K-Nearest neighbors (kNN)

Ejemplo 1: Hay los siguientes datos de flores del género Iris, donde cada flor está caracterizada por dos variables: ***Sepal Length*** y ***Sepal Width***. Además, se conoce la especie de cada flor: **Setosa**, **Versicicolor** y **Virginica**.

El objetivo es predecir la especie de una nueva flor desconocida basándose en la similitud con las flores existentes en el conjunto de datos usando el algoritmo de *K-Nearest Neighbors* (*kNN*) para predecir la especie de la nueva flor.

	Sepal Length	Sepal Width	Specie	k=1	k=3
Nueva flor	5.9	2.8		***Virginica***	***Versicicolor***

Sepal Length	Sepal Width	Species	Distancia euclidiana $d(A,B) = \sqrt{(X_B - X_A)^2}$	Ranking
5.3	3.7	Setosa	$\sqrt{(5.9-5.3)^2+(2.8-3.7)^2} = 1.08$	10
5.1	3.8	Setosa	1.2806	12
7.2	3.0	Virginica	1.3152	13
5.4	3.4	Setosa	0.7810	7
5.1	3.3	Setosa	0.9433	9
5.4	3.9	Setosa	1.2083	11
7.4	2.8	Virginica	1.5	15
6.1	2.8	Versicicolor	0.2	3
7.3	2.9	Virginica	1.4035	14
6.0	2.7	Versicicolor	0.1414	2
5.8	2.8	Virginica	0.1	1
6.3	2.3	Versicicolor	0.6403	6
5.1	2.5	Versicicolor	0.8544	8
6.3	2.5	Versicicolor	0.5	4
5.5	2.4	Versicicolor	0.5656	5

a) Realiza el gráfico de dispersión de los datos obtenidos.
b) Calcula la distancia euclidiana entre la nueva flor y cada flor del conjunto de datos y clasifícalas de menor a mayor distancia.
c) ¿Cuál sería la especie de la nueva flor si $k = 1$ o $k = 3$? Razona la respuesta.

Cuando $k = 1$, el algoritmo kNN asigna la especie de la flor más cercana a la nueva flor. En este caso, la distancia más corta corresponde a Virginica. Sin embargo, si $k = 3$, el algoritmo considera los 3 vecinos más cercanos y se asigna la especie mayoritaria, en este caso, Versicicolor.