



Konsep Clustering Analysis

Asyraf Ilmansyah Hia, B.Eng.

Data Team Lead at Telkom Indonesia



Asyraf Ilmansyah Hia, B.Eng.
Data Team Lead

Email: asyrafilmansyah@gmail.com

WA: +62 821 1038 4458

LinkedIn: Asyraf Ilmansyah Hia

<https://www.linkedin.com/in/asyraf-ilmansyah-hia-6a4636145/>



Data Team Lead
Feb 2022 - Present

- DB IHX - IndiHome & Addons
- DB SMB - PaDi UMKM



Data Scientist
Jan 2021 – Jan 2022

- EDM - Fraud Detection
- EDM - Cash Ratio Optimization
- DDB - App Ceria



Cikarang Techno
Park Group

StartUp in Data, AI &
Software Dev
Aug 2017 – Des 2020


- App - Android, iOS, Web Development
- AI - IoT & ChatBot
- Data - Business Intelligence Dashboard & Machine Learning



Bachelor of
Engineering
Jun 2012 – Jan 2017

- Electrical and Electronics Engineering

Table of Contents



Conceptual of Clustering Analysis
Praktik Code Clustering
RFM Analysis
Praktik Code RFM
Q & A

Clustering Analysis

- Clustering adalah teknik unsupervised learning (tidak memerlukan data pelatihan (training data)) untuk mengelompokkan data berdasarkan jarak, kemiripan atau maximum likelihood.
- Contoh : grouping customer berdasarkan profil pembelian, frekuensi, dll.

Memahami data secara mendalam

- Bisa lebih mudah melihat pola, tren, dan hubungan antar data secara mendalam.

Membuat segmentasi

- Mengidentifikasi kelompok pelanggan yang memiliki perilaku atau preferensi mirip untuk membantu perencanaan strategi pemasaran yang lebih efektif.

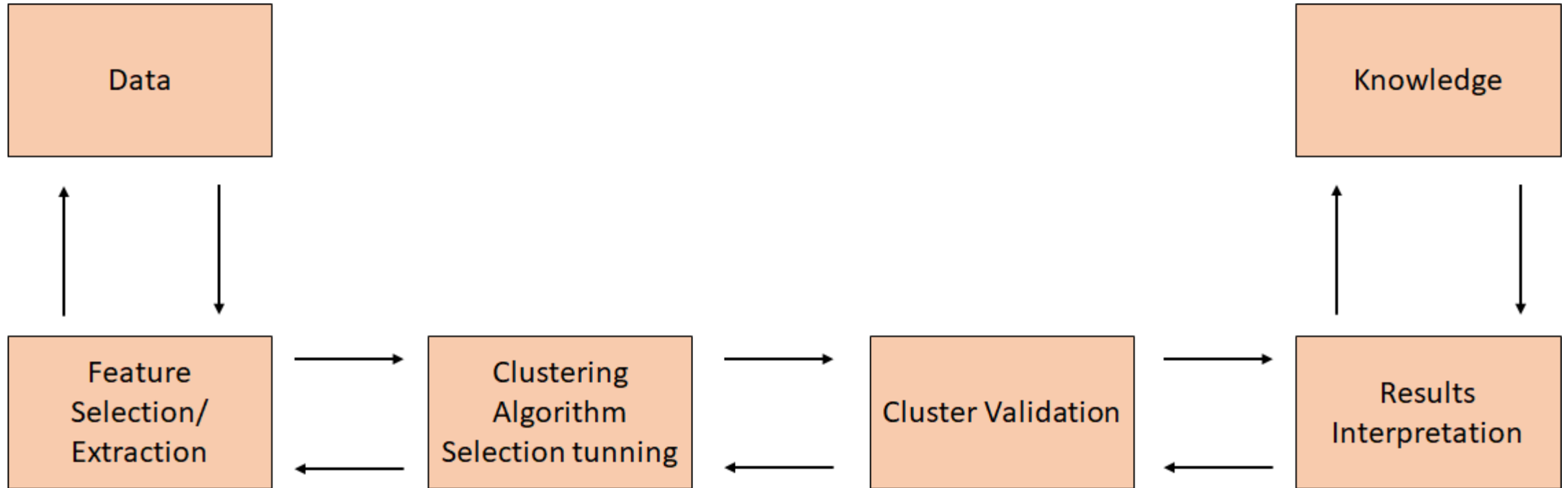
Mendeteksi anomali

- Mengidentifikasi titik data yang tidak sesuai dengan pola data pada umumnya. Menemukan data yang aneh atau menyimpang, seperti mendeteksi penipuan dan pemantauan jaringan.

