Teknik *Data Mining* : Algoritma *K-Means Clustering*

Agus Nur Khomarudin
agusnurkhomarudin@gmail.com
https://agusnkhom.wordpress.com

Basis data/database secara sederhana dapat diartikan sebagai gudang data. Tumpukan data pada basis data dapat diolah dengan memanfaatkan teknologi data mining untuk menghasilkan pengetahuan menarik/bermanfaat yang selama ini tidak diketahui secara manual. Salah satu teknik data mining adalah clustering. Algoritma K-Means Clustering sebagai salah satu metode yang mempartisi data ke dalam bentuk satu atau lebih cluster atau kelompok, sehingga data yang memiliki karakteristik yang sama dikelompokkan dalam satu cluster yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam kelompok yang lain. Kelompok atau cluster yang didapat merupakan pengetahuan/informasi yang bermanfaat bagi pengguna kebijakan dalam proses pengambilan keputusan.
Kata Kunci: *Data Mining*, *Clustering*, Algoritma *K-Means Clustering*

Pendahuluan

Perkembangan teknologi informasi yang semakin canggih saat ini, telah menghasilkan banyak tumpukan data. Pertambahan data yang semakin banyak akan menimbulkan pertanyaan besar, yaitu “apa yang dapat dilakukan dari tumpukan data tersebut?”. Untuk menjawab pertanyaan tersebut, dapat diterapkan sebuah teknologi basis data yang dikenal dengan *data mining*.

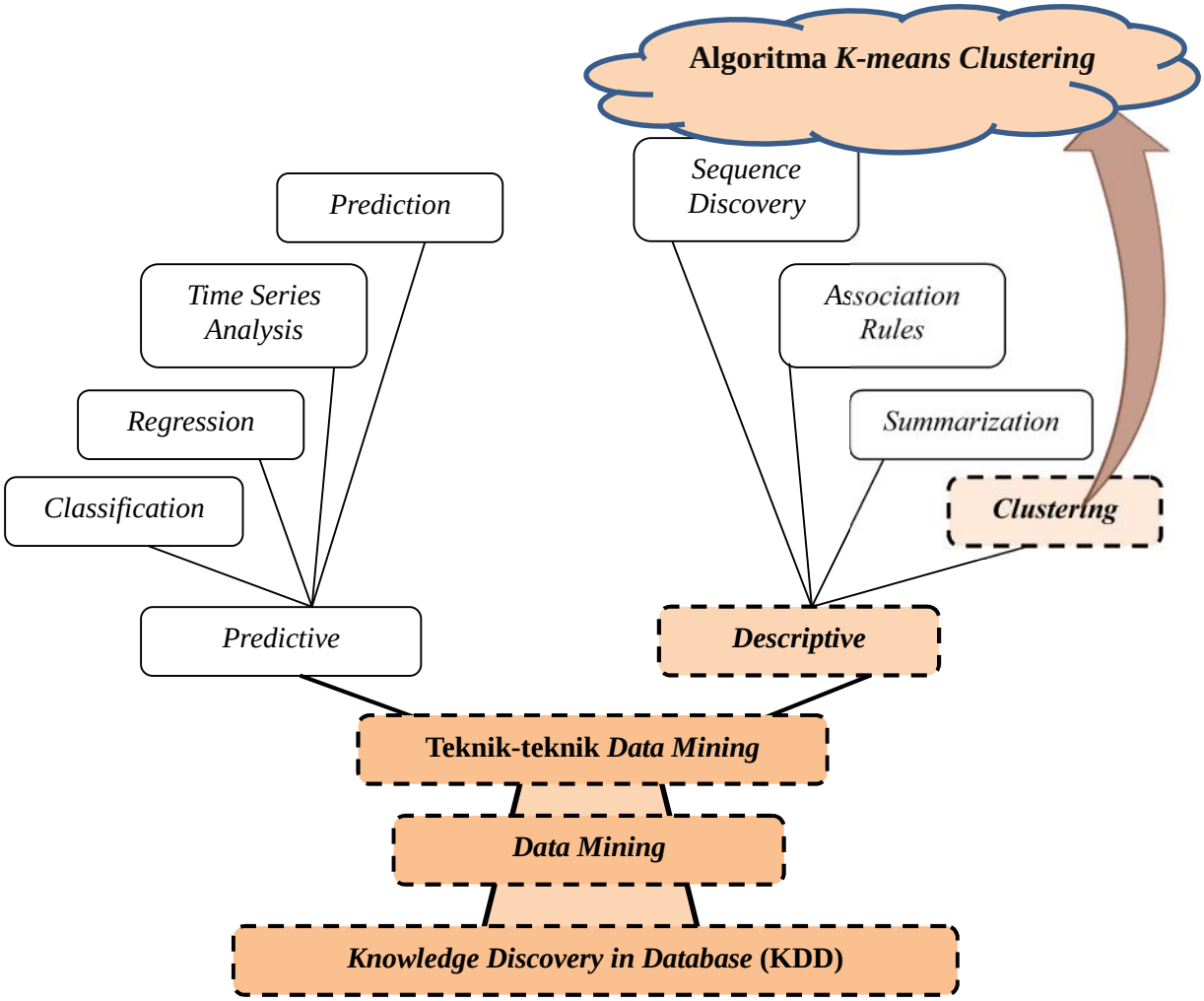
Data mining dapat diterapkan untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual. Terdapat beberapa teknik yang digunakan dalam *data mining*, salah satu teknik *data mining* adalah *clustering*. Terdapat dua jenis metode *clustering* yang digunakan dalam pengelompokkan data, yaitu *hierarchical clustering* dan *non-hierarchical clustering*.

