MODUL DATA MINING

KLASTERING
PERTEMUAN 3 (ONLINE)

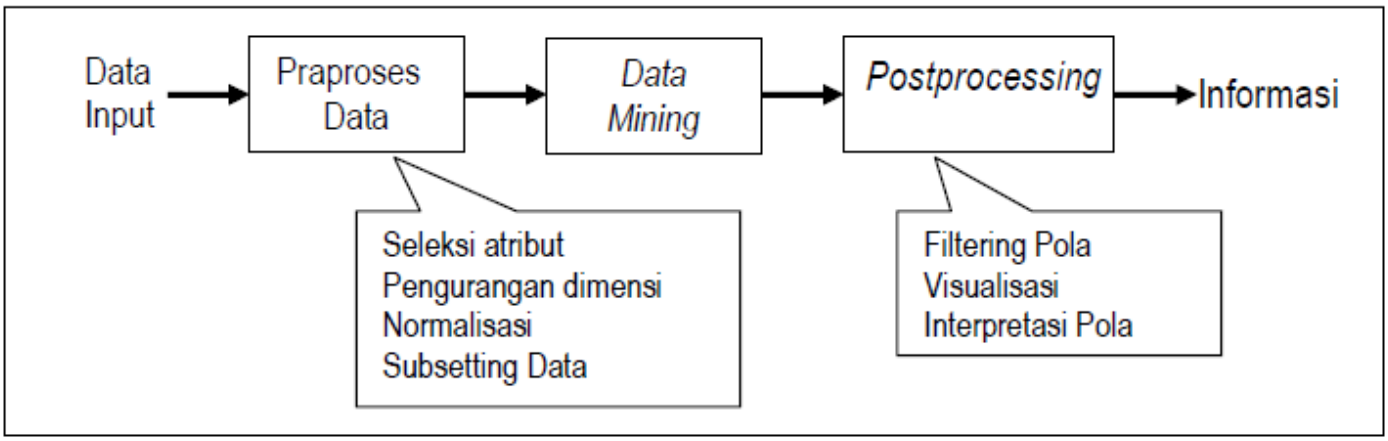
Disusun Oleh
Syefira Salsabila

Perkembangan teknologi telah membawa dampak yang sangat besar di berbagai sisi kehidupan manusia. Pengolahan data yang merupakan aset bagi perusahaan ataupun organisasi sudah menjadi kegiatan yang sangat penting untuk menunjang aktivitas dan kemajuan bagi suatu perusahaan atau organisasi. Data Mining merupakan salah satu perkembangan teknologi yang sangat berguna untuk membantu perusahaan atau organisasi dalam mengolah data dan menggali informasi yang sangat di butuhkan untuk pengembangan perusahaan atau organisasi.

Data mining membantu analisis untuk memperkirakan tren dan sifat-sifat perilaku bisnis yang sangat berguna untuk mendukung pengambilan keputusan penting dalam menunjang aktivitas dan pengembangan perusahaan. Analisis dengan data mining dilakukan dengan otomatisasi sehingga dapat mengurangi penggunaan waktu dan biaya yang tinggi. Data Mining mengeksplorasi basis data untuk menemukan pola-pola yang tersembunyi, mencari informasi untuk memprediksi yang mungkin saja terlupakan oleh para pelaku bisnis karena terletak di luar ekspektasi mereka.Perkembangan *data mining* (DM) yang pesat tidak dapat lepas dari perkembangan teknologi informasi yang memungkinkan data dalam jumlah besar terakumulasi.

Data mining muncul setelah banyak dari pemilik data baik perorangan maupun organisasi mengalami penumpukan data yang telah terkumpul selama beberapa tahun, misalnya data pembelian, data penjualan, data nasabah, data transaksi, email dan sebagainya. Kemudian muncul pertanyaan dari pemilik data tersebut, apa yang harus dilakukan terhadap tumpukan data tersebut. Misalnya, perangkat lunak *data mining* bisa membantu perusahaan ritel untuk menemukan pelanggan yang memiliki ketertarikan tertentu. Istilah ini umumnya dipersempit artinya yaitu hanya untuk menggambarkan perangkat lunak yang merepresentasikan data dengan cara-cara yang baru. Namun sebenarnya perangkat lunak data mining tidak hanya berfungsi mengubah presentasi tersebut, melainkan juga menemukan relasi tak dikenal antar-data.

Data mining adalah eksplorasi dan analisis secara otomatis atau semi otomatis terhadap data besar dengan tujuan untuk menemukan pola baru dan bermakna yang mungkin masih belum diketahui. *Data mining* merupakan bagian integral dari *Knowledge Discovery in Databases* (KDD). Keseluruhan proses KDD, mulai dari data masukan sampai menjadi informasi ditunjukkan oleh Gambar 1.



