WeeK 9
K-Means Clustering

Agus Tri Widodo

1. Clustering

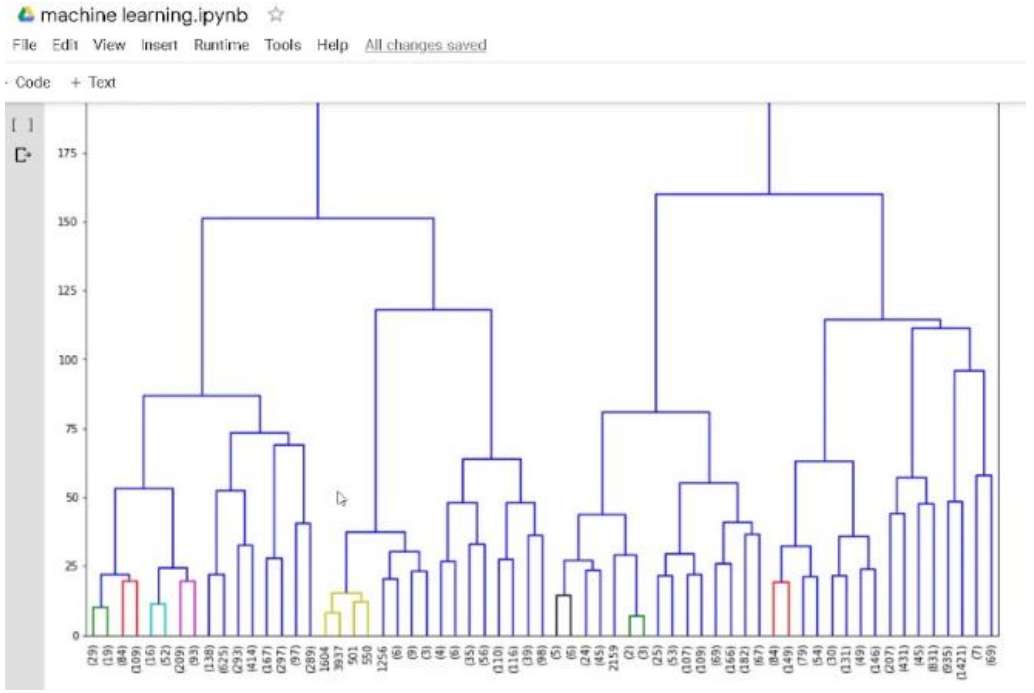
Mengelompokkan item data ke dalam sejumlah kecil grup sedemikian sehingga masing-masing grup mempunyai sesuatu persamaan yang esensial.

Pengklusteran merupakan pengelompokan record, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan.