K-means clustering sebagai salah satu metode data *clustering* non-hirarki mempartisi data yang ada ke dalam bentuk satu atau lebih *cluster* atau kelompok, sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu *cluster* yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam kelompok yang lain. Kelompok atau *cluster* yang didapat merupakan pengetahuan/informasi yang bermanfaat bagi pengguna kebijakan dalam proses pengambilan keputusan.

Pembahasan

1. Peta Konsep/*Mind Mapp*

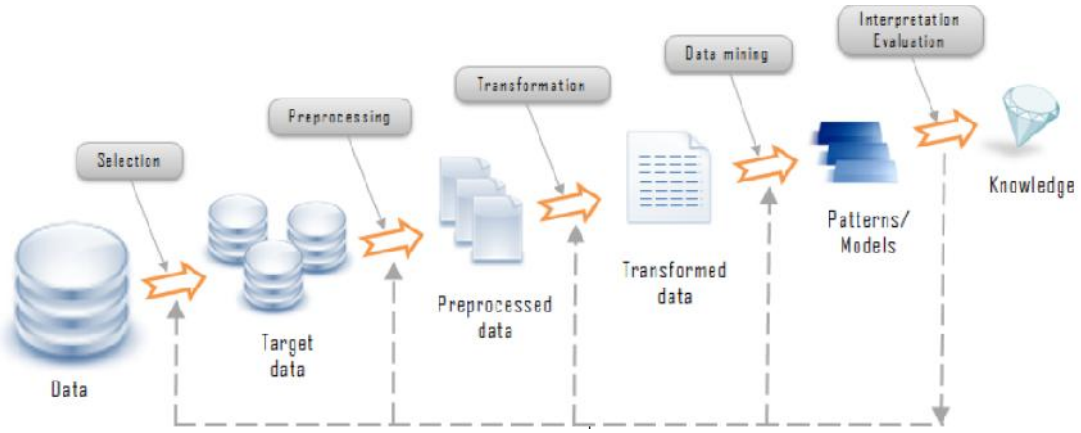
Peta konsep di bawah ini bertujuan untuk memudahkan kita dalam memahami materi yang dibahas dalam makalah ini. Adapun peta konsep makalah ini adalah sebagai berikut :



Gambar 1 Peta Konsep/*Mind Mapp* Makalah

2. *Knowledge Discovery in Database* (KDD)

Knowledge Discovery in Database (KDD) adalah proses menemukan pengetahuan yang berguna dari sebuah data yang bervolume besar, dan sering disebut sebagai *data mining*. KDD adalah proses yang terorganisir untuk mengidentifikasi pola-pola yang berlaku, berguna dan mudah dipahami dari kumpulan data yang besar dan kompleks. *Data mining* adalah inti dari proses KDD, yang melibatkan dalam menyimpulkan algoritma yang menjelajahi data, mengembangkan model dan menemukan pola-pola yang sebelumnya tidak diketahui. Model ini digunakan untuk memahami fenomena dari data, analisis dan prediksi. Proses dalam *Knowledge Discovery in Database* (KDD) dapat diilustrasikan pada gambar 2 berikut :



Gambar 2 Proses *Knowledge Discovery in Database*

Data yang dikumpulkan dari berbagai sumber heterogen yang terintegrasi ke dalam penyimpanan data tunggal yang disebut sebagai data target. Data yang relevan diputuskan untuk dianalisis dan diperoleh dari pengumpulan data. Kemudian, itu adalah pra-diproses dan berubah menjadi format standar yang sesuai. *Data mining* adalah langkah yang paling inti dalam algoritma/teknik intelligensi yang diterapkan untuk mengekstrak pola atau aturan yang bermakna. Akhirnya, pola dan aturan yang ditafsirkan tersebut menjadi pengetahuan atau informasi yang baru dan berguna.

