UNIVERSIDAD DE LA CIUDAD DE AGUASCALIENTES

MAESTRÍA EN CIENCIA DE DATOS

INFERENCIA ESTADÍSTICA PARA CIENCIA DE DATOS

“Análisis inferencial de tráfico en servidores web”

Alumno:

E23S-18014: MITSIU ALEJANDRO CARREÑO SARABIA

Periodo Enero 2024 - Junio 2024, Aguascalientes, Ags

Resumen

En el presente trabajo se realizan diversos análisis de la rama de la estadística inferencial a datos de conexiones y solicitudes a servidores web reales que alojan un conjunto de sitios disponibles desde internet. El propósito de este análisis es evaluar ciertas estimaciones y aproximaciones a características de interés como, ¿Cuál es la cantidad de tráfico que recibe un sitio usualmente? ¿Cuáles son las horas de mayor y menor demanda? ¿Cuál es el rango estimado de peticiones usualmente? Contestando a estas preguntas se espera tener una visión más clara del comportamiento de los clientes así como de las dinámicas involucradas en procesar y ofrecer determinados servicios y contenido en internet.

Introducción

Mediante la estadística inferencial es posible obtener generalizaciones, y aproximaciones sobre una población a partir de información obtenida de una muestra de la misma. Para ello se aplican diversas técnicas estadísticas a través de las cuales se estiman valores desconocidos intentando que estas sean precisas y confiables para que sean útiles en la toma de decisiones y predicciones basadas únicamente en la muestra representativa de la población en lugar de medir todos los elementos de la población.

A pesar de que la estadística inferencial es un herramienta versátil, aplicable a un amplio rango de campos y estudios, su uso usualmente está enfocado en investigaciones y predicciones de carácter social (predicciones basadas en encuestas, obtener estadísticas de una población, o en procesos democráticos) pero también es posible aplicarlo al análisis computacional [1].

En el siguiente estudio se analizaron bitácoras de conexión a servidores web conectados a internet, cabe mencionar que el servidor analizado provee recursos a 40 dominios distintos, y los datos comprenden un periodo de 15 días, a pesar de ello se lograron registrar más de un millón de conexiones, por cada conexión se guarda la siguiente información:

- ***remote_addr:*** Dirección IPv4 del cliente que inició la conexión.
- ***remote_usr:*** Nombre de usuario (solo aplica si el usuario está autenticado).
- ***date_time:*** Hora y fecha en formato “[dd-MM-YYYY:HH:MM:SS-timezone]”.

- ***date:*** Fecha en formato “dd/MM/YYYY”.
- ***time:*** Hora en formato “HH:MM:SS”.
- ***request:*** Descripción del recurso solicitado por el cliente.
- ***req_method:*** Método HTTP mediante el cual se solicitó el recurso.
- ***req_uri:*** Dirección URI del recurso solicitado.
- ***http_ver:*** Versión HTTP bajo la que se estableció la conexión cliente-servidor.
- ***status:*** Código de estatus HTTP que resolvió el servidor.
- ***body_bytes_sent***: Cantidad de bytes enviados en la respuesta.
- ***http_referer:*** Valor de la cabecera http_referer con el que se conectó el cliente.
- ***user_agent:*** Valor de la cabecera user_agent con el que se conectó el cliente.
- ***dec_req_uri:*** Es una réplica decodificada del valor req_uri.
- ***clean_path:*** Filtrado del valor req_uri únicamente con el recurso solicitado (sin parametros).
- ***clean_query_list:*** Listado de python de los parámetros query.
- ***domain:*** Nombre del dominio listado en el campo http_referer.
- ***fdate:*** Valor date pero en tipo de dato fecha.
- ***dateunixtimest:*** Valor time pero en formato unixtimestamp.
- ***ftime:*** Valor time pero en formato estandar.
- ***fabstime:*** Valor time transformada en decimal.
- ***fdatetime:*** Valor date_time en formato estandar.

Al dataset se le aplicaron transformaciones de preprocesamiento en el cuál se expandieron los datos string iniciales además de generar múltiples formatos para valores como fecha y hora entre otros, en la sección de anexos 1, hay un pequeño extracto de los datos analizados.

