Científica

vol. 29, no. 2, July-December. 2025, pp. 01-20
ISSN 2594-2921, Instituto Politécnico Nacional, MÉXICO
e290209 | DOI: 10.46842/ipn.cien.v29n2a09

ARTICLE

Mecanismos para la detección de anomalías de ciberseguridad a través del análisis de grandes cantidades de datos

Mechanisms for the Detection of Cybersecurity Anomalies through Big Data Analysis

https://cientifica.site

Pablo Andrés **Martínez Velasco** [1]
Luis Eduardo **Bautista Villalpando** [2]
Edgar Oswaldo **Díaz** [3]
Juan **Muñoz López** [4]

Universidad Autónoma de Aguascalientes,
Centro de Ciencias Básicas, Aguascalientes, MÉXICO
[1] ORCID: 0009-0002-0607-672X / pamartinez@live.com.mx
[2] ORCID: 0000-0001-8665-1776 / eduardo.bautista@edu.uaa.mx

Instituto Nacional de Estadística y Geografía,
Aguascalientes, MÉXICO
[3] ORCID: 0000-0002-3695-7715 / oswaldo.diaz@inegi.org.mx
[4] ORCID: 0000-0003-2532-0134 / juan.munoz@edu.uaa.mx

Recibido 09/08/2025, aceptado 19/12/2025.

Resumen

Esta investigación se centra en la convergencia de la Ciencia de Datos y la Ciberseguridad. El objetivo es el de proponer una solución innovadora que fortalezca la protección de organizaciones en México frente a amenazas cibernéticas. Este trabajo se enfoca en los ataques Denegación de Servicios Distribuidos (DDoS por sus siglas en inglés) a través de diversos estudios y comparaciones usando la metodología más adecuada que permite procesar bases de datos relacionadas de ciberataques desde una perspectiva de modelos de aprendizaje automático tanto supervisados como semi-supervisados. Estos algoritmos han sido entrenados, evaluados y configurados cuidadosamente evitando el sobre ajuste (overfitting) y el ruido de datos corruptos. Así, el modelo resultante es capaz de clasificar eficientemente registros de ataques DDoS y archivos de registros (logs) normales con una alta precisión, demostrando su potencial para ser implementado en ambientes de producción con el objetivo de fortalecer la ciberseguridad evitando la fuga de datos e inhabilitación de servicios.

Palabras clave: aprendizaje automático, aprendizaje semi-supervisado, aprendizaje supervisado, ciberseguridad, ciencia de datos, big data.

Abstract

This research centers on the convergence of Data Science and Cybersecurity. The aim is to propose an innovative solution to strengthen the protection of organizations in Mexico against cyber threats. This work focuses on Distributed Denial of Service (DDoS) attacks through various studies and comparisons, employing the most suitable methodology to process cyberattack-related datasets from the perspective of both supervised and semi-supervised machine learning models. These algorithms were carefully trained, evaluated, and tuned to avoid overfitting and noise from corrupt data. Consequently, the resulting model is capable of efficiently classifying DDoS attack records and normal log files with accuracy, demonstrating its potential for implementation in production environments to enhance organizational cybersecurity, thereby preventing data breaches and service disruptions.

Index terms: machine learning, semi-supervised learning, supervised learning, big data science, cybersecurity.

I. Introducción

La generación de la información digital en las décadas del 2010 al 2020 se incrementó a niveles acelerados. En 2018 se generaron 2.5 quintillones de bytes de datos diarios aumentando con el uso de la tecnología de IoT [1]. En el año 2018 el tamaño de la esfera de datos alcanzó los 18 zettabytes y se espera que para el 2025 alcance los 175 zettabytes [2].

En este contexto, las técnicas clásicas para procesar datos, en su mayoría técnicas estadísticas sobre la actividad en línea de usuarios anónimos o de información sensible, son insuficientes para cubrir la demanda de procesamiento en tiempo y forma [3]. Esto ha resultado en la creación de nuevas y especializadas técnicas de gestión de datos como el uso de la ciencia de datos (Data Science por su traducción al inglés).

De acuerdo con "World Economic Forum's Global Risk Report" de 2020 [4], los ciberataques en infraestructuras críticas se encontraron en el puesto número cinco. En el año 2023, los ciberataques a infraestructuras ocuparon el puesto número ocho, lo cual, no deja de ser uno de los factores críticos a corto y largo plazo considerándose uno de los principales temas de interés dentro del área de tecnología [5]. Como lo menciona Alessandro Profumo para El Economista: "La cibercriminalidad ha costado más de 6 billones de dólares a la economía del mundo" [6].

II. Revisión de la literatura

A. Ciencia de Datos

El término Ciencia de Datos se acuñó por primera vez en 1960 por Peter Naur. Usado como sinónimo de ciencias computacionales, en 1974 se le relacionó con los métodos de procesamiento de datos por Naur y finalmente apareció publicado en un artículo de la Federación Internacional de Sociedades de Clasificación (The International Federation of Classification Societies) en 1996. Desde entonces este concepto se adoptó de forma internacional para describir esta área interdisciplinaria [7].

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 [8] desde; el planteamiento del problema, adquisición de datos, preparación de datos, análisis exploratorio, modelado de datos, visualización y comunicación e implantación y mantenimiento.

B. Aprendizaje supervisado y semisupervisado

El 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 con base a los resultados esperados y ajusta sus parámetros internos hasta lograr una predicción satisfactoria [9].

Normalmente se divide un grupo de datos en dos conjuntos, uno para el entrenamiento y otro para la evaluación o prueba (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% y se obtiene información del algoritmo como el porcentaje de predicción o tablas como la matriz de confusión.

Científica, vol. 29, no. 2, pp. 01-20, July-December 2025,
ISSN 2594-2921, Instituto Politécnico Nacional, México
e290209 | DOI: 10.46842/ipn.cien.v29n2a09

El Aprendizaje Semi-supervisado también llamado "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 [9].

El enfoque basado en el auto entrenamiento 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 el que se entrena el modelo. El proceso continúa iterando las veces necesarias hasta conseguir una cantidad aceptable de datos etiquetados con una confianza alta [10].

C. Ciencia de Datos de la Ciberseguridad

En nuestra era moderna digital, el problema de la Ciberseguridad ya no solo está relegada a aquellos expertos en ciencias computacionales, sino que es un asunto que engloba a todas las personas [11]. La figura 1 presenta la evolución de la toma de decisiones contra riesgos desde la comprensión, detección y predicción hasta acciones prescriptivas.

