

**UNIVERSIDAD AUTÓNOMA DE MADRID  
ESCUELA POLITÉCNICA SUPERIOR**



**Grado en Grado en Ingeniería Informática**

**TRABAJO FIN DE GRADO**

**Implementación Eficiente de Clasificadores  
Sencillos para Paquetes de Red**

**Autor: Alberto Cañas Gutiérrez de Cabiedes  
Tutor: Daniel Perdices Burrero**

**agosto 2025**

**Todos los derechos reservados.**

Queda prohibida, salvo excepción prevista en la Ley, cualquier forma de reproducción, distribución comunicación pública y transformación de esta obra sin contar con la autorización de los titulares de la propiedad intelectual.

La infracción de los derechos mencionados puede ser constitutiva de delito contra la propiedad intelectual (*arts. 270 y sgts. del Código Penal*).

**DERECHOS RESERVADOS**

© 2025 por UNIVERSIDAD AUTÓNOMA DE MADRID

Francisco Tomás y Valiente, n<sup>o</sup> 1

Madrid, 28049

Spain

**Alberto Cañas Gutiérrez de Cabiedes**

**Implementación Eficiente de Clasificadores Sencillos para Paquetes de Red**

**Alberto Cañas Gutiérrez de Cabiedes**

IMPRESO EN ESPAÑA – PRINTED IN SPAIN

*A mis padres, a mis hermanos, a mis amigos y a nuestra Señora la Virgen María madre de Las Tablas  
y Schoenstatt.*

*“No puedes defender. No puedes prevenir. Lo único que puedes hacer es detectar y responder” –  
Bruce Schneier*



# AGRADECIMIENTOS

---

En primer lugar, me gustaría agradecer a mis padres su apoyo y cariño a lo largo de toda mi etapa de estudiante así como por financiar mis estudios universitarios. También me gustaría agradecer a mis hermanos su cariño y apoyo. Sin el apoyo de mi familia no habría podido llegar tan lejos.

Quiero agradecer a mis compañeros Aitana Ayala, Diego Grande, Laura López, Cecilia Jiménez, Antonio Quintana, Jesús Quintana y en especial a Luis Arranz quien ha sido mi compañero en la mayoría de las asignaturas de prácticas, su apoyo, su cariño y los ratos de estudio juntos en la biblioteca de la EPS. Sin ellos mi estancia en la universidad habría sido mucho más difícil. También quiero agradecer a mis amigos que no forman parte de la universidad por su apoyo y su tiempo dedicado a mí durante todos estos años.

Por último me gustaría agradecer a todos mis profesores de la Escuela Politécnica Superior, de quienes he aprendido una gran cantidad de conocimientos que me han nutrido como persona y como profesional. Me gustaría agradecer especialmente a mis profesores de Programación I, Programación II, Sistemas Operativos, Inteligencia Artificial, Redes I, Redes II y Ciberseguridad por transmitirme los conocimientos que he necesitado para poder plantear, emprender y ejecutar este Trabajo de Fin de Grado. Me gustaría agradecer también de manera especial a mi tutor de TFG Daniel Perdices Burrero por el interés dedicado a este trabajo y por su guía sin la cual no habría podido desarrollarlo.



# RESUMEN

---

Por Hacer

# PALABRAS CLAVE

---

Por, Hacer





# ABSTRACT

---

TO DO

# KEYWORDS

---

TO, DO



# ÍNDICE

---

<b>1</b>	<b>Introducción</b>	<b>1</b>
1.1	Motivación .....	1
1.2	Objetivos .....	2
1.3	Estructura del documento .....	2
<b>2</b>	<b>Estado del Arte</b>	<b>5</b>
2.1	Conjunto de Datos .....	5
2.2	Métodos de Filtrado de Paquetes .....	6
2.3	Artículos Similares y Relacionados .....	8
<b>3</b>	<b>Implementación</b>	<b>9</b>
3.1	Arquitectura de la propuesta .....	9
3.2	Componentes .....	9
	<b>Bibliografía</b>	<b>13</b>



# LISTAS

---

## Lista de algoritmos

## Lista de códigos

3.1	PCAP a Dict .....	11
3.2	Entrenamiento Árbol .....	12

## Lista de cuadros

## Lista de ecuaciones

## Lista de figuras

2.1	Flujo eBPF .....	8
-----	------------------	---

## Lista de tablas

## Lista de cuadros



# INTRODUCCIÓN

---

## 1.1. Motivación

El ransomware es una forma específica de malware cuyo objetivo principal es bloquear el acceso a los datos de una organización o individuo, generalmente mediante técnicas de cifrado. Una vez que los archivos han sido cifrados, el atacante demanda el pago de un rescate (habitualmente en criptomonedas) a cambio de proporcionar la clave necesaria para restaurar la información secuestrada [?].

