DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

1. Cara kerja Algoritma

Secara garis besar, algoritma DBSCAN terdiri dari tiga tahap utama: parameter initialization, cluster expansion, dan noise identification.

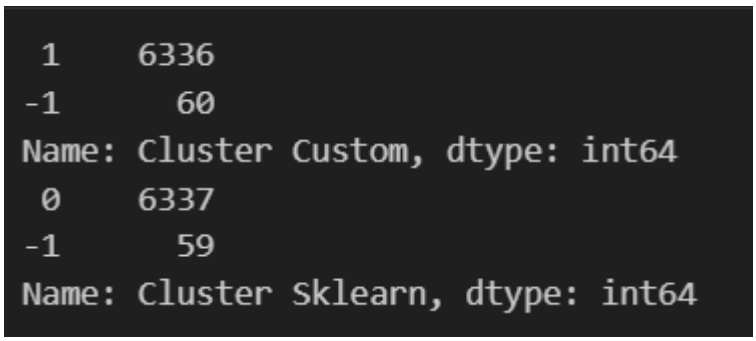
Pada tahap pertama, dua parameter utama, yaitu eps (epsilon) dan minPts, akan ditentukan. eps adalah jarak maksimum yang digunakan untuk mendefinisikan tetangga dari suatu titik, sedangkan minPts adalah jumlah minimum titik yang harus ada dalam radius eps agar suatu area dapat dianggap sebagai cluster.

Pada tahap kedua, setiap titik data akan diperiksa untuk melihat apakah titik tersebut merupakan titik inti (core point), yaitu titik yang memiliki setidaknya minPts tetangga dalam radius eps. Jika suatu titik adalah titik inti, maka algoritma akan memperluas cluster dengan menambahkan semua tetangga langsung ke dalam cluster tersebut. Proses ini akan terus berlanjut untuk setiap tetangga yang baru ditambahkan, hingga tidak ada lagi titik yang dapat ditambahkan ke dalam cluster.

Pada tahap ketiga, titik-titik yang tidak dapat dimasukkan ke dalam cluster mana pun akan diberi label sebagai noise. Titik-titik ini tidak termasuk dalam cluster mana pun karena tidak memenuhi syarat sebagai titik inti atau tidak berada dalam jangkauan eps dari titik inti mana pun.

Algoritma DBSCAN akan berhenti setelah semua titik dalam dataset telah diklasifikasikan baik sebagai anggota suatu cluster atau sebagai noise.

2. Perbandingan Hasil Evaluasi Model



```
outlier_custom = X_EDA[X_EDA['Cluster Custom'] == -1]
outlier_sklearn = X_EDA[X_EDA['Cluster Sklearn'] == -1]

concatenated = pd.concat([outlier_sklearn, outlier_custom])

difference = concatenated.drop_duplicates(keep=False)
difference
```

✓ 0.0s

	Ship Mode	Segment	City	State	Region	Category	Sub-Category	Product Name	Sales	Quantity
1939	Standard Class	Consumer	Broomfield	Colorado	West	Office Supplies	Binders	Wilson Jones Active Use Binders	8.736	4

Hanya ada satu outlier yang ada di model saya dan tidak ada di model scikit learn. Perbedaan kecil ini disebabkan oleh perbedaan pemilihan titik awal.