Análisis inferencial de tráfico en servidores web

Mitsiu Alejandro Carreño Sarabia
E23S-18014

ANTECEDENTES

Se cuentan con registros de conexiones a un servidor web el cual aloja 40 dominios, cada conexión cuenta con 21 características entre las que destacan:

	remote_addr	date_time	req_uri	status	body_bytes_sent
0	185.213.174.190	[27/Jun/2023:07:12:12 -0600]	/	502.0	575.0
1	185.213.174.190	[27/Jun/2023:07:12:12 -0600]	/index.php?s=/index/think%5Capp/invokeMethod&method[0]=think%5Cview%5Cdriver%5CPhp&method[1]=display&vars[0]=%3C?php%20echo%20md5(%271f3870be274f6c49b3e31a0c6728957f%27);	502.0	575.0
2	185.213.174.190	[27/Jun/2023:07:12:13 -0600]	/index.php?s=/admin/think%5Capp/invokeMethod&method[0]=think%5Cview%5Cdriver%5CPhp&method[1]=display&vars[0]=%3C?php%20echo%20md5(%271f3870be274f6c49b3e31a0c6728957f%27);	502.0	575.0
3	185.213.174.190	[27/Jun/2023:07:12:14 -0600]	/index.php?s=/api/think%5Capp/invokeMethod&method[0]=think%5Cview%5Cdriver%5CPhp&method[1]=display&vars[0]=%3C?php%20echo%20md5(%271f3870be274f6c49b3e31a0c6728957f%27);	502.0	575.0
4	185.213.174.190	[27/Jun/2023:07:12:14 -0600]	/index.php?s=/home/think%5Capp/invokeMethod&method[0]=think%5Cview%5Cdriver%5CPhp&method[1]=display&vars[0]=%3C?php%20echo%20md5(%271f3870be274f6c49b3e31a0c6728957f%27);	502.0	575.0

JUSTIFICACIÓN

El análisis de peticiones ofrece grandes beneficios

Entender el uso real de las plataformas (insights)

- Adaptar a las necesidades reales
- Toma de decisiones de desarrollo basada en datos

Detectar comportamiento anómalo

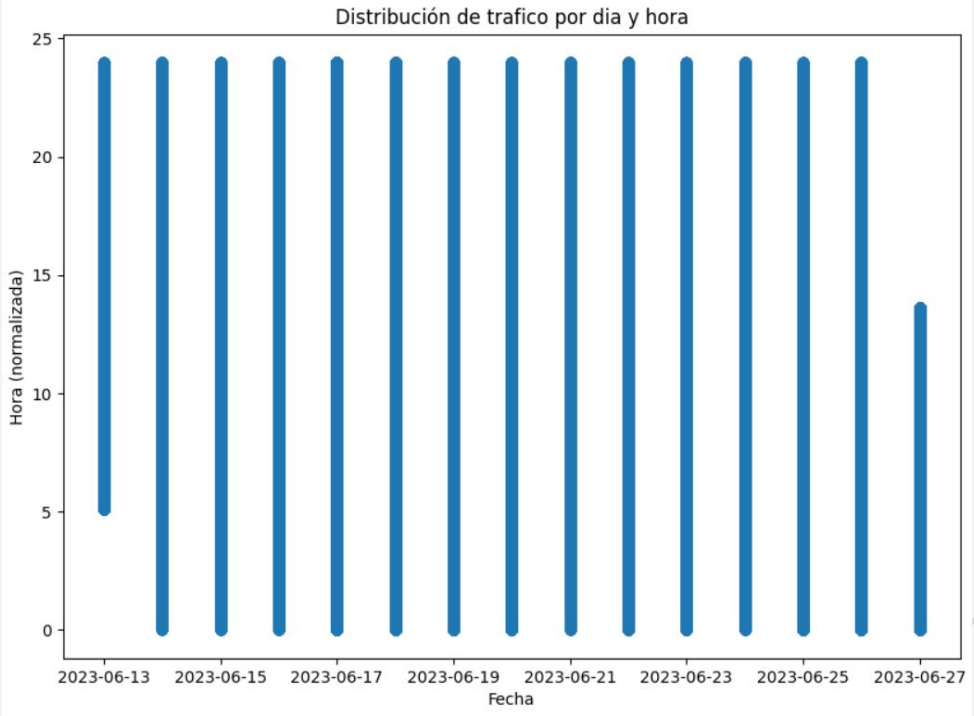
- Detección de servicios caídos
- Ataques de denegación de servicio
- Conexiones anómalas y/o maliciosas