Esta clase de ataques se ha convertido en una amenaza persistente y cada vez más frecuente en el panorama actual de la ciberseguridad. Según el informe anual de la empresa especializada en ciberseguridad Cyberint [?], el pasado año 2024 se reportaron, a nivel global, 5.414 ataques de ransomware alrededor del mundo lo que supone un incremento del 11 % respecto al año 2023. Es también destacable el hecho de que un tercio del total de los ataques se realizaron en el último trimestre del año, lo que puede indicar una tendencia de aumento de ataques de esta clase en el presente año 2025.

Dado que este tipo de ataques no solo persisten, sino que van en aumento, resulta imprescindible desarrollar mecanismos capaces de detectarlos y prevenirlos sin que ello afecte negativamente al rendimiento del sistema ni se convierta en un cuello de botella. Para poder desplegar un sistema de detección y respuesta ante ransomware sin afectar al rendimiento del sistema, es fundamental analizar dónde implementar dicho sistema, tanto a nivel lógico como físico.

Un aspecto fundamental a considerar es en qué espacio del sistema operativo debe desplegarse la medida de seguridad. En el artículo de Parola et al.(2023) [?], se compara el rendimiento de la gestión de paquetes en tres escenarios: desde el espacio de usuario mediante el flujo tradicional, desde el espacio de usuario utilizando la herramienta DPDK —que opera mediante técnicas de polling para evitar la intervención del kernel— y, finalmente, desde el propio espacio del kernel.

Aunque DPDK presenta la menor latencia, el estudio concluye que, en términos de equilibrio entre rendimiento y consumo de recursos, el enfoque más eficiente es el que se implementa directamente

en el espacio del kernel. Este equilibrio sugiere que podría resultar especialmente interesante explorar el desarrollo de mecanismos de detección de ransomware a nivel de kernel, ya que permitiría una supervisión más cercana al sistema sin comprometer significativamente el rendimiento.

Otro aspecto relevante a considerar es el nodo físico de la red en el que debe implementarse el sistema de detección. La práctica más común consiste en desplegar el software de detección directamente en el servidor que se desea proteger —por ejemplo, el servidor A—. Sin embargo, teniendo en cuenta el crecimiento sostenido de las redes, con una tasa de crecimiento compuesta del 24 % según un informe de Cisco (2021) [?], este enfoque puede resultar cada vez menos eficiente.

El aumento del tráfico puede derivar en una mayor exposición a ataques, lo que incrementa la carga computacional que el servidor A debe asumir para protegerse a sí mismo. Por ello, puede resultar conveniente delegar parte de esta responsabilidad a otros componentes de la infraestructura de red, como las Tarjetas de Interfaz de Red (NIC, por sus siglas en inglés). Dentro de esta categoría, destacan especialmente las SmartNICs, que incorporan unidades de procesamiento dedicadas (DPUs) capaces de ejecutar tareas de seguridad de forma autónoma, descargando así al servidor principal.

## 1.2. Objetivos

El objetivo principal de este trabajo es evaluar el impacto que la aplicación de un sistema de filtrado de tráfico basado en árboles de decisión puede tener sobre el rendimiento de red. Dicho sistema está diseñado para identificar patrones de comunicación asociados a posibles ataques de tipo ransomware.

Para llevar a cabo esta evaluación, se implementará el árbol de decisión en lenguaje C utilizando tecnologías eBPF/XDP, lo que permitirá su integración directa en el plano de datos de red, reduciendo la latencia y mejorando la eficiencia del procesamiento.

Como objetivos secundarios, se analizará la viabilidad de desplegar este mecanismo de filtrado en dispositivos distintos a los servidores convencionales, como las SmartNICs.