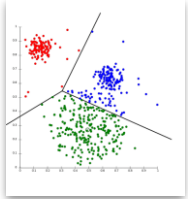
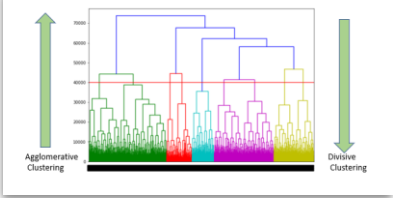
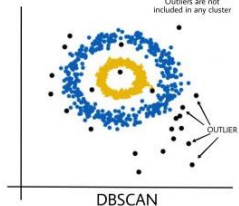
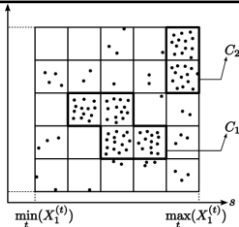
Reduksi dimensi

- Mengelompokkan / Mengaggregasi variabel serupa untuk mengurangi jumlah variabel saat melakukan analisis tanpa kehilangan terlalu banyak informasi.

Unsupervised Learning Analysis Process

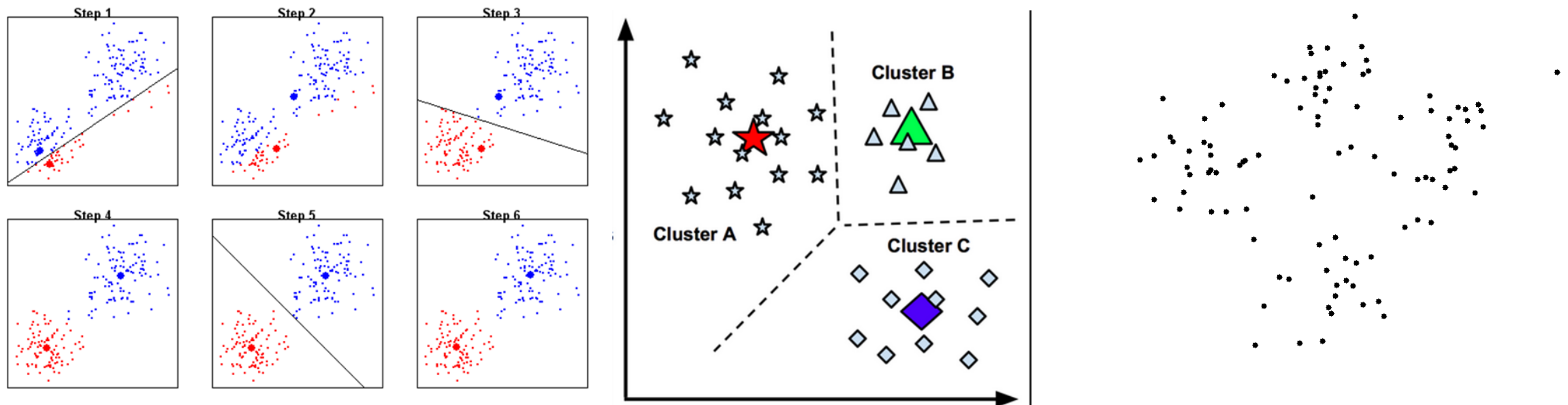


Metode Clustering

Metode	Ciri Umum	Contoh Gambar
Partitioning	<ul style="list-style-type: none"> Membagi data ke dalam sejumlah cluster yang ditentukan sebelumnya (misal, k). Algoritma yang paling dikenal adalah K-means dan K-medoids. Proses iteratif untuk meminimalkan fungsi objektif, seperti total jarak dalam cluster. Setiap titik data tepat termasuk dalam satu cluster. Efisien untuk dataset besar, tetapi kualitas clustering sangat bergantung pada pemilihan nilai k awal dan titik awal. 	
Hierarchical	<ul style="list-style-type: none"> Membangun hierarki atau pohon cluster secara bertahap, bisa agglomerative (bottom-up) atau divisive (top-down). Tidak memerlukan penentuan jumlah cluster pada awal algoritma. Hasilnya bisa divisualisasikan menggunakan dendrogram, yang memudahkan interpretasi dan pemilihan jumlah cluster. Lebih mudah untuk menangani bentuk cluster yang non-globular atau ukuran cluster yang berbeda. Komputasi bisa menjadi sangat intensif, terutama untuk dataset yang besar. 	
Density-Based	<ul style="list-style-type: none"> Cluster dibentuk berdasarkan area kepadatan data yang tinggi, dipisahkan oleh area kepadatan yang rendah. Populer dengan DBSCAN (Density-Based Spatial Clustering of Applications with Noise) sebagai contoh