Evaluación Hito 1
Observaciones Generales de las Presentaciones hechas por los Profesores

- **Todas las observaciones más importantes del Hito 1 se encuentran mencionadas en este informe. Se espera que todos los grupos lean estos comentarios en detalle y que indiquen en el informe del Hito 2 qué comentarios creen aplican a su grupo y cómo lo consideraron en su avance.**
- **Para los grupos que deseen conocer más detalles de su evaluación particular pueden consultarlo a los profesores, previa cita agendada por mail. También, las consultas pueden realizarse el día Viernes 13 de Septiembre durante la "Sesión de trabajo guiado para el H2".**
- **Además deben indicar en el informe 2, cuáles comentarios de sus compañeros (evaluaciones escritas publicadas en u-cursos) les parecieron más adecuados y por qué. También pueden indicar en el informe si algún comentario de sus pares no les pareció útil o adecuado.**

Lo que se evaluó en el Hito 1: *En este primer hito los alumnos deben mostrar sus avances iniciales en sus proyectos, específicamente una exploración inicial de los datos. Esta exploración debe haber permitido al grupo decidir cómo seguir hacia el hito 2. Es importante señalar que haber realizado la exploración es un requisito para tener nota aprobatoria en este Hito, independiente de los otros elementos evaluados.*

Los aspectos generales que los grupos debían abordar fueron: motivación, problema/objetivos/hipótesis, descripción de los datos y exploración. A continuación se presentan las observaciones para cada uno de estos ítemes.

1) ***Motivación:*** *¿Contexto general del tema/problema/datos de estudio?¿Por qué es de interés?*

- Es importante haber planteado una motivación clara para estudiar un conjunto de datos, esta motivación puede salir, por ejemplo, de las aplicaciones útiles que puede tener el problema.
- Un aspecto que puede ser mejorado por varios grupos fue la claridad con que se explicó la importancia del problema.

2) ***Hipótesis/Problemáticas iniciales:*** *¿Qué es lo que se desea analizar? o ¿Qué problema queremos resolver?¿Cómo se haría esto de forma preliminar?*

- En este punto lo ideal es haber sido absolutamente explícito y conciso en cuanto a cuáles son las preguntas que se quieren responder con los datos. Además, de indicar con qué hipótesis está trabajando el grupo (si es que la hay). La idea para el hito 2 es moverse en dirección a responder estas preguntas y/o validar la hipótesis. Por eso se

aconseja a los grupos que no hayan sido claros en este aspecto resolverlo antes de proceder hacia el hito 2.

- Varios grupos necesitan diferenciar entre **hipótesis** y **preguntas de investigación**. No son lo mismo, ni son intercambiables, aunque ambas sirven para iniciar la investigación basada en datos. Una **hipótesis** es una suposición o teoría sobre los datos que sirve para iniciar una investigación, por ejemplo: "Es posible localizar el epicentro de un sismo utilizando sólo la información de los mensajes publicados por usuarios en redes sociales". En este contexto, una **pregunta de investigación** sería: "¿Qué tipo de relación existe entre la actividad de los usuarios durante un sismo y su magnitud?". También se debe diferenciar de un **objetivo**, que en ese caso sería: "Calcular la correlación entre el volúmen de los mensajes y la intensidad reportada por el centro sismológico".
- Muchos grupos plantearon problemas que pueden responderse trivialmente explorando el dataset (e.g., existe una correlación entre la variable X y la variable Y). Es importante plantear problemas que pueden ser estudiados mediante técnicas vistas en el curso como clustering, reglas de asociación, clasificación y regresión.
- Otra sugerencia para los grupos que sí plantearon una hipótesis super concreta, como por (ejemplo ver si un factor x incide sobre otro y), es que consideren también otras alternativas probables para investigar. Así no se limitan a explorar un solo aspecto de sus datos.
- Por otro lado, el objetivo de su proyecto no puede ser "aplicar la herramienta X", como por ejemplo: "crear un modelo con redes neuronales". El objetivo debe tener relación con los datos, por ejemplo: comprobar si es posible predecir un cierto comportamiento en base a datos históricos. Luego, encontrar el mejor algoritmo/modelo de clasificación (que puede no ser redes neuronales).

*3) **Descripción de los datos que se van a utilizar:** Sus características más relevantes e interesantes (estadísticas, gráficos, etc).*

Este era el punto más importante para esta etapa, y por lo tanto, haber realizado una exploración a conciencia es requisito para aprobar el Hito 1.