Escalar apropiadamente la infraestructura

- Dado las arquitecturas infrastructure as a service (IaaS), es importante tener sólo las prestaciones necesarias

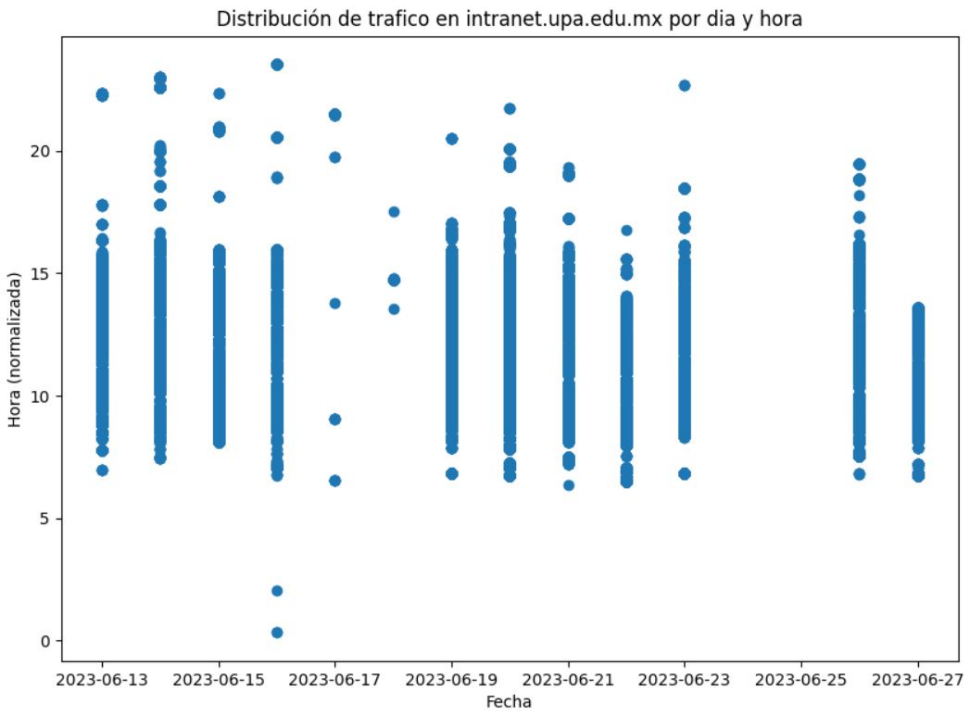
¿Existen tiempos muertos (sin conexiones) en el servidor?

El servidor se encuentra en constante actividad, 24/7, lo cual hace sentido porque maneja múltiples dominios



¿Existen tiempos muertos (sin conexiones) en el dominio intranet.upa.edu.mx?