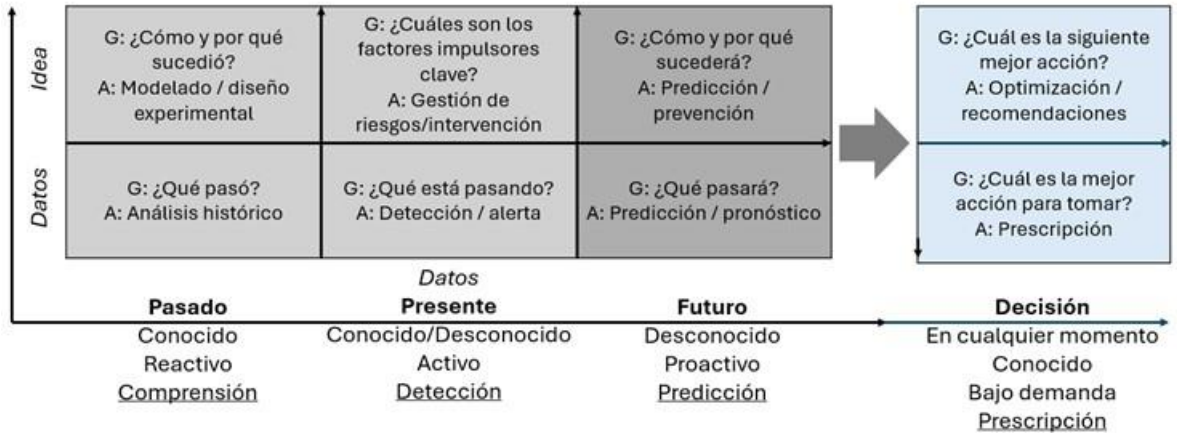

Fig. 1. Análisis de ciclo de vida de toma de decisiones desde los datos hasta la decisión [12].

Debido a lo anterior, es necesario 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 de forma rápida y efectiva ante amenazas conocidas y no tan conocidas. Es aquí donde la Ciencia de Datos, el Aprendizaje Automático y la Inteligencia Artificial (IA) juegan un rol vital en el próximo paso de Ciberseguridad, a lo cual se le conoce como: Ciencia de Datos de Ciberseguridad [13].

D. Datos sintéticos

Los datos sintéticos son utilizados en el área de la IA para el entrenamiento de modelos ya que apoyan la etapa de evaluación y generación de datos, esto con el fin de que cumplan con los criterios necesarios para el entrenamiento de algoritmos y se asegure la privacidad de los datos originales. La figura 2 muestra cómo se generan datos sintéticos a partir de datos reales siguiendo sus comportamientos y tendencias.

Científica, vol. 29, no. 2, pp. 01-20, July-December 2025,
ISSN 2594-2921, Instituto Politécnico Nacional, México
e290209 | DOI: 10.46842/ipn.cien.v29n2a09

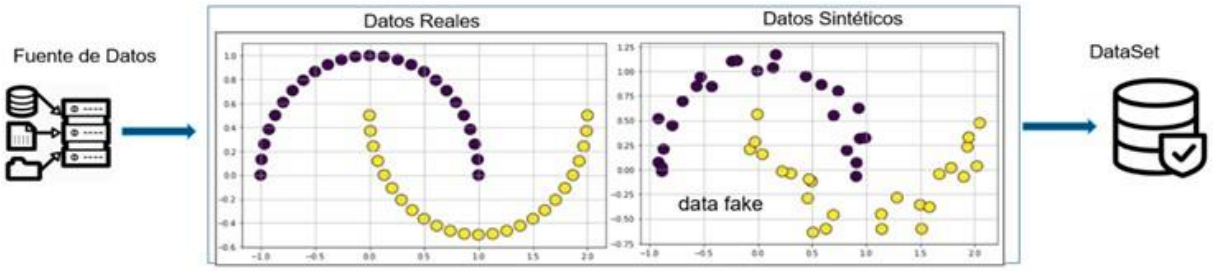

Fig. 2. Proceso de generación de datos sintéticos.

Los datos sintéticos se generan aleatoriamente partiendo de un conjunto de datos originales y siguiendo su estructura. Actualmente, existen 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 [14]. La figura 3 presenta el flujo iterativo con el que se etiquetan datos sintéticos mediante un pequeño conjunto de datos etiquetados conocido somo auto entrenamiento (self-training) para modelos semi-supervisados.

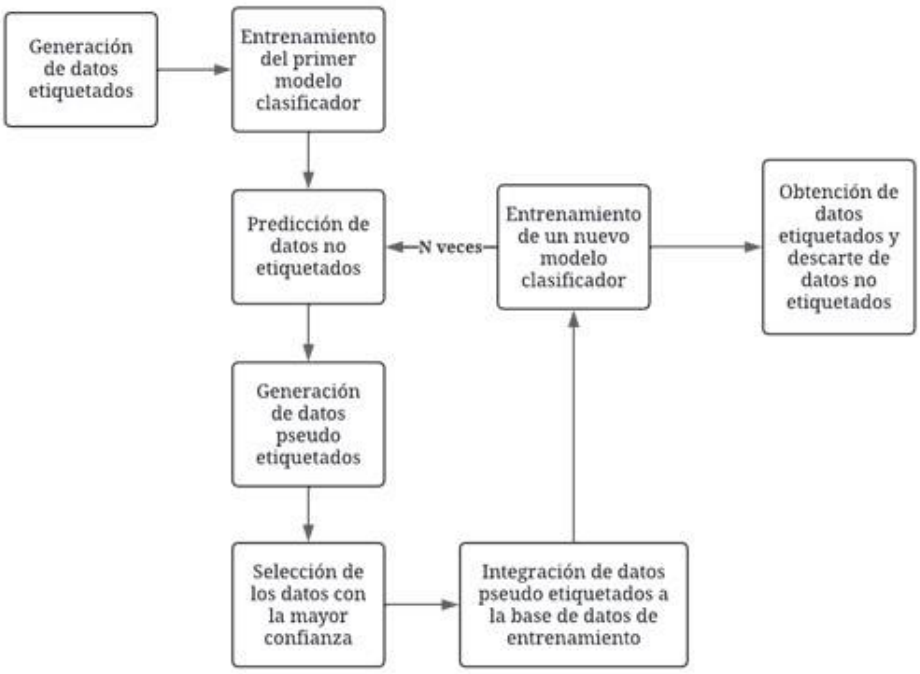

Fig. 3. Flujo para la creación de datos sintéticos.

