REGRESIÓN - CLASIFICACIÓN

Ejercicio 1
Justificar la expresión del intervalo de confianza para la recta de regresión.

Ejercicio 2
Considerar la siguiente tabla de datos sobre el rendimiento de cultivos de papas y el registro de lluvias acumuladas en el período de duración del cultivo:

Datos de entrenamiento

$x =$ Lluvia (mm)	$y =$ Rendimiento (ton/ha)
206	29
188	25
219	31
372	25
345	29
231	30
203	26
170	23
55	12
91	15
292	28
141	24
129	23
170	22
324	30

Datos de validación

$x =$ Lluvia (mm)	$y =$ Rendimiento (ton/ha)
213	30
80	16
391	25
250	26
57	9
303	28
263	28
157	25
72	13
157	23
188	26
216	25
362	28
283	33
308	30

1. Correr una regresión lineal para predecir el rendimiento y en función de la lluvia x.
2. Comparar el MSE en entrenamiento, validación y CV.
3. Determinar el grado óptimo en caso de aplicar una regresión polinomial.
4. Hallar el valor de λ óptimo para la regresión polinomial de grado 5 con regularización.

Ejercicio 3
El objetivo del ejercicio es implementar el algoritmo de descenso de gradiente para estimar los parámetros óptimos de la regresión logística a partir de los siguientes datos:

Paciente	Glucosa	Tiene diabetes (Si/No)
A	90	No
B	160	Si
C	100	No
D	200	Si
E	130	Si

1. Mostrar que la función de perdida de la regresión logística es $L(\beta) = \frac{1}{N}\sum_{i=1}^{N} \ln\left(1+e^{-y_i\beta' x_i}\right)$ donde $\beta = (\beta_0, \beta_1)$ es el parámetro que se busca.
2. Mostrar que el gradiente de L respecto de β está dado por $\frac{\partial L(\beta)}{\partial \beta_j} = \frac{1}{N}\sum_{i=1}^{N} -\frac{y_i x_{ij} e^{-y_i\beta' x_i}}{1+e^{-y_i\beta' x_i}}$
3. Implementando la formula del descenso por gradiente $\beta_{k+1} = \beta_k - \alpha\nabla L(\beta_k)$, calcular el óptimo $\hat{\beta}$ para este ejemplo

Ejercicio 4
Generar 100 observaciones a partir de una distribución normal bivariada $\mathcal{N}(\mu_1, \Sigma_1)$ con $\mu_1 = (3,1)'$ y $\Sigma_1 = I$ (matriz identidad) y etiquetarlas como 1. Generar otras 100 observaciones a partir de una

distribución gaussiana bivariada gaussiana bivariada $\mathcal{N}(\mu_2, \Sigma_2)$ con $\mu_2 = (1,3)'$ y $\Sigma_2 = I$ y etiquetarlas como 0. En conjunto, estas 200 observaciones constituyen el conjunto de entrenamiento.

1. Escribir un código R para generar este conjunto de datos. Dibujar estos datos utilizando diferentes colores para las dos clases.
2. Suponiendo que las probabilidades a priori son iguales, encontrar el clasificador de Bayes. Calcular el error de entrenamiento.
3. Entrenar un modelo de regresión lineal, utilizando la función $\mathsf{lm(y \sim x)}$, con el conjunto de entrenamiento.
4. Trazar la frontera del clasificador de Bayes y la recta obtenida por la regresión lineal.
5. Generar un conjunto de prueba de 50 observaciones y calcular el error de prueba del clasificador de Bayes y del modelo lineal.

Ejercicio 5
Considere la siguiente tabla.

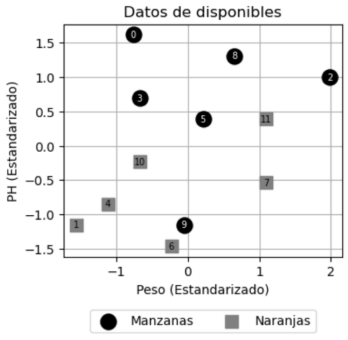
x_1	x_2	x_3	y
0	3	0	Rojo
2	0	0	Rojo
0	1	3	Rojo
0	1	2	Azul
1	0	1	Azul
1	1	1	Rojo

1. Con la distancia euclídea, ¿cuál es la predicción con $k = 1$ y con $k = 3$ para la observación de prueba $(0,0,0)$?
2. Si la frontera de decisión de Bayes en este problema es altamente no lineal, entonces ¿esperaríamos que el mejor valor para k fuera grande o pequeño? ¿Por qué?

Ejercicio 6
Se desea implementar el algoritmo de K vecinos más cercanos para clasificar Manzanas y Naranjas en base a su Peso y su PH.

El gráfico a continuación (derecha) muestra los datos disponibles estandarizados. En el lado izquierdo, se presenta el conjunto de datos segmentado en tres folds. Cada fila simboliza una observación, y su ID correspondiente está alineado con el gráfico situado a la derecha. Las columnas, por otro lado, exhiben los IDs de los 8 puntos que son parte de los dos folds que no contienen la observación de la fila en cuestión, todos ellos ordenados por su proximidad a dicha observación. Por ejemplo, en la primera fila se visualizan las observaciones de los folds 2 y 3. Están dispuestas en un orden que va desde la más cercana a la más lejana respecto a la observación con ID=0.

	ID	1ero	2do	3ro	4to	5to	6to	7mo	8vo
Fold 1	**0**	8	5	10	11	4	7	9	6
	1	4	10	6	9	5	7	11	8
	2	11	8	7	5	10	9	6	4
	3	10	5	8	4	11	9	7	6
Fold 2	**4**	1	10	9	3	0	11	8	2
	5	11	3	8	10	9	0	2	1
	6	9	10	1	3	11	8	0	2
	7	11	9	2	10	8	3	1	0
Fold 3	**8**	5	2	0	3	7	4	6	1
	9	6	4	7	1	5	3	0	2
	10	4	3	5	1	6	7	0	2
	11	5	7	2	3	0	6	4	1



Calcular el error de 3-fold cross-validation para los valores impares de K. ¿Cuál de estos valores de K elegiría?