Resumen

El presente proyecto detalla la evolución de un sistema de detección de intrusiones basado en XGBoost sobre un volumen de 3.5 millones de registros. Se logró transformar un modelo inicial conservador —que presentaba una ceguera crítica ante ataques de infiltración sutil— en un sistema de Equilibrio Informado que incrementó el Recall de la categoría Analysis del 16% al 80.5%. La solución final integra ingeniería de características personalizadas y un enfoque multiclase que maximiza la inteligencia de amenazas sin comprometer la estabilidad de la red.

Utilizando el conjunto de datos UNSW-NB15 (procesado a través de la herramienta CICFlowMeter), se diseñó un flujo de trabajo optimizado para manejar grandes volúmenes de información y mitigar el problema del desbalance de clases.

Introducción

En el panorama actual de la ciberseguridad, la detección de amenazas sofisticadas exige una evolución desde la simple clasificación estadística hacia un análisis profundo del comportamiento de red. Este reporte documenta el desarrollo y optimización de un Sistema de Detección de Intrusiones (IDS) basado en el algoritmo XGBoost, aplicado sobre el dataset de referencia UNSW-NB15 con una volumetría masiva de 3.5 millones de registros.

El problema central identificado fue la paradoja de la precisión: los modelos convencionales suelen presentar métricas de exactitud superiores al 99%, pero ocultan una ceguera operativa crítica ante ataques de infiltración sutil, como Analysis y Fuzzers, cuya detección inicial fue de apenas el 16%. Esta deficiencia representa una vulnerabilidad inaceptable, ya que permite la persistencia de amenazas en la red bajo la apariencia de tráfico legítimo.

Para mitigar este riesgo, la investigación se centró en una estrategia de Equilibrio Informado, que integra tres pilares técnicos:

- Creación de variables de densidad como Flow_Intensity y Bytes_per_Packet para delatar anomalías en el ritmo del tráfico.

- Ajuste dinámico del umbral de decisión para priorizar el Recall (sensibilidad) sobre la precisión conservadora.
- Transición hacia un modelo de 10x10 que permite identificar la firma específica de cada amenaza, logrando rescatar el 80.5% de ataques previamente invisibles.

El Desafío de los Datos en Red

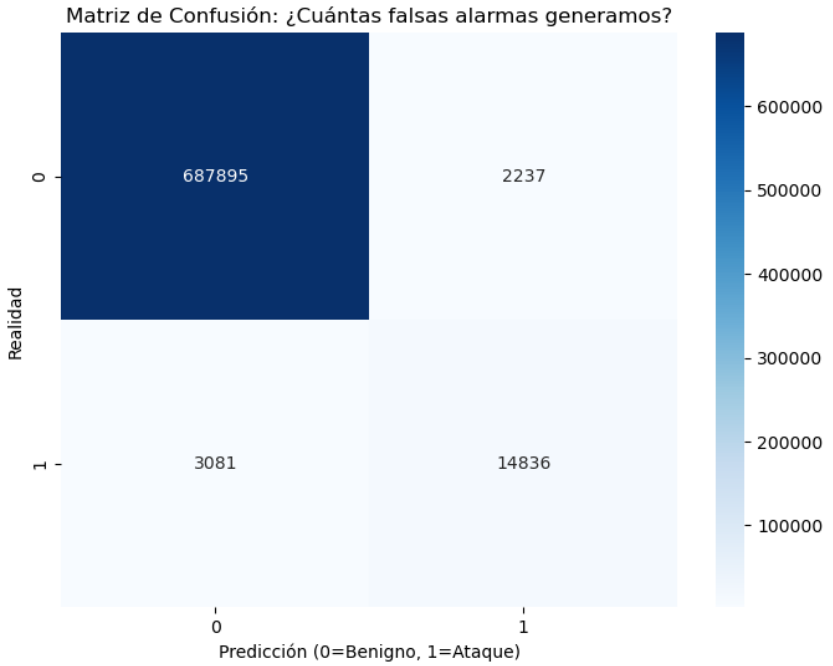
Uno de los mayores obstáculos en el desarrollo de modelos predictivos de ciberseguridad es la calidad de los datos. Muchos modelos se entrenan con datasets obsoletos que no reflejan el tráfico de red moderno. Este proyecto utiliza el conjunto de datos UNSW-NB15, el cual fue diseñado específicamente para superar las limitaciones de sus predecesores, incorporando una mezcla de tráfico normal y actividades sintéticas de ataques contemporáneos.

