BAB II

LANDASAN TEORI

2.1 Data Mining

Tan (2006) mendefinisikan data mining sebagai proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar.

Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terakit dari berbagai database besar (Turban, dkk., 2005).

Menurut Han, Kamber, dan Pei (2011:18), *data mining* adalah proses menemukan pola yang menarik dan pengetahuan dari data dalam jumlah besar.

Dari beberapa pernyataan tersebut, dapat disimpulkan bahwa *data mining* merupakan proses ekstraksi informasi dari database yang berukuran besar untuk mendapatkan pengetahuan yang tersimpan dari data tersebut.

Istilah data mining kadang disebut juga *Knowledge Discovery in Database* (KDD). Istilah *data mining* sering dipakai, mungkin karena istilah ini lebih pendek dari *Knowledge Discovery in Database*. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda, tetapi berkaitan satu sama lain. Data mining dianggap hanya sebagai suatu langkah penting dalam KDD. Proses KDD secara garis besar dapat dijelaskan sebagai berikut (Han, Kamber, dan Pei, 2011) :

1. Pembersihan data, untuk menghilangkan noise dan data yang tidak konsisten.
2. Integrasi data, di mana beberapa sumber data dapat dikombinasikan. Sebuah tren populer di industri informasi adalah untuk melakukan pembersihan dan integrasi data sebagai langkah preprocessing, dimana data yang dihasilkan akan disimpan dalam *data warehouse*.
3. Seleksi data, di mana data yang relevan dengan tugas analisis yang diambil dari database.

4. Data transformasi (dimana data diubah dan digabung ke dalam bentuk yang sesuai untuk pertambangan dengan melakukan ringkasan atau agregasi operasi) Terkadang transformasi data dilakukan sebelum proses seleksi data, khususnya dalam kasus *data warehouse*.
5. Data mining, merupakan proses esensial dimana metode cerdas diaplikasikan untuk mengekstrak data pola.
6. Evaluasi Pola, untuk mengidentifikasi pola yang benar-benar menarik yang mewakili pengetahuan.
7. Presentasi pengetahuan, dimana visualisasi dan teknik representasi pengetahuan digunakan untuk menyajikan pengetahuan hasil *data mining* kepada pengguna.

Langkah 1 sampai 4 merupakan berbagai bentuk preprocessing data, dimana data dipersiapkan untuk *data mining*. Hal ini, menunjukkan bahwa data mining sebagai salah satu langkah dalam proses KDD, karena dapat mengungkap pola-pola tersembunyi yang digunakan untuk evaluasi.

2.1.1 Fungsi *Data Mining*

Data mining dibagi menjadi dua kategori utama (Han dan Kamber, 2006 : 21- 29) yaitu:

A. Prediktif

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

B. Deskriptif

Tujuan dari tugas deskriptif adalah untuk menurunkan pola-pola (korelasi, *trend, cluster,* teritori, dan anomali) yang meringkas hubungan yang pokok dalam data. Tugas *data mining* deskriptif sering merupakan

penyelidikan dan seringkali memerlukan teknik *post-processing* untuk validasi dan penjelasan hasil.

Data mining juga memiliki beberapa fungsionalitas yaitu *Concept/Class Description*: *Characterization and Discrimination, Mining Frequent Patterns, Associations, and Correlations, Classification and Prediction, Cluster Analysis, Outlier analysis,* dan *Evolution analysis*. (Han dan Kamber, 2006 : 21 – 27)

Berikut adalah penjelasan dari masing-masing fungsi diatas:

