UNIVERSIDAD ESTATAL PENÍNSULA DE SANTA ELENA
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES

TITULO DEL TRABAJO DE TITULACIÓN

VOLCADO DE TRÁFICO WEB, RECOLECCIÓN Y ANÁLISIS DE DATOS APLICANDO ALGORITMOS DE CLASIFICACIÓN DE INTELIGENCIA ARTIFICIAL EN REDES

AUTOR

Vera Ricardo, Joseline Katiuska

TRABAJO DE INTEGRACIÓN CURRICULAR

Previo a la obtención del grado académico en
INGENIERA EN TECNOLOGÍAS DE LA INFORMACIÓN

TUTOR

Lsi. Daniel Quirumbay Yagual, MSIA.

Santa Elena, Ecuador

Año 2024

UPSE

UNIVERSIDAD ESTATAL PENÍNSULA DE SANTA ELENA
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES

TRIBUNAL DE SUSTENTACIÓN

Ing. José Sánchez Aquino, Mgt.
DIRECTOR DE LA CARRERA

Lsi. Daniel Quirumbay Yagual, Mgt.
TUTOR

Ing. Iván Coronel Suárez, Mgt.
DOCENTE ESPECIALISTA

Ing. Marjorie Coronel Suárez, Mgt.
DOCENTE GUÍA UIC

UNIVERSIDAD ESTATAL PENÍNSULA DE SANTA ELENA
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES

CERTIFICACIÓN

Certifico que luego de haber dirigido científica y técnicamente el desarrollo y estructura final del trabajo, este cumple y se ajusta a los estándares académicos, razón por el cual apruebo en todas sus partes el presente trabajo de titulación que fue realizado en su totalidad por **Vera Ricardo Joseline Katiuska**, como requerimiento para la obtención del título de Ingeniero en Tecnologías de la Información.

La Libertad, a los 04 días del mes de diciembre del año 2024

TUTOR

Firmado electrónicamente por:
DANIEL IVAN
QUIRUMBAY YAGUAL

Lsi. Daniel Quirumbay Yagual

UNIVERSIDAD ESTATAL PENÍNSULA DE SANTA ELENA
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES

DECLARACIÓN DE RESPONSABILIDAD

Yo, **Joseline Katiuska Vera Ricardo**

DECLARO QUE:

El trabajo de Titulación, "Volcado de tráfico web, recolección y análisis de datos aplicando algoritmos de clasificación de inteligencia artificial en redes", previo a la obtención del título en Ingeniera en Tecnologías de la Información, ha sido desarrollado respetando derechos intelectuales de terceros conforme las citas que constan en el documento, cuyas fuentes se incorporan en las referencias o bibliografías. Consecuentemente este trabajo es de mi total autoría.

En virtud de esta declaración, me responsabilizo del contenido, veracidad y alcance del Trabajo de Titulación referido.

La Libertad, a los 04 días del mes de diciembre del año 2024

EL AUTOR

Joseline Vera Ricardo

UNIVERSIDAD ESTATAL PENÍNSULA DE SANTA ELENA

FACULTAD DE SISTEMAS Y TELECOMUNICACIONES

CERTIFICACIÓN DE ANTIPLAGIO

Certifico que después de revisar el documento final del trabajo de titulación denominado Volcado de tráfico web, recolección y análisis de datos aplicando algoritmos de clasificación de inteligencia artificial en redes, presentado por la estudiante, Joseline Vera Ricardo, fue enviado al Sistema Antiplagio, presentando un porcentaje de similitud correspondiente al 5%, por lo que se aprueba el trabajo para que continúe con el proceso de titulación.


CERTIFICADO DE ANÁLISIS
magister

TI_VeraRicardoJoseline2

5%
Textos sospechosos

2% Similitudes
< 1% similitudes entre comillas
< 1% entre las fuentes mencionadas
3% Idiomas no reconocidos
6% Textos potencialmente generados por la IA (ignorado)

Nombre del documento: TI_VeraRicardoJoseline2.docx
ID del documento: d012d2d42bead586beb076aaabe4e35e2e6d002e
Tamaño del documento original: 6,19 MB
Autores: []

Depositante: DANIEL IVAN QUIRUMBAY YAGUAL
Fecha de depósito: 3/12/2024
Tipo de carga: interface
fecha de fin de análisis: 3/12/2024

Número de palabras: 16.898
Número de caracteres: 115.095

TUTOR

Firmado electrónicamente por:
DANIEL IVAN
QUIRUMBAY YAGUAL

Lsi. Daniel Quirumbay Yagual

UNIVERSIDAD ESTATAL PENÍNSULA DE SANTA ELENA
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES

AUTORIZACIÓN

Yo, **Joseline Vera Ricardo**

Autorizo a la Universidad Estatal Península de Santa Elena, para que haga de este trabajo de titulación o parte de él, un documento disponible para su lectura consulta y procesos de investigación, según las normas de la Institución.

Cedo los derechos en línea patrimoniales de artículo profesional de alto nivel con fines de difusión pública, además apruebo la reproducción de este artículo académico dentro de las regulaciones de la Universidad, siempre y cuando esta reproducción no suponga una ganancia económica y se realice respetando mis derechos de autor

Santa Elena, a los 04 días del mes de diciembre del año 2024

EL AUTOR

Joseline Vera Ricardo

AGRADECIMIENTO

Quiero expresar mi más profundo agradecimiento a todas las personas que hicieron posible la culminación de esta tesis. En primer lugar, agradezco a Dios por darme la fortaleza y sabiduría para superar los desafíos a lo largo de este camino.

A mi tutor, por su guía, paciencia y conocimiento compartido, que enriquecieron este trabajo y me motivaron a buscar la excelencia en cada detalle.

A mi familia, especialmente a mis padres, por su amor incondicional, sus sacrificios y su constante aliento, sin los cuales no habría sido posible llegar a estas instancias. A mis amigos quienes con sus palabras de aliento y compañía me ayudaron a mantenerme enfocado durante los momentos difíciles.

Joseline Katiuska, Vera Ricardo

DEDICATORIA

Dedico esta tesis con todo mi amor y gratitud a mi familia, en especial a mis padres, quienes han sido mi pilar de fortaleza y motivación desde el inicio de este camino. Su apoyo incondicional, sacrificios y fe en mí han sido primordial para darme el impulso de seguir adelante incluso en los momentos más desafiantes.

Finalmente, dedico este trabajo a todas las personas que creen en el poder de la educación como herramienta para transformar vidas y construir un mejor futuro. Este esfuerzo no solo es mío, sino también de quienes han caminado conmigo en este proceso.

Joseline Katiuska, Vera Ricardo

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

El trabajo de titulación con el tema "Volcado de tráfico web, recolección y análisis de datos aplicando algoritmos de clasificación de inteligencia artificial en redes", tuvo como objetivo desarrollar algoritmos en lenguaje Python, utilizando bibliotecas especializadas para realizar el rastreo de redes alámbricas e inalámbricas, volcado del tráfico web y la recolección de información, para examinar las redes empleadas en FACSISTEL. Se empleó la metodología de investigación exploratoria y diagnóstica; además de metodología ISAAF y OMSTD para el desarrollo. Se aplicaron técnicas de recolección, como la entrevista al director del departamento de TI y ficha de observación en la institución. Los resultados indicaron que el uso de algoritmos como XGBoost, Isolation Forest y Random Forest permitió el análisis satisfactorio del tráfico web del Data set; se puede concluir que, el Dashboard facilitó la interpretación visual de los datos, optimizando la toma de decisiones con respecto a la gestión del tráfico web.

Palabras claves: algoritmos de clasificación, dashboard, tráfico web.

ABSTRACT

The degree work with the topic "Web traffic dump, data collection and analysis applying artificial intelligence classification algorithms in networks", aimed to develop algorithms in Python language, using specialized libraries to track wired and wireless networks. , dumping web traffic and information collection, to examine the networks used in FACSISTEL. The exploratory and diagnostic research methodology was used; in addition to ISAAF and WHOSTD methodology for development. Collection techniques were applied, such as the interview with the director of the IT department and an observation sheet at the institution. The results indicated that the use of algorithms such as XGBoost, Isolation Forest and Random Forest allowed the satisfactory analysis of the web traffic of the Dataset; It can be concluded that the Dashboard facilitated the visual interpretation of the data, optimizing decision making regarding web traffic management.

Keywords: classification algorithms, dashboard, web traffic.

INTRODUCCIÓN

La propuesta está basada en una institución de educación superior, que posee siete facultades y alrededor de 27 carreras, las cuales se establecen debido a su expansión dentro de la matriz de la universidad y también en las instalaciones de la sede en Playas. Debido a su gran dimensión, cada una de las carreras se divide en diversos sectores de acuerdo con la facultad a la que pertenezcan. En este contexto, se escogió a la Facultad de Sistemas y Telecomunicaciones, la cual destaca la presencia de varias conexiones inalámbricas y alámbricas, tanto abiertas como privadas que coexisten en el mismo sector.

Las redes inalámbricas institucionales muestran lentitud a causa de intermitencias, saturación y solapamiento de canales, lo que puede dificultar los análisis si no se reconocen los problemas. Las redes utilizan mecanismos de protección que restringen el acceso a diversas páginas web. Adicionalmente, las redes alámbricas presentan caídas en el servicio y una elevada latencia, generando ruido en los análisis vinculados a la infraestructura. En cambio, el departamento de Tecnología de la Información de la institución se enfrenta a diario a numerosos intentos de acceso a páginas maliciosas debido al gran número de estudiantes conectados a la red universitaria. Este departamento dispone de técnicas de seguridad para reducir estas conexiones, aunque no son totalmente efectivas. Así mismo, aunque cuentan con un sistema de rastreo de páginas maliciosas, presenta deficiencias debido a que es antiguo y tiene una limitada efectividad de los sistemas.

Debido a estas problemáticas, se plantea la creación de algoritmos que permitan realizar el análisis de las redes que hay en la Facultad de Sistemas y Telecomunicaciones de la Institución de Educación Superior, realizando un mapeo de las redes dentro de las instalaciones por medio de un Data set que proporcionó la entidad, con el fin de tener una visualización de la dimensión del tráfico web que se encontrará en el sector, las cuales se monitorizarán posteriormente.

Para la elaboración de la propuesta, se utiliza la metodología de tipo diagnóstica, permitiendo identificar a través de técnicas de recolección de datos, la situación actual de las redes mediante mecanismos de monitoreo. Además, se aplica la

metodología exploratoria para recopilar trabajos relacionados con el análisis de redes a través de algoritmos de machine learning.

Por otra parte, se emplea la metodología ISAAF para ofrecer precisión integridad y eficiencia en las evaluaciones de seguridad, la cual consta de las fases: planificación, desarrollo, monitoreo, evaluación y reportes. Además, se utiliza la metodología OMSTD para trabajar con Python, estableciendo los siguientes subtemas: organización y estructura; interacción; entrada y salida de información; redistribución y despliegue.

Con respecto a la arquitectura del sistema, esta permite la gestión y análisis de los datos provenientes del Dataset, procesando los paquetes de la red educativa para almacenar y visualizar en un Dashboard. Así mismo, utiliza un proceso en Python, almacenamiento en la base de datos MySQL y brinda herramientas para la generación de reportes y gráficas.

El presente trabajo, se estructura de la siguiente forma:

El capítulo I se centra en el contexto y objetivos del proyecto, incluyendo antecedentes y justificación. Además, se describe la metodología de investigación, beneficiarios, variables de estudio y análisis de recolección de datos. Por otra parte, se elabora la metodología de desarrollo aplicada en la propuesta.

En el capítulo II se abarca el marco contextual del proyecto, abordando además el marco conceptual, marco teórico, requerimientos, arquitectura del sistema y desarrollo del trabajo.

Finalmente, se elaboran las conclusiones y recomendaciones, resumiendo los hallazgos del estudio y ofreciendo las sugerencias correspondientes para futuras investigaciones o mejoras en el sistema.

CAPÍTULO 1. FUNDAMENTACIÓN

1.1. Antecedentes

A nivel global, hay un sin número de empresas que adoptan a las redes inalámbricas como un medio fundamental de transmisión y acceso a internet para sus propios usuarios, los cuales realizan actividades mediante el uso de distintos dispositivos móviles como ordenadores portátiles, asistentes digitales, celulares, entre otros [1]. Al existir una gran cantidad de equipos intentando ingresar y tener una navegación más rápida, provoca una alta demanda de acceso a la red, la cual, si no tiene herramientas de control adecuado puede crear una deficiencia en la comunicación [1].

Así mismo, se está viviendo una evolución de las redes de comunicaciones hacia redes de velocidad alta, en donde ingresan las redes alámbricas, siendo un tipo de red informática que emplea cables para la conexión de dispositivos; uno de los inconvenientes es la congestión, que ocurre cuando la demanda de datos supera la capacidad del ancho disponible de banda, causando lentitud y pérdida de paquetes; además, la latencia se presenta debido a la distancia, problemas en el hardware o sobrecarga de switches y routers [2].

La propuesta se centra en una Institución de Educación Superior, que obtuvo la aprobación de su creación en septiembre de 1995 por el comité de gestión y presentado ante el Congreso Nacional, para luego ser aprobado por el mismo, en junio de 1996; inició con cuatro facultades, pero en la actualidad dentro de su oferta académica existen siete facultades y alrededor de 27 carreras, las cuales debido a su expansión no solo están establecidas dentro de la matriz de la universidad, sino también en las instalaciones de la sede en Playas [3].

Debido a la gran dimensión de la entidad educativa, cada una de las carreras que oferta se encuentran divididas en varios sectores de acuerdo con la facultad a la que pertenezcan. En este caso, se escogerá a la Facultad de Sistemas y Telecomunicaciones, donde se realizó la técnica de observación (**Ver Anexo 1**), la cual destaca la presencia de varias conexiones inalámbricas y alámbricas, tanto abiertas como privadas coexistiendo en el mismo sector; algunas redes pertenecientes a la universidad y otras generadas por un móvil o aparato externo,

con el fin de otorgar servicio de Internet, ya sea para el personal administrativo, docentes o comunidad estudiantil.

En la institución educativa, las conexiones de red muestran una lentitud debido a intermitencias o saturación de sus canales, lo que complica el análisis si no se entienden los problemas. De igual manera, estas redes utilizan sistemas de protección que limitan el acceso a diferentes sitios o aplicaciones, restringiendo así el flujo de datos destinados a usos educativos e investigativos.

Se realizó una entrevista dirigida al director del departamento de TI de la institución (**Ver Anexo 2**), revelando que, se enfrenta diariamente con una gran cantidad de intentos de acceso a páginas maliciosas debido al volumen de alumnos conectados a la red universitaria. Además, el departamento cuenta con métodos de seguridad para mitigar dichas conexiones, aunque no son eficaces completamente, debido a la constante aparición de amenazas nuevas que no se registran en la base de datos.

Por otro lado, el sistema de protección incluye antivirus y la base de datos de páginas maliciosas, pero el proceso de actualización es manual y depende de un encargado que no siempre lo completa de forma adecuada; en casos de infección, el tiempo de respuesta es inmediato, cortando las comunicaciones de la red afectada y empleando un servicio alternativo para mantener una conexión estable. Finalmente, agrega que, aunque cuentan con un sistema de rastreo de páginas maliciosas, presenta deficiencias debido a que es antiguo y tiene una limitada efectividad de los sistemas desarrollados por educandos, que, aunque son útiles, no ofrecen el adecuado rendimiento a largo plazo.

En la Universidad de Cantabria en España, se desarrolló el proyecto llamado “Uso de Machine-Learning en el control de congestión sobre redes 5G”, planteando que el número de usuarios conectados a redes móviles aumenta de forma rápida, donde las redes 5G surgen como una solución a este problema de demanda creciente de tráfico; el objetivo general es estudiar distintas técnicas de aprendizaje automático para predecir la congestión, utilizando la metodología aglomerativa; se pudo concluir que, se realizó el estudio de datos por parte de nodos finales de la red para predecir la congestión con técnicas de Machine Learning. No obstante, los datos

son obtenidos a través de un software de simulación y solo trabaja con redes de quinta generación [4].

Se realizó un estudio en la Universidad Distrital Francisco José de Caldas en Bogotá titulado "Algoritmo de volcado del tráfico de datos para redes inalámbricas sobre una red definida por software", donde se prevé que redes definidas por software sean las responsables de proveer una implementación de servicios en red, siendo la solución al crecimiento exponencial en el número de dispositivos conectados a una red; el objetivo principal es evaluar la aplicabilidad del concepto de volcado de tráfico de datos para las redes inalámbricas utilizando el paradigma de redes definidas por software a través del emulador Minitet y controlador ONOS; se concluyó que, el algoritmo de volcado del tráfico de datos mejora las capacidades de red en términos de QoS, teniendo mayores velocidades y soportando más cantidad de hosts sin que la red se pueda saturar [5].

En Ecuador, se realizó el proyecto de titulación con el tema "Estudio de cobertura de redes inalámbricas con Frecuencias 2.4 y 5.0 GHz en las carreras de Ingeniería de Sistemas Computacionales y Tecnología de la Información de la Universidad Estatal del Sur de Manabí", donde se evalúa la teoría de las frecuencias inalámbricas que sean adaptables para las carreras de la institución; cuyo objetivo es realizar un estudio de cobertura de redes inalámbricas con 2.4 y 5.0 GHz como frecuencia, empleando la metodología basada en tres factores: documentación, métodos y técnicas utilizadas para obtener la información; se pudo concluir que, se diseñó una arquitectura Wi-Fi que brinda mejor cobertura; sin embargo, no está analizando ni tratando la información que se encuentra en el tráfico de la red [6].

Localmente, en la Universidad Estatal Península de Santa Elena, se elaboró el trabajo de titulación "Modelo predictivo del tráfico de Internet: Caso puntos digitales gratuitos Zona 5", centrándose en el diseño y evaluación de modelos de redes neuronales para proponer el modelo predictivo empleando técnicas de inteligencia artificial; se usan arquitecturas de redes neuronales con entornos cambiantes y dinámicos, explorando distintas combinaciones de funciones de activación, con el fin de determinar la configuración adecuada que maximice los resultados; se concluyó que, el modelo predictivo cuenta con una arquitectura de

activaciones exponencial, siendo una herramienta efectiva para gestionar y anticipar el tráfico de Internet, ayudando a planificar de mejor manera el ancho de banda en cada uno de los diferentes puntos digitales [7].

Posterior a la descripción y análisis de proyectos similares, así como la explicación de las problemáticas presentadas anteriormente, se plantea desarrollar algoritmos de inteligencia artificial, cumpliendo los procesos de rastreo, obtención y análisis de la información del tráfico web de redes de un Data set proporcionado por el área de TICS de la institución educativa, aplicando diversos algoritmos de inteligencia artificial.

1.2. Descripción del Proyecto

En el presente trabajo se pretende desarrollar algoritmos de inteligencia artificial para el cumplimiento de análisis de la información del tráfico web de las redes de un Data set proporcionado por la entidad educativa. Luego se procederá con la preparación del equipo que servirá de medio para ejecutar los códigos. Dicho aparato, capturará los datos que se encuentren en el Dataset, los cuales serán analizados y clasificados para una posterior presentación a través de gráficas estadísticas.

Para una mejor comprensión de la elaboración del proyecto, se manejarán las siguientes fases adaptadas de la metodología ISSAF, bajo el estándar IEEE 802.11:

- **Fase 1. Planificación**
    - Indagación y recopilación de la información acerca de las herramientas o ambientes que se van a utilizar.
    - Levantamiento de información del sector.
- **Fase 2. Desarrollo**
    - Empezar la codificación del algoritmo.
    - Limpieza de la data set
    - Validación de los datos de la data set.
    - Aprendizaje al algoritmo.
- **Fase 3. Monitoreo y Evaluación**
    - Monitoreo de los datos del Data set.
    - Evaluación de rendimiento y seguridad

    - Evaluación de los algoritmos de Machine Learning
- **Fase 4. Reportes**
    - Visualizar la información a través de gráficas estadísticas por medio de un aplicativo web

Las herramientas y lenguajes de programación utilizados para la elaboración del proyecto, son las siguientes:

- **Python:** Para el desarrollo de los algoritmos de inteligencia artificial [8].
- **MySQL:** Se utiliza como base de datos almacenando la información extraída del tráfico web [9].

Según la Resolución RCT-FST-SO-09 No. 03-20111, el presente trabajo contribuye a la línea de investigación de Desarrollo de Software con la sub-línea de investigación Desarrollo de algoritmos y visión artificial [10], debido a que se encuentra vinculado con una combinación de medios y técnicas para la obtención y tratamiento de información, con el fin de ayudar en la toma de decisiones dentro de una organización.

1.3. Objetivos del Proyecto

Objetivo General

Desarrollar algoritmos en lenguaje Python, utilizando bibliotecas especializadas para realizar el rastreo de redes alámbricas e inalámbricas, volcado del tráfico web y la recolección de información, para examinar las redes empleadas en FACSISTEL.

Objetivos Específicos

- Establecer las herramientas adecuadas para la detección y análisis de las redes.
- Clasificar el tráfico web, aplicando algoritmos de machine learning para examinar las redes.
- Analizar el Data set a través de un Dashboard, para la presentación de los resultados mediante la generación de reportes.

1.4. Justificación del Proyecto

Actualmente las conexiones inalámbricas sirven para muchos fines, uno de estos es dar accesibilidad a datos corporativos desde lugares remotos; al emplear las redes inalámbricas se obtienen beneficios como la eficiencia, movilidad, adaptabilidad, reducción de los costos y la creación de nuevos servicios [11]. En base a esto, y junto al aumento de dispositivos electrónicos que la utilizan conlleva a que se eleve el tráfico de datos, convirtiendo a la tecnología Wi-Fi en un punto clave [11].

Por otro lado, las redes alámbricas ofrecen una mayor confiabilidad y velocidad, debido que son más estables porque poseen una conexión directa al enrutador, lo cual significa que hay menor posibilidad de que se interrumpan por cortes de energía u otros inconvenientes que pueden ocurrir con las conexiones inalámbricas; por ende, las redes cableadas tienen más estabilidad y son más seguras en áreas con mucha interferencia de otros dispositivos [12].

El tráfico en la red es inmenso, por lo que se determinan controles como el monitoreo y el respectivo estudio, debido a la importancia que tiene para la gestión de las redes de telecomunicaciones [13]. Existen hoy en día herramientas comerciales y de software libre que permiten ejecutar estos procesos de análisis, ayudando de tal forma a la detección y notificación de problemas para luego buscar soluciones al instante [13].

Después de estudiar las problemáticas existentes, se plantea la creación de algoritmos que permitan realizar el análisis de las redes que hay en la Facultad de Sistemas y Telecomunicaciones de la Institución de Educación Superior. Para esto, se realizará un mapeo de las redes dentro de las instalaciones universitarias por medio de un Data set, con la finalidad de tener una visualización de la dimensión del tráfico web que se encontrará en el sector, las cuales posteriormente serán monitorizadas.

El desarrollo de un código propio ofrece la facilidad de moldearlo con base a las funcionalidades que se requieran, esto con la ayuda de información, métodos y herramientas sin ningún coste adicional. Por tal motivo, los algoritmos no tendrán limitantes si se llegara a plantear una expansión para otras funciones; en este caso,

los algoritmos estarán diseñados para analizar los datos de las redes que se encuentren en el archivo plano.

Un analizador de red resulta ser un instrumento muy importante para conocer los problemas y tomar las decisiones respectivas para solucionarlos. Por lo que el monitoreo de las redes a través del Data set será primordial para obtener información del tráfico que existe en la red. Los resultados obtenidos del sondeo serán analizados mediante algoritmos de machine learning.

Por otro lado, con respecto a los reportes, se visualizarán los resultados analizados a través de gráficas estadísticas utilizando lenguaje Python donde se mostrará información que servirá para tomar acciones correctivas ante alguna posible anomalía presentada en el tráfico web de la red.

El presente trabajo tiene como finalidad analizar un Data set que incluye las redes que se encuentran en las áreas de FACSISTEL, donde los datos del tráfico web de la red, permitirán la creación de reportes por medio de gráficas estadísticas con información notable que facilite la toma de decisiones del personal encargado.

Este proyecto está direccionado a los objetivos del Plan de Creación de Oportunidades, de acuerdo con el eje social donde se detalla lo siguiente [14]:

Objetivos del Eje Social

Objetivo 5. Proteger a individuos, garantizando servicios y derechos, donde se erradique la pobreza e integración social.

Política 5.5.- Mejora de la conexión tecnológica y acceso a herramientas digitales en la población.

1.5. Alcance del Proyecto

El proyecto tiene como propósito desarrollar algoritmos que permitan realizar un análisis a las redes alámbricas e inalámbricas por medio de un Data set brindada por el Departamento de TI en una Institución de Educación Superior. Estos algoritmos cumplirán con la limpieza del archivo plano, la validación de la información del archivo para saber si es anómalo o normal con el fin de entrenar de forma correcta los algoritmos de Machine Learning. Los códigos de los algoritmos

analizaran los datos del archivo en formato log y presentaran graficas estadísticas de los resultados obtenidos.

Las fases que se llevarán a cabo en esta propuesta, se detallan a continuación:

La fase de planificación, lleva a cabo toda la indagación y recopilación de información acerca de las herramientas y ambientes que servirán para la realización del proyecto. También se realiza el respectivo levantamiento de información del sector, con la finalidad de tener un conocimiento previo de los datos que se obtendrán en el sondeo; para posteriormente, realizar la respectiva limpieza de los datos necesario del Data set.

La fase de desarrollo, contemplará el desarrollo de los algoritmos en Python de autoría propia con base a la información obtenida en la fase de planificación. Dicha codificación cumplirá con el proceso de limpieza y validación de la información. Con respecto a los algoritmos de Machine Learning, se realiza la codificación para el aprendizaje de los algoritmos bajo parámetros establecidos y para el procesamiento de información.

La fase de monitoreo y evaluación, se centra en la preparación de los equipos para la ejecución de los algoritmos desarrollados y los algoritmos de Machine Learning que llevarán el análisis de las redes del Data set.

En la fase de reportes, los datos analizados se mostrarán mediante gráficas estadísticas utilizando Python y la librería matplotlib, junto con información necesaria para tomar decisiones ante cualquier anomalía encontrada en el tráfico web de las redes.

El estudio se realizará sobre un Data set que incluye las redes alámbricas e inalámbricas que se encuentran en las áreas de FACSISTEL.

1.6. Metodología del Proyecto

1.6.1. Metodología de la investigación

Se utiliza la metodología exploratoria, la cual determina la mejor manera de recolectar información ante algo que no se encuentra identificado con claridad [15], en este caso, se considera este tipo de investigación debido que se hace una

recopilación de una variedad de trabajos relacionados con el análisis de redes a través de algoritmos de inteligencia artificial.

De igual manera, se emplea la metodología de tipo diagnóstica, la cual permitirá identificar escenarios puntuales acerca del estado de un fenómeno [16]. En base a los resultados de la observación y entrevista realizada en el sitio, se tiene presente con mayor profundidad la situación actual de las redes por medio de distintos mecanismos de monitoreo, con el fin de analizar las redes con algoritmos de Machine Learning.

