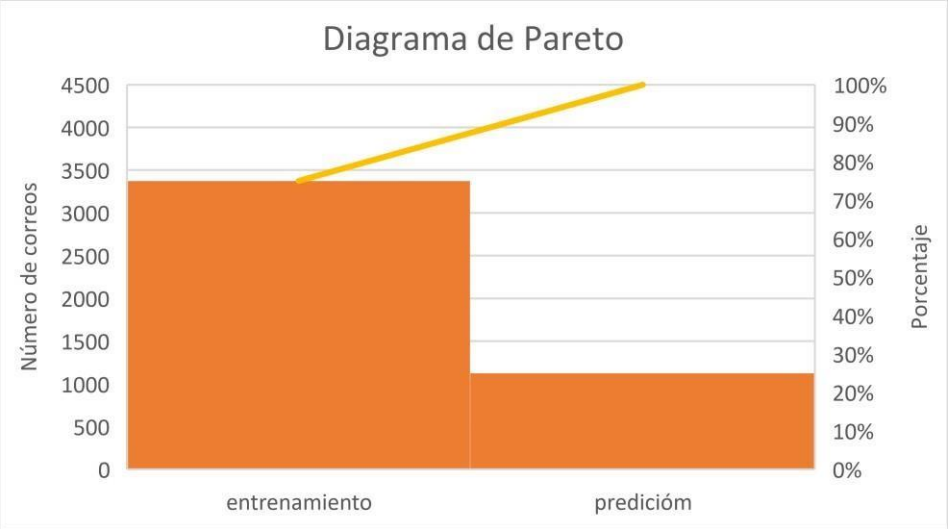
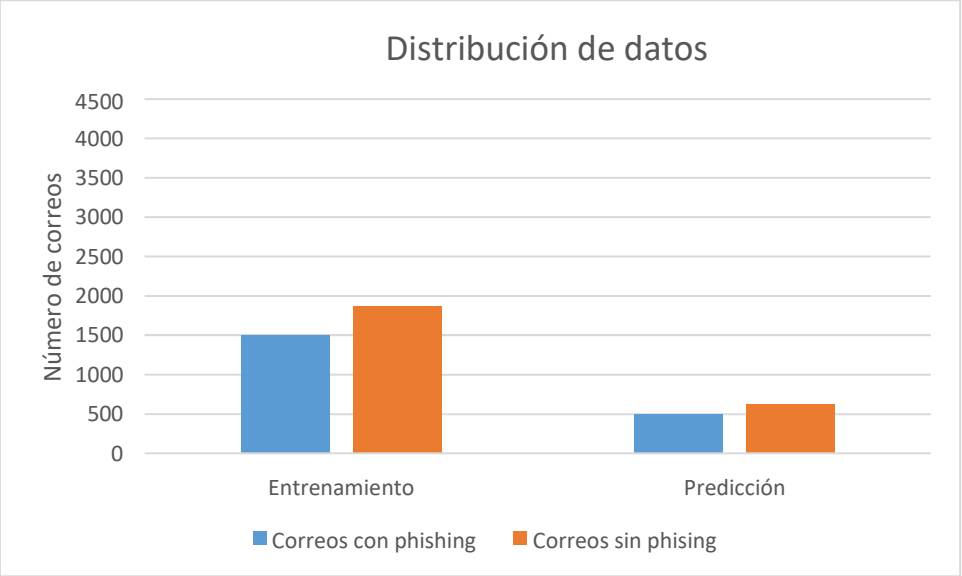


Figura 43

Distribución de datos



Como se visualiza en la Figura 42, son 1125 datos que representan el 20% de los datos para realizar la predicción distribuidas en 495 correos con phishing y 630 correos sin phishing. Al igual que como se mencionó anteriormente son 3375 correos para realizar el proceso de aprendizaje respectivo.

Segunda fase: Predicción del modelo

Como se describió en el punto anterior, se trabajó con 1125 correos para la fase de predicción, que se la realiza con el algoritmo de RamdonForest. Este proceso utiliza la librería sklearn, la misma que proporciona las métricas necesarias para la evaluar los resultados de las predicciones del modelo.

Para esto se definieron las variables Verdadero positivo (VP) que determina el número de correos marcados como phishing y si lo son realmente, Verdadero negativo (VN) aquellos que se marcaron como no phishing y no lo son, Falsos Positivos (FP) aquellos correos que se marcaron como phishing y en realidad no lo eran. Finalmente, los Falsos negativos (FN) que se marcaron como no Phishing y si lo eran. Todas estas variables son definidas por la librería que trabaja con ella en sus funciones internas. (scikit learn, 2020). A continuación, se puede visualizar la cantidad exacta de correos en cada variable, de acuerdo a las predicciones del modelo.