Pero, esta tendencia es más interesante, analizemos un poco más a fondo

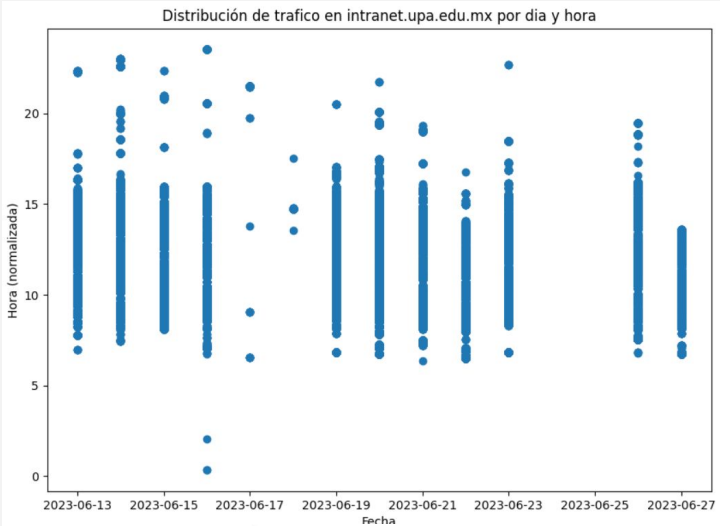


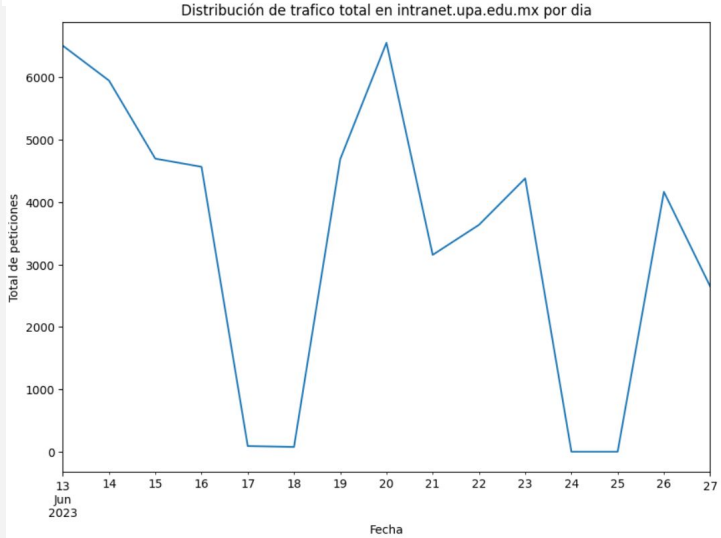
¿Existen tiempos muertos (sin conexiones) en el dominio intranet.upa.edu.mx?

La actividad se concentra entre las 7 am y las 5 pm

Sábados y domingos casi nula actividad

intranet.upa.edu.mx es un dominio con fines educacionales y su actividad está estrechamente relacionada con la actividad en la institución educativa.

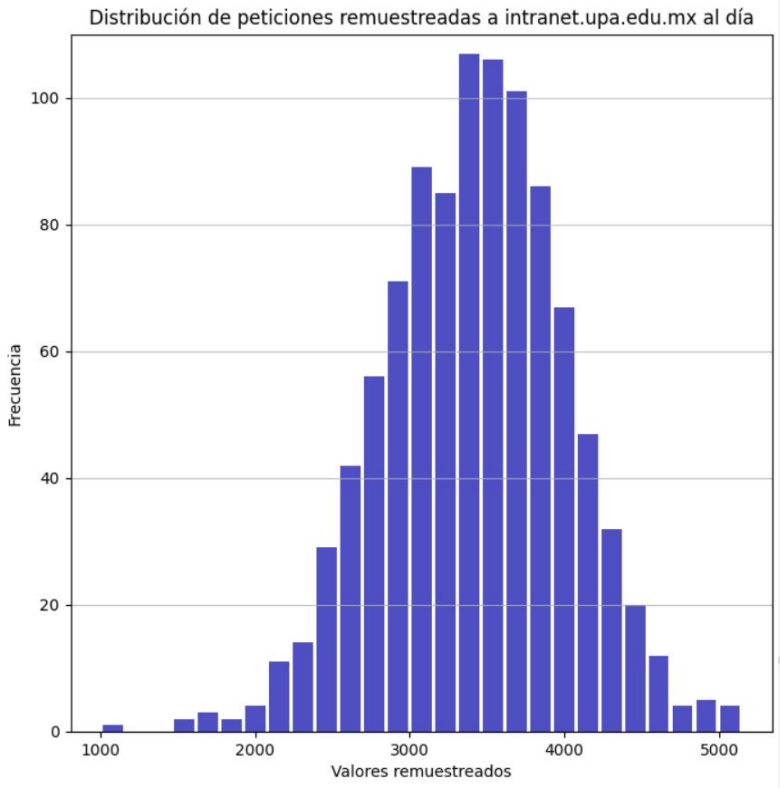




¿Cuáles son los rangos esperados de carga para el dominio intranet.upa.edu.mx?

Solo contar con 15 días es una limitante, aplicando remuestreo e intervalo de confianza del promedio al 95% es posible determinar los rangos de operación normal del dominio.