III. Estado del arte

Dentro de la literatura reciente, se han realizado investigaciones relacionadas con la mejora de la generación e implementación de algoritmos entrenados. Algunos trabajos se enfocan en las primeras etapas del entrenamiento como la selección del mejor método considerando la eficacia de diferentes algoritmos de aprendizaje automático frente a un mismo set de datos [15]. Otros explican a un nivel más específico y técnico cómo mejorar la detección de anomalías combinando las fortalezas de diferentes algoritmos para la creación de sistemas de detección multifacéticas [16], mejorar las capacidades y atender las vulnerabilidades de algoritmos robustos. SVM [17] y el aprendizaje profundo [18], son ejemplos mediante los cuales conjuntos de procesos adicionales al modelo tradicional orientan al uso de nuevas e innovadoras técnicas de aprendizaje automático.

El presente trabajo busca ir más allá, considerando además de la evaluación de algoritmos, el tipo de enfoque que se va a utilizar según el contexto, limitaciones y necesidades en el que se realiza el entrenamiento (tamaño del set de datos, extensión de atributos, plazos de tiempo, eficacia esperada, etc).

Si bien existen trabajos que describen todas las técnicas de aprendizaje automático y cómo implementarlas según el caso de uso [19] se podría agregar a la clasificación “F. Conjunto de datos experimental” (F. Experimental Dataset).

Así, los enfoques que este trabajo menciona y cómo impactan en el proceso de aprendizaje automático considerando un escenario real, dan ventajas y áreas de oportunidad desde esta perspectiva. Esto debido a que en la mayoría de las empresas y organizaciones todavía no es común contar con un lago de datos (data lake) o almacén de datos (data warehouse) para poder procesar información. Así, se buscan resolver estas limitantes mediante técnicas como el self training, generación de datos sintéticos, técnicas de llenado de datos vacíos, tipos de normalización y diferentes métodos para generar datos sintéticos enfocados en las prioridades del contexto como son; plazos de entrega, calidad del algoritmo o calidad de la información.

IV. Objetivos y preguntas de investigación

El desarrollo de este trabajo se basa en las siguientes preguntas de investigación:

1. ¿Qué áreas/aspectos dentro del sistema de información son los que poseen más riesgo?
2. ¿Cómo se puede aplicar Ciencia de Datos en Ciberseguridad en las organizaciones para mejorar los procesos de detección y prevención de ataques de ciberseguridad?

Derivado de lo anterior, se definieron cuatro objetivos de investigación:

1) 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).
2) Diseñar un instrumento de medición estadístico para detectar posibles ataques orientados a tecnologías web.
3) Identificar dominios y segmentos de red por internet maliciosos que realicen ciberataques a la organización.
4) Automatizar estrategias de acción frente a ciberataques.

Con base en trabajos anteriores realizados en la misma área, [20] se buscó recuperar los registros de las bitácoras (logs) de los servidores, se analizaron datos sintéticos y se usó el lenguaje de programación Python para su transformación. Posteriormente se utilizó un conjunto de datos de entrenamiento y de pruebas que permitió desarrollar un algoritmo de aprendizaje automático para servir de apoyo en la toma de decisiones logrando así automatizar las tareas de detección de ataques y amenazas del departamento de ciberseguridad.

Científica, vol. 29, no. 2, pp. 01-20, July-December 2025,
ISSN 2594-2921, Instituto Politécnico Nacional, México
e290209 | DOI: 10.46842/ipn.cien.v29n2a09

V. Metodología

El desarrollo de este trabajo de investigación se basa en una metodología con enfoque en la detección de anomalías de ciberseguridad mediante el análisis de grandes cantidades de datos. Así, las etapas que compone esta metodología son [21]:

1. Etapa de fuente de datos: Identifica y selecciona las fuentes de datos relevantes para el análisis.
2. Etapa de extracción y carga de datos: Extrae los datos seleccionados de las fuentes identificadas y los deposita en un entorno de almacenamiento adecuado.
3. Etapa de recuperación de información: Realiza consultas y exploración de los datos almacenados para extraer conjuntos específicos que se utilizarán en el análisis.
4. Etapa de procesamiento de datos: Aplica técnicas de procesamiento de datos para detectar patrones, tendencias y posibles anomalías.
5. Etapa de construcción del modelo: Se desarrollan y entrenan los modelos de detección de anomalías utilizando técnicas de aprendizaje automático.
6. Etapa de evaluación y validación de resultados: Se Prueba el modelo construido utilizando un conjunto de datos de prueba con el fin de validar y evaluar su desempeño.
7. Etapa de presentación de resultados: Comunicar los hallazgos del análisis a través de informes, tableros (dashboards), y visualizaciones.
8. Etapa de entrega de productos de datos: Proporciona los modelos entrenados, los scripts, la documentación y cualquier otro producto de datos generado durante el proceso.

La figura 5 muestra el flujo de las etapas de desarrollo de la presente investigación.

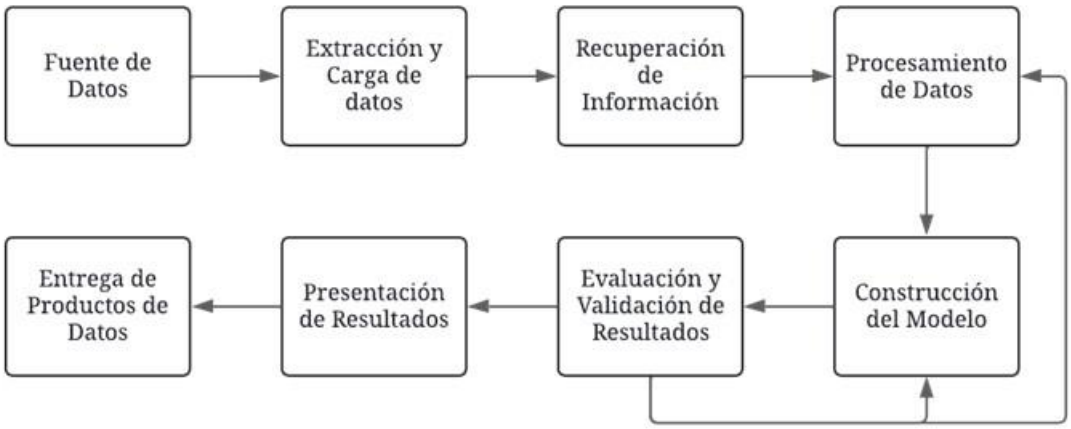

Fig. 4. Etapas de la metodología para el desarrollo de la investigación.