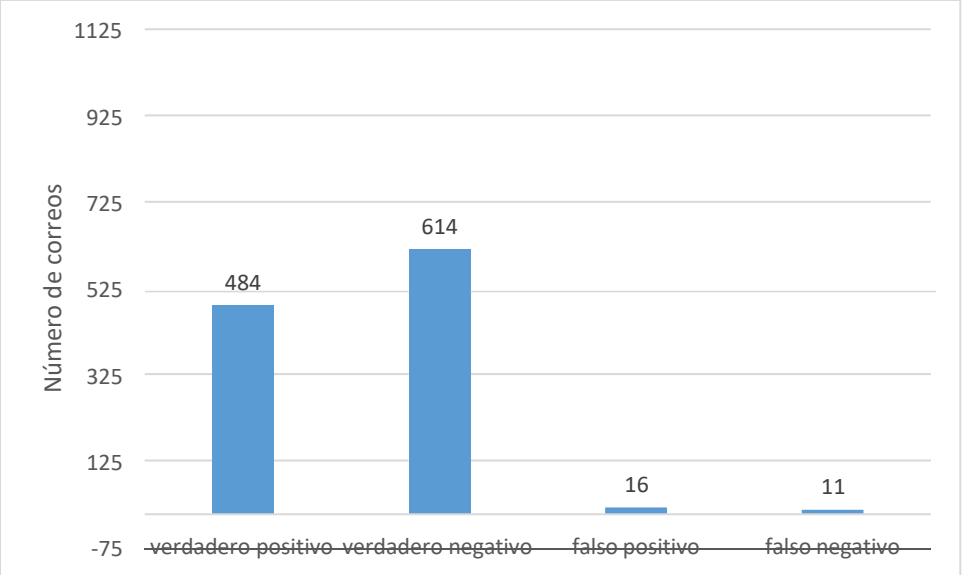
Tabla 13

Número de correos en cada variable

Variables	y_pred	y_test	Totalcorreos
VP	1	1	484
VN	0	0	614
FP	1	0	16
FN	0	1	11

Figura 44

Número de correos en cada variable



Para la valuación del desempeño del modelo en esta fase se utiliza los datos descritos en la Tabla 13, que son las variables necesarias para determinar el valor de las utiliza las métricas definidas en el CAPITULO IV. Esto permite analizar el modelo de una manera estadística.

A continuación, se puede visualizar los resultados de las ecuaciones de cada métrica, que permite saber el nivel de precisión al momento de la detección y aprendizaje de este modelo. Los valores necesarios para el desarrollo de las ecuaciones se encuentran definidas en la Tabla 13.

La Ecuación 1 Accuracy score indica la exactitud, es decir determina el número de casos correctos que el modelo definió. (IArtificial.net, 2020)

$$Accuracy = \frac{\text{VP+VN}}{\text{VP+VN+FP+FN}} \qquad \textbf{Ecuación 1}$$

$$Accuracy = \frac{484 + 614}{484 + 614 + 16 + 11} * 100$$

$$Accuracy = 97{,}6\%$$

Para la medición de la calidad del modelo de detección de phishing se aplica la métrica de precisión Ecuación 2, identificando el número de correos que verdaderamente tienen phishing. (IArtificial.net, 2020)

$$precision = \frac{VP}{VP+FP}$$ **Ecuación 2**

$$precision = \frac{484}{484 + 16} * 100$$

$$precision = 96.8\%$$

En la Ecuación 3 la métrica Recall score identifica cual es el porcentaje de los correos electrónicos que están infectados, que el modelo es capaz de identificas. (IArtificial.net, 2020)

$$recall = \frac{\text{VP}}{\text{VP+FN}}$$ **Ecuación 3**

$$recall = \frac{484}{484+11} *100$$

$$recall = 97.77\%$$

Para determinar el rendimiento del modelo se realiza una combinación de la métrica precisión y recall en un valor. Este valor se lo obtiene del F1 score Ecuación 4. (IArtificial.net, 2020)

$$F1\ score\ = \frac{PRECISION\ SCORE * RECALL SCORE}{PRECISION\ SCORE + RECALL\ SCORE} *2$$ **Ecuación 4**