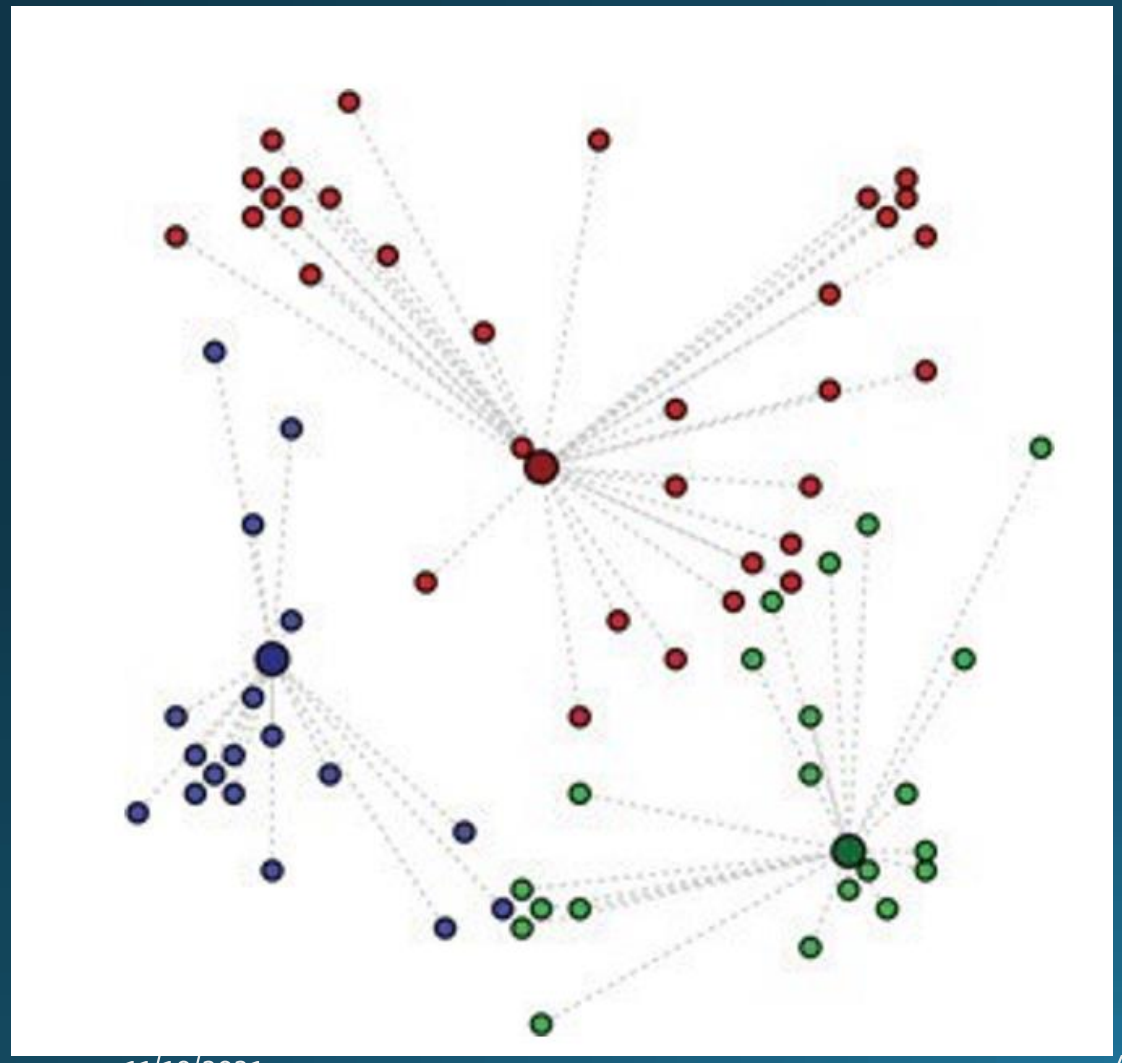
Kluster adalah kumpulan record yang memiliki kemiripan suatu dengan yang lainnya dan memiliki ketidakmiripan dengan record dalam kluster lain.

Clustering

- Proses partisi satu set objek data ke dalam himpunan bagian yang disebut dengan cluster.
- Data dalam satu cluster memiliki tingkat kemiripan yang maksimum dan data antar cluster memiliki kemiripan yang minimum.
- Partisi tidak dilakukan secara manual melainkan dengan suatu algoritma clustering.
- Digunakan dalam berbagai aplikasi seperti misalnya pada business inteligence, perbankan, pengenalan pola citra, web search, bidang ilmu biologi, dan untuk keamanan (security).



Clustering - Algoritma



- Langkah melakukan Hierarchical clustering:
  a. Identifikasi item dengan jarak terdekat
  b. Gabungkan item itu kedalam satu cluster
  c. Hitung jarak antar cluster
  d. Ulangi dari awal sampai semua terhubung

K-Means Clustering

- 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.

Data Mining

"apa yang dapat dilakukan dari tumpukan data tersebut?"

