Ejercicio 1

Julio Takimoto, Nelson Marin, Jesus Carrasco

2024-05-15

This R Markdown document is made interactive using Shiny. Unlike the more traditional workflow of creating static reports, you can now create documents that allow your readers to change the assumptions underlying your analysis and see the results immediately.

To learn more, see Interactive Documents (http://rmarkdown.rstudio.com/authoring_shiny.html).

1. Data Science

Pregunta 1:

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?.

RPTA: Se ha considerado que la pregunta es Descriptiva; dado que, lo que se busca es listar resultados reales (registro de vehículos) para identificar de manera estadística la cantidad de tipos de vehículos que pasan por la autopista.

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?

RPTA: Se ha considerado que la pregunta es Inferencial; dado que, en base al registro, se debe inferir o brindar una afirmación general para toda la categoría (genero/rango de edad) de usuarios.

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?

RPTA: Se ha considerado que la pregunta es Predictiva; dado que, en base a hechos reales (registros), se predice que la media o el comportamiento será el mismo para los siguientes casos (nuevas peticiones provenientes de una red 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?

RPTA: Se ha considerado que la pregunta es Exploratoria; dado que, no se tienen certeza del comportamiento que vaya a realizar el usuario. Sin embargo, en base a la información tratada, se puede realizar una medición / inferencia sobre su comportamiento.

Pregunta 2:

Análisis y resolución del problema mediante Data Science:

_Donde se obtendrían los datos / qué preguntas se podrían realizar para solucionar el problema:

a. Tráfico de datos hacia internet (red): Con un histórico de la información de unos meses, se podría identificar la siguiente información:

- ¿Qué estación(es) de trabajo se encuentra expuesta a la red sin autorización?
- ¿Qué estación(es) de trabajo han sido expuestas a internet recientemente?
- ¿Cuánto tiempo (días y horas) estuvo expuesto este servicio web sin autorización?
- ¿Existe algún tráfico anómalo en la red de la empresa?
- ¿Qué cantidad de tráfico se ha consumido por este servicio web?
- ¿Cuántos usuarios o peticiones accedieron al servicio web?

b. Listado o inventario de servicios web activos: con esta información, se puede evidencia todos los servicios web autorizados y se puede identificar los servicios web que no cuentan con autorización.

_Datos y gráficos se obtendrían:

Se podría realizar un par de gráficos; los cuales, se podrían presentar en los diferentes comités de gestión:

a. Comportamiento de la red a través de los meses desde el lanzamiento del servicio web no autorizado.

b. Cantidad de usuarios o cantidad de peticiones recibidas en el tiempo que el servicio web no autorizado estuvo activo.

_Como se comunicarían estos:

Se sugiere informar este caso a algún comité (especializado) de gestión; en el cual se pueda informar esta situación como un incidente de seguridad. Se podría agregar la información de tiempo, medidas tomadas,

_Solución integral del problema:

a. Realizar el análisis de información detallado líneas arriba (utilizando big data) / estaciones de trabajo.

b. Realizar el bloqueo de los accesos no autorizados y restringir el acceso a internet a ciertas estaciones de trabajo.

c. Realizar un monitoreo de tráfico de red periódico para identificar posibles anomalías.

d. Realizar un escaneo preventivo a la red de la empresa.

e. Reforzar la cultura organizativa con temas de seguridad de información.

2. Introducción al R

Pregunta 1:

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

Las dimensiones del data set son:

dim(epa_http)

```
[1] 47748    12
```

2. ¿Valor medio de la columna Bytes?

colnames(epa_http) <- c(“IP”, “Timestamp”, “Tipo”, “URL”, “Protocolo”, “Còdigo_respuesta”, “Bytes”)

mean(epa_http$Bytes, na.rm = T)

```
[1] 7352.335
```

Pregunta 2:

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

```
grep(".edu",epa_http$IP,ignore.case = T,perl = T)
```

epa_http_edu <- epa_http[grep(“.edu”,epa_http$IP,ignore.case = T,perl = T),]

```
nrow(epa_http_edu)
[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”?

epa_http$dia <- str_sub(epa_http$Timestamp,2,3)

epa_http$hora <- str_sub(epa_http$Timestamp,5,6)

epa_http$t <- str_sub(epa_http$Timestamp,5,12)

table(epa_http$hora)

```
  00   01   02   03   04   05   06   07   08   09   10   11   12   13   14   15   16   17   1
8   19   20   21   22   23   684  434  399  248  347  374  303  846 1994 3096 3209 3820 3827
4391 4716 4284 4042 2793 1820 1493 1310 1015 1117 1186
```

Es a las 14 horas: 4,716

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”?

epa_http_edu <- epa_http[grep(“.edu”,epa_http$IP,ignore.case = T,perl = T),]

epa_http_edu_txt <- epa_http_edu[grep(“.txt”,epa_http_edu$URL,ignore.case = T,perl = T),]

sum(epa_http_edu_txt$Bytes)

```
sum(epa_http_edu_txt$Bytes, na.rm = T)
[1] 3017871
```

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 = “/”?

epa_http$URL1 <- str_sub(epa_http$URL,1,1)

grep(“/”,epa_http$URL1,ignore.case = T,perl = T)

```
epa_http_url <- epa_http[grep("/",epa_http$URL,ignore.case = T,perl = T),]
nrow(epa_http_url)
[1] 47748
```

Pregunta 6:

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

```
 table(epa_http$Protocolo)
 HTTP/0.2 HTTP/1.0
 1    47747
```

Son 47,747