## 1.3. Estructura del documento

El presente documento se divide en cinco capítulos. El primer capítulo, explica la importancia de investigar acerca de métodos de filtrado de ransomware así como su impacto en el rendimiento de la red. El segundo capítulo se adentra en el conjunto de datos usado para entrenar el árbol de decisión así como en el estado del arte de las técnicas, herramientas y artículos que aportan soluciones similares o perspectivas parecidas a otros problemas. El tercer capítulo desarrolla la metodología usada en el Trabajo y en él se explica la implementación que he realizado usando eBPF/XDP, scikit-learn, Mininet y Jinja. El cuarto capítulo expone los experimentos y métricas realizadas para el posterior análisis. Por



último el quinto capítulo presenta las conclusiones y comentarios.



# ESTADO DEL ARTE

---

## 2.1. Conjunto de Datos

En este apartado se describe el conjunto de datos empleado para el entrenamiento del modelo de árbol de decisión, diseñado con el objetivo de detectar posibles comunicaciones asociadas a ransomware.

El conjunto de datos utilizado proviene del trabajo realizado por Eduardo Berrueta et al. entre los años 2015 y 2022 [?]. Este dataset contiene capturas de tráfico en formato .pcap correspondientes a setenta familias distintas de ransomware, con un volumen total superior a 60 GB con los datos comprimidos.

Debido a las limitaciones de almacenamiento y capacidad de cómputo disponibles, así como al alcance definido para este trabajo, he optado por trabajar únicamente con dos de las muestras incluidas en el conjunto de datos original, que serán los que trataremos en profundidad en este apartado.

Pero antes de tratar estos dos subconjuntos de datos en concreto, cabe hacer una mención al artículo de Eduardo Berrueta y su equipo [?]. En él explican que la motivación para crear el dataset proviene de la necesidad de estandarizar recursos a la hora de evaluar la fiabilidad de las herramientas de detección de ransomware, ya que habitualmente cada fabricante o investigador utiliza su propio conjunto de datos. Al ser estos conjuntos muy específicos y reducidos, las comparativas resultantes tienden a ser poco realistas. Con su trabajo, Berrueta pretende aportar un dataset completo para pruebas de detección de ransomware que contribuya a resolver estas problemáticas.

Para este trabajo he usado muestras de dos ransomwares tomadas en el año 2021. El primero de los casos analizados es Hive, un conocido Ransomware-as-a-Service (RaaS) que emergió en junio de 2021. Este ransomware se caracterizaba por emplear un esquema de cifrado híbrido basado en los algoritmos RSA y AES, y por implementar una estrategia de doble extorsión. En este modelo, no solo se cifraban los datos de la víctima, sino que también se exfiltraban, exigiendo posteriormente un pago tanto por la clave de descifrado como por evitar la publicación de la información sustraída [?]. El segundo de los casos analizados es CryLock, una variante evolucionada del ransomware Cryakl.

CryLock se propagaba principalmente a través de correos electrónicos de phishing o campañas de spam, y al igual que otros ransomware modernos, empleaba un esquema de cifrado híbrido basado en los algoritmos RSA y AES. Además, una de sus características distintivas era la eliminación de las copias de seguridad presentes en los dispositivos comprometidos, con el objetivo de dificultar la recuperación de los datos sin realizar el pago del rescate. [?]

## 2.2. Métodos de Filtrado de Paquetes

El filtrado de paquetes constituye una de las técnicas fundamentales para el control del tráfico en redes informáticas. Su propósito principal es inspeccionar, permitir o bloquear paquetes de datos en función de criterios previamente definidos, como direcciones IP, puertos, protocolos o patrones de comportamiento.

A lo largo de la evolución de las redes de comunicaciones, el problema del filtrado de paquetes ha sido abordado mediante una amplia variedad de enfoques, que van desde soluciones sencillas y de rápida implementación hasta mecanismos con un alto grado de complejidad técnica. El mecanismo más sencillo de implementar para el filtrado de paquetes es el uso de un firewall. En los sistemas operativos GNU/Linux, este puede configurarse mediante el comando iptables, que opera a través de Netfilter, el subsistema encargado del filtrado de paquetes a nivel de kernel. El hecho de que el firewall funcione en espacio de kernel permite que su impacto sobre el rendimiento del sistema sea reducido. Su función principal consiste en controlar el tratamiento de los paquetes de red entrantes, salientes y reenviados, permitiendo aceptarlos, denegarlos, redirigirlos o registrarlos en función de criterios como dirección IP, puertos, protocolos, entre otros. [?]