- Data Mining – Clustering

Terdapat dua jenis metode clustering yang digunakan dalam pengelompokkan data:

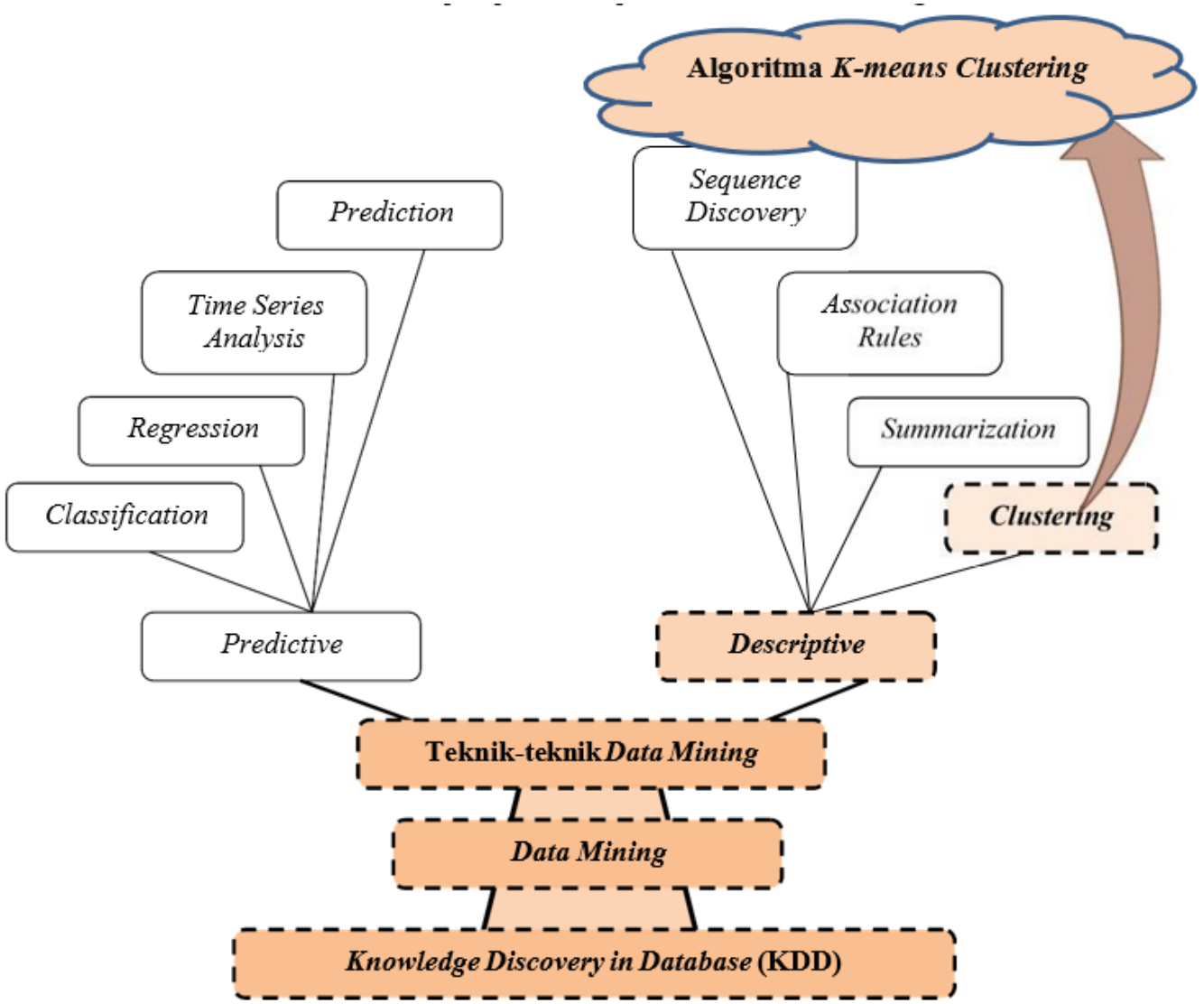
- hierarchical clustering
- non-hierarchical clustering.

K-means clustering

- 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.

