Curso Introducción a la Minería de Datos:
Instrucciones para el Hito 2

Dada la situación sanitaria, se les pide que desarrollen los trabajos coordinándose online. Se les aconseja hacer al menos una o dos reuniones semanales para conversar avances y distribuir partes del trabajo. La idea es que todos y cada uno hagan una parte del estudio, para luego editar el informe entre todos.

Lo que se espera del Hito 2: Mejorar exploración y preguntas del hito 1 (en especial debilidades señaladas en la corrección). Desarrollar una propuesta metodológica de los experimentos de minería de datos a realizar en su proyecto en relación a las preguntas y metas planteadas, y llevar a cabo **al menos un experimento** de esta propuesta. En detalle:

1. **Mejorar hito 1:**
   - Ahora que entienden mejor en qué consisten las técnicas de Minería de Datos pueden refinar sus preguntas y alinearlas con lo que permiten hacer las herramientas.
   - Incluir datos (o datasets) que puedan complementar (o aportar más valor) para responder las preguntas del Hito 1. Esto significa que algunas preguntas podrían no responderse con el análisis exploratorio inicial, lo que conllevaría a agregar más columnas (o filas).
   - Mejoren la fase exploratoria para incorporar sus nuevas preguntas y/o fuentes de datos.
   - Es posible que su dataset actual no cumpla las características mínimas para el proyecto, por lo que esta instancia permite replantear la factibilidad de continuar con estos datos y buscar un problema nuevo.
   - Incluir las sugerencias del equipo docente y los comentarios de sus compañeros para consolidar el hito 1.

2. **Propuesta metodológica experimental inicial:**
   - Describir la metodología experimental asociada para responder todas sus preguntas. Esto incluye preprocesamiento adicional en caso de ser necesario, plantear los modelos a utilizar (técnicas de Minería de Datos) y las técnicas de evaluación correspondientes. Esta propuesta es un contrato que deberán luego llevar a cabo en el Hito 3. **ARGUMENTEN TODAS SUS DECISIONES.**
   - Ejemplo (para una pregunta, debe repetir para cada pregunta): para responder la pregunta X vamos a agregar los datos por país, luego reduciremos las dimensiones usando las técnicas Z y K o combinaremos los atributos H y L mediante una suma, para luego aplicar clustering. Elegiremos la mejor solución de clustering comparando los algoritmos E, F, G y los compararemos con el

método de visualización con diferentes números de clusters. La idea es que los resultados de este experimento nos permitirán responder la pregunta X mediante las métricas A, B, C. Den argumentos para todas las componentes de sus metodología. ¿Por qué enfocarse en técnicas Z y K?, ¿Por qué evaluar con la métrica A,B?¿Qué haremos si no encontramos resultados significativos? (por ejemplo probaremos clasificación).

- Comentarios:
    - Si van a usar técnicas supervisadas (clasificación o regresión) se recomienda comparar varios modelos en sus experimentos (árboles, KNN, SVM, etc). Si van a usar técnicas de resampling por tener clases desbalanceadas (oversampling de clase minoritaria, subsampling de clase mayoritaria, SMOTE), no transformen sus datos de testing.
    - Para experimentos con técnicas de clustering es muy importante que puedan hacer un análisis cualitativo de sus clusters. Por ejemplo, pueden mirar algunos ejemplos por cluster y tratar de entender qué es lo que representan. Pueden incluso tratar de ponerle un nombre a sus clusters. Recuerde que el clustering debe evaluarse y que se han visto técnicas para hacerlo en clases.
    - A veces es posible etiquetar una muestra de sus datos de manera manual para poder aplicar técnicas de clasificación. Si no tienen etiquetas, las pueden crear ustedes mismos.
    - Pueden usar técnicas de análisis de datos que ustedes conozcan o quieran aprender pues no se enseñan en este curso. Acá existen muchas opciones, como test de hipótesis, series de tiempo, procesamiento de imágenes, procesamiento de lenguaje natural, etc..

3. **Resultado preliminar:**
    - Implementen al menos uno de los experimentos planteados en la parte anterior y discutan sus resultados. Tenga en consideración que si hacen un experimento deben probar varios parámetros dentro de los cuales está el algoritmo (por ej. Si usa clasificación debe comparar varios algoritmos diferentes en un mismo experimento) ¿Permiten estos resultados preliminares responder la pregunta correspondiente? Recuerden que para el hito 3 deberán implementar todos los experimentos.