Otro enfoque relevante para el filtrado de paquetes, especialmente orientado a la seguridad de los sistemas, es el uso de sistemas de detección de intrusiones (Intrusion Detection Systems, IDS por sus siglas en inglés). Uno de los más reconocidos en este ámbito es Snort, ampliamente utilizado desde su lanzamiento en 1998 y caracterizado por ser de código abierto [?]. Para llevar a cabo el filtrado, Snort emplea la biblioteca libpcap para capturar paquetes de red, que posteriormente son decodificados para identificar su estructura y los protocolos involucrados (Ethernet, IP, TCP/UDP, entre otros). A partir de esta información, Snort monitoriza el tráfico y genera alertas en función de un conjunto de reglas predefinidas. Además, este sistema puede operar en diferentes modos: sniffer, logger o como IDS propiamente dicho, y permite su ampliación mediante un sistema de plugins.

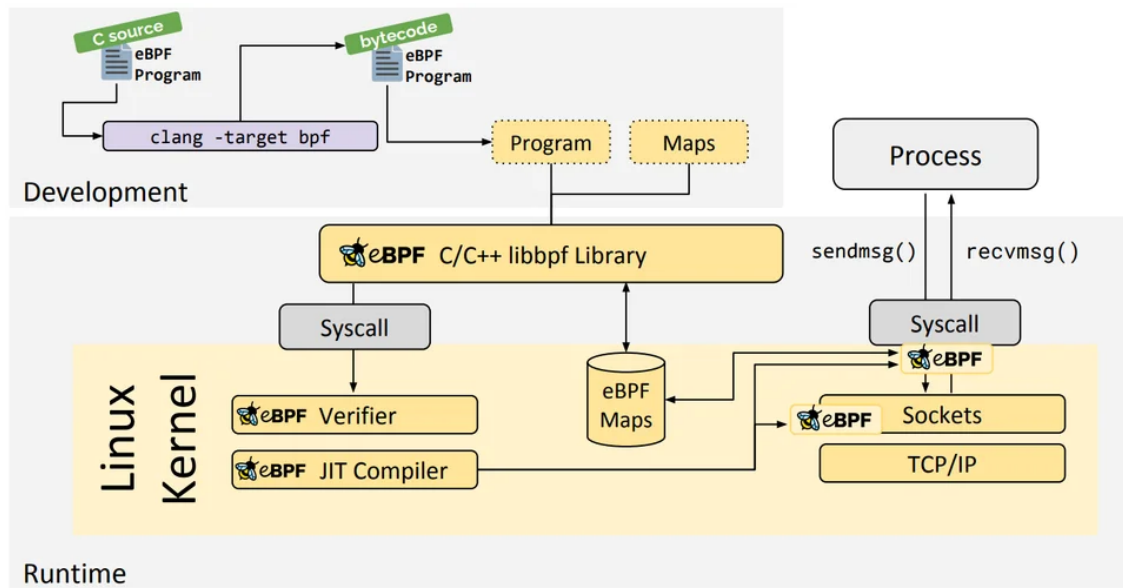
Esta evolución desde mecanismos de filtrado simples hacia un análisis más profundo del tráfico se ha visto respaldada por tecnologías como libpcap y BPF (Berkeley Packet Filter), que permiten la captura eficiente de paquetes sin comprometer significativamente el rendimiento del sistema. BPF es una interfaz que habilita la captura de paquetes a través de programas escritos en lenguaje C, los cuales deben ajustarse a un conjunto de restricciones más estricto que el lenguaje C estándar. Un

aspecto clave de BPF es su modelo de ejecución: aunque los filtros se definen desde el espacio de usuario, su ejecución se realiza en el espacio del kernel, lo que permite un procesamiento más rápido y eficiente al evitar cambios de contexto innecesarios [?].

Por otro lado, libpcap es una biblioteca inicialmente desarrollada en C, que actualmente cuenta con una versión en Python, y que se utiliza para capturar paquetes ofreciendo una API uniforme para acceder a datos de red a bajo nivel desde el espacio de usuario. Cuando el sistema lo permite, libpcap hace uso de BPF para realizar filtrado a nivel de kernel [?]. Esta biblioteca (y, por extensión, BPF) ha sido utilizada en multitud de herramientas, entre ellas Wireshark [?].