También se aplicará una de las metodologías de la investigación orientada al manejo de machine learning en la obtención de resultados, llamada CRISP-DM, la cual incluye un modelo y guía basado en seis fases, que no siguen un orden necesariamente, de modo que algunas son bidireccionales [17]:

- **Comprensión del negocio:** Se centra en la definición de los objetivos y requisitos del proyecto desde la perspectiva del negocio, identificando la necesidad del análisis de datos de las redes para optimizar su funcionamiento.
- **Comprensión de datos:** Exploración de la información disponible, a través de la identificación de anomalías o patrones en los mismos.
- **Preparación de los datos:** Se preparan los datos para el modelado, en este caso, se realizó la limpieza de datos empleando el programa RStudio, borrando duplicados y corrigiendo posibles errores, así como seleccionando las características más importantes para analizar el tráfico web.
- **Modelado:** Se seleccionan y aplican técnicas adecuadas de modelado, aplicando algoritmos de machine learning para clasificar el tráfico de red.
- **Evaluación:** Se evalúa la calidad del modelo, así como los resultados obtenidos; en este contexto, se evalúan los resultados del análisis de datos.
- **Despliegue:** Implica la implementación del modelo en un entorno real, incluyendo la presentación de los resultados en un Dashboard.

Se tiene una mejor comprensión del tráfico de la red usando herramientas como Fortinet, para la exploración, preparación y limpieza de los datos recogidos generando un modelado apropiado para el trabajo de investigación [17].

1.6.2. Beneficiarios del Proyecto

El beneficiario directo será el Departamento de Tecnologías de la Información y Comunicación de la UPSE, debido a que las medidas que se tomen ante cualquier situación presentada con respecto a las redes de la universidad, permitiendo tener un mejor rendimiento en la navegación y acceso a la red.

Los beneficiarios indirectos del presente proyecto son la comunidad académica que conforma la Facultad de Sistemas y Telecomunicaciones de la Universidad, ya que, la codificación desarrollada permitirá tener un control del tráfico web que se encuentra en las redes alámbricas e inalámbricas y a través del reporte generado del análisis se podrán tomar medidas de prevención y corrección ante los diferentes eventos que se presenten en la red.

1.6.3. Variables

Precisión de los algoritmos: Proporción de predicciones correctas realizadas por cada uno de los algoritmos en el análisis del tráfico web.

1.6.4. Análisis de recolección de datos

Ficha de observación

A través de la técnica de observación (**Ver Anexo 1**) realizada en las instalaciones de la Facultad de Sistemas y Telecomunicaciones de la Universidad, se pudo visualizar el entorno junto a la información recolectada posterior al análisis realizado, con la finalidad de tener en cuenta el rumbo al que se pretende llegar con el proyecto. Debido a que esto permitirá una monitorización adecuada de las redes, el comportamiento de estas y el tratamiento de los datos del tráfico web de la red que serán analizados.

Se determinó la existencia de varias redes en el sector, tanto libres como privadas, algunas pertenecientes a la universidad y otras generadas por otros dispositivos externos; como también se logra identificar el roaming del lugar. Al conectarse en una de las redes libres del sector, se observó el alcance de la señal recibida y aunque por ciertos lapsos de tiempo dicha señal era alta, se presentaba cierta lentitud en la comunicación. Por otro lado, en el área, algunas veces se encontraron interferencias o saturaciones en las redes; esto se pudo visualizar a través de aplicativos móviles

y PC para analizar la red, donde de igual forma se observó la existencia del montaje o solapamiento de varios canales de la red en el sector.

Luego, en la navegación web aparecieron negaciones en el ingreso de ciertas páginas e incluso el uso de algunas aplicaciones móviles, es decir, que las redes del lugar utilizan sistemas de seguridad para bloquear paquetes de datos entrantes y salientes en el tráfico de Internet. Mientras se mantuvo la conexión en una de las redes del sector, se hicieron pruebas para conocer si se podía ingresar de forma remota a otros servidores; en este caso, no se obtuvo el acceso.

Con respecto a las redes alámbricas, se pudo observar problemas relacionados a la velocidad y estabilidad de la conexión, es decir, a pesar de tener una infraestructura de cableado estructurado, en algunas zonas se experimentan caídas del servicio, afectando la continuidad en las pruebas de conectividad y en el acceso a los servidores locales. Así mismo, se detectaron latencias elevadas al transferir volúmenes grandes de datos, lo cual podría asociarse a la congestión en los enlaces o fallas de equipos en red.

Entrevista dirigida al director del departamento de TI

La entrevista con el director del departamento de TI (**Ver Anexo 2**), reveló que, el departamento se enfrenta diariamente a más de mil intentos de acceso a sitios maliciosos, debido al gran número de alumnos conectados tanto de forma alámbrica como inalámbrica. Aunque cuentan con técnicas seguras y sistema de antivirus, no son eficaces de manera completa, de modo que la base de datos trabaja de forma manual, sin tener una actualización constante.

Así mismo, en caso de alguna infección, el tiempo de respuesta es ágil, utilizando un servicio adicional de la red, manteniendo una conexión estable mientras se da solución al inconveniente. Esto es fundamental para reducir interrupciones con respecto al acceso a la red, destacando la necesidad de tener un sistema para protección actualizado en el contexto digital.

1.7. Metodología de desarrollo

Metodología ISSAF

Para la elaboración del proyecto, se establece la metodología ISSAF de Open Information Systems Security Group (OISSG), la cual tiene como objetivo ofrecer integridad, precisión y eficiencia a las evaluaciones de seguridad en diversos dominios usando diferentes criterios de prueba [18].

Bajo la adaptación de los lineamientos de la metodología, se procederá al enfoque de cuatro fases, las cuales se describen a continuación [18]:

Fase 1. Planificación: Se plantea el desarrollo del proyecto, la indagación y recopilación de la información acerca de las herramientas o ambientes que se van a utilizar junto al levantamiento de información del sitio.

Fase 2. Desarrollo: En esta fase, se procede con la codificación para los procesos de limpieza y validación, los códigos para el entrenamiento y procesamiento de los datos del archivo proporcionado.

Fase 3. Monitoreo y Evaluación: Se lleva a cabo el monitoreo y evaluación de los algoritmos de Machine Learning previamente entrenamos en la fase de desarrollo con las redes de la data set otorgada.

Fase 4. Reportes: Los resultados obtenidos después de ser analizados se presentarán mediante un reporte visualizado a través de un Dashboard.

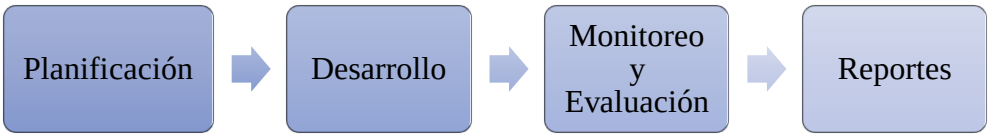


Figura 1: Fases adaptadas de la metodología ISSAF

Metodología OMSTD

Por otro lado, se utiliza la metodología OMSTD (Open Methodology for Security Tool Developers), que presenta una serie de casos de estudio y categorizados como una guía para lograr desarrollar herramientas bien construidas [19]. Es una metodología pensada para trabajar con diferentes lenguajes de programación como

Python, que se establece por bloques y en cuanto al desarrollo existen subtemas, los cuales de detallarán a continuación [19]:

Organización y estructura: En esta parte se organiza el proyecto, tomando la información de las herramientas que servirán para la creación y estructura del código.

Interacción: Aquí se estudia que tan amigable será la interacción del usuario con el entorno que se analizará y con la codificación realizada.

Entrada y salida de información: El algoritmo permitirá recoger la información del tráfico web de las redes, las cuales serán almacenados en una base de datos y luego se generará un informe en HTML.

Redistribución: Se crean algunos fragmentos de código para que puedan ejecutarse en diferentes sistemas y lograr implementarlos en cualquier entorno.

Despliegue: Se establecerá la forma correcta en la que se lleve a cabo la ejecución del algoritmo diseñado dentro del sector que será monitoreado.

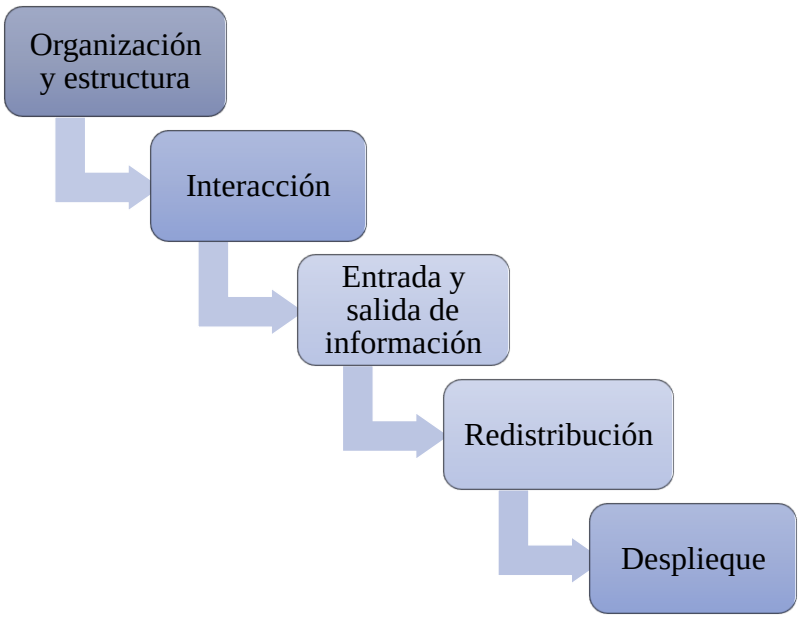


Figura 2: Metodología OMSTD

CAPÍTULO 2. PROPUESTA

2.1. Marco Contextual

2.1.1. Instituciones de Educación Superior

La educación superior pública está compuesta de varios subsistemas; en conjunto, el sistema de educación superior ofrece diferentes opciones de formación acorde a sus intereses y objetivos profesionales [20]. En principio, se conoce como educación superior al nivel donde se forman los futuros profesionales, técnicos y licenciados, quienes estudian sus carreras durante tres y cinco años, dependiendo de la clase de estudios [20].

Es por esto, que cada institución de educación superior, determina: carrera que va a ofrecer, duración de los estudios, malla curricular, tipo de estudios, tiempo que debe durar, estrategias didácticas para su desarrollo, estrategias de evaluación de ellos aprendizajes y requisitos de aprobación; esta información se contiene en el programa de estudios con base a lineamientos nacionales e internacionales, independientemente del tipo y autonomía de la institución de educación superior [21].

2.1.2. Redes en Instituciones de Educación Superior

El propósito principal de las redes en una institución es compartir los recursos de Tecnologías de la Información y Comunicación que posee la entidad educativa, por medio de procesos de trabajo conjunto entre coordinadores académicos, directivos, docentes y alumnos; los beneficios son varios y en la mayoría de los casos, depende de los recursos tecnológicos con los que cuenta la institución [22].

- **Centralizar la información:** Agrupar en un servidor determinados archivos de los trabajos de los alumnos, como: estudios, deberes, exámenes, entre otros; permitiendo que los educadores realicen seguimiento académico y control del cumplimiento de actividades, desde cualquier computadora conectada a la red institucional.
- **Compartir recursos:** Optimizar la cantidad de dispositivos como unidades de almacenamiento o impresoras, que se pueden compartir mediante la red; por ejemplo: no es necesario utilizar memorias USB para la transferencia de

información que se requiere imprimir, pues la misma, se encuentra en la red y se puede imprimir desde cualquier ordenador conectado a ella.

- **Seguridad:** El administrador de la red puede asignar distintos permisos para emplear los recursos compartidos, de acuerdo con las funciones de cada usuario o grupo de los mismos. De tal forma, los profesores tendrán acceso al programa de registro académico, pero los estudiantes no.
- **Acceso remoto:** La información puede ser consultada desde cualquier lugar y en cualquier momento, siendo posible habilitar el servicio, incluso para que los educadores se comuniquen con la red desde su casa.
- **Conectividad entre redes:** En las instituciones educativas grandes se puede dividir la información para que esta resida en diversos servidores, logrando interconectarse para permitir a los usuarios acceder a cualquiera de dichas redes, dependiendo de las funciones que ejerzan.
- **Administración centralizada:** Es posible habilitar opciones del sistema operativo de la red, facilitando la administración.

Por otro lado, se plantea que el WIFI en las entidades educativas rompe las barreras de comunicación, de forma que su portabilidad colabora entre los docentes y estudiantes, inspirando a la participación de los padres de familia al mejorar la interfaz entre la institución y el hogar; a muchos alumnos les resulta más sencillo relacionarse por medio de los dispositivos digitales, por lo que perder esto, es desaprovechar una oportunidad de mejorar la educación y bienestar de los educandos [23].

Así mismo, se destacan las características del WIFI en centros educativos [23]:

- **Capacidad:** A medida que aumenta el número de dispositivos conectados, resulta más complejo que la red inalámbrica los soporte a todos, por lo que es necesario, garantizar que dichos dispositivos conectados a un punto de acceso puedan emplear las aplicaciones sin notar la degradación del rendimiento.
- **Cobertura:** Es imperativo entender dónde se hallan las áreas críticas en el campus, para planificarlas y que garanticen la cobertura idónea.

- **Alta densidad:** Las instituciones educativas tienden a ser capaces de dar mucho soporte a diversas áreas de alta densidad, las cuales contienen miles de usuarios que se conectan de forma simultánea a la red; la alta densidad posee un impacto en las redes debido al volumen grande de dispositivos y a requisitos de capacidad que se derivan de esto.
- **Seguridad:** Cuantos más datos existen, más información requiere de seguridad, y, con el incremento del uso de tecnología en las instituciones, hay muchos datos que necesitan protección.
- **Costo:** La vida útil de las plataformas actuales es de 3 a 4 años, por lo que después de ese tiempo, resulta complicado mantener los niveles de rendimiento necesarios y la fiabilidad a la que los usuarios finales se acostumbran.

2.1.3. Base legal

Ley Orgánica de Telecomunicaciones

Título II

Redes y prestación de servicio de telecomunicaciones

CAPITULO 1. Establecimiento y explotación de redes

Art. 9.- Redes de telecomunicaciones

El gobierno central o los gobiernos autónomos descentralizados tienen la facultad de llevar a cabo las obras requeridas para que las redes e infraestructura de telecomunicaciones sean instaladas de manera ordenada y oculta. En este caso, el Ministerio responsable de las Telecomunicaciones y de la Sociedad de la Información definirá la política y regulación técnica nacional para establecer las tasas o compensaciones a abonar por los proveedores de servicios por la utilización de dicha infraestructura [24]. Para las redes inalámbricas, es necesario acatar las políticas y regulaciones de prevención o precaución, además de las de mimetización y disminución de la polución visual. [24].

Por otro lado, la **Ley Orgánica de Protección de datos personales**, destaca los siguientes artículos:

Art. 10.-Principios

j) Seguridad de datos personales. - Los encargados y responsables del manejo de la información personal deben aplicar todas las medidas de seguridad pertinentes y necesarias, considerándose estas como las aprobadas por el estado de la técnica, ya sean estas organizativas, técnicas o de cualquier otro tipo, con el fin de resguardar los datos personales de cualquier riesgo, amenaza o vulnerabilidad, considerando la naturaleza de los datos personales, el contexto y el contexto [25].

Capítulo VI. Seguridad de datos personales

Art. 37.- Seguridad de datos personales. - Según sea el caso, el encargado o responsable del tratamiento de datos personales deberá adherirse al principio de seguridad de datos personales. Para ello, deberá considerar las categorías y el volumen de datos personales, el estado de la técnica, las mejores prácticas de seguridad integral y los costos de aplicación en función de la naturaleza, el alcance, el contexto y los objetivos del tratamiento, además de reconocer la probabilidad de riesgos [25].

Art. 40.- Análisis de riesgo, amenazas y vulnerabilidades. - Análisis de riesgo, amenazas y vulnerabilidades. - Para evaluar riesgos, amenazas y vulnerabilidades, el encargado y responsable de la gestión de la información personal deberán emplear un método que tome en cuenta, entre otros aspectos [25]:

1) Los detalles específicos del tratamiento;
2) Las especificidades de los participantes implicados; y,
3) Las categorías y la cantidad de información personal que se está tratando.

Art. 41.-Determinación de medidas de seguridad aplicables.- Para establecer las medidas de seguridad autorizadas por el estado de la técnica, a las que está obligado el responsable y el responsable del manejo de la información personal, se deben considerar, entre otros aspectos, los siguientes factores [25]:

1) Los hallazgos del estudio de peligros, amenazas y vulnerabilidades;
2) El carácter de la información personal;
3) Las particularidades de los participantes implicados; y,

4) Los historiales de pérdida, modificación, divulgación o restricción de acceso a los mismos por el propietario, ya sean accidentales o deliberados, por acción u omisión, así como los historiales de transferencia, comunicación o acceso no permitido o en exceso de autorización de dichos datos.

Así mismo, el Código Orgánico Integral Penal (COIP) del Ecuador, establece en [26]:

Sección Tercera: Delitos contra la seguridad de los sistemas de información y comunicación

- **Art. 229.-** Alineación ilícita de la base de datos: El individuo que divulgue datos registrados, guardados en documentos, ficheros o base de datos, a través o dirigidas a un sistema informático, electrónico o de telecomunicaciones; materializando intencionalmente la violación de la intimidad o privacidad de los individuos, será sancionada con pena privativa de la libertad de 1 a 3 años.
- **Art. 230.-** Intercepción ilegal de datos: Se sancionará con pena privativa de la libertad de 3 a 5 años a:
    - La persona que, sin orden previa judicial, intercepte, desvíe, escuche, grabe u observe en cualquier manera una información digital en su procedencia, destino o dentro de un sistema digital.
    - El individuo que elabore, comercialice, desarrolle, implemente o transmita mensajes, certificados o páginas en línea, enlaces o ventanas emergentes o modifique el sistema de solución de nombres de dominio de páginas.
    - La persona que fabrique, produzca, posea o facilite materiales, sistemas informáticos o dispositivos electrónicos que se destinen a la comisión del delito descrito en el párrafo anterior.
- **Art. 232.-** Ataque a la integridad de sistemas informáticos: La persona que dañe, destruya, borre, altere o deteriore datos informáticos, mensajes de correo, sistemas de tratamiento de información o de telecomunicaciones, se sancionará con pena privativa de la libertad de 3 a 5 años.

- **Art. 234.-** Acceso no consentido a un sistema telemático, informático o de telecomunicaciones: La persona sin previa autorización que acceda en parte o a todo un sistema informático o de telecomunicaciones, así como mantenerse en el mismo de quien dio el derecho, para explotar de manera ilegítima, modificar un portal web o desviar tráfico de datos, será sancionada a la pena privativa de la libertad de 3 a 5 años.

2.2. Marco Conceptual

2.2.1. Redes de comunicación de datos

Las redes de comunicación de datos son infraestructuras que se crearon para poder transmitir información por medio del intercambio de datos; es decir, son arquitecturas específicas para este propósito, cuya base principal es la conmutación de los paquetes y que atienden a una clasificación correspondiente, tomando en consideración la distancia que es capaz de cubrir su arquitectura física y el tamaño presentado [27].

Redes alámbricas

Las redes alámbricas son un tipo de red informática que emplea cables para conectar los dispositivos; también, se usan para varios propósitos, incluida la conexión de impresoras, computadoras y servidores dentro de la oficina o una vivienda [28]. Así mismo, el término “alámbrico” se refiere al acceso a Internet de banda ancha por cable, donde una red cableada suele ser más rápida que redes inalámbricas, de modo que no hay interferencia de otros dispositivos en la zona; las redes alámbricas también tienen más confiabilidad porque no dependen de ondas de radios que se pueden interrumpir por obstáculos físicos, como árboles y paredes [28].

Figura 3: Redes alámbricas [27].

Las redes alámbricas son más fiables y rápidas que las redes inalámbricas, además de ser más baratas de configurar; por otra parte, sus características, son [29]:

- Las redes alámbricas usualmente conllevan a tener conexiones Ethernet, que emplean un protocolo de red y cables normalizados similares a los de teléfono fijo.
- Un sistema Ethernet emplea un cable par de cobre trenzado o un sistema de transporte que utiliza cable coaxial; las redes alámbricas de Ethernet más recientes alcanzan velocidades de hasta cinco Gb por segundo.
- El conector de ethernet que se emplea es el de par trenzado sin blindaje, el cual sirve para conectar distintos dispositivos; no obstante, es costoso y voluminoso, haciendo que sea menos práctico para usar en la vivienda.
- Los sistemas de banda ancha brindan Internet por cable, utilizando el tipo de cable coaxial que también emplea la televisión por cable.

Redes inalámbricas

Las redes inalámbricas son una conexión que se da mediante ondas electromagnéticas, o sea, posibilitando la transmisión de información sin la necesidad de un medio físico; en este escenario, el cableado estructurado; por lo tanto, los dispositivos remotos se vinculan con facilidad cuando se hallan dentro la misma red [30]. Otra de sus funciones es que permiten que múltiples terminales se comuniquen sin una conexión cableada; su origen se remonta hace aproximadamente 25 años, en 1997, cuando se comenzó a probar este patrón, fue allí, que desde entonces las conexiones WIFI son la representación más precisa de dicha tecnología [30].

Figura 4: Redes inalámbricas [28].

A continuación, se describen las características o aspectos claves de las redes inalámbricas [31]:

- **Velocidad de envío:** Habilidad para enviar datos a gran velocidad, facilitando una experiencia de usuario sin interrupciones y sin interrupciones.
- **Amplitud y cobertura:** La habilidad de la red para abarcar amplias zonas, desde hogares y oficinas hasta zonas urbanas completas, asegura una conexión constante en diferentes lugares.
- **Protección segura:** Poner en marcha los diversos protocolos de seguridad para proteger los datos y la privacidad de los clientes frente a distintas amenazas o ataques cibernéticos.
- **Adaptabilidad:** Aspecto que permite la conexión de la red en distintos lugares sin requerir cables, brindando libertas para desplazarse y agilidad en la conexión.
- **Estabilidad e interferencia:** Minimiza la interferencia y conserva una señal constante, independientemente de elementos como la congestión del espectro electromagnético o la existencia de barreras físicas.
- **Costo:** La valoración del costo inicial y el costo a largo plazo de instaurar y sostener la red inalámbrica frente a otras alternativas para la conectividad, considerando factores como los equipos, la instalación y los servicios que se asocian.

2.2.2. Tráfico web

El tráfico web es la cantidad de datos totales que se envían y reciben a consecuencia de la actividades de usuarios en un sitio web, constituyendo gran parte del tráfico de todo Internet, aunque no se centra en su totalidad, siendo uno de los puntos más relevantes para la influencia de cualquier página existente; es una de las métricas que se toman en cuenta cuando se realiza analítica de una web, enfocada en cada página en lugar de un cómputo total del dominio, para verificar el alcance y eficacia de cada una [32].

Tráfico web anómalo

El tráfico web anómalo se refiere a todas las páginas que poseen comportamientos o resultados inusuales en comparación con un patrón esperado, siendo su característica mayor, verse con normalidad en términos de estructura, contenido o comportamiento; la detección de esta clase de sitios es fundamental, de modo que permite la identificación de actividades maliciosas, como ciberataques, distribución de malware o pishing [33].

2.2.3. Detección de anomalías

Es necesario contar con tecnología dinámica que sepa diferenciar el comportamiento normal del anormal en base a la forma en que los hosts y servidores interactúan con la red; es donde los métodos estadísticos para la detección de anomalías basada en machine learning son útiles; la detección de las anomalías en la red se basa en clasificar datos diferenciados entre los comportamientos inusuales de dispositivos o aplicaciones [34].

Inteligencia Artificial

El objetivo de la inteligencia artificial (IA) es que los ordenadores realicen las mismas acciones que la mente puede realizar; existen algunas que se etiquetan como "inteligentes" y otras que no, pero todas poseen habilidades psicológicas como la percepción, la predicción, la planificación y el control motor que posibilitan a los humanos y otros animales lograr sus metas [35]. La inteligencia no es un aspecto exclusivo de la existencia sino un espacio estructurado de capacidades diferentes para procesar información, del mismo modo, la IA usa muchas técnicas para resolver una variedad de tareas [35].

Machine Learning

El aprendizaje automático (en inglés, machine learning) es uno de los enfoques principales de la inteligencia artificial [36]. El aprendizaje automático usa algoritmos para aprender de los patrones de datos, como los filtros de spam de correo electrónico que utilizan este tipo de aprendizaje con el fin de detectar qué mensajes son correo basura y separarlos de aquellos que no lo son [36]. Éste es un

ejemplo de cómo los algoritmos pueden usarse para aprender patrones y utilizar el conocimiento adquirido para tomar decisiones [37].

Machine Learning se basa en diferentes algoritmos para resolver problemas de datos; a los científicos de datos, les gusta señalar que no existe un único tipo de algoritmo que sirva para todo y que sea mejor para resolver un problema [38]. La clase de algoritmo empleado depende del problema que se desea resolver, el número de variables, el tipo de modelo que mejor convenga, entre otros aspectos; en la siguiente figura, se observan algunos de los algoritmos de machine learning [38].

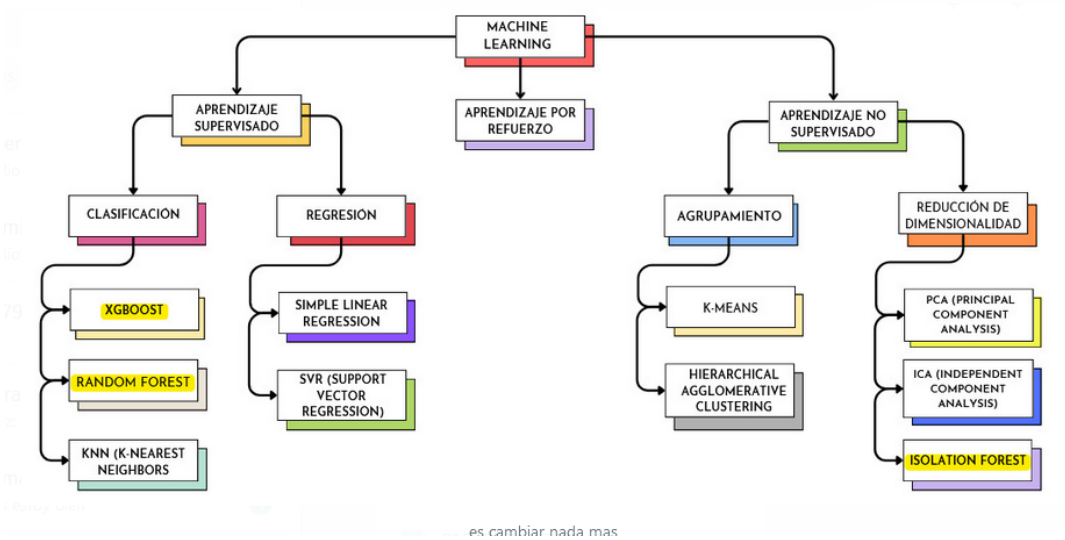


Figura 5: Algoritmos de Machine Learning.

Algoritmos no supervisados

Los algoritmos no supervisados son aquellos que tienen la capacidad de extraer, clasificar y disminuir las dimensiones, siendo de gran utilidad para mejorar los resultados de las técnicas de agrupamiento y que no necesitan una etiqueta de clase [39]. Estos algoritmos pueden describir patrones que estén ocultos como también agrupar los datos más útiles para ordenarlos en categorías [39].