utamanya. Mampu mengidentifikasi cluster dengan bentuk apapun dan bisa menangani noise atau outlier. Tidak memerlukan penentuan jumlah cluster terlebih dahulu. Efektivitasnya tergantung pada pemilihan parameter kepadatan (seperti radius dan jumlah minimum poin). 	
Grid-Based	<ul style="list-style-type: none"> Ruang data dibagi menjadi sejumlah sel yang membentuk grid. Cluster dibentuk berdasarkan kepadatan sel dalam grid, bukan berdasarkan jarak antar titik data. Cepat dan efisien karena kompleksitas waktu tergantung pada jumlah sel dalam grid, bukan pada jumlah data. Algoritma yang terkenal adalah STING (Statistical Information Grid) dan CLIQUE (Clustering In QUEst). Sangat efektif untuk dataset berdimensi besar, namun resolusi grid dapat mempengaruhi kualitas clustering. 	

K-Means Clustering

K-Means adalah algoritma unsupervised sederhana untuk melakukan clustering/segmentasi. Intinya adalah dengan mendefinisikan jumlah cluster dan titik tengah tiap cluster.



Step K-Means :

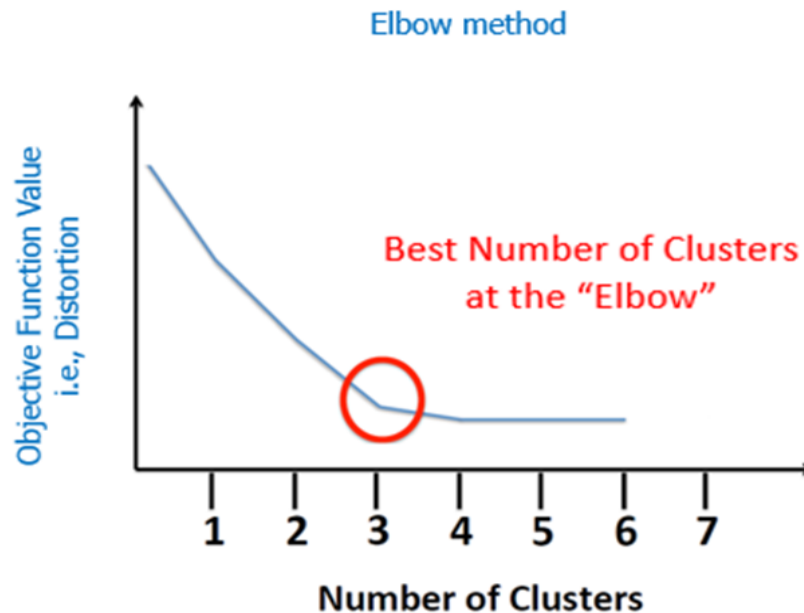
- Randomly initialize points called the cluster centroids (K).
- Iteration :
 1. Cluster assignment.
 2. Move centroid step.

Elbow Method

Why use elbow method? → **Validate the number of clusters**

- Pada K-Means :
 1. Jumlah cluster harus didefinisi.
 2. Menggunakan perhitungan matematika (jarak). Jadi data yang digunakan adalah data numerik kontinu.

Inti elbow method → Jalankan K-Means clustering pada rentang nilai K/jumlah cluster (misal dari 1-10 cluster) dan tiap nilai K dihitung sum of squared distances (SSD).



Jika grafik garis terlihat seperti lengan, maka “siku” pada lengan tersebut adalah nilai k yang terbaik.

