BAB II

LANDASAN TEORI

2.1 Penelitian Sebelumnya

Penelitian yang dilakukan Munandar, Widyarto dan Harsiti (2013) adalah menentukan konsentrasi jurusan yang akan diambil oleh mahasiswa sesuai dengan nilai akademis di Universitas Serang Raya Program Studi Teknik Informatika dengan menggunakan Fuzzy C-Means. Karena instansi pendidikan yang diteliti memiliki tiga konsentrasi jurusan yakni Pemrograman, Web, dan Multimedia maka data uji akan dijadikan tiga *cluster* dan yang dijadikan variabel ada sepuluh yakni nilai mata kuliah pemrograman internet dan HTML, algoritma pemrograman, pemrograman terstruktur, komunikasi data, strukutr data, matematika diskrit, nilai IP semester 1, nilai IP semester 2, nilai IP semester 3 dan nilai IP semester 4.

Penelitian yang dilakukan oleh Okta dan Saikhu (2010) adalah melakukan proses pengelompokan objek data set ke dalam beberapa *cluster* dengan menggunakan algoritma *Particle Swarm Optimization K-Harmonic Means* (PSOKHM). Algoritma ini merupakan penggabungan dari algoritma *Particle Swarm Optimization* dan *K-Harmonic Means.* Pada penelitian ini digunakan algoritma PSOKHM untuk melakukan clustering data, serta algoritma KHM dan PSO sebagai perbandingan evaluasi hasil *cluster* berdasarkan nilai fungsi objektif, F-Measure, dan running time. Uji coba dilakukan dengan 3 skenario terhadap 5 data set yang berbeda. Berdasarkan hasil uji coba diperoleh bahwa, jika ditinjau dari nilai objective function dan F-Measure, PSOKHM mampu memberikan hasil yang lebih baik. Sedangkan bila dilihat dari running time, PSOKHM mampu mengungguli PSO namun tidak lebih baik daripada KHM.

2.2 Pengertian Data Mining

Secara sederhana *data mining* adalah penambangan atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari sejumlah data yang sangat besar (Davies, 2004). *Data Mining* juga disebut sebagai serangkaian proses untuk menggali nilai tambah berupa pengetahuan yang selama ini tidak diketahui secara manual dari suatu kumpulan data (Pramudiono, 2007).

Data mining, sering juga disebut sebagai *Knowledge Discovery In Database* (KDD). KDD adalah kegiatan yang meliputi pengumpulan, pemakaian data, historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar (Santoso, 2007).

Data mining adalah kegiatan menemukan pola yang menarik dari data dalam jumlah besar, data dapat disimpan dalam *database*, data *warehouse*, atau penyimpanan infromasi lainnya. *Data mining* berkaitan dengan bidang ilmu-ilmu lain, seperti *database system,* data *warehousing, statisik, machine learning, information retrieval,* dan komputasi tingkat tinggi. Selain itu, *data mining* didukung oleh ilmu lain seperti *neural network,* pengenalan pola, *spatial data analysis, image database, signal processing* (Han, 2006).

Data mining didefinisikan sebagai proses menemukan pola-pola dalam data. Proses ini otomatis atau seringnya semiotomatis (Witten, 2005). Pola yang ditemukan harus penuh arti dan pola tersebut memberikan keuntungan, biasanya keuntungan secara ekonomi.

Karakteristik *data mining* sebagai berikut :

1. *Data mining* berhubungan dengan penemuan sesuatu yang tersembunyi dan pola data tertentu yang tidak diketahui sebelumnya.
2. *Data mining* biasa menggunakan data yang sangat besar. Biasanya data yang besar digunakan untuk membuat hasil lebih dipercaya.
3. *Data mining* bewrguna untuk membuat keputusan yang kritis, terutama dalam strategi (Davies, 2004).