BPF ha ido evolucionando a lo largo de los años, llegando su desarrollo hasta eBPF (Extended Berkeley Packet Filter). eBPF, como su propio nombre indica, comenzó siendo una extensión de BPF, pero con el paso del tiempo se ha convertido en una plataforma más general capaz de interactuar con distintas partes del sistema operativo, tales como redes, seguridad, trazado y observabilidad. A diferencia del BPF original —que únicamente filtraba paquetes en el espacio del socket—, eBPF permite engancharse (hook) en múltiples puntos del kernel, ejecutar programas verificados para garantizar la seguridad, utilizar mapas compartidos entre el kernel y el espacio de usuario, y beneficiarse de la compilación JIT (Just In Time) para un alto rendimiento [?]. En base a eBPF, se ha desarrollado XDP (eXpress Data Path), una tecnología para el kernel de Linux que permite ejecutar programas eBPF en una etapa más temprana del procesamiento de paquetes, lo que se traduce en una menor latencia y un mayor rendimiento. Para ello, XDP establece como hook el controlador (driver) de la tarjeta de red, o bien opera en modos alternativos como generic (pila de red) u offload (hardware si la NIC lo soporta). De este modo, el filtrado de paquetes se sitúa en la capa más baja de la pila de red. Es importante señalar que XDP requiere especificar la tarjeta de red sobre la que se ejecutará el programa eBPF [?].

Para entender lo que resta del presente trabajo es necesario entender el flujo de funcionamiento de XDP y por lo tanto es necesario conocer el flujo de eBPF, para ello nos basaremos en 2.1 obtenido de [?]. El primer paso consiste en escribir el programa en un subconjunto restringido del lenguaje C. Posteriormente, mediante herramientas como Clang y LLVM, el código se compila a bytecode. A continuación, desde el espacio de usuario, el programa se carga en el espacio del kernel y se prepara el espacio de usuario para recibir información mediante los mapas definidos en el programa, en caso de ser necesario. Estos dos últimos pasos pueden realizarse mediante un programa en C en el espacio de usuario utilizando la biblioteca libbpf. Una vez que el programa llega al kernel, este pasa por el verificador, cuyo objetivo es impedir la ejecución de acciones no permitidas, como bucles infinitos o accesos ilegales a memoria. A continuación, el compilador en tiempo real (Just-In-Time compiler, JIT) transforma el bytecode en código máquina. Tras este proceso, el programa se “engancha” (hook) a diferentes eventos de la capa de red, como, por ejemplo, el controlador (driver) de la tarjeta de red en el caso de XDP. Durante la ejecución, y siempre que sea necesario, el programa interactúa con los mapas definidos. Este flujo garantiza que el código eBPF se ejecute de forma segura y eficiente, integrándose de manera controlada en el sistema operativo.



**Figura 2.1:** Flujo de ejecución de un programa eBPF. Obtenida de [?] siguiendo la licencia Creative Commons Attribution 4.0 International License.

## 2.3. Artículos Similares y Relacionados

Como se ha expuesto en la motivación y en las subsecciones anteriores, la detección temprana de ransomware y la necesidad de contar con sistemas de filtrado de paquetes que ofrezcan un alto rendimiento y una baja latencia constituyen problemáticas que han captado la atención y el interés de la comunidad académica desde hace tiempo. Dado que el presente trabajo se centra en analizar el impacto en el rendimiento de red de sistemas de detección y filtrado de ransomware basados en árboles de decisión, resulta pertinente revisar los trabajos académicos que exploran tanto el uso de tecnologías como eBPF y XDP en el filtrado de tráfico malicioso, como aquellos que estudian la aplicación de árboles de decisión para la detección de ransomware.

# IMPLEMENTACIÓN

---

## 3.1. Arquitectura de la propuesta

## 3.2. Componentes

Los componentes de la arquitectura se dividen en tres módulos principales. El primero corresponde al árbol de decisión, entrenado mediante la biblioteca scikit-learn de Python. El segundo está constituido por el código XDP, el cual incluye una parte autogenerada a partir del árbol de decisión. Finalmente, el tercer módulo es el entorno Mininet, empleado para la ejecución de pruebas en un entorno virtual local. Cabe destacar el fichero `DecisionTree/arbolrw.py`, encargado tanto del entrenamiento del árbol de decisión como de la autogeneración del código C, apoyándose para ello en una plantilla Jinja.

