UNIVERSIDAD AUTÓNOMA DE AGUASCALIENTES

CENTRO DE CIENCIAS BÁSICAS

DEPARTAMENTO DE SISTEMAS DE INFORMACIÓN

TRABAJO PRÁCTICO

MECANISMOS PARA LA DETECCIÓN DE ANOMALÍAS DE CIBERSEGURIDAD A TRAVÉS DEL ANÁLISIS DE GRANDES CANTIDADES DE DATOS

QUE PRESENTA L.I. PABLO ANDRÉS MARTÍNEZ VELASCO PARA OPTAR POR EL GRADO DE: MAESTRÍA EN INFORMÁTICA Y TECNOLOGÍAS COMPUTACIONALES

TUTOR

Ph.D. Luis Eduardo Bautista Villalpando

COMITÉ TUTORAL

MC. Edgar Oswaldo Díaz

Dr. Juan Muñoz López

Aguascalientes, Ags, 22 de noviembre del 2024

Autorizaciones

CARTA DE VOTO APROBATORIO
INDIVIDUAL

M. en C. Jorge Martín Alférez Chávez
DECANO (A) DEL CENTRO DE CIENCIAS BÁSICAS

P R E S E N T E

Por medio del presente como *TUTOR* designado del estudiante *PABLO ANDRÉS MARTÍNEZ VELASCO* con ID *187999* quien realizó *el trabajo práctico* titulado: ***MECANISMOS PARA LA DETECCIÓN DE ANOMALÍAS DE CIBERSEGURIDAD A TRAVÉS DEL ANÁLISIS DE GRANDES CANTIDADES DE DATOS***, un trabajo propio, innovador, relevante e inédito y con fundamento en el Artículo 175, Apartado II del Reglamento General de Docencia doy mi consentimiento de que la versión final del documento ha sido revisada y las correcciones se han incorporado apropiadamente, por lo que me permito emitir el **VOTO APROBATORIO**, para que *él* pueda proceder a imprimir*lo* así como continuar con el procedimiento administrativo para la obtención del grado.

Pongo lo anterior a su digna consideración y sin otro particular por el momento, me permito enviarle un cordial saludo.

A T E N T A M E N T E
"Se Lumen Proferre"
Aguascalientes, Ags., a 08 de noviembre de 2024.

Luis Eduardo Bautista Villalpando
Tutor de *trabajo práctico*

c.c.p.- Interesado
c.c.p.- Secretaría Técnica del Programa de Posgrado

Elaborado por: Depto. Apoyo al Posgrado.
Revisado por: Depto. Control Escolar/Depto. Gestión de Calidad.
Aprobado por: Depto. Control Escolar/ Depto. Apoyo al Posgrado.

Código: DO-SEE-FO-07
Actualización: 01
Emisión: 17/05/19

INEGI

Dirección General de Integración, Análisis e Investigación
Dirección General Adjunta de Investigación

Dirección del Laboratorio de Ciencia de Datos y Métodos Modernos de Producción de Información

Asunto: CARTA DE VOTO APROBATORIO

M. en C. Jorge Martín Alférez Chávez
DECANO (A) DEL CENTRO DE CIENCIAS BÁSICAS

P R E S E N T E

Por medio del presente como Co-Tutor designado del estudiante **PABLO ANDRÉS MARTÍNEZ VELASCO** con ID 187999 quien realizó el trabajo práctico titulado: **MECANISMOS PARA LA DETECCIÓN DE ANOMALÍAS DE CIBERSEGURIDAD A TRAVÉS DEL ANÁLISIS DE GRANDES CANTIDADES DE DATOS**, un trabajo propio, innovador, relevante e inédito y con fundamento en el Artículo 175, Apartado II del Reglamento General de Docencia doy mi consentimiento de que la versión final del documento ha sido revisada y las correcciones se han incorporado apropiadamente, por lo que me permito emitir el **VOTO APROBATORIO**, para que él pueda proceder a imprimir así como continuar con el procedimiento administrativo para la obtención del grado.

Pongo lo anterior a su digna consideración y sin otro particular por el momento, me permito enviarle un cordial saludo.

A T E N T A M E N T E

"Se Lumen Proferre"
Aguascalientes, Ags., a 08 de noviembre de 2024

Edgar Oswaldo Díaz
Co-Tutor de trabajo práctico

Subdirector de Investigación en Ciencia de Datos [D]
Dirección del Laboratorio de Ciencia de Datos y Métodos Modernos de Producción de Información
oswaldo.diaz@inegi.org.mx
Tel. (449) 9105300 ext. 312134

M. en C. Jorge Martín Alvarez Chávez
DECANO (A) DEL CENTRO DE CIENCIAS BÁSICAS

P R E S E N T E

Por medio del presente como *ASESOR* designado del estudiante *PABLO ANDRES MARTINEZ VELASCO* con ID *187999* quien realizó el *trabajo práctico* titulado: *MECANISMOS PARA LA DETECCIÓN DE ANOMALÍAS DE CIBERSEGURIDAD A TRAVÉS DEL ANÁLISIS DE GRANDES CANTIDADES DE DATOS*, un trabajo propio, innovador, relevante e inédito y con fundamento en el Artículo 175, Apartado II del Reglamento General de Docencia doy mi consentimiento de que la versión final del documento ha sido revisada y las correcciones se han incorporado apropiadamente, por lo que me permito emitir el **VOTO APROBATORIO**, para que *él* pueda proceder a imprimirlo así como continuar con el procedimiento administrativo para la obtención del grado.

Pongo lo anterior a su digna consideración y sin otro particular por el momento, me permito enviarle un cordial saludo.

A T E N T A M E N T E
"Se Lumen Proferre"
Aguascalientes, Ags., a 08 de noviembre de 2024

Juan Muñoz López
Asesor de *MECANISMOS PARA LA DETECCIÓN DE ANOMALÍAS DE CIBERSEGURIDAD A TRAVÉS DEL ANÁLISIS DE GRANDES CANTIDADES DE DATOS*

c.c.p.- Interesado
c.c.p.- Secretaría Técnica del Programa de Posgrado

DICTAMEN DE LIBERACIÓN ACADÉMICA PARA INICIAR LOS TRÁMITES DEL EXAMEN DE GRADO

Fecha de dictaminación dd/mm/aaaa: 20/11/2024

NOMBRE: PABLO ANDRÉS MARTÍNEZ VELASCO **ID** 187999

PROGRAMA: MAESTRÍA EN INFORMÁTICA Y TECNOLOGÍAS COMPUTACIONALES **LGAC (del posgrado):** INGENIERÍA DE SISTEMAS DECISIONALES PARA MEJORAR PROCESOS ORGANIZACIONALES

TIPO DE TRABAJO: ( ) Tesis (X) Trabajo Práctico

TITULO: MECANISMOS PARA LA DETECCIÓN DE ANOMALÍAS DE CIBERSEGURIDAD A TRAVÉS DEL ANÁLISIS DE GRANDES CANTIDADES DE DATOS

IMPACTO SOCIAL (señalar el impacto logrado): DESARROLLO Y DOCUMENTACIÓN DEL FLUJO DE PROCESOS PARA LA CREACIÓN DE ALGORITMOS ENTRENADOS DENTRO DEL LABORATORIO DE CIENCIA DE DATOS DE LA UAA

INDICAR SI NO N.A. (NO APLICA) SEGÚN CORRESPONDA:

	Elementos para la revisión académica del trabajo de tesis o trabajo práctico:
SI	El trabajo es congruente con las LGAC del programa de posgrado
SI	La problemática fue abordada desde un enfoque multidisciplinario
SI	Existe coherencia, continuidad y orden lógico del tema central con cada apartado
SI	Los resultados del trabajo dan respuesta a las preguntas de investigación o a la problemática que aborda
SI	Los resultados presentados en el trabajo son de gran relevancia científica, tecnologica o profesional según el área
SI	El trabajo demuestra más de una aportación original al conocimiento de su área
SI	Las aportaciones responden a los problemas prioritarios del país
SI	Generó transferencia del conocimiento o tecnológica
SI	Cumple con la ética para la investigación (reporte de la herramienta antiplagio)
	El egresado cumple con lo siguiente:
SI	Cumple con lo señalado por el Reglamento General de Docencia
SI	Cumple con los requisitos señalados en el plan de estudios (créditos curriculares, optativos, actividades complementarias,estancia, predoctoral, etc)
SI	Cuenta con los votos aprobatorios del comité tutoral, en caso de los posgrados profesionales si tiene solo tutorpodrá liberar solo el tutor
SI	Cuenta con la carta de satisfacción del Usuario
SI	Coincide con el título y objetivo registrado
SI	Tiene congruencia con cuerpos académicos
SI	Tiene el CVU del Conacyt actualizado
N.A.	Tiene el artículo aceptado o publicado y cumple con los requisitos institucionales (en caso que proceda)
	En caso de Tesis por artículos científicos publicados
N.A.	Aceptación o Publicación de los artículos según el nivel del programa
N.A.	El estudiante es el primer autor
N.A.	El autor de correspondencia es el Tutor del Núcleo Académico Básico
N.A.	En los artículos se ven reflejados los objetivos de la tesis, ya que son producto de esté trabajo de investigación.
N.A.	Los artículos integran los capítulos de la tesis y se presentan en el idioma en que fueron publicados
N.A.	La aceptación o publicación de los artículos en revistas indexadas de alto impacto

Con base a estos criterios, se autoriza se continúen con los trámites de titulación y programación del examen de grado:

Si X

No

FIRMAS

Elaboró:

* NOMBRE Y FIRMA DEL CONSEJERO SEGÚN LA LGAC DE ADSCRIPCION: DRA. MARÍA DOLORES TORRES SOTO

NOMBRE Y FIRMA DEL SECRETARIO TÉCNICO: DRA. LIZETH ITZIGUERY SOLANO ROMO

*** En caso de conflicto de intereses, firmará un revisor miembro del NAB de la LGAC correspondiente distinto al tutor o miembro del comité tutoral, asignado por el Decano**

Revisó:

NOMBRE Y FIRMA DEL SECRETARIO DE INVESTIGACIÓN Y POSGRADO: DR. ALEJANDRO PADILLA DÍAZ

Autorizó:

NOMBRE Y FIRMA DEL DECANO: M. EN C. JORGE MARTÍN ALDÉREZ CHÁVEZ

Nota: procede el trámite para el Depto. de Apoyo al Posgrado

En cumplimiento con el Art. 105C del Reglamento General de Docencia que a la letra señala entre las funciones del Consejo Académico: ... Cuidar la eficiencia terminal del programa de posgrado y el Art. 105F las funciones del Secretario Técnico, llevar el seguimiento de los alumnos.

Elaborado por: D. Apoyo al Posg
Revisado por: D. Control Escolar/D. Gestión de Calidad
Aprobado por: D. Control Escolar/ D. Apoyo al Posg

Código: DO-SEE-F0-15
Actualización: 01
Emisión: 28/04/20

Agradecimientos

Agradezco a la Universidad Autónoma de Aguascalientes y a sus profesores quienes desde el nivel bachillerato me han guiado a lo largo de mi trayectoria académica y personal, gracias a todos encontré mi pasión por aprender.

Al Instituto de Ciencia y Tecnología del Estado de Aguascalientes y su programa “Talentos que Trascienden en la Ciencia y Tecnología” que gracias a su apoyo pude finalizar esta meta.

A mi tutor el Dr. Bautista quien siempre me brindó su ayuda y me facilitó el equipo y el espacio necesario para poder continuar con este trabajo, que siempre tuvo una solución para los problemas más agobiantes que se me presentaron y que con cada nueva alternativa me mostraba más del apasionante mundo de la inteligencia artificial.

A mi cotutor el Maestro Oswaldo y mi asesor el Dr. Muñoz por su orientación, paciencia y enseñanzas en un área tan compleja y multidisciplinaria como lo es la analítica y ciencia de datos, gracias por todos los retos que me llevaron a donde estoy ahora y que sé que quiero seguir superando.

A mi familia, amigos y mi psicóloga que me apoyaron incluso en los momentos que sentía que el mundo entero caía sobre mí, que sentía que no era suficiente, gracias a ustedes estoy hoy escribiendo estos agradecimientos con felicidad y paz.

Dedicatorias

A mis hermanos Manuel y Valentina, confío en que mi ejemplo de superación los inspire a seguir sus propios sueños también. Los admiro por siempre haberme demostrado que se puede seguir adelante con una sonrisa, espero en un futuro poder ver un documento como este con sus nombres.

A mi pareja Pablo Torres, con quien las penas son más llevaderas y quien me ha dado los momentos más bellos, esta solo fue una aventura más de muchas que quiero vivir a tu lado, te amo.

A Stivi, mi gata, la más afectada de todos, quien presenció cada una de mis crisis y siempre estuvo ahí para calmarme y arañarme.

Índice General

Índice de Tablas

Índice de Gráficas y Figuras

I. Resumen

La presente tesis está enfocada en la fusión de la ciencia de datos y la ciberseguridad, con el objetivo de proponer soluciones innovadoras que fortalezcan la protección de organizaciones en México frente a amenazas cibernéticas, en este caso, enfocado en los ataques Denegación de Servicios Distribuidos (DDoS por sus siglas en ingles). En el marco de esta investigación, se realizaron diversos estudios y comparaciones sobre la metodología más adecuada para procesar bases de datos relacionadas con ciberataques, explorando diferentes enfoques de modelos de aprendizaje automático, tanto supervisados como semi-supervisados.

El trabajo comenzó con un análisis exhaustivo de las estadísticas de ciberseguridad en México, evidenciando la creciente necesidad de profesionales especializados en estas áreas para enfrentar los desafíos actuales y futuros. La metodología empleada incluyó varios intentos y ajustes a lo largo del proceso de experimentación, lo que permitió profundizar en el conocimiento de métodos efectivos de preprocesamiento de datos, normalización y estructuración de los conjuntos de datos.