Berdasarkan beberapa pengertian tersebut dapat ditarik kesimpulan bahwa *data mining* adalah suatu teknik menggali informasi berharga yang terpendam atau tersembunyi pada suatu koleksi data (*database*) yang sangat besar sehingga

ditemukan suatu pola yang menarik yang sebelumnya tidak diketahui. *Data mining* sendiri berarti usaha untuk mendapatkan sedikit barang berharga dari sejumlah besar material dasar. Karena itu *data mining* sebenarnya memiliki akar yang panjang dari bidang ilmu seperti kecerdasan buatan (*artificial intelligent*), *machine learning*, statistik dan database. Beberapa metode yang sering disebut-sebut dalam literatur *data mining* antara lain *clustering*, *classification*, *association rules mining*, *neural network*, *genetic algorithm* dan lain-lain (Pramudiono, 2007).

2.3 Tahap – Tahap Data Mining

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

Tahap-tahap *data mining* ada 6 yaitu :

1. Pembersihan data (*data cleaning*).

   Pembersihan data merupakan proses menghilangkan *noise* dan data yang tidak konsisten atau data yang tidak relevan. Pada umumnya data diperoleh, baik dari *database* suatu perusahaan maupun hasil eksperimen, memiliki isian-isian yang tidak sempurna seperti data yang hilang, data yang tidak valid atau juga hanya sekedar salah ketik. Selain itu, ada juga atribut-atribut data yang tidak relevan dengan hipotesa *data mining* yang dimiliki. Data-data yang tidak relevan itu juga lebih baik dibuang. Pembersihan data juga akan mempengaruhi performansi dari teknik *data mining* karena data yang ditangani akan berkurang jumlah dan kompleksitasnya.

2. Integrasi Data (*data integraton*)

   Integrasi data merupakan penggabungan data dari berbagai *database* ke dalam satu *database* baru. Tidak jarang data yang diperlukan untuk *data mining* tidak hanya berasal dari satu *database* atau file teks. Integrasi data dilakukan pada atribut-atribut yang mengidentifikasikan entitas-entitas yang unik seperti atribut nama, jenis produk, nomor pelanggan dan lainnya. Integrasi data perlu dilakukan

secara cermat karena kesalahan pada integrasi data bisa menghasilkan hasil yang menyimpang dan bahkan menyesatkan pengambilan aksi nantinya. Sebagai contoh bila integrasi data berdasarkan jenis produk dari kategori yang berbeda maka akan didapatkan korelasi antar produk yang sebenarnya tidak ada.

3. 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*. Sebagai contoh, sebuah kasus yang meneliti faktor kecendrungan orang membeli dalam kasus *market analysis*, tidak perlu mengambil nama pelanggan, cukup dengan id pelanggan saja.

4. Transformasi Data (*data transformation*)

   Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam *data mining*. Beberapa metode *data mining* membutuhkan format data yang khusus sebelum bisa diaplikasikan. Sebagai contoh beberapa metode standar seperti analisis asoisasi dan *clustering* hanya bisa menerima input data kategorikal, Karenanya data berupa angka numerik yang berlanjut perlu dibagi-bagi menjadi beberapa interval. Proses ini sering disebut transformasi data.

5. Proses *Mining*

   Merupakansuatu proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data.

6. Evaluasi Pola (*pattern evaluation*).

   Untuk mengidentifikasi pola-pola menarik kedalam *knowledge based* yang ditemukan. Dalam tahap ini hasil dari teknik *data mining* berupa pola-pola yang khas maupun model prediksi dievaluasi untuk menilai apakah hipotesa yang ada tercapai. Bila ternyata hasil yang diperoleh tidak sesuai hipotesa ada beberapa alternatif yang dapat diambil seperti menjadikannya umpan balik untuk memperbaiki proses *data mining*, mencoba metode *data mining* yang lebih sesuai, atau menerima hasil ini sebagai suatu hal yang diluar dugaan yang mungkin

bermanfaat.

2.4 Konsep Pengelompokan (C*lustering Concept)*

Analisis kelompok (*cluster analysis*) adalah suatu analisis statistik juga merupakan salah satu teknik data mining yang bertujuan untuk mengidentifikasi sekelompok obyek yang mempunyai kemiripan karakteristik tertentu yang dapat dipisahkan dengan kelompok obyek lainnya, sehingga obyek yang berada dalam kelompok yang sama relatif lebih homogen daripada obyek yang berada pada kelompok yang berbeda (Tan, 2006) . Jumlah kelompok yang dapat diidentifikasi tergantung pada banyak dan variasi data obyek.