### 3.2.1. Árbol de decisión

El conjunto de datos empleado para el entrenamiento del árbol de decisión se ha generado a partir de un fichero CSV de 294,6 MB elaborado por el autor. Para su creación, se utilizó la versión más reciente del script `pcap2csv.py`, el cual permite la unión de dos o más ficheros PCAP en un único fichero CSV, así como la diferenciación entre paquetes de tráfico malicioso y legítimo mediante el uso de las banderas `-l` y `-m`. El script almacena únicamente información de los paquetes hasta el nivel de aplicación, sin analizar el contenido de este nivel, debido a la complejidad que supone obtener dichos datos en C a nivel de kernel.

El proceso de conversión comienza definiendo un diccionario con todas las posibles características que puede tener un paquete de red. A continuación, se leen los paquetes uno a uno y cada uno se procesa mediante una función que devuelve un diccionario en el que cada clave corresponde al nombre de una columna y cada valor al dato asociado a ese paquete en particular (véase un fragmento de esta función en [3.1](#)). En caso de que un paquete carezca de alguno de los campos —por ejemplo, los relativos a IP—, dichos valores se registran como `None` en lugar de omitirse, con el fin de mantener un formato uniforme en el CSV y evitar errores durante el procesamiento posterior.

Una vez transformado el paquete en diccionario, se añade al final la etiqueta “Yes” si corresponde a una muestra de ransomware, o “No” en caso contrario. Posteriormente, el diccionario se incorpora a un array. Cuando dicho array alcanza los 1024 elementos, su contenido se vuelca al fichero CSV y se vacía para liberar memoria. Este procedimiento se repite hasta procesar todos los paquetes, momento en el que se realiza un último volcado con los datos restantes. Se ha optado por este método de lectura y escritura para optimizar tanto el uso de memoria como la frecuencia de accesos de escritura al disco.

Los ficheros PCAP correspondientes al tráfico de ransomware fueron obtenidos, tal y como se explica en el apartado 2.1 por el trabajo de Eduardo Berrueta et al. entre los años 2015 y 2022 [?]. Por otro lado, el fichero PCAP con tráfico legítimo fue generado por el autor mediante la captura de tráfico durante una jornada laboral en un laboratorio universitario, registrando el uso normal de los dispositivos y recogiendo tráfico de protocolos como HTTP, HTTPS, DNS, SSH, entre otros. Con el objetivo de incrementar la precisión del árbol de decisión, se generó adicionalmente tráfico SMB2 mediante la copia de un fichero de gran tamaño a una unidad de almacenamiento ubicada en el CPD de la universidad.

En el proceso de preparación del dataset, se eliminaron las direcciones IP, las direcciones MAC y el timestamp de los paquetes, debido al número limitado presente tanto en las capturas de tráfico malicioso como en las de tráfico legítimo. Con ello, se evitó que el árbol de decisión se apoyara en características poco representativas o irrelevantes para un entorno real. Como métrica de calidad se mantuvo el índice de Gini, dada su fiabilidad y buen rendimiento, y como estrategia de división se empleó el criterio best. Aunque se preveía que el aumento del tamaño del dataset respecto a las pruebas podría incrementar el tiempo de entrenamiento, en la práctica no se observó un impacto significativo. En caso contrario, se habría considerado cambiar la estrategia a random. En cuanto a la profundidad máxima del árbol, se fijó inicialmente en cuatro niveles, ya que en pruebas previas con otros datasets de paquetes de red no se apreciaron mejoras de precisión al superar este valor.

Para entrenar el árbol de decisión, una vez definidas las columnas que serán eliminadas, se establece la variable  $x$  como el conjunto de columnas que forman parte de los datos de entrada y la variable  $y$  como la columna correspondiente al clasificador. A continuación, mediante la función `train_test_split`, se divide el conjunto de datos de forma aleatoria, asignando el 80 % de las filas al entrenamiento y el 20 % restante a las pruebas, con el fin de evaluar la fiabilidad del modelo. Una vez realizada esta partición, se procede a entrenar el árbol utilizando los parámetros seleccionados y los datos de entrenamiento. Finalmente, se puede calcular la exactitud del modelo, si bien es recomendable interpretar este valor con cautela, dado que no siempre refleja de forma completamente sólida la capacidad de generalización del clasificador. El código correspondiente al entrenamiento del árbol puede consultarse en 3.2