Por eso mismo, antes de comenzar a trabajar en el Hito 2 todos los grupos deben haber invertido tiempo en realizar una exploración de su dataset, ya que esto les permitiría definir qué hacer el resto del proyecto. En base a la exploración pueden decidir qué preguntas de investigación tiene sentido intentar de responder y/o qué hipótesis podrían ser testeadas en base a los datos. Esto es especialmente importante para los grupos que mostraron una exploración nula o muy superficial.

Todos los grupos deben indicar el origen de sus datasets, como por ejemplo si son parte de un dataset de concurso de algún sitio de ML. Esto influye en la cantidad de preprocesamiento que

se debe realizar (los datos de concurso vienen bastante preprocesados) y sobre las tareas a realizar.

Importante: si se apoyan en análisis realizados por terceros (por ejemplo notebooks publicados en kaggle para un dataset) debe indicar claramente la fuente y cuál es el aporte original de uds. a este análisis.

También se debe evaluar si los datos con los que cuenta el grupo son suficientes para responder preguntas interesantes de investigación o validar las hipótesis planteadas. De lo contrario se debe considerar agregar más datos, inclusive de otras fuentes o cambiar de dataset por completo.

Algunos comentarios específicos:

- Para los grupos que tienen datos temporales (e.j. por año) es importante que hagan la exploración y el análisis separado por años. Lo mismo para los grupos que tengan datos geográficos (separar por regiones).

- Si tienen datos temporales o geográficos, tienen que definir a qué nivel de granularidad van a trabajar (a nivel día, a nivel semana, etc). Ojo que ese nivel de agregación permite obtener un volumen de datos razonables para responder sus preguntas. También tienen que tener esto en cuenta si quieren incorporar datos de otras fuentes. Tienen que pensar bien como consolidar los datos en una tabla única de granularidad bien definida.

- Para los grupos que quieren hacer análisis de sentimiento o emociones. Tienen que ver cómo van a extraer variables cuantitativas de emoción a partir del texto. Pueden usar un clasificador pre-entrenado o un lexicón de palabras. Hay un paquete Weka que pueden probar: https://affectivetweets.cms.waikato.ac.nz/

- Hay grupos que proponen realizar clasificación sobre sus datos sin discutir de dónde sacarán las etiquetas (labels) de sus instancias para entrenar sus modelos. ¡No se puede entrenar un clasificador sin un set de entrenamiento! Si sus datos no tienen etiquetas, replanteense sus objetivos. Una opción es enfocarse más en clustering otra es buscar la forma de obtener etiquetas en base a alguna heurística (supervisión a distancia). Por ejemplo, quiero clasificar tweets en base a sentimiento pero no tengo labels. Solución: bajarme tweets con emojis positivos y negativos, luego borrar el emoji del tweet y asumir el label del emoji como mi etiqueta.

- Si van a trabajar con una muestra de un dataset grande tomen en consideración la forma en que muestrean los datos. Una muestra aleatoria simple puede sub-representar grupos minoritarios. Consideran realizar muestras estratificadas.

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

- La mayoría de los grupos se vio bien preparados, sin embargo hubo unos pocos grupos que llegaron con presentaciones que no abordaban todos los temas que se pedían o que eran muy cortas (incluso para 7 minutos!) y carentes de contenidos. Eso no debe pasar, ya que hay una pauta clara con los objetivos de la presentación.

- Se aconseja a los grupos aprovechar bien el tiempo de la presentación. Es decir, hacer una selección de los aspectos más importantes del trabajo (para cada punto en la Pauta) en la presentación. El resto de los detalles deben ir en el informe.

- Es importante ser claro explicando qué significa cada tabla o gráfico que se está presentando. La idea de la presentación es interesar a la audiencia y transmitir el mensaje principal.

Comentarios sobre la "evaluación de pares":

- Se deben entregar comentarios y sugerencias con la intención de ayudar a sus compañeros a construir un mejor trabajo en base a lo que Ud. ha aprendido en el curso hasta ahora.

- Para que una crítica sea útil a la persona que la recibe, ésta debe formularse en un tono constructivo y explicando el por qué.

Con respecto a los Informes:

- Recuerden ponerle nombre a sus proyectos y número de grupo.

- Algunas páginas Web tenían faltas de ortografía, problemas en el encoding (se veían con caracteres extraños), o las imágenes no aparecían. Ahora no ese aplicó descuento por esas cosas, pero la próxima vez sí ya que no es aceptable en un trabajo profesional.

- La idea del informe es que sea un relato detallado de su proyecto. Debería indicar lo mismo que la presentación pero con más detalle, incluyendo scripts, gráficos, etc. Hubo grupos que se limitaron a incluir scripts sin casi ninguna explicación. Sin embargo, lo que se busca es que el reporte documente el trabajo para que otros puedan reproducirlo y llegar a las mismas conclusiones.