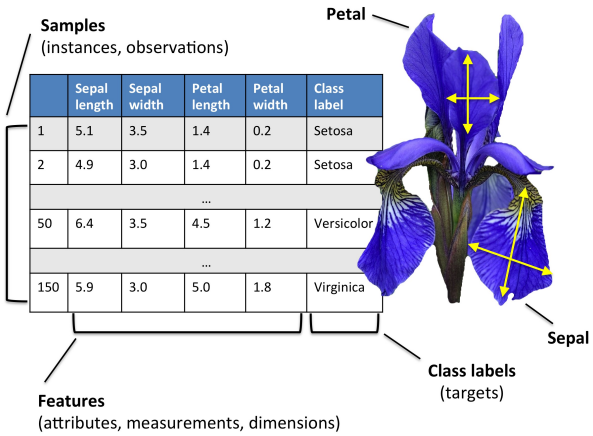
ML_CLUSTERING_KMEANS_01	Segmentación en clusters de la base de datos de flores	ML

Un botánico está interesado en distinguir las especies de algunas flores (de iris) que ha encontrado. Para ello ha registrado algunas medidas asociadas con cada flor: la longitud y anchura de los pétalos (petal), y la longitud y anchura de los sépalos (sepal). Todas las medidas están en centímetros.



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica


Iris Versicolor


Iris Virginica


Iris Setosa

Además, dispone de algunas medidas de flores en las que previamente un experto botánico ha catalogado dentro de las especies ('setosa', 'versicolor' y 'virginica').

Construir un modelo de aprendizaje maquina (machine learning) no supervisado, basado en el algoritmo de clustering KMEANS que permita agrupar los datos según el tipo de especie de la flor en función de las medidas.

SOLUCIÓN

Definir las librerías a utilizar

```
# Importar librerías a utilizar
import numpy as np
import matplotlib.pyplot as plt
# Para visualizar gráficos en 3D
from mpl_toolkits.mplot3d import Axes3D

from sklearn.cluster import KMeans
from sklearn import datasets
```

Cargamos en Python la base de datos IRIS

```
# Cargar la base de datos IRIS
iris = datasets.load_iris()
```

Asignar las variables de atributos (X) y etiquetas (y).

```
# Asignar los atributos (X) y las etiquetas (y)
X = iris.data
y = iris.target
```

Definir el estimador para clustering

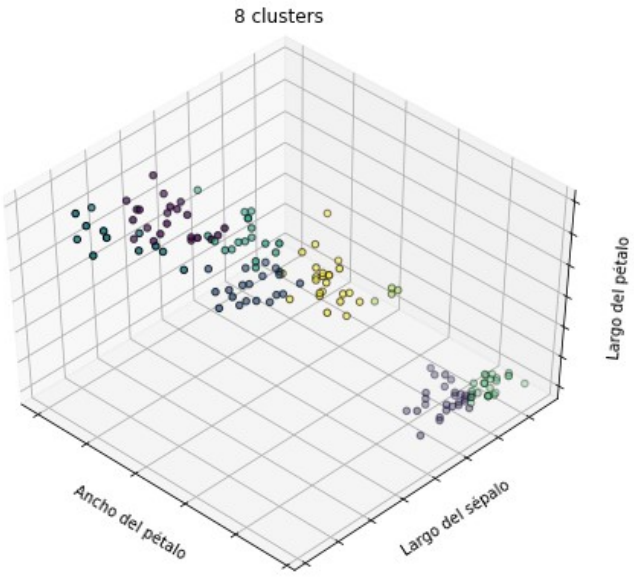
```
# Definir el conjunto de estimadores para clustering
estimators = [('k_means_iris_8', KMeans(n_clusters=8)),
              ('k_means_iris_3', KMeans(n_clusters=3)),
              ('k_means_iris_bad_init', KMeans(n_clusters=3, n_init=1,
                                               init='random'))]
```