3. *Data Mining*

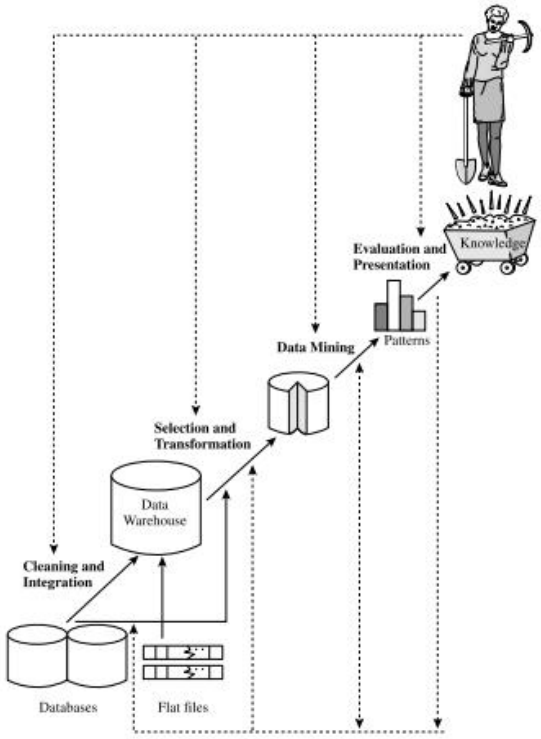
Secara sederhana, *data mining* dapat diartikan sebagai proses mengekstrak atau menggali *knowledge* yang ada pada sekumpulan data. Informasi dan *knowledge* yang didapat tersebut dapat digunakan pada banyak bidang, seperti manajemen bisnis, pendidikan, kesehatan dan sebagainya. Menurut Tacbir, *data mining* adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari *database* yang besar. Istilah *data mining* memiliki hakikat sebagai disiplin ilmu yang tujuan utamanya adalah untuk menemukan, menggali, atau menambang pengetahuan dari data atau informasi yang kita miliki. Proses menggali informasi dalam *data mining* melibatkan integrasi teknik dari berbagai disiplin ilmu, seperti teknologi *database* dan *data warehouse*, statistik, *machine learning*, komputasi dengan kinerja tinggi, *pattern recognition*, *neural network*, visualisasi data dan sebagainya.

Data mining menggunakan pendekatan *discovery-based* dimana pencocokan pola (*pattern matching*) dan algoritma-algoritma yang lain digunakan untuk menentukan relasi-relasi kunci di dalam data yang dieksplorasi. *Data mining* (penambangan data), sesuai dengan namanya, berkonotasi sebagai pencarian informasi bisnis yang berharga dari basis data yang sangat besar. Dengan tersedianya basis data dalam kualitas dan ukuran yang memadai, teknologi *data mining* memiliki kemampuan-kemampuan sebagai berikut:

a. Mengotomatisasi prediksi *trend* sifat-sifat bisnis. *Data mining* mengotomatisasi proses pencarian informasi di dalam basis data yang besar.
b. Mengotomatisasi penemuan pola-pola yang tidak diketahui sebelumnya. *Tools data mining* "menyapu" basis data, kemudian mengidentifikasi pola-pola yang sebelumnya tersembunyi dalam satu sapuan. Contoh dari penemuan pola ini adalah analisis pada data penjualan *ritel* untuk mengidentifikasi produk-produk yang kelihatannya tidak berkaitan, yang seringkali dibeli secara bersamaan oleh *customer*.

a. Tahapan dalam *Data Mining*

Sebagai suatu rangkaian proses, *data mining* dapat dibagi menjadi beberapa tahap proses yang diilustrasikan pada Gambar 3 Tahap-tahap tersebut bersifat interaktif, pemakai terlibat langsung atau dengan perantaraan *knowledge base*.



Gambar 3 Tahap-tahap *Data Mining*

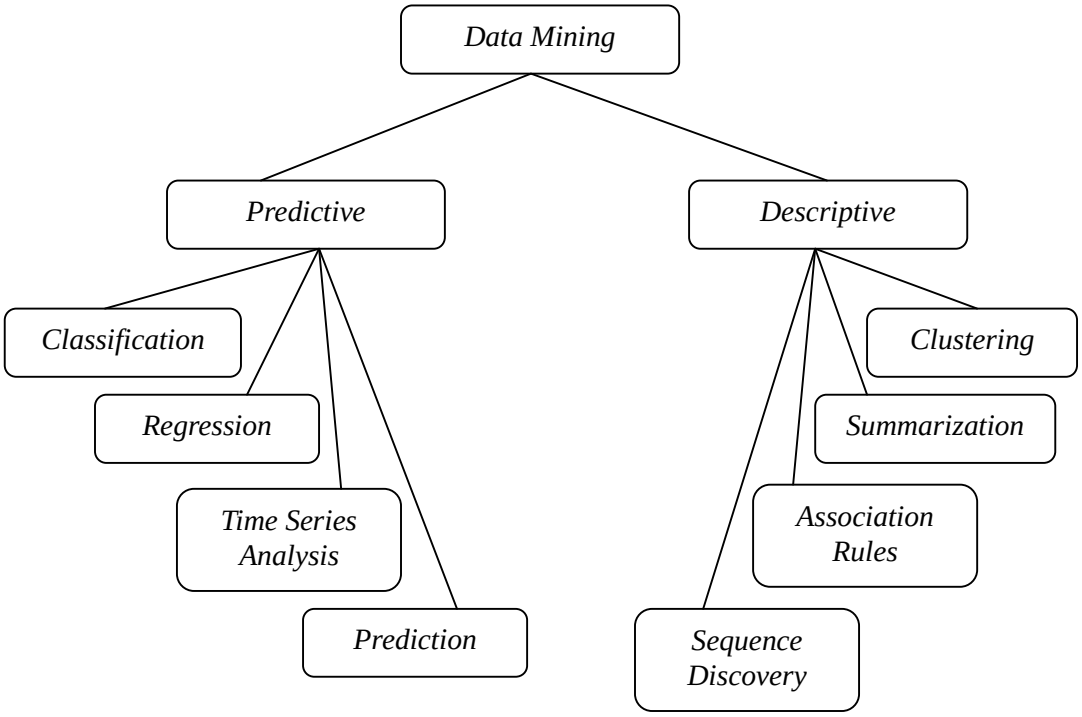
Tahap-tahap *data mining* adalah sebagai berikut:

a. Pembersihan data (*data cleaning*)
   Pembersihan data merupakan proses menghilangkan *noise* dan data yang tidak konsisten atau data tidak relevan.
b. Integrasi data (*data integration*)
   Integrasi data merupakan penggabungan data dari berbagai *database* ke dalam satu *database* baru.
c. Seleksi data (*data selection*)
   Data yang ada pada *database* sering kali tidak semuanya dipakai, oleh karena itu hanya data yang sesuai untuk dianalisis yang akan diambil dari *database*.
d. Transformasi data (*data transformation*)
   Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam *data mining*.
e. Proses *mining*
   Merupakan suatu proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data.
f. Evaluasi pola (*pattern evaluation*)
   Untuk mengidentifikasi pola-pola menarik ke dalam *knowledge based* yang ditemukan.
g. Presentasi pengetahuan (*knowledge presentation*)
   Merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna.

b. Teknik-Teknik *Data mining*

Data mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual. Perlu diingat bahwa kata *mining* sendiri berarti usaha untuk mendapatkan sedikit data berharga dari sejumlah besar data dasar. Karena itu *data mining* sebenarnya memiliki akar yang panjang dari bidang ilmu seperti kecerdasan buatan (*artificial intelligent*), *machine learning*, statistik dan basis data.

Menurut Ahmed, teknik *data mining* biasanya terbagi dalam dua kategori, prediksi dan deskripsi. Teknik prediksi menggunakan data historis untuk menyimpulkan sesuatu tentang kejadian di masa depan. Sedangkan teknik deskripsi bertujuan untuk menemukan pola dalam data yang menyediakan beberapa informasi tentang hubungan interval yang tersembunyi.



Gambar 4 Teknik *Data Mining*

Menurut Kumar dan Saurabh, terdapat beberapa teknik yang digunakan dalam *data mining*, yaitu:

1. *Classification*
   Klasifikasi adalah teknik yang paling umum diterapkan pada *data mining*. Pendekatan ini sering menggunakan keputusan pohon (*decision tree*) atau *neural network* berbasis algoritma klasifikasi. Proses klasifikasi data melibatkan *learning* dan klasifikasi. Dalam belajar (*learning*) data pelatihan (*training*) dianalisis dengan alogaritma klasifikasi. Dalam klasifikasi pengujian data dilakukan dengan menggunakan perkiraan akurasi dari aturan klasifikasi. Jika akurasi bisa diterima, maka aturan dapat diterapkan untuk data baru. Salah satu contoh yang mudah dan populer adalah dengan *decision tree* yaitu salah satu metode klasifikasi yang paling populer karena mudah untuk diinterpretasi. *Decision tree* adalah model prediksi menggunakan struktur pohon atau struktur berhirarki.

Decision tree adalah struktur *flowchart* yang menyerupai *tree* (pohon), dimana setiap simpul internal menandakan suatu tes pada atribut, setiap cabang merepresentasikan hasil tes, dan simpul daun merepresentasikan kelas atau distribusi kelas. Alur pada *decision tree* di telusuri dari simpul akar ke simpul daun yang memegang prediksi kelas untuk contoh tersebut. *Decision tree* mudah untuk dikonversi ke aturan klasifikasi (*classification rules*).

2. *Clustering*
   Clustering bisa dikatakan sebagai identifikasi kelas objek yang memiliki kemiripan. Dengan menggunakan teknik *clustering* kita bisa lebih lanjut mengidentifikasi kepadatan dan jarak daerah dalam objek ruang dan dapat menemukan secara keseluruhan pola distribusi dan korelasi antara atribut. Pendekatan klasifikasi secara efektif juga dapat digunakan untuk membedakan kelompok atau kelas objek.

3. *Predication*
   Teknik regresi dapat disesuaikan untuk predikasi. Analisis regresi dapat digunakan untuk model hubungan antara satu atau lebih *independent variables* dan *dependent variables*. Dalam *data mining independent variabel* adalah atribut-atribut yang sudah dikenal dan respon variabel apa yang kita inginkan untuk diprediksi. Akan tetapi, banyak masalah di dunia nyata bukan prediksi yang mudah. Karena itu, teknik kompleks (seperti: *logistic regression*, *decision trees* atau pohon keputusan, *neural nets* atau jaringan syaraf) mungkin akan diperlukan untuk memprediksi nilai. Model yang berjenis sama sering dapat digunakan untuk regresi dan klasifikasi. Misalnya, CART (*Classification and Regression Trees*) yaitu alogaritma pohon keputusan yang dapat digunakan untuk membangun kedua pohon klasifikasi dan pohon regresi. Jaringan saraf juga dapat menciptakan kedua model klasifikasi dan regresi.