VI. Diseño de la investigación

Dentro del Instituto Nacional de Estadística y Geografía se cuenta con el esquema AIOps (Inteligencia artificial para operaciones de TI), este esquema busca integrar la Inteligencia Artificial y el uso de Big Data para que por medio de datos históricos recabados, se analicen y se determinen patrones que entrenen modelos de aprendizaje automático.

Además, se busca sustituir herramientas de operaciones de tecnologías de información (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. Actualmente, la organización 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.

Métodos y Materiales

Se instaló el software de procesamiento NoSQL Elasticsearch 8.19 en de la infraestructura del clúster Mictlán. Este clúster de servidores pertenece al Laboratorio de Ciencia de Datos & Ciberseguridad de la Universidad Autónoma de Aguascalientes. Esto mejoró considerablemente la capacidad y el tiempo de procesamiento de grandes cantidades de datos.

Los componentes de hardware y de software de la infraestructura del clúster Mictlán son:

- 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).
- Elasticsearch 8.19 (servidor de búsqueda).
- Kibana 8.19 (interfaz gráfica para Elasticsearch).
- Docker 8.19 (contenerización de servicios).
- Servicios SSH y FTP (terminales remotas y transferencia de archivos).
- Debian GNU/Linux 12 (sistema operativo base para los nodos del clúster).
- Plataforma de servicios de analítica y ciencia de datos.

La figura 6 muestra un diagrama de los componentes de hardware y de interconexión del clúster Mictlán.

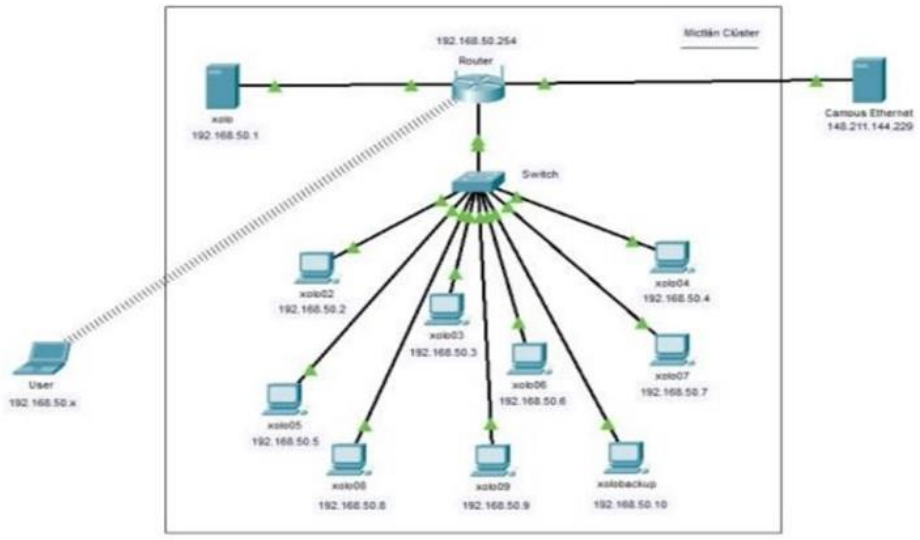

Fig. 5. Componentes de hardware y de interconexión del clúster Mictlán.

A. Diseño de experimento I

El primer grupo de datos que se analizó fue proporcionado por INEGI y generado a partir de datos del Departamento de Ciencia de Datos.

1) Transformación de los datos: Se documentó un Python Notebook para el proceso de decodificación y transformación de datos en crudo que inicialmente estaban en formato comprimido (.tar). La clasificación de los atributos que se obtuvo de la transformación de datos se estableció en ocho variables que se enlistan y definen en la tabla 1.

TABLA 1
ATRIBUTOS PLANEADOS, TIPOS DE VARIABLE Y SU DESCRIPCIÓN

Atributo	Tipo de Variable	Descripción
IP	Categórica	IP de la cual proviene la solicitud
Date	Continua	Fecha y hora en la que se realizó 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	Estatus 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

Como resultado de la limpieza y preprocesamiento de los datos se obtuvieron 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 limpios de aproximadamente 3,200,000 en total.

2) Análisis de los datos: Para el campo Date se eliminó la sección que define la hora global (-0600) dado que todos los registros tienen este valor y no aporta información valiosa a los objetivos del trabajo práctico. Para el campo Method se decidió resumir la cantidad de datos a aquellos que tengan los valores 400s y 500s de HTTP dado que denotan un error tanto en la solicitud del cliente como del servidor y es el objetivo para analizar.
3) Almacenamiento de los datos: Al evaluar los resultados se observaron picos en los valores de los 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, se concluyó 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.

B. Diseño de experimento II

Para este experimento, fue necesario el uso de datos sintéticos con el fin de generar registros suficientes y asegurar la calidad de los datos que contribuyeran al desarrollo de algoritmos entrenados satisfactoriamente. La figura 4 muestra el árbol de decisión con el que se concluyó que la mejor opción para entrenar y evaluar el algoritmo era el uso de archivos ficticios que en este caso llamamos “Datos Sintéticos” [22].

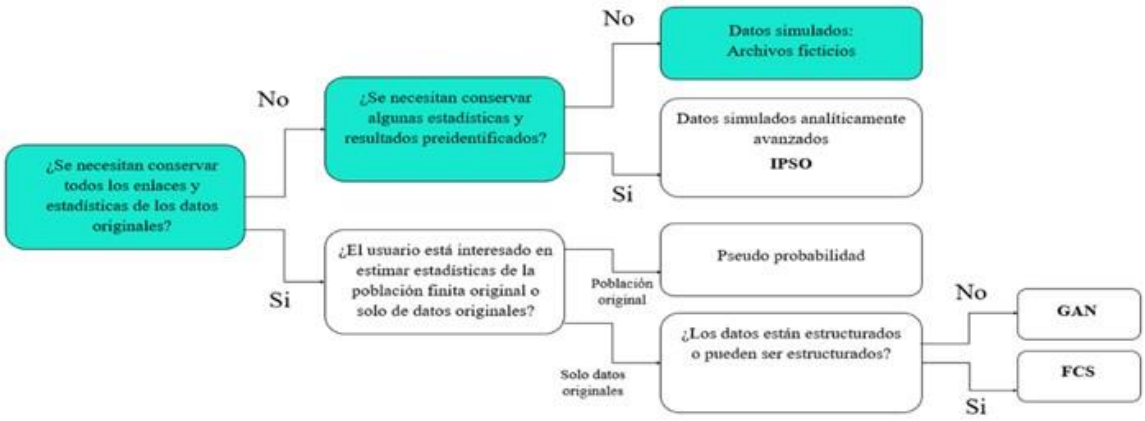

Fig. 6. Árbol de decisión de los datos a utilizar.

