Curso Introducción a la Minería de Datos:
Instrucciones para el Hito 1

El proyecto del curso se debe realizar en grupos de 5 personas. Para desarrollar sus proyectos deberán coordinarse fuera del horario de clases de forma presencial o usando plataformas como Discord, Google Meet, Google Hangouts, Whatsapp, o incluso un google doc/slides colaborativo. Se les aconseja hacer una o dos reuniones semanales para conversar del proyecto y asignarse partes del trabajo. La idea es que todos y cada uno de los integrantes del grupo hagan una parte del estudio, para luego editar el informe entre todos.

Lo que se espera del Hito 1: Cada grupo debe elegir un dataset real, ya sea público o privado para realizar su proyecto semestral de minería de datos (ver sugerencias aquí, pero incluso pueden recolectarlo Uds. mismos, u obtenerlo de otra fuente, la elección es libre). El proyecto será evaluado en 3 etapas (Hito 1, Hito 2 e Hito 3).

Para el Hito 1 los alumnos deben mostrar sus avances iniciales en sus proyectos en una **A. Presentación y un B. Informe que deben ser enviados vía Tareas en u-cursos**. Los detalles de la evaluación se presentan a continuación:

A. Presentación (presencial, 3 minutos): Las presentaciones serán evaluadas de acuerdo a la siguiente pauta (no olvidar ningún punto!):

1. **Motivación:** ¿Cuál es el contexto general del tema/problema/datos de estudio?¿Por qué podría ser interesante estudiar estos datos?

2. **Exploración de Datos:** Características más relevantes e interesantes del dataset (estadísticas de resumen, gráficos, etc.).

   Comentario: en muchos casos es necesario explorar los datos a un nivel de análisis diferente al de los datos originales. Esto se puede lograr haciendo **agregación**. Por ejemplo, supongamos que tenemos un dataset de indicadores socio-económicos para varios países a lo largo del tiempo, donde el tiempo está a nivel de mes. Si quiero explorar las diferencias y similitudes entre varios países es necesario **agregar** los indicadores de alguna forma (ej: sumar, promediar) y así poder crear una tabla donde tenga una fila por cada país. Luego podría graficar los países en un scatterplot. Adicionalmente, a veces es necesario hacer un **join** entre dos o más tablas para poder realizar su exploración. Tengan en cuenta que distintas tablas pueden tener objetos en distintos niveles de granularidad (ej: una tabla está a nivel de días y otro a nivel de semanas) y será necesario usar **agregación** para poder hacer el **join**.

3. **Preguntas y problemas:** Dada la exploración anterior y su motivación original, formular preguntas que se pueden responder mediante la minería de datos y que se puedan vincular a la problemática planteada en la motivación.

    Ejemplos:
    1. ¿Es posible predecir la variable X en función de Y, Z, K?
    2. ¿Existen grupos importantes de ejemplos que se comporten de manera similar de acuerdo a algún criterio?
    3. ¿Existen asociaciones frecuentes inesperadas entre grupos de atributos?

Comentarios:

- Los ejemplos anteriores son genéricos. Las preguntas que ustedes deben plantear deben estar conectadas con la problemática abordada por sus datos y donde puedan argumentar que responderlas tenga alguna utilidad para alguien.
- Traten de pensar en preguntas que puedan ser abordadas usando técnicas que verán en el curso: (ej: clasificación, clustering, regresión, análisis de asociación). Si no entienden la idea general de esas técnicas, por favor pregunten.
- Preguntas que puedan ser respondidas trivialmente mediante la exploración no son válidas en esta etapa. Por ejemplo: ¿existe una correlación positiva entre X e Y? De hecho, la idea es que los resultados de su exploración los ayuden a formular buenas preguntas. La gran diferencia entre una pregunta que se puede responder con exploración y otra que necesita técnicas de minería de datos, es que éstas últimas requieren la construcción de modelos que encuentren relaciones entre múltiples atributos y/o ejemplos.
- En algunos casos pueden llegar a plantear de forma complementaria una **hipótesis**. Una hipótesis (de manera muy simplificada), es una creencia actual o resultado esperado asociada a sus preguntas. Por ejemplo: la literatura sugiere que los países latinoamericanos se comportan de manera más similar a los africanos que a los europeos respecto al criterio X, entonces esa será la hipótesis para esa pregunta. En ciertos casos, el objetivo es refutar la hipótesis actual mediante el análisis de datos.
- Sean conscientes que puede que sus datos actuales no permitan responder alguna de sus preguntas. En ese caso, traten de pensar si es posible conseguir un dataset complementario que haga que esa pregunta sí pueda ser respondida. Si eso no es factible, descarten la pregunta.

