Apa itu Clustering dan Algoritma K-Prototypes

Clustering adalah proses pembagian objek-objek ke dalam beberapa kelompok (*cluster*) berdasarkan tingkat kemiripan antara satu objek dengan yang lain.
Terdapat beberapa algoritma untuk melakukan *clustering* ini. Salah satu yang populer adalah k-means.
K-means itu sendiri biasa nya hanya digunakan untuk data-data yang **bersifat numerik**.
Sedangkan untuk yang **bersifat kategorikal** saja, kita bisa menggunakan **k-modes**.
Lalu bagaimana apabila di data kita terdapat gabungan kategorikal dan numerikal variabel?
Jawabannya kita bisa menggunakan algoritma k-prototype yang merupakan gabungan dari k-means dan k-modes. Hal ini bisa dilakukan dengan menggunakan library k-modes yang didalamnya terdapat modul kprototype.
Untuk menggunakan algoritma kprototype kamu perlu memasukkan jumlah *cluster* yang dikehendaki dan juga memberikan *index* kolom untuk kolom-kolom yang bersifat kategorikal.
Untuk lebih lengkapnya kamu bisa melihat dokumentasi dari kprototype melalui *link* berikut:
https://github.com/nicodv/kmodes

Mencari Jumlah Cluster yang Optimal

Salah satu parameter penting yang harus dimasukkan pada algoritma kprototype adalah jumlah *cluster* yang diinginkan. Oleh karena itu, kamu perlu mencari jumlah *cluster* yang optimal.
Salah satu cara untuk mendapatkan nilai optimal tersebut adalah dengan menggunakan bantuan 'elbow plot'.
Elbow plot ini dapat dibuat dengan cara memvisualisasikan total jarak seluruh data kita ke pusat *cluster*-nya. Selanjutnya kita memilih titik siku dari pola yang terbentuk dan menjadikannya sebagai jumlah *cluster* kita.

Untuk melakukan hal ini kamu perlu menjalankan algoritma kprototypes dengan berbagai jumlah *cluster*. Selanjutnya kamu juga menyimpan nilai cost_ dan memvisualisasikannya dengan *line plot* atau *point plot.*