ISSN 2395-8618

Clasificación de tráfico cifrado en redes domésticas mediante análisis estadístico de flujos y aprendizaje supervisado

Duvan Estiven Uribe-Rodríguez, César Jesús Núñez-Prado*

***Resumen*—El crecimiento del tráfico cifrado en redes domésticas ha reducido considerablemente la efectividad de las técnicas tradicionales de monitoreo y clasificación basadas en inspección profunda de paquetes, dificultando la implementación de mecanismos eficientes de calidad de servicio (QoS). En este trabajo se propone un enfoque de clasificación de tráfico cifrado basado en el análisis estadístico de flujos y la aplicación de algoritmos de aprendizaje supervisado. Para ello, se utilizó un conjunto de datos público especializado en tráfico de red cifrado, sobre el cual se realizaron procesos de limpieza, transformación y selección de características estadísticas relacionadas con el comportamiento temporal y volumétrico de los flujos de red. Posteriormente, se implementaron y evaluaron distintos modelos de clasificación, incluyendo Random Forest, K-Nearest Neighbors (KNN), Support Vector Machines (SVM) y Perceptrón Multicapa (MLP), considerando escenarios con datos balanceados y desbalanceados. El desempeño de los modelos fue analizado mediante métricas derivadas de la matriz de confusión, utilizando Accuracy, Precision, Recall y F1-Macro como criterios principales de evaluación. Los resultados experimentales mostraron que el modelo Random Forest obtuvo el mejor desempeño global, alcanzando una exactitud de 96.67 % y un valor F1-Macro de 0.96 bajo el escenario con datos desbalanceados. Asimismo, el balanceo de clases permitió obtener un comportamiento más equilibrado entre categorías y mejorar el desempeño sobre clases con menor representación. Los resultados obtenidos demuestran que los modelos supervisados son capaces de identificar categorías de tráfico cifrado utilizando únicamente características estadísticas de flujo, sin necesidad de inspeccionar el contenido de los paquetes, respaldando la viabilidad del enfoque propuesto como herramienta de apoyo para mecanismos inteligentes de gestión y priorización de tráfico en redes domésticas cifradas.**

***Palabras clave*—Tráfico cifrado, clasificación de tráfico de red, aprendizaje supervisado, calidad de servicio, análisis estadístico de flujos, redes domésticas.**

Encrypted Traffic Classification in Domestic Networks Using Statistical Flow Analysis and Supervised Learning

***Abstract*—The growth of encrypted traffic in domestic networks has considerably reduced the effectiveness of traditional monitoring and classification techniques based on deep packet inspection, making the implementation of efficient Quality of Service (QoS) mechanisms increasingly difficult. This work proposes an encrypted traffic classification approach based on statistical flow analysis and supervised learning algorithms. A public dataset specialized in encrypted network traffic was used, applying data cleaning, transformation, and feature selection processes focused on temporal and volumetric flow characteristics. Subsequently, several classification models were implemented and evaluated, including Random Forest, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Multilayer Perceptron (MLP), considering both balanced and imbalanced data scenarios. Model performance was evaluated using confusion matrix-based metrics, including Accuracy, Precision, Recall, and F1-Macro as the main evaluation criteria. Experimental results showed that the Random Forest model achieved the best overall performance, reaching an accuracy of 96.67% and an F1-Macro score of 0.96 under the imbalanced data scenario. Additionally, class balancing allowed a more uniform behavior across categories and improved classification performance on minority classes. The obtained results demonstrate that supervised models are capable of identifying encrypted traffic categories using only statistical flow features without inspecting packet contents, supporting the viability of the proposed approach as a tool to assist intelligent traffic management and prioritization mechanisms in encrypted domestic networks.**

***Index Terms*—Encrypted traffic, network traffic classification, supervised learning, quality of service, statistical flow analysis, domestic networks.**

Manuscript received on 15/03/2025, accepted for publication on 29/06/2025. Corresponding author is César Jesús Núñez-Prado.

The authors are with Instituto Politécnico Nacional, Escuela Superior de Ingeniería Mecánica y Eléctrica, Mexico City, Mexico. Emails: duvanestiven321@gmail.com, cnunezp@ipn.mx.

I. Introducción

La clasificación de tráfico de red es una tarea fundamental para la gestión eficiente de recursos y la implementación de mecanismos de calidad de servicio (QoS) en redes domésticas [13]. El crecimiento sostenido de aplicaciones con alta demanda de ancho de banda, como plataformas de streaming, videojuegos en línea y videoconferencias, ha incrementado significativamente la carga sobre las infraestructuras de red domésticas, provocando problemas de congestión, latencia y degradación del servicio [4].

Tradicionalmente, la identificación de aplicaciones y servicios de red se realizaba mediante inspección profunda de paquetes y análisis de puertos. Sin embargo, la adopción generalizada de protocolos de cifrado como SSL / TLS ha limitado considerablemente la capacidad de inspeccionar el contenido de los paquetes, reduciendo la efectividad de

ISSN 2395-8618

las técnicas convencionales de clasificación de tráfico [8], [15]. Como consecuencia, los mecanismos tradicionales de monitoreo y priorización de tráfico enfrentan dificultades para distinguir aplicaciones sensibles a la latencia, afectando la implementación eficiente de políticas de QoS en redes modernas.