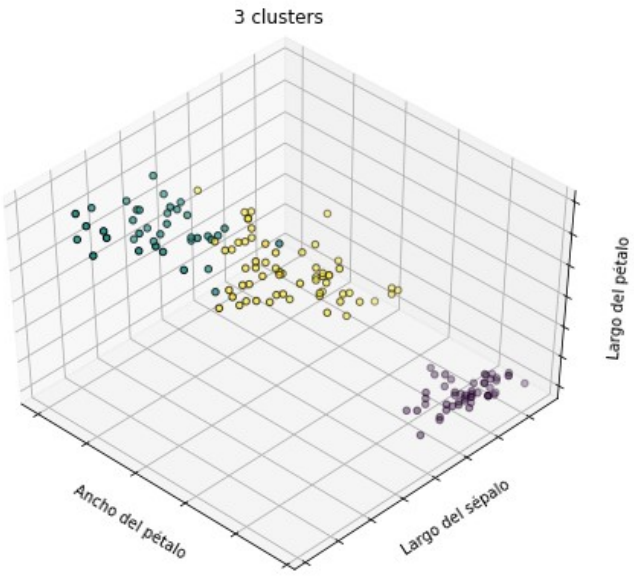
Construir los gráficos para k=8, k=3 y k=3 (con mala inicialización)

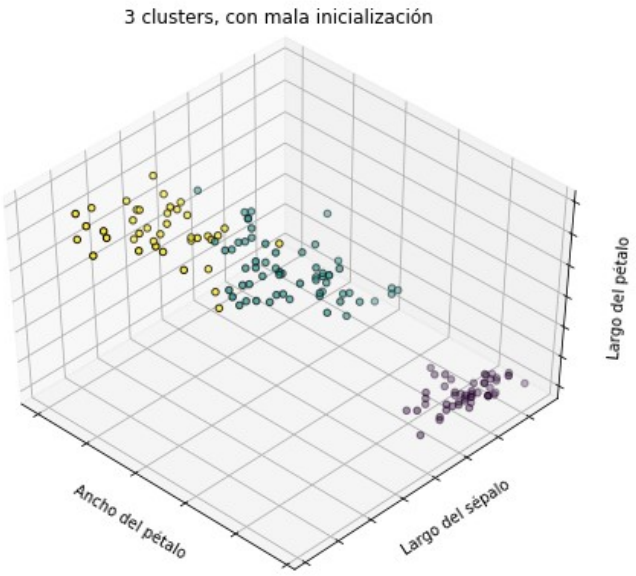
```
# Construir los gráficos para k=8, k=3 y k=3 (con mala inicialización)
fignum = 1
titles = ['8 clusters', '3 clusters', '3 clusters, con mala inicialización']
for name, est in estimators:

    fig = plt.figure(fignum, figsize=(8, 6))
    ax = Axes3D(fig, rect=[0, 0, .95, 1], elev=48, azim=134)

    # Entrenar el modelo KMEANS
    est.fit(X)

    # estimar/predecir las etiquetas sobre el conjunto X
    labels = est.labels_

    # Visualizar los puntos (ancho_pétalo, largo_sépalo, largo_pétalo)
    ax.scatter(X[:, 3], X[:, 0], X[:, 2],c=labels.astype(np.float), edgecolor='k')

    # Definir parámetros del gráfico. Ejes, títulos,...
    ax.w_xaxis.set_ticklabels([])
    ax.w_yaxis.set_ticklabels([])
    ax.w_zaxis.set_ticklabels([])
    ax.set_xlabel('Ancho del pétalo')
    ax.set_ylabel('Largo del sépalo')
    ax.set_zlabel('Largo del pétalo')
    ax.set_title(titles[fignum - 1])
    ax.dist = 12
    fignum = fignum + 1
```

