# Proyecto Final: Modelado de Tráfico de Red y Detección de Anomalías

## Introducción

El análisis del tráfico de red es fundamental para la ciberseguridad, la optimización del rendimiento y la planificación de la capacidad de las redes de computadoras. Diariamente, millones de paquetes de datos viajan a través de estas redes. ¿Se comportan de manera predecible? ¿Podemos modelar su llegada? ¿Cómo distinguimos un comportamiento normal de un posible ataque o una falla del sistema? En este proyecto, utilizarán la teoría de la probabilidad para responder a estas preguntas, actuando como analistas de datos y seguridad.

## Valor y Modalidad

**Ponderación:** 10% de la calificación final del curso. **Modalidad:** Individual.

## Objetivo General

Aplicar los conceptos de variables aleatorias discretas, continuas y conjuntas para modelar un conjunto de datos de tráfico de red, analizar su comportamiento y proponer un criterio simple para la detección de anomalías.

## Objetivos Específicos

**Modelar un fenómeno discreto:** Analizar y modelar el **número de paquetes** que llegan a un servidor en un intervalo de tiempo fijo utilizando una variable aleatoria discreta (por ejemplo, la distribución de Poisson).

**Modelar un fenómeno continuo:** Analizar y modelar el **tiempo entre la llegada de paquetes consecutivos** utilizando una variable aleatoria continua (por ejemplo, la distribución Exponencial).

**Analizar una relación conjunta:** Estudiar la relación entre dos variables, como el **tipo de protocolo** (ej. TCP, UDP) y el **tamaño del paquete**, utilizando el concepto de variables aleatorias conjuntas.

**Aplicar los modelos:** Utilizar los modelos probabilísticos desarrollados para establecer un umbral que permita identificar un comportamiento anómalo en el tráfico de red.

## Metodología y Fases del Proyecto

### Fase 1: Selección y Comprensión del Conjunto de Datos

Para este proyecto, no necesitarán capturar tráfico de red real. Utilizarán un conjunto de datos que se ha preparado para ustedes. Este archivo (network_traffic.csv) contiene información simulada sobre paquetes de red y tiene las siguientes columnas:

- timestamp: El momento exacto en que llegó el paquete (en segundos).
- packet_size: El tamaño del paquete (en bytes).
- protocol: El protocolo de transporte utilizado (**'TCP' o 'UDP'**).

El número **6** siempre corresponde a **TCP**. (Protocolo de Control de Transmisión)

El número **17** siempre corresponde a **UDP**. (Protocolo de Datagramas de Usuario)

**Fase 2: Análisis con Variables Aleatorias**

Deberá usar un lenguaje de programación como **Python** (con librerías como Pandas, NumPy, Matplotlib y SciPy), **R** o cualquier otro lenguaje para realizar el análisis.

1. **Análisis de Variable Discreta (Número de Paquetes):**

   Agrupe los datos en intervalos de 1 segundo. Cuenten cuántos paquetes llegan en cada segundo.

   Calculen la tasa promedio de llegada (λ), que será el número promedio de paquetes por segundo.

   **Hipótesis:** El número de llegadas por segundo sigue una **distribución de Poisson** con el parámetro (λ) que calcularon.

   Generen un histograma de sus datos (frecuencia del número de paquetes por segundo) y superpongan la función de masa de probabilidad de la distribución de Poisson teórica. Discutir qué tan bien se ajusta el modelo.

2. **Análisis de Variable Continua (Tiempo entre Llegadas):**

   Calcular los tiempos entre llegadas consecutivas de paquetes (la diferencia entre los timestamp consecutivos).

   Calculen la tasa promedio de llegada (λ) a partir de estos tiempos (será el inverso del tiempo promedio entre llegadas).

   **Hipótesis:** El tiempo entre llegadas sigue una **distribución Exponencial** con el parámetro (λ).

   Generen un histograma de los tiempos entre llegadas y superpongan la función de densidad de probabilidad de la distribución Exponencial teórica. Discutan el ajuste.

3. **Análisis de Variables Conjuntas (Protocolo y Tamaño):**

   Discretizar el tamaño del paquete en dos categorías: "Pequeño" (≤ 500 bytes) y "Grande" (>500 bytes).

   Crear una **tabla de contingencia** (o tabla de probabilidad conjunta) que muestre las probabilidades de las cuatro combinaciones posibles (TCP y Pequeño, TCP y Grande, UDP y Pequeño, UDP y Grande).

   Calcular:

   - La probabilidad marginal de que un paquete sea 'TCP'.
   - La probabilidad marginal de que un paquete sea 'Grande'.
   - La probabilidad condicional de que un paquete sea 'Grande' **dado que** es 'TCP', .

   Con base en sus cálculos, ¿son los eventos "el protocolo es TCP" y "el tamaño del paquete es Grande" independientes? Justifiquen su respuesta matemáticamente.

   **Fase 3: Detección de Anomalías**

   Basándose en su modelo de Poisson (número de paquetes por segundo), definir una regla simple para detectar una "anomalía". Por ejemplo:

- *Una anomalía se considera cualquier intervalo de 1 segundo en el que el número de paquetes recibidos excede el valor esperado en más de tres desviaciones estándar.*

  Calcular este umbral usando los parámetros de su distribución de Poisson ($\lambda = \mu, \sigma^2 = \lambda$) y verifiquen si alguno de los intervalos en sus datos calificaría como anómalo según su regla.

## Que debes entregar:

Deberá entregar dos archivos en un único comprimido .zip:

1. **Video:** El enlace (linck) de un video bien explicado sobre todo el código utilizado para el análisis, desde la carga de los datos hasta la generación de las gráficas y los cálculos.
2. **Informe Técnico (PDF):** Un documento de máximo 4 páginas con la siguiente estructura:
   - **Portada:** Con su nombre, cédula y título del proyecto.
   - **Introducción:** Breve descripción del problema y los objetivos.
   - **Análisis Discreto:** Explicación del modelo de Poisson, presentación de la gráfica comparativa y discusión sobre la calidad del ajuste.
   - **Análisis Continuo:** Explicación del modelo Exponencial, presentación de la gráfica comparativa y discusión sobre la calidad del ajuste.

- **Análisis Conjunto:** Presentación de la tabla de probabilidad conjunta y respuesta a las preguntas sobre independencia y probabilidades condicionales.
- **Detección de Anomalías:** Definición de su regla, cálculo del umbral y resultados.
- **Conclusiones:** Reflexión final sobre la utilidad de los modelos probabilísticos en este contexto.

## Criterios de Evaluación

| Criterio | Ponderación |
|---|---|
| **Claridad de la explicación en el video** | 30% |
| **Aplicación Correcta de Conceptos Probabilísticos** | 40% |
| **Calidad del Informe y Discusión de Resultados** | 30% |
| **Total** | **100%** |

## Fecha de Entrega

- **Fecha Límite:** domingo, 19 de octubre de 2025.
- **Hora Límite:** 11:59 am.

**Medio:** Por correo electrónico

Estudiantes del prof. Luis Rodríguez: larodri@uc.edu.ve
Estudiante de la profesora Mirba Romero: romeromirba@gmail.com