Ciencia de Datos 2021

Propuesta de Proyectos

26.abril.2021

1 Instrucciones generales

Hay dos tipos de proyectos:

- Un análisis de datos:

  El resultado debe ser una narrativa visual que se debe presentar de manera electrónica (como si fuera para un concurso de datos). Se trabaja en equipos de dos. Cada equipo tiene toda la libertad de elegir las preguntas de interés a las cuales se quieren enfocar (sugerencia: conviene acotar las preguntas de interés e ir por profundidad para que el resultado tenga más coherencia). La idea es por supuesto aprovechar lo más posible las herramientas que vimos en clase.
  Si aplica, enfócate a preguntas ligadas a Guatemala o de relevancia especial para todos (para nosotros).

  La mitad de la calificación es para evaluar los aspectos técnicos y la otra mitad es para la presentación / la narrativa.

- Discutir un artículo:

  El resultado debe ser un reporte donde se resumen las ideas principales del artículo usando la terminología del curso y que incluye al menos un ejemplo propio.

  La mitad de la calificación es para evaluar los aspectos técnicos, una cuarta parte para el ejemplo propio y lo restante para la presentación / la narrativa. Algunos de estos artículos podrían servir como inicio de su trabajo de graduación.

Es muy importante incluir siempre referencias con citas si uno se apoya en material/resultados de otras personas.

Fechas importantes:

- 7 de mayo: Elección del tipo de proyecto y elección de la base de datos/artículo.
- 14 de mayo: Entregar dos parráfos de lo que se pretende hacer.
- Semana de exámenes finales: entrega final.

Al igual que en el primer proyecto, se debe entregar

- Reporte técnico de su metodología / discusión,
- Presentación de resultados,
- Código o herramientas auxiliares utilizadas.

2 Proyectos

1. Estadísticas de salud por país WHO:
   https://www.who.int/data/gho/publications/world-health-statistics

2. Igualdad de género en casa:
   https://www.equalityathome.org/

3. Exceso de mortalidad en México
   http://www.dgis.salud.gob.mx/contenidos/basesdedatos/da_exceso_mortalidad_mexico_gobmx.html

   Comentario: lo interesante es por supuesto poder compararlo con los datos de años anteriores. Una posibilidad es a través de https://www.inegi.org.mx/app/descarga/?t=127&ag=00#microdatos.
   Lo inconveniente es que uno obtiene los datos tabulados. Requiere cierto trabajo a mano.

4. Datos informativos sobre Guatemala.
   Un punto de partida puede ser
   https://nathsmo.medium.com/sitios-de-datasets-sobre-guatemala-1647fdfb6f25

3 Artículos

1. Unmasking Clever Hans predictors and assessing what machines really learn (2019)
   https://www.nature.com/articles/s41467-019-08987-4
   Énfasis en método LRP.

   Ver también:
   https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130140
   https://www.youtube.com/watch?v=IwCi-DBqqiw

2. Minimum-Distortion Embedding (2021)
   https://arxiv.org/pdf/2103.02559.pdf

   Es casi una monografía; sugiero tomar una aplicación donde se ponen varios métodos que vimos bajo un marco común. Parte de la novedad es el algoritmo de optimización.

3. SoccerMix: Representing Soccer Actions with Mixture Models (2020)
   https://tomdecroos.github.io/reports/ecml_2020.pdf

   El código que acompaña el paper es en Python.

4 Otros recursos

- https://www.kaggle.com/
- https://paperswithcode.com/