Existen los tipos de algoritmos de análisis de conglomeraciones, los cuales permiten realizar la agrupación de unidades muestrales teniendo en cuenta las similitudes que se encuentren entre ellos en un grupo de variables de forma numérica [40]. Por otra parte, el de análisis de componentes principales es un método que tiene como fin reducir las dimensiones de los datos, no agrupan muestras individuales, sino

variables [40]. No es excluyente y se pueden complementar con otros algoritmos, colaborando en conjunto para elevar la exactitud y efectividad de los modelos de machine learning para mejorar el análisis y la comprensión de datos en diferentes contextos [40].

Algoritmos supervisados

Los algoritmos supervisados o aprendizaje supervisado, se conoce también como machine learning supervisado, siendo una subcategoría de la inteligencia artificial o machine learning, definiéndose por el uso de conjuntos de datos etiquetados para el entrenamiento de algoritmos clasificados por datos y que predicen los resultados de forma precisa [41]. A medida que se introducen los datos en el modelo, se ajusta a sus ponderaciones hasta que se establezca el modelo de manera correcta, lo cual sucede en el marco de un proceso de validación cruzada [41].

En el aprendizaje supervisado, el modelo entrena con un conjunto de datos de entrada y un conjunto que corresponde a los datos de salida que se etiquetan en pares; por lo general, el etiquetado es realizado manualmente; es por esto, que el aprendizaje supervisado asiste a las organizaciones en la solución de varios problemas reales a gran escala, tales como la clasificación de spam y crear modelos de machine learning con precisión alta [42].

Árboles de decisión

Cada familia de algoritmos de aprendizaje aplica diferentes mecanismos para la construcción de modelos analíticos como, por ejemplo, al construir un modelo de clasificación, algoritmos de árbol de decisión explota el espacio de características dividiendo incrementalmente los registros de datos en particiones cada vez más homogéneas, siguiendo una estructura jerárquica en forma de árbol [43].

Los árboles de decisión (Decision Trees) son uno de los métodos poderosos comúnmente utilizados en diversos campos, como el aprendizaje automático, el procesamiento de imágenes y la identificación de patrones [44]. DT es un modelo sucesivo que une una serie de pruebas básicas de manera eficiente y cohesiva donde una característica numérica, donde se compara con un valor umbral en cada prueba y las reglas conceptuales son mucho más fáciles de construir que los pesos

numéricos en la red neuronal de conexiones entre nodos, utilizada principalmente con fines de agrupación [44].

Random Forest

Es un algoritmo de machine learning de uso común, el cual combina la salida de diversos árboles de decisión para alcanzar un solo resultado; su facilidad de flexibilidad y uso han impulsado la adopción, manejando problemas de regresión y clasificación [45]. El algoritmo de bosque aleatorio es una ampliación del método de ensacado, ya que utiliza tanto el ensacado como la aleatoriedad de las características para generar un bosque de árboles de decisión sin correlación [45].

Los aspectos de este algoritmo, son [45]:

- Aplica varios árboles de decisión para llevar a cabo una proyección, incrementando la exactitud y solidez del modelo.
- Cada árbol se edifica mediante una elección aleatoria de las características, garantizando que los árboles se relacionen entre ellos y minimizando el sobreajuste.
- Es eficiente tanto en problemas para la clasificación, como para la regresión.
- Puede manejar grandes cantidades de datos sin perder su eficacia.

Se escogió este algoritmo por su capacidad para manejar grandes cantidades de datos, sin requerir de gran capacidad, lo que es idóneo para los entornos con procesadores con una capacidad intermedia.

XGBoost

Es un método de aprendizaje automático supervisado para clasificación y regresión, el cual se basa en árboles de decisión y supone una mejora sobre los demás métodos, como el bosque aleatorio y el refuerzo de los gradientes, funcionando bien con datasets complejos y grandes al emplear varias técnicas de optimización [46].

Las características de este algoritmo, son [46]:

- XGBoost incluye técnicas de optimización que optimizan la precisión y minimizan el sobreajuste.

- Se diseña para trabajar con volúmenes grandes de información, contando con alta eficacia en tiempo de entrenamiento y predicción.
- Es capaz de manejar de mejor forma, las relaciones no lineales y complejas entre las diferentes variables.

Se eligió el algoritmo debido a su capacidad para manejar volúmenes grandes de información de forma eficiente, además de su alta precisión. Además, brinda un rendimiento superior en problemas de regresión y clasificación.

Isolation Forest

Se trata de un procedimiento no supervisado para detectar irregularidades cuando la información no está etiquetada, es decir, no se sabe la verdadera clasificación de las observaciones; su operación se basa en el algoritmo de clasificación y regresión de Random Forest, donde se forma por diversos árboles que se denominan Isolation trees [47].

Las características de este algoritmo, son [47]:

- Se diseña de manera específica para la detección de observaciones atípicas en conjuntos grandes de datos.
- Se construyen los árboles rápidamente.
- Es fácil de comprender e implementar.

Se escogió este algoritmo por su rapidez en la detección de las anomalías en grandes cantidades de datos. Así mismo, es idóneo en entornos computacionales que poseen limitaciones de procesamiento.

TOPS

Significa Tera Operations Per Second, siendo una unidad de medida con la que se puede cuantificar el rendimiento del sistema computacional cuando se dirige a la IA, por ende, lo que mide esta unidad es la capacidad para realizar millones de operaciones de coma flotante por cada segundo [48]. Es empleada para simplificar y especificar el rendimiento de la Unidad de Procesamiento Neural de una PC [48].

GPU

Es la unidad de procesamiento de gráficos, conteniendo diversos núcleos especializados y pequeños, los cuales ofrecen un gran desempeño al trabajar juntos, dividiendo las tareas de procesamiento entre múltiples núcleos de manera simultánea; la GPU se destaca en tareas paralelas, como la renderización de imágenes durante un juego, calcular resultados en cargas grandes de trabajo de IA o en la manipulación de datos de video al crear contenido [49].

2.2.4. Sitio web

Es una estructura que se conforma de información generada en un ámbito o espacio de comunicación nuevo, el cual se crea mediante tecnologías de la información y posee dos elementos esenciales, planteando un conjunto de prestaciones que los usuarios visitantes de la web, pueden ejercitar para satisfacer sus necesidades [50].

Dashboard

Dashboard o "Tablero digital" que "es una interfaz gráfica de usuario en dónde se pueden administrar recursos informáticos y analizar información para la toma de decisiones, genéricamente, un dashboard engloba a varias herramientas que muestran información relevante para la empresa a través de una serie de indicadores de rendimiento, también denominados KPI, que son métricas utilizadas para cuantificar los resultados de una determinada acción o estrategia, en función de unos objetivos predeterminados [51].

El dashboard es una pantalla que puede mostrar un gráfico como KPI (indicador clave de rendimiento) de una organización o empresa, siendo una medida importante necesaria para tomar una decisión [52]. Existen algunas características de un panel si se diseña de forma correcta, las cuales se presentan a continuación [52]:

a. El dashboard muestra una visualización dinámica y tangible de datos que se actualizan periódicamente.
b. Permiten a los usuarios mantenerse actualizados sobre cualquier cambio en el negocio.

c. Requieren ligeros cambios en el código del programa para ser enviados, implementados y mantenidos.
d. Utilizan componentes visuales para resumir de un vistazo los datos y las excepciones que requieren acción.
e. Son transparentes para los usuarios, lo que significa que los mismos necesitan capacitación, para que sea fácil de usar el panel.
f. El dashboard combina la información de varias fuentes de datos en una presentación empresarial que es concisa y se combina en una sola. Por ejemplo, los paneles permiten la búsqueda fuentes de datos o informes existentes, brindando un contexto que se puede comparar y evaluar con más detalle.

Los KPI son una herramienta fundamental para el aumento de la competitividad, comprendiéndola como una expresión cuantificable del comportamiento o desempeño de toda una empresa o solo una de sus partes, cuya magnitud al ser comparada con algún nivel de referencia, podría señalar desviaciones sobre la cual se tomarán acciones correctivas o preventivas según sea el caso; lo que le da valor a los KPI, es el resultado de la medición del mismo, sustituyendo un valor de comparación en referencia a una meta u objetivo [53].

2.2.5. Herramientas y lenguajes de programación utilizados en la propuesta

A continuación, se presentan las herramientas que se utilizan en la presente propuesta:

Python

Es un lenguaje de programación utilizado en aplicaciones web, ciencia de datos, desarrollo de software y machine learning, el cual es eficiente y fácil de entender, además que se puede ejecutar en múltiples plataformas [54]. Para el desarrollo de los algoritmos de inteligencia artificial se necesitará de Python, un lenguaje de programación potente y de fácil aprendizaje [54]. Tanto el intérprete de Python como la extensa librería estándar están disponibles libremente en código fuente y binario para la mayor parte de plataformas desde la web del lenguaje [54].

Request

Es un estándar que sirve para realizar solicitudes HTTP cuando se desarrolla del lado del servidor en un sitio web, facilitando el trabajo de los desarrolladores y simplificando todo en una API simple [55]. Dentro de las librerías de Python tenemos a Request que se utilizará para realizar peticiones de http, permitiendo interactuar con la web sin problemas y sin tener la necesidad de agregar de forma manual cadenas de consulta a las URL o codificar datos POST [55].

Matplotlib

Es una librería de Python que permite crear gráficos en dos dimensiones, personalizando: diagramas de barras, histogramas, diagramas de sectores, de dispersión de puntos, de líneas, de áreas, de violín, de contorno, mapas de color; así como la combinación entre todos ellos [56].

2.2.6. Base de datos

Es un conjunto estructurado y ordenado que representa una realidad organizada de forma independiente acerca de las aplicaciones, lo que significa que pueden ser compartidas y utilizadas por usuarios y aplicaciones distintas [57]. Es decir, que una base de datos puede considerarse una colección de datos variables en el tiempo [57].

MySQL

Como base de datos relacional se utiliza MySQL, ya que, ofrece escalabilidad y flexibilidad, cuya finalidad es almacenar y mantener sincronizados los datos entre aplicaciones y clientes a través de objetos, ofreciendo también soporte sin conexión para los aparatos móviles y la web [9].

2.2.7. Metodologías

OMSTD

Es una metodología colaborativa y abierta, que sirve de apoyo para el desarrollo de herramientas nuevas; esta guía se dirige a los auditores de seguridad que requieren desarrollar sus propias herramientas y desean que éstas sean algo más que un script simple [58]. Los casos de estudio se dividen en los bloques siguientes [58]:

- **Desarrollo**
    - Organización y estructura (ST): Cómo son organizados los proyectos.
    - Comportamiento (BH): Es la manera de interactuar con distintos frameworks.
    - Interacción (IT): Interacción del usuario con otros sistemas o entornos.
    - Específicos de lenguaje (LS): Trucos, usos concretos y buenas prácticas de desarrollo. En este caso, Python.
    - Entrada/Salida de información (IO): Generación de informes, exportar datos e importar la información externa de XML/JSON, entre otros.
    - Redistribución (RD): Creación de paquetes, compilarlos y redistribuirlos para diversos sistemas, portarlos a distintos entornos y el uso correcto de los sistemas sobre control de versiones, entre otros.
    - Despliegue (DP): Cómo se pone en producción de forma correcta la aplicación.
- **Hacking (HH):** Intervienen casos correspondientes a hacking.
- **Cracking (CH):** Intervienen casos correspondientes a cracking.
- **Malware (MH):** Intervienen casos correspondientes de malware.
- **Forensic (FH):** Intervienen casos correspondientes de Forensic.
- **Hardening (DH):** Intervienen casos correspondientes de hardening.

ISSAF

ISSAF desarrolla una metodología detallada y extensa sobre cómo realizar pruebas de penetración; esta metodología del marco de evaluación de seguridad de los sistemas de información, cuenta con el apoyo de un grupo de seguridad de los sistemas de información abiertos (OISSG) [59]. ISSAF divide el proceso de pentesting en tres etapas [59]:

- Planificación y preparación.
- Evaluación.

- Informar, destruir y limpiar los artefactos.

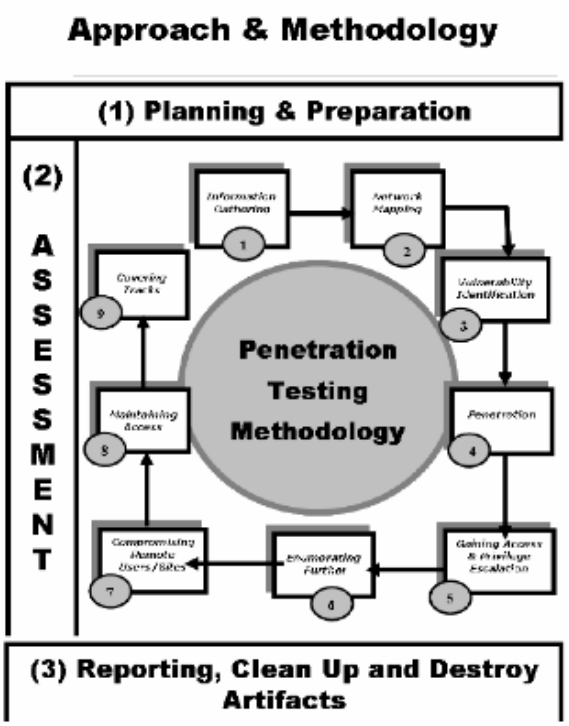


Figura 6: Metodología ISSAF [59].

2.3. Marco Teórico

2.3.1. Inteligencia Artificial de la mano con la ciencia de datos

La Inteligencia Artificial y la Ciencia de Datos son áreas interdisciplinarias que han tomado gran auge en la última década, debido a la gran cantidad de datos que la sociedad produce para el desarrollo de todas sus actividades [60]. Dichos datos se analizan por medio de métodos científicos, sistemas y procesos utilizados para extraer el conocimiento y el entendimiento optimizado de sus distintas maneras, que se valen de diferentes técnicas originarias de estadística, la minería de datos, el aprendizaje automático, entre otras; para el posterior apoyo a la toma de decisiones de negocio, gobierno y otros rubros donde la sociedad desarrolla sus actividades [60].

La IA se interpreta como la disciplina de las ciencias informáticas, que propone modelos computacionales basados en redes neuronales biológicas humanas [61]. En este contexto, se plantean distintos modelos de inteligencia artificial, que, gracias a los avances en tecnología digital, ha permitido desarrollar sistemas

inteligentes que facilitan el procesamiento. Además, comprende varios campos como el reconocimiento de voz, el procesamiento del lenguaje natural, la visión computacional, la robótica avanzada, la captura de conocimiento, la planificación y optimización, entre otros, con el objetivo de que el sistema posea la habilidad para percibir, razonar, interactuar y aprender. [61].

La minería de datos web es una tecnología usada para descubrir conocimiento interesante en todos los aspectos relacionados con la web; el enorme volumen de datos en la misma, generado por la explosión de usuarios y el desarrollo de librerías digitales, hace que la extracción de la información útil sea un gran problema [62]. Cuando los usuarios navegan por la web, se halla de forma frecuente saturado por la información, por tal motivo, integrar herramientas de minería de datos puede contribuir con la extracción de datos relevantes [62].

2.3.2. Clasificación de tráfico web mediante técnicas de Machine Learning

Para llevar a cabo la categorización del tráfico, se pueden emplear sistemas de aprendizaje supervisado, sistemas de aprendizaje no supervisado y sistemas híbridos de aprendizaje. Si se busca categorizar el tráfico mediante un sistema de aprendizaje no supervisado, es necesario disponer de una información correctamente etiquetada, ya que, si no se tiene de esta manera no será posible llevar a cabo la caracterización que permitirá la separación [63]. Dicho problema no se presenta en los sistemas de aprendizaje no supervisado, debido a que son sistemas que utilizan técnicas de representation learning, donde se pueden hallar características a partir de datos brutos, para que luego puedan ser clasificados [63].

La clasificación, una subcategoría del aprendizaje supervisado, tiene como meta anticipar las etiquetas de clase categoría de nuevas instancias, basadas en observaciones pasadas [64]. Estas etiquetas de clase son discretas, valores desordenados que se pueden entender como membresías grupales de las instancias; un ejemplo, es la detección de correo no deseado, la cual hace una tarea de clasificación binaria, donde el algoritmo de aprendizaje automático aprende un conjunto de reglas para distinguir entre dos posibles clases: mensajes que son o no son correo no deseado [64].

Uno de los algoritmos para clasificación es el árbol de decisiones, que son estructuras en forma de árbol, cuyos nodos representan una elección entre varias alternativas y cada nodo hoja representa una decisión, permitiendo explorar los posibles resultados para varias opciones, evaluando el riesgo y las recompensas para cada posible curso de acción [65]. Estas decisiones generan reglas, que luego se usan para clasificar los datos; entre los algoritmos de aprendizaje de árboles de decisión, destacan ID3, C4.5 y ASSISTANT [65].

2.3.3. Algoritmos para el aprendizaje de patrones

Los algoritmos de reconocimiento de patrones son otro ejemplo de aplicación del aprendizaje automático para solucionar problemas de visión, reconocimiento de discurso y robótica; los algoritmos de aprendizaje automático se pueden clasificar en supervisados y no supervisados; el aprendizaje supervisado corresponde a la situación en que se tiene una variable de salida, ya sea cuantitativa o cualitativa, que se desea predecir basándose en un conjunto de características [66]. Por otro lado, el aprendizaje no supervisado es una situación en la que existe un conjunto de información que contiene diferentes características de individuos, sin que ninguna de ellas se considere como variable de salida que se desee predecir [66].

El aprendizaje informático se emplea para hallar anomalías ocultas en gran cantidad de información [67]. Uno de los algoritmos utilizados es el algoritmo de Bayes ingenuo, que emplea la regla de Bayes para prever la pertenencia de una nueva muestra de cierto tipo a partir de un conjunto de datos de entrenamiento, asumiendo que hay un modelo de probabilidad inherente a los datos; este modelo también presupone que hay autonomía entre los atributos [68].

Por otra parte, los algoritmos de bosques aleatorios son una variante de los árboles de decisión, conocidos también como ensambles de árboles, generando una cantidad específica de árboles de decisión independientes, cuyos resultados se promedian; esta característica contribuye en la generalización del modelo; así mismo, las máquinas de soporte vectorial son el modelo que separa los datos de entrada y los transforma hacia un espacio de características de dimensión alta, construyendo un hiperplano de separación por motivo de la distancia entre vectores formados de soporte [67].

2.4. Requerimientos

Para la ejecución de los algoritmos desarrollados, se precisa de los requerimientos que se describen a continuación:

R01. Los algoritmos podrán ejecutarse en equipos electrónicos que cuenten con un sistema con los siguientes requisitos:

Procesador	Intel Core I3 de cuarta generación
Memoria RAM	4 GB de RAM
Disco Duro	128 GB
Tarjeta de video	Integrada al procesador
Tarjeta de red	802.11 b/g/n 2.4GHZ

Tabla 1: Requerimientos mínimos del equipo

Procesador	Intel Core 7 de octava generación o Ryzen 7 3700H
Memoria RAM	16 GB de memoria RAM
Disco Duro	500 GB
Tarjeta de video	Envidia GeForce GTX 1650
Tarjeta de red	802.11 b/g/n/ac 2.4 GHZ - 5 GHZ

Tabla 2: Requerimientos recomendados del equipo

R02. Para la ejecución de las pruebas y análisis se utiliza Jetson Orin, Co procesador Coral y Raspberry pi5, con las siguientes características:

Procesador	NVIDIA Orin
Memoria RAM	8 GB LPDDR5
Rendimiento	4 TOPS
Almacenamiento	64 GB eMMC
GPU	NVDIDIA Ampere con 256 núcleos
Conectividad	USB 3.2, HDM 2.1, PCIe Gen 4
Consumo de energía	30 W

Tabla 3: Características de Jepson Orin

TPU	Google Coral Edge TPU
Rendimiento	Hasta 4 TOPS en inferencias
Dimensiones	88 mm x 56 mm
Conectividad	Conexión USB 2.0. y GPIO
Interfaz de comunicación	USB Y GPIO
Eficiencia energética	Consume menos de 2W

Tabla 4: Características de Coprocesador Coral para Raspberry Pi 5

R03. El algoritmo creado para la limpieza tiene que iniciar con un archivo en formato .log, por ende, el usuario deberá proporcionar el mismo formato para los datos que se desean analizar.

R04. El algoritmo deberá cumplir con el proceso de eliminación de columnas que son irrelevantes para el análisis del tráfico web, eliminación de datos vacíos o inconsistentes, la conversión del archivo .log al formato CSV.

R05. El usuario tendrá que crear una cuenta en "VirusTotal" para la obtención de la API, la cual se conectará con el algoritmo que está dedicado para validar si los datos del CSV limpio son realmente normales o anómalos.

R06. El CSV generado deberá analizarse a través de las APIs creadas de VirusTotal con el fin de confirmar que la información proporcionada por el archivo .log es realmente verídica, el código deberá generar un nuevo CSV con una nueva columna de resultado donde estarán clasificados en normales y anómalos los datos del archivo.

R07. El usuario tendrá que tener instalado phpmyadmin para utilizar el gestor de base de datos MySql, con el fin de visualizar la información que se obtenga mediante el escaneo y análisis por medio de las Apis del Data set.

R08. Los algoritmos solo se pueden ejecutar en sistemas operativos Windows y Linux.

R09. El sistema deberá tener instalado "Python 3" desde la versión 3.0 en adelante, siendo recomendable utilizar la versión 3.11.3.

R10. El sistema deberá tener instalado “Visual Studio Code”, para escribir el código fuente correspondiente al Dashboard.

R11. Para ejecutar los algoritmos, se deben tener instaladas las siguientes librerías: comtypes, request, streamlit, pandas, os, matplotlib y scikit learn

R12. Se debe contar con un navegador web instalado, como: Google Chrome, Firefox, Microsoft Edge, entre otros; los mismos que servirán para visualizar las gráficas.

R13. El usuario deberá comprimir el archivo .log a un formato zip para poder subir en la página web debido al tamaño del archivo log.

R14. Al finalizar la conversión del archivo zip a log, se visualizará de forma previa los datos que leyó y depuró el código.

R15. El usuario deberá escoger uno de los tres algoritmos de clasificación de machine learning para continuar, las tres opciones a escoger son XGBoost, Isolation Forest y Random Forest.

R16. El algoritmo tiene un tiempo variable para analizar todos los datos del archivo, dependiendo del equipo donde se ejecute.

R17. En caso de existir algún inconveniente o inconsistencia con respecto al programa, el usuario deberá interrumpir el análisis que se esté realizando.

R18. En el Dashboard deberá visualizar gráficas estadísticas de los resultados del archivo y tendrá de forma opcional los detalles del algoritmo que escogió permitiendo ver el rendimiento del mismo.

2.5. Arquitectura del sistema

La siguiente figura muestra la arquitectura de cómo se implementa el sistema en la organización, interceptando paquetes en la red de la misma manera. A continuación, se mencionan las partes principales del algoritmo desarrollado.

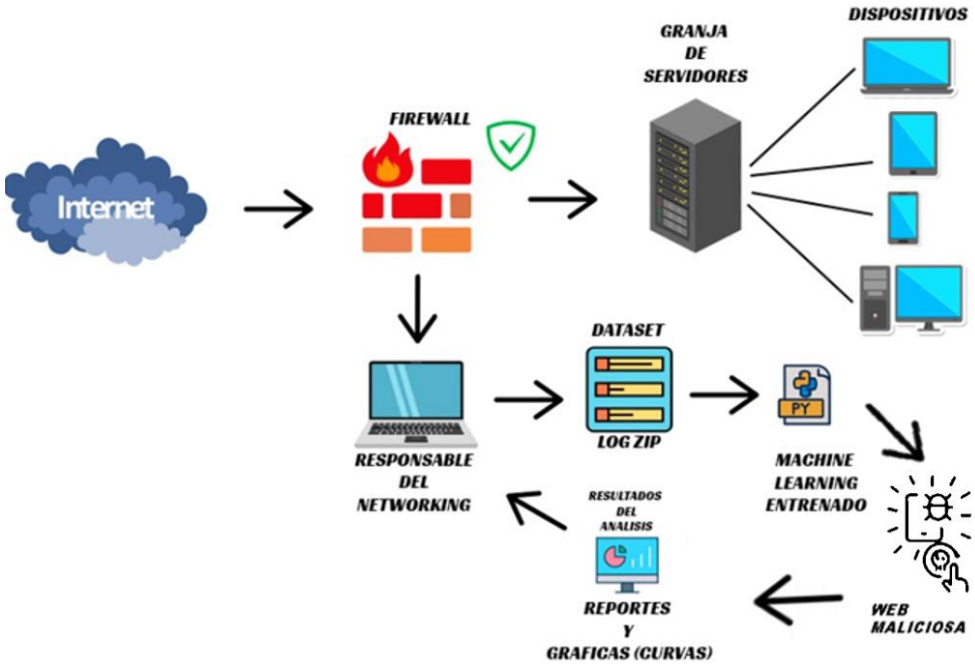


Figura 7: Arquitectura del sistema

La arquitectura del sistema muestra la configuración de un sistema de procesamiento y análisis de datos, creado para la gestión y análisis del conjunto de datos proporcionado por el departamento de TICS de la entidad educativa, donde se administrará un área de la facultad de sistemas y telecomunicaciones. Este sistema simplifica el intercepto y examen de paquetes en la red de la entidad, empleando varias etapas para procesar, almacenar y presentar la información relevante en un panel de control gráfico.

Además, se fundamenta en información proveniente de Internet, administrándose a través de un intermediario que se encarga de la conectividad, que tiene el deber de recabar y gestionar los datos situados en los servidores de la entidad educativa. Esta información se resguarda a través de un cortafuegos antes de llegar a la granja de servidores, donde se guarda de manera organizada para su posterior procesamiento.

En la fase inicial, el sistema obtiene un archivo plano en formato CSV, que alberga información sin procesar proporcionada por la entidad. Este archivo está en estado bruto, lo que significa que necesita ser procesado antes de que los datos puedan ser examinados. Después, se utiliza un script de preprocesamiento, creado en Python, para la limpieza y transformación de la información; este script lleva a cabo tareas

como la validación de campos, eliminación de registros duplicados y reorganización de los datos para que se encuentren en un formato apropiado. Por lo tanto, como consecuencia de este procedimiento se genera el archivo CSV preparado.

Una vez que se realizó este proceso, el archivo se carga en un servidor centralizado y se almacena en la base de datos MySQL, la cual facilita el almacenamiento de los datos, permitiendo realizar consultas de forma eficiente para recuperar la información. A través de la BD, se pueden llevar a cabo consultas que buscan identificar los registros normales o anómalos.

Finalmente, el sistema posee la funcionalidad para generar reportes y gráficos en el Dashboard, donde se presenta de manera visual el estado de los datos. Esta interfaz gráfica permite al equipo, monitorear de forma intuitiva los resultados del análisis de los archivos.

2.6. Desarrollo de la propuesta

Fase 1. Planificación

En esta fase, el primer paso es preparar las herramientas y ambiente necesarios para la ejecución del proyecto. Se inicia con la recopilación de información sobre las mismas, como Python, librerías de manejo de CSV y demás APIs de análisis de seguridad como VirusTotal. Luego, se comprende el contexto y los tipos de datos que se analizarán, permitiendo definir cómo abordar los archivos CSV y las IPs que se van a procesar.

