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?

Respuesta: **Descriptiva**

La pregunta es de naturaleza descriptiva porque busca describir y resumir la información disponible en el registro de vehículos, sin buscar establecer relaciones causales o hacer predicciones sobre el comportamiento futuro de los vehículos en 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?

Respuesta: **Exploratoria**

La pregunta es de naturaleza exploratoria, ya que busca explorar posibles patrones o tendencias en los datos de visualización y establecer relaciones entre la preferencia de género literario y la edad de los 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?

Respuesta: **Predictiva**

La pregunta es de naturaleza predictiva, ya que busca hacer una predicción sobre el comportamiento futuro de las peticiones que provienen de una red de telefonía específica, basada en el análisis de datos pasados. También tiene un componente exploratorio al preguntar si se ha 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: **Inferencial**

La pregunta es de naturaleza inferencial, ya que busca inferir la pertenencia de un usuario a uno o varios grupos de preferencias de productos en base a su historial de compras.

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:

Para resolver este problema, se podría seguir los siguientes pasos:

1. Obtener los registros de conexiones TCP de cada máquina de cada trabajador involucrado. Estos registros deben ser exhaustivos y contener información sobre las conexiones entrantes y salientes, la dirección IP de origen y destino, el puerto y el protocolo utilizados (TCP, UDP, etc.).
2. Limpiar y preprocesar los datos para eliminar registros irrelevantes o duplicados y para normalizar la información (por ejemplo, eliminando información redundante o convirtiendo direcciones IP en nombres de host). Se podría utilizar herramientas de procesamiento de lenguaje natural y algoritmos de detección de anomalías para identificar patrones sospechosos en los datos.
3. Formular preguntas para resolver el problema, por ejemplo: ¿Hay algún patrón en las conexiones TCP que sugiera que un trabajador está usando la red para fines no relacionados con el trabajo? ¿Se puede identificar a los trabajadores que han habilitado servicios no autorizados en la red? ¿Hay algún patrón que sugiera que los trabajadores han accedido a sitios web o servicios no autorizados?
4. Explorar los datos utilizando gráficos y visualizaciones para identificar patrones y tendencias. Por ejemplo, se podrían crear gráficos de barras para mostrar la cantidad de conexiones TCP realizadas por cada trabajador, gráficos de línea para mostrar la evolución temporal de las conexiones TCP, o mapas de calor para mostrar los patrones de conexión entre direcciones IP y puertos.
5. Utilizar técnicas de aprendizaje automático, como el clustering o la clasificación, para agrupar a los trabajadores según su comportamiento en la red y para identificar patrones sospechosos.
6. Comunicar los resultados a la dirección y a los trabajadores involucrados. Los resultados deben ser claros y precisos, y deben incluir recomendaciones para mejorar la seguridad de la red y prevenir futuras incidencias. Es importante que la comunicación sea respetuosa y que se evite acusar a los trabajadores sin pruebas concluyentes.

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)

```
#cargar librerías
install.packages("stringr")
library(readr)
library(stringr)
library(dplyr)
library(tidyr)
library(lubridate)

#cargar datos
epa_http <- read_table("Data Science/epa-http.csv", col_names = FALSE)

#nombre columnas
names(epa_http)<- c("URL", "Time", "Tipo", "Recurso", "Protocolo", "Post", " Bytes ")

#número de columnas
nrow(epa_http)
[1] 47748

#número de filas
ncol(epa_http)
[1] 7
```

2. Valor medio de la columna Bytes

```
# Convierte a numero la columna Bytes
epa_http$` Bytes `<- as.numeric(epa_http$` Bytes `)
Warning message:
NAs introduced by coercion

# Valor medio
 mean(epa_http$` Bytes `, na.rm=TRUE)
[1] 7352.335
```

Pregunta 2:

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

```
# Filtra .edu
edu_filter <- filter(epa_http, str_detect(epa_http $URL, ".edu"))

# determinando cuantas son
nrow(edu_filter)
[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"?

```
# Limpia la columna Tipo
epa_http$TipoLimp<- str_sub(epa_http$Tipo, 2)

# Filtra GET
get_filter <- filter(epa_http, str_detect(epa_http $TipoLimp, "GET"))
View(get_filter)

# Formato fecha y hora a columna Timestamps
get_filter$Timestamps <- as.POSIXct(get_filter$Time, format = "[%d:%H:%M:%OS]")

# Obtener la hora en cada registro
get_filter$Hora <- format(get_filter$Timestamps, format = "%H")

# Contabiliza numero de registros por hora
Conteo <- table(get_filter$Hora)

# Hora con el mayor numero de registros
x3 <- names(sort(Conteo, decreasing = TRUE)[1])
x3
[1] "14"
```

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"?

```
# Filtra .edu
edu_filter <- subset(epa_http, grepl("\\.edu", URL))

# Filtra .txt
txt_filter <- subset(edu_filter, grepl(".txt", Recurso))

# Suma de Bytes
 x4 <- sum(as.numeric(txt_filter$` Bytes `), na.rm = TRUE)
```

```
Warning message:
NAs introduced by coercion
 X4
[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 = "/"?

```
# Arreglo5
Arreglo5 <- subset(epa_http, Recurso == "/")
nrow(Arreglo5)
[1] 2382
```

Pregunta 6:

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

```
# Arreglo6
Arreglo6 <- subset(epa_http, !grepl("HTTP/0.2", Protocolo))
nrow(Arreglo6)
[1] 47747
```