4. *Association rule*
   Digunakan untuk mengenali kelakuan dari kejadian-kejadian khusus atau proses dimana link asosiasi muncul pada setiap kejadian. Contoh dari aturan assosiatif dari analisa pembelian di suatu pasar swalayan adalah bisa diketahui berapa besar kemungkinan seorang pelanggan membeli roti bersamaan dengan susu. Dengan pengetahuan tersebut pemilik pasar swalayan dapat mengatur penempatan barangnya atau merancang kampanye pemasaran dengan memakai kupon diskon untuk kombinasi barang tertentu.

   Penting tidaknya suatu aturan assosiatif dapat diketahui dengan dua parameter, *support* yaitu prosentasi kombinasi atribut tersebut dalam basisdata dan *confidence* yaitu kuatnya hubungan antar atribut dalam aturan asosiatif. Motivasi awal pencarian *association rule* berasal dari keinginan untuk menganalisa data transaksi supermarket, ditinjau dari perilaku *customer* dalam membeli produk. *Association rule* ini menjelaskan seberapa sering suatu produk dibeli secara bersamaan. Sebagai contoh, *association rule* "*beer* => *diaper* (80%)" menunjukkan bahwa empat dari lima *customer* yang membeli *beer* juga membeli *diaper*. Dalam suatu *association rule* X => Y, X disebut dengan *antecedent* dan Y disebut dengan *consequent rule*.

5. *Neural network*
   Jaringan saraf adalah seperangkat unit penghubung *input* dan *output* dimana setiap koneksinya memiliki bobot. Selama fase *learning*, jaringan belajar dengan menyesuaikan bobot sehingga dapat memprediksi kelas yang benar label dari setiap *input*. Jaringan saraf memiliki kemampuan yang luar biasa untuk memperoleh arti

dari data yang rumit atau tidak tepat dan dapat digunakan untuk mengambil pola-pola serta mendeteksi *tren* yang sangat kompleks untuk diperhatikan baik oleh manusia atau teknik komputer lain. Jaringan saraf sangat baik untuk mengidentifikasi pola atau *tren* pada data dan sangat cocok untuk melakukan prediksi serta memprediksi kebutuhan.

6. *Decision trees*
   Decision trees atau pohon keputusan adalah struktur *tree-shaped* yang mewakili set keputusan. Keputusan ini menghasilkan aturan untuk klasifikasi sebuah kumpulan data. Metode pohon keputusan diantaranya yaitu *Classification and regression trees* (CART) dan *Chi Square Automatic Interaction Detection* (CHAID).

7. *Nearest Neighbor Method*
   Teknik yang mengklasifikasikan setiap *record* dalam sebuah kumpulan data berdasarkan sebuah kombinasi suatu kelas k *record* yang sama dalam sebuah kumpulan data historis (dimana k lebih besar atau sama dengan 1). Terkadang disebut juga dengan teknik *K-Nearest Neighbor*.

4. *Clustering*

Madhu Yedha mendefenisikan *clustering* sebagai proses pengorganisasian objek data ke dalam *set* kelas yang saling berhubungan, yang disebut *cluster*. *Clustering* merupakan contoh dari klasifikasi tanpa arahan (*unsupervised*). Klasifikasi merujuk kepada prosedur yang menetapkan objek data set kelas. *Unsupervised* berarti bahwa pengelompokkan tidak tergantung pada standar kelas dan pelatihan atau *training*.

Menurut Deka, *Clustering* merupakan salah satu teknik *data mining* yang digunakan untuk mendapatkan kelompok-kelompok dari objek-objek yang mempunyai karakteristik yang umum di data yang cukup besar. Tujuan utama dari metode *clustering* adalah pengelompokan sejumlah data atau objek ke dalam *cluster* atau grup sehingga dalam setiap *cluster* akan berisi data yang semirip mungkin. *Clustering* melakukan pengelompokkan data yang didasarkan pada kesamaan antar objek, oleh karena itu klasterisasi digolongkan sebagai metode *unsupervised learning*. Menurut Oyelade, *clustering* dapat dibagi menjadi dua, yaitu *hierarchical clustering* dan *non-hierarchical clustering*.

Hierarchical clustering adalah suatu metode pengelompokan data yang dimulai dengan mengelompokkan dua atau lebih objek yang memiliki kesamaan paling dekat. Kemudian proses diteruskan ke objek lain yang memiliki kedekatan kedua. Demikian seterusnya sehingga *cluster* akan membentuk semacam pohon dimana ada hierarki (tingkatan) yang jelas antar objek, dari yang paling mirip sampai yang paling tidak mirip. Secara logika semua objek pada akhirnya hanya akan membentuk sebuah *cluster*. *Dendogram* biasanya digunakan untuk membantu memperjelas proses hierarki tersebut.

Berbeda dengan metode *hierarchical clustering*, metode *non-hierarchical clustering* justru dimulai dengan menentukan terlebih dahulu jumlah *cluster* yang diinginkan (dua *cluster*, tiga *cluster*, atau lain sebagainya). Setelah jumlah *cluster* diketahui, baru proses *cluster* dilakukan tanpa mengikuti proses hierarki. Metode ini biasa disebut dengan *K-Means Clustering*.

Algoritma *K-means Clustering*

K-means clustering merupakan salah satu metode *cluster analysis* non hirarki yang berusaha untuk mempartisi objek yang ada kedalam satu atau lebih *cluster* atau

kelompok objek berdasarkan karakteristiknya, sehingga objek yang mempunyai karakteristik yang sama dikelompokan dalam satu *cluster* yang sama dan objek yang mempunyai karakteristik yang berbeda dikelompokan kedalam *cluster* yang lain.

