Actividad #1 – Curso Data Science - Stage Barcelona

GRUPO #4

Integrantes:

- **Richard Crocce**
- **Kevin Moscoso**
- **Roberto Ibarra**

Contenido

Control de Cambios

Fecha	Versión	Editado por	Comentarios
17-11-2023	1.0	Grupo 04	Creación Documento
19-11-2023	1.1	Grupo 04	Actualizaciones finales, se coloca REPO como Capitulo.

(I) Data Science

1. Pregunta #1:

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

i. 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?

Tipo: Descriptiva, ya que en base a información recolectada se busca describir una distribución.
Razón: Son atributos (información detallada y particular) a partir de los cuales podemos agruparlos y contabilizar una composición de atributos simple o agrupada.

ii. 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?

Tipo: Exploratoria. El servicio video-on-demand, busca encontrar la relación entre 3 conjuntos de datos (usuarios, películas (atributo de categoría) y fecha de visualización). Ojo son datos a los cuales se quiere analizar retroactivamente.
Razón: La correlación nos permite medir la tendencia entre dos variables. En este caso son atributos de género literario (atributo de las películas y rango de edad (atributo de usuarios)

iii. 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?

Tipo: Predictiva e Inferencia
Razón: La primera pregunta necesita una respuesta a través de herramientas predictivas, ya que busca prever el comportamiento futuro de los mensajes de la red en cuestión. La segunda parte es inferencial, ya que intenta inferir si el mismo efecto se observa en otras redes de telefonía

iv. 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?

Tipo: Causal.
Razón: Busca establecer una relación causal entre el historial de compras de un usuario y su asignación a grupos específicos, explorando si el historial influye en la asignación a grupos.

2. 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.

Respuesta:

2.1. OBJETIVO:

Ubicar el “usuario(s)” que implementó de forma “no autorizada” un “servicio web” (activo o cesado) para explicarles que no expongan sin permiso los sistemas de los operadores o la dirección.

2.2. ALCANCE:

“Usuario(s)” que implementó de forma “no autorizada” un “servicio web” (activo o cesado) el cual es usado vía el servicio VPN por parte de algunos clientes.

Tipo: Exploratoria

2.3. FUENTES Y FILTROS:

Servicio VPN:

- Filtro #1: Conexiones desde el segmento VPN hacia el segmento lan (usuarios) mediante el puerto TCP 80 y 443.
- Filtro #2:
    - IP origen del segmento vpn
    - Cuenta vpn
    - Fecha y hora del inicio y cierre de la conexión de las cuentas vpn filtradas.
    - IP lan del host destino con puerto TCP 80 y 443 habilitado.
    - Fecha y hora de las conexiones hacia el puerto 80 y 443 del host destino.
- Filtro #3:
    - De los hosts con puerto TCP 80 y 443 habilitado, excluir los hosts/usuarios autorizados.

2.4. GRÁFICAS:

Gráfica #1:

- Listado de host “no autorizados” con el puerto TCP 80 y 443 habilitado.

Gráfica #2:

- Cantidad de conexiones y fechas hacia los hosts “no autorizados” con el puerto TCP 80 y 443 habilitado.
- Respuestas del tipo “200” para saber las webs no autorizadas y aún activas.
- Listado de cuentas vpn (clientes) que se conectaron a los hosts “no autorizados” con el puerto TCP 80 y 443 habilitado.

2.5. COMUNICACIÓN:

- Usuario final:
    - Envío de correo o documento formal informando al usuario final el uso indebido (no autorizado) de los recursos e información de los sistemas de los operadores o la dirección.
    - Capacitación sobre la política de seguridad de la información “Uso correcto de los recursos de la Compañía”.
- Jefe/Gerencia del usuario final:
    - Comunicación sobre el hallazgo, las acciones tomadas y los riesgos involucrados para el negocio.

2.6. RECOMENDACIONES:

- Centralización de los logs para mayor trazabilidad y visibilidad.
- Automatización de los filtros utilizados y que este sea ejecutado periódicamente (mensual) para detección proactiva de la implementación de servicios web no autorizados.
- El resultado de la ejecución de la tarea programada enviarlo al personal involucrado para la toma de acciones en caso aplique.

(II) Introducción a R y Datos Elegantes

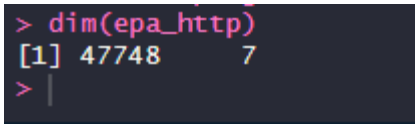
1. **Pregunta #1:**

   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.

