UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
School of Professional & Executive Development

Seminario Internacional en Herramientas y Técnicas de Detección de Ciberamenazas

Data Science Aplicado a la Ciberseguridad

Integrantes:

- Carlos Arana Orrego
- Luis Zevallos Sulca
- Roger Misagel Vega

20 noviembre 2023

1. Sección I – Data Science

El primer apartado de la práctica resolveremos las preguntas que implica trabajar con datos:

Clasificación de las preguntas:
- ❖ Descriptivas: ¿Cómo es el conjunto de datos, estadísticamente?
- ❖ Exploratorias: ¿Qué relaciones existen en los datos?
- ❖ Inferenciales: ¿Cómo se generalizan los datos a una muestra mayor?
- ❖ Predictivas: ¿Se pueden predecir nuevos valores no vistos?
- ❖ Causales: ¿Qué causa el comportamiento visto en los datos?

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?

Respuesta:

1. Descriptivo, Nos pide analizar un conjunto de datos estadísticos como el numero de llantas, modelo y marca de un vehículo para ajustar el precio del peaje, para esto nos basamos en data que con la que ya se cuenta en el dataset.
2. Inferencial, Buscamos analizar desde los datos de usuarios muestra un comportamiento general para dar una clasificación a tipos de usuarios.
3. Predictivo, Se busca analizar y predecir, comparando con otros proveedores, a través de los datos como las peticiones que provienen de una red de telefonía.
4. Causales, Se busca analizar los datos a través de preferencias de compra de los usuarios para generalizarlo grupos de usuarios.

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:

Obtención de datos

Los datos necesarios para este análisis se pueden obtener de los registros de conexiones TCP de cada máquina de cada trabajador. Estos registros suelen almacenarse en un servidor de registro o en un sistema de almacenamiento en la nube. Los datos deben incluir la siguiente información:

- Dirección IP de origen
- Dirección IP de destino
- Puerto de origen
- Puerto de destino
- Fecha y hora de la conexión
- Tratamiento de datos

Los datos se pueden tratar de la siguiente manera:

- Se eliminan los registros de conexiones que no sean de TCP.
- Se eliminan los registros de conexiones a direcciones IP internas.
- Se eliminan los registros de conexiones a direcciones IP de clientes legítimos.
- Preguntas para resolver el problema

Las siguientes preguntas se pueden utilizar para resolver el problema:

¿Cuáles son las direcciones IP de destino más comunes?
¿Cuáles son los puertos de destino más comunes?
¿A qué horas del día se producen las conexiones?
Datos y gráficos

Los siguientes datos y gráficos se pueden obtener para responder a estas preguntas:

Tabla de frecuencias: Esta tabla mostrará la frecuencia con la que se producen las conexiones a cada dirección IP de destino.
Gráfico de barras: Este gráfico mostrará la frecuencia con la que se producen las conexiones desde cada dirección IP de origen.
Gráfico de líneas: Este gráfico mostrará la hora del día a la que se producen las conexiones.
Comunicación de los resultados

Los resultados del análisis se pueden comunicar de la siguiente manera:

Se puede crear un informe que resuma los resultados clave.
Se puede crear una presentación que muestre los datos y gráficos.

2. Sección II - Introducción a R y DatosElegantes

El segundo apartado de la práctica trabajaremos con el fichero de registro de peticiones HTTP, epa-http.zip y lo cargaremos en R Studio.

Trabajaremos con los siguientes packages mencionados en las sesiones de teoría para un análisis más fácil:

- *readr*
- *stringr*
- *tidyr (separate)*
- *dplyr (mutate, count)*

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)?

Para poder obtener la información solicitada se ha cargado de manera gráfica en la variable epa_http el dataset solicitado.
Se usará el comando NROW para conocer la cantidad de filas y el NCOL para saber la cantidad de columnas.

filas <- NROW(epa_http)	*El resultado de la variable es 47748*
Colums <- NCOL(epa_http)	*El resultado de la variable es 7*

El dataset cuenta con 47748 filas y 7 columnas

2. ¿El valor medio de la columna Bytes?

Como parte del proceso en la pregunta 1, importamos el dataset y colocamos la columna X7 (columna Bytes) como un tipo numérico, de esta manera podemos usar comandos propios de este tipo de dato, como el comando mean, que devuelve el valor medio de la columna Bytes, también se remueve de la media los posibles valores nulos (NA).

```
media <- mean(epa_http$X7, na.rm=T)
```

El resultado de la variable media es: `[1] 7352.335`

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")?

Para conocer la cantidad de IPs que sean educativas necesitamos filtrar en la columna IPs de origen (X1), el fragmento de string ".edu", y mediante el comando NROW contar la cantidad de filas que contienen .edu en su columna de IP origen.

```
NcolumEDU <- NROW(filter(epa_http,grepl(".edu",X1)==TRUE))
```

Resultado NclomEDU = 6539

Pregunta 3:

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

Para conocer la hora de mayor volumen, primero filtramos las peticiones de tipo GET en la columna peticiones (X3)

```
z<-filter(df2,grepl("GET",X3)==TRUE)
```

Posteriormente con el comando count tendremos el total de peticiones por hora

```
z<-z %>% count(hora)
```

Finalmente filtramos la hora con mayor número de peticiones

```
#filtramos la hora de mayor uso
Preg3 <- z %>% filter(n==max(z$n))
```

El resultado Preg3 es que el horario con mayor numero de peticiones es el de las 14:00 horas con 4546 peticiones tipo GET

Pregunta 4:

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

Primero filtramos todas las peticiones a instituciones educativas en la variable edu

```
edu <- filter(epa_http,grepl("edu",X1)==TRUE)
```

Posteriormente filtramos los archivos que tengan un formato .txt en la columna x4 y lo colocamos en la variable TXT

```
TXT <- filter (edu,substr(edu$X4, nchar(edu$X4)-3, nchar(edu$X4))==".txt")
```

Finalmente sumamos los valores numéricos de la columna Bytes (X7) quitando los valores nulos (NA)

```
Preg4 <- sum(TXT$X7, na.rm = TRUE)
```

Resultado Preg4 106806 bytes

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 = "/"?

Según el enunciado filtramos las peticiones a la URL "/" y lo colocamos en la variable temp.

```
temp <- filter(epa_http, X4=="/")
```

Posteriormente contamos la cantidad de filas de la variable temp en la variable Preg5

```
Preg5<- NROW(temp)
```

El resultado de Preg5 es de 2382 peticiones.

Pregunta 6:

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

Según el enunciado filtramos las peticiones a la URL al protocolo “HTTP/0.2” y lo colocamos en la variable temp2.

```
temp2 <- filter(epa_http, substr(X5, 1, 8)!="HTTP/0.2")
```

Finalmente en la variable Preg6 colocamos la cantidad de filas de la variable temp2.

```
Preg6<- NROW(temp2)
```

Resultado de la variable Preg6 es de 47747 peticiones.