Visualizar los datos

La función de la máquina es determinar por sí misma si los datos que le damos como parámetros proceden de una cereza o de un albaricoque. De todos modos, para nosotros es interesante controlar su predicción.

Para conseguirlo debemos intentar determinar por nosotros mismos los dos clústeres antes pedirle a la máquina que los encuentre sola, pero ¿cómo conseguirlo fácilmente? Con ayuda de un gráfico. Creamos un archivo de script nuevo al que llamaremos clustering e introducimos las siguientes líneas de programa:

import pandas as pnd

import matplotlib.pyplot as plt

#Carga de los datos

frutas = pnd.read\_csv("datas/frutas.csv",

names=['DIAMETRO','PESO'], header=None)

#Visualización gráfica de los datos

frutas.plot.scatter(x="DIAMETRO",y="PESO")

plt.show()

La finalidad de este programa es crear una nube de puntos del conjunto de las parejas Diámetro/Peso.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

*Visualización de los datos*

A partir de este gráfico podemos determinar por nosotros mismos los dos clústeres en potencia correspondientes a las cerezas y a los albaricoques:

Gráfico

Descripción generada automáticamente con confianza baja

*Determinación manual de los clústeres*

En efecto, según nuestro conjunto de observación, el diámetro de una cereza está entre 1 y 29, lo que tiende a crear un primer clúster, como se indica en la figura anterior.