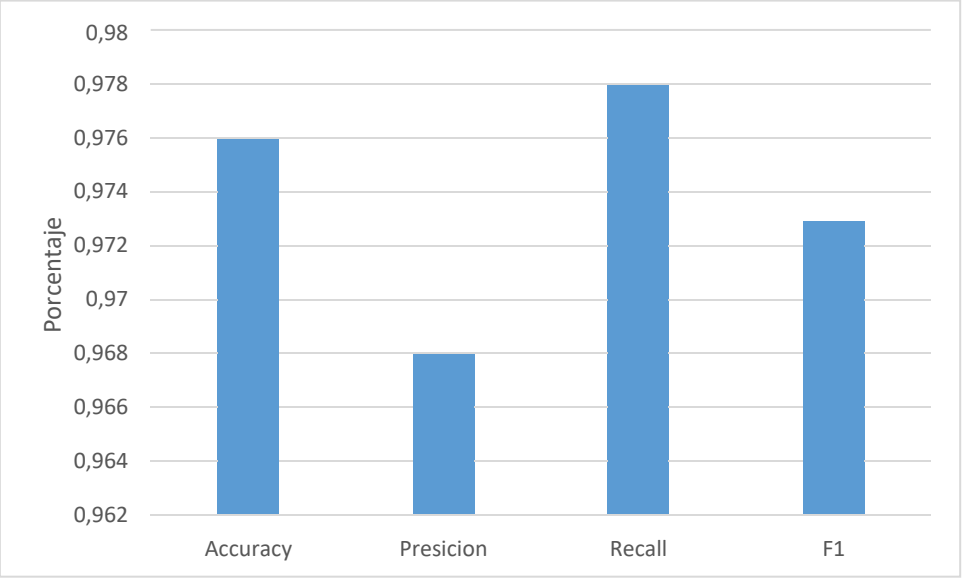
$$F1\ score = (\frac{0{,}968 * 0.97777778}{0.968+0{,}97777778} *2) *100$$

$$F1\ score = 97{,}28\%$$

En la Figura 44 se puede visualizar los resultados obtenidos en cada una de las métricas.

Figura 45

Resultados de métricas



Una vez obtenido las variables definidas y los resultados de las métricas, se puede visualizar el nivel de precisión del correo es de un 97% superando el porcentaje que se planteó cumplir al iniciar. Es necesario también tener plasmado el porcentaje de error que tiene el modelo, que da mayor validez a este análisis de resultados.

En la Ecuación 5 determina la tasa de errores (ER), es decir el porcentaje en el que modelo no clasifico los correos de manera correcta.

$$ER = \frac{\text{FN+FP}}{\text{VN+FN+FP+VP}}$$ **Ecuación 5**

$$ER = \frac{11+16}{614+11+16+484} *100$$

$$ER = 2{,}4\%$$

En resumen, de todos los resultados obtenidos, se visualiza que, de los 1125 correos

analizados, el 97,77% que corresponde a 1098 correos fueron categorizados correctamente, mientras que solo el 2.4% correspondiente a 27 correos fueron categorizados incorrectamente. Por lo tanto, se determina que el modelo cuenta con un porcentaje de predicción alto, teniendo fiabilidad en los resultados que el modelo entrega en cada análisis de correos electrónicos y detección de Phishing.

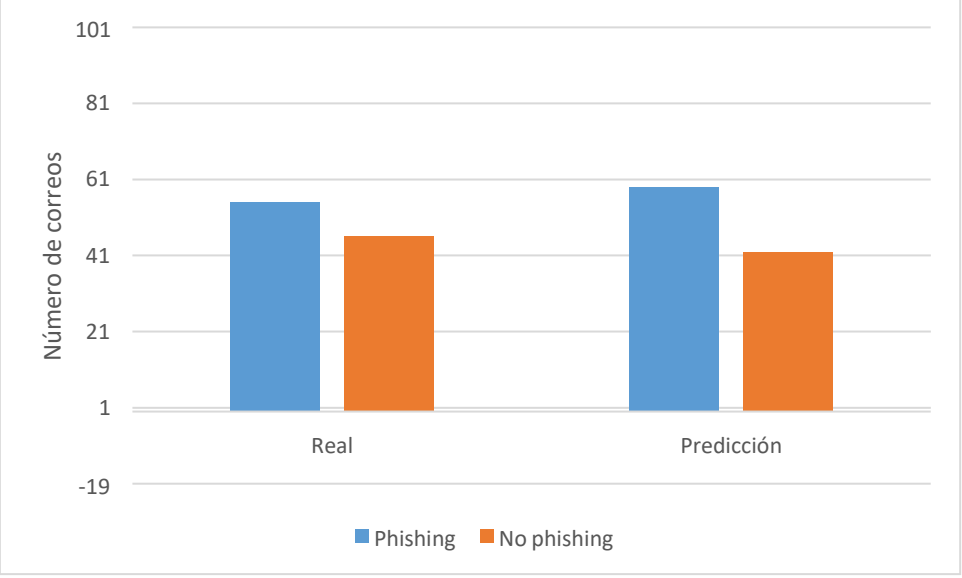
Fase 3: Aplicación del modelo

En esta fase se realiza un análisis de resultados, referente al número de correos detectados con phishing en el ambiente controlado. Para esto se tiene una cuenta de Gmail, en donde existen 101 correos, 55 son phishing y 46 no phishing.

Al analizar estos correos, el modelo categorizó 59 correos con phishing y 42 correos sin phishing. Estos resultados en términos de porcentaje dan un 96.03% de precisión. Es necesario destacar que este análisis se lo realiza no solo con correos generados ya que también se alimentó la base con correos reales, enviados a entidades comerciales públicas. En la figura 46 se puede visualizar los valores exactos descritos en este punto.