El origen de dicho tráfico puede ser generado por peticiones de usuarios reales, peticiones de bots, peticiones automatizadas y peticiones de usuarios con intenciones maliciosas cada uno de estos grupos de clientes se comportan de maneras muy específicas, por lo que se considera una área de oportunidad el

analizar estos datos para reconocer y clasificar los usuarios que interactúan con los servicios ofrecidos [2].

Cabe destacar que el tráfico de servidores web tiene claras tendencias como recursos solicitados, región geográfica de donde se solicita, hora en que se solicitó, cantidad de bytes enviados, por lo que identificar las tendencias y realizar predicciones es información relevante para gestionar de manera adecuada los recursos que requiere el servidor para su correcto funcionamiento, bajo un análisis estadístico se permite examinar la información desde múltiples dimensiones y perspectivas.

Como consecuencia del la facilidad, conveniencia y accesibilidad a servicios en internet, cada vez se genera más información y se establecen tendencias de confiar en alojar datos personales y sensibles en servidores de terceros, lo que aumenta la relevancia de evaluar qué y cómo se están accediendo a los recursos solicitados, así como desarrollar herramientas que faciliten filtrar posibles anomalías para tomar acciones correctivas.

Desarrollo

Dado que el servidor es compartido entre 40 dominios distintos es posible realizar tanto análisis del servidor completo como segmentado por dominio, esta cualidad se aprovechó para analizar al servidor como un ente y a cada dominio de manera individual según la pregunta que se intente responder.

Primero se contestó la pregunta, ¿existen tiempos muertos (sin conexiones) en el servidor?, para ello se realizó un conteo de conexiones por día y hora, anexo 2, como se puede apreciar, el servidor está contestando conexiones de manera permanente.

Al tener un comportamiento de actividad tan homogéneo, se adoptó el enfoque a un solo dominio (intranet.upa.edu.mx) y se realizó la misma pregunta, ¿existen tiempos muertos en el servidor? La gráfica del anexo 3 y 4 nos muestra una tendencia de horario de oficina, de lunes a viernes de 7 am a 5 pm aproximadamente, lo cual hace sentido ya que el dominio intranet.upa.edu.mx es un dominio con fines educacionales y su actividad está estrechamente relacionada con la actividad en la institución educativa.

Otra pregunta que se desea responder es ¿Cuáles son los rangos esperados de carga para el dominio intranet.upa.edu.mx? Para ello primero fue necesario realizar un remuestreo con reemplazo aplicando la técnica Bootstrap intentando solventar el poco rango temporal (15 días) que se contaba originalmente de los datos. Una vez remuestreado, fue posible calcular un intervalo de confianza, en este caso definido en 95%, anexo 5, de esta manera es posible establecer una serie de parámetros que delimiten la actividad normal de la anormal. En el caso específico del dominio intranet.upa.edu.mx, se puede considerar como carga normal entre 1,500 y 5,000 peticiones diarias.

Para la pregunta ¿cuál es la cantidad promedio esperada de peticiones a un dominio en específico? se empleó la técnica de inferencia por verosimilitud, para este estudio se decidió analizar el dominio dbmanager.designa.mx el cuál tiene un comportamiento muy distinto a intranet.upa.edu.mx, anexo 6, el cual afortunadamente durante el lapso capturado tuvo un valor offset, el día 16 de junio de 2023 recibió una cantidad importante de actividad para después regresar a sus valores normales. A través de la inferencia por verosimilitud se obtuvieron los siguientes datos:

Dep. Variable:	sum	**R-squared:**	0.063
Model:	OLS	**Adj. R-squared:**	-0.009
Method:	Least Squares	**F-statistic:**	0.8722
Date:	Sat, 15 Jun 2024	**Prob (F-statistic):**	0.367
Time:	08:27:53	**Log-Likelihood:**	-73.349
No. Observations:	15	**AIC:**	150.7
Df Residuals:	13	**BIC:**	152.1
Df Model:	1		
Covariance Type:	nonrobust		

Omnibus:	35.696	**Durbin-Watson:**	2.364
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	68.055
Skew:	3.079	**Prob(JB):**	1.67e-15
Kurtosis:	11.425	**Cond. No.**	19.3

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	26.4286	18.776	1.408	0.183	-14.134	66.991
mdate	-1.9286	2.065	-0.934	0.367	-6.390	2.533

Al graficarlos obtenemos el resultado del anexo 7, en el que a pesar de haber tenido un valor offset se establece una regresión lineal que captura de manera aceptable (R2 = 0.063) la tendencia en los datos y nos muestra un umbral en el que es más probable que se encuentre el dato real.