Durante el desarrollo, se detectó que, por error, se había incluido la variable `eth_type` en la lista de columnas excluidas del entrenamiento. Tras corregir esta omisión y volver a entrenar el modelo, se



observó que la precisión se mantenía, pero esta vez con un nivel menos de profundidad, alcanzando una profundidad máxima de tres niveles a pesar de que el límite establecido era de cuatro. Este resultado sugiere que la inclusión de dicha variable aporta capacidad predictiva al modelo, optimizando su estructura sin pérdida de precisión.

**Código 3.1:** Fragmento de código que transforma parte de los datos de un paquete en un diccionario fácilmente convertible a CSV.

```

39 def paqt2dict(paqt: Packet) -> Dict[str, Any]:
40     """Función que convierte un paquete en un diccionario con las características extraídas."""
41     dic = {}
42
43     try:
44         packet = Ether(paqt) if isinstance(paqt, bytes) else paqt
45     except Exception as e:
46         print(f"Error al procesar el paquete:_{e}")
47         return {}
48
49     # Tamaño total del paquete
50     dic["size"] = len(packet)
51
52     # Capa Ethernet
53     if packet.haslayer(Ether):
54         dic["eth_src"] = packet[Ether].src
55         dic["eth_dst"] = packet[Ether].dst
56         dic["eth_type"] = packet[Ether].type
57     else:
58         dic["eth_src"] = None
59         dic["eth_dst"] = None
60         dic["eth_type"] = None

```

### 3.2.2. Código XDP y Código autogenerado

Se añadió la capacidad de autogenerar el código BPF/XDP una vez entrenado el árbol de decisión. Para ello, se recorre de forma recursiva el árbol y, en cada nodo, se genera la condición correspondiente a partir del nombre de la característica (previamente convertido a su nombre real) y del valor de umbral. Esta condición se pasa por una función que, en determinados casos, la sustituye por una equivalente válida. Por ejemplo, la condición `src_port <= inf` se reemplaza por `ip_proto == IPPROTO_TCP || ip_proto == IPPROTO_UDP`, dado que el hecho de que el puerto de destino sea menor que infinito únicamente implica que el puerto existe.

Cuando la condición resultante es válida, si corresponde al hijo izquierdo del nodo se inserta en una estructura `if`, mientras que si corresponde al hijo derecho se inserta en una estructura `else`. Para la generación del código se empleó una plantilla Jinja con la estructura prediseñada en C. Durante las pruebas, se observó que el mejor resultado se obtenía extrayendo los datos de las cabeceras IP antes

**Código 3.2:** Fragmento de código de arbolrw.py encargado de entrenar el árbol de decisión.

```

82     # Especifica la ruta del archivo CSV
83     csv_file_path = '../dataset.csv'
84
85     # Lista de columnas que NO quieres usar para el entrenamiento
86     columnas_a_excluir = ["eth_src", "eth_dst", "ip_src", "ip_dst", "timestamp"]
87
88     # Cargar el CSV
89     try:
90         df = pd.read_csv(csv_file_path)
91     except FileNotFoundError:
92         print(f"Error: El archivo CSV no se encontró en la ruta: {csv_file_path}")
93         exit()
94
95     # Última columna es la clase objetivo
96     nombre_columna_objetivo = df.columns[-1]
97     y = df[nombre_columna_objetivo]
98
99     # Columnas X sin las excluidas
100    columnas_caracteristicas = [col for col in df.columns if col != nombre_columna_objetivo and col not
        in columnas_a_excluir]
101    x = df[columnas_caracteristicas]
102    x = pd.get_dummies(x)
103
104    # División de datos
105    x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)
106
107    # Entrenar el modelo
108    modelo = DecisionTreeClassifier(criterion='gini', splitter='best', max_depth=4, random_state=42)
109    modelo.fit(x_train, y_train)
110    exactitud = modelo.score(x_test, y_test)
111    print(f'La exactitud del modelo es: {exactitud:.2f}')

```

de la evaluación del árbol, de forma que ya estuvieran disponibles en la plantilla. En caso de que el paquete analizado carezca de cabecera IP, dichos valores se establecen en cero.