Se obtuvo un segundo grupo de datos en la página Kaggle llamada “DDOS Dataset” que es un compilado de otras bases de datos de logs reales etiquetados tanto como ataques DDoS como comportamiento normal de registros [23]. El archivo contaba con dos grupos de datos llamados "ddos_balanced" y "ddos_imbalanced". Cada uno de 6.6 y 4 gigas respectivamente con 33 atributos, además se optó por analizar ambos grupos de datos para comparar sus características.

a. El grupo de datos “ddos_balanced” era aquel con un concentrado de 50% registros etiquetados como ataques DDoS y 50% etiquetados como logs benignos.
b. El grupo de datos “ddos_imbalanced” era aquel con un concentrado de 20% registros etiquetados como ataques DDoS y 80% etiquetados como logs benignos.

1) Transformación de los datos: En la etapa de limpieza de datos para el grupo de datos “ddos_balanced” se filtraron 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” se filtraron 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 se consiguieron 20,500,000 registros con 33 atributos, esto permitió que se cumpliera con los requisitos de extensión de registros y de atributos para generar satisfactoriamente un algoritmo entrenado para detectar ataques DDoS.

En la etapa de selección de datos, se evaluó la significancia de cada atributo de ambos grupos de datos para filtrar los atributos y entrenar al algoritmo, ya que entrenarlo con los 33 atributos no era óptimo en tiempo y capacidades de procesamiento. La tabla 2 muestra los diferentes tipos de normalización utilizados.

TABLA 2. COMPARATIVA DE RESULTADOS DE ENTRENAMIENTO.

Método	***Descripción***
Original	El grupo de datos con los atributos sin normalizar.
Log Transformation	Normalización por Transformación Logarítmica.
Min Max	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.
Z score	Normalización por puntuación de Z.

2) Análisis de los datos: Para el grupo de datos "ddos_imbalanced" se seleccionaron registros con un valor superior al 0.5 de significancia por atributo. Para el grupo de datos "ddos_balanced" se seleccionaron registros con un valor superior al 0.5 de significancia por atributo, resultando en 9 atributos con el valor más alto para el método de Mínimos y Máximos que fue el que obtuvo mayores puntuaciones con diferencia.

Se eligió 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 con la fórmula (1).

$$z_i = x_i - \min(x) \,/\, \max(x) - \min(x) \tag{1}$$

De un total de 15 variables candidatas se eligieron 8 con la mayor calificación positiva, 1 con la mayor calificación negativa y la variable "Label" como variable de control, dando como total 10 variables (Tabla 3).

TABLA 3. VARIABLES Y SU SIGNIFICANCIA.

Número	***Variable***	***Calificación***
1	Protocol	-0.778777
2	Tot Fws Pkt	0.705499
3	TotLen Fwd Pkts	0.693646
4	Fwd Pkt Len Max	0.692462
5	Fwd Header Len	0.704898
6	Down/Up Ratio	0.87525
7	Fwd Seg Size Avg	0.684889
8	Subflow Fwd Pkts	0.705499
9	Subflow Fwd Byts	0.693646
10	Label	NA

3) Entrenamiento algoritmo: Para el entrenamiento se consideraron los algoritmos más utilizados en ataques DDoS [24]:

- Árboles de Decisión
- Bosques aleatorios
- Máquina de Vectores de Soporte (SVM)
- Redes neuronales

Los algoritmos SVM y Redes Neuronales fueron excluidos de las pruebas debido a que no se cuenta con equipo de procesamiento óptimo para implementarlos en tiempos establecidos para el proyecto. Sin embargo, se agregan a la tabla para ser considerados en caso de poder ser implementados en el futuro.

TABLA 4
COMPARATIVA DE RESULTADOS DE ENTRENAMIENTO

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

Se desarrollaron algoritmos entrenados mediante la técnica de bosques aleatorios tomando en cuenta valores ligeramente más altos en comparación con la técnica de árboles de decisión. Para cada prueba en la tabla se consideró el valor más alto obtenido con una precisión promedio del 95%. (Tabla 5).

TABLA 5
COMPARATIVA DE PRECISIÓN ALGORITMO GENERADO VS DATOS SINTÉTICOS

	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%

Científica, vol. 29, no. 2, pp. 01-20, July-December 2025,
ISSN 2594-2921, Instituto Politécnico Nacional, MÉXICO
e290209 | DOI: 10.46842/ipn.cien.v29n2a09

Para realizar una evaluación más exhaustiva se buscó generar datos sintéticos de diferentes formas e ingresarlas al algoritmo entrenado para medir su precisión, otros métodos para generar datos sintéticos fueron descartados por la imposibilidad de generar registros útiles para evaluar al algoritmo entrenado (Tabla 6).

TABLA 6
COMPARATIVA DE PRECISIÓN POR DATOS SINTÉTICOS DE DIFERENTES FUENTES

	Normal	***Scikit-learn***	***Faker***	***Tonic***	***MostAI 96%***	***MostAI 96% sin Sobre entrenamiento***
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

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

Nota: En la sección de bibliografía, se adjunta el enlace para la consulta de código utilizado.

C. Diseño de experimento III

En un tercer experimento, se buscó tener datos con comportamiento real de una página web. Para esto, se facilitaron registros por parte de los servidores de la Universidad Autónoma de Aguascalientes y se descargaron alrededor de 5 millones de registros de logs de diferentes páginas de la universidad en formatos .log.

1) Transformación de los datos: Dada la estructura diferente de los registros y a la estandarización de atributos se eliminaron registros con información incompleta obteniendo cerca de 3 millones de registros limpios. Además, con fines de complementar la información se generaron dos nuevos atributos para cada registro: Seg_Ant y Seg_AntIP.
En una segunda limpieza de registros se eliminaron aquellos con información con errores de formato y caracteres especiales, se finalizó con los siguientes atributos.

TABLA 7
ATRIBUTOS Y SU DESCRIPCIÓN DEL GRUPO DE DATOS III

Atributo	*Descripción*
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.

Además, se omitieron dos atributos relacionados a detalles del navegador utilizado por no considerarse relevantes en el análisis de información. La figura 7 muestra siete de los atributos con mayor relevancia.