Figura 46

Resultados de la aplicación del modelo de precisión

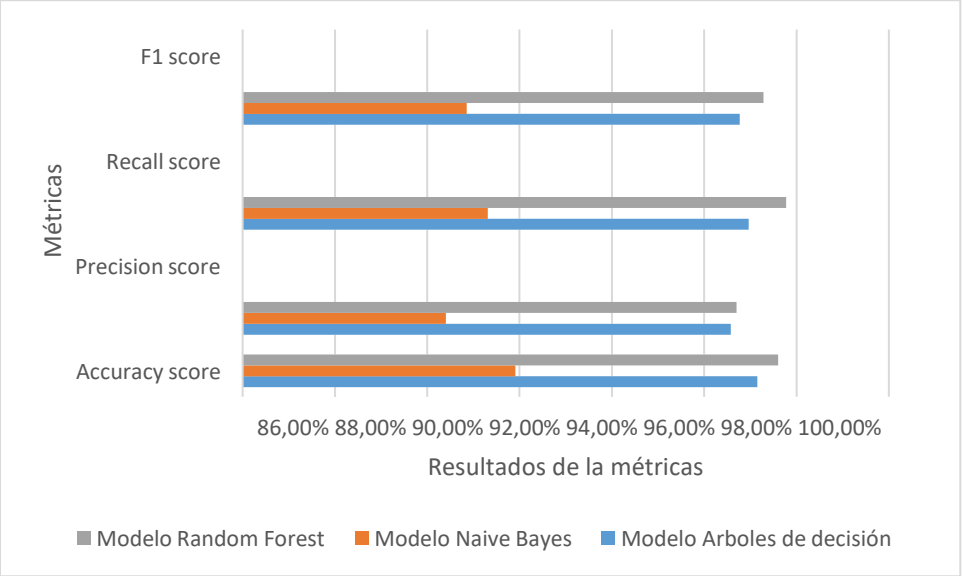


Comparación con otros modelos de minería de datos

Se realizó la comparación con otros modelos para poder determinar la eficacia del modelo seleccionado, mediante el respaldo de valores estadísticos que determinas porque el modelo RamdonForest es correcto.

Para la comparación se seleccionaron los modelos árbol de decisión y Naive Bayes, seleccionadas por el motivo que son algoritmos cuyo funcionamiento también permite el cumplimiento del objeto de este proyecto. Además de ser de los principales modelos aplicados en la minería de datos. Se obtuvo las métricas descritas en el punto anterior, para poder comparar los modelos de acuerdo al porcentaje obtenido.

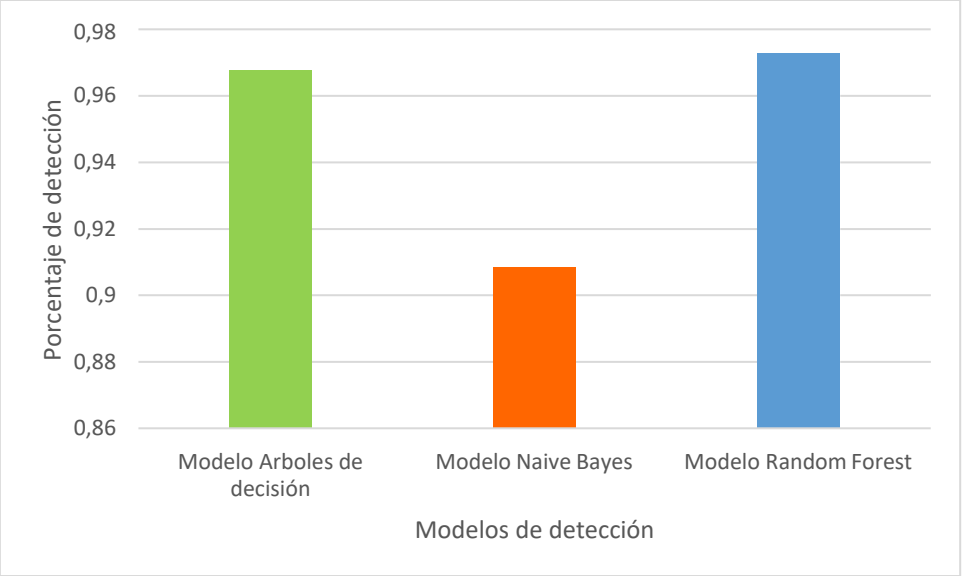
Figura 47 *Comparación con otros modelos*



En base a los resultados de cada una de las métricas se puede analizar y determinar que Random Forest tiene un mayor nivel de porcentaje de precisión en todas las métricas, pues otorgó al modelo una precisión en la detección del 97.77%, seguido por el modelo de árboles de decisión con un 96,77% y por último Naive Bayes con un porcentaje de 90,85%. Estos resultados se los puede visualizar en la figura 47 y 48.

Figura 48

Porcentaje de detección de cada modelo



Discusión

Una vez obtenido todos los resultados de las diferentes pruebas aplicadas, y realizado su respectivo análisis, se determina que el modelo construido logró un 97.77% en determinar la existencia de phishing en un correo electrónico.

