Laporan Praktikum Clustering

Anggota :
1. Tuti Purwaningsih (5054241002)
2. Muhammad Aditya Nugraha (5054241004)
3. Mohammad Akmal Fayyazi (5054241045)

Mata Kuliah : Penambangan Data

Segmentasi pelanggan merupakan salah satu metode analisis penting dalam data mining dan customer relationship management (CRM). Tujuannya adalah mengelompokkan pelanggan ke dalam segmen-segmen yang homogen berdasarkan perilaku, preferensi, dan karakteristik demografis. Dengan segmentasi yang tepat, perusahaan dapat merancang strategi pemasaran, promosi, dan retensi pelanggan secara lebih efektif.

Pada praktikum ini, digunakan dataset pelanggan yang mencakup informasi demografis berupa usia, status perkawinan, pendidikan, riwayat pembelian produk, serta respons terhadap promosi, dll. Data tersebut nantinya akan diolah dan dianalisis menggunakan empat algoritma unsupervised clustering yang berbeda-beda, yaitu : K-Means; Gaussian Mixture Model (GMM); DBSCAN; dan Spectral. Tujuan dari praktikum ini ialah untuk membandingkan keempat algoritma tersebut berdasarkan metrik kualitas clustering, visualisasi segmen, serta interpretabilitas profil pelanggan.

Pembahasan praktikum

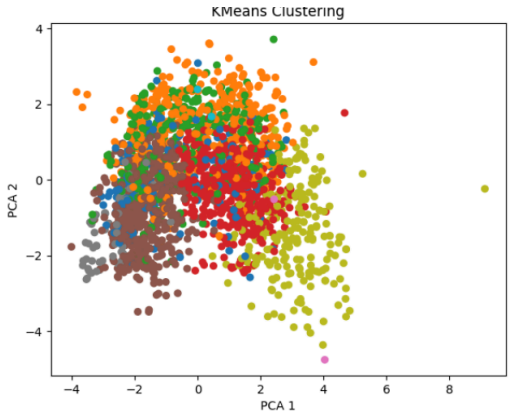
1. EDA
   Pada tahap eksplorasi data ini dilakukan untuk memahami data, seperti: bentuk data, statistik data, mengecek missing value, cek outlier

2. Prepocessing
   Prepocessing yang dilakukan, yakni:
   - menghandle missing value dengan imputasi
   - Mengkonversi ke tanggal untuk kolom (Dt_Customer)
   - Membuat beberapa fitur baru, seperti : “Age” (didapat dari mengurangi 2025 dengan tahun lahir) ; “Tenure” (didapat dari mengurangi tahun sekarang dengan tahun saat awal menjadi pelanggan pada kolom Dt_Customer); “TotalSpend” (menjumlahkan pengeluaran untuk semua jenis produk); “Frekuensi” (menjumlah semua jenis pembelian) ;”Promo_Accepted” (menjumlah semua penawaran promo yang dilakukan)
   - Menghapus semua kolom yang tidak diperlukan (sudah di ekstrak informasinya ke fitur baru)
   - Mengubah data kategorikan jadi numerik pada kolom ‘Education’ dan ‘MaritalStatus’
   - Melakukan standarisasi dengan StandarScaler

3. Implementasi Clustering
   - Kmeans

Pada Kmeans dilakukan pengujian jumlah cluster terlebih dahulu dengan menggunakan silhouette score 2-10k, hasilnya k=10 yang menghasilkan nilai tertinggi. Setelah itu, dilakukan uji hyperparameter dan menghasilkan parameter terbaik untuk digunakan adalah ‘n_cluster’ : 9, ‘init’ : ‘k-means++’, dan ‘n_init’ : 23. Setelah di clustering dengan menggunakan parameter terbaik, hasil silhouette scorenya adalah 0.13059392629505637.

Setelah diclustering, terdapat 9 cluster dengan 112 noise atau outlier.



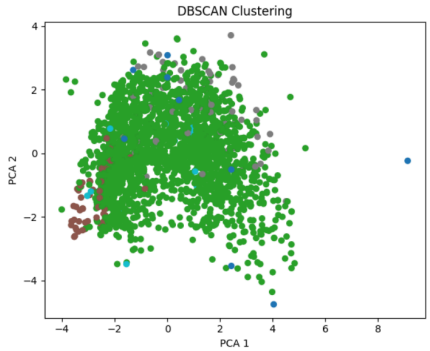
Cluster profiling (mean of key features, non-noise):

Cluster	Recency	Frekuensi	Promo_Accepted	TotalSpend
0	-0.012036	-0.204362	-0.214668	-0.264510
1	-0.010313	0.018972	-0.197699	-0.108637
2	-0.023217	-0.087321	-0.239665	-0.231484
3	0.105940	0.701901	-0.096366	0.553255
4	0.026528	-0.789640	-0.332452	-0.800817
5	0.134363	0.604257	1.035390	0.974402
6	-0.041570	-1.021852	-0.272121	-0.870902
7	-0.204441	0.784865	1.957781	1.633599
8	-1.592395	0.539114	-0.439037	-0.301933

Hasil cluster profiling atau profil segmen sudah di standarisasi.