Gambar 1. Proses *Knowledge Discovery in Databases*

Sebelum melakukan proses data mining, baiknya mengetahui terlebih dahulu apa yang bisa dilakukan oleh *data mining*, agar apa yang dilakukan nantinya memang sesuai dengan apa yang dibutuhkan serta menghasilkan sesuatu yang sebelumnya tidak diketahui dan bersifat baru serta bermanfaat bagi penggunanya sendiri. Pada dasarnya data mining mempunyai kegunaan serta tugas untuk mengspesifikasikan pola yang harus ditemukan dalam proses data mining. Secara umum tugas data mining dapat dibagi menjadi dua kategori yaitu :

a. Prediktif

Tujuan dari tugas prediktif adalah untuk memprediksi nilai dari atribut tertentu berdasarkan pada nilai dari atribut-atribut lainnya. Atribut yang diprediksi umumnya dikenal sebagai target atau variable tak bebas, sedangkan atribut-atribut yang digunakan untuk membuat prediksi dikenal sebagai variabel bebas.

b. Deskriptif

Tujuan dari tugas deskriptif adalah menurunkan pola-pola (korelasi. Trend, cluster, trayektori, dan anomali) yang meringkas hubungan yang pokok dalam data. Tugas data mining deskriptif sering disebut sebagai penyelidikan dan seringkali memerlukan teknik *postprocessing* untuk validasi dan penjelasan hasil.

Berdasarkan tugas data mining secara umum, maka data mining mempunyai tugas-tugas yang berkaitan dengan data mining itu sendiri yaitu sebagai berikut:

a. Model prediksi

Model Prediksi berkaitan dengan pembuatan model yang dapat melakukan pemetaan dari setiap himpunan variabel ke setiap targetnya, kemudian menggunakan model tersebut untuk memberikan nilai target pada himpunan baru yang didapat. Ada dua jenis model prediksi, yaitu klasifikasi dan regresi. Klasfikasi digunakan untuk memprediksi nilai dari variabel target diskret, sedangkan regresi untuk memprediksi nilai dari target variabel target kontinu. *Predictive modelling* digunakan untuk membangun sebuah model untuk target variable sebagai fungsi

dari *explanatory variable. Explanatory variable* dalam hal ini merupakan semua atribut yang digunakan untuk melakukan prediksi, sedangkan variabel target merupakan atribut yang akan diprediksi nilainya.

b. Analisis kelompok
   Analisis kelompok melakukan pengelompokkan data-data ke dalam sejumlah kelompok (cluster) berdasarkan kesamaan karakteristik masing-masing data pada kelompok-kelompok yang ada. Data-data yang masuk dalam batas kesamaan dengan kelompoknya akan bergabung dalam kelompok tersebut, dan akan terpisah dalam kelompok yang berbeda jika keluar dari batas kesamaan dengan kelompok tersebut. Tidak seperti klasifikasi yang menganalisa kelas data obyek yang mengandung label. *Clustering* menganalisa objek data tanpa memeriksa kelas label yang diketahui. Label-label kelas dilibatkan di dalam data training. Karena belum diketahui sebelumnya. Clustering merupakan proses pengelompokkan sekumpulan objek yang sangat mirip.

c. Analisis asosiasi
   Analisis asosiasi (*association analysis*) digunakan untuk menemukan pola yang menggambarkan kekuatan hubungan fitur dalam data. Pola yang ditemukan biasanya mempresentasikan bentuk aturan implikasi atau subset fitur. Tujuannya adalah untuk menemukan pola yang menarik dengan cara yang efesien. *Association analysis* digunakan untuk menemukan aturan asosiasi yang memperlihatkan kondisi-kondisi nilai atribut yang sering muncul secara bersamaan dalam sebuah himpunan data.

d. Deteksi Anomali
   Deteksi anomali (*anomaly detection*) berkaitan dengan pengamatan sebuah data dari sejumlah data yang secara signifikan mempunyai karakteristik yang berbeda dari sisa data yang lain. Data-data yang karakteristiknya menyimpang (berbeda) dari data yang lain disebut *outlier.* Algoritma deteksi anomali yang baik harus mempunyai laju deteksi yang tinggi dan laju error yang rendah. Deteksi anomali dapat diterapkan pada sisterm jaringan untuk mengetahui pola data yang memasuki pola data yang memasuki jaringan sehingga penyusupan bisa ditemukan jika pola kerja data yang datang berbeda.
   Anomaly detection merupakan metode pendeteksian suatu data dimana tujuannya adalah menemukan objek yang berbeda dari sebagian besar objek lain. *Anomaly* dapat di deteksi dengan menggunakan uji statistik yang menerapkan model distribusi atau probabilitas untuk data.