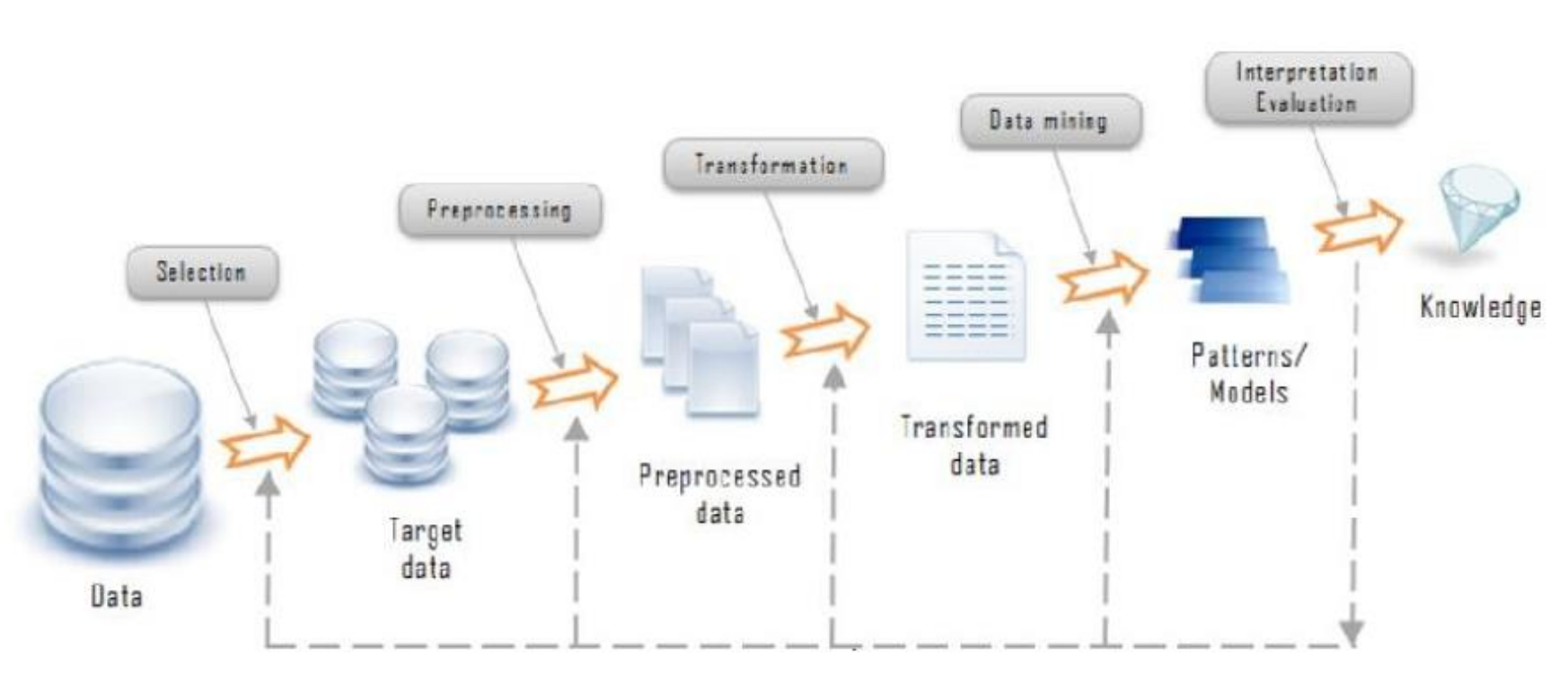
Mind Mapp

memudahkan kita dalam memahami materi yang dibahas



Knowledge Discovery in Database (KDD)

- 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 :



Proses Knowledge Discovery in Database

Dari berbagai sumber heterogen yang terintegrasi ke dalam penyimpanan data tunggal yang disebut sebagai data target

