Practica 1

Christiam Meza - Carlos Alama - Rony Orocollo

2024-05-15

```
library(stringr)
```

```
## Warning: package 'stringr' was built under R version 4.3.3
```

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.3.3
```

Pràctica 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?

R: exploratoria.- Esto se debe a que la naturaleza de los datos solicitados buscan comprender las relaciones y patrones que existen dentro del conjunto de datos. Se incluyen preguntas sobre correlaciones, asociaciones o agrupaciones de datos.

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?

R: descriptiva.- Se elige ello, debido a que los datos solicitados buscan comprender las relaciones y patrones que existen dentro del conjunto de datos. Se incluiyen preguntas sobre correlaciones, asociaciones o agrupaciones de datos.

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?

R: causales.- Es asì, ya que lo solicitado busca comprender qué factores o variables causales están influenciando el comportamiento observado en los datos. Implican la identificación de relaciones de causa y efecto; sobre todo con la consulta final.

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?

R: predictiva.- Se escoge ello, ya que se centra en si es posible predecir valores futuros o desconocidos basados en el análisis de datos históricos. Lo solicitado Implica el desarrollo y la evaluación de los datos a tener.

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.

R:Para abordar el problema de identificar a los usuarios que han estado utilizando la red empresarial para fines no autorizados, como la apertura de servicios web no autorizados, podemos seguir un enfoque basado en Data Science. Aquí se detalla un plan de acción según el esquema enseñado en clase:

a). Pregunta de Interés ¿Cuáles usuarios han utilizado la red empresarial para fines no relacionados con el trabajo, específicamente abriendo servicios web no autorizados?

b). Obtener los Datos

b1)Fuente de Datos:

- Registros de conexiones TCP de cada máquina de los trabajadores.

- Información de las conexiones legítimas y autorizadas (lugares remotos de clientes, servicios web autorizados, etc.).

- Listado de servicios web autorizados y sus respectivas configuraciones de puertos.

B2)Pasos para Obtener los Datos:

- Recopilar los registros de conexiones TCP desde los firewalls, routers y servidores de la red.

- Obtener logs de aplicaciones de seguridad que monitoricen el tráfico de red.
- Recabar listas de servicios y puertos autorizados desde el departamento de IT.

c). Explorar los Datos

c1)Tratamiento de los Datos:

- Limpiar los datos eliminando entradas duplicadas y registros corruptos.

- Estandarizar formatos de tiempo y direcciones IP.
- Enriquecer los datos con información geográfica de las IPs para identificar conexiones remotas legítimas.

c2). Exploración Inicial:

- Realizar un análisis descriptivo de los datos, como el número de conexiones por usuario, tipos de servicios web accesados, puertos utilizados, etc.

- Identificar patrones normales de uso de la red y compararlos con posibles actividades sospechosas.

c3)Preguntas Específicas:

- ¿Qué usuarios han accedido a puertos o servicios web no autorizados?
- ¿Existen patrones de conexiones fuera del horario laboral habitual?

- ¿Hay conexiones repetitivas a ciertos puertos que no están en la lista de servicios autorizados?

D). Modelar los Datos

D1)Métodos de Modelado:

- Análisis de Anomalías: Usar algoritmos de detección de anomalías (e.g., Isolation Forest, Local Outlier Factor) para identificar comportamientos fuera de lo común.

- Clustering: Aplicar clustering (e.g., K-means) para agrupar comportamientos similares y detectar usuarios con patrones de conexión atípicos.

- Reglas de Asociación: Generar reglas para identificar combinaciones de puertos y direcciones IP no autorizadas.

D2)Evaluación:

- Validar los modelos con datos históricos y conocidos de actividades autorizadas e inusuales.

- Realizar pruebas de sensibilidad y precisión para ajustar los modelos y reducir falsos positivos.

E). Comunicar y Visualizar el Resultado

Datos y Gráficos a Obtener:

- Tablas resumen de usuarios y sus respectivas conexiones a servicios no autorizados.

- Gráficos de barras que muestren la frecuencia de uso de puertos no autorizados por usuario.

- Mapas geográficos de conexiones remotas para identificar ubicaciones sospechosas.

- Gráficos de líneas que muestren patrones de conexión fuera del horario laboral.

Comunicación:

- Preparar un informe detallado con hallazgos clave, respaldado por gráficos y visualizaciones claras.

- Crear un dashboard interactivo para que los operadores puedan explorar los datos y resultados.
- Realizar una presentación a la dirección, explicando los riesgos y proponiendo medidas preventivas, como la implementación de alertas automáticas para conexiones sospechosas.

Ejemplo de Informe y Visualización

Informe:

- Resumen Ejecutivo: Descripción del problema, metodología utilizada y principales hallazgos.

- Análisis de Datos: Detalle del proceso de obtención y limpieza de datos, seguido de análisis descriptivo y modelado.

- Hallazgos Clave: Usuarios identificados con actividad sospechosa, patrones de comportamiento atípicos, y ejemplos de conexiones no autorizadas.

- Recomendaciones: Sugerencias para mejorar la seguridad de la red, como políticas de monitoreo continuo y educación de los empleados sobre el uso adecuado de los recursos.

Visualizaciones:

- Gráfico de Barras: “Número de Conexiones a Puertos No Autorizados por Usuario.”

- Mapa Geográfico: “Ubicaciones de Conexiones Remotas.”

- Gráfico de Líneas: “Actividad de Red por Hora del Día.”

Estas herramientas y visualizaciones ayudarán a la dirección a entender el alcance del problema y tomar decisiones informadas para proteger la red empresarial.

Se cambian los nombres de las cabaceras:

Se hizo la modificacion de los datos null

Introduccion a R

Pregunta 1

Se procede a importar el DataSet que es un archivo csv con nombre epa-http.csv de la siguiente ruta “C:/Users/carlosac12/e/epa-http.csv”