Salah satu metode yang diterapkan dalam KDD adalah *clustering*. *Clustering* adalah membagi data ke dalam grup-grup yang mempunyai obyek yang karakteristiknya sama. Garcia (2002) menyatakan *clustering* adalah mengelompokkan item data ke dalam sejumlah kecil grup sedemikian sehingga masing-masing grup mempunyai sesuatu persamaan yang esensial.

C*lustering* memegang peranan penting dalam aplikasi data mining, misalnya eksplorasi data ilmu pengetahuan, pengaksesan informasi dan text mining, aplikasi basis data spasial, dan analisis web. *Clustering* diterapkan dalam mesin pencari di Internet. Web mesin pencari akan mencari ratusan dokumen yang cocok dengan kata kunci yang dimasukkan. Dokumen-dokumen tersebut dikelompokkan dalam *cluster-cluster* sesuai dengan kata-kata yang digunakan.

Klasterisasi adalah proses membagi data yang tidak berlabel menjadi kelompok-kelompok data yang memiliki kemiripan. Misalkan *K* adalah jumlah klaster, *C* merupakan label klaster, dan P merupakan dataset. Klasterisasi harus memenuhi kriteria sebagai berikut:

$$C_i \neq \Phi, \forall i \in \{1,2,...,K\} \quad (1)$$

$$C_i \cap C_j = \Phi, \forall i \neq jandi, j \in \{1,2,...,K\} \quad (2)$$

$$\bigcup_{i=1}^{K} Ci = P \quad (3)$$

*Kategori clustering m*embagi *clustering* dalam dua kelompok, yaitu *hierarchical and partitional clustering*. *Partitional Clustering* disebutkan sebagai pembagian obyek-obyek data ke dalam kelompok yang tidak saling *overlap* sehingga setiap data berada tepat di satu *cluster*. *Hierarchical clustering* adalah sekelompok cluster yang bersarang seperti sebuah pohon berjenjang (hirarki).

KLASTERING

Data mining juga merupakan metode yang digunakan dalam pengolahan data berskala besar oleh karena itu data mining memiliki peranan yang sangat penting dalam beberapa bidang kehidupan diantaranya yaitu bidang industri, bidang keuangan, cuaca, ilmu dan teknologi. Dalam data mining juga terdapat metode – metode yang dapat digunakan seperti klasifikasi, clustering, regresi, seleksi variabel, dan market basket analisis. Data mining juga bisa diartikan sebagai rangkaian kegiatan untuk menemukan pola yang menarik dari data dalam jumlah besar, kemudian data – data tersebut dapat disimpan dalam database, data warehouse atau penyimpanan informasi. Ada beberapa ilmu yang mendukung teknik data mining diantaranya adalah data analisis, *signal processing*, *neural network* dan pengenalan pola.

C*lustering* merupakan pengelompokkan record, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. *Cluster*

adalah kumpulan *record* yang memiliki kemiripan satu sama lainnya dan memiliki ketidakmiripan dengan *record-record* dalam *cluster* lain. Pengclusteran berbeda klasifikasi yaitu tidak adanya variabel target dalam pengclusteran. Pengclusteran tidak mencoba untuk melakukan klasifikasi, mengestimasi, atau memprediksi nilai dari suatu variabel target. Akan tetapi, algoritma pengclusteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan (Homogen), yang mana kemiripan record dalam suatu kelompok akan bernilai maksimal, sedangkan kemiripan dengan record dalam kelompok lain akan bernilai minimal.

Beberapa contoh pengclusteran dalam dunia bisnis dan penelitian adalah sebagai berikut :

a. *Clustering* untuk mendapatkan kelompok-kelompok konsumen untuk target pemasaran dari suatu produk bagi perusahaan yang tidak memiliki dana pemasaran yang besar.
b. *Clustering* untuk tujuan audit akutansi, yaitu melakukan pemisahan terhadap perilaku finansial yang baik maupun yang termasuk mencurigakan.
c. *Clustering* terhadap ekspresi dari gen, untuk mendapatkan kemiripan perilaku dari gen dalam jumlah besar

Clustering adalah metode yang digunakan dalam data mining yang cara kerjanya mencari dan mengelompokkan data yang mempunyai kemiripan karakteristik antara data satu dengan data lainnya yang telah diperoleh. Ciri khas dari teknik data mining ini adalah mempunyai sifat tanpa arahan (*unsupervised*), yang dimaksud adalah teknik ini diterapkankan tanpa perlunya data *training* dan tanpa ada *teacher* serta tidak memerlukan target *output. Custering* digunakan untuk menganalisis data untuk memecahkan permasalahan dalam pengelompokkan data atau lebih tepatnya mempartisi dari dataset ke dalam subset. Pada teknik *clustering* targetnya adalah untuk kasus pendistribusian (objek, orang, peristiwa dan lainnya) ke dalam suatu kelompok, hingga derajat tingkat keterhubungan antar anggota *cluster* yang sama adalah kuat dan lemah antara angota *cluster* yang berbeda

