1. Definición del Problema y Recopilación de Datos

Tema Seleccionado:

Detección de Intrusiones y Análisis de Anomalías en Tráfico de Red mediante Técnicas Estadísticas

Preguntas de Investigación

1. ¿Existen diferencias estadísticamente significativas en el comportamiento de variables de flujo de red —como *src_bytes*, *dst_bytes* y *duration*— entre el tráfico normal y los distintos tipos de ataques (DoS, Probe, R2L y U2R)?

2. ¿Es posible reducir la dimensionalidad de las 41 características del tráfico de red mediante el Análisis de Componentes Principales (PCA), conservando al menos el 95 % de la varianza explicada, y cómo impacta esta reducción en la visualización y separación de los distintos tipos de ataques?

3. ¿Qué técnica de clasificación estadística, Regresión Logística o K-Vecinos más Cercanos (K-NN), ofrece una mayor sensibilidad para detectar ataques raros (como U2R) en comparación con ataques volumétricos más comunes (como DoS)?

Justificación de la Relevancia

La seguridad informática constituye uno de los desafíos más críticos en la infraestructura tecnológica actual. Los sistemas tradicionales de detección de intrusiones (IDS), basados en firmas, suelen presentar limitaciones frente a ataques nuevos o modificados, conocidos como *zero-day attacks*.

Este proyecto propone un enfoque basado en el análisis estadístico del comportamiento del tráfico de red, lo cual permite identificar patrones anómalos sin depender de firmas previamente conocidas. Este tipo de aproximación resulta especialmente relevante en entornos dinámicos donde los ataques evolucionan constantemente.

Conjunto de Datos (Dataset)

Nombre: NSL-KDD (Network Security Laboratory - Knowledge Discovery in Databases).

Fuente:

El dataset NSL-KDD es una versión refinada del clásico KDD'99, diseñada para eliminar redundancias y sesgos presentes en el conjunto original. Es ampliamente utilizado como estándar académico para la evaluación de algoritmos de detección de intrusiones. Se encuentra disponible públicamente a través del repositorio del *Canadian Institute for*

Cybersecurity de la Universidad de New Brunswick (UNB), así como en plataformas como Kaggle.

Enlaces:

- NSL-KDD Dataset en Kaggle
- NSL-KDD en el Canadian Institute for Cybersecurity (UNB)

Descripción de las Variables

El conjunto de datos está compuesto por **41 variables predictoras** y **1 variable objetivo (class)**. Las características describen distintas propiedades de cada conexión de red y se agrupan en las siguientes categorías:

- **Características básicas (Basic features):**

  Derivadas directamente de las cabeceras TCP/IP.

  - *duration*: duración de la conexión (numérica).
  - *protocol_type*: tipo de protocolo (categórica: TCP, UDP, ICMP).
  - *service*: servicio de red (categórica: http, telnet, ftp, etc.).
  - *src_bytes* / *dst_bytes*: bytes enviados desde el origen al destino y viceversa (numéricas).
  - *flag*: estado de la conexión (categórica: SF, S0, REJ, etc.).

- **Características de contenido (Content features):**

  Capturan información relacionada con el contenido de la conexión, útiles para detectar comportamientos sospechosos, como:

  - *num_failed_logins*: número de intentos fallidos de autenticación.
  - *root_shell*: indicador de acceso a privilegios de superusuario.

- **Características de tráfico (Traffic features):**

  Estadísticas calculadas sobre ventanas temporales (por ejemplo, 2 segundos), orientadas a detectar patrones como ataques de denegación de servicio (DoS).

  - *count*: número de conexiones al mismo host.
  - *serror_rate*: proporción de conexiones con errores SYN.

Variable objetivo (class):

Indica si la conexión es *normal* o corresponde a un tipo específico de ataque (por ejemplo, *neptune*, *satan*, *smurf*), los cuales pueden agruparse en las categorías DoS, Probe, R2L y U2R.