A.1 Ejemplo

Retomando el ejemplo de las cervezas del hito anterior en el cual consideramos las siguientes preguntas y problemas:

- ¿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?

Mejorar hito 1: del análisis exploratorio pudimos apreciar que nuestras preguntas sí pueden ser respondidas a través de los datos. Adicionalmente, encontramos un dataset que incluye el consumo per cápita de cerveza en el mundo, por lo que complementaremos el análisis con estos datos. Otro caso opuesto sería que una o más preguntas no pudieran ser respondidas. En estos casos se deberían replantear qué preguntas o problemas es posible responder con el análisis exploratorio, incorporando nuevas fuentes o cambiando el dataset. En este último caso, se espera que los grupos se contacten con el equipo docente para ver el caso personalmente.

Propuesta experimental:

- En el dataset existen cervezas que no tienen ranking o que algunos ranking no están en una unidad estándar. Por lo tanto, pre-procesaremos los datos para limpiar aquellos registros que no tienen nota y estandarizaremos los valores de ranking en 2 escalas (cualitativo y cuantitativo).
- Aplicaremos transformaciones al precio ya que este considera diferentes unidades, de modo de estandarizarla en una sola (CLP por ejemplo).
- Extraeremos características del texto de los reviews, por lo que representaremos de forma vectorial el texto para entender si este puede (o no) entregar mayor información para las tareas planteadas.
- Dado que nuestras preguntas 1 y 2 son de carácter predictivo, nos focalizaremos de manera particular en clasificación, donde proponemos utilizar nuestro dataset para crear un modelo que permita estimar el ranking de una cerveza a partir de un conjunto 50 características (el número es solo un ejemplo para esta propuesta, ustedes deben completar con información real).
- Para evaluar la calidad de la clasificación compararemos diversos algoritmos, utilizaremos las métricas tradicionales como F1, precision y recall, aplicando k-fold cross validation o un particionado de 80-20 para entrenamiento y testeo respectivamente. Nuestra idea de esto es no sobre-ajustar el modelo y que este aprenda de subsets de entrenamiento distintos.
- También aplicaremos técnicas de clustering para encontrar de manera natural si las características de nuestro dataset son suficientes para encontrar grupo de cervezas similares (pregunta 3).
- Probaremos múltiples combinaciones en el número de clúster así como distintos enfoques de clustering (jerárquico y particional).
- También probaremos usando distintos subconjuntos de atributos al hacer clustering para evaluar si los ejemplos se agrupan de manera distinta cuando consideramos información diferente.

- Para evaluar los clusters, utilizaremos el enfoque visual así como también la estimación de métricas tales como cohesión y separación.

Experimento preliminar:

- Aplicar técnicas de clustering a mis cervezas (K-means, DBSCAN), evaluar y discutir los resultados.

B. Entregables:

1. **Reporte BREVE (unas 8 páginas impresas, donde 5 corresponden al hito 1 y 3 al hito 2) presentados en una página Web. 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). El informe debe ser enviado en un archivo que contenga todo lo necesario para su visualización vía u-cursos.**

   Estructura sugerida:
   a. Introducción: plantear el problema y la motivación.
   b. Exploración de datos.
   c. Preguntas y problemas.
   d. Propuesta experimental.
   e. Se evaluará positivamente el incluir código fuente utilizado para generar sus estadísticas y análisis. Por ejemplo, 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. A su vez, gráficos o tablas que no sean relevantes pueden ser anexadas al final del documento. **Ojo, en este hito tienen que diseñar y escribir todos sus experimentos e implementar al menos uno.**

2. **Una presentación en formato PPT o PDF de no más de 10 slides que resuma sus mejoras al hito 1, su propuesta experimental y su resultado preliminar (no es necesario hacer un video para este hito). Los profesores evaluarán principalmente el PPT y se apoyarán en el informe para entender detalles.**

3. **Para este hito no se solicitará la grabación de un video. La presentación (PPT) que deben subir es un resumen de lo expuesto en el informe.**