Ante esta problemática, diversas investigaciones han explorado el uso de características estadísticas de flujo y algoritmos de aprendizaje automático para clasificar tráfico cifrado sin acceder al contenido de los datos. Trabajos previos han demostrado que atributos como duración del flujo, tamaño de paquetes y tiempos entre llegadas permiten diferenciar distintos tipos de servicios de red mediante técnicas de minería de datos, aprendizaje supervisado y aprendizaje profundo [2], [7], [9]. No obstante, muchos enfoques recientes se centran en arquitecturas complejas de aprendizaje profundo o escenarios específicos de detección de anomalías, mientras que el análisis comparativo de modelos supervisados tradicionales aplicados a tráfico cifrado multiclase en entornos domésticos continúa siendo un área de interés práctico.

En este contexto, el presente trabajo propone un enfoque de clasificación de tráfico cifrado orientado a redes domésticas mediante el análisis estadístico de flujos y la aplicación de algoritmos de aprendizaje supervisado. Para ello, se evaluaron modelos basados en Random Forest, K-Nearest Neighbors (KNN), Support Vector Machines (SVM) y Perceptrón Multicapa (MLP), utilizando un conjunto de datos público especializado en tráfico cifrado [3], [5], [6], [11], [14]. La metodología contempla etapas de preprocesamiento, selección de características y evaluación experimental bajo escenarios con datos balanceados y desbalanceados, con el objetivo de analizar el impacto de la distribución de clases sobre el desempeño de los modelos de clasificación.

Los resultados obtenidos muestran que los modelos supervisados son capaces de identificar categorías de tráfico cifrado utilizando únicamente características estadísticas de flujo, sin necesidad de inspeccionar el contenido de los paquetes. Asimismo, el estudio permite comparar el comportamiento de diferentes algoritmos bajo escenarios multiclase y evaluar la influencia del balanceo de datos sobre métricas de clasificación orientadas a QoS.

Las principales contribuciones de este trabajo son las siguientes:

- Se propone un enfoque de clasificación de tráfico cifrado basado exclusivamente en características estadísticas de flujo, preservando la privacidad del usuario al evitar la inspección del contenido de los paquetes.
- Se realiza un análisis comparativo entre distintos algoritmos de aprendizaje supervisado, incluyendo Random Forest, KNN, SVM y Perceptrón Multicapa, utilizando métricas orientadas a clasificación multiclase.
- Se evalúa el impacto del balanceo de clases sobre el desempeño de los modelos de clasificación en escenarios de tráfico cifrado.
- Se analiza la viabilidad del aprendizaje automático como herramienta de apoyo para mecanismos inteligentes de gestión y priorización de tráfico en redes domésticas cifradas.

El resto del artículo está organizado de la siguiente manera. La Sección II presenta el trabajo relacionado. La Sección III describe la metodología propuesta, incluyendo el conjunto de datos, las etapas de preprocesamiento y los modelos de clasificación utilizados. La Sección IV presenta los resultados experimentales y el análisis comparativo de los modelos evaluados. Finalmente, la Sección V expone las conclusiones y posibles líneas de trabajo futuro.

II. Trabajo relacionado

La clasificación de tráfico de red mediante técnicas de inteligencia artificial ha adquirido una creciente relevancia debido al incremento del tráfico cifrado y a las limitaciones de los métodos tradicionales basados en inspección profunda de paquetes [8], [10], [15]. En este contexto, diversas investigaciones han explorado estrategias basadas en análisis estadístico de flujos y aprendizaje automático para identificar servicios y comportamientos de red sin acceder al contenido de los paquetes.

Uno de los enfoques iniciales orientados al análisis estadístico de tráfico fue presentado por Gil Delgado et al. [7], quienes propusieron una metodología de clasificación basada en características extraídas de flujos de red, incluyendo duración, número de paquetes, cantidad de bytes y tiempos promedio entre llegadas. Su metodología integró herramientas como Wireshark, pkt2flow y tshark para la captura y procesamiento de información, complementándose con técnicas de reducción de dimensionalidad mediante PCA y agrupamiento utilizando K-Means. Los resultados reportaron una precisión cercana al 82% en la diferenciación entre tráfico web y tráfico P2P, demostrando la viabilidad del análisis estadístico de flujos como alternativa a la inspección tradicional de paquetes.

Posteriormente, Montaña y Montoya exploraron el uso de aprendizaje supervisado y aprendizaje profundo para el diseño de sistemas de detección de intrusos [10]. Su propuesta destacó la capacidad de los modelos de Deep Learning para reconocer patrones complejos y detectar anomalías dentro del tráfico de red. Aunque el trabajo presentó principalmente una arquitectura conceptual y no incluyó una evaluación experimental detallada, estableció una transición importante entre los enfoques estadísticos tradicionales y los modelos supervisados aplicados al análisis de tráfico.

El problema específico de la clasificación de tráfico cifrado fue abordado por Elmaghraby et al. [9], quienes desarrollaron modelos capaces de identificar distintos servicios de red utilizando únicamente metadatos y características temporales del flujo, sin necesidad de inspeccionar el contenido cifrado. Su metodología incluyó modelos basados en redes neuronales, arquitecturas LSTM bidireccionales y modelos

ISSN 2395-8618

híbridos CNN-LSTM, complementados con clasificadores tradicionales como Random Forest, SVM y KNN. Los autores reportaron valores de exactitud superiores al 95%, validando la efectividad del aprendizaje automático para la clasificación de tráfico cifrado en escenarios multiclase.

Por otra parte, Barcenas-Medina y Alcaraz-Cancio [2] exploraron el uso de modelos híbridos de aprendizaje profundo para la detección de comportamientos complejos y anomalías en tráfico de red empresarial. Su propuesta integró Redes Neuronales Convolucionales (CNN) y Redes Neuronales Recurrentes (RNN) para capturar patrones espaciales y temporales dentro de los flujos de red. Los resultados alcanzaron una precisión del 96.84%, evidenciando el potencial de los modelos híbridos de aprendizaje profundo para tareas avanzadas de análisis y clasificación de tráfico.

