Desafío - Clasificación desde la econometría

- Para realizar este desafío debes haber estudiado previamente todo el material disponibilizado correspondiente a la unidad.
- Una vez terminado el desafío, comprime la carpeta que contiene el desarrollo de los requerimientos solicitados y sube el `.zip` en el LMS.
- Desarrollo desafío:
    - El desafío se debe desarrollar de manera Individual.
    - Para la realización del desafío necesitarás apoyarte del archivo *Apoyo Desafío - Clasificación desde la econometría*.

Descripción

En esta sesión trabajaremos el dataset south african heart, el cual contiene las siguientes variables:

- `sbp`: Presión Sanguínea Sistólica.
- `tobacco`: Promedio tabaco consumido por día.
- `ldl`: Lipoproteína de baja densidad.
- `adiposity`: Adiposidad.
- `famhist`: Antecedentes familiares de enfermedades cardiácas. (Binaria)
- `types`: Personalidad tipo A
- `obesity`: Obesidad.
- `alcohol`: Consumo actual de alcohol.
- `age`: edad.
- `chd`: Enfermedad coronaria. (dummy)

Desafío 1: Preparar el ambiente de trabajo

- Cargue las librerías básicas para importación y manipulación de datos (numpy, pandas), gráficos (matplotlib y seaborn) y de modelación econométrica (statsmodels).
- Importe el archivo southafricanheart.csv que se encuentra dentro del material de apoyo.
- Realice una descripción del set importado mostrando:
    - lista con los nombres de variables importadas
    - un análisis descriptivo mediante `.describe()`
    - Distribución de categorías para las variables `famhist` y `chd.`

Desafío 2

A continuación se presenta el siguiente modelo a estimar:

$$\log\left(\frac{\Pr(\texttt{chd}=1)}{1-\Pr(\texttt{chd}=1)}\right) = \beta_0 + \beta_1 \cdot \texttt{famhist}$$

Para ello ejecute los siguientes pasos:

1. Recodifique `famhist` a dummy, asignando 1 a la categoría minoritaria.
2. Utilice `smf.logit` para estimar el modelo.
3. Implemente una función `inverse_logit` que realice el mapeo de log-odds a probabilidad.
4. Con el modelo estimado, responda lo siguiente:

    - ¿Cuál es la probabilidad de un individuo con antecedentes familiares de tener una enfermedad coronaria?
    - ¿Cuál es la probabilidad de un individuo sin antecedentes familiares de tener una enfermedad coronaria?
    - ¿Cuál es la diferencia en la probabilidad entre un individuo con antecedentes y otro sin antecedentes?
    - Replique el modelo con `smf.ols` y comente las similitudes entre los coeficientes estimados.
      Tip: Utilice $\beta/4$

Desafío 3: Estimación completa

Implemente un modelo con la siguiente forma:

$$\log\left(\frac{\Pr(\text{chd}=1)}{1-\Pr(\text{chd}=1)}\right) = \beta_0 + \sum_{j=1}^{N} \beta_j \cdot X$$

- Depure el modelo manteniendo las variables con significancia estadística al 5%.
- Compare los estadísticos de bondad de ajuste entre ambos.
- Reporte de forma sucinta el efecto de las variables en el log-odds de tener una enfermedad coronaria.

Desafío 4: Estimación de perfiles

A partir del modelo depurado, genere las estimaciones en log-odds y posteriormente transfórmelas a probabilidades con `inverse_logit`. Los perfiles a estimar son los siguientes:

- La probabilidad de tener una enfermedad coronaria para un individuo con características similares a la muestra.

- La probabilidad de tener una enfermedad coronaria para un individuo con altos niveles de lipoproteína de baja densidad, **manteniendo todas las demás características constantes.**

- La probabilidad de tener una enfermedad coronaria para un individuo con bajos niveles de lipoproteína de baja densidad, **manteniendo todas las demás características constantes.**