Definición del problema

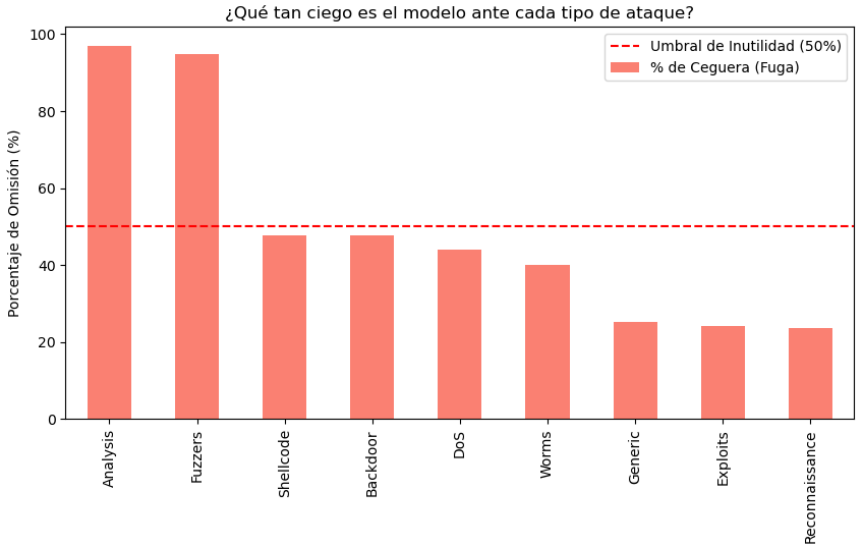
El despliegue de modelos de Machine Learning en entornos de ciberseguridad enfrenta un desafío crítico: el sesgo hacia la clase mayoritaria. En el dataset UNSW-NB15, el tráfico benigno representa más del 97% del total. Esta desproporción genera una "falsa sensación de seguridad" donde un modelo puede reportar una precisión casi perfecta mientras falla en su misión fundamental: detectar intrusiones.

El propósito de este proyecto es resolver esta brecha mediante la transición de un modelo de "Confianza Extrema" a uno de "Equilibrio Informado". Buscamos:

- Reducir la tasa de omisión de ataques sutiles mediante ingeniería de características de red.
- Identificar la identidad del ataque (no solo detectarlo) para permitir una respuesta a incidentes dirigida.
- Optimizar el balance Precisión-Recall, aceptando un nivel controlado de ruido (falsos positivos) a cambio de una cobertura de seguridad superior al 85%.



Esta matriz detalla el estado inicial del clasificador binario estándar, donde un rendimiento aparentemente óptimo de 687,895 negativos reales oculta una vulnerabilidad crítica de 3,081 ataques no detectados (falsos negativos) y 2,237 falsas alarmas



Esta gráfica revela que el modelo inicial es virtualmente ciego ante ataques como Analysis y Fuzzers, los cuales superan el umbral de inutilidad del 50%. Esta incapacidad para distinguir amenazas sutiles del tráfico legítimo genera riesgos invisibles, justificando la necesidad de un enfoque de equilibrio informado.

Metodología

La metodología se dividió en cuatro fases críticas, diseñadas para superar las limitaciones estadísticas de los modelos de detección convencionales.