Se evaluaron distintas técnicas de aprendizaje automático empleando algoritmos avanzados que fueron entrenados, evaluados y configurados cuidadosamente para evitar problemas como el sobreajuste (overfitting) y el ruido de datos corruptos. Finalmente, el modelo resultante fue capaz de clasificar eficientemente registros de ataques DDoS y logs normales con una alta precisión, demostrando su potencial para implementarse en infraestructuras tecnológicas con el fin de fortalecer la ciberseguridad de organizaciones evitando la fuga de datos e inhabilitación de servidores.

II. Abstract

This thesis is focused on the merge of data science and cybersecurity, this with the aim of proposing innovative solutions that strengthen the protection of organizations in Mexico against cyber threats, for this case focused on DDoS (Distributed Denial of Service) attacks. Within the framework of this research, various studies and comparisons were carried out by means of the most appropriate methodologies to process datasets related to cyberattacks, exploring different approaches to machine learning models, both supervised and semi-supervised.

The work began with an exhaustive analysis of cybersecurity statistics in Mexico, evidencing the growing need for specialized professionals in these areas to face current and future challenges. The methodologies used included several attempts and adjustments throughout the experimentation process that allow deepening knowledge of effective methods of data preprocessing, normalization and structuring of databases.

Different machine learning techniques were tried using advanced algorithms that were carefully trained, evaluated, and tuned to avoid problems such as overfitting and corrupted data. Finally, the resulting models were able to efficiently classify DDoS attack logs and normal logs with high accuracy, demonstrating their potential to be implemented into technological infrastructures to strengthen the cybersecurity of organizations, preventing data leakage and server crashes.

III. Introducción

¿Por qué es necesaria la Ciencia de Datos/ Análisis de datos en Ciberseguridad?

La generación de información digital en las décadas del 2010-2020's se ha incrementado a niveles acelerados. Tan solo en 2018 se generaban 2.5 quintillones de bytes en datos diario, aumentando cada vez más con el uso de la tecnología de IoT (Marr, 2018). En el año 2018 el tamaño de la esfera de datos alcanzó los 18 zettabytes y se espera que para el 2025 se alcance un tamaño de 175 zettabits (Marr, 2021).

En este contexto las técnicas clásicas para procesar datos (datos que en su mayoría son estadísticos de actividad en línea, personales anónimos o de información sensible) son insuficientes para cubrir la demanda de procesamiento en tiempo y forma (Alani, 2021), lo cual ha dado pie a la creación de nuevas y más especializadas técnicas de gestión de datos como la ciencia de datos (Data Science por su traducción al inglés).

Los ciberataques en infraestructuras críticas se encontraban en el puesto número 5 dentro del "World Economic Forum's Global Risk Report" de 2020 (World Economic Forum, 2020), y aunque en el año 2023 su urgencia fue desplazada por cuestiones ambientales no deja de ser uno de los factores críticos con el puesto 8 al corto y largo plazo considerándose como el principal tema de interés dentro del área de tecnología (World Economic Forum, 2023).

"La cibercriminalidad ha costado más de 6 billones de dólares a la economía del mundo" (Reuters, 2022).

Adicionalmente hay que considerar que muchas de estas formas de ciberataque evolucionan día con día, y los fraudes e infiltraciones se vuelven más elaborados utilizando información personal sensible filtrada en internet o dentro de las redes sociales.

Capítulo 1 Ciencia de datos

El término ciencia de datos se acuñó por primera vez por Peter Naur en 1960 únicamente usado como sinónimo de ciencias computacionales. En 1974 se le relacionó con los métodos de procesamiento de datos por Naur y finalmente en 1996 apareció en un artículo público por la Federación Internacional de Sociedades de Clasificación (The International Federation of Classification Societies). Desde entonces se adoptó este concepto de forma internacional para describir a este campo interdisciplinario (Vicario & Coleman, 2019).

La ciencia de datos se define como un campo de las ciencias computacionales que se encarga del procesamiento de datos estructurados y no estructurados para un proceso de toma de decisiones inteligentes con base en la información disponible. Incluye todo el proceso de preparación de los datos (Monnappa, 2022) desde:

- Planteamiento del problema.
- Adquisición de datos.
- Preparación de datos.
- Análisis exploratorio.
- Modelado de datos.
- Visualización y comunicación.
- Implantación y mantenimiento.

Además, comúnmente abarca áreas de la Inteligencia Artificial (Artificial Intelligence por su traducción al inglés) que se refiere a un software programado para emular la inteligencia y el aprendizaje humano, como el Aprendizaje Automático (Machine Learning por su traducción al inglés) y la Minería de Datos (Data Mining por su traducción al inglés).

1.1 Aprendizaje automático

Por la dificultad del desarrollo de sistemas expertos en los años 80, se optó por algoritmos que aprendieran de los datos generando un modelo con el que se podrían predecir las entradas futuras, entre más grande el set de datos que alimente al algoritmo, más exacto será. Así es como nació el aprendizaje automático (Rebala et al., 2019).

El aprendizaje automático es una rama de la inteligencia artificial y las ciencias computacionales que se enfoca en el uso de datos y algoritmos para automatizar soluciones a problemas complejos difíciles de solucionar con métodos de programación convencionales (Rebala et al., 2019).

El aprendizaje automático es usado en predicción y prevención de varios fenómenos, unos ejemplos serían la predicción de incendios (Surya, 2017),

predicción del mercado de valores y finanzas (Sai et al., 2008) y detección de fraude en comercios digitales (Nanduri et al., 2020).

1.2 Minería de datos

La minería de datos es el proceso iterativo de descubrir patrones entre otras correlaciones de interés dentro de conjuntos de datos de gran tamaño, convirtiendo datos crudos en conocimiento útil (Kantardzic, 2002).

1.3 Análisis de datos

Por otro lado, identificamos un área muy relacionada con la ciencia de datos, que es el análisis de datos, la cual se encarga de aplicar y/o automatizar algoritmos que den como resultados conjuntos de datos listos para examinar y buscar correlaciones, permitiendo a las organizaciones tomar decisiones más acertadas, actualizadas y verificar o refutar modelos. El análisis de datos se enfoca en la inferencia y generar conclusiones en base a la información con la que se cuenta (Monnappa, 2022).

Estas nuevas tecnologías y herramientas han impulsado mejoras en los procesos de negocio de las organizaciones en diversas áreas tales como; la industria financiera, el área médica y la ciberseguridad.

Capítulo 2 Ciberseguridad

Para definir lo que es la ciberseguridad, también es importante diferenciarla de muchos otros conceptos parecidos y que comúnmente se usan indistintamente (Alani, 2021; Cebula et al., 2014).

- Ciberseguridad es la práctica de defender equipos computacionales (computadoras, servidores, celulares, redes, entre otros) de ataques maliciosos, así como asegurar la integridad, confidencialidad y disponibilidad de la información que contengan. Este concepto incluye la seguridad de la persona (usuario) misma considerado como un objeto más de ataque y de riesgo.
- La seguridad de la información se refiere a la protección de los recursos de información y la garantía de su privacidad. Normalmente los humanos toman parte activa de este proceso.
- La seguridad de la red se relaciona a las metodologías como políticas, procedimientos y controles usados para proteger los activos dentro de la red (información almacenada, en movimiento entre equipos, software e incluso hardware).

- Los Riesgos Cibernéticos (Cyber Risks por su traducción al inglés) se definen como “riesgos operacionales en los activos de información y tecnología que como consecuencia afectan la confidencialidad, disponibilidad, y/o integridad de la información dentro de los sistemas de información”

Los ciberataques (Cyberattacks por su traducción al inglés) son considerados como un acto que amenaza las tres características esenciales de la información y los sistemas que las gestionan; la confidencialidad, integridad y la disponibilidad. Existen muchos tipos de incidentes de ciberseguridad que pueden amenazar la información, y, por lo tanto, ser ciberataques, unos de los más conocidos son (Sun et al., 2019):

- Acceso no autorizado a sistemas o áreas restringidas.
- Programas malignos.
- Ataque de negación de servicio normal y distribuida o DoS (Denial of Service) y DDoS (Distributed Denial of Service).
- Ingeniería social (Phishing, Vishing, Tailgating, Scareware, Search Engine poisoning, spam, etc.) (Ivaturi & Janczewski, 2011).
- Ataques de día cero (Zero-day attack por su traducción al inglés).

Estos conceptos no solo están relacionados a la tecnología, sino que caracterizan al ser humano ya sea como un factor de riesgo, es decir, un recurso a proteger junto con la tecnología. Considerando lo anterior, la ciberseguridad es entonces un conjunto de metodologías que incluyen desde el diseño de redes y software hasta el comportamiento humano.

En nuestra era moderna digital, el problema de la ciberseguridad ya no solo está relegado a aquellos expertos en ciencias computacionales, sino que es un asunto que engloba a todos (Singer & Friedman, 2014).

Por eso se necesitan desarrollar mecanismos de seguridad más flexibles, eficientes y dinámicos que puedan hacer frente a las amenazas y que puedan cambiar sus políticas automáticamente para reaccionar rápida y efectivamente a amenazas conocidas y no tan conocidas. Es aquí donde la ciencia de datos, el aprendizaje automático y la inteligencia artificial pueden jugar un rol vital en el próximo paso de ciberseguridad, que es, la ciencia de datos de ciberseguridad (Sarker et al., 2020).

La ciencia de datos de ciberseguridad (Cybersecurity Data Science por su traducción al inglés) se refiere a la colección y análisis de grandes cantidades de eventos de seguridad de diferentes fuentes usando tecnologías de aprendizaje automático con el fin de detectar riesgos de seguridad dentro de un sistema o una organización, así como el descubrimiento de nuevas perspectivas y patrones, todo esto a la vez que se optimizan los procesos de operaciones.

En la tabla siguiente (Fig. 1) se muestran las etapas de análisis y modelos de respuesta en ciencia de datos según el enfoque (Longbing & Philip S, 2018)

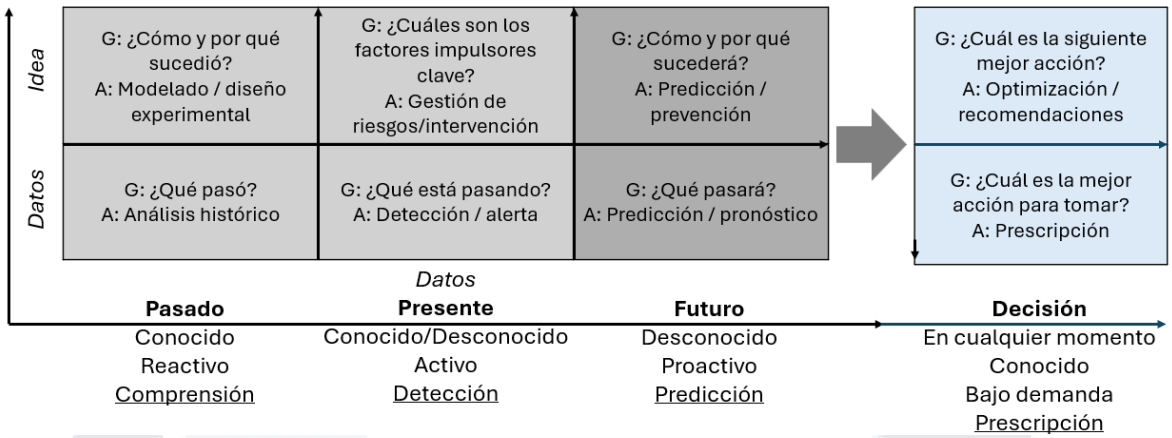


Figura 1. Análisis de toda la vida de decisión desde los datos hasta el conocimiento.

(Tomada de (Longbing & Philip S, 2018, p. 237))

Algunas de las principales soluciones de las cuales han surgido más estudios en los últimos años (más específicamente desde 2014 y teniendo un notorio incremento desde el año 2018) utilizando esta nueva técnica de seguridad (Alani, 2021) son:

- Detección y bloqueo de anomalías e infiltraciones.
- Detección y erradicación de spam, spoofing y phishing.
- Detección temprana de programas malignos y secuestro de datos (ransomware por su traducción al inglés).
- Seguridad en el código de software al momento de programar para evitar clonación de código inseguro con brechas de seguridad conocidas.
- Seguridad en la nube.

Para finalizar, es necesario mencionar futuras direcciones que se presentan como retos para los próximos académicos e ingenieros que se adentren en esta área. Algo que muchos autores omiten es que la ciencia de datos como herramienta puede utilizarse para mejorar o para burlar la ciberseguridad, si bien sería un concepto sin ahondar se debe de considerar la existencia de Ciencia de Datos en Ciberataques (Data Science Cyberattacks por su traducción al inglés) como una técnica igual de importante para evitar sistemas como malware (programa maligno) dinámico, y con respuesta en tiempo real según el entorno o programas que dejen inservible a una IA.

2.1 Desaprendizaje automático

El Desaprendizaje Automático (Machine Unlearning por su traducción al inglés) acuñado por Cao y Yang (Y. Cao & Yang, 2015), se desarrolló como una respuesta al riesgo que representan los ciberataques y como uno de los primeros intentos de respuesta a los ataques contra máquinas virtuales e inteligencia artificial en el área de seguridad informática. Esta consiste en que dentro de un sistema que cuenta con medidas de seguridad mejoradas mediante aprendizaje automático, un atacante con este conocimiento orquestará ataques estratégicos para "enseñarle" erróneamente a los algoritmos de seguridad y así facilitar que en ataques reales las defensas no respondan adecuadamente, el desaprendizaje automático "borrará" esas enseñanzas falsas propiciando algoritmos más robustos.

2.2 Aspectos relacionados con la obtención de datos

La escasez de bases de datos confiables y gratuitas en internet, junto con la reticencia de grandes empresas (como hospitales, consultoras, bancos y entidades gubernamentales) a compartir sus registros internos de incidentes de ciberseguridad, limita el acceso a datos útiles para el análisis de amenazas (Cremer et al., 2022).

