



## Comentarios y observaciones

Esta guía se enfoca en el análisis exploratorio estadístico de datos clínicos. Reforzaré el manejo de:

- Estadística descriptiva.
- Visualización.
- Test de hipótesis.
- Intervalos de confianza.
- Comparaciones entre grupos.
- Evaluación de supuestos estadísticos.

1. La guía tiene plazo de entrega el día viernes 15/11 a las 13.00 hrs SCL.
2. La entrega consiste en los jupyter notebooks que desarrollé y un reporte tipo resumen del trabajo planteado.
3. Si usa GPT o cualquier otra herramienta de IA, debe indicarlo dentro de las herramientas empleadas.
4. Toda entrega se envía en un archivo comprimido .tar.gz o .zip al correo: `david.medina@umag.cl`
5. Preguntas hasta el jueves 13.11 a las 12.00 hrs SCL.

## 1 Descripción General

Usted empleará el mismo **dataset sintético** con información clínica del **primer trimestre del embarazo**. El objetivo de esta actividad es caracterizar estadísticamente los datos y explorar posibles asociaciones entre las variables clínicas y el riesgo de desarrollar **diabetes gestacional**.

El dataset contiene:

- $N \approx 1500$  registros
- Variables clínicas (edad, IMC pregestacional, presión arterial, etc.)
- Marcadores bioquímicos (FPG, HbA1c, insulina, lípidos, HOMA-IR)
- Factores de riesgo (antecedentes familiares, GDM previa, PCOS, tabaquismo)
- Estilo de vida (actividad física, dieta)
- Etiqueta: `label_gdm`

---

Incluye además:

- Datos faltantes (MCAR y MAR)
- Presencia de outliers
- Desbalance de clases ( $\sim 17\%$  positivos)

## 2 Actividades a desarrollar

### 2.1 Análisis exploratorio

1. Resumen de dimensiones, tipos de datos y valores faltantes.
2. Estadística descriptiva (media, mediana, sd, IQR, percentiles).
3. Identificación y análisis de outliers.
4. Visualización básica:
  - Histogramas
  - Boxplots
  - Density plots
  - Barras (para variables categóricas)

### 2.2 Intervalos de confianza

1. Calcular intervalos de confianza para variables numéricas relevantes (e.g., IMC, FPG, HbA1c).
2. Interpretar los intervalos desde un punto de vista clínico.

### 2.3 Test de hipótesis

Proponga y ejecute pruebas adecuadas según tipo de variable y pregunta planteada. Ejemplos:

- Comparación de presión arterial entre GDM vs no GDM (t-test, Mann–Whitney).
- Evaluar diferencias en dieta por nivel de actividad física (ANOVA / Kruskal–Wallis).
- Comparación de proporciones (antecedentes familiares vs GDM).

Debe:

- Formular hipótesis ( $H_0$  y  $H_1$ ).
- Verificar supuestos.
- Elegir la prueba adecuada.
- Interpretar resultados.

### 2.4 Evaluación de normalidad

1. Evaluar normalidad en variables de interés (e.g., Shapiro–Wilk, Kolmogorov–Smirnov).
2. Comentar impacto en las pruebas aplicadas.

---

## 2.5 Análisis bivariado

1. Matrices de correlación para variables numéricas.
2. Gráficos bivariados (scatter + regresión, boxplots comparativos).
3. Discusión de asociaciones relevantes.

## 3 Entregables

- Informe en formato libre incluyendo:
  - Descripción general del dataset
  - EDA
  - Intervalos de confianza
  - Test de hipótesis
  - Bondad de ajuste
  - Análisis bivariado
  - Interpretación clínica
  - Conclusiones
- Código empleado (celdas o script)