Actividad Evaluable 1

Descripción

MÓDULO	**Máster en Gestión de la Ciberseguridad**
ASIGNATURA	**Data Driven Security**
Fecha Límite de Entrega	17 de diciembre de 2024, a las 23:59
Puntos	20% de la Nota Total.
Carácter	Grupo (máx. 3 personas)

Enunciado:

En esta actividad se planteará una serie de preguntas relacionadas con los temas vistos en las sesión 1. Los estudiantes deben responder a tales preguntas en un documento, de forma clara y concisa.

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. Por favor, entregad a tiempo.

1. Data Science

Data Science implica trabajar con datos con tal de obtener información / conocimiento / sabiduría.

Cuando tratamos con datos, hemos de saber qué buscamos, o qué queremos conocer.

La pregunta que nos hacemos antes de realizar un análisis de datos es lo más importante:

- La pregunta nos ayuda a guiarnos sobre la exploración de datos
- Hay que evitar creer que los datos nos inspirarán "algo", y entonces veremos de qué sirven

PREGUNTAS SOBRE LOS 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?

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.

2. Introducción a R

El segundo apartado de la práctica consiste en el análisis de un fichero de registro de peticiones HTTP, que debéis descargar el fichero adjunto: epa-http.zip, cargar en R, y realizar un análisis

Se recomienda tener cierto nivel de familiaridad y al alcance los *cheatsheet* de las distintas librerías mencionadas en las sesiones de teoría para un análisis más fácil:

- readr

- stringr

Alternativamente, recordad que podéis consultar la sección de ayuda de RStudio y buscar en la documentación los parámetros, así como ejemplos de uso (al final de cada página de documentación) para las funciones (escribiendo *?<nombre-funcion>* o presionando F1 sobre el nombre de la función.

Para las siguientes preguntas se requiere usar R. Indica en este documento para cada pregunta el resultado obtenido, describiendo a grandes rasgos el procedimiento seguido para la obtención de la respuesta, justificando cada decisión tomada a la hora de manipular los datos (descartar, agrupar, transformar, etc).

Asegúrate de entregar también el código en un fichero aparte, para poder ejecutarse directamente en un terminal limpio de R.

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. Cuales son las dimensiones del dataset cargado (número de filas y columnas)
2. Valor medio de la columna Bytes

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, ¿cuantas pertenecen a una IP claramente educativa (que contenga ".edu")?

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"?

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"?

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 = "/"?

Pregunta 6:

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