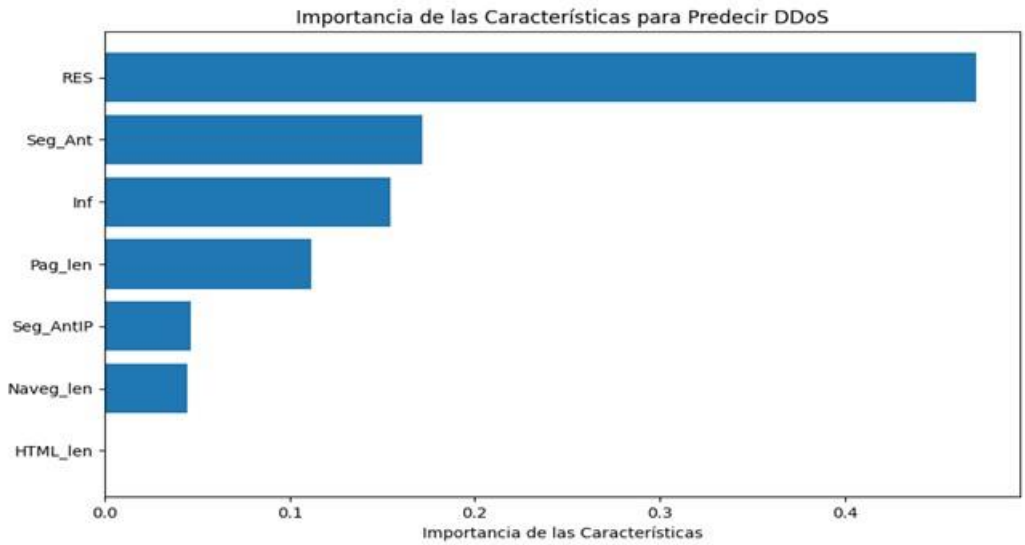

Fig. 7. Relevancia de los atributos para predecir DDOS.

2) Análisis de los datos: En la fase de análisis de datos se presentó un inconveniente, los registros no estaban etiquetados como ataques DDOS y logs normales. Así pues, se decidió que mediante la metodología de auto entrenamiento (self-training), utilizando un conjunto de datos pequeño, se etiquetarían los registros del grupo de datos actual. Se descartaron las bases de datos utilizadas anteriormente (experimento I y II) debido a que no contaban con la misma estructura de la información ni los mismos atributos.

Se obtuvo el siguiente grupo de datos de Kaggle llamada "WordPress DDos Log Dataset" [25] que contiene los mismos atributos que el grupo de datos que se estaba manejando (Tabla 7) y cuenta con las etiquetas de logs malignos y benignos.

Para la clasificación mediante Self-training se consideró un rango de confianza del 90% para integrar los registros en una nueva ronda. En total se realizaron nueve (9) rondas en las que de los 2 millones y medio de registros se clasificaron exitosamente casi 2 millones como malignos o benignos.

3) Entrenamiento del algoritmo: Para poder normalizar el grupo de datos y entrenar un modelo se realizaron los siguientes ajustes a los atributos del grupo de datos:

TABLA 8
ATRIBUTOS Y SU DESCRIPCIÓN PARA ENTRENAR AL ALGORITMO

Atributo	*Descripción*
IP	Se eliminaron los puntos de las direcciones manejando un solo número entero. Posteriormente se eliminó por considerarse no relevante para el entrenamiento.
Seg_Ant	Se mantuvo el mismo formato.
Seg_AntIP	Se mantuvo 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	Se mantuvo el mismo formato.
Inf	Se mantuvo 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. Se realizó un análisis de la relevancia de cada uno de los atributos para detectar si un registro era o no un ataque DDOS. Además, para el entrenamiento del algoritmo elegimos el método de "Árboles de Decisión" y adicionalmente se implementó un código para evitar el sobre entrenamiento o "overfitting" del modelo.

La figura 8 muestra una precisión con promedio un del 97% en 10 entrenamientos diferentes utilizando datos de prueba.

Científica, vol. 29, no. 2, pp. 01-20, July-December 2025,
ISSN 2594-2921, Instituto Politécnico Nacional, MÉXICO
e290209 | DOI: 10.46842/ipn.cien.v29n2a09

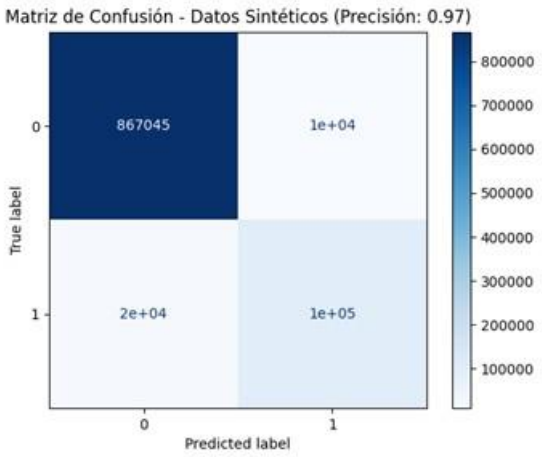

Fig. 8. Matriz de confusión con datos de prueba.

Para una evaluación más exhaustiva se generó un millón de datos sintéticos usando nuevamente la página de Mostly.AI con una precisión del 95% mostrándose los valores en la figura 9. Se obtuvo una precisión con un promedio del 97% evaluando la predicción del modelo entrenado previamente. La figura 9 muestra una precisión promedio un del 97% en 10 entrenamientos diferentes utilizando los datos sintéticos.

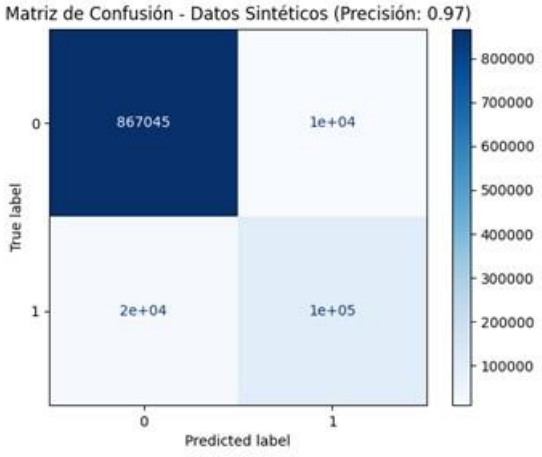

Fig. 9. Matriz de confusión con datos sintéticos.

VII. Resultados y Discusión

La tabla 9 muestra una comparación entre los experimentos desarrollados, los tipos de datos utilizados y los algoritmos implementados.