Finalmente después de haber analizado varios dominios de manera individual, aprovechamos tener información de un servidor que resuelve múltiples dominios para construir una pequeña red neuronal que tomando como input los datos “status”, “body_bytes_sent”, “fabstime”, y “req_method” intenta clasificar si el dominio al que corresponde la conexión es “moodle.ucags.edu.mx”, “dbmanager.designa.mx” o “intranet.upa.edu.mx”, una vez que se completó el predictor se calculó la curva ROC y el área bajo la curva, anexo 8, con ello podemos evaluar tanto el desempeño general del clasificador como la habilidad de clasificar cada uno de los dominios, esta herramienta nos permite ajustar el modelo del predictor hasta alcanzar los niveles de precisión deseados.

Conclusión

Al analizar la distribución del servidor en general y de un dominio en específico, anexos 2 y 3 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.

Al realizar el intervalo de confianza como el mostrado en el anexo 5, es posible establecer parámetros de operación esperados en el servidor, y ajustar la infraestructura acorde, un posible beneficio de tener esta información es abaratar costos al prevenir invertir en más infraestructura de la necesaria. Además, bajo este cálculo es posible detectar actividad anómala, principalmente la sobresaturación del servidor, actividad que puede deberse a un ataque cibernético conocido como ataque de denegación de acceso.

La inferencia por verosimilitud nos permite hacer estimaciones exactas a valores desconocidos, lo cuál permite establecer métricas exactas a pesar de no contar con toda la información de la población completa, estas métricas se pueden emplear para reportes o cálculos en donde se requiere un valor preciso comparado con los intervalos de confianza que nos ofrecen rangos.

Con el análisis ROC y de área bajo la curva, anexo 8, se obtiene información importante para evaluar el desempeño de un algoritmo clasificador, no solo de manera general, sino que nos permite también 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.

Finalmente se debe mencionar que las herramientas y técnicas de estadística inferencial nos permite trabajar con muestras que en términos prácticos es usualmente con lo que contamos.

Este estudio se centró en ciertos dominios específicos, por cuestiones de tiempo principalmente, pero es posible expandirlo para analizar cada dominio de manera individual, y evaluar qué tan distintos son entre sí, si existe algún tipo de comportamiento compartido cuando ambos dominios pertenecen al mismo ramo (educación, entretenimiento, servicio, gubernamental, etc) además quedan varios análisis que se pueden realizar al servidor en general, para entender el comportamiento no a nivel dominio, sino a nivel servidor, lo cuál también tiene el potencial de ofrecernos perspectivas interesantes sobre su uso.

Referencias

1. Ethem Utku Aktas, Mehmet Cagri Calpur, Umit Ulkem Yildirim, and Emrah Yıldırım “Inferring Dependencies Among Web Services with Predictive and Statistical Analysis of System Logs” https://www.researchgate.net/profile/Ethem-Aktas/publication/329370936_Inferring_Dependencies_Among_Web_Services_with_Predictive_and_Statistical_Analysis_of_System_Logs/links/5c05248892851c6ca1f9c83b/Inferring-Dependencies-Among-Web-Services-with-Predictive-and-Statistical-Analysis-of-System-Logs.pdf
2. Wanchun Li, Ian Gorton “Analyzing Web Logs to Detect User-Visible Failures” https://www.usenix.org/legacy/event/slaml10/tech/full_papers/Li.pdf