Menurut Daniel dan Eko, Langkah-langkah algoritma *K-Means* adalah sebagai berikut:

a. Pilih secara acak k buah data sebagai pusat *cluster*.
b. Jarak antara data dan pusat *cluster* dihitung menggunakan *Euclidian Distance*. Untuk menghitung jarak semua data ke setiap titik pusat *cluster* dapat menggunakan teori jarak *Euclidean* yang dirumuskan sebagai berikut:

$$D(i,j) = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \ldots + (X_{ki} - X_{kj})^2}$$

dimana:
D (i,j) = Jarak data ke *i* ke pusat *cluster j*
X_{ki} = Data ke *i* pada atribut data ke *k*
X_{kj} = Titik pusat ke *j* pada atribut ke *k*

c. Data ditempatkan dalam *cluster* yang terdekat, dihitung dari tengah *cluster*.
d. Pusat *cluster* baru akan ditentukan bila semua data telah ditetapkan dalam *cluster* terdekat.
e. Proses penentuan pusat *cluster* dan penempatan data dalam *cluster* diulangi sampai nilai *centroid* tidak berubah lagi.

Berikut ini adalah contoh penerapan algoritma *K-Means:*

Tabel 1 Data Mahasiswa

No	Nama	Jurusan	Kota Asal	IPK
1	Ade Supryan Stefanus	IS	Jakarta	3,16
2	Adelina Ganardi Putri Hardi	ACC	Semarang	3,22
3	Adeline Dewita	BF	Bekasi	3,29
4	Adiputra	IB	Jakarta	2,83
5	Afrieska Laura Trisyana	PR	Jakarta	3,15
6	Agam Khalilullah	IB	Banda Aceh	3,25
7	Agus Mulyana Jungjungan	IB	Bogor	3,43
8	Agusman	PR	Bekasi	3,06
9	Aidil Friadi	BF	Banda Aceh	3,36
10	Ajeng Putri Ariandhani	ACC	Bandung	3,28

Transformasi Data

Agar data di atas dapat diolah dengan menggunakan metode *k-means clustering*, maka data yang berjenis data *nominal* seperti kota asal dan jurusan harus diinisialisasikan terlebih dahulu dalam bentuk angka.

Tabel 2 Inisialisasi Data Wilayah Kota Asal

Wilayah	Frekuensi	Inisial
Jakarta	84	1
Jawa Barat	82	2
Sumatera Utara	28	3
Sulawesi	14	4
Jawa Timur	13	5
Sumatera Selatan	13	6
Bali	8	7
Kalimantan	1	8

Tabel 3 Inisialisasi Data Jurusan

Jurusan	Singkatan	Frekuensi	Inisial
Accounting	ACC	46	1
Management, concentration in International Business	IB	37	2
Public Relation	PR	35	3
Management, concentration in Banking & Finance	BF	28	4
Industrial Engineering	IE	23	5
Information Technology	IT	20	6
Management, concentration in Marketing	MKT	18	7
Visual Communication Design	VCD	12	8
Management, concentration in Hotel & Tourism Management	HTM	9	9
Electrical Engineering	EE	6	10
Business Administration	BA	4	11
International Relations	IR	2	12
Management, concentration in Human Resources Management	HRM	1	13
Information System	IS	1	14
Management	MGT	1	15

Pengolahan data

Setelah semua data mahasiswa ditransformasi ke dalam bentuk angka, maka data-data tersebut telah dapat dikelompokan dengan menggunakan algoritma *K-Means Clustering*. Untuk dapat melakukan pengelompokan data-data tersebut menjadi beberapa *cluster* perlu dilakukan beberapa langkah, yaitu:

1. Tentukan jumlah *cluster* yang diinginkan. Dalam penelitian ini data-data yang ada akan dikelompokkan mejadi tiga *cluster*.
2. Tentukan titik pusat awal dari setiap *cluster*. Dalam penelitian ini titik pusat awal ditentukan secara *random* dan didapat titik pusat dari setiap *cluster* dapat dilihat pada tabel 2.4.

Tabel 4 Titik Pusat Awal Setiap *Cluster*

Titik Pusat awal	Nama	Jurusan	Kota Asal	IPK
Cluster 1	Dally Teguh Sesario	9	3	2,94
Cluster 2	Hervina Juliana	1	1	3,18
Cluster 3	Pascal Muhammadi	1	2	3,15