En conjunto, los trabajos previos demuestran que las características estadísticas de flujo contienen información suficiente para identificar distintos tipos de tráfico cifrado sin acceder al contenido de los paquetes. Sin embargo, gran parte de las investigaciones recientes se enfocan en arquitecturas complejas de aprendizaje profundo o en escenarios orientados a detección de anomalías e intrusiones. En contraste, el presente trabajo se centra en el análisis comparativo de algoritmos supervisados aplicados a clasificación multiclase de tráfico cifrado en redes domésticas, considerando además el impacto del balanceo de clases sobre el desempeño de los modelos y su posible aplicación como apoyo para mecanismos de calidad de servicio.

III. Metodología

La metodología propuesta para la clasificación de tráfico cifrado se estructuró en cinco etapas principales: selección del conjunto de datos, preprocesamiento, selección de características, entrenamiento de modelos y evaluación experimental. El objetivo consistió en analizar la capacidad de distintos algoritmos de aprendizaje supervisado para identificar categorías de tráfico de red utilizando exclusivamente características estadísticas de flujo, sin necesidad de inspeccionar el contenido de los paquetes.

La Figura 1 presenta el flujo general de la metodología empleada en esta investigación. Inicialmente, se seleccionó un conjunto de datos público especializado en tráfico cifrado y posteriormente se aplicaron procesos de limpieza, transformación y consolidación de categorías. A continuación, se realizó un análisis de relevancia de características con el propósito de identificar los atributos estadísticos con mayor capacidad discriminativa dentro del tráfico de red.

Posteriormente, los datos preprocesados fueron utilizados para entrenar distintos modelos de clasificación supervisada, incluyendo Random Forest, K-Nearest Neighbors (KNN), Support Vector Machines (SVM) y Perceptrón Multicapa (MLP). Finalmente, los modelos fueron evaluados bajo escenarios con datos balanceados y desbalanceados mediante métricas derivadas de la matriz de confusión, permitiendo analizar el impacto de la distribución de clases sobre el desempeño de los clasificadores.

A. Conjunto de datos

El conjunto de datos utilizado en esta investigación proviene de un repositorio público de la Universidad de New Brunswick (UNB) [14], seleccionado debido a que contiene capturas reales de tráfico de red previamente etiquetadas y orientadas a tareas de clasificación de tráfico cifrado. El dataset incluye diferentes categorías asociadas a servicios de red como VoIP, Streaming, Chat, Navegación y Transferencia de archivos, permitiendo evaluar escenarios de clasificación multiclase en entornos de tráfico cifrado.

Cada instancia del conjunto de datos corresponde a un flujo de red definido mediante la denominada “5-tupla” [12], compuesta por dirección IP de origen, dirección IP de destino, puerto de origen, puerto de destino y protocolo de transporte. A partir de estos flujos se extrajeron características estadísticas relacionadas con el comportamiento temporal y volumétrico del tráfico, incluyendo duración del flujo, cantidad de paquetes, volumen de bytes transferidos y tiempos entre llegadas de paquetes.

Debido a que el contenido de los paquetes se encuentra cifrado mediante protocolos SSL/TLS, el proceso de clasificación se basó exclusivamente en atributos estadísticos visibles a nivel de flujo. Este enfoque permite preservar la privacidad del usuario al evitar la inspección directa del contenido transmitido, manteniendo al mismo tiempo información suficiente para diferenciar distintos tipos de servicios de red mediante técnicas de aprendizaje automático.

B. Preprocesamiento de datos

Como etapa inicial, se realizó un proceso de limpieza sobre el conjunto de datos con el propósito de eliminar registros inconsistentes, valores faltantes y atributos redundantes que pudieran afectar el entrenamiento de los modelos de clasificación. Asimismo, se llevó a cabo una consolidación de categorías relacionadas con distintos tipos de tráfico cifrado, permitiendo reducir redundancias y obtener una estructura multiclase más adecuada para el análisis experimental.

Posteriormente, las variables numéricas fueron transformadas y estandarizadas mediante normalización tipo Z-score, con el objetivo de reducir diferencias de escala entre características y evitar que atributos con magnitudes mayores dominaran el proceso de entrenamiento de los clasificadores.

Adicionalmente, se aplicó un proceso de selección de características para identificar los atributos estadísticos con mayor capacidad discriminativa dentro del tráfico de red. Esta etapa tuvo como finalidad reducir dimensionalidad, disminuir ruido y mejorar la capacidad predictiva de los modelos supervisados. La relevancia de las características fue evaluada mediante criterios asociados al índice de impureza de Gini, permitiendo identificar variables relacionadas con duración del