El lenguaje de programación **Python** se emplea para entrenar los algoritmos de machine learning, facilitando el análisis y procesamiento de datos. La Librería **Request** permite realizar solicitudes HTTP de forma sencilla, lo que asegura una interacción fluida con los servicios web requeridos. Por otra parte, **Matplotlib** se usa para generar los gráficos en dos dimensiones, siendo útil en la visualización de los resultados obtenidos. Como base de datos, se utiliza **MySQL**, para el almacenamiento y gestión de información de manera estructurada, lo que asegura su accesibilidad y organización para la generación de reportes.

Con relación a los algoritmos, se realiza una tabla comparativa para seleccionar los más idóneos con respecto al presente proyecto. A continuación, se describen diversos algoritmos de machine learning:

Algoritmo	**Ventajas**	**Desventajas**	**Nivel de procesamiento**
Support Vector Machine (SVM)	• Alta precisión en problemas de clasificación. • Eficaz con pequeños datasets.	• Escalabilidad limitada con grandes datasets. • Complicado ajuste de parámetros.	Alto
Naive Bayes	• Eficiente y muy rápido. • Excelente para problemas de clasificación de textos.	• Independencia entre sus características. • Precisión baja en datos relacionados.	Bajo
K Means	• Rápido para clustering. • Sencillo de interpretar.	• No es apto para regresión o clasificación. • Sensible a valores iniciales.	Bajo
Random Forest	• Es resistente a sobreajuste. • Soporta regresión y clasificación.	• Lentitud con muchos árboles. • Menos interpretabilidad.	Intermedio

K Nearest Neighbors	• Fácil de implementar. • No necesita un entrenamiento específico.	• Intensivo computacionalmente en las predicciones. • Sensible al ruido en los datos.	Bajo
XGBoost	• Precisión alta en la clasificación y regresión. • Optimiza velocidad y rendimiento.	• Costoso computacionalmente. • Necesita ajustes en los parámetros.	Intermedio
Isolation Forest	• Detección de anomalías. • Eficaz en datasets grandes.	• Precisión menor en tareas generales. • Sensible al tamaño muestral.	Intermedio

Tabla 5: Comparativa de algoritmos de machine learning

Se evaluaron distintos algoritmos de machine learning, tomando en consideración su nivel de procesamiento, así como sus ventajas y desventajas; por tal motivo, se seleccionaron tres algoritmos: **XGBoost, Random Forest e Isolation Forest**, los cuales ofrecen un balance entre eficiencia, precisión y capacidad para el manejo de múltiples datos. Así mismo, cuentan con un nivel de procesamiento intermedio, siendo ideal para un análisis riguroso sin ser tan costoso o complicado de manera computacional. Estos algoritmos destacan en sus características como precisión y escalabilidad, lo que es adecuado para el Data set actual.

Fase 2. Desarrollo

En cuanto al data set, se deben asegurar que los datos sean confiables y que se encuentren en óptimas condiciones para su procesamiento. La limpieza de datos se

realizó utilizando un Script, el cual facilita la depuración y organización de la información mediante scripts automáticos.

Para llevar a cabo la limpieza de los datos, se emplea un script que automatiza la preparación de la información contenida en el Data set de Excel que brindó la institución, tomando como segmento la facultad de sistemas y telecomunicaciones. A continuación, se detallan los pasos que fueron realizados:

1. **Importar la información:** El script comienza con la importación del Dataset, para luego leerlo y cargar los respectivos datos.
2. **Tratamiento de los datos faltantes:** El script logra identificar los datos faltantes o nulos, para luego manejarlos de manera adecuada, eliminando valores según se necesite.
3. **Formatear la información:** Se revisa que toda la información contenga un formato constante, lo que hace más sencillo el análisis.
4. **Filtrado de datos irrelevantes:** El script elimina cualquier dato que no se necesite en el análisis, dejando únicamente la información que es útil.
5. **Validación de los datos:** Finalmente, el script ejecuta una serie de comprobaciones que validan los datos para que estén listos para el análisis.

2.1. Limpieza del Data set

Para empezar el reconocimiento de las IP de manera anómalas, la institución brinda un archivo plano o Data set, el cual contiene información completa del historial de navegación de la red raíz o principal que proporciona o comparte el internet a la entidad educativa y de la cual se deberá extraer la información necesaria para posteriormente analizarla, tomando como segmento la facultad de sistemas y telecomunicaciones. Cabe mencionar que por cuestiones de privacidad se procede a la difuminación de información de IPs de la entidad de educación superior.

Date,Time,Eventtime,Tz,Logid,Type,Level,Srcip,Srcport,Sr	onid,Proto,Policyid,Poluuid,Policyname,Service,Transip,Transport,Appid,App,Appcat,Apprisk,Applist,
2024-08-16,10:12:12,1723821131881813774,-500,13,traffic,0	WAN_1G_CED,wan,,538999764.0,17,141.0,d7a125c2-eb6e-51ed-4fa5-7b2b1c6203ea,Administrativos_ha
2024-08-16,10:12:12,1723821132511820108,-500,13,traffic,0	N_1G_CED,wan,,539001388.0,17,150.0,35230bd8-9def-51ee-9098-92dc72feec09,TalentoHumano_hacia
2024-08-16,10:12:12,1723821132391802389,-500,13,traffic,0	_WAN_1G_CED,wan,,539110915.0,6,169.0,175e0fe2-f5bc-51ee-8a5b-a225ecc9beeb,Activos_Fijos_to_I
2024-08-16,10:12:11,1723821131541793728,-500,13,traffic,0	AN_1G_CED,wan,,539153528.0,6,11.0,bf2d1d48-bb2e-51ea-472e-d18b767a05be,Estudiantes2_hacia_In
2024-08-16,10:12:11,1723821131340722531,-500,20,traffic,0	_1G_CED,wan,,538394151.0,17,166.0,b5c36412-eb7a-51ee-7ae6-f73749d378c4,RedApuepse_to_Interne
2024-08-16,10:12:11,1723821131271802744,-500,13,traffic,0	G_CED,wan,,539156283.0,6,166.0,b5c36412-eb7a-51ee-7ae6-f73749d378c4,RedApuepse_to_Internet,H
2024-08-16,10:12:11,1723821131191797676,-500,13,traffic,0	1G_CED,wan,,539156282.0,6,166.0,b5c36412-eb7a-51ee-7ae6-f73749d378c4,RedApuepse_to_Internet,I
2024-08-16,10:12:11,1723821130774576746,-500,20,traffic,0	_WAN_1G_CED,wan,,538976484.0,17,4.0,883d9936-bb2c-51ea-a6cb-8c5e57f8d1d2,Oficinas2_hacia_Int
2024-08-16,10:12:11,1723821131221819148,-500,13,traffic,0	_WAN_1G_CED,wan,,539022778.0,17,141.0,d7a125c2-eb6e-51ed-4fa5-7b2b1c6203ea,Administrativos_h
2024-08-16,10:12:09,1723821129291821621,-500,13,traffic,0	AN_1G_CED,wan,,539021567.0,17,161.0,1c2eb9e6-da77-51ee-83e5-8f36d053dabc,EdificioVicerrectora
2024-08-16,10:12:09,1723821129301822746,-500,13,traffic,0	P_WAN_1G_CED,wan,,539021396.0,17,4.0,883d9936-bb2c-51ea-a6cb-8c5e57f8d1d2,Oficinas2_hacia_In
2024-08-16,10:12:09,1723821128811798697,-500,13,traffic,0	P_WAN_1G_CED,wan,,539021105.0,17,4.0,883d9936-bb2c-51ea-a6cb-8c5e57f8d1d2,Oficinas2_hacia_In
2024-08-16,10:12:09,1723821128911798836,-500,13,traffic,0	_WAN_1G_CED,wan,,539021144.0,17,141.0,d7a125c2-eb6e-51ed-4fa5-7b2b1c6203ea,Administrativos_h
2024-08-16,10:12:09,1723821129291824183,-500,13,traffic,0	AN_1G_CED,wan,,539006710.0,17,11.0,bf2d1d48-bb2e-51ea-472e-d18b767a05be,Estudiantes2_hacia_I
2024-08-16,10:12:09,1723821129241814897,-500,13,traffic,0	P_WAN_1G_CED,wan,,539136392.0,6,1.0,43ef5ac6-bb2c-51ea-17e6-d583f2394236,TICS_hacia_Internet,
2024-08-16,10:12:09,1723821128992588766,-500,20,traffic,0	_WAN_1G_CED,wan,,538588788.0,17,11.0,bf2d1d48-bb2e-51ea-472e-d18b767a05be,Estudiantes2_hac
2024-08-16,10:12:08,1723821128051811600,-500,13,traffic,0	N_1G_CED,wan,,539020432.0,17,166.0,b5c36412-eb7a-51ee-7ae6-f73749d378c4,RedApuepse_to_Inter
2024-08-16,10:12:08,1723821128051813866,-500,13,traffic,0	AN_1G_CED,wan,,539020541.0,17,164.0,56fc656c-e60e-51ee-ddaa-fff1687a8a65,Red_FacultadSistema
2024-08-16,10:12:08,1723821128051810878,-500,13,traffic,0	WAN_1G_CED,wan,,539013318.0,17,141.0,d7a125c2-eb6e-51ed-4fa5-7b2b1c6203ea,Administrativos_ha
2024-08-16,10:12:07,1723821127651801923,-500,13,traffic,0	N_1G_CED,wan,,539139158.0,6,11.0,bf2d1d48-bb2e-51ea-472e-d18b767a05be,Estudiantes2_hacia_Int

Figura 8: Data inicial a la cual se le hará el análisis

Para extraer esta información era preciso el uso de la herramienta RStudio; para conocer el proceso de tratamiento con una data más pequeña. Luego, se recurrió al uso de un Script para sanear la información requerida para el análisis.

main.py

C: > Users > Usuario > Documents > TESIS > TESIS > tesis > main.py > ...

```python
import Scrip
import time
import sys
import union3

def cargar_progreso_por_etapa(etapas):
    for i, etapa in enumerate(etapas, start=1):
        sys.stdout.write(f"\rEjecutando etapa {i}/{len(etapas)}: {etapa}... {int((i/len(etapas)) * 100)}%")
        sys.stdout.flush()
        time.sleep(1)  # Simula la carga entre etapas (ajusta esto según lo necesites)

    print("\nTodas las etapas completadas.")
def main():
    print("Iniciando el proceso de ejecución...")
    etapas = ["Convirtiendo de LOG a CSV", "Ejecutando Limpieza", "Ejecutando el Xboost"]
    cargar_progreso_por_etapa(etapas)
    Scrip.main()
    print("Iniciando Limpieza...")
    union3.main()
    print("Iniciando Xboost")


if __name__ == "__main__":
    main()
```

Figura 9: Código del Script a ejecutar

Se realiza una ejecución en modo de bajo nivel con comandos, en donde ejecuta la lectura del archivo plano.

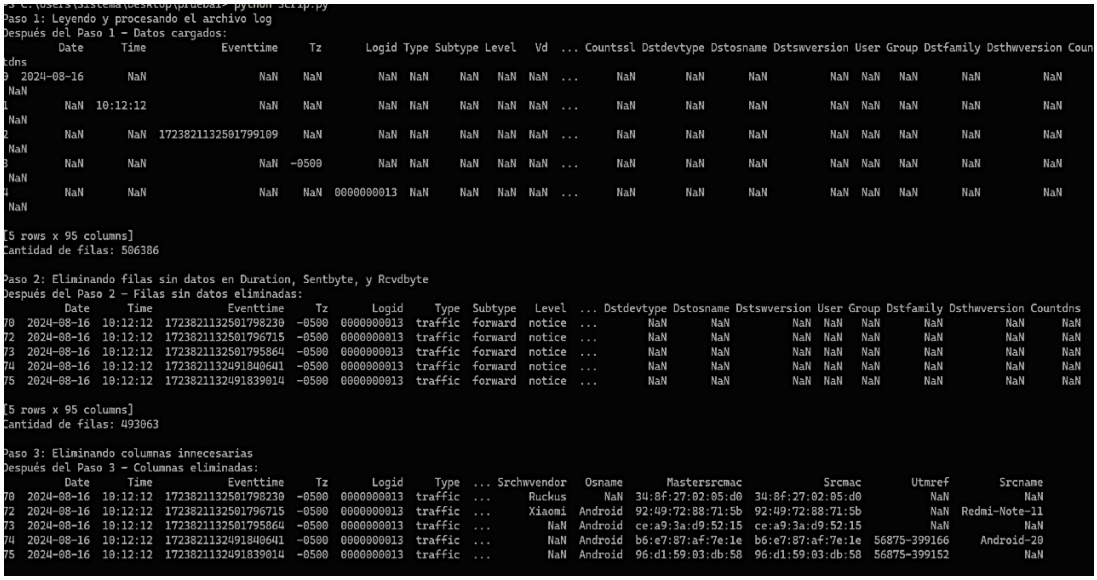


Figura 10: Ejecución del código de limpieza

Luego, se proporciona la recolección de datos necesaria para el posterior análisis de la fase de desarrollo.

Date	Time	Eventtime	Tz	Logid	Type	Level	Srcip	Srcport	Srcintf	Dstip	Dstport	Dstintf	Dsti
16/8/2024	10:12:12	1,72382E+18	-500	13	traffic	0		42382	VLAN_400		53	LACP_WAN_1G_CED	wan
16/8/2024	10:12:12	1,72382E+18	-500	13	traffic	0		45692	VLAN_70		53	LACP_WAN_1G_CED	wan
16/8/2024	10:12:12	1,72382E+18	-500	13	traffic	0		37152	VLAN_70		53	LACP_WAN_1G_CED	wan
16/8/2024	10:12:12	1,72382E+18	-500	13	traffic	0		57730	VLAN_71		443	LACP_WAN_1G_CED	wan
16/8/2024	10:12:12	1,72382E+18	-500	13	traffic	0		48476	VLAN_71		443	LACP_WAN_1G_CED	wan
16/8/2024	10:12:12	1,72382E+18	-500	13	traffic	0		43654	VLAN_70		443	LACP_WAN_1G_CED	wan
16/8/2024	10:12:12	1,72382E+18	-500	20	traffic	0		61566	VLAN_71		5228	LACP_WAN_1G_CED	wan
16/8/2024	10:12:12	1,72382E+18	-500	13	traffic	0		40784	VLAN_71		443	LACP_WAN_1G_CED	wan
16/8/2024	10:12:12	1,72382E+18	-500	13	traffic	0		51086	VLAN_71		5222	LACP_WAN_1G_CED	wan
16/8/2024	10:12:12	1,72382E+18	-500	13	traffic	0		7772	VLAN_400		443	LACP_WAN_1G_CED	wan
16/8/2024	10:12:12	1,72382E+18	-500	13	traffic	0		53858	VLAN_71		443	LACP_WAN_1G_CED	wan
16/8/2024	10:12:12	1,72382E+18	-500	13	traffic	0		46796	VLAN_70		443	LACP_WAN_1G_CED	wan
16/8/2024	10:12:12	1,72382E+18	-500	13	traffic	0		59384	VLAN_71		443	LACP_WAN_1G_CED	wan
16/8/2024	10:12:12	1,72382E+18	-500	20	traffic	0		49660	VLAN_70		443	LACP_WAN_1G_CED	wan
16/8/2024	10:12:12	1,72382E+18	-500	13	traffic	0		63916	VLAN_71		443	LACP_WAN_1G_CED	wan
16/8/2024	10:12:12	1,72382E+18	-500	13	traffic	0		56666	VLAN_71		53	LACP_WAN_1G_CED	wan
16/8/2024	10:12:12	1,72382E+18	-500	20	traffic	0		53054	VLAN_71		443	LACP_WAN_1G_CED	wan
16/8/2024	10:12:12	1,72382E+18	-500	20	traffic	0		50012	VLAN_70		443	LACP_WAN_1G_CED	wan
16/8/2024	10:12:12	1,72382E+18	-500	13	traffic	0		43840	VLAN_71		443	LACP_WAN_1G_CED	wan
16/8/2024	10:12:12	1,72382E+18	-500	13	traffic	0		40974	VLAN_71		443	LACP_WAN_1G_CED	wan
16/8/2024	10:12:12	1,72382E+18	-500	13	traffic	0		46550	VLAN_70		5222	LACP_WAN_1G_CED	wan

Figura 11: Resultado de Script de limpieza

2.2. Validación de los datos de la data set

2.2.1. Implementación de algoritmos de filtrado

En esta etapa, con el fin de garantizar la seguridad y detectar las posibles amenazas en las diversas solicitudes de usuarios, se procede con el entrenamiento de algoritmos de clasificación, permitiendo identificar patrones en los datos del tráfico web, distinguiendo entre comportamientos normales y anómalos. Para ello, se emplean tres algoritmos de machine learning: Isolation Forest, Random Forest y

XGBoost. Todos estos algoritmos se entrenan con el Data set limpiado previamente, para mejorar la precisión en la detección de las anomalías.

- **API de VirusTotal:** Esta API analiza la URL solicitada para determinar si se asocia con actividades maliciosas o si contiene algún tipo de virus; además, emplea bases de datos actualizadas y motores de antivirus para generar un informe detallado sobre seguridad de dicha página analizada.

Figura 12: Api de la Web VirusTotal

- **API de Criminal IP:** Una vez que la solicitud pasa por la primera API (VirusTotal), se somete al segundo filtro empleando la API de Criminal IP, donde la herramienta compara la IP de la página solicitada con una base de datos que contiene direcciones IP que se asocian a actividades maliciosas, como plataformas que estuvieron involucradas en ciberataques, pishing o distribución de malware.

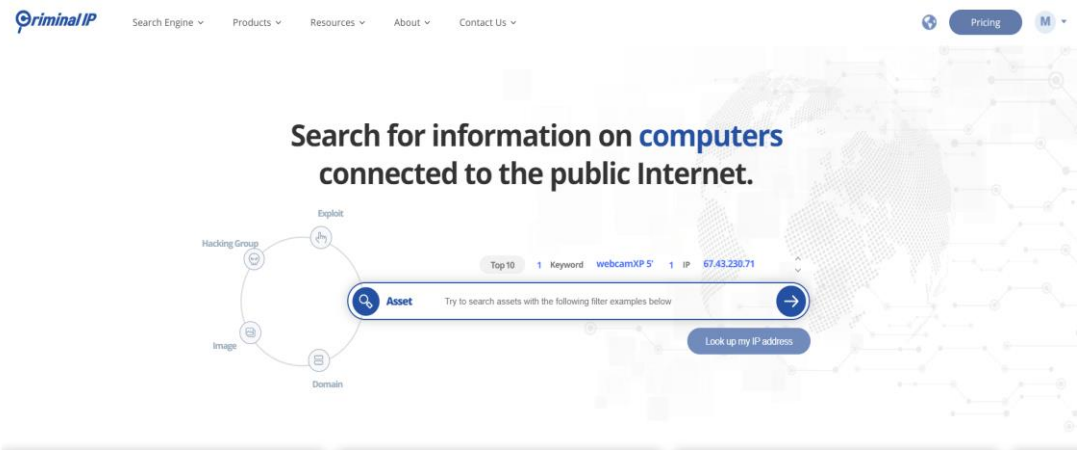


Figura 13: Api de la Web Criminal IP

2.2.2. Integración con MySQL

Una parte esencial del sistema es almacenar los datos de las solicitudes maliciosas o normales; para esto, se utiliza MySQL, siendo una base de datos relacional que permite registrar y guardar solicitudes que se han identificado como peligrosas o normales en cuanto al tráfico web. En este contexto, el programa Python envía la información sobre las respectivas solicitudes detectadas por las APIs a esta base de datos, almacenándolas, donde se incluyen los datos de la IP solicitada, fecha y el resultado del análisis.

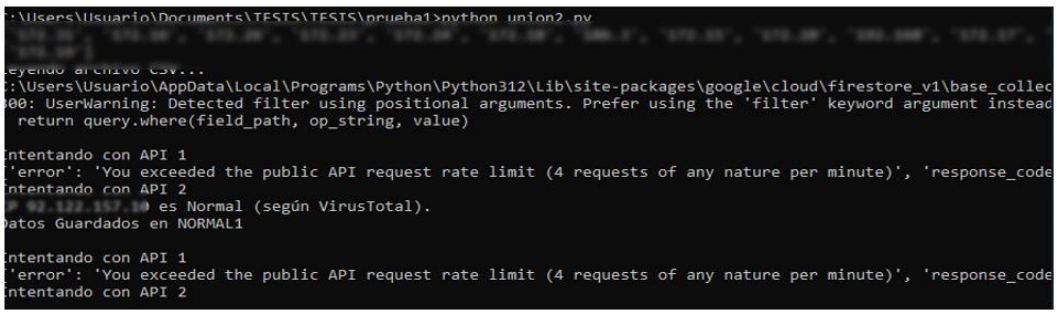


Figura 14: Ejecución del archivo plano

Se guarda la información en la base datos, siendo las IP anómalas y las que no se pudieron detectar con las debidas APIs.

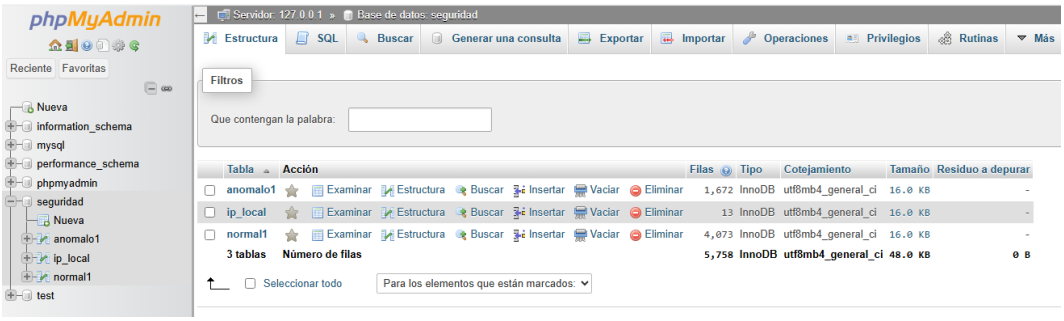


Figura 15: Base de datos MySQL

2.2.3. Proceso de ejecución de la validación

Se profundizó más en la limpieza del csv debido a que tenía aun datos que no permitían un buen entrenamiento de los algoritmos; por ende, esta vez se hizo un complemento con otros archivos. En este caso se comprobó si los datos eran

anómalos o normales a través de las APIs de virus total, realizando verificaciones para un entrenamiento adecuado de la IA.

Por tal motivo, se trabajó la codificación debido a varios factores por la cantidad de datos que proporciona el Data set. Por múltiples pruebas realizadas por el código, hubo un colapso en las APIs, ya que solo se estaban empleando dos en el código.

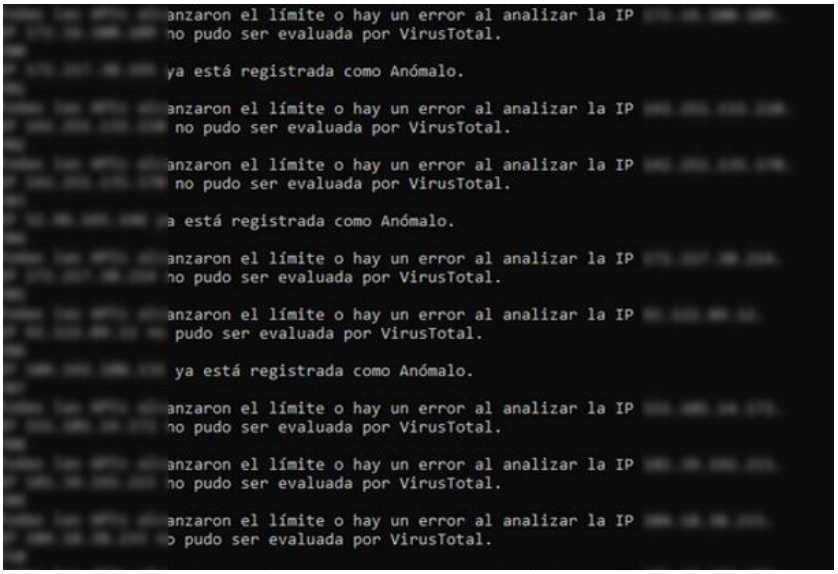


Figura 16: Saturación de APIs de virus total

Se agregaron más APIs de virus total con el fin de crear un bucle entre ellas; es decir, que empiece por una API, si con esta no se pudo continuar que pase a la siguiente API y así sucesivamente; con el fin de abarcar todos los datos del archivo CSV.

```
# Definir las API Keys
API_KEY1 = "84a6cd4ec57a4ad4729e135835dd46d29f7c67afff08573c2788a89f09f84622"
API_KEY2 = "4f278505cb5bdba811cb3c117457db459c8b3334e7f2d74d34f16d4b3bd7e4a0"
API_KEY3= "2e8ce00ad19bab9f7720ea67f9dc13319a8780215575813d4c2852295c20b219"
API_KEY4= "6e990e9e07f206bf83f7c28130ed090cd7008eac4fb6a731e7695e54fec99238"
# Estado de disponibilidad de las APIs
api1_available = True
api2_available = True
api3_available = True
api4_available = True
```

Figura 17: Implementación de nuevas APIs de virus total

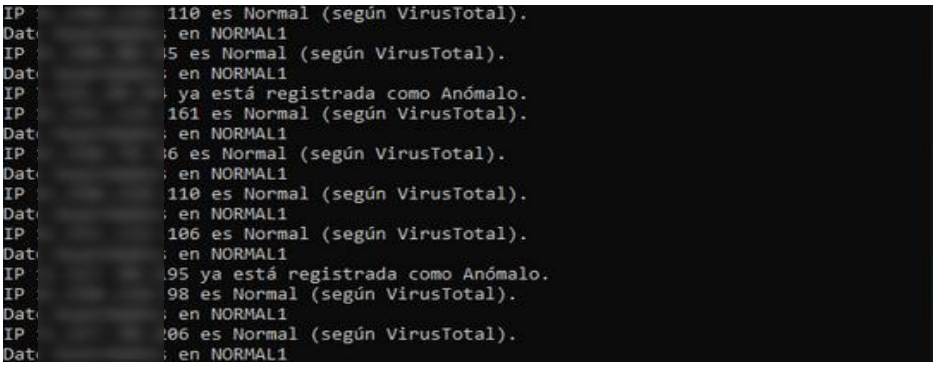


Figura 18: Ejecución del código para la verificación

Ahora, se tomaron en cuenta las IPs locales de la universidad, ya que, no se pretende analizar una conexión interna, por ende, se realizó esto en la codificación.

```
for index, row in data.iterrows():
    src_addr = row['Srcip']
    dst_addr = row['Dstip']
    ip_to_check = None

    if any(src_addr.startswith(ip) for ip in guia):
        ip_to_check = dst_addr
    elif any(dst_addr.startswith(ip) for ip in guia):
        ip_to_check = src_addr
    else:
        print(f"Ambas IPs {src_addr} y {dst_addr} son locales. Saltando...")
        resultados.append('Local')
        continue
```

Figura 19: Codificación para saltar el análisis en caso de que ambas IPs sean locales

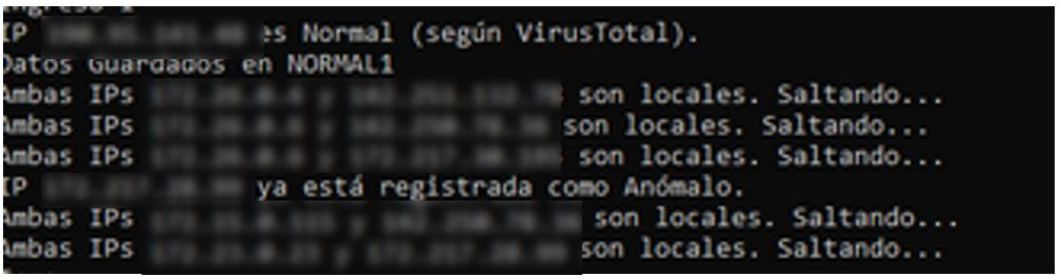


Figura 20: Ejecución del código

Se tuvo aun el problema de que no está leyendo bien las IPs locales; sin embargo, se dejó terminar el análisis para tener una mejor visualización con el nuevo CSV que se va a generar con una nueva columna con el resultado del análisis.