Metode *clustering* merupakan metode pengelompokan data, observasi, atau kasus menjadi kelas objek-objek yang serupa. Sedangkan *cluster* didefinisikan sebagai kumpulan data yang sama satu sama lain, dan tidak sama dengan data di lain *cluster.* Metode *clustering* mencari segmen keseluruhan data menjadi subgrup-subgrup yang relatif homogen atau biasa disebut sebagai *cluster*. Teknik *cluster* mempunyai dua metode dalam pengelompokkannya yaitu *hierarchical clustering* dan *non-hierarchical clustering*. *Hierarchical clustering* merupakan suatu metode pengelompokkan data yang cara kerjanya dengan mengelompokkan dua data atau lebih yang mempunyai kesamaan atau kemiripan, kemudian proses dilanjutkan ke objek lain yang mimiliki kedekatan dua, proses ini terus berlangsung hingga *cluster* membentuk semacam *tree* dimana ada hirarki atau tingkatan yang jelas antar objek dari yang paling mirip hingga yang paling tidak mirip. Namun secara logika semua objek pada akhirnya hanya akan membentuk sebuah *cluster*.

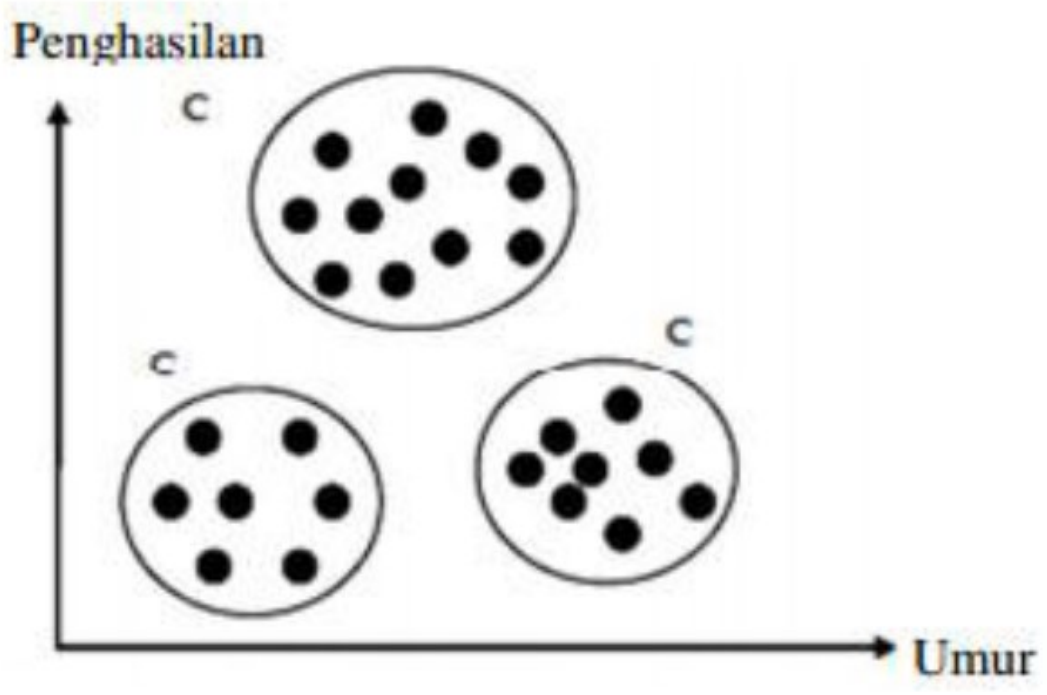
Sedangkan *non-hierarchical clustering* pada teknik ini dimulai dengan menentukan jumlah *cluster* yang diinginkan (dua *cluster,* tiga *cluster,* empat *cluster* atau lebih), setelah jumlah yang *cluster* yang diingikan maka proses *cluster* dimulai tanpa mengikuti proses hierarki. Algoritme *clustering* menempatkan data-data yang sama pada satu kelompok *(cluster),* sedangkan data yang tidak sama pada kelompok *(cluster)* yang lain. Contoh metode clustering ini adalah algoritma *K-means,* algoritma *Agglomerative Hierarchical Clustering*, *Divisive Hierarchical Clustering*.

Metode K-Means bertujuan untuk membuat *cluster* objek berdasarkan atribut menjadi *k* partisi. Cara kerja metode ini adalah mula – mula ditentukan *cluster* yang akan dibentuk, pada elemen pertama dalam tiap *cluster* dapat dipilih untuk dijadikan sebagai titik tengah (*centroid*), selanjutnya akan dilakukan pengulangan langkah – langkah hingga tidak ada objek yang dapat dipindahkan lagi. Salah satu penerapan metode k-means untuk menghasilkan informasi mengenai pengelompokkan penyakit “AKUT” dan “TIDAK AKUT” yang banyak diderita oleh pasien. Yang kemudian hasil tersebut dapat dijadikan bahan atau dasar penyuluhan kesehatan oleh Dinas Kesehatan setempat.