3. Tempatkan setiap data pada *cluster*. Dalam penelitian ini digunakan metode *hard k-means* untuk mengalokasikan setiap data ke dalam suatu *cluster*, sehingga data akan dimasukan dalam suatu *cluster* yang memiliki jarak paling dekat dengan titik pusat dari setiap *cluster*. Untuk mengetahui *cluster* mana yang paling dekat dengan data, maka perlu dihitung jarak setiap data dengan titik pusat setiap *cluster*. Sebagai contoh, akan dihitung jarak dari data mahasiswa pertama ke pusat *cluster* pertama:

$$D\ (1,1) = \sqrt{(14-9)^2 + (1-3)^2 + (3{,}16-2{,}94)^2} = 5{,}390$$

Dari hasil perhitungan di atas didapatkan hasil bahwa jarak data mahasiswa pertama dengan pusat *cluster* pertama adalah 5,390.
Jarak data mahasiswa pertama ke pusat *cluster* kedua:

$$D\ (1,2) = \sqrt{(14-1)^2 + (1-1)^2 + (3{,}16-3{,}18)^2} = 13{,}000$$

Dari hasil perhitungan di atas didapatkan hasil bahwa jarak data mahasiswa pertama dengan pusat *cluster* kedua adalah 13.
Jarak data mahasiswa pertama ke pusat *cluster* ketiga:

$$D\ (1,3) = \sqrt{(14-1)^2 + (1-2)^2 + (3{,}16-3{,}15)^2} = 13{,}038$$

Dari hasil perhitungan di atas didapatkan hasil bahwa jarak data mahasiswa pertama dengan pusat *cluster* ketiga adalah 13.038.
Berdasarkan hasil ketiga perhitungan di atas dapat disimpulkan bahwa jarak data mahasiswa pertama yang paling dekat adalah dengan *cluster* 1, sehingga data mahasiswa pertama dimasukkan ke dalam *cluster* 1. Hasil perhitungan selengkapnya untuk 5 data mahasiswa pertama dapat di lihat pada tabel 2.5.

Tabel 5 Contoh Hasil Perhitungan Setiap Data ke Setiap *Cluster*

No	Nama	Jurusan	Kota Asal	IPK	Jarak Ke			Jarak terdekat ke *Cluster*
					C1	C2	C3	
1	Ade Supryan Stefanus	14	1	3,16	5,390	13,000	13,038	1
2	Adelina Ganardi Putri Hardi	1	5	3,22	8,251	4,000	3,001	3
3	Adeline Dewita	4	2	3,29	5,111	3,164	3,003	3
4	Adiputra	2	1	2,83	7,281	1,059	1,450	2
5	Afrieska Laura Trisyana	3	1	3,15	6,328	2,000	2,236	2

4. Setelah semua data ditempat ke dalam *cluster* yang terdekat, kemudian hitung kembali pusat *cluster* yang baru berdasarkan rata-rata angggota yang ada pada *cluster* tersebut.
5. Setelah didapatkan titik pusat yang baru dari setiap *cluster*, lakukan kembali dari langkah ketiga hingga titik pusat dari setiap *cluster* tidak berubah lagi dan tidak ada lagi data yang berpindah dari satu *cluster* ke *cluster* yang lain.

Dalam penelitian ini, iterasi *clustering* data mahasiswa terjadi sebanyak 7 kali iterasi. Pada iterasi ke-7 ini, titik pusat dari setiap *cluster* sudah tidak berubah dan tidak ada lagi data yang berpindah dari satu *cluster* ke *cluster* yang lain.

Dari hasil *cluster* 1, terlihat bahwa karakteristik mahasiswa pada *cluster* 1 didominasi oleh mahasiswa yang berasal dari jurusan *Information Technology* dan *Marketing*. Sedangkan, berdasarkan kota asal didominasi oleh mahasiswa yang berasal dari wilayah kota asal DKI Jakarta dan Jawa Barat, sehingga dapat disimpulkan bahwa rata-rata mahasiswa pada *cluster* 1 yang berasal dari wilayah kota asal DKI Jakarta dan Jawa Barat mengambil jurusan *Infromation Technology* dan *Marketing*.

Tabel 6 Hasil Analisa *Clustering*

Hasil *Cluster* 1	Hasil *Cluster* 2	Hasil *Cluster* 3
Cluster 1 terdiri dari 70 orang, yang berasal dari jurusan IT = 19 orang MKT = 15 orang VCD = 12 orang HTM = 9 orang EE = 6 orang BA = 4 orang IR = 2 orang MGT = 1 orang IS = 1 orang HRM = 1 orang	*Cluster* 2 terdiri dari 132 orang, yang berasal dari aktifis ACC = 39 orang IB = 30 orang BF = 22 orang PR = 21 orang IE = 20 orang	*Cluster* 3 terdiri dari 41 orang, yang berasal dari jurusan: PR = 14 orang ACC = 7 orang IB = 7 orang BF = 6 orang E-3 = 3 orang MKT = 3 orang IT = 1 orang
Dan berasal dari Wilayah: DKI Jakarta = 30 orang Jawa Barat = 20 orang Sumatera Utara = 12 orang Sulawesi = 2 orang Jawa Timur = 2 orang Sumatera Selatan = 2 orang Bali = 1 orang Kalimantan = 1 orang	Dan berasal dari Wilayah: Jawa Barat = 62 orang DKI Jakarta = 54 orang Sumatera Utara = 16 orang	Dan berasal dari Wilayah: Sulawesi = 12 orang. Jawa Timur = 11 orang Sumatera Selatan = 11 orang Bali = 7 orang
Dengan rata-rata nilai IPK 3.2	Dengan rata-rata nilai IPK 3.25	Dengan rata-rata nilai IPK 3.31

Kemudian, dari hasil *cluster* 2 di atas dapat dilihat bahwa karakteristik mahasiswa pada *cluster* 2 didominasi oleh mahasiswa yang berasal dari jurusan *Accounting* dan *International Business*. Sedangkan, berdasarkan kota asal didominasi oleh mahasiswa yang berasal dari wilayah kota asal DKI Jakarta dan Jawa Barat, sehingga dapat disimpulkan bahwa rata-rata mahasiswa pada *cluster* 2 yang berasal dari wilayah kota asal DKI Jakarta dan Jawa Barat mengambil jurusan *Infromation Technology* dan *Marketing*.