ISSN 2395-8618

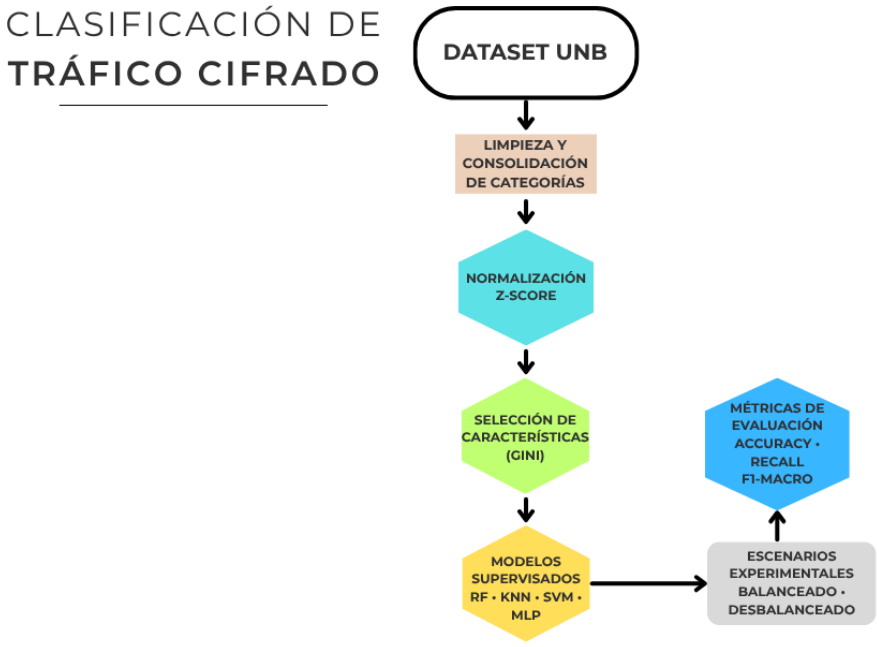

Fig. 1. Metodología empleada.

flujo, volumen de información transmitida y tiempos entre llegadas de paquetes como algunos de los atributos más representativos para la clasificación de tráfico cifrado.

C. Modelos de clasificación

Con el propósito de analizar el comportamiento de distintos enfoques de aprendizaje supervisado sobre tráfico cifrado, se implementaron cuatro algoritmos ampliamente utilizados en tareas de clasificación de tráfico de red:

- Random Forest (RF),
- K-Nearest Neighbors (KNN),
- Support Vector Machines (SVM),
- Perceptrón Multicapa (MLP).

El modelo Random Forest fue seleccionado debido a su capacidad para manejar relaciones no lineales y reducir el sobreajuste mediante la agregación de múltiples árboles de decisión [3]. Por otra parte, K-Nearest Neighbors permitió evaluar un enfoque de clasificación basado en proximidad geométrica entre flujos de red con características estadísticas similares [6]. Asimismo, Support Vector Machines fue empleado para construir fronteras de decisión no lineales utilizando funciones kernel [5], mientras que el Perceptrón Multicapa permitió modelar relaciones complejas mediante redes neuronales multicapa [10], [11].

En el caso de KNN, se evaluaron diferentes valores del parámetro k, específicamente $k = 3$, $k = 5$ y $k = 7$, con el propósito de analizar la sensibilidad del modelo respecto al número de vecinos considerados durante el proceso de clasificación. Para SVM se utilizó un kernel de tipo RBF (*Radial Basis Function*), debido a su capacidad para modelar relaciones no lineales dentro del espacio de características. Por otra parte, el modelo MLP fue implementado mediante una arquitectura multicapa completamente conectada entrenada utilizando aprendizaje supervisado basado en retropropagación.

Todos los modelos fueron entrenados utilizando las características estadísticas extraídas de los flujos de red y posteriormente evaluados bajo escenarios con datos balanceados y desbalanceados, con el objetivo de analizar su capacidad de generalización y desempeño en tareas de clasificación multiclase de tráfico cifrado.

La Tabla I resume la configuración general utilizada para cada uno de los modelos evaluados.

D. Diseño experimental

El diseño experimental se estructuró en dos escenarios de evaluación con el propósito de analizar el impacto de la distribución de clases sobre el desempeño de los modelos de

ISSN 2395-8618

TABLE I
CONFIGURACIÓN GENERAL DE LOS MODELOS DE CLASIFICACIÓN.

Modelo	Configuración principal
Random Forest	Ensamble de árboles de decisión
KNN	$k = 3, 5, 7$
SVM	Kernel RBF
MLP	Red neuronal multicapa fully-connected

clasificación y evaluar su capacidad de generalización bajo diferentes condiciones de entrenamiento.

En el primer escenario experimental se utilizaron los datos en su distribución original, conservando el desbalance natural entre categorías de tráfico presentes en el conjunto de datos. Este escenario permitió analizar el comportamiento de los clasificadores bajo condiciones más cercanas a entornos reales de red, donde ciertas categorías poseen una representación significativamente mayor respecto a otras. Para este caso, la evaluación se realizó mediante una partición entrenamiento-prueba de tipo *holdout* utilizando una proporción 80/20.

En el segundo escenario experimental se empleó una versión balanceada del conjunto de datos con el objetivo de reducir el sesgo hacia las clases mayoritarias y analizar el comportamiento de los modelos bajo una distribución más uniforme entre categorías. Este escenario permitió evaluar con mayor detalle la capacidad de clasificación sobre clases con menor representación y analizar el impacto del balanceo sobre métricas orientadas a problemas multiclase. Para esta etapa se utilizó validación cruzada *k-fold* con $k = 5$, permitiendo obtener una estimación más robusta del rendimiento de los modelos y reducir la dependencia respecto a una única partición de los datos.

Ambos escenarios fueron utilizados para comparar el desempeño de los algoritmos supervisados en tareas de clasificación multiclase de tráfico cifrado, así como para analizar la influencia de la distribución de clases sobre métricas globales y por categoría.

E. Métricas de evaluación

El desempeño de los modelos de clasificación fue evaluado mediante métricas derivadas de la matriz de confusión, incluyendo Accuracy, Precision, Recall y F1-Score. Debido a la naturaleza multiclase del problema y al posible desbalance entre categorías de tráfico, se utilizó el valor F1-Macro como principal criterio de comparación, ya que esta métrica asigna el mismo peso a todas las clases independientemente de su frecuencia dentro del conjunto de datos.

Adicionalmente, se emplearon matrices de confusión para realizar un análisis cualitativo del comportamiento de los clasificadores, permitiendo identificar patrones de error y observar la capacidad de cada modelo para diferenciar las distintas categorías de tráfico cifrado.