Tujuan dari pengelompokan sekumpulan data obyek ke dalam beberapa kelompok yang mempunyai karakteristik tertentu dan dapat dibedakan satu sama lainnya adalah untuk analisis dan interpretasi lebih lanjut sesuai dengan tujuan penelitian yang dilakukan.

Ada banyak metode pengelompokan yang sudah dikembangkan para ahli. Masing-masing metode mempunyai karakter, kelebihan, dan kekurangan. Farley dan Raftery (1998) menyarankan membagi metode *clustering* menjadi dua kelompok utama yakni *hierarchical* dan *partitioning*, sedangkan menurut Han and Kamber (2001) menyarankan mengelompokkan metode menjadi tambahan tiga kategori utama : *density-based methods*, *model-based clustering* dan *grid-based methods*.

1. Pengelompokan Hierarki (*Hierarchical Clustering*)

   Dalam metode pengelompokan hirarki suatu data dapat memiliki cluster lebih dari satu. Pada dasarnya teknik hierarki ini menjelaskan bahwa satu data tunggal bisa dianggap sebuah kelompok (*cluster*), dua atau lebih kelompok kecil bisa bergabung menjadi sebuah kelompok besar dan seterusnya hingga semua data menjadi sebuah kelompok.

   Dalam metode ini terdapat dua tipe dasar yaitu *agglomerative* (pemusatan) dan *divisive* (penyebaran). Dalam metode *agglomerative*, setiap obyek atau observasi dianggap sebagai sebuah cluster tersendiri. Dalam tahap selanjutnya, dua cluster yang mempunyai kemiripan

digabungkan menjadi sebuah cluster baru demikian seterusnya. Sebaliknya, dalam metode *divisive* kita beranjak dari sebuah cluster besar yang terdiri dari semua obyek atau observasi. Selanjutnya, obyek atau observasi yang paling tinggi nilai ketidakmiripannya kita pisahkan demikian seterusnya. Dalam *agglomerative* ada lima metode yang cukup terkenal, yaitu : s*ingle linkage, complete linkage*, a*verage linkage*, w*ard's method*, c*entroid method*.

2. Pengelompokan Partisi (*Partitioning Clustering*)

   Metode partisi ini dimana harus menentukan jumlah k partisi yang diinginkan lalu setiap data dites untuk dimasukkan pada salah satu partisi sehingga tidak ada data yang *overlap* dan satu data hanya memiliki satu cluster. Contohnya: algoritma *K-Means*.

   Berbeda dengan Hierarki yang membagi data sesuai jarak dan menerapkan pada dendogram, metode partisi ini mencoba memindah suatu data dari *cluster* yang satu ke *cluster* yang lainnya secara rekursif dari partisi awal hingga mencapai *cluster yang* optimal secara global.

3. Pengelompokan Berbasis Kepadatan (*Density-based Clustering*)

   Dalam pengelompokan berbasis kepadatan, *cluster* didefinisikan sebagai daerah dengan kepadatan lebih tinggi dari sisa kumpulan data. Jadi metode ini menerapkan bahwa suatu data termasuk suatu kelompok/*cluster* jika memiliki anggota yang banyak dan jika ada jika data yang kepadatannya kecil dianggap sebagai *noise* atau pemisah antar *cluster*. Algoritma yang terkait dengan metode ini contohnya *DBSCAN*.