Es necesario que al obtener bases de datos relacionados a la ciberseguridad de organizaciones se proceda con cautela, discreción y anonimato ya que se trata de información sensible que como se mencionó anteriormente puede ser usada para fines de mejora de la protección de la infraestructura tecnológica y también para encontrar áreas vulnerables u oportunidades de ataques futuras, además de que contienen información privada de usuarios internos y externos.

2.3 Ciberseguridad en México

La seguridad cibernética de México es de 37.66 puntos de 100 y ubicando a México en la posición 92 de 172 a nivel mundial en 2023 (antes el puesto 84 en 2022) por debajo de países como Pakistán, Jamaica y una gran parte de países sudamericanos como Uruguay, Panamá y Costa Rica. (National Cyber Security Index, 2023b) Dentro de este ranking los aspectos a evaluar y su respectiva calificación en México son:

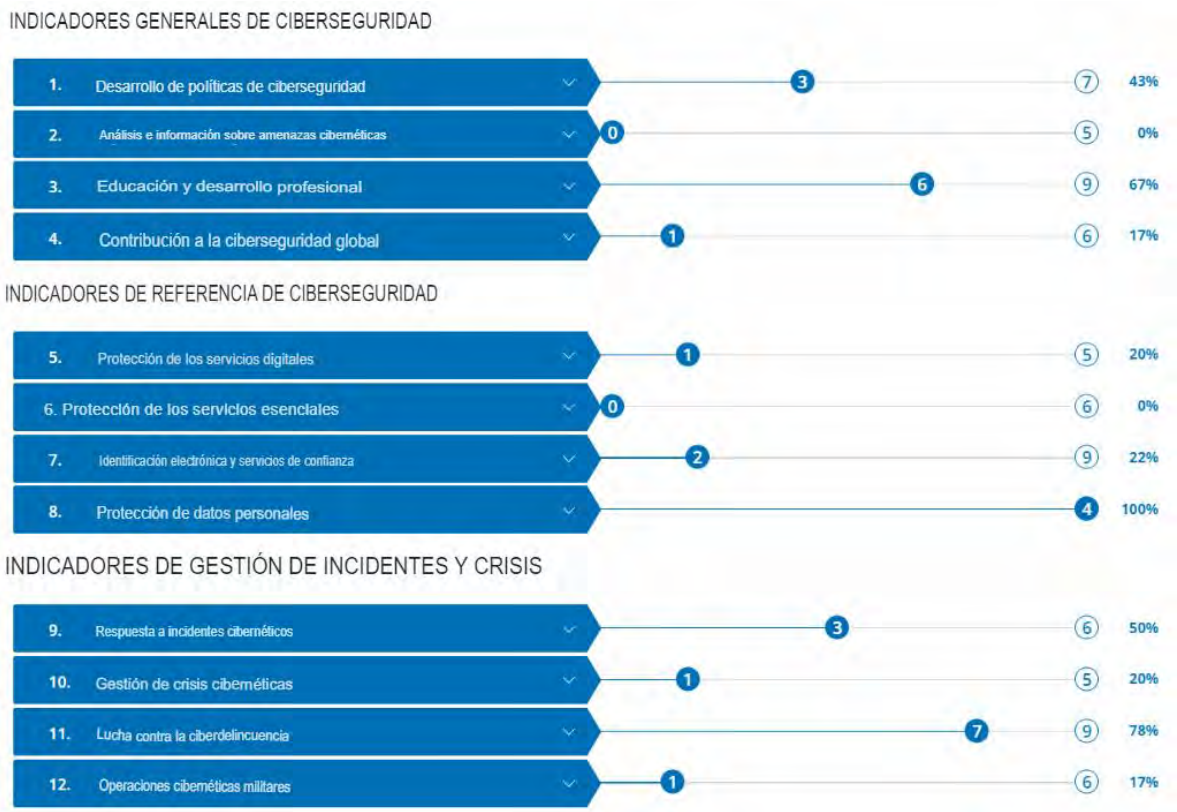


Figura 2. NCSI Porcentaje de cumplimiento en México.

(Tomado de (National Cyber Security Index, 2023a))

Actualmente, se vive una crisis de falta de profesionistas en ciberseguridad en nuestro país, esto cobra un especial riesgo dentro de las empresas gubernamentales. Uno de los principales orígenes es la falta de inversión tanto en educación como en infraestructura relacionada a la protección de activos informáticos como de prevención de ataques cibernéticos (Jabbour, 2022).

"La ola de ataques ransomware que se suscitó en 2021 registró un crecimiento del 619% en México, cuatro veces más que lo detectado en 2020, además de que estaba planificada" (Guarneros, 2022). Sin tener que remontarnos al pasado lejano descubriremos que han comenzado a suscitarse eventos con magnitudes cada vez más graves dentro de empresas gubernamentales como Pemex (Chávez et al., 2019) Lotería Nacional (Guillen, 2021) y la Sedena (Paredes, 2022).

Dadas las cifras anteriores el uso de técnicas de ciencias de datos en el ámbito de la ciberseguridad en México es una importante área de oportunidad para mejorar acciones preventivas y correctivas en la seguridad de la información sobre todo en organizaciones que son el objetivo de ataques cibernéticos en la industria pública y privada.

Capítulo 3 Planteamiento del problema

El objetivo principal de ese trabajo práctico es renovar y cubrir posibles vulnerabilidades actuales revisando trabajos de ciberseguridad anteriores y datos recabados en el departamento de centros de datos de la organización, automatizando y haciendo más eficiente el proceso de acción frente a ciberataques.

Capítulo 3.1 Objetivos de la intervención

3.1.1 Objetivos específicos

- Identificar riesgos actuales en la seguridad de la información dentro del Laboratorio de ciencia de datos en el Instituto Nacional de Estadística y Geografía (INEGI).
- Diseñar un instrumento de medición estadístico para detectar posibles ataques orientados a tecnologías web.
- Identificar dominios y segmentos de red por internet maliciosos que realicen ciber ataques.
- Automatizar estrategias de acción frente a ciberataques.

3.1.2 Preguntas de investigación

- ¿Qué áreas/aspectos son los que poseen más riesgo dentro del sistema de información?
- ¿Cómo se puede aplicar ciencia de datos en ciberseguridad para mejorar los procesos de detección y prevención de ataques?

3.1.3 Justificación y análisis de viabilidad

- Justificación:
    - Reforzar la seguridad de la información de una empresa gubernamental.
- Viabilidad:
    - Financiero: No se requiere inversión dado que la infraestructura y recursos ya existen.
    - Humanos: Un área especializada para estas tareas ya existe dentro de INEGI.
    - Materiales: Dado que es manejo de datos y estadística no hay necesidad de materiales físicos, en cuando a los datos en si estos ya están recabados.

- Temporales: Existen proyectos anteriores de maestría llevados a cabo en este periodo de tiempo, hay experiencia empírica de que el tiempo bien administrado es suficiente.

Capítulo 3.2 Plan previo de evaluación de resultados

Dado el problema de investigación, la población será única, se realizará una sola evaluación y el principal rasgo es la satisfacción. Se tomará un punto inicial a partir de una encuesta y después se verá el progreso con una encuesta final tras la implementación.

3.2.1 Aspectos a evaluar

- Desempeño.
- Usabilidad.
- Optimización del proceso.
- Utilidad de la información presentada.

3.2.2 Instrumento

- Doble, uno de diagnóstico y otro de evaluación final
  - Diagnóstico: Evaluación del modelo mediante división de registros para entrenamiento (70%) y para revisión de precisión (30%) utilizando el grupo de datos inicial y datos sintéticos generados.
  - Evaluación final: Revisión de la satisfacción de los stakeholders conforme al modelo, la documentación y los descubrimientos.

Capítulo 3.3 Solución propuesta

En base a en trabajos anteriores realizados en la misma área (Ostos Ríos et al., 2020) se busca recuperar los registros de los logs de los servidores.

En la primera etapa analizaremos datos sintéticos facilitados por el departamento de ciencia de datos de INEGI y usaremos el lenguaje de programación Python para transformarlos del formato inicial a un documento con formato csv (texto con valores separados por comas), con el que posteriormente podremos empezar a utilizar como un conjunto de datos de entrenamiento y de pruebas que permitirá desarrollar el algoritmo que planeamos implementar para la toma de decisiones logrando así automatizar las tareas de detección de ataques y amenazas del departamento de seguridad.

3.3.1 Definición de tipo de estudio, investigación

Descriptivo-Correlacional

No es la primera vez que se explora el terreno y se está familiarizado con conceptos y tecnologías y ya se cuenta con información para establecer correlaciones. Dichas correlaciones van a ser identificadas y utilizadas en la creación del software entrenado para identificar ataques.

Capítulo 4 Fundamentación teórica

4.1 LibreOffice

LibreOffice es un software libre creado por “The Document Foundation” una organización sin fines de lucro e impulsado por su comunidad que cree en el Software Libre según los principios de su Manifiesto de la Próxima Década. Es uno del software de ofimática más usados, robustos, de código abierto y compatible con otros formatos de procesador de texto, herramienta de creación de presentaciones y hoja de cálculo (LibreOffice, 2024).

4.2 Lenguaje python

Python es un lenguaje orientado a objetos de alto nivel con semántica dinámica. Sus estructuras de datos integradas de alto nivel junto a su escritura y vinculación dinámica la hacen muy atractiva para el Rapid Application Development (Desarrollo Rápido de Aplicaciones), asimismo lo convierte en uno de los preferidos para el manejo de grandes cantidades de datos y unión con bases de datos en diferentes formatos como Excel, SQL e incluso en la web (Python, 2022).

4.3 Apache superset

Es una plataforma de exploración y visualización de datos de código abierto con una interfaz de creación de dashboards completa y rápida, un editor SQL para la gestión de la información con soporte para varios formatos de bases de datos en el mercado, uso de API’s y arquitectura en la nube (Apache Software Foundation, 2024).

4.4 Metodología

Para implementar una metodología enfocada en la detección de anomalías de ciberseguridad mediante el análisis de grandes cantidades de datos, fueron realizadas las siguientes fases (Villasenor Garcia et al., 2022):

- Fase fuente de datos: Identificar y seleccionar las fuentes de datos relevantes para el análisis. Estas fuentes pueden incluir logs de servidores, registros de red, datos de sistemas de detección de intrusiones (IDS), tráfico de red, y otras fuentes relacionadas con la ciberseguridad. Su importancia es garantizar que los datos sean confiables, relevantes y suficientemente diversos para detectar comportamientos anómalos.
- Fase de extracción y carga de datos: Extraer los datos seleccionados de las fuentes identificadas y cargarlos en un entorno de almacenamiento adecuado (como un data warehouse o un lago de datos). Esta fase incluye la limpieza y preprocesamiento básico de los datos para eliminar inconsistencias o valores erróneos. Su importancia es asegurar la calidad y la accesibilidad de los datos para el análisis posterior.
- Fase de recuperación de información: Realizar consultas y exploración de los datos almacenados para extraer conjuntos específicos que se utilizarán en el análisis. Esta fase también puede incluir la agregación y transformación de datos para adecuarlos al modelo de análisis. Su importancia es preparar los datos en un formato que permita un análisis eficiente y efectivo, optimizando la relevancia de la información.
- Fase de procesamiento de datos: Aplicar técnicas de procesamiento de datos para detectar patrones, tendencias y posibles anomalías. Esto puede incluir la normalización, reducción de dimensionalidad, y técnicas de análisis exploratorio de datos. Su importancia es transformar los datos brutos en un estado donde los algoritmos de detección de anomalías puedan trabajar de manera óptima.
- Fase de construcción del modelo: Desarrollar y entrenar modelos de detección de anomalías utilizando técnicas de aprendizaje automático, como modelos supervisados, no supervisados, o híbridos. Se seleccionan algoritmos específicos que se adapten a la naturaleza de los datos y el tipo de anomalías que se desean detectar. Su importancia es la de crear un modelo capaz de identificar comportamientos anómalos y potenciales amenazas de ciberseguridad.
- Fase de evaluación y validación de Resultados: Probar el modelo construido utilizando un conjunto de datos de prueba o validación para evaluar su desempeño en términos de precisión, recall, F1 score, etc. También se realizan ajustes y optimizaciones si es necesario. Su importancia es la de validar que el modelo sea robusto y confiable antes de su implementación en un entorno de producción.

- Fase de presentación de resultados: Comunicar los hallazgos del análisis a través de informes, dashboards, y visualizaciones que resuman las anomalías detectadas y su posible impacto en la ciberseguridad de la organización. Su Importancia es la de asegurar que los resultados sean comprensibles y útiles para los stakeholders, facilitando la toma de decisiones.
- Fase de entrega de productos de Datos: Proporcionar los modelos entrenados, scripts, documentación, y cualquier otro producto de datos generado durante el proyecto a los equipos encargados de la implementación y monitoreo continuo. Su importancia es la de garantizar una transferencia de conocimientos y herramientas que permita la operación continua del sistema de detección de anomalías en ciberseguridad.

4.5 Datos sintéticos

Los datos sintéticos se usan en el área de la IA para el entramiento de modelos ya que apoyan en la etapa de evaluación y generación de datos, esto con el fin de que cumplan los criterios necesarios para el entrenamiento de algoritmos y se asegure la privacidad de los datos originales.

Estos datos se generan aleatoriamente según los datos originales y siguiendo su estructura, actualmente hay varios programas y paquetes de software que permiten generar datos sintéticos, aunque para generar datos significativos se ha comenzado a utilizar la IA y el aprendizaje automático (Becerra Pozas, 2023).

Específicamente para este trabajo práctico, fue significativa la implementación de datos sintéticos para generar los registros suficientes y asegurar la calidad de los datos que contribuyeran en la generación de algoritmos entrenados satisfactoriamente.