Srcip	Srcport	Srcintf	Dstip	Dstport	Dstintf	Dstintfrole	Dstinetsvc	Sessionid	Proto	Policyid	Poluuid	Policyname
[redacted]	56205	VLAN_121	[redacted]	443	LACP_WAN_1G_CED	wan		5,39E+09	17	1410	d7a125c2-eb6e-51ed-4fa5-7b2b1c6203ea	Administrativos_hacia_Internet
[redacted]	59197	VLAN_123	[redacted]	443	LACP_WAN_1G_CED	wan		5,39E+09	17	1500	35230bd8-9def-51ee-9098-92dc72feec09	TalentoHumano_hacia_Internet
[redacted]	58583	VLAN	[redacted]	443	LACP_WAN_1G_CED	wan		5,384E+09	17	1660	b5c36412-eb7a-51ee-7ae6-f73749d378c4	RedApuepse_to_Internet
[redacted]	59832	VLAN	[redacted]	80	LACP_WAN_1G_CED	wan		5,392E+09	6	1660	b5c36412-eb7a-51ee-7ae6-f73749d378c4	RedApuepse_to_Internet
[redacted]	41736	VLAN	[redacted]	80	LACP_WAN_1G_CED	wan		5,392E+09	6	1660	b5c36412-eb7a-51ee-7ae6-f73749d378c4	RedApuepse_to_Internet
[redacted]	62630	VLAN_121	[redacted]	443	LACP_WAN_1G_CED	wan		5,39E+09	17	1410	d7a125c2-eb6e-51ed-4fa5-7b2b1c6203ea	Administrativos_hacia_Internet
[redacted]	64720	VLAN126	[redacted]	443	LACP_WAN_1G_CED	wan		5,39E+09	17	1610	1c2eb9e6-da77-51ee-83e5-8f36d053dabc	EdificioVicerrectorado_to_Internet
[redacted]	61740	VLAN_121	[redacted]	443	LACP_WAN_1G_CED	wan		5,39E+09	17	1410	d7a125c2-eb6e-51ed-4fa5-7b2b1c6203ea	Administrativos_hacia_Internet
[redacted]	45047	VLAN	[redacted]	443	LACP_WAN_1G_CED	wan		5,39E+09	17	1660	b5c36412-eb7a-51ee-7ae6-f73749d378c4	RedApuepse_to_Internet
[redacted]	56109	Vlan127	[redacted]	443	LACP_WAN_1G_CED	wan		5,39E+09	17	1640	56fc656c-e60e-51ee-ddaa-fff1687a8a65	Red_FacultadSistemas_to_Internet
[redacted]	63619	VLAN_121	[redacted]	443	LACP_WAN_1G_CED	wan		5,39E+09	17	1410	d7a125c2-eb6e-51ed-4fa5-7b2b1c6203ea	Administrativos_hacia_Internet
[redacted]	52552	VLAN_123	[redacted]	443	LACP_WAN_1G_CED	wan		5,39E+09	17	1500	35230bd8-9def-51ee-9098-92dc72feec09	TalentoHumano_hacia_Internet
[redacted]	60630	VLAN_121	[redacted]	443	LACP_WAN_1G_CED	wan		5,39E+09	17	1410	d7a125c2-eb6e-51ed-4fa5-7b2b1c6203ea	Administrativos_hacia_Internet
[redacted]	48200	VLAN_121	[redacted]	443	LACP_WAN_1G_CED	wan		5,389E+09	17	1410	d7a125c2-eb6e-51ed-4fa5-7b2b1c6203ea	Administrativos_hacia_Internet
[redacted]	53736	VLAN	[redacted]	443	LACP_WAN_1G_CED	wan		5,39E+09	17	1660	b5c36412-eb7a-51ee-7ae6-f73749d378c4	RedApuepse_to_Internet
[redacted]	53652	VLAN_121	[redacted]	80	LACP_WAN_1G_CED	wan		5,392E+09	6	1410	d7a125c2-eb6e-51ed-4fa5-7b2b1c6203ea	Administrativos_hacia_Internet
[redacted]	50781	VLAN126	[redacted]	80	LACP_WAN_1G_CED	wan		5,391E+09	6	1610	1c2eb9e6-da77-51ee-83e5-8f36d053dabc	EdificioVicerrectorado_to_Internet
[redacted]	55726	VLAN_121	[redacted]	443	LACP_WAN_1G_CED	wan		5,387E+09	17	1410	d7a125c2-eb6e-51ed-4fa5-7b2b1c6203ea	Administrativos_hacia_Internet
[redacted]	64168	VLAN_121	[redacted]	443	LACP_WAN_1G_CED	wan		5,39E+09	17	1410	d7a125c2-eb6e-51ed-4fa5-7b2b1c6203ea	Administrativos_hacia_Internet
[redacted]	65456	Vlan127	[redacted]	443	LACP_WAN_1G_CED	wan		5,39E+09	17	1640	56fc656c-e60e-51ee-ddaa-fff1687a8a65	Red_FacultadSistemas_to_Internet
[redacted]	63064	VLAN126	[redacted]	443	LACP_WAN_1G_CED	wan		5,39E+09	17	1610	1c2eb9e6-da77-51ee-83e5-8f36d053dabc	EdificioVicerrectorado_to_Internet
[redacted]	65373	VLAN_121	[redacted]	443	LACP_WAN_1G_CED	wan		5,39E+09	17	1410	d7a125c2-eb6e-51ed-4fa5-7b2b1c6203ea	Administrativos_hacia_Internet
[redacted]	50519	VLAN_121	[redacted]	443	LACP_WAN_1G_CED	wan		5,39E+09	17	1410	d7a125c2-eb6e-51ed-4fa5-7b2b1c6203ea	Administrativos_hacia_Internet

Figura 21: Carga de datos en el CSV

En el proceso de pruebas, hubo varios problemas. No solo era el colapso de APIs, sino también de la base de datos en línea Firestore; por tal motivo como en la figura 20 algunos datos no fueron analizados, se optó tomar la base de datos MySQL; para esto se tuvo que comenzar nuevamente el análisis.

```
:\Users\Usuario\Documents\TESIS\TESIS\prueba1>python union2.py
[redacted]

eyendo archivo CSV...
:\Users\Usuario\AppData\Local\Programs\Python\Python312\Lib\site-packages\google\cloud\firestore_v1\base_collec
00: UserWarning: Detected filter using positional arguments. Prefer using the 'filter' keyword argument instead
  return query.where(field_path, op_string, value)

ntentando con API 1
'error': 'You exceeded the public API request rate limit (4 requests of any nature per minute)', 'response_code
ntentando con API 2
[redacted] es Normal (según VirusTotal).
atos Guardados en NORMAL1

ntentando con API 1
'error': 'You exceeded the public API request rate limit (4 requests of any nature per minute)', 'response_code
ntentando con API 2
```

Figura 22: Ejecución del código

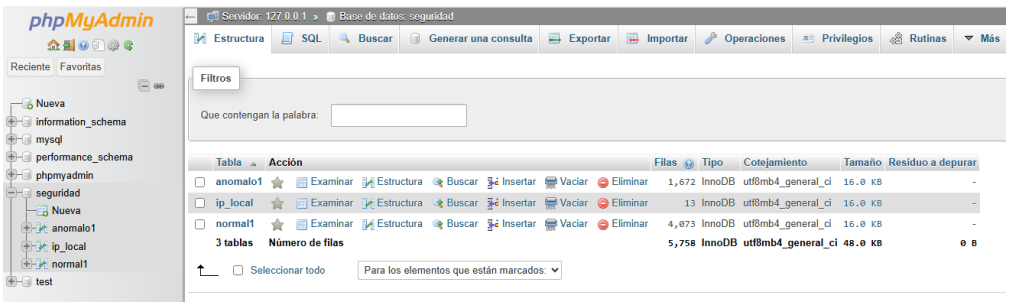
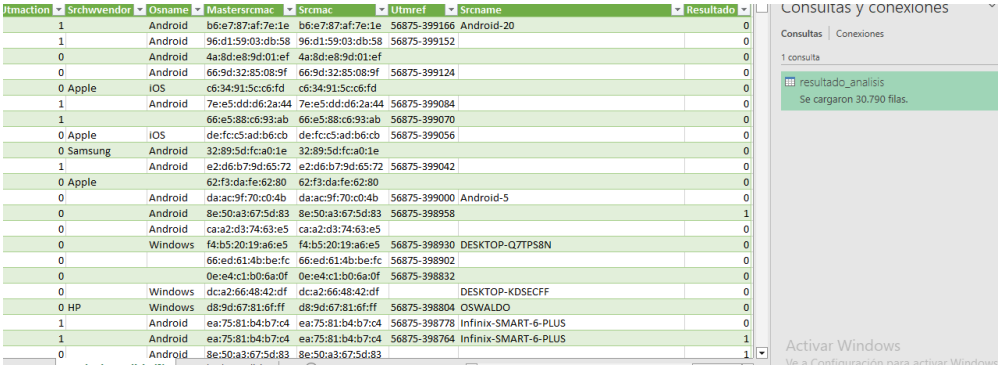


Figura 23: Almacenamiento de las IPs analizadas del CSV

Al final, se obtuvo el nuevo CSV con todos los datos analizados y directamente con la conversión de 0 para los datos normales y 1 para los anómalos.



Utmaction	Srchwvendor	Osname	Mastersrcmac	Srcmac	Utmref	Srcname	Resultado
	1	Android	b6:e7:87:af:7e:1e	b6:e7:87:af:7e:1e	56875-399166	Android-20	0
	1	Android	96:d1:59:03:db:58	96:d1:59:03:db:58	56875-399152		0
	0	Android	4a:8d:e8:9d:01:ef	4a:8d:e8:9d:01:ef			0
	0	Android	66:9d:32:85:08:9f	66:9d:32:85:08:9f	56875-399124		0
	0 Apple	iOS	c6:34:91:5c:c6:fd	c6:34:91:5c:c6:fd			0
	1	Android	7e:e5:dd:d6:2a:44	7e:e5:dd:d6:2a:44	56875-399084		0
	1		66:e5:88:c6:93:ab	66:e5:88:c6:93:ab	56875-399070		0
	0 Apple	iOS	de:fc:c5:ad:b6:cb	de:fc:c5:ad:b6:cb	56875-399056		0
	0 Samsung	Android	32:89:5d:fc:a0:1e	32:89:5d:fc:a0:1e			0
	1	Android	e2:d6:b7:9d:65:72	e2:d6:b7:9d:65:72	56875-399042		0
	0 Apple		62:f3:da:fe:62:80	62:f3:da:fe:62:80			0
	0	Android	da:ac:9f:70:c0:4b	da:ac:9f:70:c0:4b	56875-399000	Android-5	0
	0	Android	8e:50:a3:67:5d:83	8e:50:a3:67:5d:83	56875-398958		1
	0	Android	ca:a2:d3:74:63:e5	ca:a2:d3:74:63:e5			0
	0	Windows	f4:b5:20:19:a6:e5	f4:b5:20:19:a6:e5	56875-398930	DESKTOP-Q7TPS8N	0
	0		66:ed:61:4b:be:fc	66:ed:61:4b:be:fc	56875-398902		0
	0		0e:e4:c1:b0:6a:0f	0e:e4:c1:b0:6a:0f	56875-398832		0
	0	Windows	dc:a2:66:48:42:df	dc:a2:66:48:42:df		DESKTOP-KDSECFF	0
	0 HP	Windows	d8:9d:67:81:6f:ff	d8:9d:67:81:6f:ff	56875-398804	OSWALDO	0
	1	Android	ea:75:81:b4:b7:c4	ea:75:81:b4:b7:c4	56875-398778	Infinix-SMART-6-PLUS	0
	1	Android	ea:75:81:b4:b7:c4	ea:75:81:b4:b7:c4	56875-398764	Infinix-SMART-6-PLUS	1
	0	Android	8e:50:a3:67:5d:83	8e:50:a3:67:5d:83			1

Figura 24: CSV generado con la nueva columna de 0 normal y 1 anómalo

Se elaboró un solo código de toda la limpieza que se tuvo que realizar al Data set.

```
5
6   def cargar_progreso_por_etapa(etapas):
7       for i, etapa in enumerate(etapas, start=1):
8           sys.stdout.write(f"\rEjecutando etapa {i}/{len(etapas)}: {etapa}... {int((i/len(etapas)) * 100)}%")
9           sys.stdout.flush()
10          time.sleep(1)  # Simula la carga entre etapas (ajusta esto según lo necesites)
11
12      print("\nTodas las etapas completadas.")
13  def main():
14      print("Iniciando el proceso de ejecución...")
15      etapas = ["Convirtiendo de LOG a CSV", "Ejecutando Limpieza", "Ejecutando el Xboost"]
16      cargar_progreso_por_etapa(etapas)
17      Scrip.main()
18      print("Iniciando Limpieza...")
19      union3.main()
20      print("Iniciando Xboost")
21
22
23  if __name__ == "__main__":
24      main()
```

Figura 25: Código unificado para la limpieza partiendo del archivo log

Esta fase aborda el proceso de diseño y desarrollo del sistema de detección de tráfico web normal y anómalo, detallando la codificación de los algoritmos que permiten el rastreo, así como el análisis y el almacenamiento de las solicitudes provenientes de las computadoras conectadas a la red de la universidad. El análisis se realiza empleando un Data set que proporcionó el Departamento de TI de la institución, el cual contiene información sobre las IPs de las solicitudes, que fueron previamente recopiladas por medio de un script ejecutado. Sin embargo, solo se considera a la facultad de sistemas y telecomunicaciones.

2.3. Diseño del sistema de detección

Se diseñó el sistema que permite capturar y analizar el tráfico web en la red de la institución de educación superior, sin necesidad de redirigir las peticiones a una

Dataset. Por ende, las IPs de los ordenadores conectados fueron extraídos en un archivo plano. Cabe recalcar, que el sistema de detección está diseñado para recoger todas las solicitudes enviadas desde las computadoras de una red hacia las direcciones IP externas, pasando cada solicitud por la IP principal, donde se encuentra alojado el programa Python, que se encarga del procesamiento y análisis del tráfico web.

A continuación, se muestra un diagrama de red de computadoras conectadas, así como un dispositivo con la IP 192.168.0.200, los cuales realizan peticiones que se envían a la IP principal. En el diagrama se puede observar la estructura del flujo de datos que se debe seguir, desde la solicitud de un dispositivo hacia un sitio web, pasando por el router, hasta llegar al servidor principal, donde el tráfico es analizado por los diferentes filtros de seguridad.

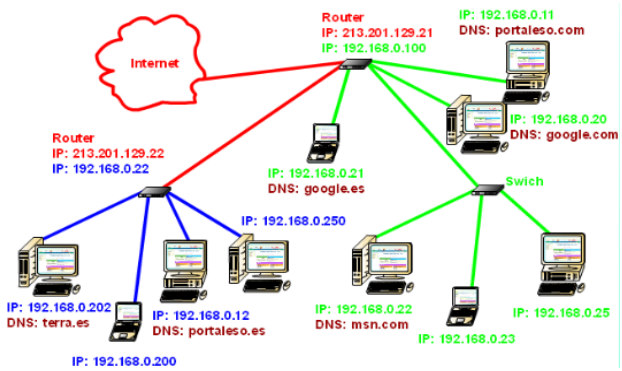


Figura 26: Análisis lógica de la red

2.4.Entrenamiento de los algoritmos

Los algoritmos de entrenamiento implican alimentar el modelo con una gran cantidad de datos extraídos del conjunto de información para aprender los patrones y relaciones dentro de ellos. Durante este proceso, el algoritmo ajusta parámetros internos, intentando minimizar errores en la predicción y clasificación de la IP a analizar.

Además, se realizan múltiples iteraciones a medida que el modelo evalúa y refina sus predicciones, acercándose progresivamente a la precisión deseada del análisis

de red. A medida que avanza el entrenamiento, el rendimiento se evalúa con datos de prueba para evitar el sobreajuste, que se programa en el siguiente código.

2.4.1. Entrenamiento del algoritmo Isolation Forest

El algoritmo Isolation Forest tiene como objetivo detectar anomalías examinando puntos de datos que pueden distinguirse de los sobresalientes. Para entrenar este modelo, se carga el conjunto de datos que ya se ha limpiado y se modifica el Isolation Forest para que aprenda a reconocer patrones comunes en la información. Es un modelo beneficioso para identificar anomalías, generando estructuras de árboles de decisión que resaltan datos que no actúan como la mayoría. Esta capacitación permitirá asignar calificaciones de anomalías a los registros, facilitando una clasificación rápida.

```python
# Función para entrenar el modelo Isolation Forest
def entrenar_modelo_isolation_forest(data):
    # Convertir IPs a subredes
    data['Srcip'] = data['Srcip'].apply(lambda x: ip_to_subnet(x, prefix=24))
    data['Dstip'] = data['Dstip'].apply(lambda x: ip_to_subnet(x, prefix=24))

    # Manejar valores NaN
    from sklearn.impute import SimpleImputer
    imputer = SimpleImputer(strategy='mean')
    data[['Srcport', 'Dstport', 'Sentbyte', 'Rcvdbyte', 'Sentpkt', 'Rcvdpkt']] = imputer.fit_transform(
        data[['Srcport', 'Dstport', 'Sentbyte', 'Rcvdbyte', 'Sentpkt', 'Rcvdpkt']]
    )

    # Seleccionar las características para entrenamiento
    X = data[['Srcip', 'Dstip', 'Srcport', 'Dstport', 'Sentbyte', 'Rcvdbyte', 'Sentpkt', 'Rcvdpkt']]
    y = data['Resultado']

    # Entrenar modelo Isolation Forest
    model = IsolationForest(n_estimators=100, max_samples='auto', contamination=0.1, random_state=42)
    model.fit(X, y)

    # Guardar modelo entrenado
    joblib.dump(model, 'modelo_isolation_forest.pkl')
```

Figura 27: Código del entrenamiento de Isolation Forest

A través de su ejecución, se genera una vista previa en formato CMD, tal como se muestra a continuación:

```
C:\Users\Usuario\Documents\TESIS\TESIS\tesis>python isolation.py
Iniciando el análisis de IPs...
Modelo Isolation Forest cargado exitosamente.
Archivo resultado_vulnerable_isolation.csv generado con éxito en la misma carpeta que el script.
```

Figura 28: Ejecución de Isolation Forest

2.4.2. Entrenamiento del algoritmo Random Forest

El modelo Random Forest es un algoritmo de clasificación que genera múltiples árboles de decisión y elige el resultado final en función de la concordancia de estos

árboles, aumentando así la precisión de los mismos. En este entrenamiento, el algoritmo obtuvo datos clasificados, es decir, normales o anormales, para crear los árboles de decisión. Esto reduce los errores en la clasificación del tráfico web.

```python
# Función para entrenar el modelo Random Forest
def entrenar_modelo_random_forest(data):
    # Convertir IPs a subredes
    data['Srcip'] = data['Srcip'].apply(lambda x: ip_to_subnet(x, prefix=24))
    data['Dstip'] = data['Dstip'].apply(lambda x: ip_to_subnet(x, prefix=24))

    X = data[['Srcip', 'Dstip', 'Srcport', 'Dstport',
              'Sentbyte', 'Rcvdbyte', 'Sentpkt', 'Rcvdpkt']]
    y = data['Resultado']
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

    model = RandomForestClassifier(n_estimators=100, max_depth=10, random_state=42)
    model.fit(X_train, y_train)

    joblib.dump(model, 'modelo_random_forest.pkl')
```

Figura 29: Codificación para el entrenamiento de Random Forest

Mediante su ejecución, se tendrá una previsualización en forma y modelado de CMD o código a bajo nivel.

```
C:\Users\Usuario\Documents\TESIS\TESIS\tesis>python forest.py
Iniciando el análisis de IPs...
Modelo Random Forest cargado exitosamente.
Archivo resultado_vulnerable_forest.csv generado con éxito en la misma carpeta que el script.
```

Figura 30: Ejecución del código de Random Forest

2.4.3. Entrenamiento del algoritmo XGBoost

XGBoost es un modelo de escala que modifica múltiples árboles secuencialmente para optimizar el rendimiento y reducir los errores residuales. En este entrenamiento, el modelo recibe el conjunto de datos listo, lo que mejora el proceso de aprendizaje para rectificar las predicciones incorrectas en cada fase.

```python
# Función para entrenar el modelo XGBoost
def entrenar_modelo_xgboost(data):
    # Convertir IPs a subredes
    data['Srcip'] = data['Srcip'].apply(lambda x: ip_to_subnet(x, prefix=24))
    data['Dstip'] = data['Dstip'].apply(lambda x: ip_to_subnet(x, prefix=24))

    X = data[['Srcip', 'Dstip', 'Srcport', 'Dstport',
              'Sentbyte', 'Rcvdbyte', 'Sentpkt', 'Rcvdpkt']]
    y = data['Resultado']
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
    model = xgb.XGBClassifier(n_estimators=100, learning_rate=0.1, max_depth=5, subsample=0.8, colsample_bytree=0.8,scale_pos_weight=len(y_train[y_train == 0]) / len(y_train[y_t
    model.fit(X_train, y_train)

    joblib.dump(model, 'modelo_xgboost_optimizado.pkl')
```

Figura 31: Codificación para el entrenamiento de XGBoost

Mediante su ejecución, se obtendrá una representación en formato CMD o código de bajo nivel.

```
C:\Users\Usuario\Documents\TESIS\TESIS\tesis>python xboost1.py
Iniciando el análisis de IPs...
Modelo XGBoost optimizado cargado exitosamente.
Archivo resultado_vulnerable_xboost.csv generado con éxito en la misma carpeta que el script.
```

Figura 32: Ejecución del código de XGBoost

Se comprueban los resultados que se obtuvieron de los tres algoritmos, para ver la efectividad que tienen cada uno en el entrenamiento, en comparación con el resultado de las APIs.

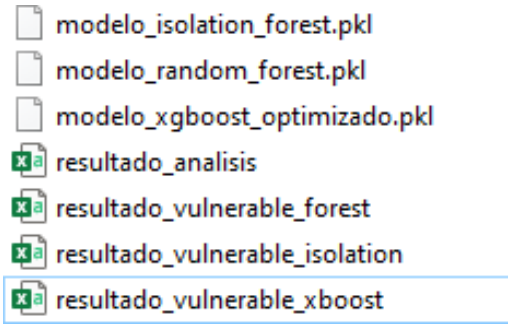


Figura 33: Resultado del entrenamiento de los algoritmos de Machine Learning

Se genera el nuevo archivo en formato CSV con la columna adicional que presenta los resultados del entrenamiento de los diversos algoritmos de Machine Learning.

Srchwvendor	Osname	Mastersrcmac	Srcmac	Utmref	Srcname	Resultado	Anomalia
	Android	b6:e7:87:af:7e:1e	b6:e7:87:af:7e:1e	56875-399166	Android-20	0	Normal
	Android	96:d1:59:03:db:58	96:d1:59:03:db:58	56875-399152		0	Normal
	Android	4a:8d:e8:9d:01:ef	4a:8d:e8:9d:01:ef			0	Normal
	Android	66:9d:32:85:08:9f	66:9d:32:85:08:9f	56875-399124		0	Normal
Apple	iOS	c6:34:91:5c:c6:fd	c6:34:91:5c:c6:fd			0	Normal
	Android	7e:e5:dd:d6:2a:44	7e:e5:dd:d6:2a:44	56875-399084		0	Normal
		66:e5:88:c6:93:ab	66:e5:88:c6:93:ab	56875-399070		0	Normal
Apple	iOS	de:fc:c5:ad:b6:cb	de:fc:c5:ad:b6:cb	56875-399056		0	Normal
Samsung	Android	32:89:5d:fc:a0:1e	32:89:5d:fc:a0:1e			0	Normal
	Android	e2:d6:b7:9d:65:72	e2:d6:b7:9d:65:72	56875-399042		0	Normal
Apple		62:f3:da:fe:62:80	62:f3:da:fe:62:80			0	Normal
	Android	da:ac:9f:70:c0:4b	da:ac:9f:70:c0:4b	56875-399000	Android-5	0	Normal
	Android	8e:50:a3:67:5d:83	8e:50:a3:67:5d:83	56875-398958		1	Anómalo
	Android	ca:a2:d3:74:63:e5	ca:a2:d3:74:63:e5			0	Normal
	Windows	f4:b5:20:19:a6:e5	f4:b5:20:19:a6:e5	56875-398930	DESKTOP-Q7TPS8N	0	Normal
		66:ed:61:4b:be:fc	66:ed:61:4b:be:fc	56875-398902		0	Normal
		0e:e4:c1:b0:6a:0f	0e:e4:c1:b0:6a:0f	56875-398832		0	Normal
	Windows	dc:a2:66:48:42:df	dc:a2:66:48:42:df		DESKTOP-KDSECFF	0	Normal
HP	Windows	d8:9d:67:81:6f:ff	d8:9d:67:81:6f:ff	56875-398804	OSWALDO	0	Normal
	Android	ea:75:81:b4:b7:c4	ea:75:81:b4:b7:c4	56875-398778	Infinix-SMART-6-PLUS	0	Normal
	Android	ea:75:81:b4:b7:c4	ea:75:81:b4:b7:c4	56875-398764	Infinix-SMART-6-PLUS	1	Anómalo

Figura 34: CSV nuevo con columna de resultados del entrenamiento de los algoritmos

Fase 3. Monitoreo y Evaluación

En esta fase se llevará a cabo el monitoreo del Data set de las solicitudes del tráfico web, evaluando la efectividad del sistema. Se analizarán los datos obtenidos para la identificación de patrones de tráfico anómalos y normales, así como la precisión en la detección de amenazas.

3.1. Monitoreo de los datos

El monitoreo continuo de red es fundamental para la detección de anomalías en el tráfico web, garantizando que la red de la institución esté protegida contra ciber amenazas.

El siguiente diagrama de flujo, corresponde al proceso que se debe llevar a cabo desde las computadoras de la red, pasando por los filtros de seguridad y culminando con el almacenamiento de las solicitudes en la base de datos.