4. Pengelompokan Berbasis Model (*Model-based Clustering*)

   Metode pengelompokan berbasis model (*model-based clustering*) berusaha untuk mengoptimalkan kecocokan antara data yang diberikan dengan beberapa model matematika. Metode tersebut sering didasarkan pada asumsi bahwa data yang dihasilkan oleh campuran distribusi

probabilitas yang mendasarinya. Algoritma yang terkait adalah *expectation maximization, conceptual clustering,* dan *neural network.*

5. Pengelompokan Berbasis Grid (*Grid-based Clustering*)

   Metode pengelompokan berbasis grid (*grid-based clustering*) menggunakan multiresolusi grid strukur data untuk proses pengelompokan/*clustering*. Proses pengelompokan dilakukan dengan cara menempatkan sejumlah objek atau data ke dalam ruang atau sel dengan jumlah terbatas yang membentuk sebuah grid dalam satu kelompok. Keuntungan menggunakan metode ini adalah waktu proses yang sangat cepat, jika dengan menggunakan metode yang lain tergantung dengan banyaknya jumlah data maka jika menggunakan metode ini tergantung pada hanya jumlah sel pada setiap dimensi yang terbatas. Algoritma yang terkait adalah *sting (statistical information grid), wavecluster,* dan *clique.*

2.5 Fitur Data Yang Diolah

Terdapat beberapa metode *data mining* yang dapat digunakan untuk mengotomatiskan proses pengelompokan (*clustering)* terhadap berbagai kasus. Adapun dalam skripsi ini, model klaster yang digunakan adalah model klaster dengan menggunakan metode *K-Harmonic Means*(KHM). Sistem pengklasteran / *database* mahasiswa dikembangkan dengan memanfaatkan model klastering KHM dengan 5 buah atribut / fitur yang berupa variable biner (ya/tidak).

Adapun lima atribut (kriteria) yang dipakai dalam proses klastering tersebut penyusun peroleh dengan kuesioner diantaranya :

1. Kompetensi dasar *Database*.

   Kompetensi dasar *Database* diambil dari kuesioner yang meliputi :

   a. Mahasiswa paham arti *database*
   b. Mahasiswa mengetahui perbedaan *Flat File Databases, Relational Databases,* dan *Distributed Databases.*
   c. Mahasiswa memahami pengertian dan penggunaan normalisasi

tabel

d. Mahasiswa bisa melakukan normalisasi pertama (1NF).
e. Mahasiswa bisa melakukan normalisasi kedua (2NF).
f. Mahasiswa bisa melakukan normalisasi ketiga (3NF).
g. Mahasiswa bisa melakukan normalisasi *Boyce & Codd*.
h. Mahasiswa bisa melakukan normalisasi keempat (4NF).
i. Mahasiswa bisa melakukan normalisasi kelima (5NF).

2. Kompetensi *Database* SQL DDL.

   Kompetensi *Database* SQL DDL diambil dari kuesioner yang meliputi :

   a. Mahasiswa bisa membuat database menggunakan perintah SQL: CREATE DATABASE.
   b. Mahasiswa bisa membuat table menggunakan perintah SQL: CREATE TABLE.
   c. Mahasiswa bisa membuat memasukkan data kedalam tabel menggunakan perintah SQL: INSERT INTO.
   d. Mahasiswa bisa mengganti nilai data pada sebuah kolom dengan perintah SQL: UPDATE …. SET.
   e. Mahasiswa bisa menghapus baris table menggunakan perintah SQL: DELETE FROM.
   f. Mahasiswa bisa membuat PRIMARY KEY pada salah satu kolom dalam table dengan perintah ALTER TABLE … ADD PRIMARY KEY.
   g. Mahasiswa bisa menghapus PRIMARY KEY pada table menggunakan perintah ALTER TABLE … DROP PRIMARY KEY.
   h. Mahasiswa bisa membuat referential integrity dengan membuat FOREIGN KEY.
   i. Mahasiswa paham kapan dan bagaimana referential integrity menggunakan ON DELETE CASCADE ON UPDATE CASCADE.

j. Mahasiswa paham kapan dan bagaimana referential integrity menggunakan ON DELETE RESTRICT ON UPDATE RESTRICT.
k. Mahasiswa dapat menghapus table menggunakan perintah SQL: DROP TABLE.
l. Mahasiswa dapat menghapus database menggunakan perintah SQL: DROP DATABASE.