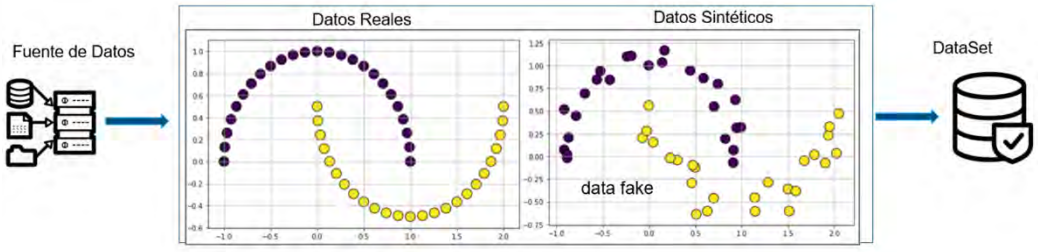


Figura 3. Creación de datos sintéticos.

4.6 Metodología de extracción, transformación y carga (ETL)

Para el manejo de datos, utilizamos la metodología ETL, la cual permite obtener repositorios de datos para su procesamiento y análisis. Esta metodología define un proceso para extraer datos desde fuentes no optimizadas para análisis.

Su abreviación implica un proceso de 3 fases que no siempre incluye otras cuestiones como el transporte de datos o los cambios que las nuevas tecnologías implican en el proceso (Miller, 2022).

Sus ventajas son (Jhawar & Tejada, 2022):

- Coordinar el trabajo desde varias fuentes de información.
- Transformar y canalizar los datos.
- Ejecución paralela de las 3 fases del proceso para ahorrar tiempo.
- Poca necesidad de procesamiento para el sistema de destino (a diferencia del ELT.

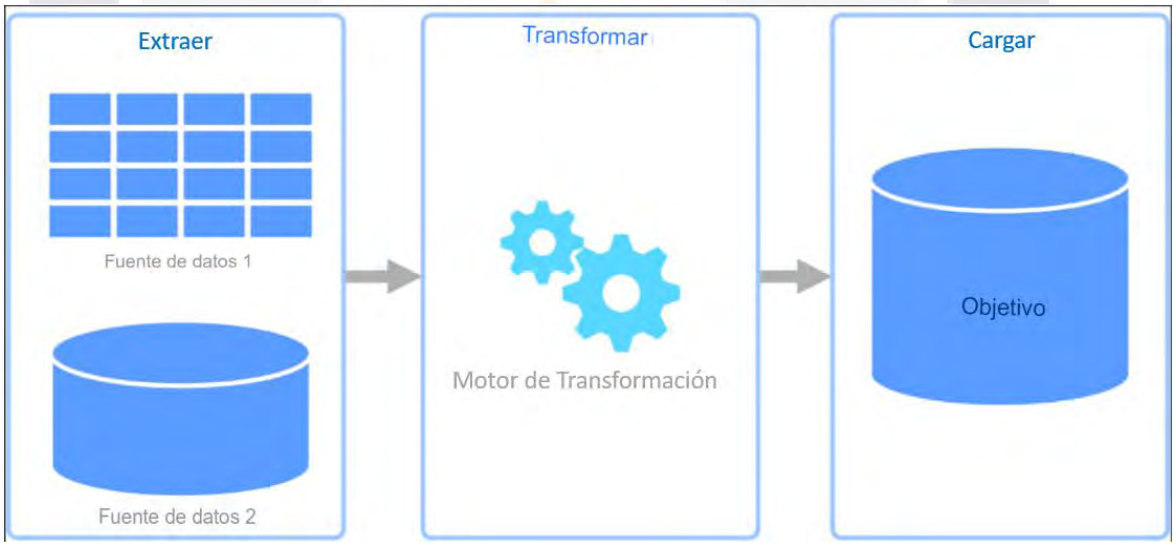


Figura 4. Proceso ETL.

(Tomado de (Jhawar & Tejada, 2022))

En este caso práctico la metodología ETL se llevó a cabo de la siguiente manera:

1. Recuperación de los datos en bruto directo de los servidores dedicados (Data Lake).
2. Limpieza y transformación de las bases de datos de forma semiautomática utilizando herramientas de hojas de cálculo y programas de Python.
3. Integración de datos con ayuda de Python.
4. Almacenamiento de todas las bases de datos procesadas para su posterior uso (Data Warehouse).

5. Visualización de la información para un análisis exploratorio y búsqueda de patrones.
6. Uso de las bases de datos para el entrenamiento de algoritmos entrenados.
7. Análisis de resultados y de ser necesario se realiza una nueva iteración desde el punto 2.

4.7 Aprendizaje supervisado

Es un tipo de modelo predictivo para el aprendizaje automático en el que se conocen los resultados de salida. El algoritmo aprende el comportamiento de los datos en base a los resultados esperados y se ajustan sus parámetros internos hasta lograr una predicción satisfactoria (Juan José Beunza et al., 2020).

Normalmente se divide un grupo de datos en dos conjuntos, uno para el entrenamiento y otro para la evaluación o testing. El algoritmo es entrenado con el primer conjunto considerando alrededor del 70% de la información y posteriormente se evalúa su precisión clasificando el segundo conjunto que es alrededor de un 30%, se comparan los resultados predichos contra la etiqueta real de los datos y se obtiene información del algoritmo como el porcentaje de predicción o tablas como la matriz de confusión.

4.8 Aprendizaje semi-supervisado

También llamados “Algoritmos supervisados de clasificación” son aquellos que contando con una pequeña muestra clasificada de los datos realiza un modelo predictivo para clasificar una cantidad masiva de nuevos datos sin etiquetar con un porcentaje esperado de confianza (Juan José Beunza et al., 2020).

El enfoque basado en el auto entrenamiento o *self training* integra la recursividad al proceso de clasificación del aprendizaje semi supervisado. Una vez obtenidos nuevos datos etiquetados, en base al porcentaje de confianza de dicha clasificación se integra una parte menor de los datos (aquellos con un porcentaje de confianza alto) al grupo de datos con la que se entrena el modelo. El proceso continúa iterando las veces necesarias hasta conseguir una cantidad aceptable de datos etiquetados con una confianza alta (Rosenberg et al., 2004).

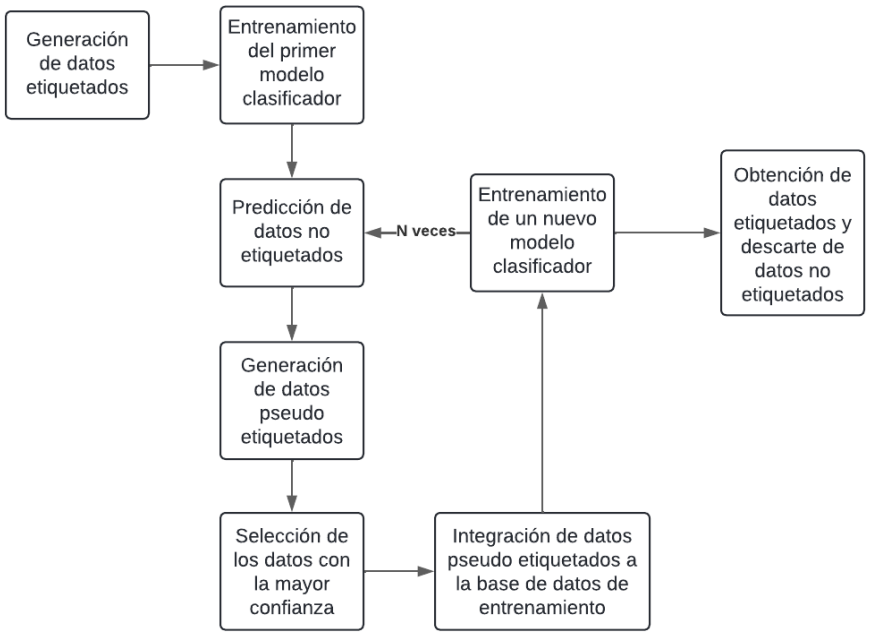


Figura 5. Proceso iterativo del *Self Training*.

4.9 Elasticsearch y kibana

Kibana fue lanzado por Elasticsearch en 2014 y que a lo largo de los años se ha posicionado como una de las herramientas de visualización y análisis de datos más populares por su facilidad de uso y flexibilidad. Acepta el procesamiento de datos NoSQL en cantidades consideradas como Big Data a la vez que dentro de toda la interfaz Elasticsearch es posible administrar los flujos y reglas de todo el procesamiento de los datos (Elasticsearch B.V., 2023).

4.10 Apache Spark

Apache Spark es un motor unificado de analíticas para procesar datos a gran escala que integra módulos para SQL, streaming, aprendizaje automático y procesamiento de grafos. Spark se puede ejecutar de forma independiente o en Apache Hadoop, Apache Mesos, Kubernetes, la nube y distintas fuentes de datos (Apache Software Foundation, 2023).

4.11 Apache Hadoop

La biblioteca (*library)* de software Apache Hadoop es un marco que permite el procesamiento distribuido de grandes conjuntos de datos en grupos de computadoras utilizando modelos de programación simples. Está diseñado para

escalar desde servidores individuales a miles de máquinas, cada una de las cuales ofrece computación y almacenamiento locales.

En lugar de depender del hardware para brindar alta disponibilidad, la biblioteca en sí está diseñada para detectar y manejar fallas en la capa de la aplicación, por lo que brinda un servicio de alta disponibilidad sobre un grupo de computadoras, cada una de las cuales puede ser propensa a fallas.

El uso de esta herramienta se busca para una etapa posterior del proyecto, si bien no se verá usado explícitamente en el documento se menciona en la documentación para reflejar la intención de integrarlo.

4.12 Jupyter notebook

Los Jupyter notebook (Cuadernos Jupyter en español) es una aplicación web de código abierto creada por el Proyecto Jupyter una organización sin ánimo de lucro.

Esta es una aplicación web creada para la elaboración y presentación de documentos híbridos entre código y reporte. Combina cuadros de código y contenido multimedia para la documentación de proyectos en su mayoría enfocados a procesamiento de datos como Python, R, Julia, Scala, entre otros (Proyect Jupyter, 2023).

4.13 Google Colaboratory

Colaboratory o Colab es un entorno que permite programar, documentar y ejecutar en el lenguaje Python sin necesidad de preconfigurar y con la ventaja de poder compartir los trabajos de manera rápida y sencilla.

El formato de notebooks (cuadernos en español) permite combinar código, texto y multimedia en un solo documento y permite la edición conjunta de varias personas como los documentos en la nube. Es un cuaderno de Jupyter alojado en el entorno Colab (Google, 2023).

4.14 Tar

El formato .tar se usa principalmente en entornos UNIX y su nombre proviene de la abreviatura de “Tape Archiver” en español “Archivador de Cinta”.

Dentro de este formato se almacenan conjuntos de archivos en uno solo sin necesidad de comprimirlos (aunque para algunos softwares si se detecte como archivos comprimidos) y gracias a esto se pueden transferir por internet fácilmente.

Para este problema se utiliza la librería “tarfile” de Python que nos permite leer, recorrer y manipular archivos .tar de forma sencilla y práctica (Python Software Foundation, 2023).

4.15 Normalización de datos

La normalización se entiende como el ajuste de los rangos de valores de diferentes atributos entre si mediante la estandarización de sus valores límites superiores e inferiores. Los datos se normalizaron para evitar que los valores muy bajos o altos de algunos atributos afectaran a la calidad del algoritmo.

Capítulo 5 Diseño de la intervención

Dentro de la organización se cuenta con el esquema **AIOps** (Inteligencia artificial para operaciones de TI) el cual busca integrar la Inteligencia Artificial y el uso de Big Data para que mediante datos históricos recabados se analice y determinen patrones y así se entrenen modelos con aprendizaje automático.

También buscamos sustituir herramientas de operaciones de TI manuales e independientes en una sola plataforma inteligente que cierre la brecha entre el complejo entorno de TI y las expectativas de los usuarios.

Se cuenta con la política “**Zero-Trust Security**” la cual busca recabar toda la información pertinente de identidades, dispositivos, datos, aplicaciones, infraestructura y redes dentro de la organización para la toma de decisiones en seguridad.

En la siguiente imagen podemos apreciar el flujo con el cuál se decidió el tipo de datos a utilizar para entrenar y evaluar al algoritmo, se concluyó que se utilizarían Archivos ficticios, en este caso se les llamarán “Datos Sintéticos” (*Synthetic Data for Official Statistics*, 2023).

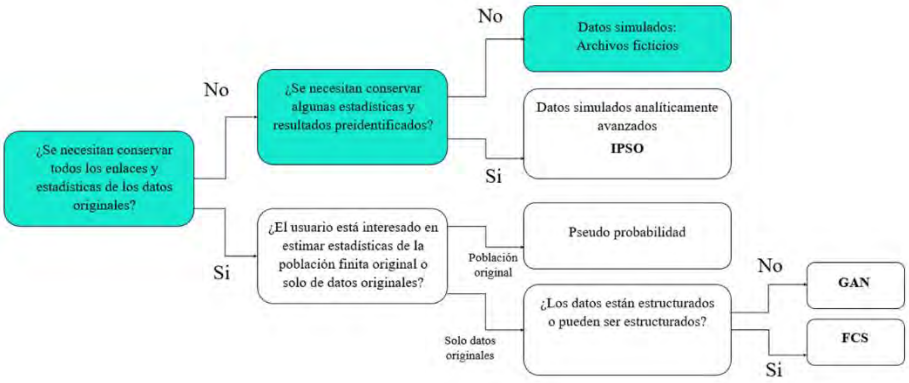


Figura 6. Árbol de Decisión de los datos a utilizar.

Instalamos Elasticsearch dentro de la infraestructura Mictlán, un clúster de servidores dentro del Laboratorio de Ciencia de Datos & Ciberseguridad de la Universidad Autónoma de Aguascalientes. Esto mejora enormemente la capacidad y el tiempo de procesamiento de grandes cantidades de datos.