IV. RESULTADOS

En esta sección se presentan los resultados obtenidos durante la evaluación experimental de los modelos de clasificación implementados. Inicialmente, se describe el análisis exploratorio del conjunto de datos y el impacto del proceso de preprocesamiento sobre la estructura final utilizada para el entrenamiento. Posteriormente, se analizan los resultados obtenidos bajo escenarios con datos desbalanceados y balanceados, utilizando métricas orientadas a clasificación multiclase.

Asimismo, se incluye un análisis comparativo del desempeño de los distintos algoritmos supervisados evaluados, considerando tanto métricas globales como matrices de confusión con el propósito de identificar patrones de clasificación y comportamiento frente a las diferentes categorías de tráfico cifrado.

A. Análisis exploratorio del conjunto de datos

Como etapa inicial, se realizó un análisis exploratorio del conjunto de datos con el propósito de identificar la distribución de las categorías de tráfico y evaluar la calidad de la información utilizada durante el entrenamiento de los modelos de clasificación. La Figura 2 muestra la distribución original de las clases presentes en el dataset de la Universidad de New Brunswick (UNB). Se observa un desbalance considerable entre categorías, además de redundancia en algunas etiquetas asociadas al uso de VPN y tráfico Non-VPN. Categorías como BROWSING, VOIP y VPN presentan una cantidad significativamente mayor de muestras respecto a clases como MAIL, STREAMING y CHAT, lo que podría generar sesgos durante el entrenamiento y favorecer a las clases mayoritarias durante el proceso de clasificación.

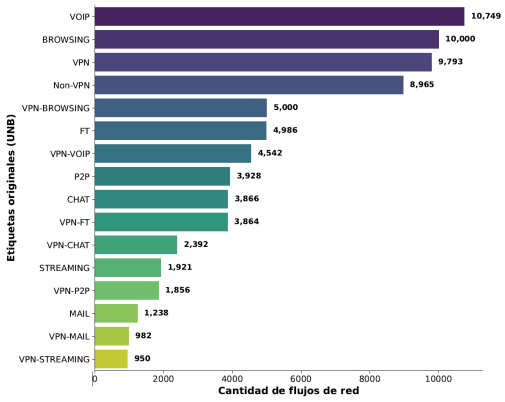

Fig. 2. Distribución de clases en el conjunto de datos original.

A partir de este análisis inicial, se realizó un proceso de limpieza y consolidación de categorías con el objetivo de eliminar redundancias y agrupar clases relacionadas con tipos específicos de servicios de red. Asimismo, se eliminaron registros inconsistentes y atributos que no

ISSN 2395-8618

aportaban información discriminativa relevante para el proceso de clasificación.

La Tabla II resume el proceso de depuración aplicado sobre el conjunto de datos. Después del preprocesamiento, el número total de categorías fue reducido y el conjunto final presentó una estructura más adecuada para el entrenamiento de modelos supervisados orientados a clasificación multiclase.

TABLE II
RESUMEN DEL PROCESO DE LIMPIEZA Y DEPURACIÓN DE DATOS.

Estado del dataset	Registros	Categorías
Original (crudo)	75,032	14+
Post-limpieza	56,274	5

Posteriormente, se realizó un análisis de relevancia de características utilizando criterios basados en impureza de Gini con el propósito de identificar las variables con mayor capacidad discriminativa dentro del tráfico cifrado. La Figura 3 presenta las 15 características más relevantes obtenidas durante esta etapa.

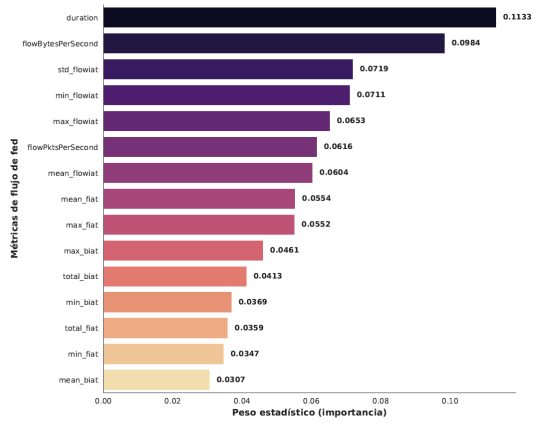

Fig. 3. Top 15 características de red más relevantes evaluadas mediante impureza de Gini.

Los resultados muestran que variables asociadas a duración del flujo, volumen de información transmitida y tiempos entre llegadas de paquetes poseen una elevada relevancia estadística para la diferenciación entre categorías de tráfico cifrado. Particularmente, atributos como *duration*, *flowBytesPerSecond* y métricas relacionadas con *flow inter-arrival time* destacaron como algunas de las variables más representativas dentro del proceso de clasificación.

En conjunto, los resultados obtenidos durante esta etapa confirman que las características estadísticas de flujo contienen información suficiente para distinguir distintos tipos de tráfico de red sin necesidad de inspeccionar directamente el contenido cifrado de los paquetes, respaldando la viabilidad del enfoque basado en aprendizaje supervisado utilizado en esta investigación.

B. Resultados del experimento con datos desbalanceados

En el primer escenario experimental, los modelos de clasificación fueron entrenados utilizando la distribución original del conjunto de datos, conservando el desbalance natural entre categorías de tráfico. Este escenario permitió evaluar el comportamiento de los clasificadores bajo condiciones más cercanas a entornos reales de red, donde ciertas categorías presentan una representación significativamente mayor respecto a otras.

La Tabla III resume los resultados obtenidos por los modelos evaluados utilizando métricas de Accuracy, Precision, Recall y F1-Macro. Los resultados evidencian diferencias importantes entre los algoritmos analizados, particularmente en métricas orientadas a clasificación multiclase.