1. *Concept/Class Description*: *Characterization and Discrimination*

   Data characterization adalah ringkasan dari semua karakteristik atau fitur dari data yang telah diperoleh dari target kelas. Data yang sesuai dengan kelas yang telah ditentukan oleh pengguna biasanya dikumpulkan di dalam *database*. Misalnya, untuk mempelajari karakteristik produk perangkat lunak dimana pada tahun lalu seluruh penjualan telah meningkat sebesar 10%, data yang terkait dengan produk-produk tersebut dapat dikumpulkan dengan menjalankan sebuah *query SQL*. Sedangkan, *data discrimination* adalah perbandingan antara fitur umum objek data target kelas dengan fitur umum objek dari satu atau satu set kelas lainnya. target diambil melalui *query database*. Misalnya, pengguna mungkin ingin membandingkan fitur umum dari produk perangkat lunak yang pada tahun lalu penjualannya meningkat sebesar 10% tetapi selama periode yang sama seluruh penjualan juga menurun setidaknya 30%.

2. *Mining Frequent Patterns, Associations, and Correlations*

   Frequent Patterns adalah pola yang sering terjadi di dalam data. Ada banyak jenis dari *frequent patterns*, termasuk di dalamnya pola, sekelompok *item set*, *sub-sequence*, dan sub-struktur. Sebuah *frequent patterns* biasanya mengacu pada satu set item yang sering muncul bersama-sama dalam suatu kumpulan data transaksional, misalnya seperti susu dan roti.

   Associations Analysis adalah pencarian aturan-aturan asosiasi yang menunjukan kondisi-kondisi nilai atribut yang sering terjadi bersama-sama

dalam sekumpulan data. Analisis asosiasi sering digunakan untuk menganalisa *Market Basket Analysis* dan data transaksi.

3. *Classification and Prediction*

Klasifikasi adalah proses untuk menemukan model atau fungsi yang menggambarkan dan membedakan kelas data atau konsep dengan tujuan memprediksikan kelas untuk data yang tidak diketahui kelasnya. Model yang diturunkan didasarkan pada analisis dari training data (yaitu objek data yang memiliki label kelas yang diketahui). Model yang diturunkan dapat direpresentasikan dalam berbagai bentuk seperti *If-then* klasifikasi, *decision tree*, naïve bayes, dan sebagainya.

Teknik *classification* bekerja dengan mengelompokkan data berdasarkan *data training* dan nilai atribut klasifikasi. Aturan pengelompokan tersebut akan digunakan untuk klasifikasi data baru ke dalam kelompok yang ada.

Dalam banyak kasus, pengguna ingin memprediksikan nilai-nilai data yang tidak tersedia atau hilang (bukan label dari kelas). Dalam kasus ini nilai data yang akan diprediksi merupakan data *numeric*. Disamping itu, prediksi lebih menekankan pada identifikasi *trend* dari distribusi berdasarkan data yang tersedia.

4. *Cluster Analysis*

Cluster adalah kumpulan objek data yang mirip satu sama lain dalam kelompok yang sama dan berbeda dengan objek data di kelompok lain. Sedangkan, *Clustering* atau Analisis *Custer* adalah proses pengelompokkan satu set benda-benda fisik atau abstrak kedalam kelas objek yang sama. Tujuannya adalah untuk menghasilkan pengelompokan objek yang mirip satu sama lain dalam kelompok-kelompok. Semakin besar kemiripan objek dalam suatu *cluster* dan semakin besar perbedaan tiap *cluster* maka kualitas analisis *cluster* semakin baik.

Dari tugas – tugas data mining yang telah di jelaskan , perbandingan antara *Classification* dan *Clustering* menurut Han dan Kamber (2006) lebih spesifik digambarkan sebagai berikut :