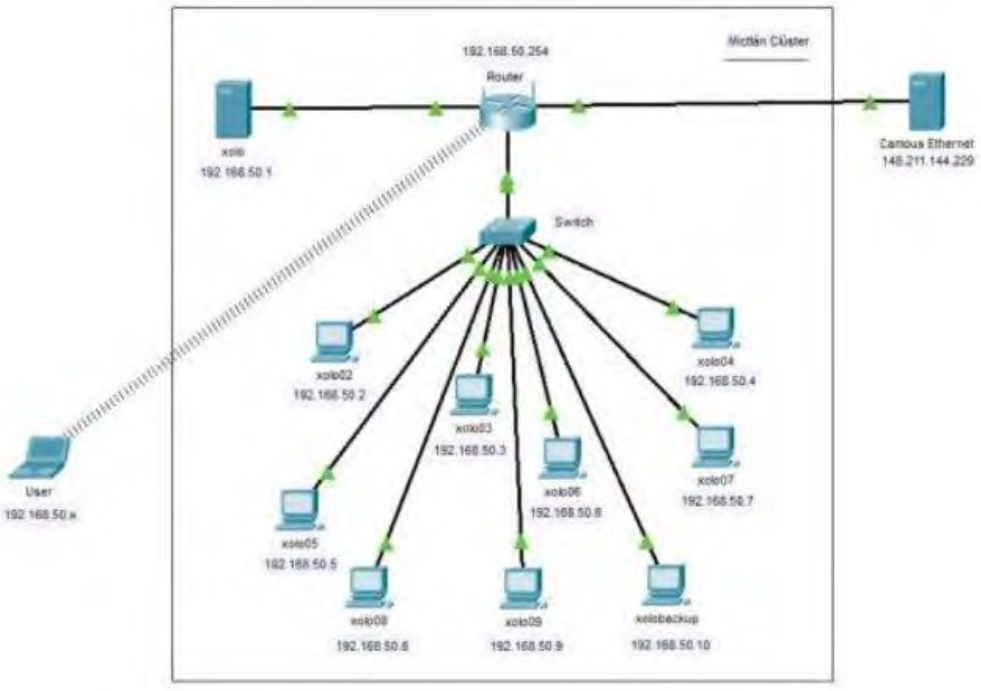


Figura 7. Gráfico de la estructura de Mictlan.

La descripción de la infraestructura del clúster Mictlan es:

- Tres servidores Sun Ultra 20 (nodos del clúster).
- Siete servidores Dell OptiPlex 3020-M (nodos del clúster).
- Switch o conmutador (interconexiones físicas entre los nodos y router).
- Router modelo RT-AC1200 (suministro de red Wi-Fi y conectividad a Internet).

El software que incluye es:

- Elasticsearch (servidor de búsqueda).
- Kibana (interfaz gráfica para Elasticsearch).
- Docker (contenerización de servicios).
- Servicios SSH y FTP (terminales remotas y transferencia de archivos).
- Debian GNU/Linux (sistema operativo base para los nodos del clúster).
- Plataforma de servicios de analítica y ciencia de datos.

Capítulo 6 Resultados de la intervención

6.1 Análisis del grupo de datos número uno

El primer grupo de datos que analizamos fue uno facilitado por INEGI, generado a partir de datos del departamento de ciencia de datos.

6.1.1 Transformación de los datos

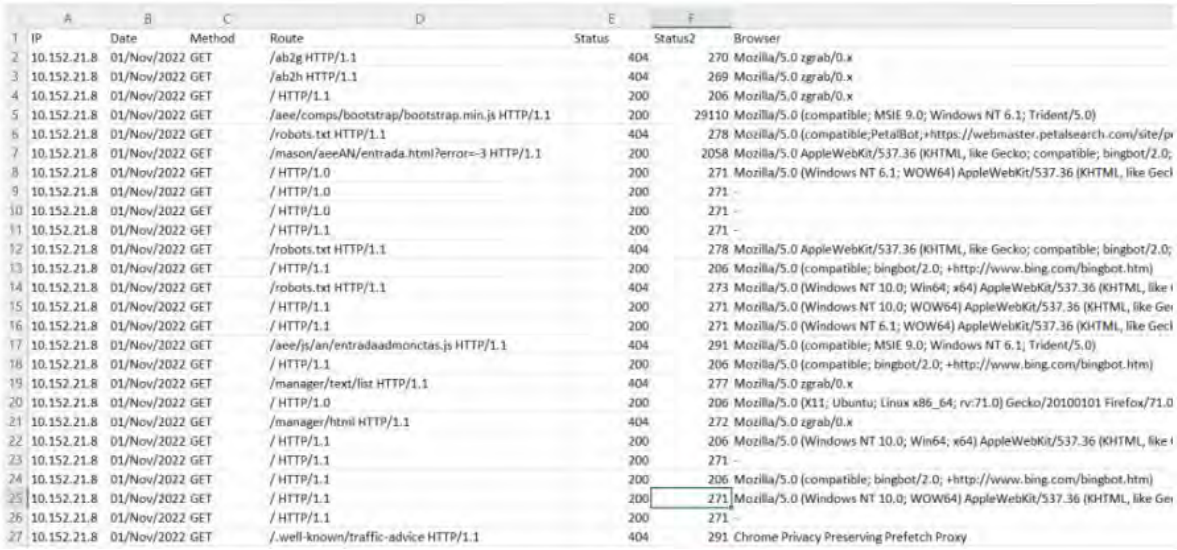
Documentamos en un Notebook el proceso de la decodificación y transformación de los datos en crudo que inicialmente estaban en formato .tar y que software de compresión y descompresión de archivos catalogaban como .gz.

Para la clasificación de atributos que se obtuvieron de la transformación de datos establecimos ocho variables a obtener las cuales se enlistan y definen en la siguiente tabla:

Atributo	Descripción
IP	IP de la cual viene la solicitud
Date	Fecha y hora en la que se realizo la solicitud en formato (día/mes/año:hora:minuto:segundo en horario de México
Method	El Método de petición HTTP utilizado
Route	Ruta solicitada
Status	Estátus HTTP devuelto
Status2	Complemento del estatus HTTP
Browser	Navegador utilizado para la solicitud
Next_Time	Variable Propuesta para detectar la diferencia de tiempo entre una solicitud y otra

Tabla 1. Atributos generados y su descripción.

Se adjunta el enlace para la consulta de código utilizado para el formato, limpieza, análisis y adaptación del grupo de datos: https://github.com/edgarOswaldoDiaz/mlops_zerotrust/tree/1c52bc7268565499075990fb76e9ce652b91903b/C%C3%B3digo/Base%20de%20Datos%201



	A	B	C	D	E	F	
1	IP	Date	Method	Route	Status	Status2	Browser
2	10.152.21.8	01/Nov/2022	GET	/ab2g HTTP/1.1	404	270	Mozilla/5.0 zgrab/0.x
3	10.152.21.8	01/Nov/2022	GET	/ab2h HTTP/1.1	404	269	Mozilla/5.0 zgrab/0.x
4	10.152.21.8	01/Nov/2022	GET	/ HTTP/1.1	200	206	Mozilla/5.0 zgrab/0.x
5	10.152.21.8	01/Nov/2022	GET	/aee/comps/bootstrap/bootstrap.min.js HTTP/1.1	200	29110	Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.1; Trident/5.0)
6	10.152.21.8	01/Nov/2022	GET	/robots.txt HTTP/1.1	404	278	Mozilla/5.0 (compatible;PetalBot,+https://webmaster.petalsearch.com/site/p
7	10.152.21.8	01/Nov/2022	GET	/mason/aeeAN/entrada.html?error=-3 HTTP/1.1	200	2058	Mozilla/5.0 AppleWebKit/537.36 (KHTML, like Gecko; compatible; bingbot/2.0;
8	10.152.21.8	01/Nov/2022	GET	/ HTTP/1.0	200	271	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecl
9	10.152.21.8	01/Nov/2022	GET	/ HTTP/1.0	200	271	-
10	10.152.21.8	01/Nov/2022	GET	/ HTTP/1.0	200	271	-
11	10.152.21.8	01/Nov/2022	GET	/ HTTP/1.1	200	271	-
12	10.152.21.8	01/Nov/2022	GET	/robots.txt HTTP/1.1	404	278	Mozilla/5.0 AppleWebKit/537.36 (KHTML, like Gecko; compatible; bingbot/2.0;
13	10.152.21.8	01/Nov/2022	GET	/ HTTP/1.1	200	206	Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)
14	10.152.21.8	01/Nov/2022	GET	/robots.txt HTTP/1.1	404	273	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like
15	10.152.21.8	01/Nov/2022	GET	/ HTTP/1.1	200	271	Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Ge
16	10.152.21.8	01/Nov/2022	GET	/ HTTP/1.1	200	271	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecl
17	10.152.21.8	01/Nov/2022	GET	/aee/js/an/entradaadmonctas.js HTTP/1.1	404	291	Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.1; Trident/5.0)
18	10.152.21.8	01/Nov/2022	GET	/ HTTP/1.1	200	206	Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)
19	10.152.21.8	01/Nov/2022	GET	/manager/text/list HTTP/1.1	404	277	Mozilla/5.0 zgrab/0.x
20	10.152.21.8	01/Nov/2022	GET	/ HTTP/1.0	200	206	Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:71.0) Gecko/20100101 Firefox/71.0
21	10.152.21.8	01/Nov/2022	GET	/manager/html HTTP/1.1	404	272	Mozilla/5.0 zgrab/0.x
22	10.152.21.8	01/Nov/2022	GET	/ HTTP/1.1	200	206	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like
23	10.152.21.8	01/Nov/2022	GET	/ HTTP/1.1	200	271	-
24	10.152.21.8	01/Nov/2022	GET	/ HTTP/1.1	200	206	Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)
25	10.152.21.8	01/Nov/2022	GET	/ HTTP/1.1	200	271	Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Ge
26	10.152.21.8	01/Nov/2022	GET	/ HTTP/1.1	200	271	-
27	10.152.21.8	01/Nov/2022	GET	/.well-known/traffic-advice HTTP/1.1	404	291	Chrome Privacy Preserving Prefetch Proxy

Figura 8. Datos en formato csv.

Como resultado de la limpieza y preprocesamiento de los datos obtuvimos 273 lotes de 10,000 registros cada uno, con una exclusión de 454,142 registros vacíos o incompletos dando como resultado final 2,730,000 registros limpiados de aproximadamente 3,200,000 en total.

Tabla 2. Porcentaje de registros limpios grupo de datos uno.

Registros en lotes	2,730,000	85%
Registros no coincidentes	454,142	14%
Registros con errores	15,096	1%

6.1.2 Análisis de los datos

Cuando obtuvimos los datos limpios y listos para procesar, se definieron las características de cada atributo para ahondar en los posibles métodos de análisis.

Tabla 3. Atributos replanteados, tipo de variable y su descripción.

Atributo	Tipo de Variable	Descripción
IP	Categórica	IP de la cual viene la solicitud
Date	Continua	Fecha y hora en la que se realizo la solicitud en formato (día/mes/año:hora:minuto:segundo en horario de México
Method	Categórica	El Método de petición HTTP utilizado
Route	Categórica	Ruta solicitada
Status	Continua	Estátus HTTP devuelto
Status2	Continua	Complemento del estatus HTTP
Browser	Categórica	Navegador utilizado para la solicitud
Next_Time	Numérica	Variable Propuesta para detectar la diferencia de tiempo entre una solicitud y otra

Para el campo Date eliminamos la sección que define la hora global (-0600) dado que todos los registros tienen este valor y no aporta a los objetivos del trabajo práctico.

Para el campo Method decidimos resumir la cantidad de datos a aquellos que tengan los valores 400’s y 500’s de HTTP dado que denotan un error tanto en la solicitud del cliente como del servidor y es lo que se busca analizar.

6.1.3 Almacenamiento de datos

Dashboards obtenidos del análisis de datos en Elasticsearch:

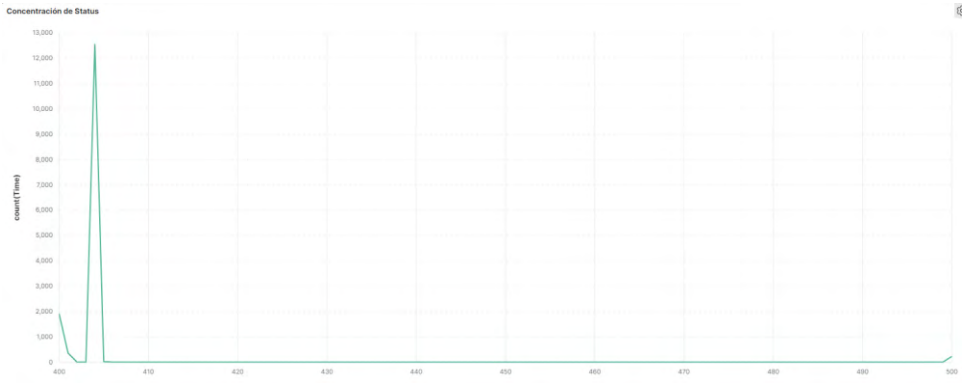


Figura 9. Concentración por Estatus en Elasticsearch.

