Tugas Bagian 2

SK-5222 Penambangan Data dalam Sains

Deadline: 21 Mei 2022 pukul 23.59

Jawablah pertanyaan dibawah ini dalam format laporan. File program (code) untuk soal nomor 1 harus dikumpulkan juga.

1. Diberikan 10 buah titik data sebagai berikut:

Titik	Koordinat x	Koordinat y
p1	4	5.2
p2	2.1	3.9
p3	3.4	3.1
p4	2.7	2
p5	0.8	4.1
p6	4.6	2.9
p7	4.3	1.2
p8	2.2	1
p9	4.1	4.1
p10	1.5	3

- Lakukan klasterisasi dari data tersebut dengan menggunakan algoritma k-means dengan jumlah partisi $\boldsymbol{K = 2}$ sebanyak **10 kali**.
- Tentukan sentroid awal (secara random) yang **berbeda** setiap melakukan klasterisasi.
- *Stopping criteria* untuk klasterisasi bisa ditentukan sendiri (tidak harus sampai tidak ada perubahan sentroid)

Pertanyaan:

a. Tuliskan hasil akhir kluster yang didapat untuk setiap klasterisasi.
b. Hitung nilai ***average* SSE** untuk masing-masing hasil klusterisasi
c. Hitung nilai ***average Sillhouette Coefficient*** untuk masing-masing hasil klusterisasi
d. Dari hasil SSE dan Sillhouette Coefficient, menurut Anda, hasil klasterisasi mana yang memberikan hasil terbaik? Berikan alasannya.
e. Apakah algoritma K-means sudah memberikan hasil yang baik? Apa yang dapat dilakukan agar hasil klasterisasi lebih baik?

2. Diberikan ***confusion matrix*** sebagai berikut

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Total
#1	1	1	0	11	4	676	693
#2	27	89	333	827	253	33	1562
#3	326	465	8	105	16	29	949
Total	354	555	341	943	273	738	3204

Hitung nilai ***entropy*** dan ***purity*** untuk matriks tersebut. **Berikan analisis untuk hasil yang didapat.**