Actividad Evaluable 1

Descripción

MÓDULO	**Seminario Internacional en Herramientas y Técnicas de Detección de Ciberamenazas**
ASIGNATURA	**Data Science Aplicado a la Ciberseguridad**
Fecha Límite de Entrega	17 de Abril de 2023, a las 23:59
Puntos	20% de la Nota Total.
Carácter	Grupo (max 2 personas)

Enunciado:

En esta actividad se planteará una serie de preguntas relacionadas con los temas vistos en las sesiones 1 y 2. Los estudiantes debe responder a tales preguntas en este mismo documento, de forma clara y concisa. Este documento debe ser exportado a PDF, y entregado a través de la página de la asignatura, antes de la fecha límite de entrega.

Se considerará tanto la corrección de las soluciones como su presentación y el código utilizado para la obtención de los resultados.

Parte de esta actividad implica ejecutar código R. Tal código debe ser entregado en un fichero de código R (extensión *.R*), éste debe poderse ejecutar directamente sobre un terminal nuevo en R o en RStudio. El código es imprescindible para la corrección del ejercicio.

Las entregas tardías serán marcadas como "tarde", y pueden NO ser evaluadas. Por favor, entregad a tiempo.

1. Data Science

Pregunta 1:

De las siguientes preguntas, clasifica cada una como descriptiva, exploratoria, inferencia, predictiva o causal, y razona brevemente (una frase) el porqué:

1. Dado un registro de vehículos que circulan por una autopista, disponemos de su marca y modelo, país de matriculación, y tipo de vehículo (por número de ruedas). Con tal de ajustar precios de los peajes, ¿Cuántos vehículos tenemos por tipo? ¿Cuál es el tipo más frecuente? ¿De qué países tenemos más vehículos?

   Respuesta:

   En este caso, las preguntas son descriptivas porque el objetivo de cada una es recopilar información esclarecedora de las características de los vehículos que circulan por la carretera para ajustar los precios de los alojamientos, lo que se visualiza cuando preguntan por la cantidad de vehículos, desglosado por tipo y qué países tienen más coches, entre otras cosas.

2. Dado un registro de visualizaciones de un servicio de video-on-demand, donde disponemos de los datos del usuario, de la película seleccionada, fecha de visualización y categoría de la película, queremos saber ¿Hay alguna preferencia en cuanto a género literario según los usuarios y su rango de edad?

   Respuesta:

   La pregunta es de tipo inferencial porque implica analizar un grupo de datos para determinar si existe una preferencia. Además, trata de inferir una conexión entre el rango de edad del usuario y la categoría de película que está viendo, lo que implica que algunas categorías de películas son más populares que otras entre diferentes rangos de edad.

3. Dado un registro de peticiones a un sitio web, vemos que las peticiones que provienen de una red de telefonía concreta acostumbran a ser incorrectas y provocarnos errores de servicio. ¿Podemos determinar si en el futuro, los próximos mensajes de esa red seguirán dando problemas? ¿Hemos notado el mismo efecto en otras redes de telefonía?

Respuesta:

Esta pregunta es predicativa porque está especulando sobre una situación futura. Las consultas buscan determinar si los próximos mensajes de esa red telefónica seguirán causando problemas, lo que implica hacer una predicción sobre los problemas en el futuro.

4. Dado los registros de usuarios de un servicio de compras por internet, los usuarios pueden agruparse por preferencias de productos comprados. Queremos saber si ¿Es posible que, dado un usuario al azar y según su historial, pueda ser directamente asignado a un o diversos grupos?

Respuesta:

Esta pregunta es exploratoria ya que el objetivo es descubrir o investigar posibles resultados o efectos de una situación o problema. La pregunta valida si es posible realizar una asignación y podría conducir a la formulación de nuevas hipótesis o preguntas más específicas.

Pregunta 2:

Considera el siguiente escenario:

Sabemos que un usuario de nuestra red empresarial ha estado usando esta para fines no relacionados con el trabajo, como por ejemplo tener un servicio web no autorizado abierto a la red (otros usuarios tienen servicios web activados y autorizados). No queremos tener que rastrear los puertos de cada PC, y sabemos que la actividad puede haber cesado. Pero podemos acceder a los registros de conexiones TCP de cada máquina de cada trabajador (hacia donde abre conexión un PC concreto). Sabemos que nuestros clientes se conectan desde lugares remotos de forma legítima, como parte de nuestro negocio, y que un trabajador puede haber habilitado temporalmente servicios de prueba. Nuestro objetivo es reducir lo posible la lista de posibles culpables, con tal de explicarles que por favor no expongan nuestros sistemas sin permiso de los operadores o la dirección.

Explica con detalle cómo se podría proceder al análisis y resolución del problema mediante Data Science, indicando de dónde se obtendrían los datos, qué tratamiento deberían recibir, qué preguntas hacerse para resolver el problema, qué datos y gráficos se obtendrían, y cómo se comunicarán estos.

AQUÍ TU RESPUESTA

Respuesta:

Lo primero que se debería realizar es tener bien en claro el problema que se quiere resolver. En este caso, el problema es la presencia de un usuario que no hace buen uso de la red empresarial.

Procederemos a recopilar datos relevantes que permitan identificar y analizar la actividad del usuario en la red empresarial. Estos datos pueden incluir registros de inicio de sesión, registros de acceso a carpetas y archivos, y registros de actividades en la red.