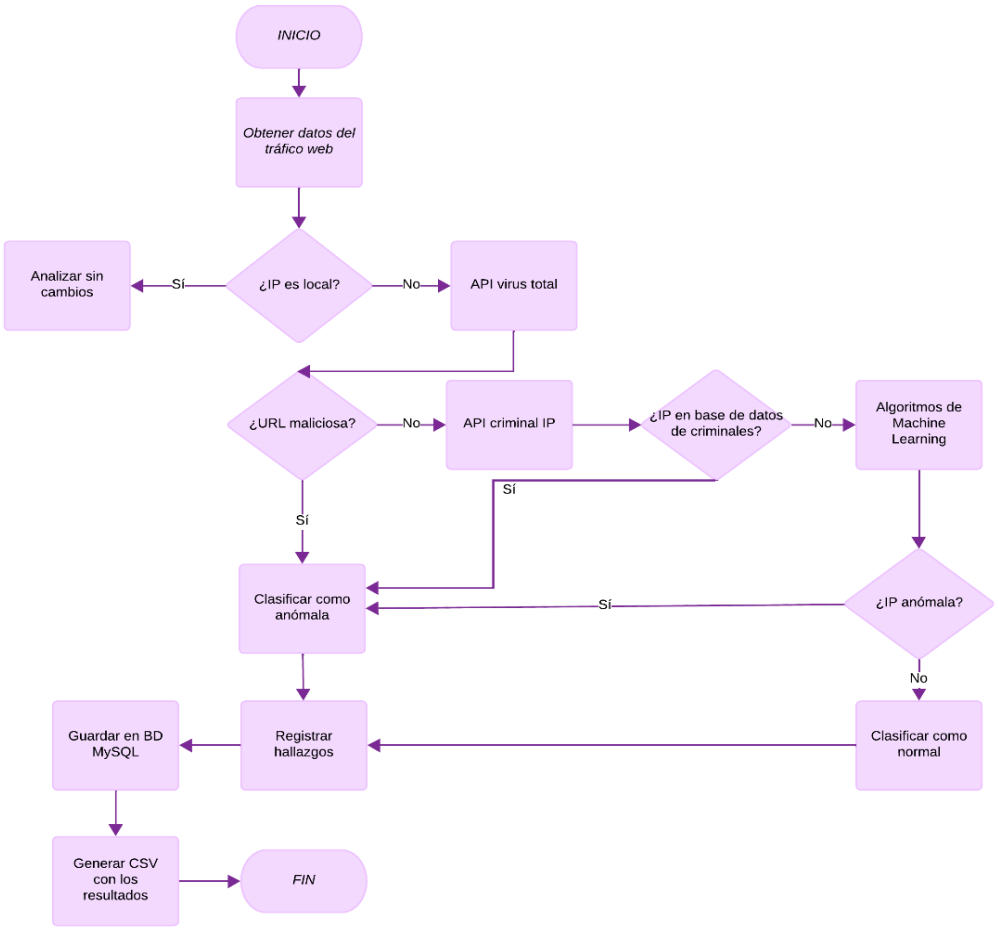


Figura 35: Diagrama de la ejecución

3.2. Evaluación de rendimiento y seguridad

Se realiza una evaluación acerca del rendimiento del sistema para la detección de tráfico web, incluyendo el análisis de solicitudes marcadas incorrectamente como peligrosas y solicitudes peligrosas que no se detectaron.

El siguiente diagrama de flujo, corresponde al proceso que se lleva a cabo desde la recopilación de registros de incidentes, pasando por las diversas solicitudes y registrando los hallazgos correspondientes.

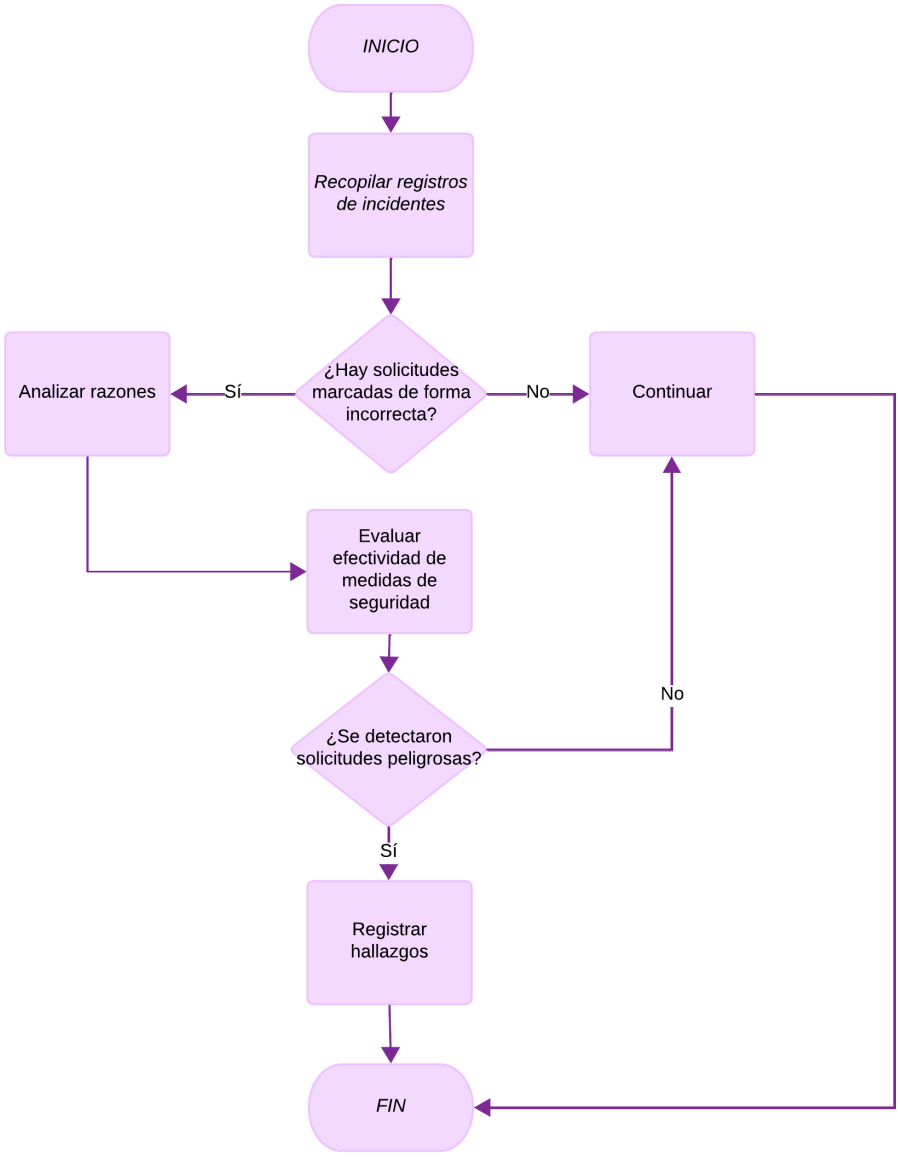


Figura 36: Diagrama de la ejecución de hallazgos

3.3. Evaluación de los algoritmos de Machine Learning

Para evaluar los algoritmos y saber que están analizando de forma correcta los datos del archivo, se toma como referencia el resultado de las Apis que se obtuvo en la parte de la validación de los datos. En este caso se realiza una comparativa entre los resultados de las Apis y los resultados que arrojen los modelos entrenados de Machine Learning con la finalidad de conocer si la capacidad de predecir no varía en gran dimensión ante un dato real como es el que nos proporciona la columna con los resultados de las Apis.

Para realizar la comparación se utiliza los modelos entrenados que tienen como una de las variables principales el resultado de las Apis en su entrenamiento y los modelos entrenados en base a la reconstrucción de variables tomadas del archivo.

Se visualiza la evaluación de los modelos a través de la Curva ROC la cual está presentada por una línea azul que muestra el desempeño que tiene el algoritmo de Machine Learning, representando la variación que hay entre la sensibilidad y la especificidad del modelo a medida que se vaya ajustando el umbral de clasificación. También presenta una línea punteada, la cual es una línea de referencia que muestra un modelo que realiza predicciones; por ende, si la curva azul se encontrara más cerca indicaría que el modelo de Machine Learning no posee la capacidad predictiva.

3.3.1. Resultados del algoritmo XGBoost

Como se muestra en la Figura 37 el área bajo la curva es de 0.75 lo que nos indica que el modelo entrenado con la columna del resultado de las Apis está prediciendo aceptablemente. En cambio, el área bajo la curva del modelo entrenado con la reconstrucción de variables en la Figura 38 nos indica que su capacidad de predecir es mas alta, por lo que se puede comprobar que al otorgarle más variables en el entrenamiento del algoritmo acierta un poco más en detectar patrones para identificar las IPs anómalas y normales

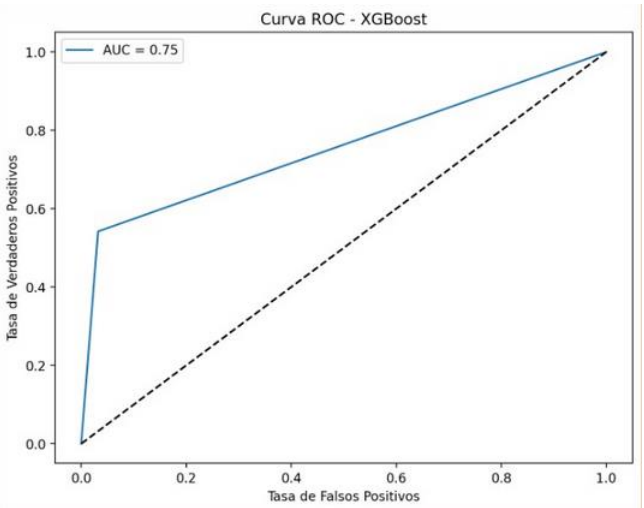


Figura 37: Gráfica ROC de XGBoost con modelo entrenado con APIs

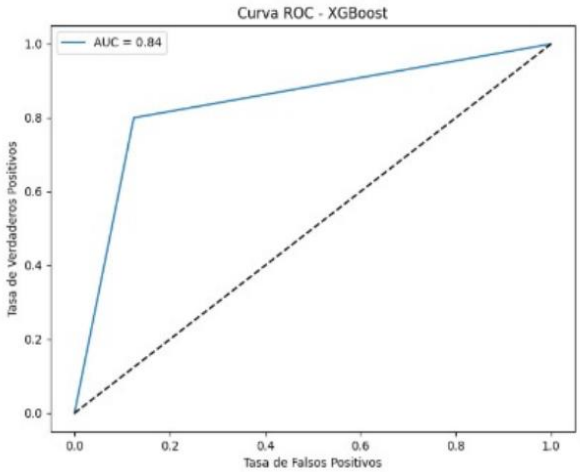


Figura 38: Gráfica ROC de XGBoost con modelo entrenado con reconstrucción

3.3.2. Resultados del algoritmo Random Forest

Para el algoritmo de Random Forest el área bajo la curva es de 0.67 por ende su capacidad de predecir es limitada aun con el valor del resultado de las apis, por otro lado, el área bajo la curva con el modelo de la reconstrucción nos indica que, aunque tenga más datos para identificar patrones solo posee el 52% de clasificar de forma adecuada los datos.

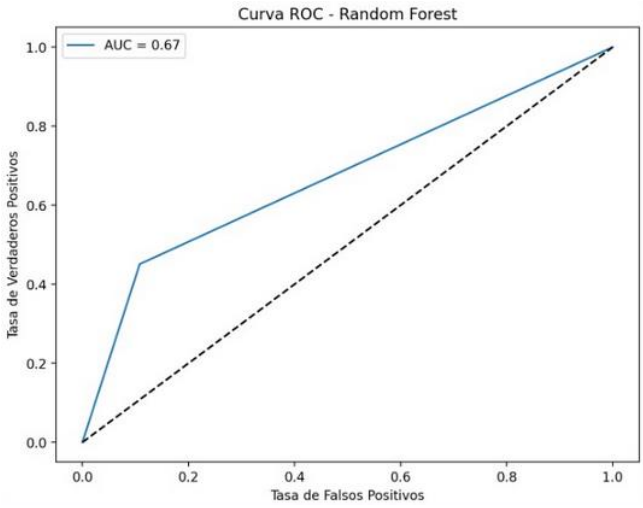


Figura 39: Gráfica ROC de Random Forest con modelo entrenado con APIs

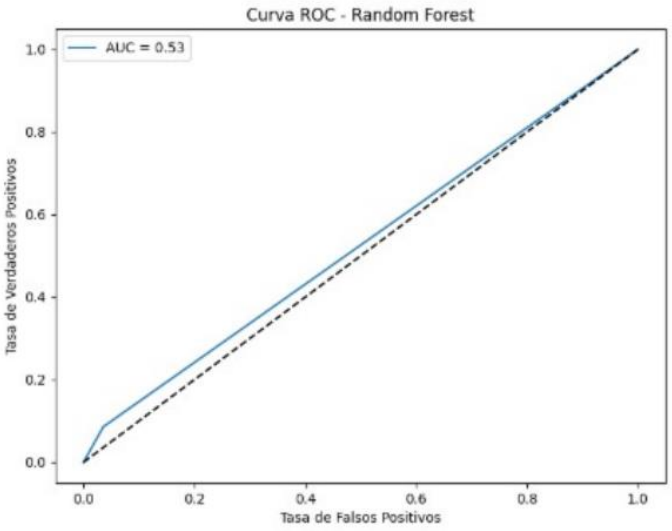


Figura 40: Gráfica ROC de Random Forest con modelo entrenado por reconstrucción

3.3.3. Resultados del algoritmo Isolation Forest

En el caso de Isolation Forest el resultado que se obtuvo con el modelo entrenado por la columna de las Apis arrojo un área bajo la curva menor que la de un modelo aleatorio, lo que nos demuestra que su capacidad de predecir es muy baja aun con un valor certero como el de las Apis. Sin embargo, con el modelo entrenado con la reconstrucción se puede observar que la probabilidad de predecir correctamente aumento no tan exageradamente, pero si hay un leve cambio.

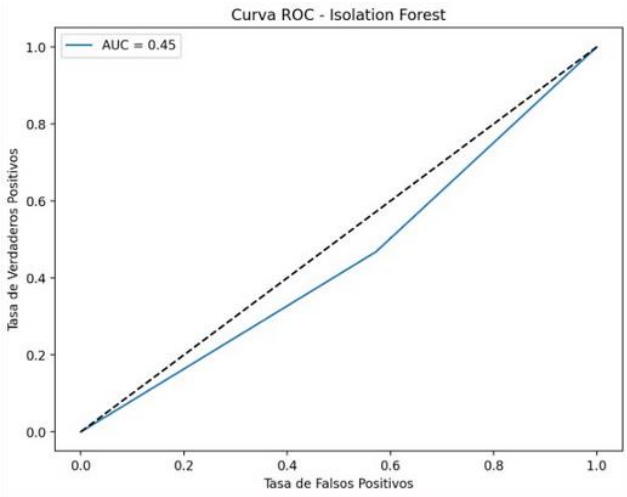


Figura 41: Gráfica ROC de Isolation Forest con modelo entrenado con APIs

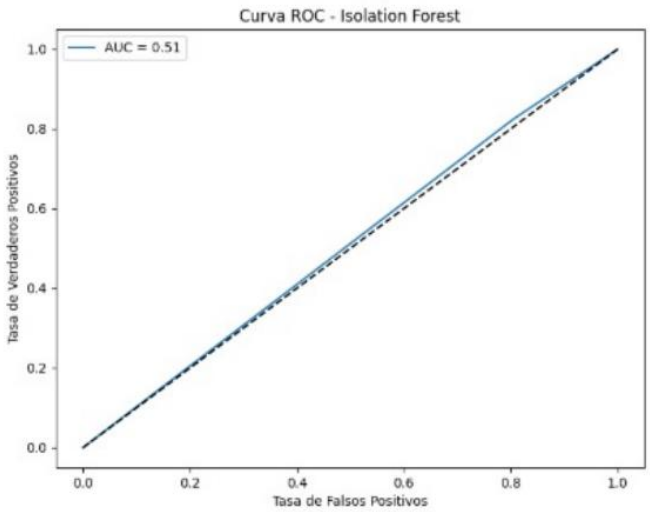


Figura 42: Gráfica ROC de Isolation Forest con modelo entrenado por reconstrucción

Los resultados que se obtuvieron de la comparativa nos indican que el modelo en donde se reconstruyeron las variables, posee mejor capacidad de predecir y clasificar de forma adecuada cada instancia de los datos analizados. Por lo que al ser robusto debido a la reconstrucción tiene una ventaja ante el modelo enfocado solo por al resultado generado del análisis de IPs a través de las Apis.

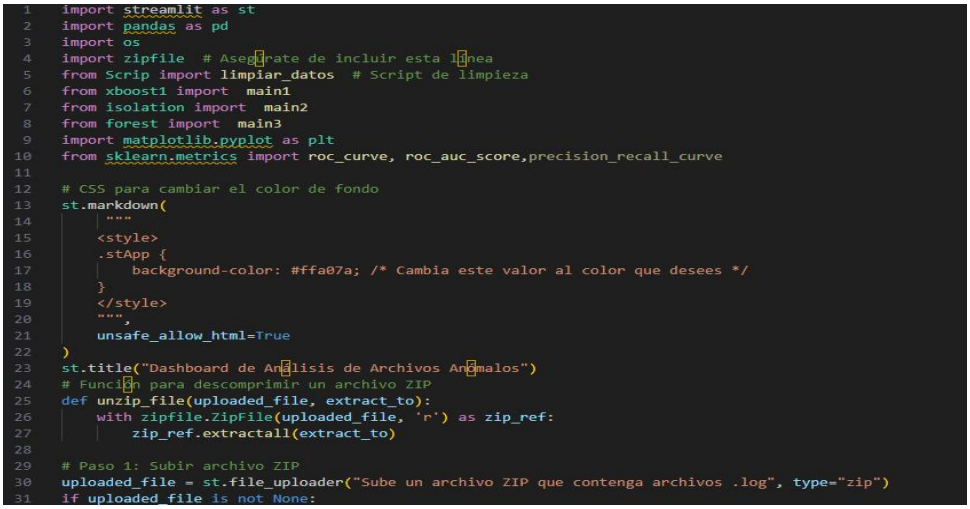
El margen de error que presenta el modelo es mínimo por lo que da esa confianza de que cualquier archivo que requiera ser analizado el modelo entrenado tanto de XGBoost, Random Forest e Isolation Forest pueden realizarlo sin ningún problema ofreciendo información relevante después del análisis. Se demuestra que el modelo de Machine Learning que tiene mayor capacidad de detectar las clases correctamente es el Xgboost.

Fase 4. Reportes

Los resultados obtenidos después de ser analizados se presentarán de manera visual mediante un Dashboard, el cual servirá para mostrar los datos de forma gráfica y comprensible, ayudando a la toma de decisiones. Por medio de los gráficos interactivos, los usuarios podrán analizar los respectivos resultados eficientemente, permitiendo una mejor interpretación del tráfico web anómalo y normal.

A continuación, se detalla el proceso de creación para el Dashboard:

1. Se realiza la codificación en el lenguaje HTML, llamando a los enlaces necesarios para brindar la funcionalidad y los estilos CSS. De la misma forma, se añaden las librerías con respecto a los gráficos que se visualizarán.



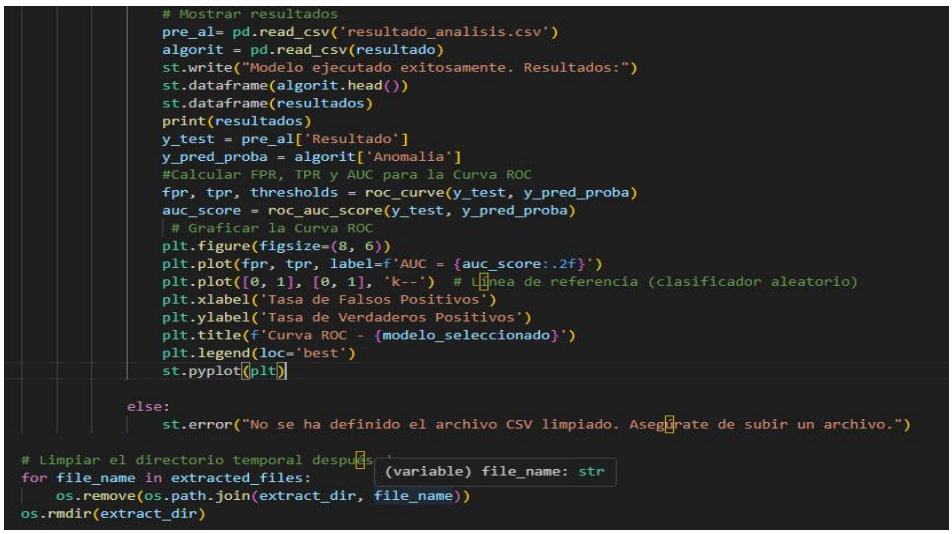
```
import streamlit as st
import pandas as pd
import os
import zipfile  # Asegúrate de incluir esta línea
from Scrip import limpiar_datos  # Script de limpieza
from xboost1 import  main1
from isolation import  main2
from forest import  main3
import matplotlib.pyplot as plt
from sklearn.metrics import roc_curve, roc_auc_score,precision_recall_curve

# CSS para cambiar el color de fondo
st.markdown(
    """
    <style>
    .stApp {
        background-color: #ffa07a; /* Cambia este valor al color que desees */
    }
    </style>
    """,
    unsafe_allow_html=True
)
st.title("Dashboard de Análisis de Archivos Anómalos")
# Función para descomprimir un archivo ZIP
def unzip_file(uploaded_file, extract_to):
    with zipfile.ZipFile(uploaded_file, 'r') as zip_ref:
        zip_ref.extractall(extract_to)

# Paso 1: Subir archivo ZIP
uploaded_file = st.file_uploader("Sube un archivo ZIP que contenga archivos .log", type="zip")
if uploaded_file is not None:
```

Figura 43: Código del dashboard

2. Se agrega la llamada al archivo de JavaScript donde se hallará la lógica de las gráficas acerca del tráfico web.



```
            # Mostrar resultados
            pre_al= pd.read_csv('resultado_analisis.csv')
            algorit = pd.read_csv(resultado)
            st.write("Modelo ejecutado exitosamente. Resultados:")
            st.dataframe(algorit.head())
            st.dataframe(resultados)
            print(resultados)
            y_test = pre_al['Resultado']
            y_pred_proba = algorit['Anomalia']
            #Calcular FPR, TPR y AUC para la Curva ROC
            fpr, tpr, thresholds = roc_curve(y_test, y_pred_proba)
            auc_score = roc_auc_score(y_test, y_pred_proba)
             # Graficar la Curva ROC
            plt.figure(figsize=(8, 6))
            plt.plot(fpr, tpr, label=f'AUC = {auc_score:.2f}')
            plt.plot([0, 1], [0, 1], 'k--')  # Línea de referencia (clasificador aleatorio)
            plt.xlabel('Tasa de Falsos Positivos')
            plt.ylabel('Tasa de Verdaderos Positivos')
            plt.title(f'Curva ROC - {modelo_seleccionado}')
            plt.legend(loc='best')
            st.pyplot(plt)

        else:
            st.error("No se ha definido el archivo CSV limpiado. Asegúrate de subir un archivo.")

# Limpiar el directorio temporal después
for file_name in extracted_files:
    os.remove(os.path.join(extract_dir, file_name))
os.rmdir(extract_dir)
```

(variable) file_name: str

Figura 44: Código del análisis del resultado

3. En la etiqueta de cuerpo, se crea una etiqueta para el título principal del Dashboard, así como los dos contenedores div para mostrar ambas gráficas de tráfico web con curvas ROC.

```python
import streamlit as st
import pandas as pd
import os
import zipfile
from Scrip import limpiar_datos
from xboost1 import main1
from isolation import main2
from forest import main3
import matplotlib.pyplot as plt
from sklearn.metrics import roc_curve, roc_auc_score

# CSS para cambiar el color de fondo
st.markdown(
    """
    <style>
    .stApp {
        background-color: #ffa07a;
    }
    </style>
    """,
    unsafe_allow_html=True
)

st.title("Dashboard de Análisis de Archivos Anómalos")

# Función para descomprimir un archivo ZIP
def unzip_file(uploaded_file, extract_to):
    with zipfile.ZipFile(uploaded_file, 'r') as zip_ref:
        zip_ref.extractall(extract_to)

# Paso 1: Subir archivo ZIP
```

Figura 45: Código para ejecución del Dashboard

4. Ahora bien, por parte del archivo se distribuirá la gráfica de manera precisa para cada una de las interacciones que se realizarán del modelado anómalo y que brinden las APIS que se entrenaron previamente.

```python
                # Asumiendo que 'algorit' tiene la columna de etiquetas verdaderas y probabilidades
                y_test = algorit['EtiquetaReal']  # Etiquetas verdaderas
                if modelo_seleccionado == "XGBoost":
                    y_pred_proba = algorit['ProbabilidadAnómalo']  # Probabilidades de XGBoost
                elif modelo_seleccionado == "Isolation Forest":
                    y_pred_proba = algorit['PuntuaciónAnómalo']  # Puntuaciones de Isolation Forest
                elif modelo_seleccionado == "Random Forest":
                    y_pred_proba = algorit['ProbabilidadAnómalo']  # Probabilidades de Random Forest

                # Calcular FPR, TPR y AUC para la Curva ROC
                fpr, tpr, thresholds = roc_curve(y_test, y_pred_proba)
                auc_score = roc_auc_score(y_test, y_pred_proba)

                # Graficar la Curva ROC
                plt.figure(figsize=(8, 6))
                plt.plot(fpr, tpr, label=f'AUC = {auc_score:.2f}')
                plt.plot([0, 1], [0, 1], 'k--')  # Línea de referencia (clasificador aleatorio)
                plt.xlabel('Tasa de Falsos Positivos')
                plt.ylabel('Tasa de Verdaderos Positivos')
                plt.title(f'Curva ROC - {modelo_seleccionado}')
                plt.legend(loc='best')
                st.pyplot(plt)

            else:
                st.error("No se ha definido el archivo CSV limpiado. Asegúrate de subir un archivo.")

# Limpiar el directorio temporal después de procesar
for file_name in extracted_files:
    os.remove(os.path.join(extract_dir, file_name))
os.rmdir(extract_dir)
```

Figura 46: Código del grafico ROC

5. Llamando el archivo en el navegador, se visualizan ambas gráficas, con el tráfico web anómalo y normal mediante cada API utilizada y brindando la siguiente ejecución:

Figura 47: Ejecución de Dashboard

El usuario tiene tres opciones de algoritmos de Machine Learning para realizar el análisis del archivo, al seleccionar el modelo se procede con la ejecución del código que al finalizar muestra la curva ROC, métricas, matriz de confusión, histograma y gráficas de dispersión.

Presentación de los resultados del XGBoost

En la Figura 48 se observa que el área bajo la curva (AUC) del algoritmo de XGBoost indica que tiene la capacidad alta de diferenciar entre las clases positivas y negativas. Al tener un AUC del 0.93, tiene un 93% de probabilidad de que haga una clasificación correcta de una instancia positiva ante una negativa.

La curva azul se encuentra por encima de la línea diagonal lo que nos demuestra que el algoritmo tiene capacidad predictiva. La gráfica también evidencia que XGBoost logra un equilibrio al inicio entre la sensibilidad y la especificidad, pero a medida que sube la tasa de falsos positivos la sensibilidad disminuye levemente.