Tabel 2.1 Perbedaan Klasifikasi dan Klustering

Classification	**Clustering**
1. Menganalisis label kelas dari data objek.	1. menganalisis data objek tanpa ada label kelas.
2. Label kelas ada atau terlihat jelas pada training data.	2.label kelas tidak ada atau tidak terlihat pada training data.
3. Bertujuan untuk mengelompokan pada kelas – kelas yang telah ditentukan.	3.bertujuan untuk mengelompokan dan menentukan label kelas dari tiap cluster yang telah terbentuk
4. Proses klasifikasi berdasarkan pada menemukan sebuah model atau fungsi yang menggambarkan dan membedakan data kelas atau konsep, dengan tujuan untuk dapat menggunakan model untuk memprediksi objek kelas yang kelas label nya blm diketahui. Model tersebut berdasarkan pada analisis dari ***training data*** (data objek yang kelas label nya telah diketahui.)	4. Proses Clustering berdasarkan pada prinsip: objek yang ada di dalam satu cluster memiliki kemiripan yang tinggi dari pada yang lainnya, tetapi sangat berbeda dengan objek yang ada pada cluster lainnya.

5. *Outlier analysis*

   Outlier merupakan objek data yang tidak mengikuti perilaku umum dari data. *Outlier* dianggap sebagai noise atau pengecualian. Analisis *data outlier* dapat dianggap sebagai *noise* atau pengecualian. Analisis *data outlier* dinamakan *Outlier Mining*. Teknik ini berguna dalam *fraud detection* dan *rare events analysis*.

6. *Evolution analysis*

Analisis evolusi data menjelaskan dan memodelkan *trend* dari objek yang memiliki perilaku yang berubah setiap waktu. Teknik ini dapat meliputi karakterisasi, diskriminasi, asosiasi, klasifikasi, atau *clustering* dari data yang berkaitan dengan waktu.

Dari buku Data Mining Technique yang dikarang oleh Berry and Linoff, proses terjadinya data mining dapat dideskripsikan sebagai virtous cycle. Didasari oleh pengembangan berkelanjutan dari proses bisnis serta didorong oleh penemuan knowledge ditindaklanjuti dengan pengambilan tindakan dari penemuan.

2.1.2 Langkah-langkah Data mining

1. ***Identity The Business Problem***

   Yang pertama dan juga dasar dari virtous cycle adalah mengetahui masalah bisnis yang kita hadapi. Karena kita tidak bisa mengolah data jika kita tidak tau yang sedang kita hadapi. Kita harus mengetahui masalah-masalah apa yang sedang dihadapi. Dengan mengetahui masalah yang dihadapi kita dapat menentukan data-data mana saja yang kita butuhkan untuk dapat dilakukan tahap analisa.

2. ***Mine The Data For Actionable Information***

   Setelah mengetajui identifikasi masalah, kita memperolah data-data mana saja yang diperlukan untuk analisa. Barulah kita melakukan analisa terhadap data-data tersebut. Dan dari analisa tersebut analisis akan dapat memperolah

sebuah knowledge baru dan barulah dapat diambil suatu keputusan atau kebijaksanaan.

3. ***Take The Action***

   Dan dari keputusan/kebijaksanaan yang didapat dari proses data mining itu barulah kita terapkan dengan aksi berupa tindakan-tindakan yang kongkrit/nyata dalam proses bisnis.

4. ***Measure Results***

   Setelah diambil tindakan-tindakan dan keputusan, kita memonitori hasil tersebut. Apakah sudah sesuai(memuaskan) dengan target2 yang ingin kita capai, apakah bisa mengatasi masalah-masalah yang dihadapi.

2.1.3 Teknik-teknik/Jenis-jenis DataMining

1. ***Market Basket Analysis***

   Himpunan data yang dijadikan sebagai objek penelitan pada area data mining. Market basket analysis adalah proses untuk menganalisis kebiasaan pelanggan dalam menyimpan item-item yang akan dibeli ke dalam keranjang belanjaannya. Market basket analysis memanfaatkan data transaksi penjualan untuk dianalisis sehingga dapat ditemukan pola berupa item-item yang cenderung muncul bersama dalam sebuah transaksi. Selanjutnya pola yang ditemukan dapat dimanfaatkan untuk merancang strategi penjualan atau pemasaran yang efektif, yaitu dengan menempatkan item-item yang sering dibeli bersamaan ke dalam sebuah area yang berdekatan, merancang tampilan item-item di katalog, merancang kupon diskon (untuk diberikan kepada pelanggan yang membeli item tertentu), merancang penjualan item-item dalam bentuk paket, dan sebagainya. Dengan menggunakan teknologi data mining, analisis data secara manual tidak diperlukan lagi.

