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.
Grupo	Omar Rojas Izaguirre Rita Rosas Postigo Carlos Llanos Meza

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**, ya que su objetivo principal es resumir y organizar los datos disponibles sobre los vehículos que circulan por una autopista y en base a ello obtener respuestas.

2. **Exploratoria**, aquí se necesita explorar los datos para identificar patrones y relaciones entre variables para resolver la pregunta sobre la relación entre género literario y edad.

3. **Inferencial**, se enfoca en hacer predicciones y proyecciones a partir de los datos un conjunto de datos disponibles, en éste caso sobre la muestra de peticiones para predecir una respuesta en un conjunto más grande.

4. **Causal**, el objetivo de este análisis es identificar si una variable afecta directamente a otra, en este caso es acerca de la posibilidad de asignar a un usuario a uno o varios grupos según su historial de por internet.

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

Para proceder con la resolución de este problema mediante Data Science, habría que hacernos las siguientes preguntas:

- *¿Qué equipos de usuarios presentan actividad anómala en sus conexiones?*
- *¿Qué actividad realizan los usuarios identificados como sospechosos?*
- *¿Cuáles son los puertos o direcciones IP más utilizadas por los usuarios sospechosos? ¿estos usuarios tenían autorización para esa actividad?*

A continuación, se describe el procedimiento a seguir:

a) Recopilar datos de los registros de conexión TCP de las máquinas de los trabajadores para identificar conexiones que se han realizado en cada PC y hacia dónde se han realizado.

b) Con los datos recopilados sería necesario filtrarlos (exceptuando los usuarios autorizados), limpiarlos y adecuarlos para el análisis, retirando registros duplicados y agrupando la información por usuario.

c) Como siguiente paso, habría que realizar un análisis de los datos para conocer la distribución de las conexiones y detectar patrones o anomalías en las conexiones por usuario o direcciones IP a las que se han conectado los usuarios. Se podría incluir gráficos que muestren la frecuencia de las conexiones por usuarios, puertos o direcciones IP más utilizados a las que se conectan estos usuarios sospechosos.

d) Con los patrones de actividad sospechosa identificados podríamos tener direcciones IP desconocidas o inusuales a las que se ha conectado usuarios y con ello evaluar en mayor detalle la actividad de los usuarios identificados.

e) Finalmente, debemos comunicar los resultados a través de una presentación ejecutiva utilizando gráficos y visualizaciones para mostrar los patrones de conexiones y los resultados del análisis, a las Jefaturas de los usuarios identificados para que se tomen las medidas oportunas y asimismo establecer planes de acción para mejorar la seguridad en la red.

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

```
> dim(epa_http)
[1] 47747     7
```

2. Valor medio de la columna Bytes

```
> summary(epa_http)
    source              hora              recurso          cod_retorno       tamaño
 Length:47747       Length:47747       Length:47747       Min.   :200.0   Min.   :      0
 Class :character   Class :character   Class :character   1st Qu.:200.0   1st Qu.:    231
 Mode  :character   Mode  :character   Mode  :character   Median :200.0   Median :   1260
                                                          Mean   :227.1   Mean   :   7352
                                                          3rd Qu.:200.0   3rd Qu.:   3223
                                                          Max.   :501.0   Max.   :4816896
                                                                          NA's   :5331

 metodo_limpio protocolo_limpio
 GET :46019    HTTP/0.2:     1
 HEAD:  106    HTTP/1.0:47746
 POST: 1622
```

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")?

```
> count(epa_http[grepl("edu",epa_http$source),])
# A tibble: 1 × 1
      n
  <int>
1  6539
```

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"?

```
> datos_get <- epa_http[grepl("GET", epa_http$metodo_limpio ), ]
> datos_get$hora_format <- as.POSIXlt(datos_get$hora,format="[%d:%H:%M:%OS]")
> datos_get$hora_2 <- datos_get$hora_format$hour
> count(datos_get, hora_2, sort = TRUE)
# A tibble: 24 × 2
   hora_2     n
    <int> <int>
 1     14  4546
 2     13  4202
 3     15  4122
 4     16  3950
 5     12  3707
 6     11  3689
 7     10  3140
 8      9  3008
 9     17  2694
10      8  1911
# ℹ 14 more rows
# ℹ Use `print(n = ...)` to see more rows
```

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"?

```
> sum(fichero_txt$tamaño,na.rm = TRUE)
[1] 3017871
```

Pregunta 5:

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

```
> count((filter(epa_http,recurso=="/")))
# A tibble: 1 × 1
      n
  <int>
1  2382
```

Pregunta 6:

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

```
> count(epa_http[grepl("HTTP/1.0",epa_http$protocolo_limpio),])
# A tibble: 1 × 1
      n
  <int>
1 47746
```