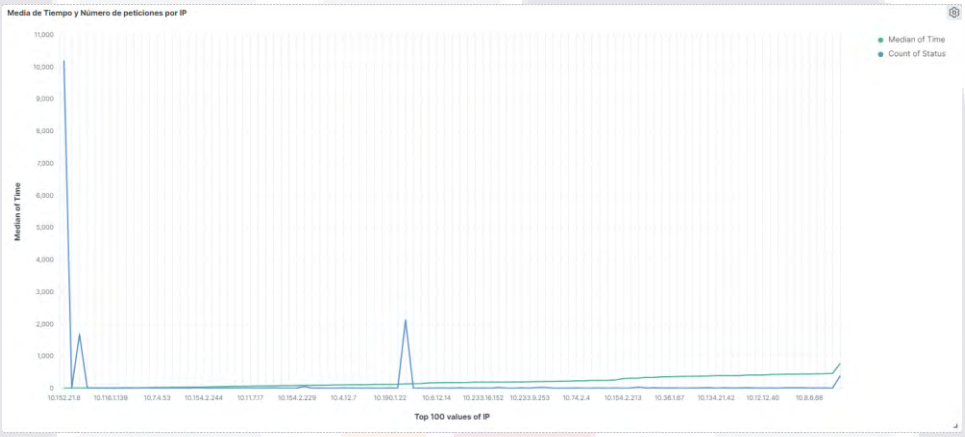


Figura 10. Tiempo por IP en Elasticsearch.

Tabla 4. Media de tiempo y cantidad de transacciones por estatus e IP en Elasticsearch.

Media de Tiempo y Cantidad de Transacciones por Estatus e IP

Status	Top 10 values of IP	Median of Time	Count of Status
400	10.152.21.244	403	21
400	10.152.21.248	5	1,687
400	10.152.21.8	1	206
401	10.152.21.8	0	370
404	10.12.187.29	15,837	1
404	10.233.27.122	8,972	1
404	10.233.6.144	3,458	2
404	10.4.8.66	3,298	1
404	10.154.2.169	2,882	1
404	10.154.2.1	2,672	1
404	10.154.2.210	2,649.5	2
404	10.154.2.44	2,368	1
404	10.233.30.44	2,327	1
404	10.154.2.249	2,186.5	2
405	10.200.10.51	5,538.5	30
405	10.152.21.8	894	1
500	10.4.8.16	891	1
500	10.154.3.33	575	1

Al evaluar los resultaron observamos picos de datos que representan un comportamiento típico de un ataque DDoS (alta concentración de solicitudes con errores 400 y 500 de HTML) sin embargo después de un análisis posterior para comenzar con el entrenamiento del algoritmo concluimos que el grupo de datos actual no cumplía con los requisitos tanto de extensión de registros para el entrenamiento como de atributos para conseguir un modelo entrenado satisfactorio.

6.2 Análisis del grupo de datos número dos

Obtuvimos un segundo grupo de datos en la página Kaggle llamada “DDOS Dataset” que es un compilado de otras bases de datos de logs reales y generados etiquetados tanto como ataques DDoS como comportamiento normal de registros (Devendra, 2019).

El archivo contaba con dos grupos de datos llamados "ddos_balanced" y "ddos_imbalanced" de 6.6 GB y 4 GB respectivamente y con 33 atributos ambos, optamos por analizar ambos grupos de datos para comparar sus características.

El grupo de datos “ddos_balanced” era aquel con un concentrado de 50% registros etiquetados como ataques DDoS y 50% etiquetados como logs benignos.

El grupo de datos “ddos_imbalanced” era aquel con un concentrado de 20% registros etiquetados como ataques DDoS y 80% etiquetados como logs benignos.

6.2.1 Transformación de los datos

En la etapa de limpieza de datos para el grupo de datos “ddos_balanced” filtramos aquellos registros vacíos e incompletos, el grupo de datos restante se dividió en 128 lotes de 100,000 registros cada uno para facilitar su manejo.

Para el grupo de datos “ddos_imbalanced” filtramos aquellos registros vacíos e incompletos, el grupo de datos restante se dividió en 77 lotes de 100,000 registros cada uno para facilitar su manejo.

En total conseguimos 20,500,000 registros con 33 atributos lo cual cumplía con los requisitos de extensión de registros y de atributos para generar satisfactoriamente un algoritmo entrenado para detectar ataques DDoS.

En la etapa de normalización evaluamos la significancia de cada atributo de ambos grupos de datos para filtrar los atributos para entrenar al algoritmo, ya que entrenarlo con los 33 atributos no era óptimo en tiempo y necesidades de procesamiento.

Normalizamos mediante los siguientes métodos matemáticos:

- Original: El grupo de datos con los atributos sin normalizar.

- Log Transformation: Normalización por Transformación Logarítmica.
- MinMax: Normalización por Mínimos y Máximos.
- Normal1: Normalización por primera forma normal (1FN).
- Normal2: Normalización por segunda forma normal (2FN).
- RobustScaler: Normalización por escalado de características.
- Zscore: Normalización por puntuación de Z.

6.2.2 Análisis de los datos

Para el grupo de datos “ddos_imbalanced” obtuvimos la siguiente tabla de valores destacando aquellos con un valor superior al 0.5 de significancia por atributo.

Tabla 5. Significancia por atributo y método de normalización “ddos_imbalanced”.

Atributos	Original	Log Transformation	MinMax	Normal1	Normal2	RobustScaler	Zscore
Protocol	-0.25387	-0.218498	-0.52859	0.172463	0.172078	-0.22553	-0.007444
Flow Duration	-0.117671	0.058405	-0.11811	0.261185	0.195994	-0.131307	-0.002611
Tot Fwd Pkts	0.015266	-0.128297	-0.019539	0.187251	0.166668	0.016484	0.000051
Tot Bwd Pkts	-0.005793	-0.040166	0.105435	0.327461	0.292412	-0.008041	-0.001239
TotLen Fwd Pkts	0.017182	-0.074151	-0.062721	0.091678	0.107727	0.012457	0.000056
TotLen Bwd Pkts	-0.003221	-0.13597	0.002015	-0.060382	-0.047841	-0.00294	-0.000822
Fwd Pkt Len Max	0.236058	-0.037134	0.561735	0.140507	0.153251	-0.093394	0.001505
Flow Byts/s	NaN	-0.216315	0.008853	-0.176513	-0.169577	0.061116	-0.000255
Flow Pkts/s	NaN	-0.105794	-0.195272	-0.068537	-0.057203	0.059882	-0.000555
Fwd IAT Tot	-0.133862	-0.085922	-0.134288	-0.220648	-0.230271	0.095031	-0.003231
Bwd IAT Tot	-0.101785	0.028389	-0.102038	0.118235	0.074317	-0.118581	-0.002362
Fwd Header Len	0.012779	0.074301	-0.014265	0.309188	0.272512	0.014014	0.000051
Bwd Header Len	-0.004675	0.149853	0.157747	0.379961	0.338228	-0.00754	-0.000603
Fwd Pkts/s	-0.054054	-0.128974	-0.04685	-0.106386	-0.098822	-0.045909	-0.001767
Bwd Pkts/s	0.042232	0.003584	0.044898	0.048364	0.050603	0.083148	0.000237
FIN Flag Cnt	0.010778	0.010661	0.01066	-0.015443	-0.014164	0.01066	0.000045
SYN Flag Cnt	0.147762	0.152791	0.152791	0.030404	0.045006	0.152791	0.000946
RST Flag Cnt	-0.12875	-0.129251	-0.129251	-0.024693	-0.023446	-0.129251	-0.001725
PSH Flag Cnt	-0.240204	-0.241078	-0.241077	-0.084486	-0.081523	-0.241077	-0.003831
ACK Flag Cnt	0.377886	0.380384	0.380382	0.345242	0.301603	-0.371403	0.003464
URG Flag Cnt	-0.076395	-0.076603	-0.076603	-0.046904	-0.052852	-0.076603	-0.002848
CWE Flag Count	0.382897	0.382828	0.382828	0.309824	0.289143	0.382828	0.001626
ECE Flag Cnt	0.031494	0.031029	0.031029	0.225851	0.20668	0.031029	0.000302
Down/Up Ratio	0.059484	0.135806	0.807268	0.329511	0.297641	0.059027	0.001253
Pkt Size Avg	-0.048952	-0.153887	0.036422	-0.12529	-0.104351	-0.007839	-0.00053
Fwd Seg Size Avg	0.27237	-0.062284	0.501802	-0.037521	0.001594	0.066951	0.001487
Bwd Seg Size Avg	-0.143323	-0.164988	-0.091253	-0.134736	-0.121123	0.021812	-0.002469
Subflow Fwd Pkts	0.015266	-0.128297	-0.019539	0.187251	0.166668	0.016484	0.000051
Subflow Fwd Byts	0.017182	-0.074151	-0.062721	0.091678	0.107727	0.012457	0.000056
Subflow Bwd Pkts	-0.005793	-0.040166	0.105435	0.327461	0.292412	-0.008041	-0.001239
Subflow Bwd Byts	-0.003221	-0.13597	0.002015	-0.060382	-0.047841	-0.00294	-0.000822
Init Fwd Win Byts	-0.156028	-0.334079	-0.051284	-0.056911	-0.061118	0.200417	-0.004186
Init Bwd Win Byts	0.006658	0.276503	0.006222	0.400171	0.403318	-0.168204	0.000164
Fwd Act Data Pkts	0.016433	-0.162227	-0.142294	-0.035115	-0.033377	0.016841	0.000052
Suma Correlaciones	0.18413	-1.189668	1.261842	3.029739	2.792073	-0.325171	-0.027193

Para el grupo de datos “ddos_balanced” obtuvimos la siguiente tabla de valores destacando aquellos con un valor superior al 0.5 de significancia por atributo en amarillo y en verde los 9 atributos con el valor más alto para el método de Mínimos y Máximos que fue la que obtuvo mayores puntuaciones con diferencia.

Tabla 6. Significancia por atributo y método de normalización “ddos_balanced”

Atributos	Original	Log Transformation	MinMax	Normal1	Normal2	RobustScaler	Zscore
Protocol	-0.440563	-0.38611	-0.778777	0.158156	0.156644	-0.308778	-1.26E-02
Flow Duration	-0.193802	0.342221	-0.12752	0.598986	0.582026	0.018141	-3.42E-03
Tot Fwd Pkts	-0.009048	-0.095938	0.705499	0.177835	0.160436	-0.012012	-4.80E-04
Tot Bwd Pkts	-0.008368	0.086984	0.666537	0.2457	0.222525	-0.010133	-2.69E-04
TotLen Fwd Pkts	0.070651	0.1611	0.693646	-0.22428	-0.201106	-0.041882	6.29E-03
TotLen Bwd Pkts	-0.005655	-0.027472	0.553082	-0.179267	-0.145461	0.00102	-3.29E-04
Fwd Pkt Len Max	0.526297	0.221869	0.692462	-0.225546	-0.191855	-0.046085	1.36E-02
Flow Byts/s	NaN	-0.386106	0.596612	-0.46067	-0.472428	-0.003407	-6.71E-04
Flow Pkts/s	NaN	-0.430007	-0.344007	-0.147157	-0.1153	0.00113	-2.38E-03
Fwd IAT Tot	-0.268332	0.034423	-0.263053	-0.47477	-0.494541	0.025727	-5.53E-03
Bwd IAT Tot	-0.135363	0.31048	-0.06247	0.486795	0.426113	0.035386	-2.14E-03
Fwd Header Len	-0.002297	0.327972	0.704898	0.256251	0.228181	-0.000073	6.26E-05
Bwd Header Len	-0.004453	0.443122	0.668261	0.290013	0.26001	0.011807	-1.76E-05
Fwd Pkts/s	-0.101778	-0.428293	-0.095314	-0.189035	-0.168334	-0.010301	-2.13E-03
Bwd Pkts/s	-0.083988	-0.287698	-0.06321	-0.001286	0.016711	0.001852	-1.84E-03
FIN Flag Cnt	0.065592	0.06542	0.065418	-0.00745	-0.006885	0.065418	1.94E-03
SYN Flag Cnt	-0.198059	-0.194536	-0.194536	-0.098433	-0.086781	-0.194536	-5.74E-03
RST Flag Cnt	-0.313667	-0.314367	-0.314366	-0.057194	-0.052022	-0.314366	-7.73E-03
PSH Flag Cnt	-0.472179	-0.473259	-0.473258	-0.170959	-0.163017	-0.473258	-1.31E-02
ACK Flag Cnt	0.771516	0.773498	0.13407	0.277491	0.248856	-0.359963	1.74E-02
URG Flag Cnt	-0.141451	-0.141756	-0.141756	-0.087063	-0.098227	-0.141756	-4.18E-03
CWE Flag Count	-0.002496	-0.002502	-0.002502	-0.001285	-0.00132	-0.002502	-2.65E-17
ECE Flag Cnt	-0.313629	-0.314324	-0.314324	-0.110753	-0.101453	-0.314324	-8.94E-03
Down/Up Ratio	0.201993	0.414759	0.87525	0.25589	0.235473	0.201108	5.73E-03
Pkt Size Avg	0.062903	-0.00003	0.656874	-0.312235	-0.296493	0.043723	5.57E-03
Fwd Seg Size Avg	0.540041	0.199058	0.684889	-0.296148	-0.274993	0.008944	1.38E-02
Bwd Seg Size Avg	-0.240646	-0.08653	0.535161	-0.29714	-0.283039	0.052379	-3.37E-03
Subflow Fwd Pkts	-0.009048	-0.095938	0.705499	0.177835	0.160436	-0.012012	-4.80E-04
Subflow Fwd Byts	0.070651	0.1611	0.693646	-0.22428	-0.201106	-0.041882	6.29E-03
Subflow Bwd Pkts	-0.008368	0.086984	0.666537	0.2457	0.222525	-0.010133	-2.69E-04
Subflow Bwd Byts	-0.005655	-0.027472	0.553082	-0.179267	-0.145461	0.00102	-3.29E-04
Init Fwd Win Byts	-0.310255	-0.627295	-0.310788	-0.107945	-0.115657	-0.030969	-8.87E-03
Init Bwd Win Byts	-0.298666	0.327341	-0.040978	0.017138	0.003217	-0.039281	-5.97E-03
Fwd Act Data Pkts	-0.206161	-0.218446	0.676424	-0.225741	-0.215758	-0.283273	-3.76E-03
Suma Correlaciones	-1.464283	-0.581748	8.000988	-0.890114	-0.908084	-2.183271	-0.023887197

Elegimos continuar con el grupo de datos "ddos_balanced" por la considerable significancia que reflejaba en el comportamiento de sus datos mediante la normalización de Mínimos y Máximos (fórmula 1).