TABLE III
RESULTADOS DEL EXPERIMENTO CON DATOS DESBALANCEADOS.

Modelo	Accuracy	Precision	Recall	F1-Macro
Random Forest	**0.9667**	**0.97**	**0.95**	**0.96**
KNN ($k = 3$)	0.8564	0.83	0.82	0.82
KNN ($k = 5$)	0.8506	0.79	0.81	0.80
KNN ($k = 7$)	0.8445	0.78	0.80	0.79
MLP	0.8447	0.82	0.79	0.80
SVM (RBF)	0.7088	0.72	0.57	0.59

El modelo Random Forest obtuvo el mejor desempeño global, alcanzando una exactitud de 96.67% y un valor F1-Macro de 0.96. Estos resultados indican una elevada capacidad para diferenciar categorías de tráfico cifrado incluso bajo escenarios con distribución desigual entre clases. Este comportamiento puede atribuirse a la capacidad de los métodos de ensamble para manejar relaciones no lineales y combinar múltiples árboles de decisión, permitiendo capturar patrones estadísticos complejos presentes dentro de los flujos de red. Asimismo, Random Forest mostró una mayor robustez frente al desbalance entre categorías, manteniendo un desempeño competitivo incluso sobre clases con menor representación.

Por otra parte, los modelos KNN presentaron una disminución progresiva del rendimiento conforme aumentó el valor de k, evidenciando sensibilidad frente al desbalance entre categorías y a la proximidad estadística entre ciertos tipos de tráfico. A medida que el número de vecinos considerados aumentó, el modelo tendió a favorecer categorías con mayor representación dentro del espacio de características, reduciendo su capacidad para discriminar adecuadamente clases minoritarias.

En el caso del Perceptrón Multicapa, el modelo alcanzó resultados competitivos, aunque inferiores respecto a Random Forest en métricas globales de clasificación. Este comportamiento sugiere que, si bien las redes neuronales multicapa son capaces de modelar relaciones complejas dentro del tráfico cifrado, los atributos estadísticos utilizados en esta investigación favorecieron enfoques de ensamble más robustos frente a variaciones y ruido presentes en los datos. Finalmente, SVM presentó el desempeño más bajo dentro

ISSN 2395-8618

de este escenario experimental, particularmente en Recall y F1-Macro, indicando dificultades para separar adecuadamente algunas categorías bajo una distribución desbalanceada y con patrones estadísticos parcialmente solapados.

La Figura 4 muestra la matriz de confusión obtenida por el modelo Random Forest utilizando el conjunto de datos desbalanceado.

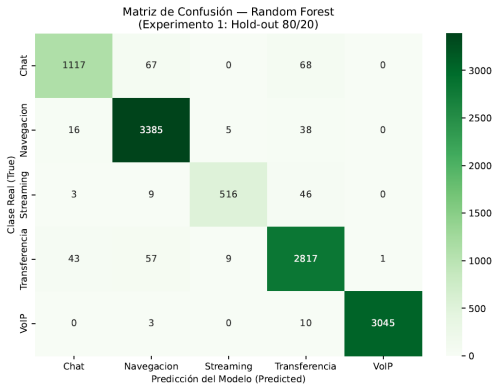

Fig. 4. Matriz de confusión del modelo Random Forest utilizando el conjunto de datos desbalanceado.

El análisis de la matriz de confusión evidencia que las categorías con mayor representación fueron clasificadas con elevada precisión, particularmente Navegación y VoIP. La categoría VoIP presentó uno de los mejores desempeños individuales, alcanzando valores cercanos a 1.00 tanto en precisión como en sensibilidad. Por otra parte, las principales confusiones se observaron entre las categorías Chat y Transferencia, así como entre Streaming y Navegación, lo cual sugiere similitudes en algunos patrones estadísticos asociados al comportamiento temporal y volumétrico de los flujos de red.

En términos generales, los resultados obtenidos en este escenario experimental muestran que los modelos supervisados basados en características estadísticas de flujo son capaces de identificar categorías de tráfico cifrado con un elevado nivel de desempeño, incluso bajo condiciones de desbalance entre clases.

C. Resultados del experimento con datos balanceados

En el segundo escenario experimental se utilizó una versión balanceada del conjunto de datos con el propósito de reducir el sesgo hacia las clases dominantes y analizar el comportamiento de los modelos bajo una distribución más uniforme entre categorías. A diferencia del experimento anterior, este escenario permitió evaluar con mayor detalle la capacidad de generalización de los clasificadores sobre clases con menor representación dentro del tráfico cifrado.

La Tabla IV presenta los resultados obtenidos por los diferentes modelos utilizando métricas de Accuracy, Precision, Recall y F1-Macro.

TABLE IV
RESULTADOS DEL EXPERIMENTO CON DATOS BALANCEADOS.

Modelo	Accuracy	Precision	Recall	F1-Macro
Random Forest	**0.9066**	**0.91**	**0.91**	**0.91**
KNN ($k = 3$)	0.7916	0.80	0.79	0.79
KNN ($k = 5$)	0.7836	0.79	0.78	0.78
KNN ($k = 7$)	0.7785	0.79	0.78	0.78
MLP	0.7765	0.79	0.78	0.78
SVM (RBF)	0.6040	0.63	0.60	0.60

Los resultados muestran que Random Forest mantuvo el mejor desempeño global dentro del escenario balanceado, alcanzando una exactitud de 90.66% y un valor F1-Macro de 0.91. Aunque la exactitud global fue ligeramente inferior respecto al escenario desbalanceado, el modelo presentó un comportamiento más uniforme entre categorías, mejorando particularmente la capacidad de clasificación sobre clases con menor representación. Este comportamiento sugiere que el balanceo de datos permitió reducir el sesgo hacia categorías dominantes y favorecer una distribución de errores más equilibrada durante el proceso de entrenamiento.