TABLA 9
COMPARACIÓN ENTRE LOS EXPERIMENTOS DESARROLLADOS

	Experimento I	Experimento II	Experimento III
Tipo de datos utilizados	Nativos con formato comprimido (.tar)	Nativos con formato .csv y sintéticos	Nativos con formato .log y sintéticos
Algoritmos	Árboles de decisión	Bosques aleatorios	Árboles de decisión
Ventajas	Se enfatizó en la velocidad de procesamiento de grandes cantidades de datos al procesarlos en formato .tar con pocos atributos para entrenar al algoritmo y al priorizar aquellos registros que demostraran un comportamiento anómalo (error 500 y 400).	Se contó con una gran cantidad de registros etiquetados tanto benignos como malignos además de una amplia variedad de atributos de los cuales se obtuvieron un conjunto de nueve atributos de alta relevancia lo cual asegura un alcance realista para generar un algoritmo de entrenamiento eficaz.	Se procesaron registros reales de un sitio web similar al que se implementaría el algoritmo con una extensión de registros y atributos suficiente para obtener un entrenamiento eficaz. La generación de datos sintéticos, así como la evaluación del algoritmo entrenado demostraron una alta eficacia en la identificación de ataques DDOS.
Desventajas	La cantidad limitada de atributos y el enfoque en los registros con errores puede llevar a generar un algoritmo sesgado por la poca información global con la que se entrena, afectando la eficacia para la detección de ataques DDOS reales.	A pesar de contar con registros etiquetados como malignos y benignos no se reflejaba el comportamiento real de logs de un sitio web lo cual sesgaba el algoritmo. Asimismo, considerando la implementación de este método se observó que los registros del sitio web no eran los mismos por lo que en caso de éxito en el entrenamiento al momento de implementarlo no contaría con los datos necesarios reduciendo su eficacia enormemente.	Al limitar los atributos a los utilizados en un sitio web se acorta su utilización para otras páginas. Sin embargo, considerando que por ahora se busca crear un algoritmo prototipo que puliera los procesos para generar otros algoritmos posteriormente no se considera una problemática real. La cantidad de registros reales pudo haber sido más amplia para mejorar el alcance del algoritmo, esto se solucionó generando datos sintéticos.
Observaciones	Para que este experimento fuera exitoso sería necesario contar con registros con mayor cantidad de atributos que posean alta relevancia y considerar registros anómalos diferentes a aquellos con error 400 y 500 para ampliar el alcance del algoritmo entrenado.	Para que este experimento fuera exitoso sería necesario distribuir los registros de manera que reflejen el comportamiento de ataques DDOS reales, así como considerar los atributos que se clasificarían una vez implementado en el sitio web	Para mejorar la calidad de este experimento sería necesario contar con una estructura formal de los registros a procesar en todos los sitios web que se busca implementar (contar con un almacén de datos -data warehouse). La calidad de un algoritmo entrenado es directamente proporcional a la calidad de los datos con los que se entrena.

Científica, vol. 29, no. 2, pp. 01-20, July-December 2025,
ISSN 2594-2921, Instituto Politécnico Nacional, MÉXICO
e290209 | DOI: 10.46842/ipn.cien.v29n2a09

Así, se consideraron tres enfoques diferentes considerando los aspectos principales de: Velocidad de entrenamiento, cantidad de datos y calidad del entrenamiento.

Experimento 1: Enfoque en anomalías. En este experimento se sacrifica alcance y se corre un mayor riesgo de obtener un algoritmo sesgado con la ventaja de necesitar menos registros y atributos. Además, se segura un entrenamiento mucho más rápido al no evaluar otros registros benignos o variantes de registros para otros tipos de ataques de ciberseguridad.

Experimento 2: Enfoque en demasía. En este experimento se sacrifica tiempo y simplicidad para considerar la totalidad del universo de uno o más conjuntos de datos bajo la premisa de que a mayor cantidad de datos es mejor la calidad del algoritmo. Sin embargo, se corre el riesgo de ampliar excesivamente los tiempos de procesamiento y entrenamiento si se desea respetar siempre el comportamiento y continuidad de la información.

Experimento 3: Enfoque en definición. En este experimento se realiza un algoritmo a la medida según la información que recaba la página y/o páginas web en las que se busca implementar el algoritmo. Se establece desde el inicio el alcance y los conjuntos de datos a utilizar corriendo el riesgo de que, si no se tienen procesos formales de procesamiento de información internos, no se obtendrá la calidad necesaria en la información para poder entrenar eficazmente el algoritmo y por consecuencia su implementación.

Para nuestro caso de estudio, el mejor enfoque es el más equilibrado en cuanto a velocidad de entrenamiento (prueba en varios modelos y tiempos establecidos) y la cantidad de datos necesarios (logs y atributos limitados de la página web). Para la elección del entrenamiento del algoritmo se optó por utilizar el método de árboles de decisión por su ligero aumento en la exactitud. Con los atributos normalizados y adaptados para aprendizaje automático se obtuvo un promedio del 96% en la precisión del algoritmo, esto compuesto de un 97% por medio datos reales y de un 95% por medio datos sintéticos.

VIII. Conclusiones

El desarrollo de este trabajo de investigación permite identificar importantes lecciones sobre la relevancia de la ciencia de datos en el ámbito de la Ciberseguridad. Además, permite entender la complejidad inherente de 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. La obtención de los datos, su limpieza y su uso, son actividades críticas para obtener resultados precisos y evitar eventualidades a lo largo del entrenamiento y evaluación de los modelos.

Se constató que, para llevar a cabo un proyecto exitoso en ciencia de datos, es necesario tener claridad sobre la procedencia de los datos y la forma del tratamiento de estos hasta el resultado final del proyecto. Además, se deben considerar diversas metodologías de análisis y preprocesamiento de datos y explorar diferentes enfoques antes de determinar el más adecuado para cada caso en particular.

El uso de técnicas avanzadas para evaluar y mejorar el rendimiento del modelo, 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 sobre ajuste, permiten asegurar una mayor calidad y robustez en el algoritmo final.

Científica, vol. 29, no. 2, pp. 01-20, July-December 2025,
ISSN 2594-2921, Instituto Politécnico Nacional, México
e290209 | DOI: 10.46842/ipn.cien.v29n2a09

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.

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 es 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.

ANEXOS

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

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

CRediT (Contributor Roles Taxonomy)

Contribuciones de los autores: Conceptualización: LE**BV**; Metodología: E**OD**; Software: PA**MV**; Investigación: PA**MV**; Redacción y preparación del borrador original: PA**MV**; Redacción, revisión y edición: J**ML**; Supervisión: LE**BV**; Análisis formal: E**OD**; Administración del proyecto: LE**BV**; Adquisición de fondos: LE**BV**.

