Examen OPI - Data Science (5 horas)

El examen está diseñado para resolverse en menos de 5 horas por alguien familiarizado con las técnicas, temas y datos. Por la naturaleza de las preguntas se les podría dedicar mucho más tiempo, pero sólo nos interesa conocer tu dominio de los temas así como tu capacidad de comprensión y planteamiento del problema. Busca el 80-20 en cada pregunta (el 20% del trabajo que te da 80% del valor en la solución)

Instrucciones:

- Resuelve la sección 1
- Selecciona la sección 2.a ó 2.b dependiendo del nivel con el que te sientas más cómodo. Sólo debes resolver una (no se darán puntos adicionales por resolver las dos)
- Envía tus respuestas a más tardar 48 horas después de recibir el correo
- Las respuestas podrán ser enviadas en un PDF, un notebook, o con un link a tu branch de GitHub
- Se dará un punto adicional si se envían en GitHub
- En promedio la gente sólo resuelve el 70% del examen. No te preocupes, nos interesa conocer tu forma de plantear los problemas.

Sección 1: Bebés (1 hora) Obligatorio

1. Para cada AGEB de la delegación Álvaro Obregón estima cuántos bebés de 0 a 6 meses de edad habitan ahí el día de hoy. Explica tu razonamiento en menos de 300 palabras. Enlista tus fuentes y presenta los resultados. (Hint: revisa el CPV 2010, puede ser útil)

Sección 2.a: Ecobici (4 horas) Intermedio

En la página de datos abiertos de Ecobici (https://www.ecobici.cdmx.gob.mx/es/informacion-del-servicio/open-data) baja los datos de movilidad de los últimos 3 meses y contesta las siguientes preguntas:

1. ¿En qué horarios hay mayor afluencia y en qué estaciones? Da una breve descripción de por qué crees que es así
2. A partir de un análisis temporal:
    a. ¿En qué estaciones puedes observar una tendencia de uso a la alta?
    b. ¿Puedes categorizar las estaciones con base en su tendencia de uso?
    c. Demuestra tus conclusiones gráficamente
3. Por cada estación de Ecobici, identifica cómo están correlacionadas las entradas-salidas entre las otras estaciones (Hint: Puedes usar un heatmap para mostrar la correlación o matrices de origen destino).
4. Usa un método de aprendizaje no supervisado para encontrar “perfiles de uso” de las estaciones. Lo que debes de hacer es categorizar a las estaciones en diferentes grupos a partir de su comportamiento de entradas y salidas. Explica qué método usaste y por qué. De los grupos que encontraste describe las características que puedes inferir de estos a partir de lo descubierto en el inciso anterior.
5. BONUS: En el sitio de Ecobici te puedes registrar para obtener URLs que regresan información sobre cada estación (Número de Slots, Latitud, Longitud). Usa la información de algunas estaciones para explicar el comportamiento de la relaciones que encontraste en la pregunta 3. Explica cómo los atributos geográficos te pueden ayudar a entender las relaciones.(O puedes bajar un Json de aquí)

Sección 2.b: QQP (4 horas) Avanzado :

1. Descarga la Base de datos histórica de Quién es Quién en los Precios de Profeco desde datos.gob.mx en formato csv.
2. Análisis exploratorio
    a. ¿Cuántos registros hay?
    b. ¿Cuántas categorías?
    c. ¿Cuántas cadenas comerciales están siendo monitoreadas?
    d. ¿Cómo podrías determinar la calidad de los datos? ¿Detectaste algún tipo de inconsistencia o error en la fuente?
    e. ¿Cuáles son los productos más monitoreados en cada entidad?
    f. ¿Cuál es la cadena comercial con mayor variedad de productos monitoreados?
3. Análisis
    a. Genera una canasta de productos que te permita comparar los precios geográfica y espacialmente. Justifica tu elección y procedimiento
    b. ¿Cuál es el estado más caro y en qué mes?
    c. ¿Cuál es la ciudad más cara del país?, ¿Cuál es la más barata?
4. Series de tiempo
    a. ¿Cuáles son los principales riesgos de hacer análisis de series de tiempo con estos datos?
    b. ¿Qué pruebas harías para mitigarlos y/o sustentar un modelo de series de tiempo? (menciónalas, no es necesario que las hagas)

    Para la canasta elegida, compara la serie de tiempo mensual de los precios en cada estado del país y responde lo siguiente:
    c. ¿Cuál es el estado que tiene la mayor variación? ¿A qué crees que se deba?
    d. ¿Hay algún patrón estacional de año con año?
    e. ¿Qué otros estados tienen una dinámica similar? Justifica tu respuesta con gráficos o tablas
5. Geoespacial
    a. Para la Zona Metropolitana de León, identifica las principales zonas comerciales (clusters). ¿Cuántas hay? ¿Cómo las definiste?
    b. ¿Cuál de estas zonas comerciales tiene mayor variedad de productos y categorías disponibles?
    c. ¿Cuál es la más barata?
    d. ¿A cuál irías para comprar una bata de laboratorio?
6. Predicción de precios
    a. Implementa un algoritmo predictivo para los niveles de precios de tu canasta de productos para el siguiente periodo disponible y todos los estados del país. Justifica la elección del algoritmo así como los parámetros del modelo.
7. BONUS: Visualización
    a. Genera un mapa interactivo que nos permita identificar la oferta de categorías en la zona metropolitana de León y el nivel de precios en cada una de ellas.