2. ***Memory-Based Reasoning***

   Metode klasifikasi yang digabungkan dengan penalaran berbasis memori. proses menggunakan satu set data untuk membuat model dari prediksi atau asumsi-asumsi yang dapat dibuat tentang objek baru yang diperkenalkan. Ada dua komponen dasar untuk metode MBR. Yang pertama adalah kesamaan

fungsi, yang mengukur bagaimana anggota yang sama dari setiap pasangan object satu sama lain. Yang kedua adalah fungsi kombinasi, yang digunakan untuk menggabungkan hasil dari himpunan tetangga untuk sampai pada keputusan.

3. ***Cluster Detection***

   Ada dua pendekatan untuk clustering. Pendekatan pertama adalah dengan mengasumsikan bahwa sejumlah cluster sudah tersimpan dalam data, tujuannya adalah untuk memecah data ke dalam cluster. Pendekatan lain, disebut clustering agglomerative, dengan asumsi keberadaan setiap jumlah yang telah ditetapkan cluster tertentu, setiap item keluar di cluster sendiri, dan proses terjadi berulang-ulang yang berupaya untuk menggabungkan cluster, meskipun proses komputasi sama.

4. ***Link Analysis***

   Proses mencari dan membangun hubungan antara object dalam kumpulan data juga mencirikan sifat yang terkait dengan hubungan antara dua object. Link Analysis berguna untuk aplikasi analitis yang mengandalkan teori grafik untuk mengambil kesimpulan. Selain itu Link Analysis berguna untuk proses optimasi.

5. ***Rule Induction***

   Ekstraksi aturan sebab-akibat dari data secara statistic. identifikasi aturan bisnis yang tersimpan di dalam data. Metode berhubungan dengan induksi aturan yang digunakan untuk proses penemuan. Salah satu pendekatan untuk penemuan aturan adalah menggunakan pohon keputusan.

6. ***Neural Networks***

   model prediksi non linear yang melakukan pembelajaran melalui latihan dan menyerupai struktur jaringan nerual yang terdapat pada mahluk hidup. Mampu menurunkan pengertian dari data yang kompleks dan tidak jelas dan dapat digunakan pula untuk mengekstrak pola dan mendeteksi tren2 yang sangat kompleks untuk dibicarakan baik oleh manusia maupun teknik komputer lainnya.

2.1.4 Tugas Data Mining (Six Tax Data Mining)

Classification

- Menyusun data menjadi kelompok-kelompok yang telah ditentukan, yang melibatkan dengan memeriksa atribut-atribut dari suatu objek tertentu dan menetapkannya ke kelas yang telah didefinisikan.

Estimation

- proses untuk menempatkan beberapa nilai numerik secara terus suatu objek, estimasi juga dapat digunakan sebagai bagian dari proses klasifikasi.

Prediction

- berbeda dengan Estimation dan Classification, Prediction adalah upaya-upaya untuk mengklasifikasikan suatau objek berdasarkan dari behaviour yang akan ditentukan(diharapkan) dari candidate behavior.

Affinity Grouping

- proses yang mengevaluasi hubungan atau asosiasi antara unsur-unsur data berupa attribute atau behaviour data yang menunjukkan beberapa tingkat afinitas antar objek.

Clustering

- sama seperti klasifikasi tetapi kelompok yang tidak/belum di tentukan standarnya, sehingga secara algoritma data tersebut akan dikelompokan berdasarkan data yang serupa dengan data yang di submit.