4. **Calidad general de la presentación:** Preparación del grupo y claridad en la exposición.

   Comentarios:
   - Es extremadamente importante que construyan una narrativa para su presentación. **Traten de contar una historia**, que cada parte se conecte con la anterior. Eviten mostrar cosas de manera aislada. Es muy probable que tengan que replantear lo que van a contar en una etapa anterior en base a lo que encontraron en una fase siguiente. Por ejemplo: puede que al diseñar las preguntas se den cuenta que faltó hacer algo en la exploración. En ese caso refinan su exploración para que su historia final quede más sólida.

A.1 Ejemplo

Supongamos que tenemos una dataset que contiene reviews de cervezas, donde cada cerveza es evaluada con una nota por cada usuario que la consumió. Adicionalmente, nuestro dataset contiene columnas que describen el tipo de cerveza (pale ale, ipa, etc), su grado alcohólico, país de origen, si es de cebada o trigo, marca, año de elaboración, entre otros.

Motivación: podríamos comentar que actualmente el mercado de cervezas artesanales está creciendo, donde más y más personas están produciendo sus propios productos dado el consumo por año per cápita, la facilidad de entrar al mercado, etc.

Exploración de datos: dado que cada cerveza tiene una puntuación por usuario, podríamos mostrar el promedio de ranking por cerveza, así como una distribución usando boxplot. También cuánto cada usuario ha hecho un review a una cerveza, o el número de cervezas por país y su respectivo ranking de manera agregada. En el análisis exploratorio, uno podría, por ejemplo, observar que algunas cervezas podrían tener bajo o alto ranking por alguna razón que tal vez (a simple vista), no tengamos muy clara, o estadísticamente no sea fácil de deducir.

Preguntas y problemas: en base a nuestra motivación y análisis exploratorio, podrían surgir ciertas inquietudes que se ven representadas en los datos. Por ejemplo, nos haría re-pensar que tal vez hay características en los datos que nos permita, conocer por ejemplo, que ciertas cervezas con específicas cualidades podrían tener mejor (o peor) aprobación en el mercado. Por ende, preguntas que se pueden desprender serían algunas como estas:

- ¿Existen características específicas de las cervezas que permitan tener mejor o peor aprobación del público?
- ¿Sería posible conocer el ranking (aproximado) de una nueva cerveza que entra al mercado considerando sus características?
- ¿Es posible encontrar grupos de cervezas (rating en común o similares) a partir de las cualidades de cada cerveza?

B. Reporte BREVE (equivalente a aproximadamente 5 páginas impresas) presentados en una página Web.

El informe debe contener la información de la presentación de manera más detallada.
Al final del informe se debe mencionar cuál fue la contribución exacta de cada miembro al proyecto (ej. John Doe estuvo a cargo de la limpieza de datos y del análisis presentado en las tablas xx y xx, también redactó la sección xx del informe).

Estructura sugerida:

1) Introducción: plantear el problema y la motivación.

2) Exploración de datos.

3) Preguntas y problemas.

4) Se evaluará positivamente el incluir código fuente utilizado para generar sus estadísticas y análisis (e.g. por ej. generar la página usando jupyter notebook, o markdown R, o poner enlaces a sus scripts. Mientras más reproducible el trabajo, mejor. El código fuente no se cuenta dentro del largo de las 5 páginas siempre que esté colapsado o incluido en anexos).