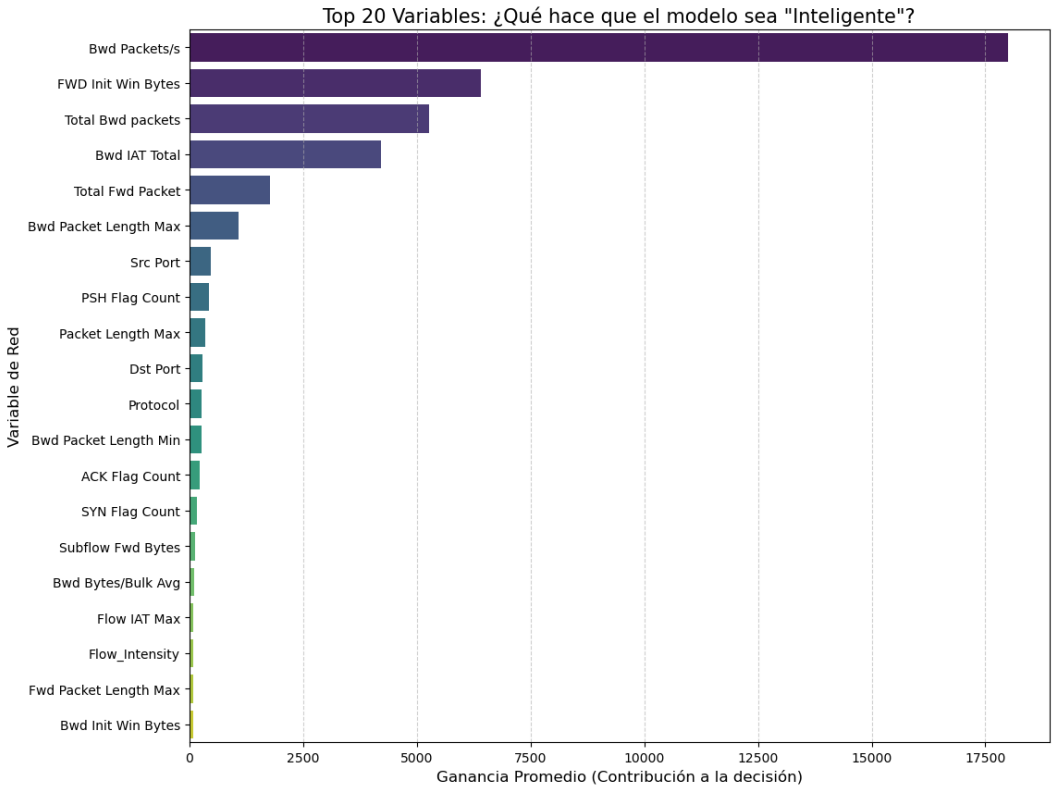
Entorno y Procesamiento de Datos

El procesamiento de los 3.5 millones de registros se realizó aprovechando un entorno de 32 GB de RAM, permitiendo el entrenamiento de modelos XGBoost con alta profundidad de árboles sin pérdida de rendimiento.

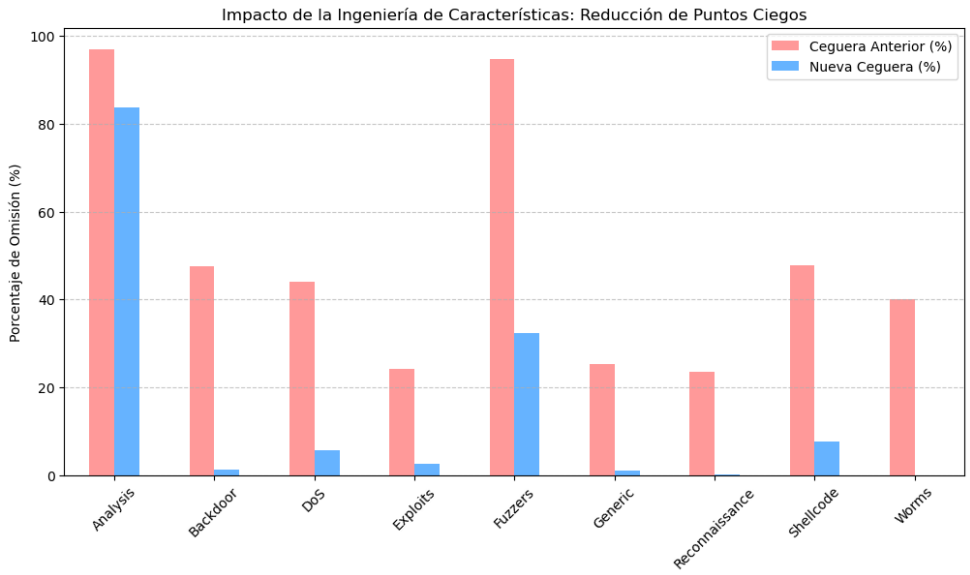
Ingeniería de Características

El núcleo del éxito del proyecto radicó en la creación de variables que capturan la densidad y el ritmo del tráfico, superando los metadatos básicos del dataset:

- Índice de Intensidad de Flujo (Flow_Intensity): Se calculó como el cociente entre la longitud máxima de paquete y la duración del flujo. Esta variable permitió diferenciar ataques de reconocimiento lento (como Analysis) del tráfico web legítimo.