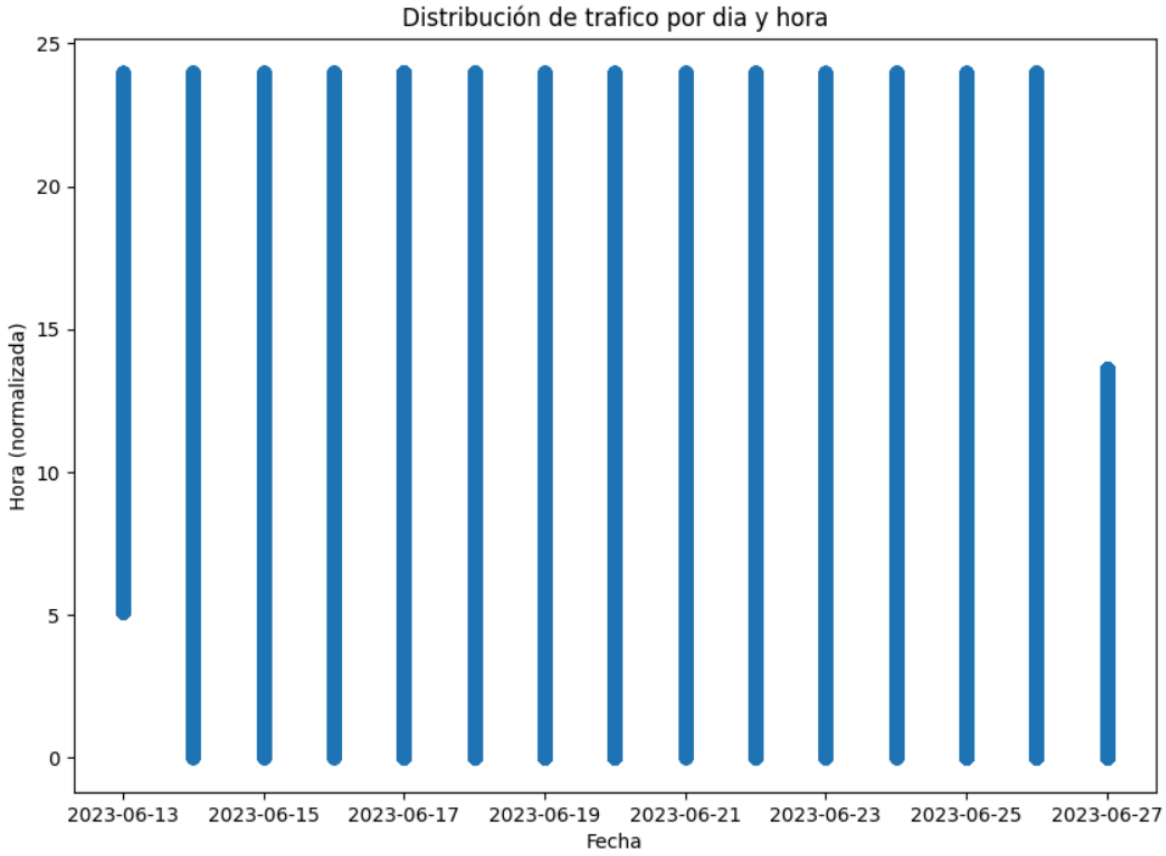
Anexos

1. Extracto de datos usados en el estudio

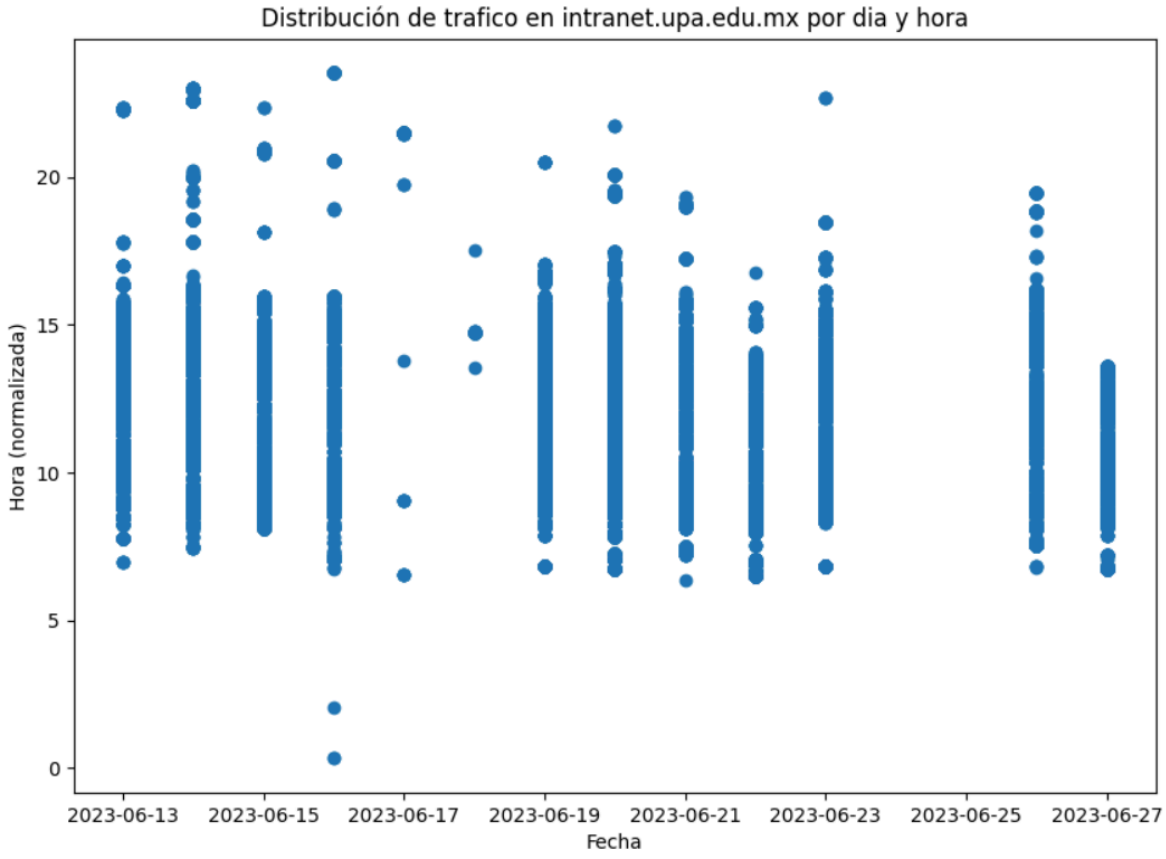
remote addr	date	time	req_uri
185.213.174.190	27/Jun/2023	07:12:12	/index.php?s=/index/think\app/invokeMethod&me
201.182.22.38	20/Jun/2023	07:14:17	/upa.php/fotos/by_matricula_min/UP200062