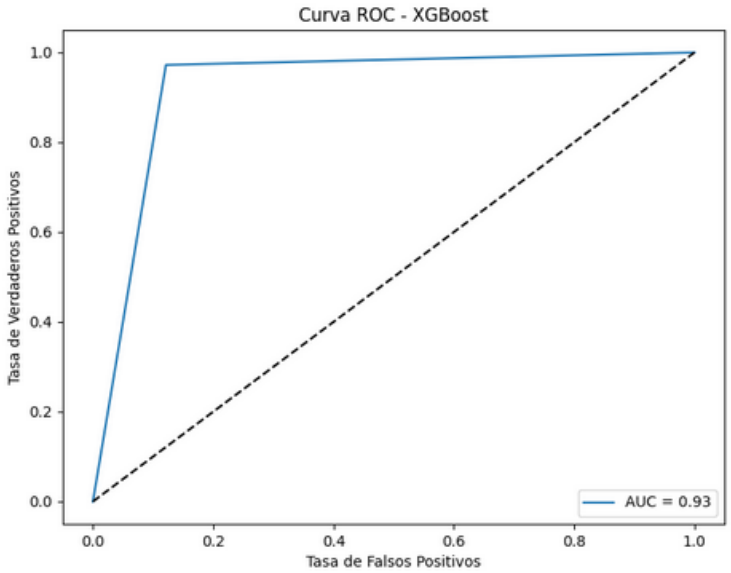


Figura 48: Curva ROC del XGBoost

También se muestra una tabla con los valores de las métricas de los resultados del algoritmo de Xgboost, en este caso a precisión que arroja es de 0.99 y una sensibilidad de 0.87 que son datos que permiten conocer la precisión que tiene el modelo con el archivo, se añade de igual forma una visualización de la cantidad de datos normales y anómalos encontró.

Métrica	Valor
Accuracy	0.8789
Precision	0.9964
Recall	0.8789
F1 Score	0.9322
N° Normal	397,752
N° Anomalo	56,552

Figura 49: Métricas del XGBoost

Además, se presenta la matriz de confusión del algoritmo XGBoost, mostrando los resultados del modelo de clasificación aplicado al tráfico de red, mostrando un rendimiento óptimo en la detección de casos normales (397.707), con una alta tasa

de falsos positivos (1579). También tiene una tasa baja de falsos negativos (45) en resultados anormales, lo que indica la precisión del identificador de eventos.

	Predicción 0 (Normal)	Predicción 1 (Anómalo)
Clase 0 (Normal)	397,707	1,579
Clase 1 (Anómalo)	45	54,973

Figura 50: Matriz de confusión de XGBoost

Se presenta un histograma que como se observa hay una notable diferencia entre los datos anómalos y normales. Las instancias etiquetadas como normales predominan de una forma uniforme en un rango largo del tiempo mientras que aquellos etiquetados como anómalos son menos frecuentes, lo cual al no tener uniformidad en ambas clases se deduce que no tiene patrones temporales evidentes en la aparición de eventos anómalos

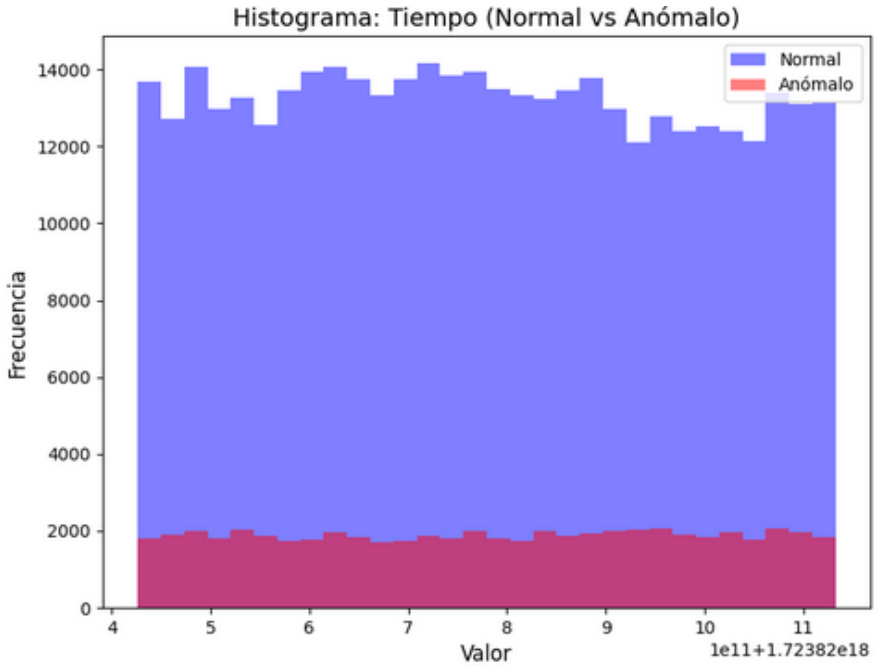


Figura 51: Histograma del XGBoost

Se visualiza una gráfica de dispersión, para los bytes recibidos, donde se muestra la relación entre el tiempo (hora) y los bytes recibidos en el tráfico web. La información se diferencia por clases, siendo normal o anómala, las cuales se representan por medio de diferentes colores para facilitar la interpretación. Los

puntos dispersos muestran que el tráfico normal se agrupa en ciertos rangos de bytes recibidos, mientras que el anómalo se dispersa más, lo que indica comportamientos inusuales en el volumen de los datos que se reciben.

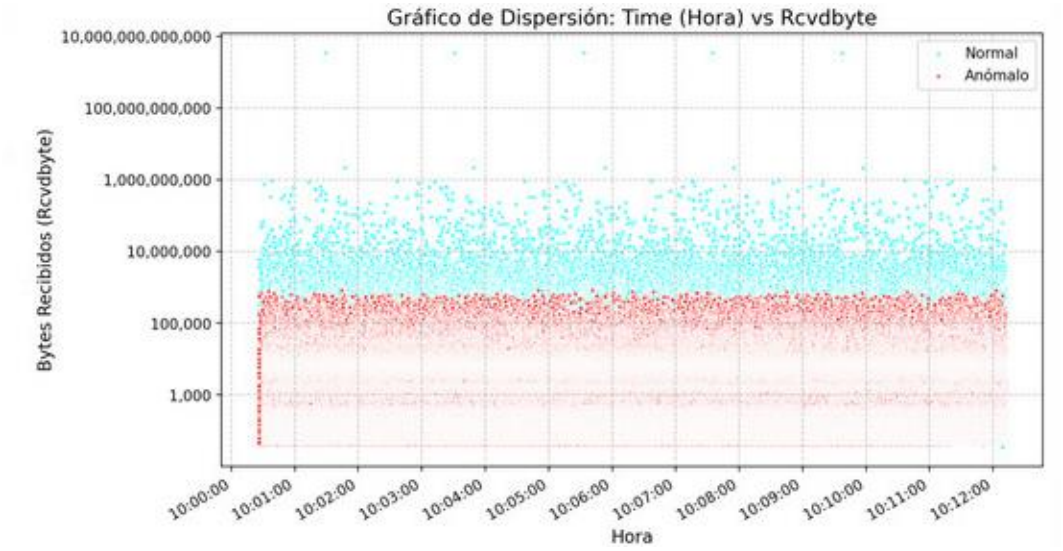


Figura 52: Gráfica de dispersión bytes recibidos

Finalmente, se presenta el diagrama de dispersión de bytes transmitidos, que muestra la correlación entre el tiempo (time) y los bytes transmitidos en el tráfico web. Los puntos se enfocan en las categorías "normal" y "anormal", lo que facilita el análisis del comportamiento del tráfico en función de los datos transmitidos.

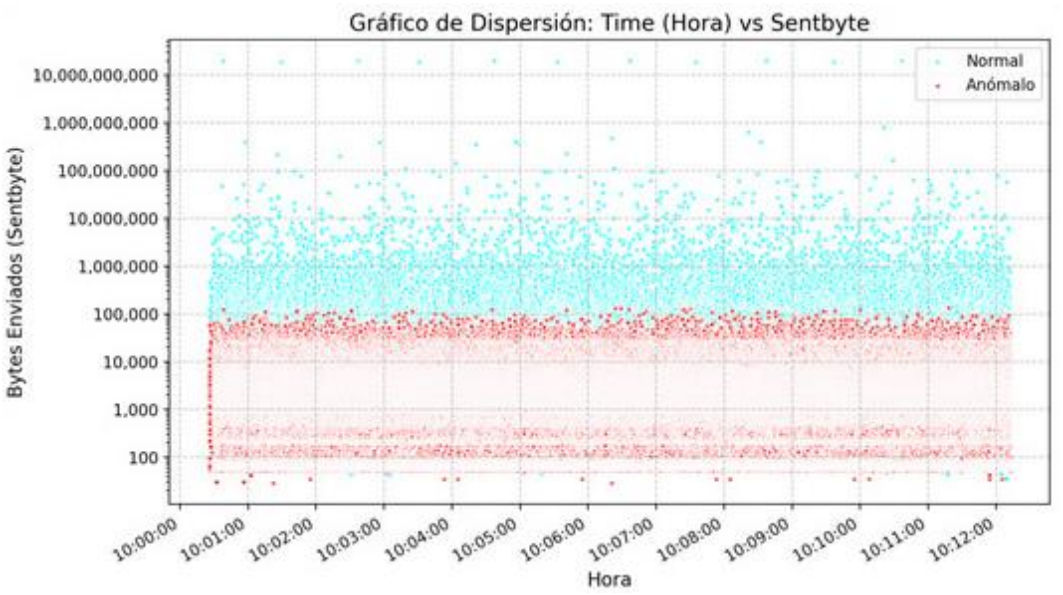


Figura 53: Gráfica de dispersión bytes enviados

Presentación de los resultados del Isolation Forest

Como se muestra en la Figura 53 el AUC es 0.52 lo que nos indica que el modelo no posee un desempeño mejor que el azar, ya que el AUC de 0.5 es el equivalente a una predicción aleatoria y la curva ROC del Isolation Forest por poco coincide con la línea diagonal de azar, por ende, no está diferenciando de forma correcta entre las clases positivas y negativas.

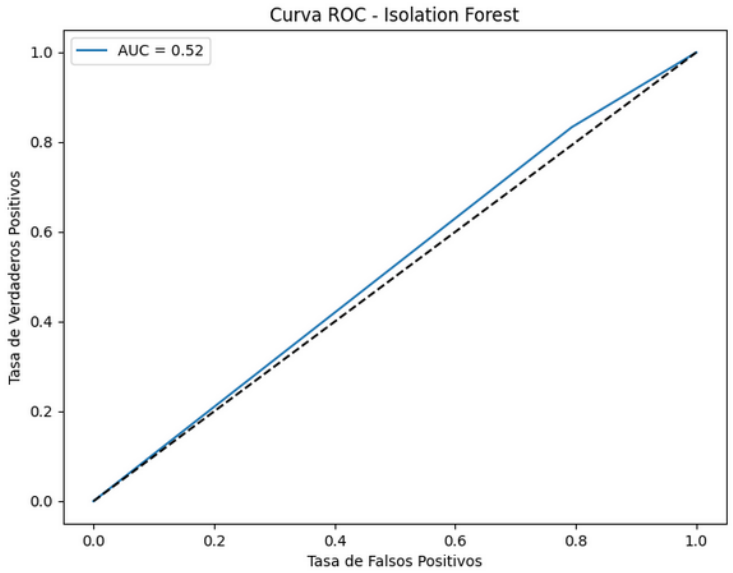


Figura 54: Curva ROC de Isolation Forest

Las métricas de los resultados del algoritmo de Isolation Forest, en este caso a precisión que arroja es de 0.54 y una sensibilidad de 0.46 por lo que está demostrando no tener una certeza de que lo que está definiendo como anómalo y normal sea realmente valores pertenecientes a esas etiquetas, en este caso detecto más datos anómalos con un total de 368102 que normales con un total de 86.202.

Métrica	Valor
Accuracy	0.4694
Precision	0.5478
Recall	0.4694
F1 Score	0.4191
N° Normal	86,202
N° Anomalo	368,102

Figura 55: Métricas del Isolation Forest

En la matriz de confusión del algoritmo, los resultados del modelo de clasificación aplicado al tráfico de red, mostrando un rendimiento un poco deficiente en la detección de casos normales (54471), con una alta tasa de falsos positivos (158764). También tiene una tasa baja de falsos negativos (31731) en resultados anormales, lo que indica la precisión del identificador de eventos.

	Predicción 0 (Normal)	Predicción 1 (Anómalo)
Clase 0 (Normal)	54,471	158,764
Clase 1 (Anómalo)	31,731	209,338

Figura 56: Matriz de confusión de Isolation Forest

El histograma de Isolation Forest muestra que los datos normales ahora se presentan con mayor frecuencia y están en una distribución uniforme por el rango del eje horizontal que en este caso está representado por el tiempo. Por otro lado, se tiene los datos anómalos los cuales se representan de forma uniforme de igual forma y es constante en el mismo rango del tiempo, cuyo comportamiento que no hay como tal un intervalo de tiempo específico donde se centre las anomalías al menos no directamente con el tiempo.

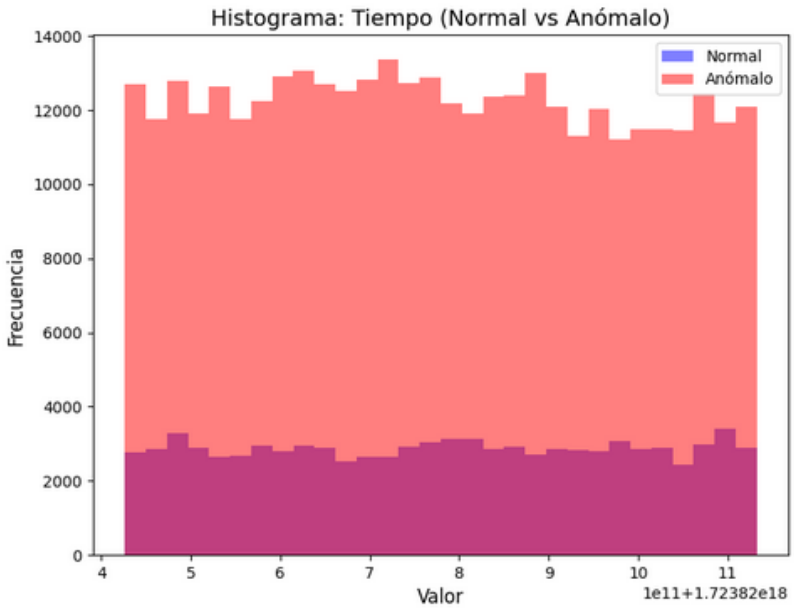


Figura 57: Histograma del Isolation Forest

Los datos mostrados en la gráfica de dispersión etiquetados como normales se encuentran distribuidos de forma amplia en niveles altos de bytes, mostrando una mayor densidad, mientras que los datos etiquetados como anómalos se concentran en valores más bajos, lo que genera una clara distinción entre ambos grupos. En el rango del tiempo, no se identifican patrones específicos que afecten la distribución de los datos normales, aunque las anomalías se mantienen dentro de un rango más limitado y en niveles de menor magnitud.

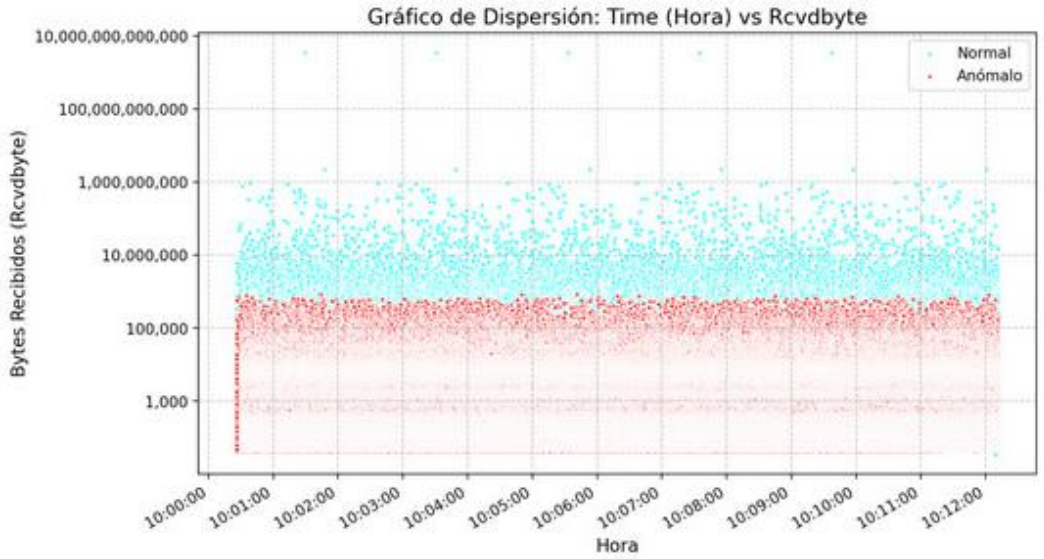


Figura 58: Gráfica de dispersión bytes recibidos

Los datos normales presentan una alta densidad y se distribuyen principalmente en niveles superiores de bytes enviados, mientras que los anómalos se concentran en valores más bajos, evidenciando una clara diferenciación entre ambas categorías. No se observan patrones temporales significativos que influyan en la distribución de los datos normales, mientras que las anomalías permanecen en un rango constante de menor magnitud. Este comportamiento podría estar relacionado con eventos que limitan el volumen de datos enviados, lo que sugiere la influencia de condiciones específicas en el sistema.

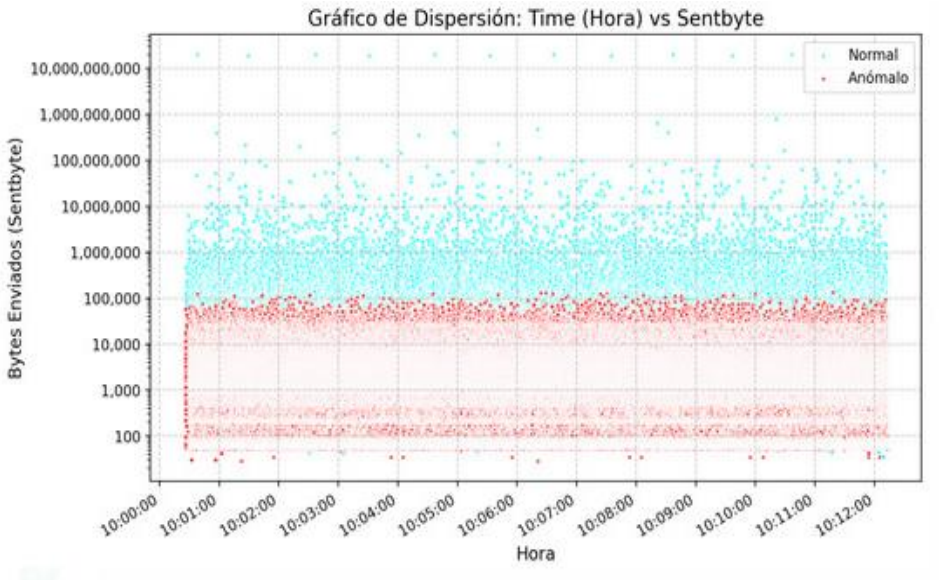


Figura 59: Gráfica de dispersión bytes enviados

Presentación de los resultados del Random Forest

Como se muestra en la Figura 59 el AUC es 0.53 lo que nos indica que el modelo al igual que el modelo de Isolation Forest no tiene un desempeño mejor que un modelo aleatorio, ya que el AUC de 0.5 es el equivalente a una predicción aleatoria y la curva ROC de Random Forest casi llega a coincidir con la línea diagonal punteada que es la representación de un modelo al azar, por lo que demuestra que no está detectando de forma correcta entre las clases positivas y negativas.

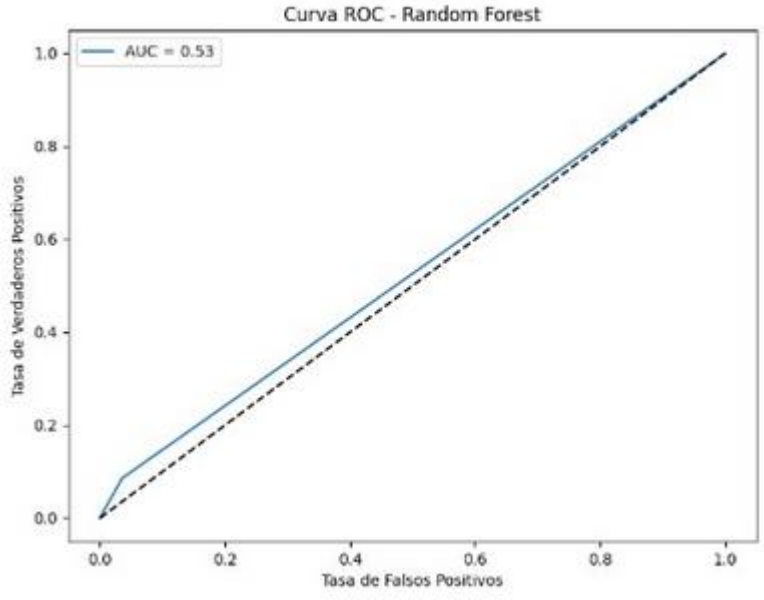


Figura 60: Curva ROC de Random Forest

Las métricas de los resultados del algoritmo de Isolation Forest, en este caso la precisión que arroja es de 0.87 y la sensibilidad de 0.90 por lo que está demostrando estar definiendo como anómalo y normal de forma adecuada en las etiquetas correspondientes, en este caso detecto más datos anómalos con un total de 17911 que normales con un total de 436.393.

Métrica	Valor
Accuracy	0.9025
Precision	0.8784
Recall	0.9025
F1 Score	0.8896
N° Normal	436,393
N° Anomalo	17,911

Figura 61: Métricas del Random Forest

En la matriz de confusión del algoritmo, los resultados del modelo de clasificación aplicado al tráfico de red, mostrando un rendimiento un poco deficiente en la

detección de casos normales (407243), con una alta tasa de falsos positivos (2747). También tiene una tasa baja de falsos negativos (29150) en resultados anormales, lo que indica la precisión del identificador de eventos.

	Predicción 0 (Normal)	Predicción 1 (Anómalo)
Clase 0 (Normal)	407,243	2,747
Clase 1 (Anómalo)	29,150	15,164

Figura 62: Matriz de confusión de Random Forest

En el histograma de Random Forest la categoría normal domina claramente en términos de frecuencia, con una distribución uniforme en todo el rango de valores en cambio los datos anómalos tienen una frecuencia considerablemente menor, manteniéndose constantes en niveles bajos a lo largo del mismo rango temporal. La uniformidad en la distribución de ambas categorías sugiere que no hay intervalos específicos donde las anomalías sean más frecuentes, lo que podría indicar que estos eventos ocurren de manera aleatoria o están asociados a factores no temporales.

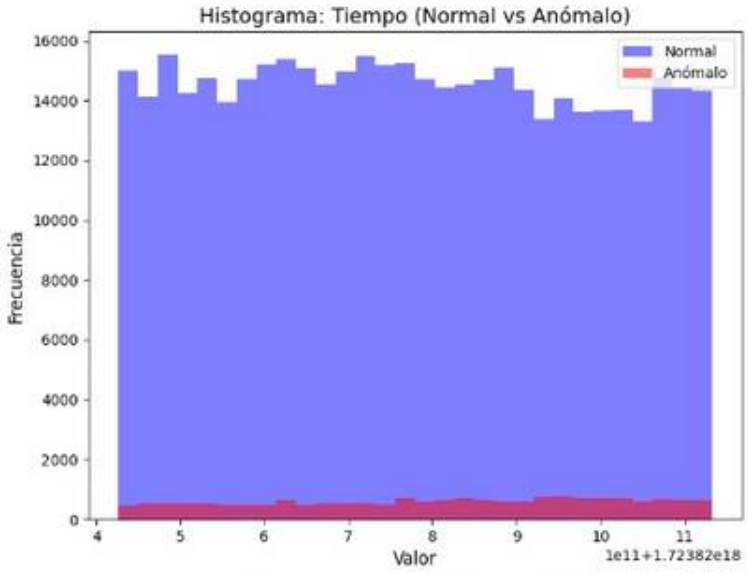


Figura 63: Histograma del Random Forest

Los datos normales se concentran en niveles altos de bytes recibidos, con una distribución uniforme a lo largo del tiempo, lo que indica una operación consistente en esta categoría. Por otro lado, los datos anómalos se distribuyen principalmente en niveles más bajos de bytes, también con una distribución homogénea en el tiempo, lo que refleja una separación clara entre ambas categorías. La escala

logarítmica del eje Y permite apreciar con mayor claridad las diferencias en magnitud entre los eventos normales y anómalos. La consistencia temporal sugiere que las anomalías no dependen de un momento específico, sino de condiciones particulares que limitan la cantidad de datos recibidos.

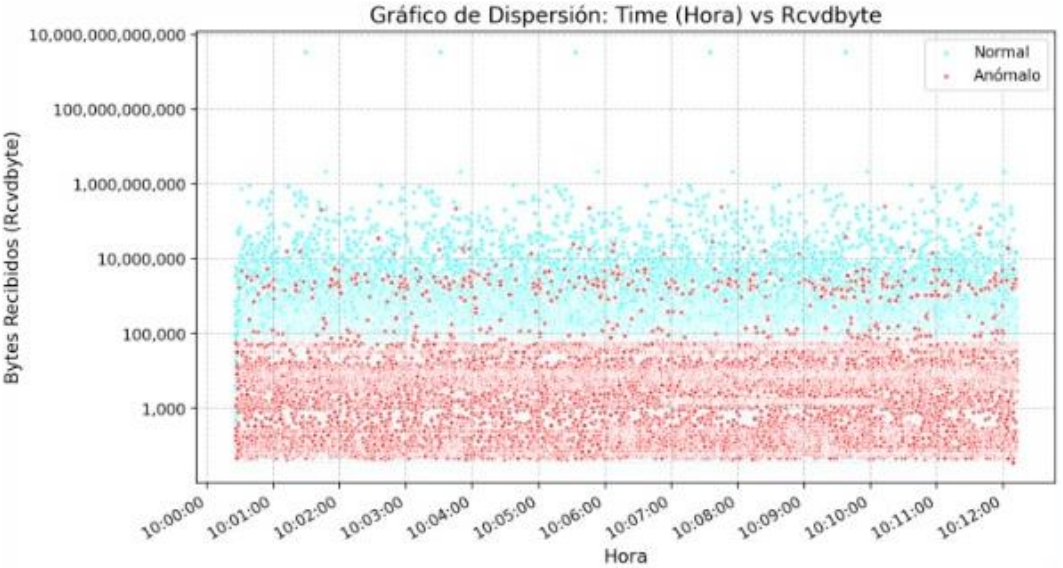


Figura 64: Gráfica de dispersión bytes recibidos

Los datos normales se agrupan principalmente en niveles altos de bytes enviados, manteniendo una distribución uniforme a lo largo del tiempo, lo que refleja un comportamiento constante en esta categoría. En contraste, los datos anómalos se concentran en niveles significativamente más bajos de bytes enviados, con una distribución igualmente uniforme en el tiempo. Esta separación evidente entre ambas categorías, destacada gracias a la escala logarítmica del eje Y, sugiere que las anomalías podrían estar asociadas con eventos que restringen la cantidad de datos transmitidos. No se observan patrones temporales específicos para ninguno de los conjuntos de datos, indicando que las anomalías no dependen de momentos particulares en el tiempo.