Sedangkan, dari hasil *cluster* 3 di atas dapat dilihat bahwa karakteristik mahasiswa pada *cluster* 3 didominasi oleh mahasiswa yang berasal dari jurusan *Public Relation*, *Accounting* dan *International Business*. Sedangkan, berdasarkan kota asal didominasi oleh mahasiswa yang berasal dari wilayah kota asal Sulawesi, Jawa Timur dan Sumatera Selatan, sehingga dapat disimpulkan bahwa rata-rata mahasiswa pada *cluster* 3 yang berasal dari wilayah kota asal Sulawesi, Jawa Timur dan Sumatera Selatan mengambil jurusan *Public Relation*, *Accounting* dan *International Business*.

Penutup

K-means clustering merupakan salah satu metode *cluster analysis* non hirarki yang berusaha untuk mempartisi objek yang ada kedalam satu atau lebih *cluster* atau kelompok objek berdasarkan karakteristiknya, sehingga objek yang mempunyai karakteristik yang sama dikelompokan dalam satu *cluster* yang sama dan objek yang mempunyai karakteristik yang berbeda dikelompokan kedalam *cluster* yang lain. Cluster yang dihasilkan dapat memberikan pengetahuan baru dan menarik, yang dapat digunakan dalam mendukung keputusan.

Referensi

[1.] Baradwaj, B. K. and Pal, S. (2011). "Mining Educational Data to Analyze Student's Performance." *International Journal of Advanced Computer Science and Applications*. 2. 64.

[2.] Begum, S. H. (2013). "Data Mining Tools and Trends - An Overview." *International Journal of Emerging Research in Management & Technology*. 2278-9359. 6.

[3.] Daniel Riano Kaparang dan Eko Sediyono. (2013). "Penentuan Alih Fungsi Lahan Marginal Menjadi Lahan Pangan Berbasis Algoritma K-means di Wilayah Kabupaten Boyolali." *JdC*. 2. 20.

[4.] Deka Dwinavinta Candra Nugraha, Zumrotun Naimah, Makhfuzi Fahmi dan Novi Setiani. (2014). "Klasterisasi Judul Buku dengan Menggunakan Metode K-Means." *Seminar Nasional Aplikasi Teknologi Informasi*. ISSN: 1907-5022. G-2.

[5.] Ediyanto, Muhlasah Novitasari Mara dan Neva Satyahadewi. (2013). "Pengklasifikasian Karakteristik dengan Metode K-means Cluster Analysis." *Buletin Ilmiah Mat. Stat. dan Terapannya (Bilmaster)*. 2. 134.

[6.] Johan Oscar Ong. (2013). "Implementasi Algoritma K-means Clustering untuk Menentukan Strategi Marketing President University." *Jurnal Ilmiah Teknik Industri*. 12. 13-20.

[7.] Mujib Ridwan, Hadi Suyono dan M. Sarosa. (2013). "Penerapan Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier." *Jurnal EECCIS*. 7. 60-61.

[8.] Oyelade, O. J., Oladipupo, O. O. and Obagbuwa, I. C. (2010). "Application of K-Means Clustering algorithm for Prediction of Student's Academic Performance." *International Journal of Computer Science and Information Scurity*. 7. 292.

[9.] Seddawy, A. B. E., Khedr, A. and Sultan, T. (2012). "Adapted Framework for Data Mining Technique to Improve Decision Support System in an Uncertain Situation." *International Journal of Data Mining & Kenowledge Management Process*. 2. 5.

[10.] Sudirman dan Nur Ani. (2012). "Implementasi Teknik Data Mining Dengan Algoritma K-means Clustering dan Fungsi Kernel Polynominal untuk Klasterisasi Objek Data." *Prosiding Seminar Nasional Efisiensi Energi untuk Peningkatan Daya Saing Industri Manufaktur & Otomotif Nasional*. B - 50.

[11.] Yedla, M., Pathakota, S. R. and Srinivasa, T. M. (2010). "Enhancing K-means Clustering Algorithm with Improved Initial Center." *International Journal of Computer Science and Information Technologies*. 1. 121.

Biografi Penulis

Agus Nur Khomarudin. Lahir di Ngawi-Jawa Timur, 02 Agustus 1990, Menyelesaikan S1 di Prodi Pendidikan Teknik Informatika dan Komputer STAIN Bukittinggi pada April 2013, dan menyelesaikan S2 di Universitas Putra Indonesia YPTK Padang pada Oktober 2014. Sekarang menjadi Dosen Tetap sejak Agustus 2016 pada Prodi Pendidikan Teknik Informatika dan Komputer Institut Agama Islam Negeri Bukittinggi, Sumatera Barat, Indonesia.