

*El principal objetivo de esta guía es afianzar los contenidos de clasificación e integrarlos con conocimientos y trabajos prácticos previos: vamos a mezclar todo en la coctelera!. Para eso, vamos a volver a nuestro primer amor: los datos de alturas. En este caso la idea es predecir a partir del dato de altura de un individuo su género implementando distintos plug-in en la regla óptima de Bayes.*

Te recordamos que en esta [página](#) podrás acceder nuevamente a los datos de alturas que generaste en un TP anterior y que vas a utilizar para resolver esta guía. La idea sigue siendo que cada uno trabaje con sus propios datos y por eso les pedimos que ingresen un **número de identificación**, ya sea el número de libreta o los 5 últimos del DNI para obtener datos de forma personalizada.

## 1. La base.

1. Descargar de esta [página](#) un conjunto de  $n = 500$  observaciones, con todas las variables y leer el archivo en R. Trabajaremos con las variables *altura* y *genero* (codificada como F o M). Graficar un plot que pueda dar información sobre la relación entre estas dos variables.
2. Con la **regla de la mayoría** vamos a *aprender* a clasificar el género de un individuo como femenino (1) o masculino (0) cuando su altura  $x = 165$  mediante el método de vecinos. Para ello, considerar los  $k = 10$  vecinos más cercanos y calcular la proporción de 1's. Según este resultado, ¿cómo clasificarías al género de un nuevo individuo con altura igual a 165 cm, F o M? Repetir con  $x = 175$ .
3. Con la **regla de la mayoría** vamos a *aprender* a clasificar el género de un individuo como femenino (1) o masculino (0) cuando su altura  $x = 165$  mediante el método de promedios móviles. Para ello, considerar una ventana de tamaño  $h = 1,5$  alrededor del punto de interés y calcular la proporción de 1's. Según este resultado, ¿cómo clasificarías al género de un nuevo individuo con altura igual a 165 cm, F o M? Repetir con  $x = 175$ .

## 2. El cuerpo: Regla óptima de Bayes - Método Discriminativo.

4. Clasificar el género de un individuo en F o M conociendo su altura mediante la regla de la mayoría utilizando el método de vecinos más cercanos. Para ello, implementar la función `ClasificoVecinos(X, Y, xNuevo, k=10)` que tenga por input un conjunto de valores de X, sus correspondientes valores de Y, un nuevo valor  $x$  para el que se quiere realizar la clasificación y la cantidad  $k = 10$  de vecinos que vamos a utilizar para calcular la regla de la mayoría. (...por ahora te damos el  $k$ ...)

5. Clasificar el género de un individuo en F o M conociendo su altura mediante la regla de la mayoría utilizando el método de promedios móviles. Para ello, implementar la función `ClasificoMovil(X, Y, xNuevo, h=1)` que tenga por input un conjunto de valores de X, sus correspondientes valores de Y, un nuevo valor x para el que se quiere realizar la clasificación y una ventana  $h = 1$  para calcular la regla de la mayoría. (...por ahora te damos la ventana  $h...$ )

### 3. El aditivo aromático: Regla óptima de Bayes - Método Generativo.

Recordemos que cuando la covariable es una variable continua la regla de Bayes óptima puede escribirse como:

$$g^{op}(x) = \begin{cases} 1 & \text{si } f_1(x)\mathbb{P}(Y = 1) \geq f_0(x)\mathbb{P}(Y = 0) \\ 0 & \text{si c. c.} \end{cases}$$

donde  $X|Y = 0 \sim f_0$  y  $X|Y = 1 \sim f_1$  son las densidades condicionales. Como hemos mencionado en clase, en los métodos generativos la regla de Bayes se implementa en la práctica estimando las densidades  $f_0$  y  $f_1$  y la probabilidad  $P(Y = 1)$ .

6. Volvamos a la **Guía 7 Predicciones ítems 6 y 7**, donde se realizó un histograma de alturas para cada género y se le superpuso una curva a cada uno de ellos. ¿Qué curva se le superpuso a cada histograma? ¿Con qué parámetros?

Realizar nuevamente los histogramas de alturas para cada cada sexo y a cada uno de ellos superponerle la curva como en la Guía 7.

**¿Qué relación guardan estas curvas con las densidades  $f_0$  y  $f_1$ ?**

7. ¿Cuál es la proporción de individuos de género femenino en tu conjunto de datos?  
**¿Cómo estimarías  $P(Y = 1)$  a partir de tus datos? ¿Cuánto te da la estimación propuesta?**
8. Haciendo un plug-in en  $g^{op}(x)$  con las estimaciones de  $f_1(x)$ ,  $f_0(x)$ ,  $\mathbb{P}(Y = 1)$  y  $\mathbb{P}(Y = 0)$ , vamos a *aprender* a clasificar el género de un individuo como femenino (1) o masculino (0) cuando su altura  $x = 165$ . Usando el método generativo con tus datos, ¿cómo clasificarías a alguien nuevo con altura igual a 165 cm, F o M? Repetir el ítem anterior para  $x = 175$ .
9. Clasificar el género de un individuo en F o M conociendo su altura mediante el método generativo. Para ello, implementar la función `ClasificoGenerativo(X, Y, xNuevo)` que tenga por input un conjunto de valores de X, sus correspondientes valores de Y, un nuevo valor x para el que se quiere realizar la clasificación.

### 4. A batir!

10. Ahora vamos a testear las reglas. En el archivo [alturas.testeo.csv](#) se encuentran 31 datos de altura que separamos para testear como funcionan los tres métodos implementados.

Para ello, aplicar a este conjunto de datos cada una de las tres reglas implementadas en los ítems anteriores, calcular el Error de Clasificación Empírico de cada clasificador sobre estos datos y completar la información en el [archivo compartido de resultados](#). ¿Cuál de ellas te parece que clasifica mejor?

## 5. Bonus Track

**Leyendo el fondo de la copa...**

11. Graficar `xNuevo` (en el eje de abscisas) tomando valores entre 160 y 170 con un paso de 0.01 y en el de ordenadas el valor con el que clasifica a cada valor la regla `ClasificoVecinos` que implementaste con tus datos (sugerimos representar con línea). Interpretar el criterio con el que clasifica esta regla.
12. Repetir el ítem anterior con `ClasificoMovil` y `ClasificoGenerativo` y superponer con otro color al gráfico anterior. Interpretar y comparar el criterio con el que clasifica cada regla.