Al mostrar que el modelo superó el porcentaje esperado, también se debe mencionar que, se pudo comprobar la veracidad de la hipótesis definida al inicio del proyecto, donde se mencionaba que la aplicación de un modelo de precisión reducirá el número de correos electrónicos con phishing. Aquí es importante mencionar, el envío del mensaje alertando al usuario de presencia de phishing en su correo electrónico, provocando que el usuario no tenga necesidad de abrir el correo o verificar la veracidad de su contenido, más pueda eliminarlo sin miedo a perder información importante, y así se realiza la mitigación del ataque, además que el modelo se encarga de enviar estos correos fraudulentos a la bandeja de spam. Pese a esto también se debe mencionar que existe un 2,4% de error que categoriza correos de manera incorrecta y genera cierto grado de

incertidumbre.

La precisión del modelo puede variar de acuerdo al enfoque que se tenga, al igual que el algoritmo seleccionado para la implementación. Los algoritmos que se seleccionaron en este proyecto se los determinó en base a los estudios primarios.

CAPÍTULO VI

COCLUSIONES Y RECOMENDACIONES

Conclusiones

Para la selección correcta de las características, que definen si un correo es legítimo o no, se necesita en primera instancia partir de una revisión sistemática de varios estudios primarios que hayan caracterizado la presencia del phishing en un correo electrónico, y así recopilar aquella información común y relevante en cada uno de estos estudios. Estos datos se complementan con la identificación de nuevas características que se logra al realizar un análisis exhaustivo de correos que contiene o no phishing obtenidos de fuentes confiables.

La aplicación de la metodología CRISP-DM, facilitó el diseño e implementación del modelo, ya que cada una de sus fases permite cumplir con el correcto proceso de minería de datos, además de dar un enfoque claro de los objetivos del modelo y del negocio.

Python es un lenguaje de programación de alto nivel y código abierto, con una basta cantidad de librerías, herramientas y funciones que permiten desarrollar todo tipo de aplicaciones.

En base al análisis de los resultados en las pruebas de todos de los modelos seleccionados Árbol de decisiones, Naive Bayes y Random Forest, se pudo concluir que este último es el modelo tiene mayor precisión al momento de categorizar y predecir un correo electrónico, a causa de que este modelo tiene un grado más alto de complejidad con respecto al aprendizaje.

El porcentaje de precisión del modelo, depende de la correcta selección de características, así como la cantidad de datos utilizados durante el entrenamiento y predicción. La selección de la herramienta de minería de datos es muy importante, ya que influye en los resultados obtenidos.

Al aplicar el modelo en el ambiente controlado en la plataforma de Gmail, que

contiene tanto correos reales, como correos simulados con phishing, se observa que el porcentaje de precisión disminuye. Este debido a que existen correos de entidades comerciales que por motivos de marketing utilizan varias características que puede provocar que el correo se interprete como phishing.

Recomendaciones

Debido a que los phishers se encuentran en constante búsqueda de nuevas vulnerabilidades para evitar la detección de sus ataques, se recomienda realizar un constante análisis de correos con phishing para determinar características nuevas o que no se hayan tomado en cuenta durante la generación de este modelo y así mejorar el dataset de entrenamiento.

Para le selección adecuada de la herramienta de minería de datos, se recomienda en primer lugar hacer una revisión sistemática de la literatura, y así tener una guía de que modelos según su funcionamiento pueden ayudar a cumplir el objetivo planteado, seguido por realizar las pruebas y comparaciones necesarias para ver el comportamiento de cada modelo, y así elegir el que mejor se acople y ayude a obtener los resultados deseados.

Para proyectos futuros que partan de este modelo, se recomienda la aplicación de Inteligencia Artificial para automatizar de manera efectiva cada uno del proceso. Esto con el fin de generar un nuevo modelo no supervisado que se pueda adaptar e implementar dentro un servidor de correos electrónicos, donde se obtenga resultados en un ambiente no controlado, dándole mayor validez al modelo.

REFERENCIAS BIBLIOGRÁFICAS

Cual es la importancia de los metadatos. (s.f.). Recuperado el 11 de JULIO de 2019, de http://culturacion.com/cual-es-la-importancia-de-los-metadatos/

25 EMIS. (6 de Marzo de 2019). Obtenido de https://www.emis.com/php/company-profile/EC/Compania_de_Transporte_de_Carga_Pesada_Nevisacargo_SA_es_4903574.html

A. Guayasmín, W. F. (2018). *Formalistic Modelling Based on Pattern Recognition Applied to the Knowledge and Human Talent Sector in Ecuador.* Bogotá: ICAI Workshops (ICAIW).

Ali, N. S. (2009). *Examining the Efficacy of Online Self-Paced Interactive Video-Recordings inNursing Skill Competency Learning: Seeking Preliminary Evidence Through an Action Research.* Medical Science.