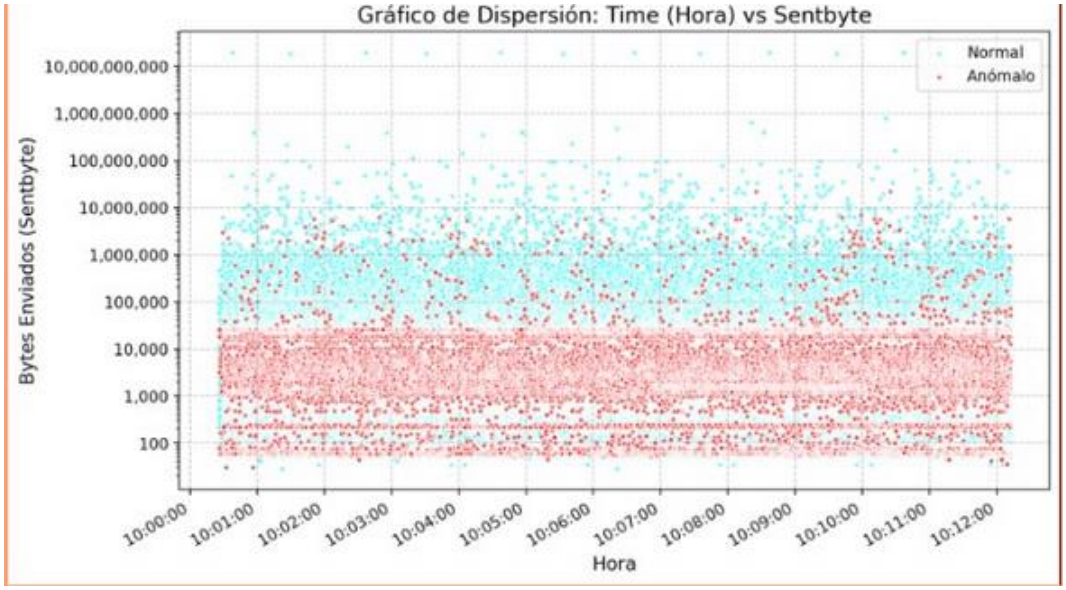


Figura 65: Gráfica de dispersión bytes enviados

CONCLUSIONES

- Se escogieron las herramientas adecuadas para la detección y análisis de las redes, siendo Python quien permitió entrenar los algoritmos de machine learning, facilitando el procesamiento de los datos; mientras que, las librerías como Request, ayudaron a realizar solicitudes HTTP y Matplotlib en la visualización de resultados. Las APIs como VirusTotal fueron claves para analizar y verificar las IPs y la base de datos MySQL se utilizó para asegurar una gestión adecuada en la información.

- La aplicación de algoritmos de aprendizaje automático para clasificar el tráfico web, permitió obtener mejoras en cuanto a la detección de anomalías en las redes; por parte del entrenamiento, algoritmos como XGBoost, Isolation Forest y Random Forest, realizaron un ajuste en los parámetros para hallar modelos normales y anómalos, lo que optimizó la eficacia en el análisis. Así mismo, se logró clasificar los datos eficientemente, generando indicadores de anomalías que facilitan la identificación precisa de irregularidades, mejorando la gestión de la red.

- El análisis del Dataset utilizando un Dashboard, simplificó la representación de los resultados a través de curvas ROC de los algoritmos XGBoost, Random Forest e Isolation Forest. XGBoost revela un rendimiento óptimo con un AUC de 0.93, frente a los demás algoritmos. De la misma forma, las matrices de confusión facilitaron la identificación de tasas de falsos positivos y negativos, ofreciendo una evaluación exacta de los algoritmos. Finalmente, los informes y gráficos presentan una comprensión del tráfico web, lo que facilita la toma de decisiones.

RECOMENDACIONES

- Es crucial elegir cuidadosamente herramientas que brinden una captura precisa del tráfico web sin perjudicar el desempeño del sistema, las cuales deben tener la habilidad de manejar gran cantidad de datos, reconocer los patrones de tráfico relevantes y proporcionar alternativas de análisis y filtrado. Por esto, es recomendable elegir soluciones completas que simplifiquen la organización y el acceso a la información, posibilitando un análisis más ágil.

- Se recomienda elegir algoritmos de machine learning apropiados para la categorización del tráfico web, como algoritmos supervisados, que requieren un conjunto de datos etiquetados para el entrenamiento de los modelos. Es esencial realizar una evaluación de diferentes algoritmos, modificando sus parámetros para obtener los resultados más idóneos con relación a la capacidad y precisión de generalización, garantizando la eficacia de cada uno al clasificar los diversos tipos de tráfico web.

- Es recomendable incluir mayor cantidad de datos en el dashboard interactivo, para una presentación comprensible de los resultados, siendo intuitivo y sencillo de utilizar, permitiendo que los usuarios exploren la información de forma dinámica, además de crear informes automáticos basados en los resultados obtenidos, lo que facilitará una correcta toma de decisiones y proporcionará información relevante oportunamente.

BIBLIOGRAFÍA

[1] CEPAL, "Tecnologías digitales para un nuevo futuro," 2022. Accessed: Sep. 09, 2024. [Online]. Available: https://repositorio.cepal.org/server/api/core/bitstreams/879779be-c0a0-4e11-8e08-cf80b41a4fd9/content

[2] Upcommons, "Redes de Acceso Alámbricas," 2022. Accessed: Sep. 11, 2024. [Online]. Available: https://upcommons.upc.edu/bitstream/handle/2117/94358/05AMCA05de15.pdf?sequence=5&isAllowed=y

[3] R. Duque, "UPSE." Accessed: Sep. 09, 2024. [Online]. Available: https://www.gbif.org/publisher/f99f2c09-1353-4f24-aab2-f4cbaef67eb1

[4] A. Fernández, "Uso de Machine-Learning en el control de congestión sobre redes 5G," UNIVERSIDAD DE CANTABRIA, 2020. Accessed: Sep. 09, 2024. [Online]. Available: https://repositorio.unican.es/xmlui/bitstream/handle/10902/19021/425971.pdf?sequence=1&isAllowed=y

[5] J. Sánchez and S. Romero, "Algoritmo de volcado del tráfico de datos para redes inalámbricas sobre una red definida por software," UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS, Bogotá, 2019.

[6] J. Vera, "ESTUDIO DE COBERTURA DE REDES INALÁMBRICAS CON FRECUENCIAS 2.4 Y 5.0 GHZ EN LAS CARRERAS DE INGENIERÍA EN SISTEMAS COMPUTACIONALES Y TECNOLOGÍA DE LA INFORMACIÓN DE LA UNESUM," UNESUM, Jipijapa, 2019. Accessed: Sep. 09, 2024. [Online]. Available: https://repositorio.unesum.edu.ec/bitstream/53000/1177/1/UNSUM-ECUADOR-SISTEMAS-2018-03.pdf

[7] A. Ramírez, "Modelo Predictivo del tráfico de Internet: caso puntos digitales gratuitos Zona 5," UPSE, Santa Elena, 2024. Accessed: Sep. 09, 2024. [Online]. Available:

https://repositorio.upse.edu.ec/bitstream/46000/11211/1/UPSE-MTI-2024-0004.pdf

[8] Python, “Python.” Accessed: Sep. 09, 2024. [Online]. Available: https://www.python.org/

[9] MySQL, “MySQL.” Accessed: Nov. 13, 2024. [Online]. Available: mysql.com

[10] Facsistel, “UPSE.” Accessed: Sep. 09, 2024. [Online]. Available: https://facsistel.upse.edu.ec/index.php?option=com_content&view=article&id=58&Item

[11] Incibe, “Protección en movilidad y conexiones inalámbricas,” 2022. Accessed: Sep. 09, 2024. [Online]. Available: https://www.incibe.es/sites/default/files/contenidos/dosieres/metad_proteccion_movilidad_y_conexiones_inalambricas.pdf

[12] Sotein, “Introducción a las redes alámbricas o cableadas.” Accessed: Sep. 11, 2024. [Online]. Available: https://sotein.com.co/redes-alambricas/#:~:text=Las%20redes%20al%C3%A1mbricas%20ofrecen%20mayor,ocurrir%20con%20las%20conexiones%20inal%C3%A1mbricas.

[13] UNLP, “Monitoreo de redes herramientas,” 2021. Accessed: Sep. 09, 2024. [Online]. Available: https://www.trabajosocial.unlp.edu.ar/uploads/docs/monitoreo_de_redes.pdf

[14] Ecuador, “Plan de Creación de Oportunidades 2021 - 2025.” Accessed: Sep. 09, 2024. [Online]. Available: https://www.planificacion.gob.ec/wp-content/uploads/2021/09/Plan-de-Creacio%CC%81n-de-Oportunidades-2021-2025-Aprobado.pdf

[15] E. Rus, “Investigación exploratoria.” Accessed: Sep. 10, 2024. [Online]. Available: https://economipedia.com/definiciones/investigacion-exploratoria.html

[16] J. Rodríguez, M. Villamizar, and S. Correa, “Aplicación de una metodología diagnóstica,” vol. 15, no. 5, 2020.

[17] IBM, “Conceptos básicos de ayuda de CRISP-DM.” Accessed: Sep. 10, 2024. [Online]. Available: https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

[18] V. Alvarez, “PROPUESTA DE UNA METODOLOGÍA DE PRUEBAS DE PENETRACIÓN ORIENTADA A RIESGOS,” *UEES*, 2018, Accessed: Sep. 10, 2024. [Online]. Available: http://repositorio.uees.edu.ec/bitstream/123456789/2525/1/ALVAREZ%20INTRIAGO%20VILMA%20KARINA.pdf

[19] OMSTD, “Metodología OMSTD.” Accessed: Sep. 10, 2024. [Online]. Available: https://omstd.readthedocs.io

[20] Gobierno de México, “Instituciones de educación superior.” Accessed: Sep. 11, 2024. [Online]. Available: https://www.gob.mx/sep/acciones-y-programas/instituciones-de-educacion-superior

[21] Euroinnova, “Institución de educación superior.” Accessed: Sep. 11, 2024. [Online]. Available: https://www.euroinnova.com/blog/institucion-de-educacion-superior#:~:text=De%20acuerdo%20con%20el%20sistema,se%20imparten%20las%20carreras%20profesionales.

[22] Eduteka, “REDES DE DATOS EN INSTITUCIONES.” Accessed: Sep. 15, 2024. [Online]. Available: https://eduteka.icesi.edu.co/articulos/RedEscolarDatos

[23] FS, “Solución Wifi para educación de FS: Wifi para escuelas, universidades y colegios.” Accessed: Sep. 15, 2024. [Online]. Available: https://community.fs.com/es/blog/fs-education-wifi-solution-wifi-for-schools-universities-colleges.html

[24] Telecomunicaciones, “LEY ORGÁNICA DE TELECOMUNICACIONES,” 2016. Accessed: Sep. 13, 2024. [Online].

Available: https://www.telecomunicaciones.gob.ec/wp-content/uploads/downloads/2016/05/Ley-Org%C3%A1nica-de-Telecomunicaciones.pdf

[25] ASAMBLEA NACIONAL, "LEY ORGÁNICA DE PROTECCIÓN DE DATOS PERSONALES," 2021. Accessed: Sep. 13, 2024. [Online]. Available: https://www.finanzaspopulares.gob.ec/wp-content/uploads/2021/07/ley_organica_de_proteccion_de_datos_personales.pdf

[26] Gobierno del Ecuador, "CÓDIGO ORGÁNICO INTEGRAL PENAL, COIP," Quito, 2021. Accessed: Nov. 22, 2024. [Online]. Available: https://www.defensa.gob.ec/wp-content/uploads/downloads/2021/03/COIP_act_feb-2021.pdf

[27] Universidad Internacional de Valencia, "¿Qué son las redes de datos?"

[28] Sotein, "Características de las redes alámbricas." Accessed: Sep. 15, 2024. [Online]. Available: https://sotein.com.co/redes-alambricas/#:~:text=Las%20redes%20al%C3%A1mbricas%20es%20un,oficina%20peque%C3%B1a%20o%20un%20hogar.

[29] Lifeder, "Redes alámbricas: características, tipos, ventajas y desventajas." Accessed: Sep. 15, 2024. [Online]. Available: https://www.lifeder.com/redes-alambricas/

[30] Termired, "Guía completa sobre Redes Inalámbricas: qué son, tipos y su importancia." Accessed: Sep. 15, 2024. [Online]. Available: https://termired.com/redes-inalambricas-que-es/

[31] Microsegur, "Características de las redes inalámbricas." Accessed: Sep. 15, 2024. [Online]. Available: https://microsegur.com/caracteristicas-de-las-redes-inalambricas/

[32] A. Herrero, "Tráfico web." Accessed: Sep. 15, 2024. [Online]. Available: https://neoattack.com/neowiki/trafico-web/

[33] A. Salinas, “Desarrollo de algoritmo para detección temprana de anomalías en tráfico URL, en redes distribuidas para la Facultad de Sistema y Telecomunicaciones (FACSISTEL) de la Universidad Estatal Península de Santa Elena,” UPSE, Santa Elena, 2023. Accessed: Sep. 15, 2024. [Online]. Available: https://repositorio.upse.edu.ec/bitstream/46000/10299/1/UPSE-TTI-2023-0047.pdf

[34] ManageEngine, “Detección de anomalías en el tráfico de red: Una técnica de monitoreo de tráfico a prueba de fallos.” Accessed: Sep. 15, 2024. [Online]. Available: https://www.manageengine.com/latam/netflow/deteccion-de-anomalias-de-trafico-de-red.html

[35] M. Boden, *Inteligencia Artificial*. 2017. Accessed: Sep. 24, 2023. [Online]. Available: https://books.google.es/books?hl=es&lr=&id=LCnYDwAAQBAJ&oi=fnd&pg=PT3&dq=inteligencia+artificial+&ots=drYqBZhKq6&sig=9A8A2LdV75hNuQr-6b6OIMwGsB4#v=onepage&q=inteligencia%20artificial&f=false

[36] L. Rouhiainen, *Inteligencia artificial 101 cosas que debes saber hoy sobre nuestro futuro inteligencia artificial*, 1st ed. 2018. Accessed: Sep. 24, 2023. [Online]. Available: https://planetadelibrosec0.cdnstatics.com/libros_contenido_extra/40/39308_Inteligencia_artificial.pdf

[37] D. Quirumbay, C. Castillo, and I. Coronel, “Una revisión del Aprendizaje profundo aplicado a la ciberseguridad,” *Revista Científica y Tecnológica UPSE*, vol. 9, no. 1, 2022, doi: https://doi.org/10.26423/rctu.v9i1.671.

[38] B. Mahesh, “Machine Learning Algorithms-A Review,” *International Journal of Science and Research*, pp. 1–7, Jan. 2019, doi: 10.21275/ART20203995.

[39] M. J. Ramos Salinas, F. M. Villegas Pancca, B. B. Cordova Chipa, S. P. Cano Quito, and P. M. Lezama Gonzales, “Análisis de patrones de morbilidad por anemia mediante algoritmos no supervisados: un enfoque basado en datos de

establecimientos de salud a nivel nacional," *Revista de investigación de Sistemas e Informática*, vol. 16, no. 2, pp. 15–24, Dec. 2023, doi: 10.15381/risi.v16i2.25776.

[40] V. Pedrero, K. Reynaldos Grandón, J. Ureta Achurra, and E. Cortez Pinto, "Generalidades del Machine Learning y su aplicación en la gestión sanitaria en Servicios de Urgencia," *Rev Med Chil*, vol. 149, no. 2, pp. 248–254, Feb. 2021, doi: 10.4067/S0034-98872021000200248.

[41] IBM, "¿Qué es el aprendizaje supervisado?" Accessed: Sep. 15, 2024. [Online]. Available: https://www.ibm.com/es-es/topics/supervised-learning

[42] AWS, "¿Cuál es la diferencia entre el aprendizaje supervisado y el no supervisado?" Accessed: Sep. 15, 2024. [Online]. Available: https://aws.amazon.com/es/compare/the-difference-between-machine-learning-supervised-and-unsupervised/

[43] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," pp. 1–11, Apr. 2021, doi: 10.1007/s12525-021-00475-2/Published.

[44] J. Bahzad and A. M. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.

[45] IBM, "¿Qué es el random forest?" Accessed: Sep. 15, 2024. [Online]. Available: https://www.ibm.com/mx-es/topics/random-forest

[46] ArcGIS Pro, "Cómo funciona el algoritmo XGBoost." Accessed: Sep. 15, 2024. [Online]. Available: https://pro.arcgis.com/es/pro-app/latest/tool-reference/geoai/how-xgboost-works.htm

[47] J. Amat, "Detección de anomalías: Isolation Forest." Accessed: Sep. 15, 2024. [Online]. Available: https://cienciadedatos.net/documentos/66_deteccion_anomalias_isolationforest

[48] Y. Fernández, “Qué son los TOPS, qué miden y para qué se usa esta unidad de medida en la inteligencia artificial.”

[49] Intel, “CPU o GPU: opciones interesantes para sus necesidades informáticas.” Accessed: Nov. 22, 2024. [Online]. Available: https://www.intel.la/content/www/xl/es/products/docs/processors/cpu-vs-gpu.html

[50] J. Alonso, “El sitio web como unidad básica de información y comunicación. Aproximación teórica: definición y elementos constitutivos,” pp. 1–23, 2008, Accessed: Sep. 24, 2023. [Online]. Available: https://idus.us.es/bitstream/handle/11441/33488/El%20sitio%20web%20como%20unidad%20basica%20de%20informacion%20y%20comunicacion.pdf?sequence=1&isAllowed=y

[51] J. A. Orjuela, A. J. Osorio, and Y. Patiño, “Importancia del balanced scorecard (bsc) para medir el desempeño estratégico de las empresas”, Accessed: Sep. 24, 2023. [Online]. Available: https://digitk.areandina.edu.co/bitstream/handle/areandina/5092/Trabajo%20de%20grado.pdf?sequence=1&isAllowed=y

[52] Y. Gusnadi and A. Hermawan, “Designing Employee Performance Monitoring Dashboard Using Key Performance Indicator (KPI),” vol. 2, pp. 1–8, Jan. 2020, Accessed: Oct. 01, 2023. [Online]. Available: https://jurnal.kdi.or.id/index.php/bt/article/view/107/63

[53] O. Ríos Jacobo, “Entendiendo a los Indicadores Clave de Desempeño (KPI),” in *Key Performance Indicators (KPI)*, 2019, pp. 23–66. Accessed: Sep. 29, 2023. [Online]. Available: https://gc.scalahed.com/recursos/files/r161r/w24174w/S8_desarrollo_aplicacion_gestion.pdf

[54] Python Software Foundation, “El tutorial de Python .” Accessed: Jun. 02, 2023. [Online]. Available: https://docs.python.org/es/3/tutorial/

[55] C. Rakesh and V. Bala, *Python Requests Essentials*. Packt Publishing Ltd, 2015.

[56] Aprendeconalf, “La librería Matplotlib.” Accessed: Nov. 22, 2024. [Online]. Available: https://aprendeconalf.es/docencia/python/manual/matplotlib/

[57] Á. Gómez, J. Regalado, J. Gutiérrez, O. Quimis, K. Marcillo, and J. MArcillo, *FUNDAMENTOS SOBRE LA GESTIÓN DE BASE DE DATOS*. 2017. Accessed: Sep. 24, 2023. [Online]. Available: https://books.google.es/books?hl=es&lr=lang_es&id=H0VBDwAAQBAJ&oi=fnd&pg=PA7&dq=base+de+datos&ots=fXm-YIwC0z&sig=n7RD8W0MybwFh1eLLUFiuyxEgsM#v=onepage&q=base%20de%20datos&f=false

[58] “OMSTD.” Accessed: Sep. 24, 2023. [Online]. Available: https://omstd.readthedocs.io/start.html

[59] Pymesec, “ISSAF.” Accessed: Sep. 23, 2024. [Online]. Available: https://pymesec.org/issaf/

[60] S. Pérez, H. Facchini, and J. Tramontina, “Análisis de Tiempo Real de Tráfico de Redes de Datos mediante Técnicas de Inteligencia Artificial”, Accessed: Sep. 24, 2023. [Online]. Available: https://www.researchgate.net/publication/362830143

[61] J. Márquez Díaz, “Inteligencia artificial y Big Data como soluciones frente a la COVID-19,” vol. 30, no. 8–9, pp. 448–455, Nov. 2020, doi: 10.1016/J.PUROL.2020.04.015.

[62] J. García, J. Molina, A. Berlanga, M. Patricio, Á. Bustamante, and W. Padilla, *Ciencia de datos. Técnicas analíticas y aprendizaje estadístico*. 2018. Accessed: Sep. 30, 2023. [Online]. Available: https://d1wqtxts1xzle7.cloudfront.net/64031156/Ciencia_de_datos_2018-libre.pdf?1595869926=&response-content-disposition=inline%3B+filename%3DCiencia_de_datos_Tecnicas_analiticas_y_a.pdf&Expires=1696214108&Signature=XkkKmgdukVu0fq1Xyti5~l0Y4yL8A5JereCGCL~1sjESl5ERBwrAaz08VDapHxcsVD2WR5mNDEp-SCiO1n0K10glwJurHpeP2b-B8vjDXgaVN~C-~QqmEgPjZmnYaBOCM5lLqfFa~mSJmjy6i8FzVvmk4e5asbaHVSZw9pa

JwjcP~6QVVS0Cf4wgTZo8OWaz2OSxRwcBoflwnx8kb3~8cC8w3GcKcue57hajQ99gS9FCMmTP-G~baa3DVMiAg-k8skDmf~TLelBHpAOx~mlhfxTb~TVBS4L15kqGQ4ANhN2VgkW8UCRwSpS8f-fGbCjGTzhpQfrNBm5G4rc7stxSUw__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

[63] A. León, J. Martinez, I. Ardila, and D. Mosquera, "Inteligencia artificial para el control de tráfico en redes de datos: Una Revisión," *Entre ciencia e Ingeniería* , vol. 16, pp. 1–8, May 2022, Accessed: Sep. 24, 2023. [Online]. Available: http://www.scielo.org.co/pdf/ecei/v16n31/1909-8367-ecei-16-31-17.pdf

[64] V. Mirjalili and S. Raschka, "Python Machine Learning." Accessed: Oct. 02, 2023. [Online]. Available: https://books.google.es/books?hl=es&lr=&id=5EtOEAAAQBAJ&oi=fnd&pg=PT5&dq=librer%C3%ADas+de+python&ots=eqN2RBZHHa&sig=eiTu2hhPA9Qy9F0elMqh6LINuSY#v=onepage&q=librer%C3%ADas%20de%20python&f=false

[65] J. Zárate Valderrama, N. Bedregal Alpaca, and V. Cornejo Aparicio, "Modelos de clasificación para reconocer patrones de deserción en estudiantes universitarios," *Ingeniare. Revista chilena de ingeniería*, vol. 29, no. 1, pp. 168–177, Mar. 2021, doi: 10.4067/S0718-33052021000100168.

[66] J. Ponce, A. Torres, F. Quezada, and A. Silva, "Inteligencia Artificial," pp. 1–223, Mar. 2014, Accessed: Sep. 24, 2023. [Online]. Available: https://www.researchgate.net/publication/269466259_Inteligencia_Artificial

[67] A. Urbina, A. Téllez, and R. Cruz, "Patrones que identifican a estudiantes universitarios desertores aplicando minería de datos educativa," vol. 23, pp. 1–15, 2021, doi: 10.24320/redie.2021.23.e29.3918.

[68] D. Quirumbay, B. Soria, and V. Cruz, "Efficient clustering of e-mails by applying supervised machine learning algorithms," *Journal of Applied Research and Technology* , vol. 22, no. 1, 2024, Accessed: Dec. 01, 2024.

[Online]. Available: https://jart.icat.unam.mx/index.php/jart/article/view/2383/1130

ANEXOS

Anexo 1: Ficha de observación para identificar la accesibilidad que hay en la navegación web de las redes dentro de la universidad

FICHA DE OBSERVACIÓN			
Objetivo	Identificar la accesibilidad que hay en la navegación web de las redes dentro de la universidad.		
INDICADORES	ESCALA		
	Nunca	Algunas veces	Siempre
En el sector suele haber una o más redes inalámbricas			
Se encuentran redes Wifi-privadas en el sector			
Se logra identificar el roaming del sector			
Al estar conectado en una de las redes Wifi de la Universidad logra tener estabilidad en la señal de la red			
Se detectan interferencias o saturación de las redes en el sector			
Presenta montaje o solapamiento de varios canales de red en el sector			
Se encuentra un balanceo de carga en la conectividad de la red			
Existen restricciones en la navegación de sitios web empleando las redes de la U			
Permite tener acceso remoto utilizando las redes de la U			
La conexión alámbrica presenta estabilidad en todo momento			

Se detectan caídas intermitentes del servicio con respecto a la red alámbrica			
La latencia en las redes alámbricas es alta al momento de transferir grandes volúmenes de datos			

Anexo 2: Entrevista dirigida al director del departamento de TI, Ing. Fabricio Ramos

1. **En el departamento de TI ¿Con qué cantidad de páginas maliciosas lidian diariamente?**

En lo general, diariamente por la cantidad de alumnos que se conectan a nuestra red, ya sea de manera alámbrica o inalámbrica, se lidia con una cantidad muy desfavorable de conexiones a páginas no previstas al ámbito educacional. Decirle de forma concreta una cantidad específica de sitios maliciosos en los que se quieren conectar, no tendría un número en particular, pero si superan más de mil páginas con intentos no favorables para los usuarios.

2. **¿Tienen algún método para mitigar estos tipos de entradas a páginas webs por parte del alumnado?**

Tenemos métodos de seguridad que nos permiten identificar este tipo de anomalías en los usuarios, aunque no nos brindan un cien por ciento de lo que se trata de identificar todos estos datos; siempre se encuentra mucha variedad nueva que no se halla en la base de datos en particular, ya que, mantenemos limpiezas de redes continuas o páginas con contenido no tan particular como para ser detectada a primera orden y poder impedir el acceso como nuevas tecnologías.

3. **¿Poseen algún tipo de protección sobre páginas no registradas que contengan virus?**

Si se cuenta con un sistema de antivirus para identificar páginas maliciosas, pero de manera local e ingresar con una licencia para educación que nos permite ver que páginas son legibles y de esta manera poder abastecer la base de datos de identificación este proceso se maneja de manera manual ya que, tenemos un encargado que se encarga de esto y que muchas veces no cumple con el llenado completo.

4. **Cuándo ocurre alguna infección no prevista, ¿Cuánto tiempo afecta al Internet?**

El tiempo de respuesta que manejamos en un incidente de tráfico y bloqueo de red por alguna página maliciosa o con virus, es casi inmediato dependiendo mucho de

la dificultad de este tráfico, que puede ser por descargas masivas de un virus. Esto proviene de páginas con contenido prohibido; en estos casos, se cortan las comunicaciones de le red y se identifica el sector de donde se originó este incidente; en caso de inconvenientes externos, contamos con otro servicio de red que se establece en la misma conexión para mantener la universidad con conexión estable hasta poder conocer la anomalía y restablecer los servicios de la red primaria.

5. **¿Se cuenta con un sistema de rastreo de estas páginas maliciosas?**

Si se cuenta con un sistema de rastreo, pero como anteriormente se mencionó, existe una base de datos manual y un poco obsoleta, también contamos con otros sistemas creados por estudiantes que nos han sido de utilidad, pero no completamente satisfactorios; su índice de regularidad se deteriora con el tiempo al igual que su respuesta de datos, aunque ya mantiene una base de datos actual decae en rendimiento absoluto para un tráfico de red estable para toda la universidad.