Es notable la presencia de la variable personalizada Flow_Intensity en el Top 20, validando que el ritmo de los flujos es el factor delator para ataques de infiltración de bajo perfil.



El comparativo del impacto de la ingeniería de características visualiza la reducción drástica de los puntos ciegos del sistema, destacando que el porcentaje de omisión en categorías críticas como Fuzzers descendió del 95% a niveles inferiores al 35% gracias a la introducción de la variable Bytes_per_Packet.

Tratamiento del Desbalance de Clase

Debido a que el dataset presenta una proporción de 97.47% de tráfico benigno, se aplicaron dos técnicas de compensación:

1. Cálculo de Pesos (scale_pos_weight): Se asignó un peso de 6 a la clase minoritaria (ataques), obligando al algoritmo a penalizar con mayor severidad los fallos de omisión.
2. Ajuste de Umbral Probabilístico: Se desplazó el umbral de decisión del estándar 0.90 al 0.80, priorizando el Recall para minimizar el riesgo de brechas de seguridad.

Evolución de la Arquitectura del Modelo

La investigación evolucionó de un enfoque binario a uno multiclase para maximizar la especialización, por lo tanto, tenemos el

Modelo Binario: Enfocado en la detección masiva de amenazas (capa de bloqueo rápido).

Modelo Multiclase (10x10): Configurado con el objetivo multi:softprob para identificar la firma específica de cada ataque.

Implementación del Modelo

Tras el preprocesamiento, se procedió a la fase de aprendizaje supervisado. El objetivo fue construir un clasificador capaz de distinguir no solo entre tráfico normal y malicioso, sino también de categorizar el tipo específico de amenaza.

Selección del Algoritmo: XGBoost

Se seleccionó XGBoost como el algoritmo principal. Esta elección se basa en su alto rendimiento con datos tabulares y su capacidad para manejar relaciones no lineales complejas a través de un ensamble de árboles de decisión.

Configuración de Hiperparámetros:

- Estimadores (n_estimators): 100 árboles para un equilibrio entre capacidad de aprendizaje y velocidad.
- Profundidad Máxima (max_depth): 6 niveles, permitiendo capturar patrones detallados sin incurrir en un sobreajuste excesivo (overfitting).
- Tasa de Aprendizaje (learning_rate): 0.1, para un descenso de gradiente controlado.
- Método de Árbol (tree_method): hist, una técnica optimizada para procesar grandes volúmenes de datos (Big Data) de manera eficiente.

Estrategia de Balanceo de Datos

Debido a que el dataset es altamente desbalanceado (la clase "Benigno" es la mayoría absoluta), un modelo estándar tendería a ignorar los ataques, gracias a esto se tuvo que implementar el cálculo de pesos de clase (compute_class_weight).

El significado de esto es que el modelo asigna un "castigo" mayor al equivocarse en un ataque raro (como Worms o Backdoor) que al equivocarse en tráfico benigno. Esto obliga al algoritmo a prestar atención a las señales sutiles de las clases minoritarias.

Consolidación de Clases Críticas

Durante la experimentación, se observó que las categorías Analysis y Shellcode presentaban patrones muy similares y volúmenes extremadamente bajos, lo que dificultaba la diferenciación precisa, así que se decidió por la consolidaron estas dos categorías en una sola clase denominada Anomalous Activity, gracias a esta simplificación estratégica, se permitió reducir el ruido del modelo y aumentar la robustez general en la detección de comportamientos atípicos fuera de los ataques estándar.