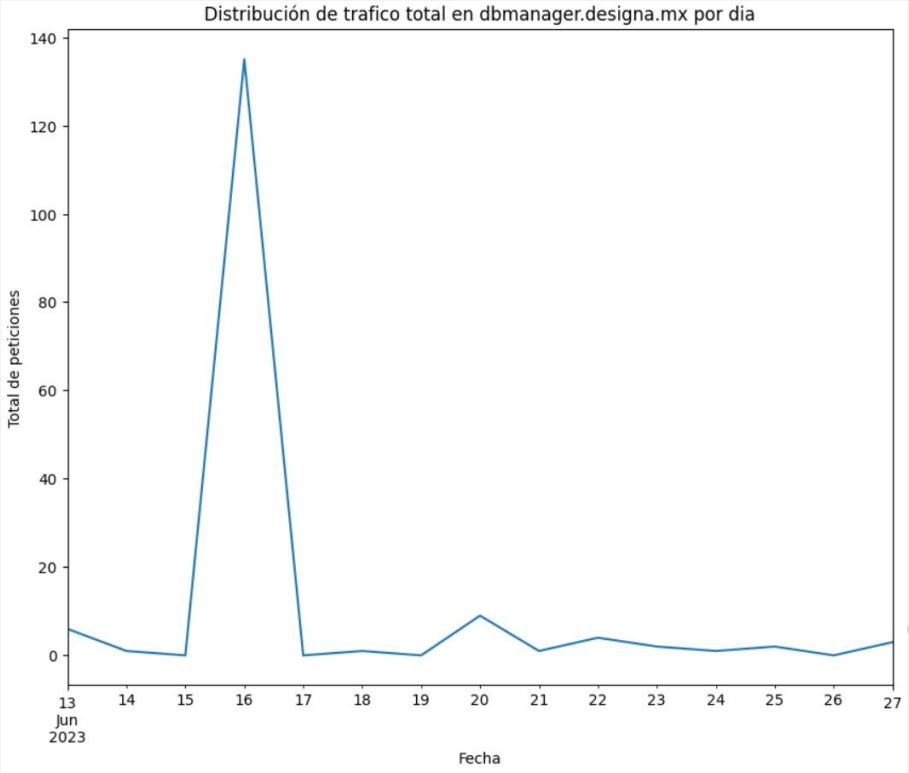
Esto permite establecer límites que se pueden **monitorear y alertar** en caso de que se sobrepasen, también puede tener un impacto en la **inversión de infraestructura**



¿Cuál es la cantidad promedio esperada de peticiones a un dominio en específico?

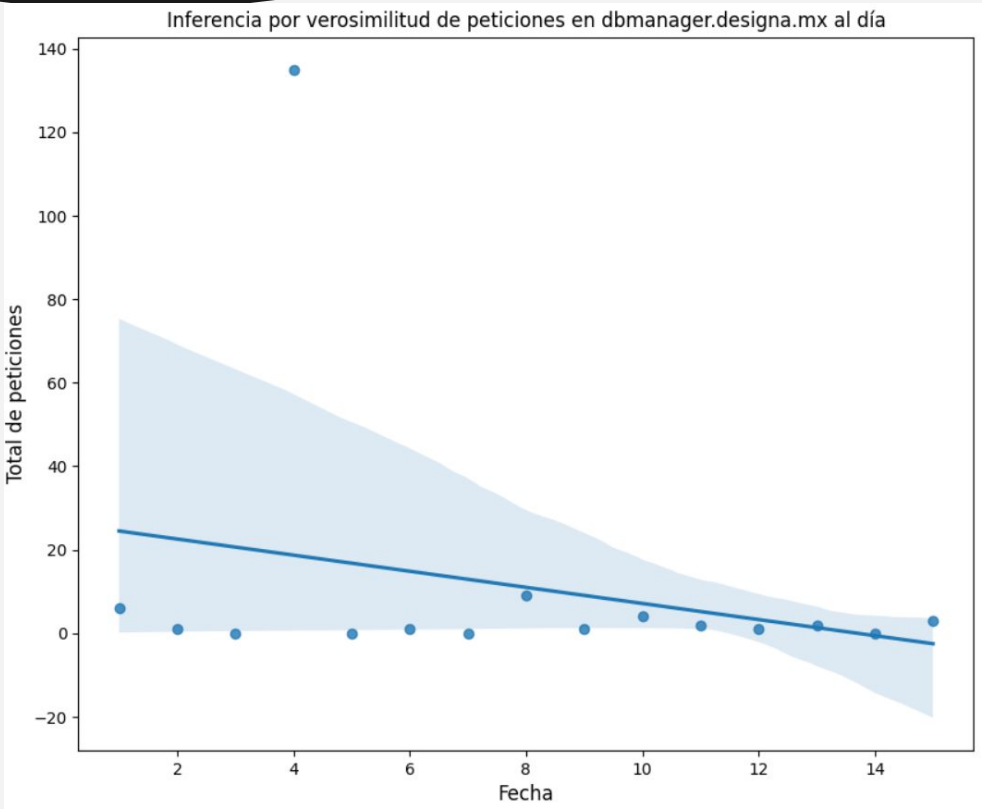
Cambiando de dominio a “dbmanager.designa.mx” notamos un comportamiento muy diferente, y tenemos un offset.

