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

Dada la situación del COVID-19, se les pide encarecidamente que desarrollen los trabajos coordinándose online, ya sea por Google Meet, Google Hangouts, Whatsapp, o incluso un google doc/slides colaborativo. Investiguen datasets que les parezcan interesantes online y elijan uno para trabajar. Se les aconseja hacer una o dos reuniones semanales para conversar y asignarse 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: Mejorar el hito 1 y desarrollar una propuesta experimental.

1. **Mejorar hito 1:**
    - Ahora que entienden mejor en qué consisten las técnicas de DM pueden refinar sus preguntas.
    - 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 experimental inicial:**
    - Describir la metodología experimental asociada para responder todas sus preguntas. Esto incluye el preprocesamiento de datos (de ser necesario posterior al Hito 1), plantear los modelos a utilizar (técnicas de DM) y las técnicas de evaluación correspondientes. Esta propuesta es un contrato. En el hito 3 deberán llevar a cabo la metodología propuesta. **ARGUMENTEN TODAS SUS DECISIONES.**
    - Por ejemplo: 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 el algoritmo de clustering G. 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?.

- 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.
    - 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 que no se enseñen en este curso. Acá existen muchas opciones, como test de hipótesis, series de tiempo, procesamiento de imágenes, procesamiento de lenguaje natural, etc..

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 debieran re-plantear que 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, 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 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 teniendo en cuenta su ranking (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.

B. 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 ANTES de las presentaciones.

Estructura sugerida:

1) Introducción: plantear el problema y la motivación.
2) Exploración de datos.

3) Preguntas y problemas.
4) Propuesta experimental.
5) 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 no tienen que correr sus experimentos, sólo diseñarlos y escribirlos.**