Ali, N. S. (2009). *Examining the Efficacy of Online Self-Paced Interactive Video-Recordings inNursing Skill Competency Learning: Seeking Preliminary Evidence Through an Action Research.* Medical Science.

Alvarez, M. A. (19 de Noviembre de 2003). *Desarrolloweb.com*. Obtenido de https://desarrolloweb.com/articulos/1325.php

Andronicus A. Akinyelu, A. O. (2014). *Clasificación del correo electrónico de phishing mediante la técnica de aprendizaje automático de bosque aleatorio.* Journal of Applied Mathematics .

Anti-Phishing Working Group, I. (2019). *APWG*. Obtenido de https://apwg.org/

APWG. (2019). Obtenido de https://docs.apwg.org/reports/apwg_trends_report_q1_2019.pdf

APWG. (2019). Obtenido de https://docs.apwg.org/reports/apwg_trends_report_q2_2019.pdf

Arán, J. M. (2 de Septiembre de 2019). *welivesecurity by ESSET*. Recuperado el 7 de 11 de 2019, de https://www.welivesecurity.com/la-es/2019/09/02/por-que-ataques-phishing-tan-efectivos/

Atighetchi, M. &. (2009). *Attribute-based Prevention of Phishing Attacks.* Cambridge, England.: Eighth International Symposium on Network Computing and Applications (IEEE).

Awi, Y. L. (2014). *A study of factors affecting consumer's repurchase intention toward XYZ restaurant, Myanmar.* International Conference on Trends in Economics, Humanities, and Management.

B. Espinoza, J. S. (2019). Phishing Attack Detection: A Solution Based on the Typical Machine Learning Modeling Cycle. *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, 202-207. doi:10.1109/CSCI49370.2019.00041

Bassi, S. (2019). *Google sites*. Obtenido de https://sites.google.com/site/sbassi/leyendoxmlenpython:dom2#:~:text=minidom.,accesible%20desde%20nuestro%20programa%20Python.

Benavides E., F. W. (2020). Classification of Phishing Attack Solutions by Employing Deep Learning Techniques: A Systematic Literature Review. *Developments and Advances in*

Defense and Security. Smart Innovation,, 152. doi:https://doi.org/10.1007/978-981-13-9155-2_5

Bertolín, J. A. (2008). *Seguridad de la informacion redes, informatica y sistemas de informacion.* Paraninfo.

BI-Spain.com - Teoría, B. I. (2009). *CRISP: Una Metodología Data Mining para inexpertos.* Obtenido de https://www.bi-spain.com/articulo/283/crisp-una-metodologia-data-mining-para-inexpertos

Blanco, A. G. (08 de Enero de 2018). *BBVA*. Obtenido de https://www.bbva.com/es/ataques-ingenieria-social-evitarlos/

Bright, M. (2011). *MillerSmiles*. Obtenido de http://www.millersmiles.co.uk/

Carrion, S. (20 de Julio de 2018). *EXTRA*. Obtenido de https://www.extra.ec/actualidad/tarjetas-fraude-clonacion-delito-bancos-AL2287266

Chuenchujit, T. (2016). *A TAXONOMY OF PHISHING RESEARCH.* Obtenido de https://www.ideals.illinois.edu/bitstream/handle/2142/90570/CHUENCHUJIT-THESIS-2016.pdf?sequence=1&isAllowed=y

CHUENCHUJIT, T. (2016). *A TAXONOMY OF PHISHING RESEARCH.* Obtenido de https://www.ideals.illinois.edu/bitstream/handle/2142/90570/CHUENCHUJIT-THESIS-2016.pdf?sequence=1&isAllowed=y

Cortina, V. G. (2015). *Aplicación de la metodología CRISP-DM a un proyecto de minería de datos en el entorno universitario.* Madrid.

Cortina, V. G. (Ocubre de 2015). *e-archivo.* Obtenido de APLICACIÓN DE LA METODOLOGÍA CRISP-DM A UN PROYECTO DE MINERIA DE DATOS EN EL ENTORNO UNIVERSITARIO: https://e-archivo.uc3m.es/bitstream/handle/10016/22198/PFC_Victor_Galan_Cortina.pdf

Cortina, V. G. (Octubre 2015). *APLICACIÓN DE LA METODOLOGÍA CRISP-DM A UN PROYECTO DE MINERÍA DE DATOS EN EL ENTORNO UNIVERSITARIO.*

Creswell. (2015). *Educational research. Planning, conducting and evaluating quantitative and qualitative research.* USA.

D. Oña, L. Z. (2019). "Phishing Attacks: Detecting and Preventing Infected E-mails Using Machine Learning Methods. *3rd Cyber Security in Networking Conference (CSNet)*, 161-163. doi:10.1109/CSNet47905.2019.9108961.