Si queremos obtener un valor puntual (promedio) podemos aplicar la inferencia por verosimilitud.



¿Cuál es la cantidad promedio esperada de peticiones a un dominio en específico?

De esta manera obtenemos una regresión lineal que nos permite estimar un valor exacto para un valor x dado.



CURVAS ROC

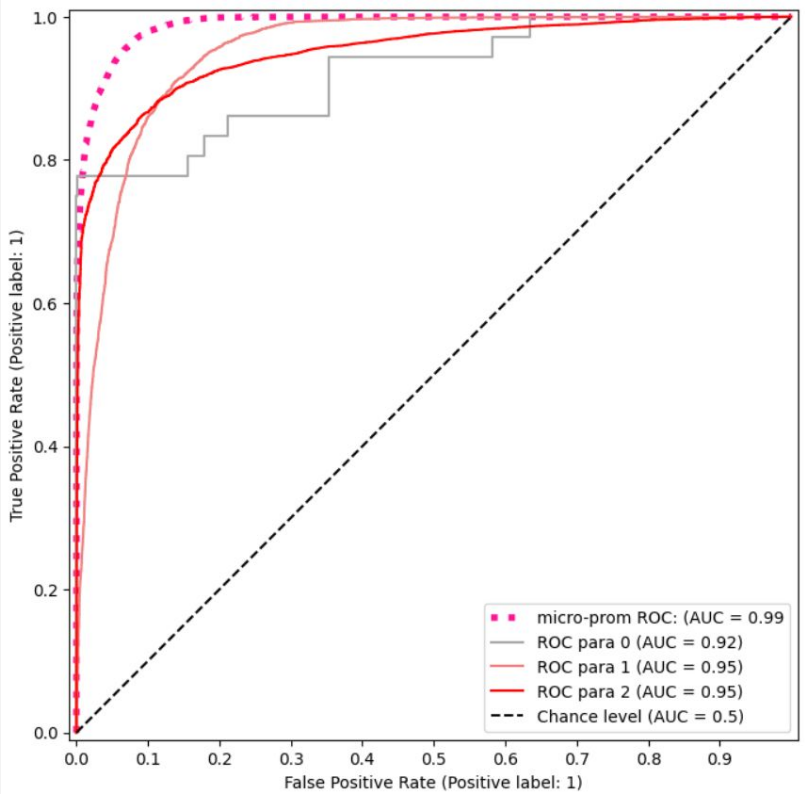
Las curvas ROC nos permiten evaluar clasificadores

Red neuronal

“status”,
“body_bytes_sent”,
“fabstime”, y
“req_method”

Dominios comparados

- moodle.upa.edu.mx
- dbmanager.designa.mx
- intranet.upa.edu.mx



CONCLUSIONES

01

Patrones y tendencias

Al analizar la distribución del servidor en general y de un dominio en específico, fue posible validar que a pesar de que el servidor constantemente está procesando y contestando conexiones todo el tiempo, **cada dominio tiene sus propios patrones de uso**.

02

Intervalos de confianza

Permiten establecer **rangos de operación** normal así como **planes de contingencia** cuando dichos límites se superan.

03

Inferencia por verosimilitud

Podemos generar un valor exacto a pesar de solo contar con una muestra de la población, esto puede ser útil en reportes o cálculos donde se requiere un valor preciso comparado con los intervalos de confianza que nos ofrecen rangos.

04

Curvas ROC

Nos permiten evaluar clasificadores, ya sea para entender de manera granula **cuales son las categorías que mejor y peor clasifica**, esto puede emplearse para mejorar el algoritmo clasificador o realizar investigaciones adicionales sobre las características específicas o grupales de las categorías peor clasificadas