**Rangkuman Unsupervised Learning - Clustering**

**Konsep Dasar dan Proses Clustering**

**Clustering** adalah teknik unsupervised learning untuk mengelompokkan data ke dalam grup berdasarkan kesamaan fitur tanpa memerlukan label. Proses ini bermanfaat untuk menemukan pola atau struktur tersembunyi dalam data.

**Tahapan dalam Proses Clustering**

1. **Pengumpulan Data**: data relevan dikumpulkan untuk dianalisis, misalnya data pelanggan, seperti usia, pendapatan, dan riwayat pembelian.
2. **Pemilihan Fitur**: fitur penting dipilih berdasarkan relevansi dengan tujuan analisis, misalnya usia dan pendapatan.
3. **Pra-pemrosesan Data**: data dibersihkan, dihapus duplikasinya, dan dinormalisasi agar setiap fitur memiliki skala yang sama.
4. **Pemilihan Metode Pengukuran Jarak**: digunakan metode seperti jarak Euclidean untuk mengukur kesamaan antara objek data.
5. **Pemilihan Algoritma Clustering**: algoritma seperti K-Means, DBSCAN, atau hierarchical clustering dipilih berdasarkan kebutuhan analisis.
6. **Penerapan Algoritma Clustering**: algoritma diterapkan untuk membagi data ke dalam cluster berdasarkan kemiripan fitur.
7. **Evaluasi Hasil Clustering**: metrik seperti silhouette score digunakan untuk menilai kualitas cluster.
8. **Interpretasi dan Tindakan**: hasil clustering dianalisis untuk menemukan pola dan diambil tindakan, seperti perancangan strategi pemasaran berdasarkan segmentasi pelanggan.

**Hierarchical Clustering (HC)**

HC adalah teknik clustering yang membentuk struktur hierarki antara cluster. Ada dua pendekatan dalam HC:

* **Agglomerative** (bottom-up): setiap objek dimulai sebagai cluster terpisah dan digabungkan hingga menjadi satu cluster besar.
* **Divisive** (top-down): dimulai dengan satu cluster besar yang dibagi menjadi cluster lebih kecil.

Hierarchical clustering digambarkan dalam dendrogram untuk menunjukkan proses penggabungan atau pemecahan cluster.

**Non-hierarchical Clustering (NHC)**

**Non-hierarchical Clustering (NHC)** adalah metode clustering yang membagi data menjadi cluster tanpa membentuk struktur hierarkis. Ini berbeda dari hierarchical clustering, yang membangun struktur pohon (dendrogram) untuk menggambarkan hubungan antar cluster. NHC berfokus pada pembentukan cluster terpisah berdasarkan kriteria tertentu, seperti jarak atau kepadatan, tanpa memperhatikan hubungan hierarkis.

**Metode Non-hierarchical Clustering**

Berikut adalah beberapa metode NHC yang sering digunakan:

**K-Means Clustering**

**K-Means Clustering** adalah metode yang membagi data menjadi kkk cluster berdasarkan jarak terdekat dari centroid (titik pusat cluster). Ini adalah salah satu metode yang paling umum digunakan karena kesederhanaannya dan efisiensinya.

**Cara Kerja**

1. **Inisialisasi:** Pilih jumlah cluster kkk dan inisialisasi centroid secara acak atau menggunakan metode seperti K-Means++.
2. **Penugasan Cluster:** Alokasikan setiap titik data ke centroid terdekat berdasarkan jarak Euclidean.
3. **Pembaruan Centroid:** Hitung ulang centroid sebagai rata-rata titik data dalam cluster.
4. **Iterasi:** Ulangi langkah penugasan dan pembaruan hingga centroid stabil atau konvergen.

**Kelebihan**

* **Simplicity:** mudah diimplementasikan dan dipahami.
* **Efisiensi:** cepat dalam komputasi, terutama untuk dataset besar.

**Kekurangan**

* **Pemilihan Jumlah Cluster:** membutuhkan penentuan jumlah cluster kkk sebelumnya.
* **Sensitivitas terhadap Outliers:** dapat terpengaruh oleh outliers yang memengaruhi posisi centroid.

**K-Medoids Clustering**

**K-Medoids Clustering** adalah varian dari K-Means yang menggunakan titik data nyata sebagai pusat cluster (medoids), bukan rata-rata. Ini membuatnya lebih robust terhadap noise dan outliers.

**Cara Kerja**

1. **Inisialisasi:** Tentukan jumlah cluster kkk dan pilih medoids secara acak.
2. **Penugasan Cluster:** Alokasikan setiap titik data ke medoid terdekat.
3. **Pemilihan Medoid Baru:** Pilih medoid baru untuk meminimalkan total jarak dalam cluster.
4. **Iterasi:** Ulangi langkah penugasan dan pemilihan medoid hingga tidak ada perubahan signifikan.

**Kelebihan**

* **Robust terhadap Outliers:** lebih tahan terhadap outliers.
* **Kualitas Cluster:** medoids memberikan pusat yang lebih representatif.

**Kekurangan**

* **Kompleksitas:** proses pemilihan medoid lebih rumit dan lambat dibandingkan K-Means.
* **Pemilihan Jumlah Cluster:** masih memerlukan penentuan jumlah cluster kkk sebelumnya.

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

**DBSCAN** adalah metode clustering berbasis kepadatan yang mengidentifikasi cluster berdasarkan kepadatan titik. Ini tidak memerlukan jumlah cluster yang ditentukan sebelumnya dan dapat menemukan cluster dengan bentuk arbitrer.

**Cara Kerja**

1. **Penentuan Parameter:** Tentukan parameter jarak maksimum (ε) dan jumlah minimum titik (MinPts).
2. **Pengelompokan:** Identifikasi titik-titik yang cukup dekat satu sama lain untuk membentuk cluster. Titik yang tidak memenuhi kriteria dianggap sebagai noise.
3. **Ekspansi Cluster:** Tambahkan titik tetangga yang memenuhi kriteria kepadatan.

**Kelebihan**

* **Tidak Memerlukan Jumlah Cluster:** tidak perlu menentukan jumlah cluster sebelumnya.
* **Menangani Noise:** dapat mengidentifikasi noise dan cluster dengan bentuk arbitrer.

**Kekurangan**

* **Pemilihan Parameter:** memerlukan pemilihan parameter yang tepat (ε dan MinPts).
* **Skalabilitas:** kurang efisien pada dataset yang sangat besar.

**Mean Shift Clustering**

**Mean Shift Clustering** adalah metode berbasis kepadatan yang mencari titik-titik yang berkumpul di sekitar jendela pencarian (kernel) tanpa memerlukan jumlah cluster awal.

**Cara Kerja**

1. **Inisialisasi:** Tentukan ukuran jendela pencarian atau kernel.
2. **Pindah Titik:** Hitung rata-rata posisi titik di sekitar jendela dan geser titik ke posisi rata-rata tersebut.
3. **Konsolidasi:** Titik-titik yang berpindah ke posisi yang sama dikelompokkan bersama untuk membentuk cluster.

**Kelebihan**

* **Tidak Memerlukan Jumlah Cluster:** tidak perlu menentukan jumlah cluster sebelumnya.
* **Menangani Bentuk Arbitrer:** efektif dalam menemukan cluster dengan bentuk arbitrer.

**Kekurangan**

* **Pemilihan Ukuran Jendela:** ukuran jendela pencarian dapat memengaruhi hasil clustering.
* **Kompleksitas:** memerlukan waktu komputasi lebih lama pada dataset besar.