Data, P. (s.f.). *Power Data.* Recuperado el 11 de JULIO de 2019, de https://www.powerdata.es/metadatos

Dirección General de Modernización Administrativa, P. e. (2012). *Metodología de Análisis y Gestión de Riesgos de los Sistemas de Información (Versión 3). Libro I.* Madrid: Ministerio de Hacienda y Administraciones Públicas.

Dunlop, M. G. (2010). *Gold Phish: Using Images for Content-Based Phishing Analysis. .* Barcelona, España.: Fifth International Conference on Internet Monitoring and Protection.

Esteve, A. J. (2018). *Influencia de las fake news en marcas.* Obtenido de https://riunet.upv.es/bitstream/handle/10251/107501/L%C3%93PEZ%20-%20Influencia%20de%20las%20fake%20news%20en%20marcas%20y%20medios%20en%20la%20Web.pdf?sequence=1&isAllowed=y#:~:text=Tldextract%3A%20es%20un%20m%C3%B3dulo%20Python,o%20el%20dominio%20superi

Foundation., P. S. (2020). *Python*. Obtenido de https://docs.python.org/3/library/urllib.request.html

Foundation., P. S. (s.f.). *Tutorial de Python 3.6.3 documentation*. Obtenido de http://docs.python.org.ar/tutorial/3/real-index.html

GEOIDEP. (s.f.). *Que son los metadatos.* Recuperado el 11 de JULIO de 2019, de https://www.geoidep.gob.pe/metadatos/que-son-los-metadatos

Guayaquil, E. (06 de MAYO de 2017). *EXTRA.EC*. Obtenido de https://www.extra.ec/actualidad/motel-delitos-revelacion-ilegal-de-base-dedatos-intendencia-de-policia-XB1300264

Gyuris, E. N. (2018). Obtenido de https://www.ridaa.unicen.edu.ar/xmlui/bitstream/handle/123456789/1960/Nagy%20Gyuris%20Emilio.pdf?sequence=1&isAllowed=y

Harán, J. M. (2 de Septiembre de 2019). *welivesecurity by ESET*. Obtenido de Harán, J., & Harán, J. (2019). Por qué los ataques de phishing siguen siendo tan efectivos | WeLiveSecurity. WeLiveSecurity. Retrieved 24 October 2019, from https://www.welivesecurity.com/la-es/2019/09/02/por-que-ataques-phishing-tan-efectivos/

Hernández Dominguez, A., & Сторчак, С. (2019). SYSTEM FOR THE DETECTION OF PHISHING ATTACKS USING EMAIL. *Telemática*.

Hernández, Fernández y Baptista. (2014).

IArtificial.net. (09 de 10 de 2020). Obtenido de https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/

Ibid. (s.f.).

IsoTools. (2013). *IsoTools*. Obtenido de https://www.isotools.org/pdfs-pro/iso-27001-sistema-gestion-seguridad-informacion.pdf

Juan Miguel Moine,Ana Silvia Haedo. (s.f.). *Una herramienta para la evaluación y comparación de.* Obtenido de http://sedici.unlp.edu.ar/bitstream/handle/10915/50428/Documento_completo.pdf-PDFA.pdf?sequence=1&isAllowed=y

kaspersky. (2019). *kaspersky*. Obtenido de kaspersky: https://latam.kaspersky.com/blog/kaspersky-lab-registra-un-alza-de-60-en-ataques-ciberneticos-en-america-latina/13266/

KDD. (s.f.). Knowledge Discovery in.

LA HORA. (28 de Febrero de 2020). Obtenido de https://www.lahora.com.ec/tungurahua/noticia/1102308887/delitos-ciberneticos-frecuentes-en-ecuador-

learnpython.org. (2019). *learnpython.org*. Obtenido de https://www.learnpython.org/es/Pandas%20Basics

LOGICALIS. (25 de 03 de 2015). *Predictive Analytics: los principales modelos del análisis predictivo*. Obtenido de https://blog.es.logicalis.com/analytics/predictive-analytics-los-principales-modelos-del-analisis-predictivo

Martínez, M. B. (2018). *Minería de Datos.* Obtenido de http://bbeltran.cs.buap.mx/NotasMD.pdf

matiasdaza. (21 de Noviembre de 2017). *GitHub*. Obtenido de https://github.com/matiasdaza/Convert-mbox-to-csv/blob/master/mbox-to-csv.py

Medero, G. S. (s.f.). *DELITOS EN INTERNET:CLASES DE FRAUDES Y ESTAFAS.* Madrid: Dialnet.