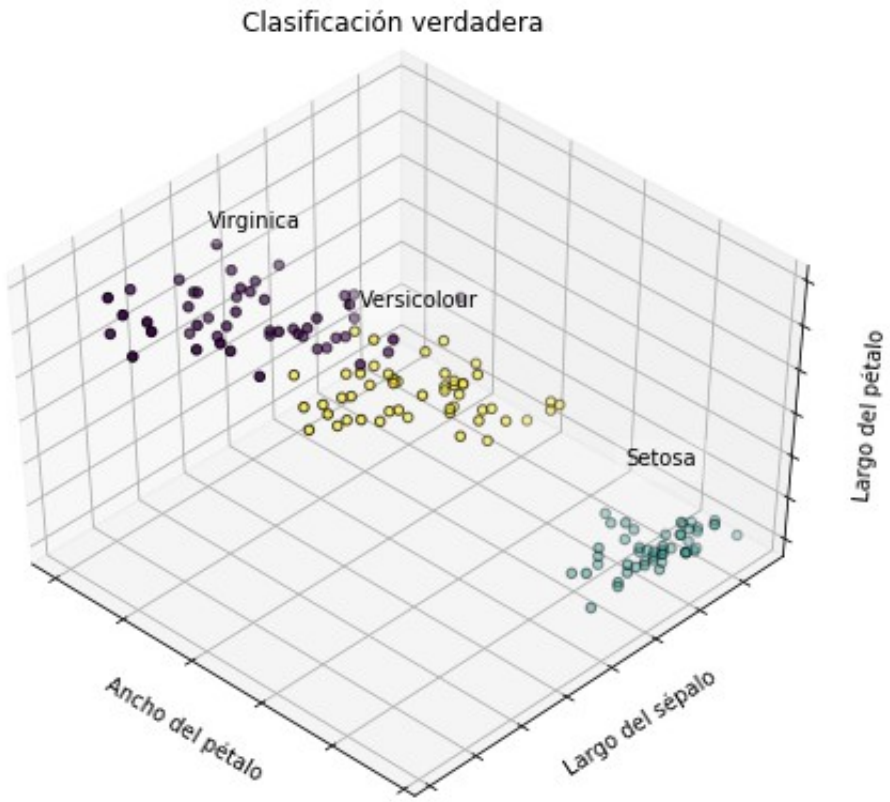
Construir el gráfico correcto (utilizando las etiquetas 'y')

```
# Visualizar el gráfico correcto
fig = plt.figure(fignum, figsize=(8, 6))
ax = Axes3D(fig, rect=[0, 0, .95, 1], elev=48, azim=134)

for name, label in [('Setosa', 0),
                    ('Versicolour', 1),
                    ('Virginica', 2)]:

    # Visualizar el texto de la especie de la planta en el punto medio del cluster
    ax.text3D(X[y == label, 3].mean(),
              X[y == label, 0].mean(),
              X[y == label, 2].mean() + 2, name,
              horizontalalignment='center',
              bbox=dict(alpha=.2, edgecolor='w', facecolor='w'))

# Reorder the labels to have colors matching the cluster results
y = np.choose(y, [1, 2, 0]).astype(np.float)
ax.scatter(X[:, 3], X[:, 0], X[:, 2], c=y, edgecolor='k')

ax.w_xaxis.set_ticklabels([])
ax.w_yaxis.set_ticklabels([])
ax.w_zaxis.set_ticklabels([])
ax.set_xlabel('Ancho del pétalo')
ax.set_ylabel('Largo del sépalo')
ax.set_zlabel('Largo del pétalo')
ax.set_title('Clasificación verdadera')
ax.dist = 12

fig.show()
```


8 clusters
Largo del pétalo
Ancho del pétalo
Largo del sépalo


3 clusters
Largo del pétalo
Ancho del pétalo
Largo del sépalo


3 clusters, con mala inicialización
Largo del pétalo
Ancho del pétalo
Largo del sépalo


Clasificación verdadera
Virginica
Versicolour
Setosa
Largo del pétalo
Ancho del pétalo
Largo del sépalo