```
##
## ── Column specification ────────────────────────────────────────────────────────
## cols(
##   X1 = col_character(),
##   X2 = col_character(),
##   X3 = col_character(),
##   X4 = col_character(),
##   X5 = col_character(),
##   X6 = col_double(),
##   X7 = col_character()
## )
```

```
## Warning: 21 parsing failures.
##  row col  expected    actual                                 file
## 7527  -- 7 columns 6 columns 'C:/Users/carlosac12/e/epa-http.csv'
## 7528  -- 7 columns 6 columns 'C:/Users/carlosac12/e/epa-http.csv'
## 7529  -- 7 columns 6 columns 'C:/Users/carlosac12/e/epa-http.csv'
## 7549  -- 7 columns 6 columns 'C:/Users/carlosac12/e/epa-http.csv'
## 7550  -- 7 columns 6 columns 'C:/Users/carlosac12/e/epa-http.csv'
## .... ... ......... ......... ...................................
## See problems(...) for more details.
```

Para un mejor entendimiento de la informaciòn, se procede a cambiar los nombres de las cabeceras del dataSet epa_http a “IPS”, “timestamp”, “peticion”, “URL”, “protocolo”, “code_respuesta” y “bytes_reply”

```
names(epa_http) <- c("IPS", "timestamp", "peticion", "URL", "protocolo", "code_respuesta", "b
ytes_reply")
epa_http
```

```
## # A tibble: 47,748 × 7
##    IPS             timestamp peticion URL   protocolo code_respuesta bytes_reply
##    <chr>           <chr>     <chr>    <chr> <chr>              <dbl> <chr>
##  1 141.243.1.172   [29:23:5… "\"GET"  /Sof… "HTTP/1.…            200 1497
##  2 query2.lycos.c… [29:23:5… "\"GET"  /Con… "HTTP/1.…            200 1325
##  3 tanuki.twics.c… [29:23:5… "\"GET"  /New… "HTTP/1.…            200 1014
##  4 wpbfl2-45.gate… [29:23:5… "\"GET"  /     "HTTP/1.…            200 4889
##  5 wpbfl2-45.gate… [29:23:5… "\"GET"  /ico… "HTTP/1.…            200 2624
##  6 wpbfl2-45.gate… [29:23:5… "\"GET"  /log… "HTTP/1.…            200 935
##  7 140.112.68.165  [29:23:5… "\"GET"  /log… "HTTP/1.…            200 2788
##  8 wpbfl2-45.gate… [29:23:5… "\"GET"  /log… "HTTP/1.…            200 124
##  9 wpbfl2-45.gate… [29:23:5… "\"GET"  /ico… "HTTP/1.…            200 156
## 10 wpbfl2-45.gate… [29:23:5… "\"GET"  /log… "HTTP/1.…            200 2788
## # ℹ 47,738 more rows
```

Tras el anàlisis de la informaciòn, se ha encontrado data vacìa y con el signo “-”. se renombra y se cambia a null a continuaciòn:

```
epa_http[is.na(epa_http)] <- "null"
epa_http[epa_http == "-"] <- "null"
epa_http
```

```
## # A tibble: 47,748 × 7
##    IPS             timestamp peticion URL   protocolo code_respuesta bytes_reply
##    <chr>           <chr>     <chr>    <chr> <chr>              <dbl> <chr>
##  1 141.243.1.172   [29:23:5… "\"GET"  /Sof… "HTTP/1.…            200 1497
##  2 query2.lycos.c… [29:23:5… "\"GET"  /Con… "HTTP/1.…            200 1325
##  3 tanuki.twics.c… [29:23:5… "\"GET"  /New… "HTTP/1.…            200 1014
##  4 wpbfl2-45.gate… [29:23:5… "\"GET"  /     "HTTP/1.…            200 4889
##  5 wpbfl2-45.gate… [29:23:5… "\"GET"  /ico… "HTTP/1.…            200 2624
##  6 wpbfl2-45.gate… [29:23:5… "\"GET"  /log… "HTTP/1.…            200 935
##  7 140.112.68.165  [29:23:5… "\"GET"  /log… "HTTP/1.…            200 2788
##  8 wpbfl2-45.gate… [29:23:5… "\"GET"  /log… "HTTP/1.…            200 124
##  9 wpbfl2-45.gate… [29:23:5… "\"GET"  /ico… "HTTP/1.…            200 156
## 10 wpbfl2-45.gate… [29:23:5… "\"GET"  /log… "HTTP/1.…            200 2788
## # ℹ 47,738 more rows
```

Se procede convertir la columna ‘timestamp’ a formato de fecha y hora:

```
epa_http$timestamp <- as.POSIXct(epa_http$timestamp, format = "[%d:%H:%M:%S]")
epa_http
```

```
## # A tibble: 47,748 × 7
##    IPS   timestamp           peticion URL   protocolo code_respuesta bytes_reply
##    <chr> <dttm>              <chr>    <chr> <chr>              <dbl> <chr>
##  1 141.… 2024-05-29 23:53:25 "\"GET"  /Sof… "HTTP/1.…            200 1497
##  2 quer… 2024-05-29 23:53:36 "\"GET"  /Con… "HTTP/1.…            200 1325
##  3 tanu… 2024-05-29 23:53:53 "\"GET"  /New… "HTTP/1.…            200 1014
##  4 wpbf… 2024-05-29 23:54:15 "\"GET"  /     "HTTP/1.…            200 4889
##  5 wpbf… 2024-05-29 23:54:16 "\"GET"  /ico… "HTTP/1.…            200 2624
##  6 wpbf… 2024-05-29 23:54:18 "\"GET"  /log… "HTTP/1.…            200 935
##  7 140.… 2024-05-29 23:54:19 "\"GET"  /log… "HTTP/1.…            200 2788
##  8 wpbf… 2024-05-29 23:54:19 "\"GET"  /log… "HTTP/1.…            200 124
##  9 wpbf… 2024-05-29 23:54:19 "\"GET"  /ico… "HTTP/1.…            200 156
## 10 wpbf… 2024-05-29 23:54:19 "\"GET"  /log… "HTTP/1.…            200 2788
## # ℹ 47,738 more rows
```

1.1 Cuales son las dimensiones del dataset cargado (número de filas y columnas)?

el nùmero de regitros es 47748 el nùmero de columnas es 7

1.2 Valor medio de las columnas Bytes

