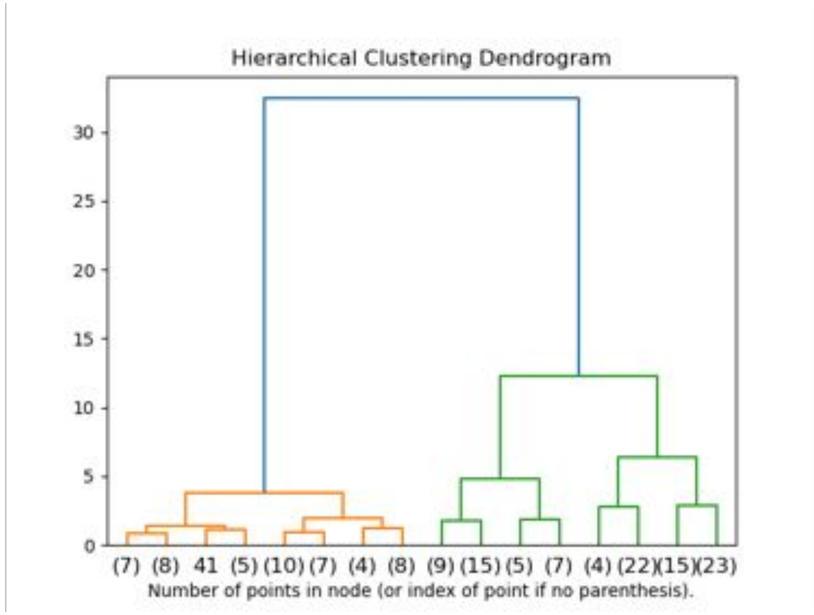
Agglomerative Clustering

Por que?

Algoritmo que usa dos próprios dados para decidir quantos cluester serão utilizados no algoritmo K-means. Ele gera o dendrograma para poder ser analisado.



Matematicamente como acontece?

Ele utiliza da distância entre dois pontos para agrupar os itens correlatos e depois desenvolver o dendrograma

$$D^2 = (x_b - x_a)^2 + (y_b - y_a)^2$$

$$\sqrt{D^2} = \sqrt{(x_b - x_a)^2 + (y_b - y_a)^2}$$

$$D = \sqrt{(x_b - x_a)^2 + (y_b - y_a)^2}$$

Vantagens e desvantagens

As vantagens deste algoritmo é que é ele consegue saber qual a quantidade de cluster deve ser usado para o problema e pois não precisa de supervisão

E uma das desvantagens é que não existe a revisão da quantidade de cluster que serão usados cada vez depois que o modelo está sendo aplicado.

Na prática

Na prática, a biblioteca Scikit Learn juntamente com a SciPy fazem o trabalho de criar o algoritmo K-means e criar o dendrograma, respectivamente.