KMeans

Procedure 方法

Kmeans adalah sebuah algoritma machine learning unsupervised yang berfungsi untuk membuat cluster dari data yang diberikan.

1. **Inisialisasi model dengan jumlah cluster yang diinginkan**
   Jumlah cluster yang optimal dapat ditentukan dengan elbow method atau silhouette method. Cluster yang sedikit berarti suat cluster lebih kompleks variansi didalamnya, sebaliknya cluster yang banyak dapat menjadi overfit.
2. **Inisialisasi Centroid**
   Centroid adalah titik pusat dari kluster. Centroid dapat dinisialisasi secara random, atau dengan kmeans++ random sesuai distribusi jarak antar titik agar centroid lebih tersebar dengan baik.
3. **Mengelompokan Data**
   Tiap titik data dihitung jaraknya ke semua centroid, dan akan memasuki cluster yang memiliki centroid terdekat.
4. **Update Centroid**
   Centroid di-update dengan mengambil nilai rata-rata dari seluruh data didalamnya.
5. **Ulangi 3-4 hingga cluster stabil**
   Hasil clustering bisa dievaluasi dengan silhouette score.

VS Sklearn

Hasil di notebook DoE menghasilkan clustering yang cukup mirip. Perbedaan yang dihasilkan mungkin diakibatkan proses pemilihan centroids yang acak.

Potential Improvements

- Menaikan performance pada perhitungan jarak dengan beberapa teknik optimasi perhitungan