Desciption - proses yang menggambarkan apa yang telah terjadi dan di identifikasi atau proses yang menjelaskan hasil akhir dari jalannya proses data mining.

2.2 Clustering

Klasterisasi merupakan alat utama dalam berbagai aplikasi dalam analisa data statistik, data mining, information retrieval, pengolahan citra dan sebagainya. Klasterisasi bertujuan melakukan pengelompokan obyek/data ke dalam beberapa klaster/kelompok sehingga obyek/data dalam satu klaster memiliki tingkat kesamaan yang maksimum dan data antar klaster memiliki kesamaan yang minimum.(Wikipedia).

C*lustering* (pengelompokan) melakukan pemisahan /pemecahan /segmentasi data kedalam sejumlah *cluster* (kelompok) menurut karakteristik tertentu yang diinginkan. Dalam pekerjaan clustering label dari setiap data belum diketahui, Diharapkan nantinya dapat diketahui kelompok data untuk kemudian diberikan label sesuai keinginan. Bidang penerapan teknik clustering: kedokteran, pendidikan, kesehatan, psikologi, hukum, statistik, astronomi, klimatologi dan sebagainya.

- Pendidikan, teknik clustering dalam pendidikan bisa diambil contoh dalam pengelompokan mahasiswa berdasarkan kategori yang dikehendaki.
- Kedokteran, teknik clustering dapat digunakan untuk mengelompokkan jenis-jenis penyakit berbahaya berdasarkan karakteristik / sifat-sifat penyakit pasien.
- Kesehatan, dapat digunakan untuk pengelompokan pasien berdasarkan ukuran tubuh / gain mass.

Klasterisasi data kategorikal sudah mulai berkembang, walaupun perkembanganya masih jauh lebih sedikit dibanding klasterisasi pada tipe data numerik. Data kategorikal secara alami tidak bisa di perlakukan sebagai data numerik karena ada beberapa operasi dalam data numerik yang tidak bisa di lakukan dalam data kategorikal seperti mean dan median. Sebagai contoh atribut data kategorikal adalah atribut berdomain jenis kelamin (pria, wanita), domain agama (Islam, Kristen, Katolik, Hindu dan sebagainya), dan domain etnis (mongoloid, kaukasoid, negroid). Klasterisasi merupakan alat utama dalam berbagai aplikasi dalam analisa data statistik, data mining, information retrieval, pengolahan citra dan sebagainya. Klasterisasi bertujuan melakukan pengelompokan obyek/data ke dalam beberapa klaster/kelompok sehingga obyek/data dalam satu klaster memiliki tingkat kesamaan yang maksimum dan data antar klaster memiliki kesamaan yang minimum (Tan, et al. 2006; Han and Kamber. 2006).

Inti dari clustering untuk partisi satu set objek dalam menguraikan dan cluster homogen, sehingga benda milik cluster yang sama lebih mirip satu sama lain dibandingkan mereka yang termasuk kelompok yang berbeda (Jain et al, 1999).

Berdasar pendekatan dalam penetapan keanggotaan dalam klaster, metode klasterisasi secara umum dapat dibagi menjadi dua yaitu *hard clustering* dan *fuzzy clustering.* Pada *Hard clustering,* tiap obyek/data hanya di alokasikan ke dalam satu satu klaster baik selama operasi klasterisasi maupun dalam output klasterisasi. Sedang pada *Fuzzy clustering,* selama operasi klasterisasi tiap obyek/data di alokasikan ke dalam beberapa klaster dan di beri derajat keanggotaan dengan nilai antara 0 dan 1. Output *Fuzzy Clustering* dapat di ubah menjadi *hard clustering* dengan memilih nilai keanggotaan tertinggi. (Jain, et al. 1999).