Por otra parte, los modelos KNN mostraron un comportamiento más estable bajo el escenario balanceado, especialmente en métricas relacionadas con desempeño multiclase. La reducción en la diferencia de representación entre categorías favoreció la construcción de vecindades más homogéneas dentro del espacio de características, permitiendo mejorar la discriminación entre clases con patrones estadísticos similares. Sin embargo, conforme aumentó el valor de k, el desempeño volvió a disminuir ligeramente, evidenciando nuevamente sensibilidad respecto al número de vecinos considerados durante el proceso de clasificación.

En el caso del Perceptrón Multicapa, el modelo mantuvo un comportamiento competitivo y relativamente estable entre ambos escenarios experimentales. Este resultado indica que las redes neuronales multicapa fueron capaces de capturar relaciones complejas dentro de las características estadísticas del tráfico, aunque sin superar el desempeño alcanzado por los métodos de ensamble. Por otra parte, SVM presentó una reducción en la diferencia de desempeño entre categorías dominantes y minoritarias; sin embargo, continuó mostrando los valores globales más bajos entre los modelos evaluados, posiblemente debido al solapamiento parcial entre algunas categorías dentro del espacio de características.

La Figura 5 muestra la matriz de confusión obtenida por el modelo Random Forest utilizando el conjunto de datos balanceado.

El análisis de la matriz de confusión evidencia una reducción en los errores asociados a categorías minoritarias, particularmente en las clases Chat y Streaming. Asimismo, la categoría VoIP continuó presentando uno de los mejores desempeños individuales, alcanzando valores cercanos a 1.00 tanto en precisión como en sensibilidad. Las principales confusiones persistieron entre categorías con patrones

ISSN 2395-8618

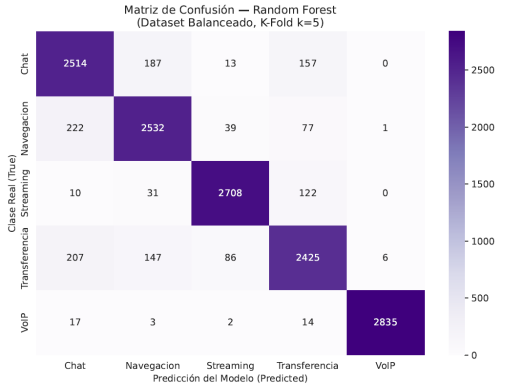

Fig. 5. Matriz de confusión del modelo Random Forest utilizando el conjunto de datos balanceado.

estadísticos similares, como Chat y Transferencia; sin embargo, el comportamiento general de la clasificación mostró una distribución de errores más equilibrada respecto al escenario desbalanceado.

En términos generales, los resultados obtenidos indican que el balanceo de clases contribuye positivamente al entrenamiento de clasificadores supervisados en tareas multiclase relacionadas con tráfico cifrado. Aunque algunos modelos presentaron una ligera disminución en Accuracy global, las métricas orientadas al desempeño por categoría, particularmente F1-Macro y Recall, mostraron un comportamiento más estable y equilibrado entre clases.

D. Comparación entre experimentos

La comparación entre ambos escenarios experimentales permitió analizar el impacto del balanceo de clases sobre el desempeño de los modelos de clasificación. Mientras que el experimento con datos desbalanceados conservó la distribución original del tráfico de red, el escenario balanceado buscó reducir el sesgo hacia las categorías con mayor representación y evaluar el comportamiento de los clasificadores bajo una distribución más uniforme entre clases.

La Tabla V resume los principales resultados obtenidos por los modelos evaluados en ambos escenarios experimentales.

Los resultados muestran que el escenario con datos desbalanceados produjo valores más altos de Accuracy global, particularmente en el modelo Random Forest. Sin embargo, este comportamiento estuvo influenciado principalmente por las categorías mayoritarias, como Navegación y VoIP, las cuales poseen una mayor cantidad de muestras dentro del conjunto original. En este contexto, métricas globales como Accuracy pueden favorecer modelos con mejor desempeño sobre clases dominantes, aun cuando las categorías minoritarias presenten mayores tasas de error.

Por otra parte, el escenario balanceado permitió obtener una distribución de errores más uniforme entre categorías, favoreciendo el desempeño sobre clases con menor representación. Este comportamiento fue especialmente visible en métricas orientadas a clasificación multiclase, como F1-Macro y Recall, las cuales presentaron una mayor estabilidad respecto al escenario desbalanceado. Estos resultados evidencian que el balanceo de clases contribuye positivamente a la capacidad de generalización de los modelos cuando existen diferencias importantes en la representación de las categorías.

Asimismo, el análisis de las matrices de confusión evidenció una reducción en los falsos negativos asociados a categorías minoritarias, indicando una mejor capacidad de discriminación entre distintos tipos de tráfico cifrado. Aunque algunos modelos presentaron una ligera disminución en Accuracy global, el balanceo de clases permitió reducir el sesgo hacia categorías dominantes y mejorar el comportamiento general de los clasificadores sobre escenarios multiclase.

Entre los modelos evaluados, Random Forest mostró el comportamiento más consistente en ambos escenarios experimentales, manteniendo los valores más altos de Accuracy y F1-Macro. Este comportamiento sugiere que los métodos de ensamble presentan una mayor robustez frente a variaciones en la distribución de clases y una mejor capacidad para capturar relaciones complejas dentro de las características estadísticas de flujo. Por otra parte, los modelos basados en proximidad, como KNN, mostraron una mayor sensibilidad frente a cambios en la distribución de los datos y al solapamiento estadístico entre categorías.