Una vez obtenidos los datos, se podría usar técnicas para analizar los datos de texto y extraer información relevante. Con estos datos transformados, se puede realizar un análisis exploratorio para identificar patrones y tendencias en la actividad del usuario en la red empresarial con las siguientes preguntas:

- ¿Cuáles son los patrones de inicio de sesión del usuario en diferentes momentos del día?
- ¿Hay algún patrón que sugiera que el usuario está tratando de acceder a carpetas o archivos para aquellos que no tienen permiso?

De aquí, se puede identificar patrones y anomalías en la actividad del usuario, y alertar a los administradores de la red empresarial sobre posibles violaciones de seguridad.

Seguidamente, con los resultados del análisis exploratorio y las inferencias predictivas, se pueden implementar medidas de seguridad adicionales para evitar la presencia de usuarios sin permisos en la red empresarial. Esto puede incluir la implementación de políticas de seguridad más estrictas, la configuración de permisos de acceso más precisos, y la monitorización constante de la actividad de los usuarios.

Podríamos generar gráficos para los siguientes datos en mención:

- Registros de inicio de sesión:

  Con los datos de inicio de sesión se puede crear gráficos de barras o de líneas que muestran la cantidad de inicios de sesión por usuario en un período de tiempo determinado. Estos gráficos ayudan a identificar patrones y tendencias en el comportamiento de inicio de sesión del usuario.

- Registros de acceso a carpetas y archivos:

  Se pueden utilizar los datos de acceso a carpetas y archivos para crear gráficos de barras o de dispersión que muestran la cantidad de accesos a carpetas y archivos

por usuario en un período de tiempo determinado, así como la relación entre la cantidad de accesos y otros factores, como la hora del día o la duración de la sesión. Estos gráficos identifican patrones y tendencias en el acceso a carpetas y archivos, lo que puede ser útil para detectar actividades sospechosas, como intentos de acceso no autorizado.

- Registros de actividad de red:

  Con los datos de actividad de red se pueden crear gráficos de barras que muestran la cantidad de tráfico de red por usuario en un período de tiempo determinado y gráficos de dispersión que muestre la relación entre la cantidad de tráfico de red y la hora del día.

- Análisis de texto:

  Se podrían analizar los comentarios de los usuarios para identificar patrones de comportamiento o intentos de acceso no autorizado.

2. Introducción a R y Datos Elegantes

El segundo apartado de la práctica consiste en el análisis de un fichero de registro de peticiones HTTP, que debéis descargar (fichero adjunto: logs-http.zip), cargar en R, y realizar un análisis

Se recomienda tener cierto nivel de familiaridad y al alcance los cheatsheet de los distintos packages mencionados en las sesiones de teoría para un análisis más fácil:

- readr

- stringr

- tidyr (separate)

- dplyr (mutate, count)

Alternativamente, recordad que podéis consultar la sección de ayuda de RStudio y buscar en la documentación los parámetros así como ejemplos de uso (al final de cada página de documentación) para las funciones (escribiendo *?<nombre-funcion>* o presionando F1 sobre el nombre de la función.

Para las siguientes preguntas se requiere usar R. Indica en este documento para cada pregunta el resultado obtenido, describiendo a grandes rasgos el procedimiento seguido para la obtención de la respuesta, justificando cada decisión tomada a la hora de manipular los datos (descartar, agrupar, transformar, etc).

Asegúrate de entregar también el código en un fichero aparte, para poder ejecutarse directamente en un terminal limpio de R.

Pregunta 1:

Una vez cargado el Dataset a analizar, comprobando que se cargan las IPs, el Timestamp, la Petición (Tipo, URL y Protocolo), Código de respuesta, y Bytes de reply.

1. Cuales son las dimensiones del dataset cargado (número de filas y columnas)

   Respuesta:

   47748 filas y 7 columnas

2. Valor medio de la columna Bytes

   Respuesta:

   El valor medio es 7352

Consejo: probad distintos parámetros para las funciones de carga de datos o directamente usad el asistente visual de RStudio para cargar datos en el panel de Entorno (Environment).

Pregunta 2:

De las diferentes IPs de origen accediendo al servidor, ¿cuántas pertenecen a una IP claramente educativa (que contenga ".edu")?

Respuesta:

Hay 6524 IPs que tienen .edu

Pregunta 3:

De todas las peticiones recibidas por el servidor cual es la hora en la que hay mayor volumen de peticiones HTTP de tipo "GET"?

Respuesta:

La hora es las 14 horas (2:00 pm) con 4546 peticiones

Pregunta 4:

De las peticiones hechas por instituciones educativas (.edu), ¿Cuántos bytes en total se han transmitido, en peticiones de descarga de ficheros de texto ".txt"?

Respuesta:

El resultado es 2705408 bytes

Pregunta 5:

Si separamos la petición en 3 partes (Tipo, URL, Protocolo), usando str_split y el separador " " (espacio), ¿cuantas peticiones buscan directamente la URL = "/"?

Respuesta:

El resultado es 2382

Pregunta 6:

Aprovechando que hemos separado la petición en 3 partes (Tipo, URL, Protocolo) ¿Cuántas peticiones NO tienen como protocolo "HTTP/0.2"?

Respuesta:

Existen 47747 peticiones que no tienen protocolo http/0.2