2.3 Clustering K – Modes

K-Modes merupakan pengembangan dari metode K-Means agar dapat di gunakan untuk klasterisasi data kategorikal. K-Modes menggunakan sebuah ukuran jarak *(dissimilarity)* berupa kecocokan suatu nilai atribut tiap dimensi terhadap titik pusat klaster, menggantikan mean dengan modus, dan menggunakan metode berbasis frekuensi untuk memutakhirkan modus dalam proses meminimalkan jarak (*dissimilarity*) dari seluruh data ke pusat klaster masing masing. Sebagai contoh atribut data kategorikal adalah atribut berdomain jenis kelamin (pria, wanita), domain agama (Islam, Kristen, Katolik, Hindu dan sebagainya), dan domain etnis (mongoloid, kaukasoid, negroid), dan disini data yang akan daya clustering dengan K-mode adalah Diagnosa dan Kecamatan, (Huang, 1998).

Perbedaan K – Modes terhadap K – Means sebagai berikut :

1. Menggunakan ukuran perbedaan sederhana cocok untuk benda kategoris,
2. Menggantikan sarana cluster dengan modes, dan
3. Menggunakan metode frekuensi berbasis untuk menemukan modes dari sekumpulan nilai.

Misalkan x, y adalah untuk objek kategoris digambarkan oleh atribut m kategoris. ukuran kesamaan antara x dan y dapat didefinisikan oleh ketidakmiripan total kategori atribut yang sesuai dari dua benda. semakin kecil jumlah ketidakmiripan adalah, dua lebih mirip objek. ukuran ini sering disebut sebagai pencocokan sederhana (kaufan dan rousseeuw 1990).

Rumus yang digunakan seperti pada persamaan di bawah ini :

$$d(x,y) = \sum_{j=1}^{r} \boldsymbol{\epsilon}\left(Xj,\ Yj\right) \quad \text{.....(2.1)}$$

Keterangan :

r = Jumlah fitur

∈(.) = Nilai pencocokan , seperti pada persamaan dibawah ini :

$$\in (x_j, y_j) = \begin{cases} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases} \quad \text{..... (2.2)}$$

Andaikan X adalah set data yang nilai fiturnya bertipe kategorikal, A_1, A_2, A_3, A_4, …, A_r. Modes dari X = { X_1, X_2, …, X_n} adalah vector Q = [q_1, q_2, …, q_r] yang meminimalkan seperti pada persamaan dibawah ini :

$$D(X,Q) = \sum_{i=1}^{n} d(X_i, Q) \quad \text{.....(2.3)}$$

Untuk persamaan diatas vektor Q merupakan vector yang bukan bagian dari X.

Notasi $n_{C_{k,j}}$ dalah jumlah obyek memiliki K kategori $\mathbf{C_{k,j}}$ di atribut $\mathbf{A_j}$ dan $f_r(A_j = c_{k,j}|X) = \frac{n_{c_{k,j}}}{n}$ frekuensi relative dari categori $\mathbf{c_{k,j}}$ di X.

Yang perlu ditekankan dalam masalah mode adalah bahwa mode dari set data X tidak bersifat unik. Misalnya, mode dari set {[a,b], [a,c], [c,b],[b,c]} bisa didapat [a,b] atau [a,c].

Ketika persamaan (1) digunakan sebagai ukuran ketidaksamaan untuk objek kategoris, fungsi biaya (1) menjadi.

$$J = \sum_{l=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{r} Wi,j \; \boldsymbol{\epsilon} \left(xi,j \; , \; qi,j\right) \quad \text{..... (2.4)}$$

Dimana untuk ∈ (.) adalah nilai pencocokan seperti pada persamaan (2) antara vektor dengan mode cluster yang diikuti , sedangkan $w_{i,I} \in W$ adalah nilai keanggotaan data dapat setiap cluster , nilainya [0, 1] , didpatkan dari persamaan berikut :