```
bytes_reply_numeric <- as.numeric(epa_http$bytes_reply[epa_http$bytes_reply != "null"])
media_bytes_reply <- mean(bytes_reply_numeric, na.rm = TRUE)
```

El valor medio es 7352.3349129

Pregunta 2

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

```
conteo_edu <- sum(grepl("\\.edu", epa_http$IPS))
```

Las que pertenecen a una IP claramente educativa que contenga “.edu” son 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_rows <- epa_http[grepl("GET", epa_http$peticion), ]
GET_rows <- GET_rows[!is.na(GET_rows$bytes_reply), ]
GET_rows$bytes_reply <- as.numeric(GET_rows$bytes_reply)
```

```
## Warning: NAs introduced by coercion
```

```
max_bytes_reply <- max(GET_rows$bytes_reply)
hora_max_bytes_reply <- GET_rows$timestamp[which.max(GET_rows$bytes_reply)]
```

La fecha y hora donde hay mayor volumen de peticiones de tipo GET es 2024-05-30 13:04:57

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”?

```
filas_filtradas <- epa_http[grepl("\\.edu", epa_http$IPS) &
                            grepl("\\.txt", epa_http$URL) &
                            !is.na(epa_http$bytes_reply), ]
epa_http$bytes_reply <- as.numeric(epa_http$bytes_reply)
```

```
## Warning: NAs introduced by coercion
```

```
filas_filtradas <- epa_http[grepl("\\.edu", epa_http$IPS) &
                            grepl("\\.txt", epa_http$URL) &
                            !is.na(epa_http$bytes_reply), ]
suma_bytes_reply <- sum(epa_http$bytes_reply, na.rm = TRUE)
```

Se han transmitido segùn las condiciones indicadas en la premisa son 3.1186399^{8}

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 = “/”?

- Actualmente ya tenemos dividio el campo peticiòn en 3 partes con nombre de columna peticion, URL y protocolo, se procede a cambiar el nombre de peticion a Tipo a continuaciòn:

```
names(epa_http)[names(epa_http) == "peticion"] <- "Tipo"
epa_http
```

```
## # A tibble: 47,748 × 7
##    IPS      timestamp           Tipo  URL   protocolo code_respuesta bytes_reply
##    <chr>    <dttm>              <chr> <chr> <chr>              <dbl>       <dbl>
##  1 141.243… 2024-05-29 23:53:25 "\"G… /Sof… "HTTP/1.…            200        1497
##  2 query2.… 2024-05-29 23:53:36 "\"G… /Con… "HTTP/1.…            200        1325
##  3 tanuki.… 2024-05-29 23:53:53 "\"G… /New… "HTTP/1.…            200        1014
##  4 wpbfl2-… 2024-05-29 23:54:15 "\"G… /     "HTTP/1.…            200        4889
##  5 wpbfl2-… 2024-05-29 23:54:16 "\"G… /ico… "HTTP/1.…            200        2624
##  6 wpbfl2-… 2024-05-29 23:54:18 "\"G… /log… "HTTP/1.…            200         935
##  7 140.112… 2024-05-29 23:54:19 "\"G… /log… "HTTP/1.…            200        2788
##  8 wpbfl2-… 2024-05-29 23:54:19 "\"G… /log… "HTTP/1.…            200         124
##  9 wpbfl2-… 2024-05-29 23:54:19 "\"G… /ico… "HTTP/1.…            200         156
## 10 wpbfl2-… 2024-05-29 23:54:19 "\"G… /log… "HTTP/1.…            200        2788
## # ℹ 47,738 more rows
```

Se procede con el càlculo del total de peticiones segùn la pregunta:

```
total_peticiones <- sum(epa_http$URL == "/")
```

Las petciones que buscan directamente la URL = “/” son 2382

Pregunta 6

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

```
cantidad_peticiones_no_HTTP_0_2 <-sum(epa_http$protocolo != "HTTP/0.2")
```

Las peticiones NO tienen como protocolo “HTTP/0.2” son 47748

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.