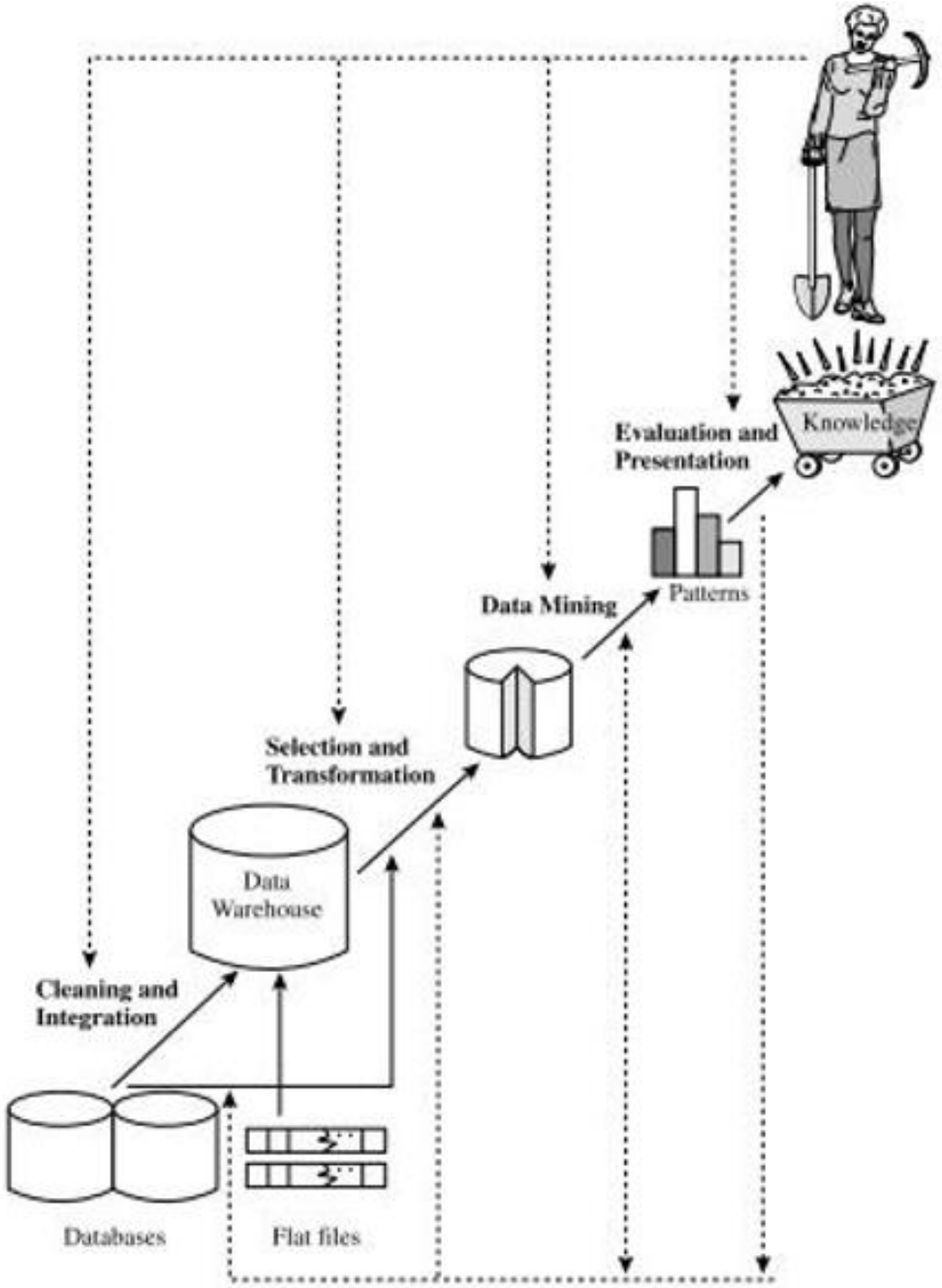
Data Mining

- 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 Data Minning

- 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.
- 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.