Hierarchical Clustering

Hierarchical Clustering adalah metode analisis kelompok yang berusaha untuk membangun sebuah hirarki kelompok data.

Agglomerative Hierarchical Clustering (Bottom-Up)	Divisive Hierarchical Clustering (Top-Down)
<div><div>1. Inisialisasi:</div><div><div>• Mulai dengan menganggap setiap titik data sebagai cluster sendiri, sehingga jika ada N titik data, maka terdapat N cluster pada awalnya.</div></div><div>2. Komputasi Jarak Antar Cluster:</div><div><div>• Hitung jarak antar semua pasangan cluster. Jarak antar cluster dapat diukur dengan berbagai cara, seperti jarak terdekat (single linkage), jarak terjauh (complete linkage), rata-rata jarak (average linkage), atau jarak antara centroid (centroid linkage).</div></div><div>3. Penggabungan Cluster:</div><div><div>• Cari dua cluster yang paling dekat berdasarkan perhitungan jarak di tahap sebelumnya, dan gabungkan menjadi satu cluster baru.</div></div><div>4. Update Matriks Jarak:</div><div><div>• Setelah penggabungan, update matriks jarak untuk mencerminkan jarak antara cluster baru dengan cluster lainnya.</div></div><div>5. Ulangi:</div><div><div>• Ulangi langkah 3 dan 4 sampai semua titik data tergabung dalam satu cluster tunggal.</div></div><div>6. Pembuatan Dendrogram:</div><div><div>• Proses penggabungan ini dapat divisualisasikan menggunakan dendrogram yang menunjukkan pada jarak atau level dimana penggabungan terjadi, memberikan insight visual mengenai struktur data.</div></div></div>	<div><div>1. Inisialisasi:</div><div><div>• Mulai dengan satu cluster yang berisi semua titik data, sehingga seluruh dataset dianggap sebagai satu cluster besar.</div></div><div>2. Memilih Cluster untuk Dibagi:</div><div><div>• Pada awalnya, seluruh dataset adalah target pembagian. Selanjutnya, pilih cluster untuk dibagi berdasarkan kriteria tertentu, seperti ukuran cluster atau heterogenitas.</div></div><div>3. Mencari Sub-cluster:</div><div><div>• Dalam cluster yang dipilih, identifikasi sub-cluster menggunakan teknik clustering lain (misalnya, K-means) atau berdasarkan jarak tertentu dalam cluster tersebut.</div></div><div>4. Pembagian Cluster:</div><div><div>• Bagi cluster terpilih menjadi dua atau lebih sub-cluster berdasarkan hasil langkah sebelumnya.</div></div><div>5. Ulangi:</div><div><div>• Ulangi langkah 2 sampai 4 untuk setiap cluster yang telah dibagi hingga setiap titik data menjadi cluster sendiri atau sampai memenuhi kriteria tertentu (misal, jumlah cluster minimum atau threshold heterogenitas).</div></div><div>6. Konstruksi Dendrogram (Opsional):</div><div><div>• Meskipun lebih umum dalam agglomerative, dendrogram juga bisa dibuat untuk divisive, menunjukkan pembagian cluster dari atas ke bawah.</div></div></div>

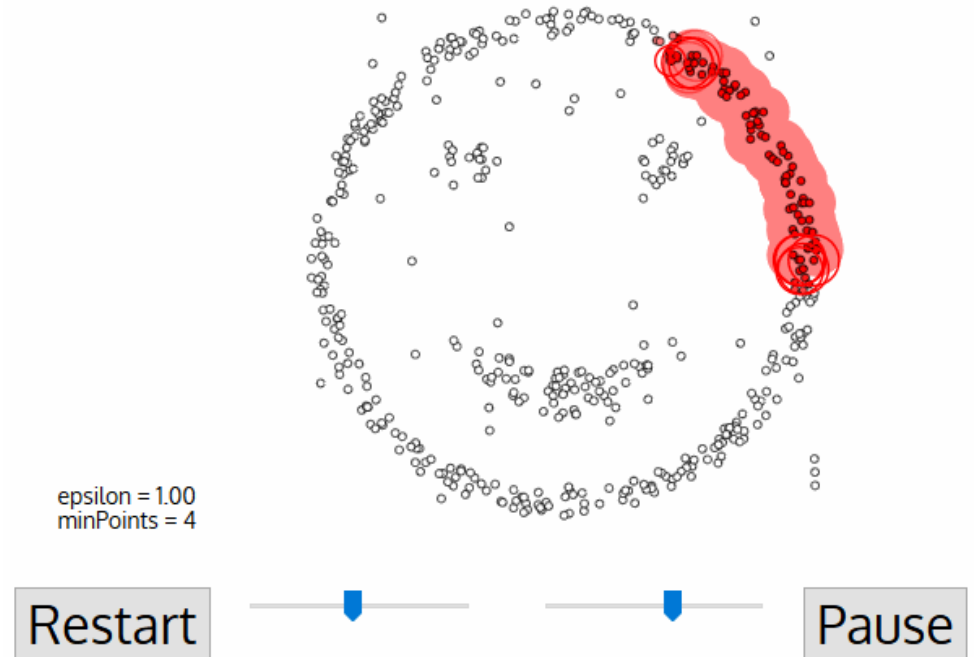
Contoh Perhitungan dalam Hierarchical Clustering