method	status	user_agent	domain	fabstime
GET	502	Mozilla/5.0 (Windows NT 10.0; Win64; x64)		7.2
GET	200	Mozilla/5.0 (Windows NT 10.0; Win64; x64;	intranet.upa.edu.mx	7.233333

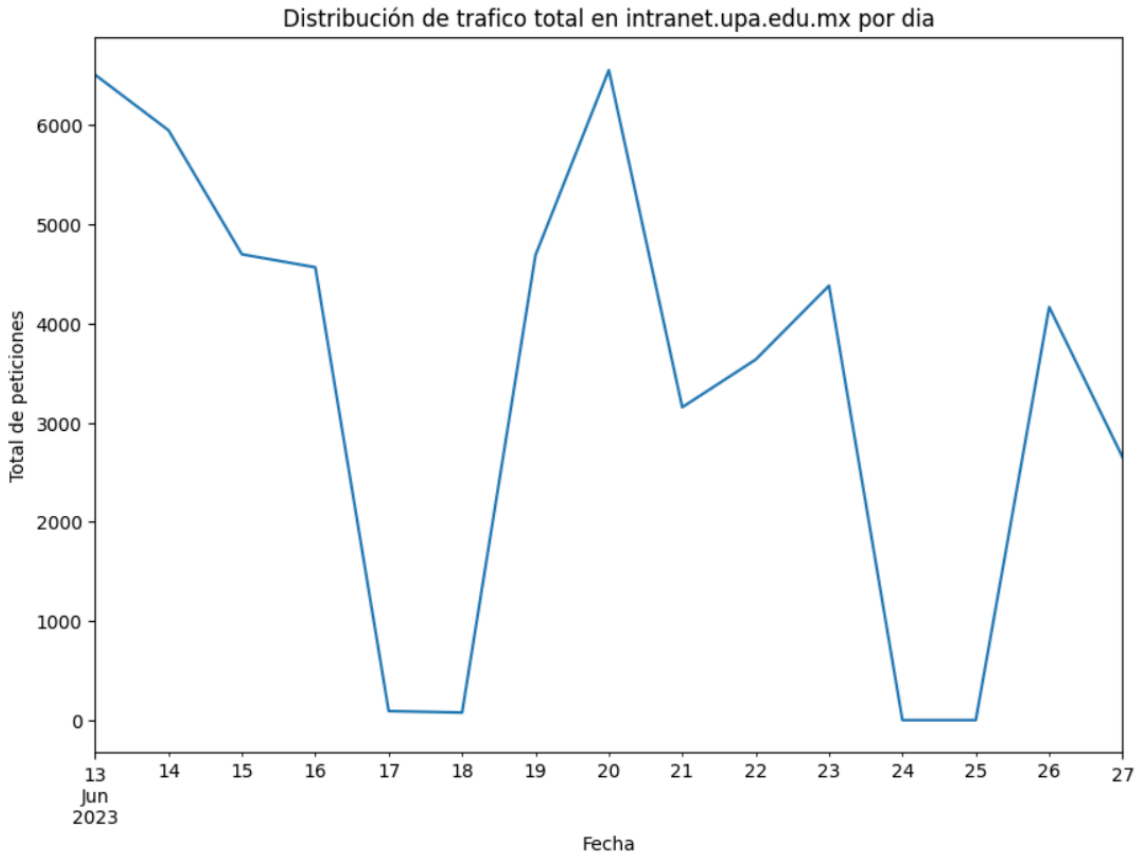
2. Distribución de tráfico en todos los dominios por hora y día