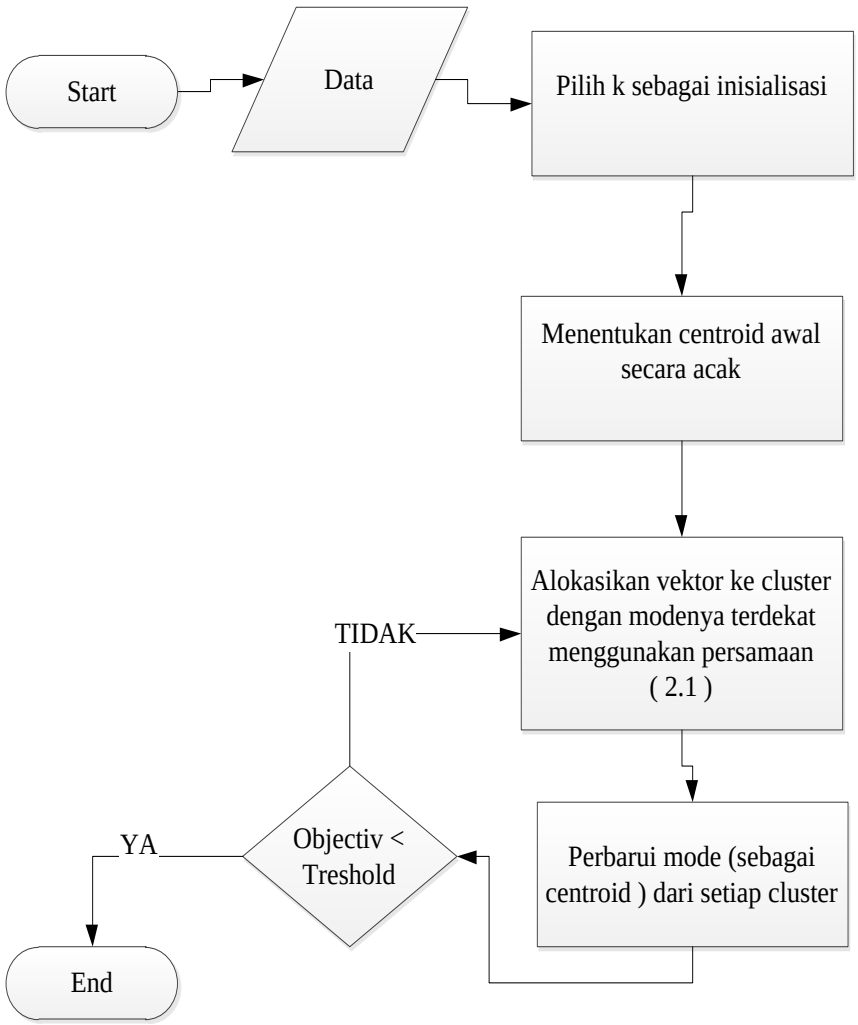
$$w_{i,l} = \begin{cases} 1 & jika\ d(X_i, Q_l) < d(X_i, Q_t),\ untuk\ l = 1,\ldots,k,\ untuk\ t = 1,\ldots,k \\ 0 & untuk\ t \neq l \end{cases} \quad \text{.....(2.5)}$$

Keterangan :

k = Jumlah cluster.

n = Jumlah data dalam setiap cluster.

2.3.1 Algoritma K – Modes



Gambar 2.1 Flowchart Algoritma K – Modes.

1. Pilih k sebagai inisialisasi centroid (modes) , satu untuk setip cluster.
2. Menentukan centroid awal secara acak.
3. Alokasikan vektor ke cluster dengan modenya terdekat menggunakan persamaan (2.1).
4. Perbarui mode (sebagai centroid) dari setiap cluster dengan nilai kategori yang sering muncul pada setiap cluster.
5. Ulangi langkah 2 dan 3 selama masih memenuhi syarat :
   a. Masih ada vector yang berpindah cluster atau
   b. Perubahan fungsi nilai objektif masih diatas nilai threshold yang ditentukan.