Source: <http://www.afif.lecture.ub.ac.id/files/2014/05/Slide-12-Klasterisasi-Hierarchical-Clustering.pdf>

DBSCAN

(Density-Based Spatial Clustering of Applications with Noise)

- DBSCAN adalah pengelompokan aplikasi spasial berbasis kepadatan dengan metode pengelompokan kebisingan (DBSCAN).
- Algoritma DBSCAN didasarkan pada gagasan intuitif tentang “cluster” dan “noise”. Ide utamanya adalah bahwa untuk setiap titik dalam sebuah cluster, lingkungan dengan radius tertentu harus memuat setidaknya sejumlah titik minimum.
- DBSCAN adalah algoritma clustered berbasis kepadatan yang mirip dengan mean-shift, namun dengan beberapa keunggulan penting.
- DBSCAN melibatkan pencarian area dengan kepadatan tinggi dalam domain dan memperluas area ruang fitur di sekitarnya sebagai cluster.



How DBSCAN works?

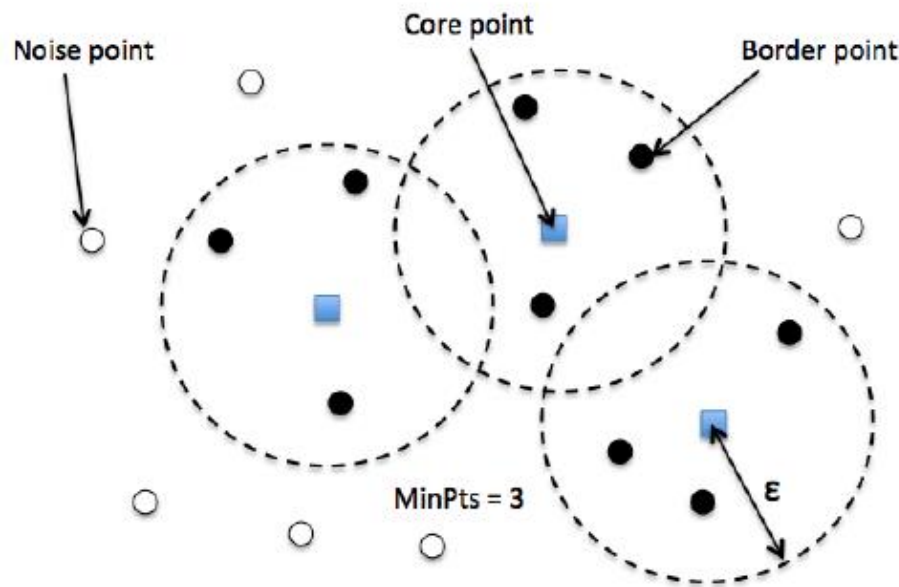
1. DBSCAN dimulai dengan titik data awal sembarang yang belum dikunjungi. Lingkungan titik ini diekstraksi menggunakan jarak epsilon ϵ (Semua titik yang berada dalam jarak ϵ adalah titik lingkungan).
2. Jika terdapat jumlah titik yang cukup (menurut minPoints) dalam lingkungan ini maka proses pengelompokan dimulai dan titik data saat ini menjadi titik pertama dalam cluster baru. Jika tidak, titik tersebut akan diberi label sebagai noise (nantinya titik noise ini mungkin menjadi bagian dari cluster). Dalam kedua kasus tersebut, titik tersebut ditandai sebagai “dikunjungi (visited)”.
3. Untuk titik pertama dalam cluster baru ini, titik-titik dalam lingkungan jarak ϵ juga menjadi bagian dari cluster yang sama. Prosedur untuk membuat semua titik di lingkungan ϵ menjadi milik cluster yang sama kemudian diulangi untuk semua titik baru yang baru saja ditambahkan ke grup cluster.
4. Proses langkah 2 dan 3 ini diulangi sampai semua titik dalam cluster ditentukan yaitu semua titik dalam lingkungan ϵ cluster telah dikunjungi dan diberi label.
5. Setelah kita selesai dengan cluster saat ini, titik baru yang belum dikunjungi diambil dan diproses, yang mengarah pada penemuan cluster atau noise lebih lanjut. Proses ini berulang hingga semua titik ditandai sebagai telah dikunjungi. Karena pada akhir proses ini semua titik telah dikunjungi, setiap titik akan ditandai sebagai milik cluster atau sebagai noise.