3. Kompetensi *Database* SQL DML.

   Kompetensi *Database* SQL DML diambil dari kuesioner yang meliputi:

   a. Mahasiswa bisa membuat memasukkan data kedalam tabel menggunakan perintah SQL: INSERT INTO.
   b. Mahasiswa bisa mengganti nilai data pada sebuah kolom dengan perintah SQL: UPDATE …. SET.
   c. Mahasiswa bisa menghapus baris table menggunakan perintah SQL: DELETE FROM.
   d. Mahasiswa bisa mengambil semua data dalam table menggunakan SQL: SELECT.
   e. Mahasiswa bisa melakukan seleksi data menggunakan klausa WHERE.
   f. Mahasiswa melakukan pengurutan data menggunakan klausa ORDER BY.
   g. Mahasiswa bisa melakukan query 2 tabel menggunakan operasi LEFT JOIN.
   h. Mahasiswa bisa melakukan query 2 tabel menggunakan operasi RIGHT JOIN.
   i. Mahasiswa bisa melakukan query 2 tabel menggunakan operasi INNER JOIN.
   j. Mahasiswa bisa melakukan seleksi hasil query menggunakan operator LIKE.
   k. Mahasiswa bisa melakukan seleksi hasil query menggunakan operator IN.

l. Mahasiswa bisa melakukan seleksi hasil query menggunakan operator BETWEEN.
m. Mahasiswa bisa melakukan penggabungan hasil query menggunakan UNION.

4. Kompetensi *Database* Agregasi.

   Kompetensi *Database* Agregasi yang diambil dari kuesioner yang meliputi :

   a. Mahasiswa dapat melakukan seleksi data dengan fungsi SUM.
   b. Mahasiswa dapat melakukan seleksi data dengan fungsi COUNT.
   c. Mahasiswa dapat melakukan seleksi data dengan fungsi MAX.
   d. Mahasiswa dapat melakukan seleksi data dengan fungsi MIN.
   e. Mahasiswa dapat melakukan seleksi data dengan fungsi AVG.
   f. Mahasiswa dapat melakukan seleksi data berakumulasi menggunakan GROUP BY.
   g. Mahasiswa dapat melakukan seleksi agregasi menggunakan HAVING.

5. Kompetensi *Database Advance*

   Kompetensi *Database Advance* diambil dari kuesioner yang meliputi :

   a. Mahasiswa dapat membuat table virtual menggunakan printah SQL: CREATE VIEW … AS SELECT.
   b. Mahasiswa dapat membuat index tabel menggunakan perintah SQL: CREATE INDEX.
   c. Mahasiswa dapat mengosongkan isi table menggunakan perintah SQL: TRUNCATE.
   d. Mahasiswa dapat menampilkan hasil query secara unik menggunakan perintah SQL: DISTINCT.
   e. Mahasiswa dapat menghapus index table menggunakan perintah SQL: DROP INDEX.

f. Mahasiswa dapat menghapus view menggunakan SQL: DROP VIEW.
g. Mahasiswa dapat membuat query bersarang (sub-query).

2.6 Algoritma *Clustering*

Clustering adalah proses pengelompokan objek data ke dalam kelas-kelas berbeda yang disebut *cluster* sehingga objek yang berada pada *cluster* yang sama semakin mirip dan berbeda dengan objek pada cluster yang lain. Teknik *clustering* banyak diterapkan pada berbagai bidang antara lain pengenalan pola, *machine learning, data mining, information retrieval*, dan *bioinformatics*.

2.6.1 Algoritma K-Means

K-Means (KM) adalah salah satu algoritma paling populer yang digunakan untuk proses *clustering* karena efisiensinya pada saat berurusan dengan data yang banyak. Meskipun algoritma tersebut mudah diimplementasikan dan dapat bekerja dengan cepat pada banyak situasi, algoritma KM memiliki beberapa kelemahan, diantaranya hasil klaster sensitif terhadap penentuan awal (inisialisasi) centroid dan mungkin hasilnya dapat mengarah kepada lokal optima.

Algoritma c*lustering* dengan *K-Means* :

1. Tentukan Nilai K sebagai jumlah kelompok / *cluster*.
2. Inisialisasi posisi centroid awal dimana sebanyak K *centroid* secara acak dari data yang ada.
3. Hitung Jarak data terhadap masing-masing centroid. Misalnya menggunakan rumus jarak *euclidean*.
4. Cari jarak terdekat dan masukkan X kedalam *cluster* sesuai dengan *centroid* tersebut.
5. Cari *centroid* baru sebanyak K dari rata-rata dalam kelompok/*cluster*
6. Lakukan langkah nomor 3 - 5 hingga posisi anggota *cluster* tidak berubah.

