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?

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?

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?

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?

AQUÍ TU RESPUESTA

1. Descriptiva: Describe las características de los vehículos sin tratar de buscar alguna relación y algún patrón entre ellos.
2. Exploratoria: Porque se intenta descubrir la existencia de una relación entre la categoría de la película y la preferencia del género literario de los usuarios en función de la edad.
3. Predictiva: Porque se intenta predecir si los mensajes próximos de una red en mención seguirá causando problemas, basado en el registro histórico.
4. Inferencial: Debido a que se busca inferir la preferencia que tiene un usuario a uno o diversos grupos en función de su historial de compras previas.

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 donde se obtendrían los datos, qué tratamiento deberían recibir, qué preguntas hacerse para resolver el problema, qué datos y gráficos se obtendrían, y cómo se comunicarían estos.

AQUÍ TU RESPUESTA

Primero se tienen que recopilar toda la información posible como los registros de conexiones TCP de cada PC de cada trabajador en el periodo en que se tiene sospechas que se ha realizado una actividad no autorizada. Se debe considerar la fecha y hora de conexión, los puertos, Ip de origen e Ip de destino.

Luego, de tener toda la información, se debe analizar los datos: Analizar las conexiones TCP donde se visualice alguna actividad sospechosa o anómala. Buscar patrones de actividad en los registros que sugieran que se ha utilizado algún servicio web no autorizado.
Preguntas:
Cuántas conexiones únicas ha hecho cada máquina?
A qué destinos se han conectado las máquinas frecuentemente?
Hay algún patrón de conexión común entre las máquinas sospechosas?

Luego exportar la información en CSV para poder importarlo en RStudio usando funcion read.csv()
Para tratar de los datos, se podrían utilizar funciones de la librería dplyr, que permite manipular y filtrar los datos de forma fácil usando funciones como filter(), mutate(), group_by()
Para los gráficos usaremos ggplot()

Se incluirá los datos y las gráficas obtenidas en un documento PDF.

De obtener algunos patrones de actividad y conexión sospechosa, se podría identificar una lista de los posibles culpables.

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)

Hallar columnas y filas

dim(info_data)

El resultado de la pregunta 1 es 47748 filas y 7 columnas

2. Valor medio de la columna Bytes

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, ¿cuantas pertenecen a una IP claramente educativa (que contenga ".edu")?

```
edu_ips <- str_count(info_data$Directions, "\\.edu")

Hallar el # de ips con extension edu
total_ips <- sum(edu_ips)
total_ips
```

Contienen edu un total de 6524

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"?

```
get_method <- filter(info_data, str_detect(info_data$Method2, "GET"))
View(get_method)
method_data <- select(get_method,Method2,Formatted_date )
method_data$Hora <- method_data$Formatted_date$hour
hour_max <- count(method_data, Hora, sort = TRUE)
hour_max
```

El resultado es 2:00 pm con 4546 peticiones

Pregunta 4:

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

```
Edu <- info_data[grepl("\\.edu", info_data$Directions), ]

Txt <- Edu[grepl("\\.txt", Edu$Resource), ]
sum(Txt$Size, na.rm=TRUE)
```

Bytes transmitidos: 2705408

```
txt_data <- filter(Edu, str_detect(Edu$Resource, "\\.txt$"))
sum(na.omit(txt_data$Size))
```

bytes transmitidos:106806

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 = "/"?

```
info_data$Resource <- as.factor(info_data$Resource)
Peticiones <- info_data[grepl("^/$", info_data$Resource), ]
summary(Peticiones)
```

Resultado es 2382

Pregunta 6:

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

```
http1_total <- sum(!str_detect(info_data$Protocol, "HTTP/0.2"))
http1_total
```

El resultado es 47747