Sudah dijelaskan sebelumnya bahwa proses analisis *cluster* metode yang digunakan untuk membagi data menjadi subset data berdasarkan kesamaan atau kemiripan yang telah ditentukan sebelumnya. Jadi analisis *cluster* secara umum dapat dikatakan bahwa:

a. Data yang terdapat dalam satu *cluster* memiliki tingkat kesamaan yang tinggi, dan
b. Yang terdapat dalam suatu *cluster* yang berbeda memiliki tingkat kesamaan yang rendah

Sebagai contoh dapat dilihat pada gambar 2 dibawah ini :



Gambar 1. Grafik Clustering

Pada gambar 2 dapat dilihat kita misalkan data tersebut merupakan data konsumen sederhana yang terdapat dua atribut didalamnya, yaitu umur dan penghasilan. Pada data yang berdasarkan dua atribtu tersebut kemudian dibagi menjadi tiga *cluster* yaitu *cluster* C1 yang terdiri dari konsumen usia muda dan berpenghasilan rendah, *cluster* C2 terdiri dari konsumen usia muda dan tua berpenghasilan tinggi, dan *cluster* C3 terdiri dari konsumen usia tua dan berpenghasilan relatif rendah.

K-Means

Algoritma *K-Means* pertama kali diperkenalkan oleh J. MacQueen pada tahun 1967, salah satu algoritma *clustering* sangat umum yang mengelompokkan data sesuai dengan karakteristik atau ciri-ciri bersama yang serupa. Grup data ini dinamakan sebagai *cluster*. Data di dalam suatu *cluster* mempunyai ciri-ciri (karakteristik, atribut, properti) serupa.

Pada algoritma *clustering*, *data* akan dikelompokkan menjadi *cluster-cluster* berdasarkan kemiripan satu *data* dengan yang lain. Prinsip dari *clustering* adalah memaksimalkan kesamaan antar anggota satu *cluster* dan meminimumkan kesamaan antar anggota *cluster* yang berbeda. Kategori algoritma *clustering* yang banyak dikenal adalah *Hierarchical Clustering*. *Hierarchical Clustering* adalah salah satu algoritma *clustering* yang dapat digunakan untuk meng-*cluster* dokumen *(document clustering*). Dari teknik *hierarchical clustering*, dapat dihasilkan suatu kumpulan partisi yang berurutan, dimana dalam kumpulan tersebut terdapat:

a. *Cluster – cluster* yang mempunyai poin – poin individu. *Cluster – cluster* ini berada di level yang paling bawah.
b. Sebuah *cluster* yang didalamnya terdapat poin – poin yang dipunyai semua *cluster* didalamnya. *Single cluster* ini berada di *level* yang paling atas.

K-Means merupakan suatu algoritma yang digunakan dalam pengelompokkan secara pertisi yang memisahkan data ke dalam kelompok yang berbeda – berda. Algoritma ini mampu meminimalkan jarak antara data ke *cluster*nya. Pada dasarnya penggunaan algoritma ini dalam proses *clustering* tergantung pada data yang didapatkan dan konklusi yang ingin dicapai di akhir proses. Sehingga dalam penggunaan algoritma k-means terdapat aturan sebagai berikut:

a. Berapa jumlah *cluster* yang perlu dimasukkan
b. Hanya memiliki atribut bertipe numeric

Dalam statistik dan mesin pembelajaran *K-means* merupakan metode analisis kelompok yang mengarah pada permatisian N objek pengamatan dalam K kelompok (cluster) dimana setiap objek pengamatan dimiliki oleh sebuah kelompok dengan mean (rata-rata) terdekat. K-means merupakan salah satu metode pengelompokkan data nonhierarki (sekatan) yang berusaha mempartisi data yang ada ke dalam bentuk dua atau lebih kelompok. Metode ini mempartisi data kedalam sebuah kelompok sehingga data berkarakteristik sama dimasukkan kedalam satu kelompok yang sama dan data

yang berkarakteristik berbeda dikelompokkan ke dalam kelompok yang lain. Adapun tujuan pengelompokkan data ini adalah untuk meminimalkan fungsi objektif yang diset dalam proses pengelompokkan, yang pada umumnya berusaha meminimalkan variasi di dalam suatu kelompok dan memaksimalkan variasi antar kelompok.