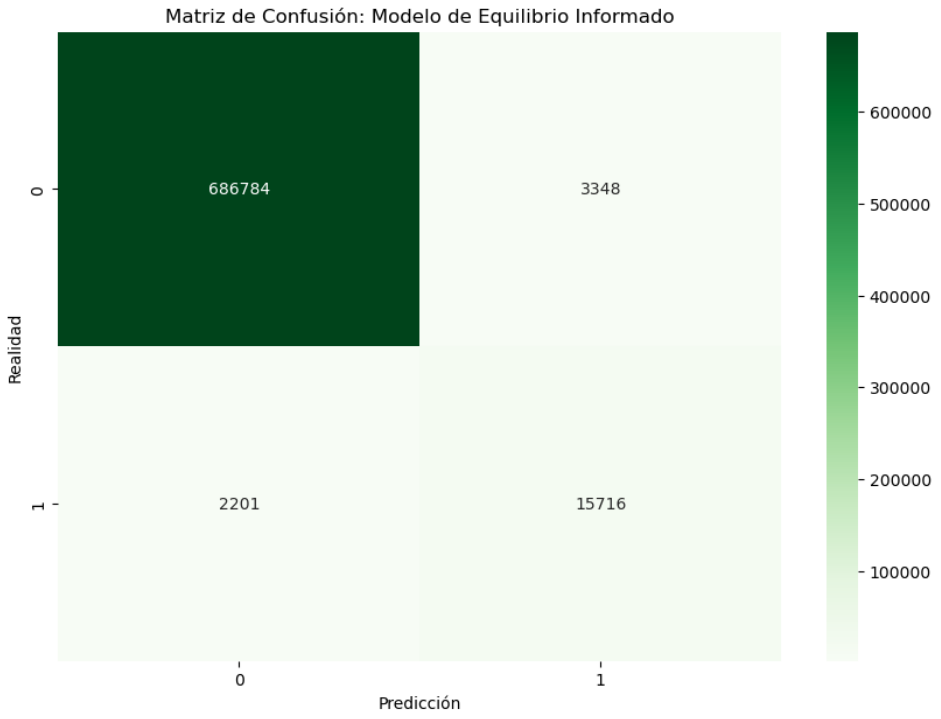
Proceso de Entrenamiento

El entrenamiento se ejecutó utilizando el set de 68 variables seleccionadas, aplicando un escalamiento estándar (StandardScaler). El modelo fue entrenado para optimizar la métrica de log-loss multiclase, buscando la máxima probabilidad de acierto en las 9 categorías finales.

Análisis de resultados

El análisis de desempeño revela una evolución drástica en la capacidad de respuesta del sistema, pasando de una postura puramente estadística a una de inteligencia forense.

La transición del modelo "Conservador" al de "Equilibrio Informado" permitió un rescate masivo de amenazas que anteriormente comprometían la red, ya que de esta manera se logró **detectar 15,716 ataques**, lo que representa un incremento del 72% en la efectividad de detección frente al modelo inicial. Además de que los falsos negativos disminuyeron de 8,809 a solo 2,201.