Microsystem. (s.f.). *Microsystem*. Obtenido de https://www.microsystem.cl/plataforma/rapidminer/

Monkey.org. (2020). *Monkey.org*. Obtenido de Monkey.org: https://monkey.org/~jose/phishing/

Mundial, B. (s.f.). *Banco Mundial*. Obtenido de https://www.bancomundial.org/es/what-we-do

Nahorney, B. (2015). *The MessageLabs Intelligence Annual Security Report: 2009 Security Year in Review*. Obtenido de http://www.symantec.com/content/en/us/enterprise/other_resources/intelligence-report-06-2015.en-us.pdf [

Nazario, J. (2005). *The online phishing corpus*. Obtenido de https://monkey.org/~jose/phishing/

Noviscargo. (s.f.). *Noviscargo*. Obtenido de http://nevisacargo.com/index.html

Ortega, J. (s.f.). *El Comercio*. Obtenido de https://www.elcomercio.com/actualidad/seguridad/mafias-movilizan-celulares-robados-paises.html

PCevallor. (18 de Septiembre de 2018). *Exacto revista Digital*. Obtenido de http://exactodigital.com/justicia-ecuatoriana-sentencia-primera-vez-caso-pirateria-senal-tv-paga-satelital/

Pentaho. (29 de Octubre de 2019). Obtenido de Business Intelligence, Data Warehouse, Monterrey, México : https://gravitar.biz/pentaho/

PERALTA, F. (2014). *Proceso de Conceptualización del Entendimiento del Negocio para Proyectos de Explotación de Información.* evista Latinoamericana de Ingenieria de

Software. 2. 273. 10.18294/relais.2014.273-306. .

PhishTank. (2011). *PhishTank.* Obtenido de http://www.phishtank.com

R.M., M., F, T., & McCluskey, L. (2015). Tutorial and critical analysis of phishing websites methods. *Computer Science Review, vol. 17*, 1-24.

RAMOS, M. D. (2011). *SEGURIDAD INFORMATICA.* Paraninfo.

Redalyc.org. (s.f.). Recuperado el 29 de Octubre de 2019, de https://www.redalyc.org/pdf/1939/193930080003.pdf

Robledano, Á. (23 de Septiembre de 2019). *OpenWebinars*. Obtenido de https://openwebinars.net/blog/que-es-python/

Rodriguez, M. L. (19 de Agosto de 2013). *Guía de tesis*. Obtenido de https://guiadetesis.wordpress.com/tag/investigacion-bibliografica-y-documental/

Román, J. V. (2 de Agosto de 2016). *Singular*. Obtenido de CRISP-DM: La metodología para poner orden en los proyectos: https://www.sngular.com/es/data-science-crisp-dm-metodologia/

Romero, J. (11 de Junio de 2019). *Técnicas y algoritmos de Minería de Datos*. Obtenido de https://jorgeromero.net/tecnicas-y-algoritmos-de-mineria-de-datos/

R-project. (2019). *R: The R Project for Statistical Computing*. Obtenido de https://www.r-project.org/

Rusydiana, A. S.-F. (2016). *"How far has our Waqf been researched."* Jurnal Etikonomi 15:1-12.

S., S., M, H., P, K., & J., D. (2010). Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,*, 373-382.

Sami Smadi∗, Nauman Aslam∗, Li Zhang∗, Rafe Alasem† and M A Hossain‡. (2015). *Detection of Phishing Emails using Data Mining.* 9th International Conference on Software, Knowledge, Information Management and Applications.

Schmidt, E. R. (2019). *El módilo Python3 de la semana*. Obtenido de https://rico-schmidt.name/pymotw-3/imaplib/

scikit learn. (2020). Obtenido de https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html

Seta, L. D. (13 de Noviembre de 2018). Obtenido de https://dosideas.com/noticias/java/314-introduccion-a-los-servicios-web-restful

Shreeram, V. S. (2010). *Anti-phishing detection of phishing attacks using genetic algorithm. .* Ramanathapuram, India.: IEEE International Conference on Communication Control and Computing Technologies (ICCCCT).

Sonowal, G. (14 de Mayo de 2020). *Detección de correo electrónico de suplantación de identidad basada en la selección de funciones de búsqueda binaria.* India. Obtenido de Detección de correo electrónico de suplantación de identidad basada en la selección de funciones de búsqueda binaria: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7275664/

Tabares, D. S. (2015). Técnicas de detección y control de phishing. *Cuaderno Activa, 7*(1), 75-81, 1.

Timón, C. E. (2017). Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source que permiten su uso . En C. E. Timón.

Twilio Docs. (2020). Obtenido de https://www.twilio.com/docs/libraries/python

UNESCO. (2019).

Uniwebsidad. (2020). Obtenido de https://uniwebsidad.com/libros/python/capitulo-10/modulos-de-sistema

Velasquez, M. E. (2013). PHISHING. *Revista de Información, Tecnología y Sociedad*.

verizon. (2016). Obtenido de https://enterprise.verizon.com/resources/reports/dbir/2019/results-and-analysis/

Weka 3: software de aprendizaje automático en Java. (s.f.). Obtenido de

https://www.cs.waikato.ac.nz/ml/weka/

William W. Cohen, MLD, CMU. (8 de MAYO de 2015). *Conjunto de datos de correo electrónico de Enron*. Obtenido de https://www.cs.cmu.edu/~./enron/