En términos generales, los resultados obtenidos demuestran que los algoritmos de aprendizaje supervisado son capaces de clasificar tráfico cifrado utilizando únicamente características estadísticas de flujo, sin necesidad de inspeccionar el contenido de los paquetes. Esto respalda la viabilidad del enfoque propuesto como herramienta de apoyo para mecanismos inteligentes de gestión y priorización de tráfico en redes domésticas cifradas.

V. Conclusiones y Trabajo a Futuro

La clasificación de tráfico cifrado representa actualmente uno de los principales desafíos dentro de la gestión de redes modernas, debido a las limitaciones impuestas por protocolos de seguridad como SSL/TLS sobre las técnicas tradicionales de inspección profunda de paquetes. En este trabajo se demostró que es posible identificar distintas categorías de servicios de red mediante el análisis estadístico de flujos y algoritmos de aprendizaje supervisado, sin necesidad de acceder al contenido de los paquetes transmitidos.

Los resultados experimentales mostraron que los modelos supervisados evaluados son capaces de diferenciar categorías de tráfico cifrado utilizando exclusivamente atributos estadísticos asociados al comportamiento temporal y volumétrico de los flujos de red. Entre los modelos analizados, Random Forest presentó el mejor desempeño global, alcanzando una exactitud de 96.67% y un valor F1-Macro de 0.96 bajo el escenario con datos desbalanceados. Asimismo, el análisis comparativo evidenció que el balanceo de clases contribuye a obtener un comportamiento más

ISSN 2395-8618

TABLE V
COMPARACIÓN ENTRE EXPERIMENTOS CON DATOS DESBALANCEADOS Y BALANCEADOS.

Modelo	**Experimento 1**		**Experimento 2**	
	Exactitud	**F1-Macro**	**Exactitud**	**F1-Macro**
Random Forest	**0.9667**	**0.96**	**0.9066**	**0.91**
KNN ($k = 3$)	0.8564	0.82	0.7916	0.79
KNN ($k = 5$)	0.8506	0.80	0.7836	0.78
KNN ($k = 7$)	0.8445	0.79	0.7785	0.78
MLP	0.8447	0.80	0.7765	0.78
SVM (RBF)	0.7088	0.59	0.6040	0.60

uniforme entre categorías y favorece el desempeño sobre clases con menor representación.

Los resultados obtenidos también confirmaron que variables relacionadas con duración del flujo, volumen de transferencia y tiempos entre llegadas de paquetes contienen suficiente capacidad discriminativa para identificar distintos tipos de tráfico cifrado. En conjunto, los hallazgos de esta investigación respaldan la viabilidad del aprendizaje supervisado basado en análisis estadístico de flujos como herramienta de apoyo para mecanismos inteligentes de gestión y priorización de tráfico en redes domésticas.

Aunque los resultados obtenidos fueron favorables, este trabajo presenta algunas limitaciones. El estudio se realizó utilizando un conjunto de datos público previamente etiquetado y no sobre tráfico capturado en tiempo real dentro de una red doméstica operativa. Asimismo, la investigación se centró exclusivamente en tareas de clasificación de tráfico y no incluyó la implementación directa de políticas dinámicas de calidad de servicio sobre dispositivos de red.

Como trabajo futuro, se plantea evaluar arquitecturas más avanzadas de aprendizaje profundo orientadas al análisis temporal del tráfico de red, incluyendo modelos híbridos basados en CNN y LSTM. Asimismo, sería relevante implementar el sistema en un entorno real de red doméstica para analizar su desempeño en tiempo real y evaluar su integración con mecanismos dinámicos de calidad de servicio.

Finalmente, futuras investigaciones podrían incorporar conjuntos de datos más recientes, nuevas categorías de tráfico y técnicas automáticas de selección de características para mejorar la capacidad de generalización de los modelos. También resulta de interés explorar enfoques orientados a sistemas embebidos y dispositivos edge, permitiendo el despliegue de clasificadores inteligentes directamente en routers domésticos y dispositivos de bajo consumo.

REFERENCES

[1] E. Alpaydin, *Introduction to Machine Learning*, 4th ed. MIT Press, 2020.

[2] J. Barcenas-Medina and M. Alcaraz-Cancio, "Hybrid deep learning models for complex network behavior identification," *Sensors*, vol. 23, no. 4, p. 1782, 2023.

[3] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[4] Cisco, "Annual internet report (2018–2023) white paper," 2020.

[5] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[6] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[7] J. G. Delgado, *Clasificación de tráfico mediante análisis estadístico de datos*. Instituto Politécnico Nacional, 2019.

[8] T. Dierks and E. Rescorla, "The transport layer security (tls) protocol version 1.2," 2008, rFC 5246.

[9] A. Elmaghraby *et al.*, "Encrypted traffic classification using machine learning and deep learning techniques," *IEEE Access*, vol. 10, pp. 112 345–112 359, 2022.

[10] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[11] S. Haykin, *Neural Networks and Learning Machines*, 3rd ed. Pearson, 2009.

[12] IETF, "Cisco ios netflow services export version 9," 2004, rFC 3954.

[13] J. F. Kurose and K. W. Ross, *Computer Networking: A Top-Down Approach*, 8th ed. Pearson, 2021.

[14] U. of New Brunswick, "Canadian institute for cybersecurity datasets," 2024, https://www.unb.ca/cic/datasets/.

[15] E. Rescorla, "The transport layer security (tls) protocol version 1.3," 2018, rFC 8446.