Actividad Evaluable 3

Descripción

MÓDULO	**Máster en Gestión de la Ciberseguridad**
ASIGNATURA	**Data Driven Security**
Fecha Límite de Entrega	26 de enero de 2025, a las 23:59
Puntos	25% de la Nota Total. **Mínimo se requiere un 5 para poder ir al examen final.**
Carácter	Grupo (max 3 personas)

Enunciado:

En esta actividad se planteará una serie de preguntas relacionadas con los temas Introducción a R, Datos Elegantes y Gráficos, Bots Crawlers y Scrapping y Introducción a Machine Learning.

El estudiante debe responder a tales preguntas antes de la fecha límite de entrega. Se considerará tanto la corrección de las soluciones como su presentación.

La entrega debe incluir el documento RMarkdown (.Rmd) y el documento renderizado en formato HTML (opcional incluir el PDF). Para la entrega, se espera que el contenido este disponible a través de un repositorio de código público creado en Github específicamente para esta actividad, antes de la fecha límite de entrega.

Se considerará tanto la corrección de las soluciones como su presentación y el código utilizado para la obtención de los resultados.

Parte de esta actividad implica ejecutar código R. Éste debe poderse ejecutar directamente sobre un terminal nuevo en R o en RStudio. El código es imprescindible para la corrección del ejercicio.

Las entregas tardías serán marcadas como "tarde", y pueden NO ser evaluadas.

2. Análisis de logs de servidor usando R (parte II)

Obtención y carga de los Datos:

Queremos programar un script con el que podamos hacer una investigación forense sobre un fichero de logs de un servidor de tipo Apache. Los datos del registro del servidor están en el formato estándar e incluyen miles de registros sobre las distintas peticiones gestionadas por el servidor web.

Nuestro programa ha de ser capaz de obtener las respuestas de forma dinámica a las siguientes preguntas utilizando instrucciones de código en R:

1. **Descomprimir el fichero comprimido que contiene los registros del servidor, y a partir de los datos extraídos, cargar en data frame los registros con las peticiones servidas.**

2. **Incluid en el documento un apartado con la descripción de los datos analizados: fuente, tipología, descripción de la información contenida (los diferentes campos) y sus valores.**

Limpieza de los Datos

3. **Aprovechando que los datos a analizar son los mismos de la primera práctica, para esta entrega es imprescindible que los datos estén en formato de "datos elegantes".**

Esto incluye la **correcta codificación de cada columna de los datos en el tipo adecuado según su naturaleza** (numérico, cadena de caracteres, lógico, timestamp, factor, etc.).

Además, para esta entrega se valorará también que los datos estén limpios y sin elementos extraños como por ejemplo espacios o signos de puntuación no necesarios.

Exploración de Datos

4. **Identificar el *número único de usuarios* que han interactuado directamente con el servidor de forma segregada según si los usuarios**

han tenido algún tipo de error en las distintas peticiones ofrecidas por el servidor.

Hacer la distinción (*break down*) del número de usuarios en función de si estos han tenido algún tipo de error durante las interacciones con el servidor y el tipo de error, es decir, ofrecer el número de usuarios que no han tenido ningún error en una de las peticiones gestionadas por el servidor y el caso contrario, el número de usuarios que sí han experimentado algún error para una petición servida según la tipología del error.

Describir en el documento los distintos tipos de errores existentes en la muestra de datos y el número único de usuarios.

Análisis de Datos

5. **Analizar los distintos tipos de peticiones HTTP (GET, POST, PUT, DELETE) gestionadas por el servidor, identificando la frecuencia de cada una de estas. Repetir el análisis, esta vez filtrando previamente aquellas peticiones correspondientes a recursos ofrecidos de tipo imagen.**

Visualización de Resultados

6. **Generar al menos 2 gráficos distintos que permitan visualizar alguna característica relevante de los datos analizados.**

   Estos deberán representar por lo menos 1 o 2 variables diferentes del data frame. Describid el gráfico e indicad cualquier observación destacable que se pueda apreciar gracias a la representación gráfica.

7. **Generar un gráfico que permita visualizar el número de peticiones servidas a lo largo del tiempo.**

Pista: Es imprescindible haber codificado correctamente el tipo de cada columna, y en particular, el de la columna usada para representar el momento en que se sirve la respuesta. Con un formato adecuado, será más fácil potencialmente extraer nuevas columnas derivadas de la original que puedan ayudar en la representación de la información (día, hora, etc.)

Clústering de datos

8. **Utilizando un algoritmo de aprendizaje no supervisado, realizad un análisis de clústering con k-means para los datos del servidor.**

    - Para este análisis debéis repetir la ejecución del modelado con distintos valores de *k* (número de clústeres) con al menos 2 valores diferentes de k.
    - A fin de retener algo de información sobre el recurso servido, generad una columna numérica derivada de esta con el número de caracteres de la URL servida.

    Pistas:

    - *La función one_hot de la librería "mlops" permite convertir fácilmente una columna factor a distintas columnas numéricas que representan el valor de la variable factor de forma que pueda usarse en algoritmos que trabajan únicamente con valores numéricos.*

```
library(mltools)
library(data.table)
epa_http_one_hot <- one_hot(as.data.table(epa_http), sparsifyNAs = TRUE)
```

    - *Descartad todas aquellas columnas que no sean numéricas y por lo tanto no pueden usarse directamente con el algoritmo k-means*

    - *La función de k-means no funciona con NAs, aseguraros que los datos que tienen NA son utilizados siempre que sea posible, por ejemplo, imputando el número de bytes servidos si por ejemplo sabemos que en el caso de 404 se deniega el acceso al recurso (0 bytes servidos).*

9. **Representad visualmente en gráficos de tipo scatter plot el resultado de vuestros clústering y interpretad el resultado obtenido** (describid las características de los distintos grupos) con los 2 valores distintos de k probados en el apartado anterior en función de los valores de las variables y el número de clúster asignado.

    Pista:

    Para hacer la interpretación, podéis probar generando gráficos con las distintas combinaciones de variables en los ejes x e y. Alternativamente, buscad X casos aleatorios para cada clúster e intentad descifrar la segmentación ofrecida por k-means.