El gobierno de un país mantiene la información de todos sus habitantes, a partir de la cual desea analizar los datos que han almacenado durante el último tiempo. Específicamente, se ha generado una muestra de 29.306 personas (dataset) y se desea investigar posibles patrones de comportamiento (los datos se encuentran en un único archivo llamado "**data.csv**"). Los datos consisten en 15 variables (e.g., dimensiones iniciales), que incluyen edad, estado civil, ocupación, etc. A partir de lo anterior, se debe formar grupos de trabajo de 5 integrantes, que desarrollen métodos y análisis en base a tres entregables durante el semestre:

- Descripción y selección de variables.
- Análisis de clusters (clustering).
- Predicción de variable.

Descripción y selección de variables: El objetivo es entender conceptos básicos de las técnicas de selección de variables para un problema sencillo utilizando métodos y bibliotecas del lenguaje de programación Python.

La actividad consiste en obtener información preliminar relevante acerca de los datos proporcionados. Para esto, se pide realizar un informe que cubra las siguientes tareas:

1. Descripción de datos: **Describir cada una de las variables de los datos proporcionados, incluyendo el tipo de variable, su descripción, distribución (utilice una tabla ó gráfico, en caso de ser necesario), y algún otro aspecto relevante sobre la variable (30% de la nota final del entregable). Se evaluará el uso adecuado de los gráficos.** Ud. deberá interpretar las variables que se proporcionan, salvo las siguientes cuyo significado no es evidente:
    1. fnlwgt: Valor asignado a una persona en base a distintos rasgos. Se espera que dentro de una misma región, personas con un mismo valor tengan un comportamiento similar.
    2. capital.gain: Dinero ganado en el último año por la compra de acciones, propiedades, u otros.
    3. capital.loss: Dinero perdido en el último año por la compra de acciones, propiedades, u otros.
2. Exploración: **Análisis preliminar de datos con técnicas de visualización** y medidas de asociación entre variables. Esto podría incluir **distribución de variables relevantes**, **relaciones entre variables**, **agregación de variables**, **propiedades de sub-poblaciones** y **análisis estadísticos simples**. En este análisis inicial, se debe entregar las primeras observaciones de los datos (10%).
3. Análisis de Calidad: **Describir todo tipo de problema detectado con los datos**, por ejemplo: datos incompletos, datos erróneos, frecuencia de los errores, etcétera (20%).
4. Limpieza de datos: **Reportar las soluciones tomadas para mejorar los problemas de calidad de datos** detectados previamente, por ejemplo, eliminación de datos (20%).
5. Selección de variables: Decidir en base a los aspectos anteriores, **las variables y datos que se utilizarán en este proyecto, junto con su justificación** (20%).

Este entregable consiste de DOS partes:

- Informe: Un reporte respondiendo a cada una de las preguntas requeridas.

Archivo de texto:

- Asignación de cada integrante del grupo a un modelo de clustering (K-means, DBSCAN, Fuzzy C-means, Jerárquico). Caso contrario los modelos serán seleccionados en forma aleatoria.
- Asignación de cada integrante del grupo a un modelo predictivo (Regresión Logística, K-NN, Naive Bayes, árbol de decisión, SVM, ó modificar un algoritmo de clustering para predicción). Caso contrario los modelos serán seleccionados en forma aleatoria.