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Figura 11. Fórmula de normalización Mínimos y Máximos

De 15 variables candidatas elegimos 8 con la mayor calificación positiva, 1 con la mayor calificación negativa y la variable "Label" de control dando como total 10 variables:

1. Protocol.
2. Tot Fws Pkt.
3. TotLen Fwd Pkts.
4. Fwd Pkt Len Max.
5. Fwd Header Len.
6. Down/Up Ratio.
7. Fwd Seg Size Avg.
8. Subflow Fwd Pkts.
9. Subflow Fwd Byts.
10. Label.

6.2.3 Entrenamiento del algoritmo

Para el entrenamiento consideramos los algoritmos más utilizados en ataques DDoS (J. T. Mejía Viteri et al., 2022):

- Árboles de decisión.
- Bosques aleatorios.
- Máquina de soporte de vectores – SVM (Support vector machine, por su traducción en inglés).
- Redes neuronales.

Para el caso de SVM y Redes Neuronales se omitieron debido a que no se pudo cumplir con las necesidades de procesamiento para poder realizarlas. Aun así, se agregan en la tabla para considerarse en caso de poder realizarse en el futuro.

Tabla 7. Comparativa de resultados de entrenamiento

X	Árboles de Decisión	Bosques Aleatorios	SVM	Redes Neuronales
True Positive	1882193	1882197		
False Positive	52	48		
False Negative	1311	1312		
True Negative	1940499	1940498		
Accuracy	99.96	99.97		
2da Prueba	99.96	99.97		
3ra Prueba	99.96	99.96		

Al final trabajamos con algoritmos entrenados mediante la técnica de Bosques Aleatorios tomando en cuenta los valores ligeramente más altos en comparación con la técnica de Árboles de Decisión. Para cada prueba en la tabla consideramos el valor más alto obtenido, pero en promedio se obtuvo una precisión del 95%.

Para realizar una evaluación más exhaustiva buscamos generar datos sintéticos de diferentes formas e ingresarlas al algoritmo entrenado para medir su precisión:

Tabla 8. Comparativa de precisión algoritmo generado vs datos sintéticos.

X	Normal	Scikit-learn	Faker	Tonic	MostAI 96%	MostAI 96% Sobre
True Positive	1784033	161790	0	2199620	2992246	2987690
False Positive	113520	478425	640390	1726471	166383	170939
False Negative	88305	368275	0	2169985	2004940	1524580
True Negative	1852531	271510	639610	1703924	1233744	1714104
Accuracy	95	34	50	50	66-74	73-79

En la tabla se muestra:

- La precisión promedio del algoritmo entrenado (Normal).
- Datos sintéticos generados con la librería Scikit-learn de Python.
- Datos sintéticos generados con la librería Faker de Python.
- Datos sintéticos generados en el portal dedicado a ciencia de datos Tonic.
- Datos sintéticos generados en el portal dedicado a ciencia de datos MostAI.
- Datos sintéticos generados en MostAI evitando el sobreentrenamiendo.

El caso especial con el portal MostAI (Platzer Michael et al., 2017) y por el que se siguió trabajando con este, es que no generaba registros aleatorios sin considerar el comportamiento de los datos (aspecto crucial en el entrenamiento para ataques DDoS) como serían las librerías de Python o el portal Tonic, si no que generaba otro algoritmo entrenado para crear un nuevo grupo de datos que reflejara lo más exacto posible al comportamiento de los datos reales, por esto es que en la tabla anterior se puede observar reflejado un porcentaje al lado del nombre, el 96% hace referencia al porcentaje de exactitud con el que el grupo de datos sintéticos se comportaba de la misma forma que los datos originales.

Los otros métodos para generar datos sintéticos fueron descartados por la imposibilidad de generar registros útiles para evaluar al algoritmo entrenado.

Tabla 9. Registros de comparativa, algoritmo generado vs datos sintéticos generados erróneamente.

	92.1		95	95
X	MostAI	MostAI Normal	MostAI	Sobreentren
True Positive	3744236	3966298	1558088	1558070
False Positive	222149	87	2367825	2367843
False Negative	2640227	3833476	1537886	1537861
True Negative	1193388	139	2336201	2336226
Accuracy	63	51	50	50

Una nueva problemática que observamos al intentar mejorar la precisión mediante datos sintéticos de MostAI fue que tenía la tendencia a mostrar una precisión del 50%, al analizar las posibles causas encontramos que era debido a la forma en la que estaban distribuidos los registros dentro del grupo de datos original, esta se encontraba dividida exactamente a la mitad, la primera mitad eran registros etiquetados como benignos y la segunda como malignos, por lo tanto el entrenamiento tanto del algoritmo entrenado como para la generación de los datos sintéticos eran erróneos ya que no se reflejaba el comportamiento real de uno o varios ataques de denegación de servicios.

Se adjunta el enlace para la consulta de código utilizado para el formateo, limpieza y adaptación del grupo de datos y posterior entrenamiento del algoritmo: https://github.com/edgarOswaldoDiaz/mlops_zerotrust/tree/1c52bc7268565499075990fb76e9ce652b91903b/C%C3%B3digo/Base%20de%20Datos%202

6.3 Análisis del grupo de datos número tres

En una tercera iteración buscando tener datos con un comportamiento actual de una página web real se facilitaron registros de parte de los servidores de la Universidad Autónoma de Aguascalientes, se descargaron casi 5 millones de registros de logs de diferentes páginas de la universidad en formatos .log.

6.3.1 Transformación de los datos

Dada la estructura diferente de los registros y a la estandarización de atributos eliminamos registros con información incompleta obteniendo cerca de 3 millones de registros limpios.

Además, con fines de complementar la información generamos 2 nuevos atributos para cada registro: Seg_Ant y Seg_AntIP.

En una segunda limpieza de registros eliminamos aquellos con información con errores de formato y caracteres especiales, finalizamos con los siguientes atributos:

- IP: La IP relacionada al registro.
- Seg_Ant: El tiempo de diferencia en segundos entre el log del registro y el anterior.
- Seg_AntIP: El tiempo de diferencia en segundos entre el log del registro y el anterior con la misma IP.
- Fecha: Fecha completa del registro del log en formato “date”.
- HMTL: Instrucción en HTML del log.
- Res: Código de respuesta HTML.
- Inf: Cantidad de información en bytes del log.
- Pag: Dirección accedida.
- Naveg: Detalles del navegador utilizado.

Omitimos dos (2) atributos relacionados a detalles del navegador utilizado por no considerarse relevantes en el análisis de información.

Debido a que en algunos registros no contábamos con la información para obtener los datos en los atributos “Seg_Ant” y “Seg_AntIP” optamos por llenar esos registros vacíos mediante el promedio obtenido de los registros que si tenían esos datos.

6.3.2 Análisis de los datos

En la fase de análisis de datos se presentó una nueva problemática, los registros no estaban etiquetados como ataques DDOS y logs normales.

Decidimos que, mediante la metodología de auto entrenamiento (self-training), utilizando un set de datos más pequeño se etiquetarían los registros del grupo de datos actual. Descartamos las bases de datos utilizadas anteriormente debido a que no contaban con la misma estructura de la información ni los mismos atributos.

Obtuvimos el siguiente grupo de datos de Kaggle llamada “WordPress DDos Log Dataset” (Ajiboye Toluwalase, 2023) que contenía los mismos atributos que el grupo de datos que se estaba manejando y contaba con las etiquetas de logs malignos y benignos.

Para la clasificación mediante Self-training consideramos un rango de confianza del 90% en adelante para integrar los registros en una nueva ronda.

En total realizamos 9 rondas en las que de los 2 millones y medio de registros se clasificaron exitosamente casi 2 millones como malignos o benignos.

6.3.3 Entrenamiento del algoritmo

Para poder normalizar el grupo de datos y entrenar un modelo hicimos los siguientes ajustes a los atributos del grupo de datos:

- IP: Eliminamos los puntos de las direcciones manejando un solo número entero. Posteriormente se eliminó por considerarse no relevante para el entrenamiento.
- Seg_Ant: Mantuvimos el mismo formato.
- Seg_AntIP: Mantuvimos el mismo formato.
- Fecha: Se eliminó por considerarse no relevante para el entrenamiento.
- HTML: Se guardó como valor entero reflejando la longitud del texto. Se renombró como “HTML_len”.
- Res: Mantuvimos el mismo formato.
- Inf: Mantuvimos el mismo formato.
- Pag: Se guardó como valor entero reflejando la longitud del texto. Se renombró como “Pag_len”.
- Naveg: Se guardó como valor entero reflejando la longitud del texto. Se renombró como “Naveg_len”.

El grupo de datos se normalizó mediante Mínimos y Máximos.

Realizamos un análisis de la relevancia de cada uno de los atributos para detectar si un registro era o no un ataque DDOS:

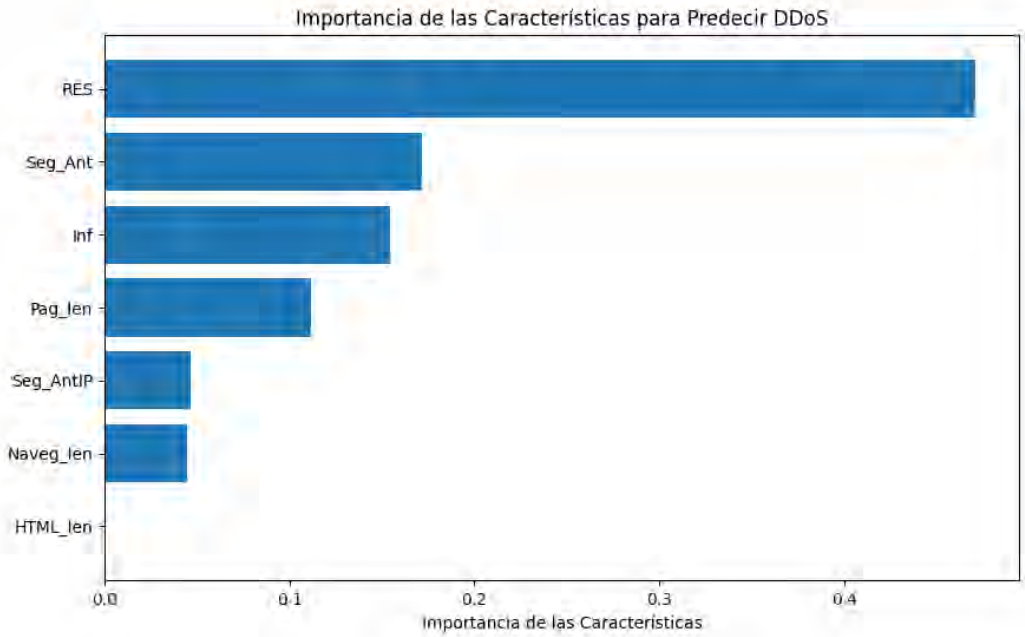


Figura 12. Relevancia de los atributos para predecir DDOS.

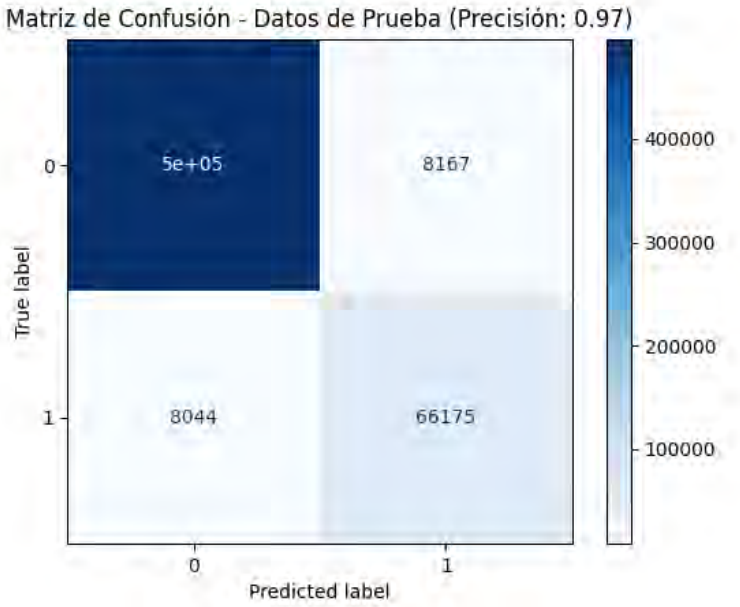
Para el entrenamiento del algoritmo elegimos el método de “Árboles de Decisión”, además implementamos código para evitar el sobre entrenamiento o “overfitting” del modelo.

Obtuvimos una precisión con promedio un del 97% en 10 entrenamientos diferentes.

```
PS C:\Users\ADMIN> & C:/Users/ADMIN/AppData/L
Precisión del modelo: 0.97
Modelo guardado en: C:\Users\ADMIN\Desktop\MI
Precisión en los datos sintéticos: 0.97
```

Figura 13. “Resultados del entrenamiento y evaluación final”

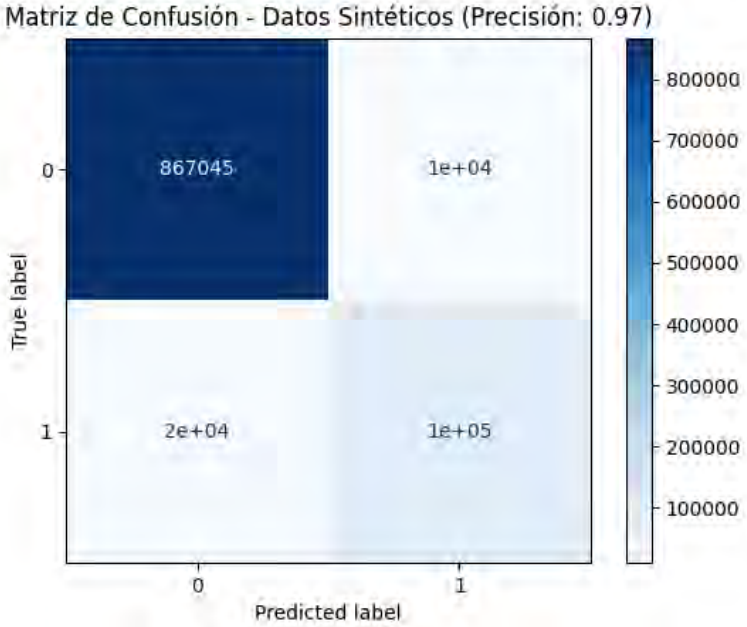
Tabla 10. Matriz de confusión con datos de prueba.