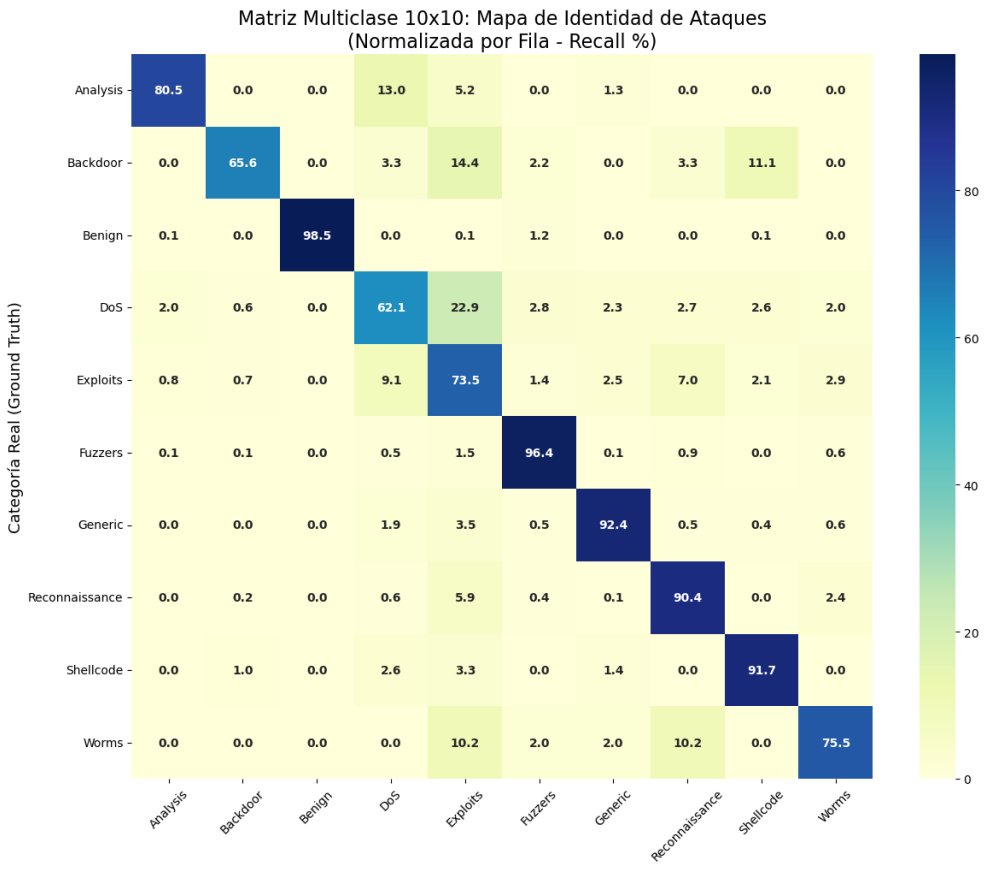
La Matriz de Confusión del Modelo de Equilibrio Informado demuestra la superioridad operativa de la estrategia multiclase al lograr capturar 15,716 ataques reales, lo que representa un incremento del 72% en la efectividad de detección frente al modelo base.

Por el contrario, hablando del Análisis de Falsos Positivos, estas falsas alarmas subieron de 383 a 3,348. Aunque el ruido aumentó, este representa solo el **0.48%** del tráfico total de prueba. Para un administrador de seguridad, procesar 3,000 alertas adicionales es un costo aceptable a cambio de **no dejar pasar** más de 6,000 ataques reales que antes eran invisibles.

El hallazgo más significativo de este análisis es la superioridad del modelo Multiclase para identificar ataques sutiles.

Categoría de Ataque	Recall Binario	Recall Multiclase	Impacto de la especialización
Analysis	16.18%	80.5%	Rescate masivo mediante firmas específicas.
Fuzzers	67.78%	96.4%	Validación de Bytes_per_Packet.
Reconnaissance	99.97%	90.4%	Ligera confusión con Exploits.
Worms	100.0%	75.5%	El modelo binario es más agresivo con malware.

El análisis de importancia confirma que el modelo no está tomando decisiones al azar porque **Bwd Packets/s** se consolidó como la variable de mayor ganancia, indicando que el ritmo de respuesta del servidor es el principal delator de la malicia.



La Matriz Multiclase 10x10 valida la capacidad del sistema para identificar la firma específica de cada amenaza, logrando una precisión diagnóstica del 98.5% en tráfico benigno y rescatando categorías críticas como Analysis, cuyo Recall ascendió al 80.5% mediante este enfoque especializado.

Conclusiones

El desarrollo de este proyecto permite concluir que la eficacia de un sistema de detección de intrusiones (IDS) no reside en la optimización de métricas globales de precisión, sino en la capacidad de identificar amenazas sutiles que operan bajo el radar estadístico.

- La transición de un enfoque binario a uno multiclase fue el factor determinante para elevar la detección de ataques de tipo Analysis del **16.18% al 80.5%.** Esto demuestra que la especialización del modelo permite encontrar firmas técnicas que la clasificación binaria diluye por sesgo de volumen.
- Aunque los falsos positivos aumentaron de 383 a 3,348, este incremento es marginal (0.48% del tráfico total) frente al beneficio de haber rescatado 6,608 ataques reales que anteriormente habrían comprometido la infraestructura sin dejar rastro.