Pada dasarnya algoritma k-means hanya mengambil sebagian dari banyaknya komponen yang didapatkan untuk kemudian dijadikan pusat *cluster* awal, pada penentuan pusat *cluster* ini dipilih secara acak dari populasi data. Kemudian algoritma k-means akan menguji masing – masing dari setiap komponen dalam populasi data tersebut dan menandai komponen tersebut ke dalam salah satu pusat *cluster* yang telah didefinisikan sebelumnya tergantung dari jarak minimum antar komponen dengan tiap – tiap pusat *cluster*. Selanjutnya posisi pusat *cluster* akan dihitung kembali sampai semua komponen data digolongkan ke dalam tiap – tiap *cluster* dan terakhir akan terbentuk *cluster* baru.

Algoritma K-Means pada dasarnya melakukan 2 proses yakni proses pendeteksian lokasi pusat cluster dan proses pencarian anggota dari tiap-tiap cluster. Proses clustering dimulai dengan mengidentifikasi data yang akan dikluster, Xij (i=1,...,n; j=1,...,m) dengan n adalah jumlah data yang akan dikluster danm adalah jumlah variabel. Pada awal iterasi, pusat setiap kluster ditetapkan secara bebas (sembarang),Ckj (k=1,...,k; j=1,...,m). Kemudian dihitung jarak antara setiap data dengan setiap pusat cluster digunakan formula Euclidean. Suatu data akan menjadi anggota dari cluster ke-k apabila jarak data tersebut ke pusat cluster ke-k bernilai paling kecil jika dibandingkan dengan jarak ke pusat cluster lain. Proses dasar algoritma k-means dapat dilihat di bawah ini :

a. Tentukan jumlah klaster yang ingin dibentuk dan tetapkan pusat cluster *k*.
b. Menggunakan jarak *euclidean* kemudian hitung setiap data ke pusat cluster.

$$d_{ij} = \sqrt{\sum_{k=1}^{n} \left\{ x_{ik} - x_{jk} \right\}^2} \qquad \ldots. (2.1)$$

Keterangan:

d_{ij} = jarak antara data ke-i dan data ke-j

n = dimensi data

x_{ik} = koordinat data ke-i pada dimensi k

x_{jk} = koordinat data ke-j pada dimensi k

c. Kelompokkan data ke dalam cluster dengan jarak yang paling pendek dengan persamaan

$$Min \sum_k^k = d_{ik} = \sqrt{\sum_j^m (C_{ij} - C_{kj})^2} \quad (2.2)$$

d. Hitung pusat cluster yang baru menggunakan persamaan

$$C_{kj} = \frac{\sum_{i=1}^{p} x_{ij}}{p} \quad (2.3)$$

Dengan :

x_{ij} € Kluster ke – k

p = banyaknya anggota kluster ke - k

e. Ulangi langkah dua sampai dengan empat sehingga sudah tidak ada lagi data yang berpindah ke kluster yang lain.

Implementasi Algoritma K-Means

K-means clustering merupakan metode *clustering* non-hirarki. Data-data yang memiliki karakteristik yang sama dikelompokkan dalam satu *cluster*/kelompok dan data yang memiliki karakteristik yang berbeda dikelompokkan dengan *cluster*/kelompok yang lain, sehingga data yang berada dalam satu *cluster*/kelompok memiliki tingkat variasi kecil. *K-means* adalah algoritma *clustering* yang mempartisi himpunan D menjadi *k cluster* data. Algoritma *K-means* mengklaster semua titik data pada D sedemikian sehingga titik data *xi* menjadi satu-satunya *k* partisi. Dengan kata lain, satu titik data hanya masuk ke dalam satu *cluster*.