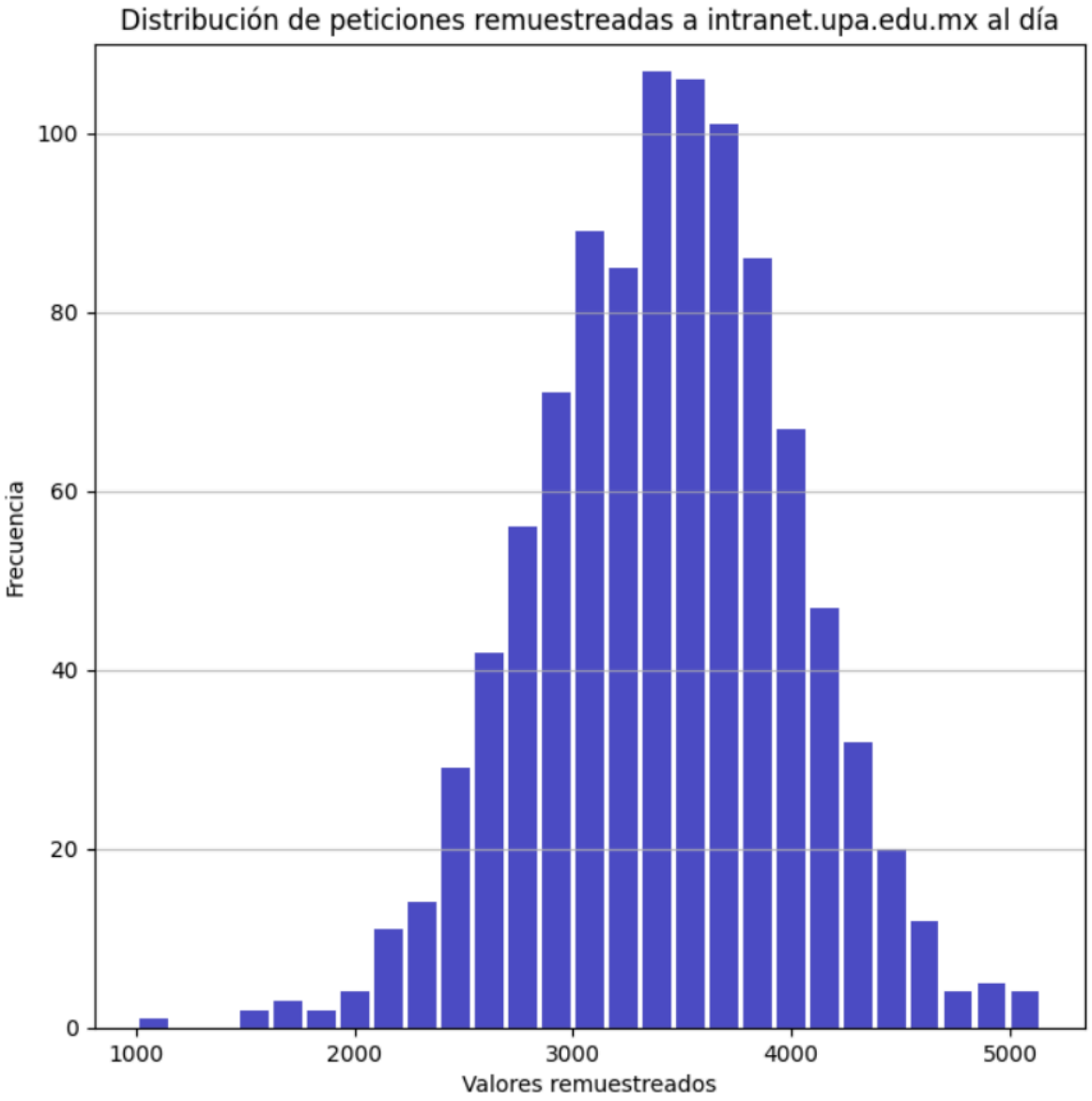
3. Distribución de tráfico en “intranet.upa.edu.mx” por hora y día