DBSCAN Algorithm

The algorithm follows the logic:

1. Identify a core point and make a group for each one, or for each connected group of core points (if they satisfy the criteria to be core point).
2. Identify and assign border points to their respective core points.

The following figure summarize very well this process and the commented notation.

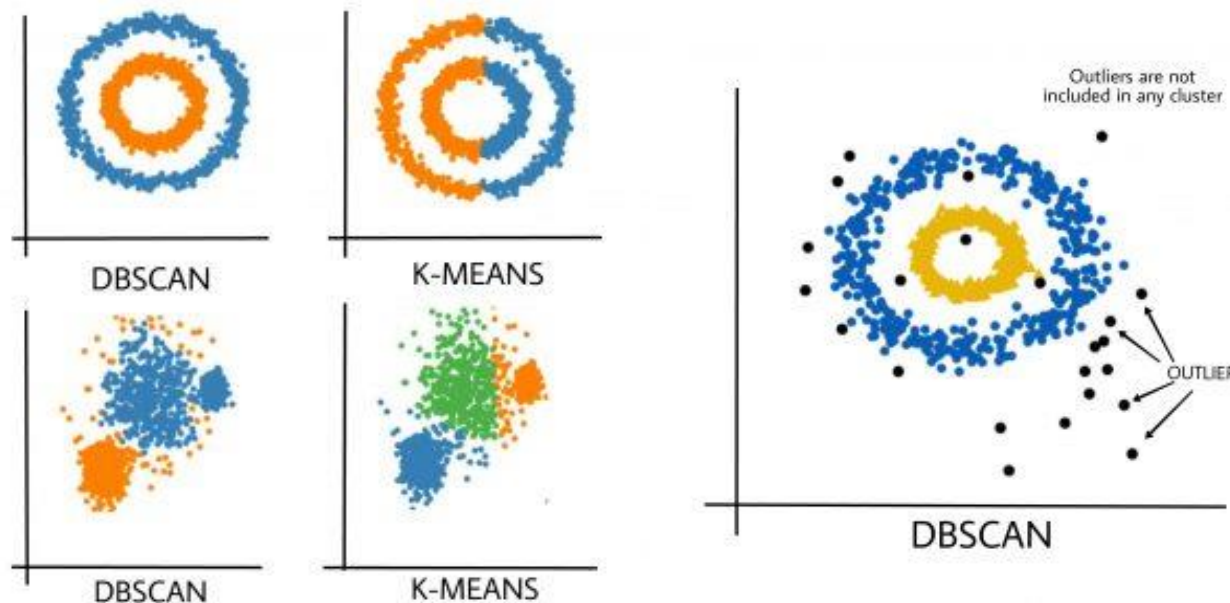


- Core Point: A point is a core point if it has more than MinPts points within ϵ .
- Border Point: A point which has fewer than MinPts within ϵ but it is in the neighborhood of a core point.
- Noise or outlier: A point which is not a core point or border point.

