GESTIÓN DE DATOS: TRABAJO FINAL

El trabajo final consiste en aplicar las diversas técnicas y los recursos vistos en la clase a un problema real, usando principalmente R. Los datos utilizados pueden provenir de una fuente o de varias fuentes; pueden ser de alguna empresa local, o pueden ser datos disponibles en internet, pero que estén en una forma inicialmente no pre-procesada.

Dependiendo de la naturaleza de los datos, estos datos deben ser almacenados en algún sistema de nube que así lo permita. Por ejemplo, si se trata de datos estructurados, se puede crear un servidor usando Amazon RDS o administrando directamente una base de datos como MySQL en una instancia de Amazon EC2. Si se trata de datos no estructurados, se pueden almacenar usando Atlas MongoDB, por ejemplo, o una instancia de Amazon EC2 donde se debe previamente instalar MongoDB. De ser necesario, se puede también utilizar alguna otra base de datos NoSQL, de preferencia usando algún servicio de Amazon.

Luego de almacenados, los datos deben ser pre-procesados usando R. Se puede, por ejemplo, realizar una o varias consultas SQL desde R para obtener "data frames" en R que representen todos los datos o parte de los mismos. Si se estuviese utilizando, por ejemplo, MongoDB, se puede también convertir los datos, o parte de los mismos, al formato de "data frames", según sea necesario. Se puede, incluso, juntar datos provenientes de ambas (o más) fuentes, de ser necesario.

Los datos en formato de "data frames" pueden luego ser explorados y, luego de un análisis de la naturaleza de los posibles datos perdidos, se puede realizar una imputación de dichos datos perdidos. Igualmente, se debe realizar un análisis de valores atípicos univariados o multivariados, escogiendo los atributos que se considere necesarios (justificando la elección).

Finalmente, se debe visualizar los aspectos que se considere más relevantes de los datos. En particular, sería conveniente utilizar un reporte interactivo realizado, por ejemplo, usando Shiny en R, y subirlo a la nube (en shinyapps.io o algún otro servidor). Se puede usar algún otro sistema para la creación de reportes interactivos, pero se debe preferir el uso de R.

Lo descrito engloba de manera general lo que se espera que tenga el trabajo final, pero se puede añadir mayores análisis, consultas, y detalles, de considerarse pertinente.

La nota del trabajo final tendrá dos componentes:

- Reporte escrito (60 %)
- Presentación oral (40%)

Reporte escrito

El reporte escrito debe contener las siguientes partes

- Introducción (2 puntos). Se debe brindar una introducción general al tema que se va a tratar. Por ejemplo, qué es lo que se va a hacer, cuáles son los datos que se va a utilizar, cuál es su naturaleza, cuál es la importancia de dichos datos, etc.

- Metodología (6 puntos). Se debe brindar una descripción del procedimiento seguido. Se debe, por ejemplo, indicar las herramientas utilizadas para el almacenamiento y procesamiento, el flujo de los datos, un muy pequeño resumen de las técnicas que se va a utilizar, etc. Se puede utilizar esquemas, en caso sean útiles para mejor ilustración.

- Resultados (6 puntos). Debe contener los resultados obtenidos usando la metodología propuesta. Se puede incluir capturas de pantalla y gráficos que muestren lo que se ha obtenido. Los gráficos deben estar debidamente comentados y justificados para poder obtener un puntaje completo.

- Conclusiones y Recomendaciones (2 puntos). Se debe brindar algunas conclusiones de sobre lo que se ha realizado en el trabajo y los resultados obtenidos, así como recomendaciones generales para mejorar el trabajo o para tener en cuenta en futuros trabajos.

- Bibliografía (1 punto). Se debe incluir una bibliografía de fuentes confiables que han sido utilizadas durante el desarrollo del trabajo.

- Anexo (3 puntos). El anexo debe contener el detalle del código utilizado, debidamente comentado. Debe ubicarse al final del reporte escrito. Alternativamente, se puede crear un repositorio (por ejemplo, en github) y se puede brindar la dirección del repositorio.

Presentación oral

La presentación oral debe contener los puntos más importantes del trabajo final. Puede seguir el mismo esquema que el desarrollado en el reporte escrito, o puede contener un esquema diferente (principalmente en el orden de presentación de metodología y resultados) si ayuda a la presentación. Si se considera necesario, se puede incluir una demostración “en vivo” de parte de los resultados.

Todos los miembros del equipo deben realizar una parte de la presentación. Se considerará la claridad de la presentación, así como el dominio del tema. La nota de la presentación oral será de un 70% grupal y de un 30% individual.

Al final de la presentación, cada grupo debe formular al menos 1 pregunta al grupo que está presentando. La respuesta puede ser dada por cualquier miembro del grupo y puede ser complementada por algún otro integrante.

El tiempo de la presentación será entre 15 y 20 minutos por grupo, y el tiempo para las preguntas será de aproximadamente 5 minutos. La presentación se debe realizar con las cámaras activadas.