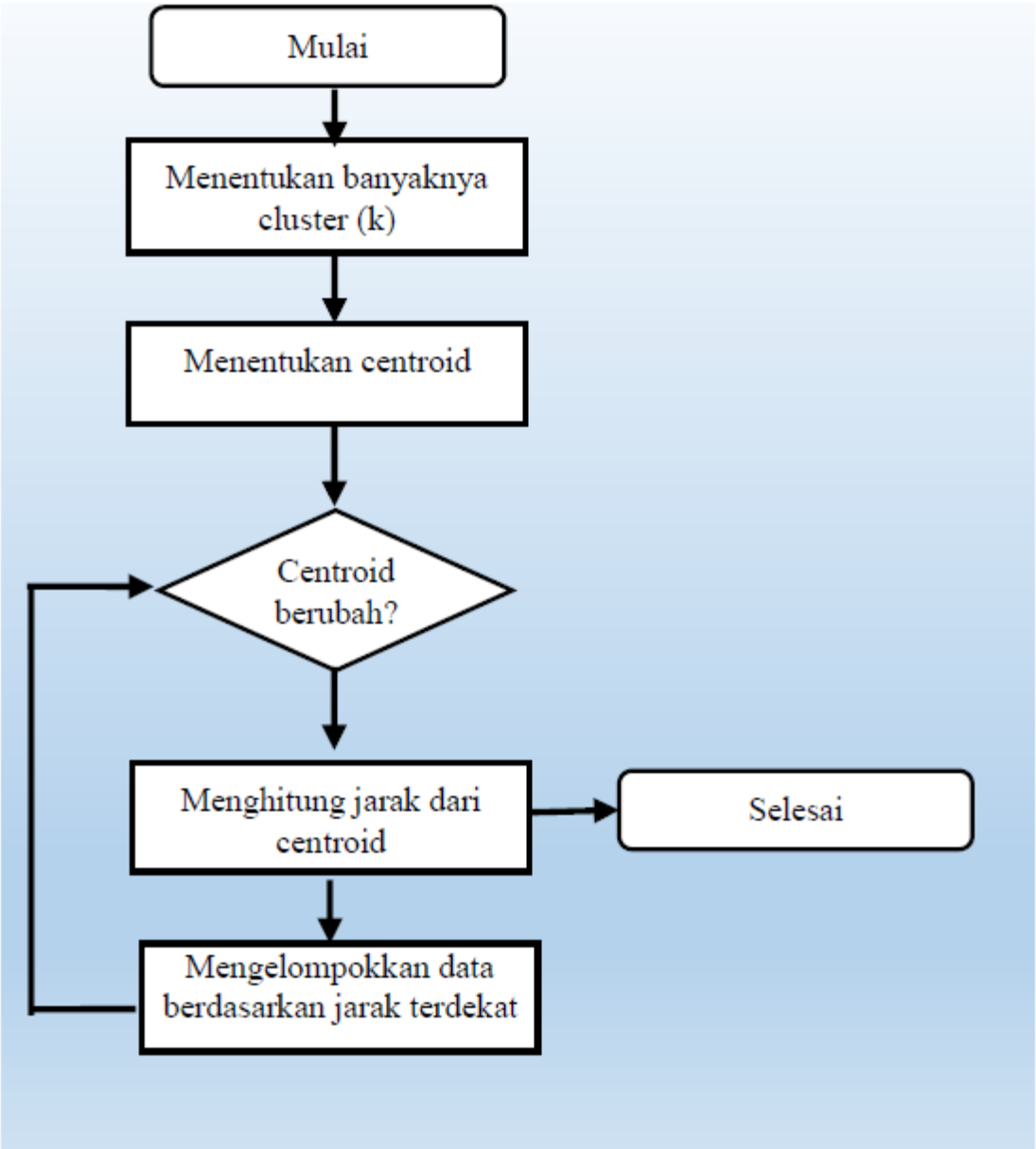
Adapun langkah-langkah dari algoritma *K-means* adalah sebagai berikut:

a. Tentukan *K* data sebagai *centroid*, *K* adalah jumlah *cluster* yang diinginkan (ditentukan oleh peneliti).
b. Tiap titik (data) kemudian dicari *centroid* terdekatnya.
c. Setiap himpunan titik (data) yang menjadi *centroid* disebut *cluster.*
d. Hitung kembali *centroid* dari setiap *cluster.*
e. Ulangi langkah 1-4 sampai *centroid* tidak berubah.

Contoh Kasus

Dapat dilihat pada gambar 3 di bawah merupakan diagram alur dari metode k-means yang digunakan dalam pengelompokkan penyakit di Puskesmas Kajen Pekalongan, pada umumnya kinerja metode k-means secara berurutan adalah sebagai berikut :

a. Menentukan banyaknya cluster (k)
b. Menentukan centroid
c. Apakah nilai centroidnya berubah? jika ya, hitung jarak data dari centroid. Jika tidak, selesai.
d. Mengelompokkan data berdasarkan jarak terdekat



Gambar 3. Alur Implementasi Algoritma K-Means

Data penelitian yang sedang dilakukan merupakan data penyakit pasien Puskesmas Kajen Pekalongan sebanyak 1000 data yang akan dikelompokkan ke dalam penyakit “AKUT(C1)” dan penyakit “TIDAK AKUT(C2)” pengelompokkan tersebut berdasarkan atribut umur, kode penyakit dan lama mengidap penyakit, yang kemudian atribut tersebut akan diolah menggunakan algoritma k-means.

Penerapan *data mining* dengan teknik *clustering* dan algoritma *k-means* yang dilakukan menghasilkan sebuah informasi mengenai Jenis penyakit yang sering diderita pasien pengguna askin berdasarkan hubungan antara diagnosa penyakit dengan jumlah

pasien pengguna askin, dimana dari informasi tersebut didapat jumlah diagnosa yang besar yaitu pada jenis penyakit Degeneratif sebesar dengan jumlah diagnosa 3602 dengan angka tertinggi pada bulan desember sebesar 386 jumlah diagnosa.