Tahapan dalam Data Mining

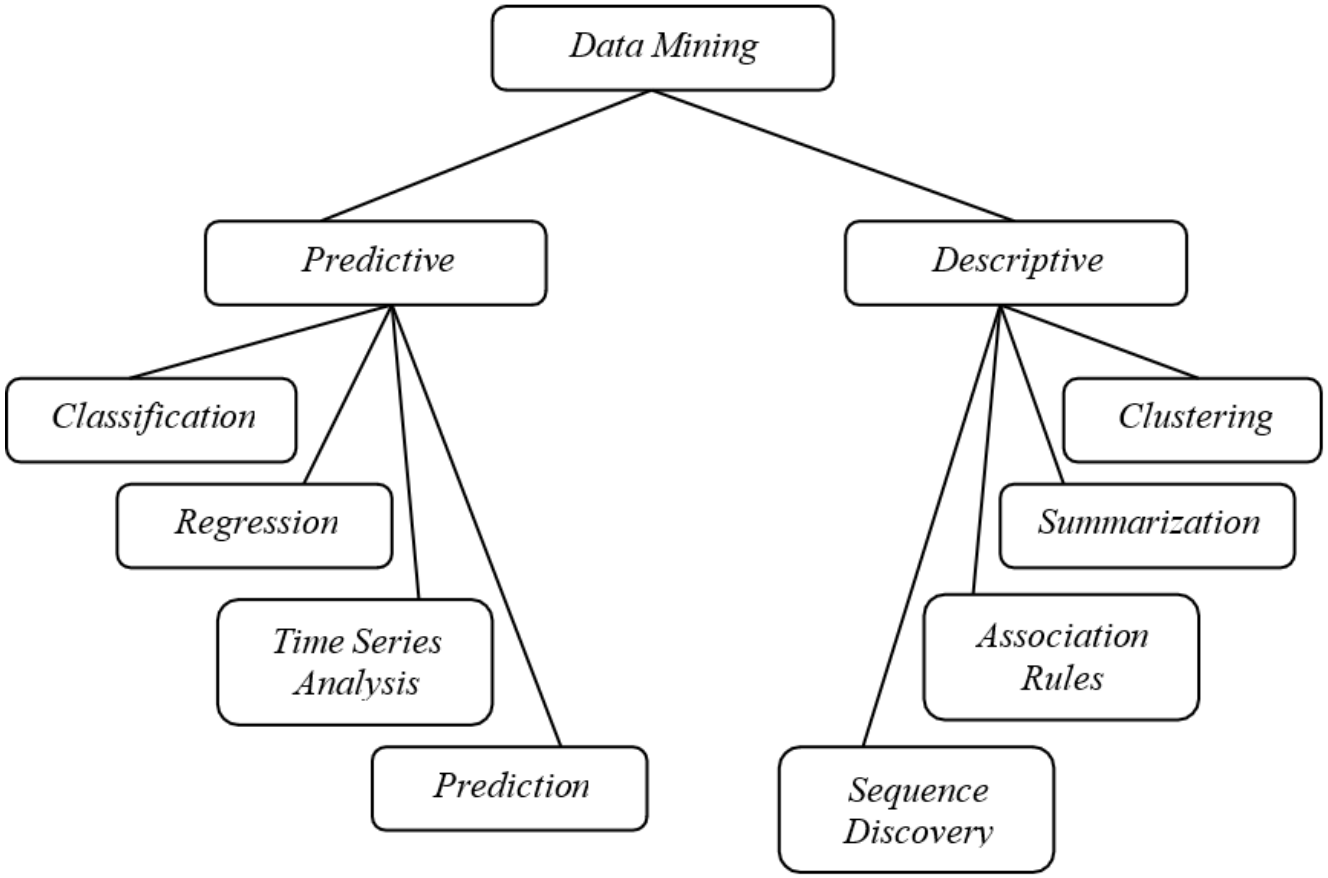


Tahap-tahap Data Mining

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)
   Mengidentifikasi pola-pola menarik ke dalam knowledge based yang ditemukan.
g. Presentasi pengetahuan (knowledge presentation)
   Visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna.

Teknik-Teknik Data mining

Data mining sebenarnya memiliki akar yang panjang dari bidang ilmu seperti kecerdasan buatan (artificial intelligent), machine learning, statistik dan basis data.



Teknik Yang Digunakan Dalam Data Mining

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.

2. Decision tree

   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).

3. 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.

4. 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.

Teknik Yang Digunakan Dalam Data Mining

5. 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.

6. 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.

7. 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).

8. 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.

Algoritma K-means Clustering

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 Xki = Data ke i pada atribut data ke k

Xkj = 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 clusterterdekat.
e. Proses penentuan pusat cluster dan penempatan data dalam cluster diulangi sampai nilai centroid tidak berubah lagi.


Thanks