2.4 Penelitian Sebelumnya

K – Modes merupakan metode yang memodifikasi dari metode K – Means yang nilainya diperoleh dari modes sedangkan K- Means diperoleh dari mean. Metode K – Modes sangat cocok digunakan untuk penelitian yang berupa kategorikal bukan numeric.

Penelitian yang berjudul “Analisa Pola Diagnosa Penyakit pada Rekam Medis Elektronik dengan Metode K-Mode” oleh Irwan . Adapun data di dapatkan 14789 dataset yang akan di clustering. Untuk diagnosa akan dikelompokkan menjadi 22 kategori dari 434 kategori, yang pengelompokkan tersebut berdasrkan kode internasional Icd-10 versi 2010 from WHO. Dataset awal memiliki 4 fitur yaitu tanggal, bulan, kecamatan, diagnosa. Dari keempat fitur ini ada 2 jenis data yaitu data numeric dan data kategorikal. Untuk tanggal dan bulan bertipe data numeric yang biasanya digunakan untuk metode fuzzy C-Means sedangkan diagnosa dan kecamatan bertipe kategorikal, sehingga untuk proses cluster pada sistem ini menggunakan metode K-Mode. Pada tabel_dataset terdapat 4 field yaitu Id_dataset, tgl_kunj, Xddx, Xreg. Dari 4 field ini hanya 2 attribut yang akan dilakukan proses clustering yaitu *Xddx (Diagnosa) dan Xreg (Kecamatan).* Sedangkan untuk tgl_kunj digunakan untuk proses analisis/ penggalian informasi lebih lanjut dari hasil perhitungan clustering K-mode.Hasil penelitian menunjukkan bahwa dapat mengelompokkan data emr untuk dianalisa lebih lanjut

dari hasil perhitungan clustering, dengan nilai keakurasian hingga 80% dari nilai purity dan entropy untuk validitas hasil clusteringnya.

Penelitian selanjutnya dilakukan oleh I Putu Agus Hendra Krisnawan dalam kasusnya yang berjudul "Rancang bangun system pengelompokan pelanggan potensial menggunakan metode K- Means untuk promosi paket wisata (Studi kasus PT. Bali Sinar Mentari) "dalam perusahaan jasa, pemberian promosi merupakan salah satu faktor penting dalam membantu penjualan jasa kepada pelanggan. Permasalahan yang timbul adalah manajer mengalami kesulitan dalam melakukan pemilihan pelanggan serta dalam pengelompokan pelanggan guna mengetahui pelanggan mana saja yang tepat untuk diberikan promosi. Pengelompokan pelanggan ini dilakukan dengan melihat pola data transaksi paket wisata yang telah ada sebelumnya dengan seleksi berdasarkan hotel dan paket wisata dan selanjutnya akan dianalisa menggunakan metode pengelompokan data *K-Means.*

Peneliatan yang dilakukan oleh narwati yang berjudul " Pengelompokan Mahasiswa Menggunakan Algoritma K – Means " Penelitian ini membahas pengelompokan mahasiswa berdasarkan data akademik menggunakan teknik clustering. Data akademik tersebut adalah hasil evaluasi tes masuk penerimaan mahasiswa baru (PMB) berupa nilai tes potensial akademik (TPA) dan nilai tes bahasa inggris, data atribut identitas diri mahasiswa seperti nama, nim, asal sekolah, kota asal, usia, jenis kelamin serta nilai Indeks Prestasi Komulatif (IPK) . Dengan menggunakan data hasil tes masuk, dan pencapaian indeks prestasi komulatif pada semester 8 , maka dapat diketahui minat belajar dari mahasiswa apakah tetap pada nilai test awal masuk atau ada perubahan yang signifikan. Hasil penelitian ini menghasilkan pola dari prestasi mahasiswa yang klusternya tetap, turun dan naik. Pola mahasiswa tersebut dapat terlihat dari asal program studi, asal kota dan asal SMA.