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

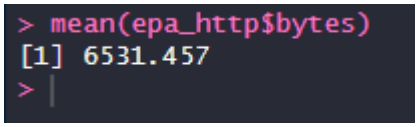
   Dedicidimos no agrupar por Petición (Tipo, URL y Protocolo) ya que más adelante se usarán esos datos para agrupar valores.

   Filas: 47748
   Columnas: 7



   ii. Valor medio de la columna Bytes

   Los valores NA les hemos colocado valores 0 con el fin de poder considerarlos a la hora de ponderar. Cabe resaltar que al hacer esto, pasamos en aumentar 21 registros que tenian valor NA a 0. Con esto el promedio simple es penalizado al incluirse 21 registros adicionales.
   Media: 6531.457



2. **Pregunta #2:**

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

Hemos considerado que el dominio “.edu” pueda ser el top level domain o un sub level domain “.edu.” . Tambien una validación de incasesensitive.

Conteo: 6539.

```
> print(conteo)
[1] 6539
>
```

3. **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"?

Primero se identifica si hay datos únicos para la misma hora. Son 25 agrupaciones por hora. Teniendo las 23 horas dos registros.

1	2023-11-30 14:00:00	4546
1	2023-11-30 14:00:00	4546
2	2023-11-30 13:00:00	4202
3	2023-11-30 15:00:00	4122
4	2023-11-30 16:00:00	3950
5	2023-11-30 12:00:00	3707
6	2023-11-30 11:00:00	3689
7	2023-11-30 10:00:00	3140
8	2023-11-30 09:00:00	3008
9	2023-11-30 17:00:00	2694
10	2023-11-30 08:00:00	1911
11	2023-11-30 18:00:00	1770
12	2023-11-30 19:00:00	1435
13	2023-11-30 20:00:00	1215
14	2023-11-30 22:00:00	1065
15	2023-11-30 23:00:00	1058
16	2023-11-30 21:00:00	984
17	2023-11-30 07:00:00	814
18	2023-11-30 00:00:00	659
19	2023-11-30 01:00:00	424

20	2023-11-30 02:00:00	386
21	2023-11-30 05:00:00	355
22	2023-11-30 04:00:00	327
23	2023-11-30 06:00:00	296
24	2023-11-30 03:00:00	201
25	2023-11-29 23:00:00	62

Bajo el anterior análisis, es indistinto agrupar por Fecha+Hora que por Hora. Por lo cual el mayor valor de horas es **14:00 horas.**

```
> print(resultados, n = nrow(resultados))
# A tibble: 25 x 2
   timestamp_agrupado  cantidad_peticiones
   <chr>                             <int>
 1 2023-11-30 14:00:00                4546
 2 2023-11-30 13:00:00                4202
 3 2023-11-30 15:00:00                4122
 4 2023-11-30 16:00:00                3950
```

4. **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"?

TotalBytes: 106806

```
> p4 <- p4[grepl(".txt$", p4$uri),]
> c4 <- sum(p4$bytes)
> c4
[1] 106806
>
```

5. **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 = "/"?

Peticiones: 2382

```
#Pregunta 5
p5  <- dplyr::filter(epa_http,stringr::str_like(uri,"/",ignore_case = TRUE))
n5 <- count(p5)
n5
```

```
1  2382
> view(p5)
> p5  <- dplyr::filter(epa_http,stringr::str_like(uri,"/",ignore_case = TRUE))
> n5 <- count(p5)
> n5
# A tibble: 1 × 1
      n
  <int>
1  2382
>
```

6. **Pregunta #6:**

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

Peticiones sinprotocolo HTTP/0.2 = dim(epa_http= - dim(p6) = 1

```
#Pregunta 6
p6  <- dplyr::filter(epa_http,!stringr::str_like(protocolo,"%HTTP/0.2%",ignore_case = TRUE))
n6Tot <- count(epa_http)
n6Diff <- count(p6)

print(as.numeric(n6Tot) - as.numeric(n6Diff))

dim(p6)
dim(epa_http)
```

```
> dim(p6)
[1] 47747     8
> dim(epa_http)
[1] 47748     8
>
```

(III) Repositorio en GITHUB

https://github.com/vanisheriii/barcelona-data-science.git

< Fin del documento >