Financiamiento: Esta investigación recibió el apoyo de la Universidad Autónoma de Aguascalientes como parte del proyecto interno de investigación PIINF23-2.

Declaración de disponibilidad de datos: Los datos se encuentran en el artículo.

Agradecimientos: Los autores agradecen a la Universidad Autónoma de Aguascalientes y al Instituto Nacional de Estadística (INEGI) por el acceso a la información necesaria para llevar a cabo esta investigacion.

Conflicto de interés: Los autores declaran que no existe conflicto de interés.

REFERENCIAS

[1] B. Marr, “How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read,” 2018. [Online]. Available: 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

[2] B. Marr, “How Much Data Is There In the World? | Bernard Marr,” 2021. [Online]. Available: https://bernardmarr.com/how-much-data-is-there-in-the-world/

[3] M. M. Alani, “Big data in cybersecurity: a survey of applications and future trends,” *J Reliab Intell Environ*, vol. 7, no. 2, pp. 85–114, 2021, doi: 10.1007/s40860-020-00120-3

[4] W. E. Forum, “The Global Risks Report 2020 Insight Report 15th Edition,” 2020. [Online]. Available: https://www3.weforum.org/docs/WEF_Global_Risk_Report_2020.pdf

[5] W. E. Forum, Global Riks Report 2023 18th Edition. 2023. [Online]. Available: https://www3.weforum.org/docs/WEF_Global_Risks_Report_2023.pdf

[6] Reuters, "Cibercrimen ha costado 6 millones de dólares a las economías del mundo," 2022. [Online]. Available: https://www.eleconomista.com.mx/tecnologia/Cibercrimen-ha-costado-6-millones-de-dolares-a-las-economias-del-mundo-20220511-0022.html

[7] G. Vicario, S. Coleman, "A review of data science in business and industry and a future view," Appl Stoch Models Bus Ind, 2019, doi: 10.1002/ASMB.2488

[8] A. Monnappa, "Data Science vs. Big Data vs. Data Analytics," 2022. [Online]. Available: https://www.simplilearn.com/data-science-vs-big-data-vs-data-analytics-article#what_is_data_science

[9] J. J. Beunza, E. P. Sanz, N. E. C. Moreno, "Manual práctico de inteligencia artificial en entornos sanitarios," Elsevier Health Sciences, pp. 35–39, 2020, [Online]. Available: https://books.google.com.mx/books?id=88nSDwAAQBAJ&dq=algoritmos+supervisados&lr=&hl=es&source=gbs_navlinks_s

[10] C. Rosenberg, M. Hebert, H. Schneiderman, "Semi-Supervised Self-Training of Object Detection Models," 2004. [Online]. Available: https://kilthub.cmu.edu/articles/journal_contribution/Semi-Supervised_Self-Training_of_Object_Detection_Models/6560834?file=12043121

[11] P. W. Singer, A. Friedman, "Cybersecurity: What Everyone Needs to Know," Oxford University Press, p. 224, 2014, [Online]. Available: https://books.google.com/books?id=9VDSAQAAQBAJ&pgis=1

[12] C. Longbing, Y. Philip S, "Data Science Thinking: The Next Scientific, Technological and Economic Revolution," Springer, 2018, [Online]. Available: http://www.springer.com/series/15063

[13] I. H. Sarker, A. S. M. Kayes, S. Badsha, H. Alqahtani, P. Watters, A. Ng, "Cybersecurity data science: an overview from machine learning perspective," *J Big Data*, vol. 7, no. 1, pp. 1–29, 2020, doi: 10.1186/s40537-020-00318-5

[14] J. L. Becerra Pozas, "¿Qué son los datos sintéticos? Datos generados para ayudar a su estrategia de IA," CIO, 2023.

[15] N. Elmrabit, F. Zhou, F. Li, H. Zhou, "Evaluation of Machine Learning Algorithms for Anomaly Detection," in 2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security), 2020, pp. 1–8. doi: 10.1109/CyberSecurity49315.2020.9138871

[16] A. Yaseen, "The Role of Machine Learning in Network Anomaly Detection for Cybersecurity." [Online]. Available: https://journals.sagescience.org/index.php/ssraml/article/view/126

[17] T. Shon, J. Moon, "A hybrid machine learning approach to network anomaly detection," Inf Sci (N Y), vol. 177, no. 18, pp. 3799–3821, 2007, doi: 10.1016/j.ins.2007.03.025

[18] T. T. Teoh, G. Chiew, E. J. Franco, P. C. Ng, M. P. Benjamin, Y. J. Goh, "Anomaly detection in cyber security attacks on networks using MLP deep learning," in 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE), 2018, pp. 1–5. doi: 10.1109/ICSCEE.2018.8538395

[19] S. Wang, J. F. Balarezo, S. Kandeepan, A. Al-Hourani, K. G. Chavez, B. Rubinstein, "Machine Learning in Network Anomaly Detection: A Survey," IEEE Access, vol. 9, pp. 152379–152396, 2021, doi: 10.1109/ACCESS.2021.3126834

[20] L. E. Ostos Ríos, L. E. Bautista Villalpando, J. Muñoz López, E. Oswaldo Díaz, "Análisis de grandes cantidades de datos por medio de técnicas de máquinas de aprendizaje para la Ciberseguridad," 2020.

[21] E. A. Villasenor Garcia *et al*., "Data Lake Strategy for Data Science Workflows," in 2022 11th International Conference On Software Process Improvement (CIMPS), IEEE, 2022, pp. 219–223. doi: 10.1109/CIMPS57786.2022.10035694

[22] United Nations Economic Commission for Europe, *Synthetic Data for Official Statistics. United Nations*, 2023. doi: 10.18356/9789210021708

[23] D. Devendra, "DDoS Dataset," 2019. [Online]. Available: https://www.kaggle.com/datasets/devendra416/ddos-datasets

[24] J. T. Viteri, M. I. Valero, A. Torres, N. M. Torres, *Seguridad contra ataques DDoS en los entornos SDN con Inteligencia Artificial*, 3rd ed., vol. 7. RMC, 2022.

[25] A. Toluwalase, "WordPress DDos Log Dataset," 2023. [Online]. Available: https://www.kaggle.com/datasets/ajiboyetoluwalase/wordpress-ddos-log-dataset?select=Wordpress+DDOS+attack+Logs.txt

Científica, vol. 29, no. 2, pp. 01-20, July-December 2025,
ISSN 2594-2921, Instituto Politécnico Nacional, MÉXICO
e290209 | DOI: 10.46842/ipn.cien.v29n2a09