Perbandingan DBSCAN dengan K-Means

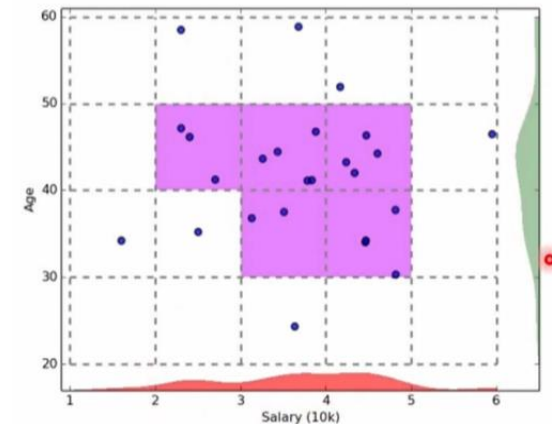
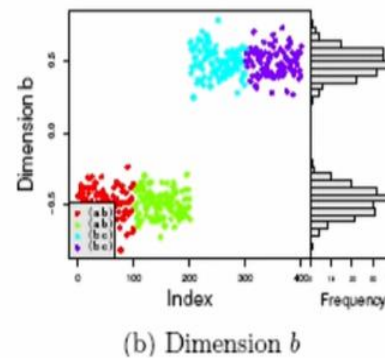
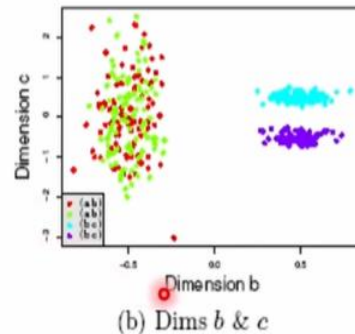
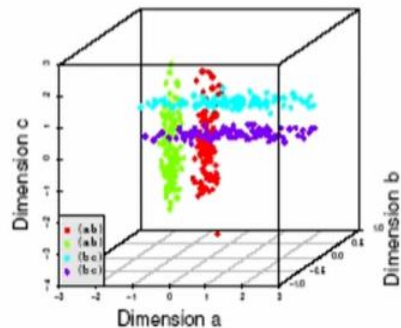
1. K-Means hanya membentuk cluster berbentuk bola. Algoritme ini gagal jika datanya tidak berbentuk bola (yaitu varians yang sama ke segala arah).
2. Algoritma K-Means sensitif terhadap outlier. Penciran dapat menyebabkan distorsi yang sangat besar pada cluster di K-Means.
3. Algoritma K-Means mengharuskan seseorang untuk menentukan jumlah cluster, sebuah biara, dll.

Pada dasarnya, algoritma DBSCAN mengatasi semua kelemahan algoritma K-Means yang disebutkan di atas. Algoritma DBSCAN mengidentifikasi wilayah padat dengan mengelompokkan titik-titik data yang berdekatan berdasarkan pengukuran jarak.



Grid-Based Clustering

- Grid-based clustering mengubah ruang data menjadi jumlah sel grid terbatas lalu melakukan operasi pada sel tersebut. Metode ini memiliki kecepatan proses yang konstan dan tidak bergantung pada jumlah objek data.
- Clique (Clustering in Quest) adalah salah satu algoritma yang menggunakan konsep grid-based clustering, yang secara otomatis mengidentifikasi subruang data dimensi tinggi yang memungkinkan pengelompokan yang lebih baik daripada ruang asli.



Perbandingan Grid-Based dengan K-Means

Grid-Based	K-Means
mengelompokkan data dengan membagi ruang data menjadi grid	mengelompokkan data ke dalam sejumlah kluster, dengan setiap kluster diwakili oleh sebuah centroid.
Jumlah cluster ditentukan berdasarkan grid	Jumlah cluster ditentukan secara manual
complexity tergantung pada number of populated grid	Complexity tergantung pada number of data
kurang sensitive terhadap outlier	Sensitive terhadap outlier

Praktik Code Clustering

Segmentasi

Segmentasi adalah proses membagi-bagi pasar/konsumen ke dalam kelompok-kelompok dengan karakteristik sama. Segmentasi dapat menjadi sebuah cara yang powerful untuk mengidentifikasi kebutuhan konsumen.

Metode umum yang sering dipakai untuk segmentasi user/customer adalah:

Demographic Information



User dikelompokkan berdasarkan gender, usia, status pernikahan, status kepemilikan rumah, atau pendidikan

Geographical Information



User dikelompokkan berdasarkan domisili tempat tinggal atau tempat user bekerja.

