Proyecto Final de Economía Computacional

Bienvenidos al proyecto final. El trabajo debe ser entregado en los mismos equipos que entregaron las tareas. La fecha final de entrega es el **28 de mayo**. Pueden escoger 1 de los 4 temas que les puse. En cada tema, les doy una problemática a resolver. Ustedes tienen decidir:

- El flujo de trabajo que deben seguir
- Qué algoritmo(s) utilizar para resolver la problemática
- El entregable para los clientes del proyecto.

Reglas de asignación: Jugaremos a first come first serve. Hay 3 proyectos y 7 equipos. En un proyecto habrá 3 equipos y en los otros 2 habrá sólo 2 equipos. Para elegir su tema, deben contestar al anuncio que haga en Canvas. Si se quedan sin espacio, deben mandar de vuelta su sencond best.

Evaluación: Dado que no les haré preguntas específicas en los proyectos, los voy a calificar con base en 3 criterios:

a) Robustez técnica de los algoritmos empleados (50%): Asegurense de justificar las decisiones que tomen al modelar y de mostra que estas son conscientes.
b) Limpieza del código (20%)
c) Narrativa del proyecto (30%): Una parte crucial de un buen data scientist es su capacidad de comunicación efectiva. En este sentido, les evaluaré la redacción, EDA y narrativa del documento que entreguen. Imagínense que yo soy su cliente en cada caso

Proyectos posibles:

1. **El impacto de Prospera/Focalización:**
   Datos: ENIGH 2018
   Objetivo: Eres un Data Scientist trabajando para el gobierno. Ante una escasez de recursos públicos, te piden hacer una evaluación del impacto del programa PROSPERA sobre:
   a) La alimentación del hogar: “Alguna vez se preocupó porque faltara comida?”
   b) La probabilidad de que los miembros del hogar tuvieran un empleo informal.
   c) La probabilidad de los integrantes terminaran la secundaria.

   Con base en tu análisis, se plantea focalizar el programar únicamente a los hogares cuyo impacto es alto y significativo en al menos 2 variables. De que tamaño quedaría el programa? Cuál sería el impacto total (vs sin focalizar)? Cuál sería el impacto promedio?

2. **Amazon Reviews:**

   Datos: Review_subset.csv
   Objetivo: Eres Data Scientist para Amazon. A la empresa le gustaría que elabores un reporte que:
   a) Identifique cuáles son los productos que mejor pronóstican cada calificación de reviews (1 a 5). Para esto, debes elaborar un modelo de predicción de la calificación con base en el tipo de producto y el texto de los reviews. Elaboren sobre cuál modelo eligieron y muestren todo el flow.
   b) Cuáles son los temas que más se hablan en cada review

c) Si se puede reemplazar el score número con un score de NLP (i.e. un análisis de sentimiento)

3. **Wheelie Wonka Bike Station ride prediction**

   Datos: hubway_stations.csv, hubway_trips.csv, weather.csv

   Objetivo: Wheelie Wonka es una empresa de bike sharing en Boston. Quieren tu ayuda para que los usuarios vean en la cantidad de bicicletas disponibles en cada estación en tiempo 'real'. Para ello, te dispones a predecir primero la duración del viaje. En tu entrega se te pide:

   a) Generar un modelo que pronostique la duración de los viajes (muestra porqué tu modelo es campeón).
   b) Mostrar patrones geográficos en los viajes observados en la base de datos. ¿Cuáles son los factores que hacen que un viaje duré más o menos?
   c) Con el modelo, genera una solución de negocio para estimar cuántas bicicletas habrá en cada estación por cada 10 minutos.