Algoritma *k-means* membagi data ke dalam *k*-buah *cluster* yang telah ditentukan. Ada beberapa cara yang dapat digunakan untuk menghitung jarak yaitu antara lain dengan menggunakan *Euclidean distance, Manhattan distance, dan Chebisev distance.* Masing-masing cara perhitungan jarak tersebut dijelaskan sebagai berikut ini.

a. *Euclidean Distance*
   Formula untuk menghitung jarak antar dengan *Euclidean Distance* untuk dua titik dalam satu, dua dan tiga dimensi secara berurutan ditunjukan pada persamaan (2.2), (2.3), dan (2.4) sebagai berikut.

$$\sqrt{(x-y)^2} = |x-y| \quad \ldots. (2.2)$$

$$d(p,q) = \sqrt{(p_1-q_1)^2 + (p_2-q_2)^2} \ldots(2.3)$$

$$d(p,q) = \sqrt{\begin{matrix}(p_1-q_1)^2 + (p_2-q_2)^2 \\ +(p_3-q_3)^2\end{matrix}} \ldots(2.4)$$

b. *Manhattan Distance (Taxicab distance)*

$$d1(p,q) = \left\| p-q \right\| 1 = \sum_{i=1}^{n} \left| p_i - q_i \right| \ldots(2.5)$$

c. *Chebisev Distance (Maximum Metric)*
   Untuk menentukan jarak dengan menggunakan *Chebisev Distance* dilakukan dengan cara mengambil nilai maksimum dari setiap koordinat dimensinya. Jika dinyatakan dalam persamaan matematika, maka *Chebisev Distance* dapat dilihat pada persamaan (2.6)

$$D_{cheb}(p,q) = \max\left(\left|p_i - q_i\right|\right) \ldots . (2.6)$$

Algoritma K-Means merupakan algoritme klasterisasi yang mengelompokkan data berdasarkan titik pusat klaster (*centroid*) terdekat dengan data. Tujuan dari K-Means adalah pengelompokkan data dengan memaksimalkan kemiripan data dalam satu klaster dan meminimalkan kemiripan data antar klaster. Ukuran kemiripan yang digunakan dalam klaster adalah fungsi jarak. Sehingga pemaksimalan kemiripan data didapatkan berdasarkan jarak terpendek antara data terhadap titik centroid.

Tahapan awal yang dilakukan pada proses klasterisasi data dengan menggunakan algoritma K-Means adalah pembentukan titik awal *centroid* cj Pada umumnya pembentukan titik awal *centroid* dibangkitkan secara acak. Jumlah *centroid* cj yang dibangkitkan sesuai dengan jumlah klaster yang ditentukan di awal. Setelah k *centroid* terbentuk kemudian dihitung jarak tiap data xi dengan *centroid* ke-j sampai k, dinotasikan dengan d (xi,cj). Terdapat beberapa ukuran jarak yang digunakan sebagai ukuran kemiripan suatu *instance* data, salah satunya adalah jarak *Euclid*.

Agglomerative Hierarchical Clustering

Metode ini menggunakan strategi disain *Bottom-Up* yang dimulai dengan meletakkan setiap obyek sebagai sebuah *cluster* tersendiri (*atomic cluster*) dan selanjutnya menggabungkan *atomic cluster – atomic cluster* tersebut menjadi *cluster* yang lebih besar dan lebih besar lagi sampai akhirnya semua obyek menyatu dalam sebuah *cluster* atau proses dapat pula berhenti jika telah mencapai batasan kondisi tertentu. Metode *Agglomerative Hierarchical Clustering* yang digunakan pada penelitian ini adalah metode *AGglomerative NESting* (AGNES). Adapun ukuran jarak yang digunakan untuk menggabungkan dua buah obyek cluster adalah *Minimum Distance.*

Langkah dalam Algoritma Agglomerative Hierarchical Clustering:

a. Hitung Matrik Jarak antar data
b. Ulangi langkah 3 dan 4 higga hanya satu kelompok yang tersisa
c. Gabungkan dua kelompok terdekat berdasarkan metode pengelompokan (*Single Linkage, Complete Linkage, Average Linkage)*
d. Perbarui Matrik Jarak antar data untuk merepresentasikan kedekatan diantara kelompok baru dan kelompok yang masih tersisa.

e. Selesai

Metode Pengelompokan Hierarki Aglomeratif

Beberapa metode pengelompokan secara hierarki Aglomeratif:

a. Single Linkage (Jarak Terdekat)

$$d_{uv} = \min\{d_{uv}\}, d_{uv} \in D$$

b. Complete Linkage (Jarak Terjauh)

$$d_{uv} = \max\{d_{uv}\}, d_{uv} \in D$$

c. Average Linkage (Jarak rata-rata)

$$d_{uv} = average\{d_{uv}\}, d_{uv} \in D$$