1. Import csv:

import pandas as pd

import numpy as np

import scipy

import matplotlib as mpl

import matplotlib.pyplot as plt

dataframe = pd.read\_csv('iris.data.csv', header=None, names=['longitud\_sepalo', 'ancho\_sepalo', 'longitud\_petalo',

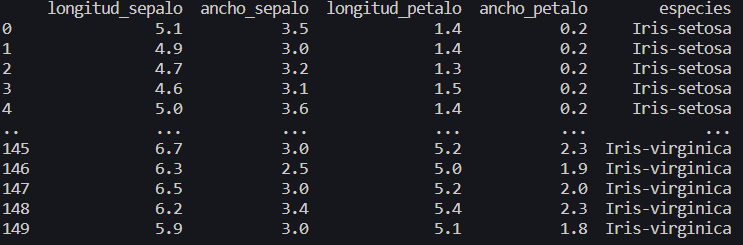
'ancho\_petalo', 'especies'])

x = dataframe

y = dataframe['especies']

target\_names = np.array(['Iris-setosa', 'Iris-versicolor', 'Iris-virginica'])

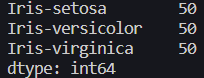
print(dataframe)



1. Explore data

# frequency table of the variable 'species'

print(dataframe.groupby('especies').size())



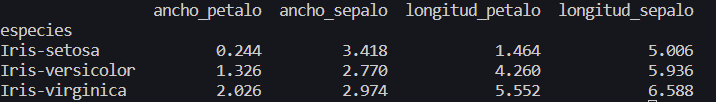
# cross table

cols = ['longitud\_sepalo', 'ancho\_sepalo', 'longitud\_petalo', 'ancho\_petalo', 'especies']

dataframe = dataframe[cols]

cross\_tab = pd.pivot\_table(dataframe, index='especies', aggfunc='mean')

print(cross\_tab)



1. Clustering

from scipy import stats

from scipy.stats import mode

from scipy.cluster.vq import kmeans, vq

x = dataframe.iloc[:, :-1].values

centroids, labels = kmeans(x, 3)

centers = centroids

labels, \_ = vq(x, centroids)

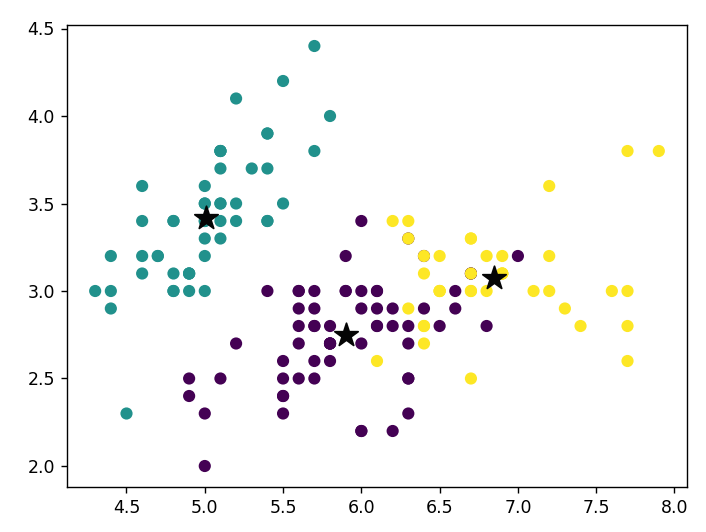


1. Visualize clusters

plt.scatter(x[:,0], x[:,1], c=labels)

plt.scatter(centers[:,0], centers[:,1], marker='\*', s=200, c='#050505')

plt.show()



1. Evaluate clustering results

# Suma de cuadrados intra-cluster (SSW)

def ssw(x, labels, centroids):

ssw = 0

for i in range(len(centroids)):

cluster = x[labels == i]

ssw += np.sum((cluster - centroids[i])\*\*2)

return ssw

print('SSW:', ssw(x, labels, centroids))

**#SSW: 78.94084142614602**

# Suma de cuadrados inter-cluster (SSB)

def ssb(x, labels, centroids):

ssb = 0

mean = np.mean(x, axis=0)

for i in range(len(centroids)):

cluster = x[labels == i]

ssb += len(cluster) \* np.sum((centroids[i] - mean)\*\*2)

return ssb

print('SSB',ssb(x, labels, centroids))

**#SSB 601.8835585738539**