- DBSCAN

  Pada DBSCAN dicari eps terbaiknya dengan menggunakan K-distance graph. Didapatkan eps terbaiknya adalah 5.77356782670634. Setelah diclustering, terdapat 4 cluster dengan 12 noise atau outlier. Selain itu nilai silhouette scorenya adalah 0.2527060839458219.



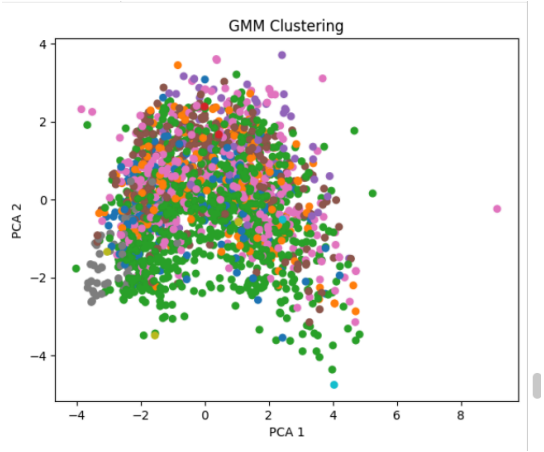
Cluster profiling (mean of key features, non-noise):

Cluster	Recency	Frekuensi	Promo_Accepted	TotalSpend
0	0.002847	0.019413	0.005418	0.018949
1	-0.041570	-1.021852	-0.272121	-0.870902
2	-0.011502	0.256257	0.104173	0.234838
3	0.047117	-0.263170	-0.361436	-0.440130

Hasil cluster profiling atau profil segmen sudah di standarisasi.

- GMM

  Pada GMM dicoba n_component dan cov_type nya. Mencoba n_component dari 2-10 dan cov_typenya ‘full’, ‘tied’, ‘diag’, ‘sheparical’. Hasil terbaiknya adalah dengan n_component : 10 dan cov_type : ‘full’. Setelah diclustering, terdapat 10 cluster dengan 112 noise atau outlier. Untuk nilai silhouette scorenya yakni 0.1397369466483639.



Cluster profiling (mean of key features, non-noise):

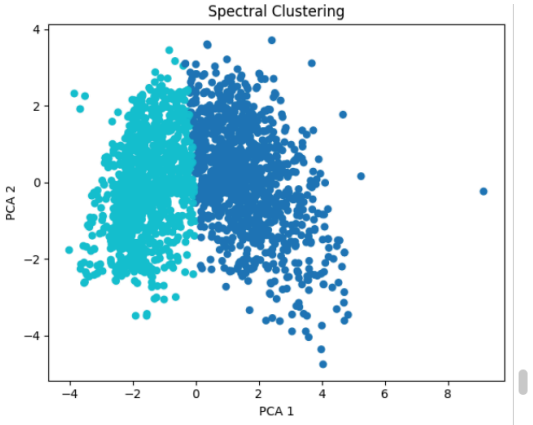
Cluster	Recency	Frekuensi	Promo_Accepted	TotalSpend
0	0.059054	-0.245961	-0.239276	-0.325238
1	0.060888	0.011325	-0.205837	-0.102135
2	0.009625	0.027376	0.019022	0.035650
3	-1.592395	0.539114	-0.439037	-0.301933
4	-0.011502	0.256257	0.104173	0.234838
5	-0.024909	-0.031226	-0.140027	-0.109255
6	0.026025	0.053354	0.028738	0.090419
7	-0.007301	-1.074829	-0.288585	-0.893385
8	0.355964	-0.296884	-0.193299	-0.478256
9	0.134363	0.604257	1.035390	0.974402

Hasil cluster profiling atau profil segmen sudah di standarisasi.

- Spectral

  Pada spectral juga mencoba parameter terbaik untuk dataset ini. Mencoba affinities dengan ‘nearest_neighbors’ dan ‘rbf’, selain itu dicoba juga beberapa cluster. Hasilnya affinity terbaik adalah ‘rbf’ dan n_clusternya

adalah 2 dengan silhouette score 0.44449267183200913. Setelah diclustering, terdapat 2 cluster dengan 12 noise atau outlier.



Cluster profiling (mean of key features, non-noise):

Cluster	Recency	Frekuensi	Promo_Accepted	TotalSpend
0	0.007808	0.779376	0.325730	0.810972
1	-0.004266	-0.755331	-0.317798	-0.784169

Hasil cluster profiling atau profil segmen sudah di standarisasi.

4. Kesimpulan

   Berdasarkan nilai silhouette score sebagai nilai kualitas metric dan juga interpretabilitas, algoritma clustering terbaik adalah spectral. Silhouette scorenya memiliki nilai tertinggi yaitu 0.44449267183200913 dibanding algoritma 3 lainnya yang tidak sampai 0.3. Berdasarkan hasil visualisasi, spectral juga mudah dipahami polanya karena perbedaan warnanya terlihat dengan jelas.