Para una evaluación más exhaustiva generamos un millón de datos sintéticos usando nuevamente la página de Mostly.AI con una precisión del 95%.

Obtuvimos una precisión con un promedio del 97% evaluando la predicción del modelo entrenado previamente.

Tabla 11. Matriz de Confusión con Datos Sintéticos



Capítulo 7 Conclusiones y trabajos futuros

El desarrollo de esta tesis permitió identificar importantes lecciones sobre la relevancia de la ciencia de datos en el ámbito de la ciberseguridad, así como la complejidad inherente a la implementación de modelos de aprendizaje automático robustos y eficientes para la detección de ciberataques. Uno de los principales hallazgos fue la importancia de realizar un adecuado procesamiento de las bases de datos, desde su obtención hasta su limpieza y uso, ya que esto es fundamental para obtener resultados precisos y evitar eventualidades a lo largo del entrenamiento y evaluación de los modelos.

Asimismo, se constató que, para llevar a cabo un proyecto exitoso en ciencia de datos, es crucial tener claridad sobre la procedencia de los datos, así como el resultado final del proyecto. Se deben considerar diversas metodologías de análisis y preprocesamiento de datos, y explorar diferentes enfoques antes de determinar el más adecuado para la situación en cuestión. Además, fue esencial el uso de técnicas avanzadas para evaluar y mejorar el rendimiento del modelo, como la normalización de los datos, la selección de atributos, la creación de datos sintéticos, la optimización de código y memoria y el uso de mecanismos que previenen el sobreajuste, lo que nos permitió asegurar una mayor calidad y robustez en el algoritmo final.

Además, la creación e implementación de un laboratorio de ciencia de datos colaborativo multidisciplinario dentro de la Universidad Autónoma de Aguascalientes que provea de servicios de análisis, procesamiento y productos de inteligencia artificial es crucial para la continuación de proyectos como este, abonando al conocimiento humano y la creación de profesionistas altamente especializados en las áreas actuales más innovadoras en cuanto a las áreas de Ciencia, Tecnología, Ingeniería y Matemáticas (STEM por sus siglas en inglés).

Como trabajos futuros, se sugiere continuar explorando modelos y técnicas que optimicen la detección de ciberataques, no solo limitándose a ataques DDoS, sino considerando otras vulnerabilidades emergentes que afectan a las organizaciones. Esto incluye la aplicación de métodos más sofisticados de aprendizaje automático (como las redes neuronales), así como el desarrollo de técnicas de ciberdefensa basadas en el análisis predictivo en conjunto con la inteligencia artificial.

Todo esto con ayuda del laboratorio de ciencia de datos de la UAA y su trabajo para implementar benchmarking para ofrecer en el futuro servicios de ciencia y análisis de datos.

Finalmente, queda claro que la ciberseguridad es un campo en constante evolución y crecimiento, con una gran área de oportunidad en México. La generación de conocimiento e investigación será clave para que las organizaciones mexicanas públicas y privadas puedan estar mejor preparadas frente a las amenazas del futuro, asegurando así la seguridad de sus datos y la continuidad de sus operaciones.

Bibliografía

Ajiboye Toluwalase. (2023, septiembre 20). *WordPress DDos Log Dataset*. Kaggle. https://www.kaggle.com/datasets/ajiboyetoluwalase/wordpress-ddos-log-dataset?select=Wordpress+DDOS+attack+Logs.txt

Alani, M. M. (2021). Big data in cybersecurity: a survey of applications and future trends. *Journal of Reliable Intelligent Environments*, *7*(2), 85–114. https://doi.org/10.1007/S40860-020-00120-3/TABLES/6

Apache Software Foundation. (2023, septiembre 23). *Apache Spark*. https://spark.apache.org.

Apache Software Foundation. (2024, agosto 14). *Apache Superset Introduction*. https://superset.apache.org/docs/intro/

Becerra Pozas, J. L. (2023). ¿Qué son los datos sintéticos? Datos generados para ayudar a su estrategia de IA. *CIO*.

Cao, Y., & Yang, J. (2015). *Towards Making Systems Forget with Machine Unlearning*. https://doi.org/10.1109/SP.2015.35

Cebula, J. J., Popeck, M. E., & Young, L. R. (2014). A Taxonomy of Operational Cyber Security Risks Version 2 CERT ® Division. *Software Engineering Institute*, *2*. http://www.sei.cmu.edu

Chávez, G., Ibarra, M., Ortega, O., & Sigler, É. (2019, noviembre 12). *Los hackers reclamaron 4.9 mdd a Pemex para liberar su información*. 2019. https://expansion.mx/tecnologia/2019/11/12/los-hackers-han-reclamado-4-9-mdd-pemex-liberar-informacion

Cremer, F., Sheehan, B., Fortmann, M., Kia, A. N., Mullins, M., Murphy, F., & Materne, S. (2022). Cyber risk and cybersecurity: a systematic review of data availability. *The Geneva Papers on Risk and Insurance - Issues and Practice 2022 47:3*, *47*(3), 698–736. https://doi.org/10.1057/S41288-022-00266-6

Devendra. (2019, enero 15). *DDoS Dataset*. Kaggle. https://www.kaggle.com/datasets/devendra416/ddos-datasets

Elasticsearch B.V. (2023, septiembre 23). *Kibana*. https://www.elastic.co/es/kibana.

Google. (2023). *Te damos la bienvenida a Colaboratory* (1). Google. https://colab.research.google.com/#scrollTo=Nma_JWh-W-IF

Guarneros, F. (2022). *En 2021, el ransomware a empresas creció más de 600% en México*. Expansión. https://expansion.mx/tecnologia/2022/01/27/en-2021-el-ransomware-a-empresas-crecio-mas-de-600-en-mexico

Guillen, B. (2021). *Avaddon: Los hackers que robaron información a la Lotería Nacional filtran 800 archivos confidenciales*. EL PAÍS México. https://elpais.com/mexico/2021-06-09/los-hackers-que-robaron-informacion-a-la-loteria-nacional-filtran-800-archivos-confidenciales.html

Ivaturi, K., & Janczewski, L. (2011). *Association for Information Systems AIS Electronic Library (AISeL) A Taxonomy for Social Engineering attacks*. http://aisel.aisnet.org/confirm2011

J. T. Mejía Viteri, M. I. Gonzales Valero, A. del R. Fernández Torres, & N. M. Crespo Torres. (2022). *Seguridad contra ataques DDoS en los entornos SDN con Inteligencia Artificial* (3a ed., Vol. 7). RMC.

Jabbour, G. (2022). *Ciberseguridad en México: Dependencias gubernamentales son blanco fácil para los delincuentes*. Expansión. https://expansion.mx/tecnologia/2022/08/18/ciberseguridad-en-mexico-falta-de-inversion

Jhawar, R., & Tejada, Z. (2022, noviembre 18). *Extracción, transformación y carga de datos (ETL)*. Azure Architecture Center | Microsoft Learn. https://learn.microsoft.com/es-es/azure/architecture/data-guide/relational-data/etl

Juan José Beunza, Enrique Puertas Sanz, & Nuin Emilia Condés Moreno. (2020). Manual práctico de inteligencia artificial en entornos sanitarios. *Elsevier Health Sciences*, 35–39. https://books.google.com.mx/books?id=88nSDwAAQBAJ&dq=algoritmos+supervisados&lr=&hl=es&source=gbs_navlinks_s

Kantardzic, M. (2002). *Data Mining Learning from Large Data Sets - Introduction*. http://ieeexplore.ieee.org/xpl/bkabstractplus.jsp?bkn=5265979

LibreOffice. (2024, agosto 13). *¿Quiénes somos?* https://es.libreoffice.org/acerca-de/quienes-somos/

Longbing, C., & Philip S, Y. (2018). Data Science Thinking: The Next Scientific, Technological and Economic Revolution. *Springer*. http://www.springer.com/series/15063

Marr, B. (2018, mayo 21). *How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read*. Forbes. https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=68339dbb60ba

Marr, B. (2021). *How Much Data Is There In the World? | Bernard Marr*. Bernard Marr & Co. https://bernardmarr.com/how-much-data-is-there-in-the-world/

Miller, K. (2022, noviembre). *ETL Process Overview*. ETL Database | Stitch. https://www.stitchdata.com/etldatabase/etl-process/

Monnappa, A. (2022). *Data Science vs. Big Data vs. Data Analytics*. Simplilearn. https://www.simplilearn.com/data-science-vs-big-data-vs-data-analytics-article#what_is_data_science

Nanduri, J., Jia, Y., Oka, A., Beaver, J., & Liu, Y. W. (2020). Microsoft uses machine learning and optimization to reduce e-commerce fraud. *Interfaces*, *50*(1), 64–79. https://doi.org/10.1287/inte.2019.1017

National Cyber Security Index. (2023a). *NCSI :: Mexico*. e-Governance Academy Foundation. https://ncsi.ega.ee/country/mx/

National Cyber Security Index. (2023b). *NCSI :: Ranking*. e-Governance Academy Foundation. https://ncsi.ega.ee/ncsi-index/?order=-rank

Ostos Ríos, L. E., Bautista Villalpando, L. E., Muñoz López, J., & Oswaldo Díaz, E. (2020). *Análisis de grandes cantidades de datos por medio de técnicas de máquinas de aprendizaje para la Ciberseguridad* [Trabajo Práctico]. Universidad Autónoma de Aguascalientes.

Paredes, A. (2022, septiembre 29). *Hackers rompen seguridad digital de Sedena; extraen miles de documentos, revela Loret*. El Universal. https://www.eluniversal.com.mx/nacion/hackers-rompen-seguridad-digital-de-sedena-extraen-miles-de-documentos-revela-loret

Platzer Michael, Kalcher Klaudius, & Boubela Roland. (2017, enero 1). *MOSTLY.AI*. https://mostly.ai/

Proyect Jupyter. (2023, noviembre 16). *Jupyter*. https://jupyter.org

Python. (2022, noviembre 22). *What is Python?* Python. https://www.python.org/doc/essays/blurb/

Python Software Foundation. (2023, noviembre 13). *tarfile — Leer y escribir archivos tar*. https://docs.python.org/es/3/library/tarfile.html.

Rebala, G., Ravi, A., & Churiwala, S. (2019). Machine Learning Definition and Basics. *An Introduction to Machine Learning*, 1–17. https://doi.org/10.1007/978-3-030-15729-6_1

Reuters. (2022). *Cibercrimen ha costado 6 millones de dólares a las economías del mundo*. El Economista. https://www.eleconomista.com.mx/tecnologia/Cibercrimen-ha-costado-6-millones-de-dolares-a-las-economias-del-mundo-20220511-0022.html

Rosenberg, C., Hebert, M., & Schneiderman, H. (2004). *Semi-Supervised Self-Training of Object Detection Models*. https://kilthub.cmu.edu/articles/journal_contribution/Semi-Supervised_Self-Training_of_Object_Detection_Models/6560834?file=12043121

Sai, K., Kranthi, V., & Reddy, S. (2008). Stock Market Prediction Using Machine Learning. *International Research Journal of Engineering and Technology*, 1032. www.irjet.net

Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: an overview from machine learning perspective. *Journal of Big Data*, *7*(1), 1–29. https://doi.org/10.1186/S40537-020-00318-5/FIGURES/3

Singer, P. W., & Friedman, A. (2014). Cybersecurity: What Everyone Needs to Know. *Oxford University Press*, 224. https://books.google.com/books?id=9VDSAQAAQBAJ&pgis=1

Sun, N., Zhang, J., Rimba, P., Gao, S., Zhang, L. Y., & Xiang, Y. (2019). Data-Driven Cybersecurity Incident Prediction: A Survey. *IEEE Communications Surveys and Tutorials*, *21*(2), 1744–1772. https://doi.org/10.1109/COMST.2018.2885561

Surya, L. (2017). Risk Analysis Model That Uses Machine Learning to Predict the Likelihood of a Fire Occurring at A Given Property. En *International Journal of Creative Research Thoughts* (Vol. 5). https://ssrn.com/abstract=3785655

Synthetic Data for Official Statistics. (2023). United Nations. https://doi.org/10.18356/9789210021708

Vicario, G., & Coleman, S. (2019). A review of data science in business and industry and a future view. *Applied Stochastic Models in Business and Industry*. https://doi.org/10.1002/ASMB.2488

Villasenor Garcia, E. A., Coronado Iruegas, A. A., Pimentel Alarcon, A. E., Suarez Ponce De Leon, R. R., Figueroa Martinez, A., Esquer Martinez, A., Silva Cuevas, V., Cabrera Zamora, I. G., & Diaz, E. O. (2022). Data Lake Strategy for Data Science Workflows. *2022 11th International Conference On Software Process Improvement (CIMPS)*, 219–223. https://doi.org/10.1109/CIMPS57786.2022.10035694

World Economic Forum. (2020). *The Global Risks Report 2020 Insight Report 15th Edition*. https://www3.weforum.org/docs/WEF_Global_Risk_Report_2020.pdf

World Economic Forum. (2023). *Global Riks Report 2023 18th Edition*. https://www3.weforum.org/docs/WEF_Global_Risks_Report_2023.pdf