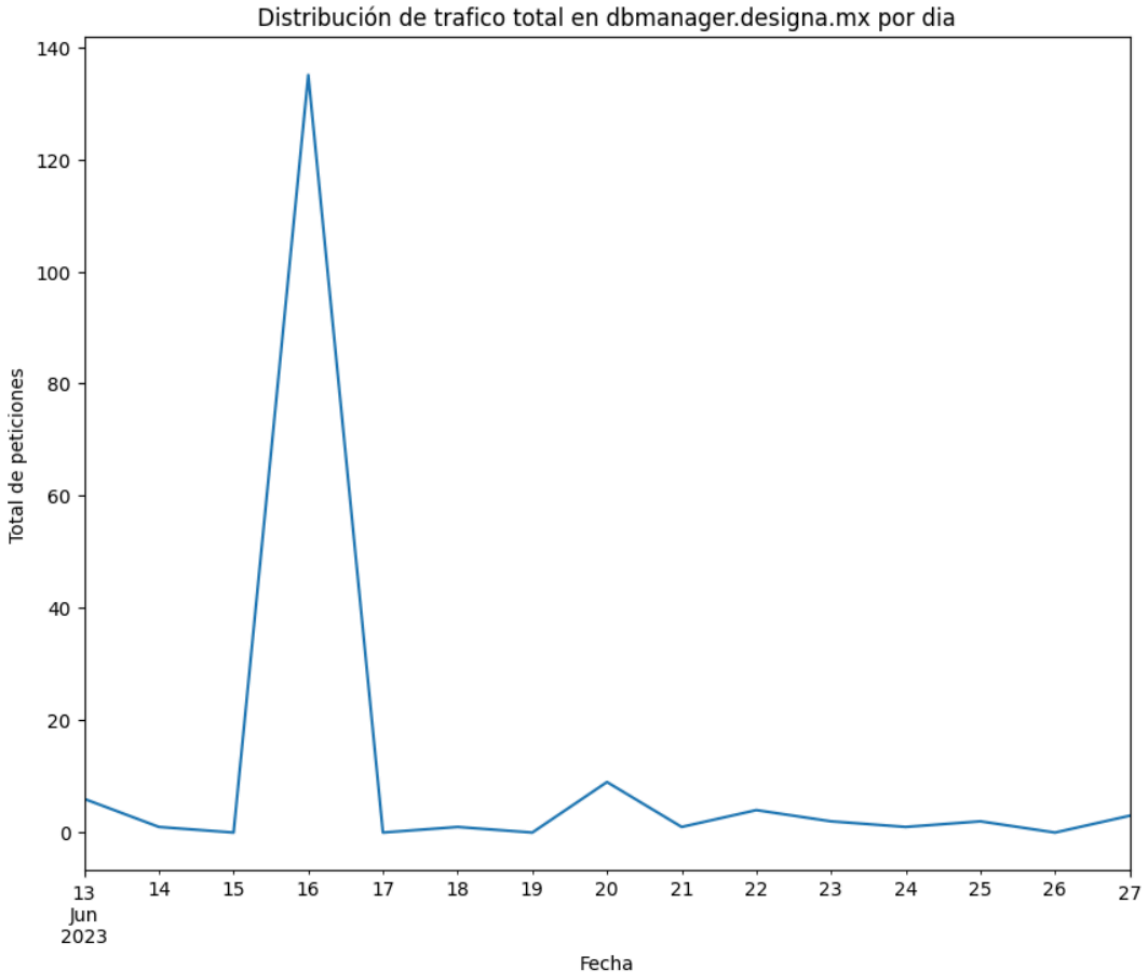
4. Distribución de trafico total en intranet.upa.edu.mx por dia