Behavioral Data



User dikelompokkan berdasarkan kebiasaan mengonsumsi atau menggunakan atau menghabiskan suatu produk, pendapatan, atau layanan.

Psychographics



USER BEHAVIOR

User dikelompokkan berdasarkan kelas sosial, lifestyle, dan kepribadian.

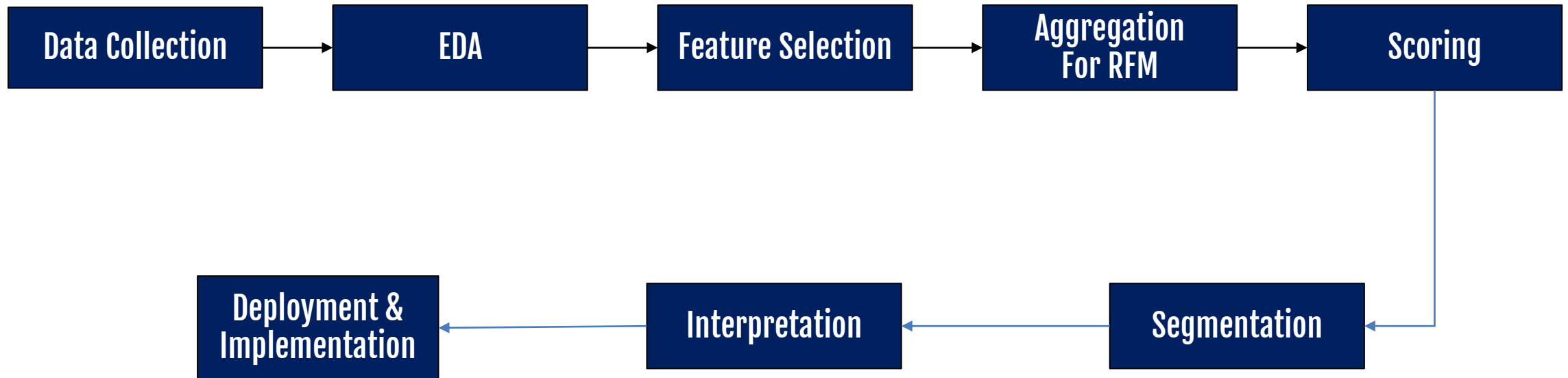
Recency Frequency Monetary (RFM)

RFM (Recency, Frequency, Monetary) merupakan metrik untuk segmentasi berdasarkan behaviour pengguna.

- Recency adalah jarak waktu antara hari ini atau tanggal maksimal di data dan terakhir kali customer bertransaksi.
- Frequency adalah jumlah transaksi customer.
- Monetary adalah total uang yang dikeluarkan customer dari seluruh transaksinya.



Workflow RFM



Praktik Code RFM

Sumber

- [https://en.wikipedia.org/wiki/Elbow_method_\(clustering\)](https://en.wikipedia.org/wiki/Elbow_method_(clustering))
- <https://neptune.ai/blog/customer-segmentation-using-machine-learning>
- <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>
- <https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python>
- <https://revou.co/kosakata/clustering>
- <https://id.scribd.com/document/264595456/Data-Mining-Partitioning-Methods>
- https://en.wikipedia.org/wiki/K-means_clustering
- <http://www.afif.lecture.ub.ac.id/files/2014/05/Slide-12-Klasterisasi-Hierarchical-Clustering.pdf>
- <https://www.kaggle.com/code/tanmay111999/clustering-pca-k-means-dbscan-hierarchical>
- <https://www.kaggle.com/discussions/getting-started/417081>
- <https://www.kaggle.com/code/egazakharenko/clustering-algorithms-from-scratch-using-python>
- <https://www.sciencedirect.com/topics/computer-science/affinity-propagation>
- https://programmersought.com/article/54505340232/#_120
- https://pyclustering.github.io/docs/0.9.0/html/d2/d4f/classpyclustering_1_1cluster_1_1clique_1_1clique.html
- <https://github.com/franciscorpuz/Grid-and-Density-based-clustering/blob/main/CLIQUE.ipynb>
- <http://eprints.umg.ac.id/1553/3/BAB%20II.pdf>



Terima kasih