2.6.2 Algoritma K-Harmonic Means

K-Harmonic Means (KHM) pertama kali diperkenalkan oleh Zhang, Hsu, dan Dayal (1999) dari HP Laboratories Palo Alto. KHM dikembangkan untuk menangani masalah utama dalam K-Means yang hasil klasteringnya sangat sensitif dengan inisialisasi data yang dijadikan sebagai centroid awal. Hasil yang sering berbeda (lokal optima) dari proses klasteringnya (pada set data yang sama) disebabkan oleh inisialisasi centroid yang berbeda. KHM juga salah satu metode *clustering* berbasis partisi yang menggunakan rata-rata harmonic (*harmonic average*) jarak dari setiap titik data ke centroid sebagai komponen dalam fungsi kinerja (fungsi objektif). KHM secara signifikan meningkatkan kualitas hasil klastering dibandingkan dengan metode seperti K-Means. Kualitas yang lebih baik tersebut adalah bahwa hasil klaster yang didapat berusaha mendekati hasil yang global optima (hasil cluster yang didapat selalu sama) atau juga disebut dengan konvergen. Algoritma KHM sangat mirip dengan KM, hal yang menjadi pembeda antara KHM dengan KM adalah pencarian *centroid* baru yang didapatkan dengan implementasi perhitungan rata-rata harmonik (*harmonic average*).

Rata-rata harmonik (*harmonic average* / HA) adalah kebalikan dari rata-rata aritmetik dalam set nilai. Untuk HA dari K nilai dinyatakan oleh persamaan berikut :

Rumus 2.1 Rata-rata harmonik :

$$HA(\{ a_i \mid i=1,..,K \}) = \frac{K}{\sum_{i=1}^{K} \frac{1}{a_i}} \qquad \text{……..……..} \quad (2.1)$$

Algoritma c*lustering* dengan *K-Harmonic Means* :

1. Tentukan Nilai K sebagai jumlah kelompok / *cluster*.
2. Inisialisasi posisi centroid awal dimana $C = \{c_j \mid j = 1, \ldots, K\}$ sebanyak K *centroid* secara acak dari data yang ada.
3. Hitung Jarak data terhadap masing-masing centroid. Misalnya menggunakan rumus jarak *euclidean* seperti persamaan berikut :

Rumus 2.2 Jarak *Euclidean* :

$$d_{i,j} = ||x_i - c_j||_2 = \sqrt[2]{(x_i - c_j)^2} \quad \text{.....} \quad (2.2)$$

Dimana X = { x_i | i=1…..N }, N adalah jumlah data yang akan diklaster dengan metode KHM.

4. Cari jarak terdekat $d_{i,min}$ dan masukkan X kedalam *cluster* sesuai dengan kelompok/*centroid* tersebut.
5. Cari *centroid* baru sebanyak K dengan persamaan berikut :

   Rumus 2.3 Mencari *centroid* dengan formula rekursif KHM :

$$m_k = \frac{\sum_{i=1}^{N} \frac{1}{d_{i,k}^{3} \left(\sum_{j=1}^{K} \frac{1}{d_{i,j}^{2}} \right)^{2}} \cdot x_i}{\sum_{i=1}^{N} \frac{1}{d_{i,k}^{3} \left(\sum_{j=1}^{K} \frac{1}{d_{i,j}^{2}} \right)^{2}}} \quad \text{..... (2.3)}$$

   Catatan : $d_{i,k} = d_{i,min} = 0$, maka vektor m_k di set menjadi 0.
6. Lakukan langkah nomor 3 - 5 hingga posisi anggota *cluster* tidak berubah.

(Zhang, Shu, Dayal. 1999. K-Harmonic Means-A Data Clustering Algorithm. *Technical Report HPL-1999-124*, Hewlett-Packard Laboratories).