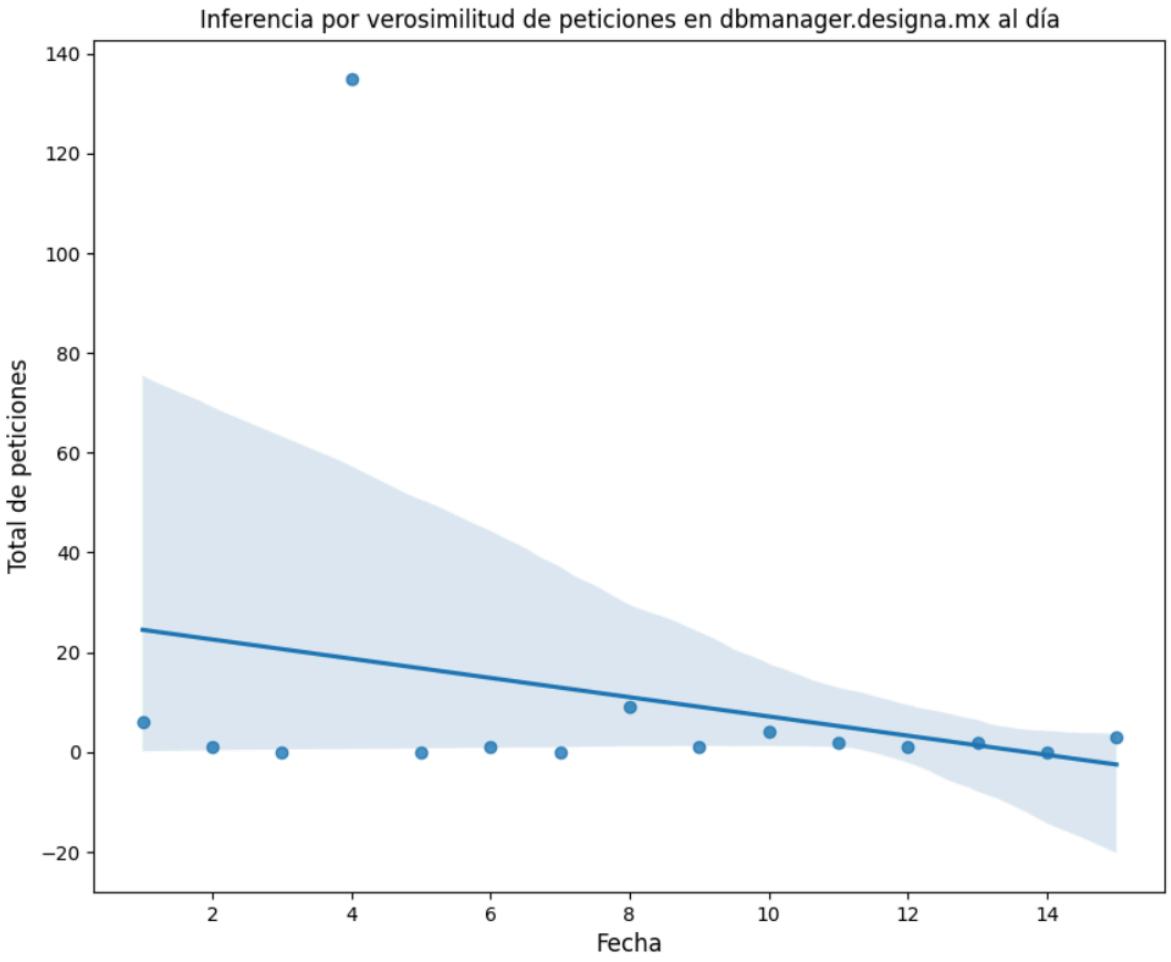
5. Remuestreo bootstrap con un intervalo de confianza de 95%



6. Distribución de tráfico total en dbmanager.designa.mx por día



7. Inferencia por verosimilitud de peticiones en dbmanager.designa.mx al día



8. Curva ROC de un predictor de tres dominios web basado en el resto de sus características