BAB I

PENDAHULUAN

1.1 LATAR BELAKANG

Saat ini, konsep data *mining* semakin dikenal sebagai *tools* penting dalam manajemen informasi karena jumlah informasi yang semakin besar jumlahnya. Data *mining* sendiri sering disebut sebagai *knowledge discovery in database* (KDD) adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola hubungan dalam data berukuran besar. *Output* dari data *mining* ini dapat digunakan untuk pengambilan keputusan di masa depan (Edward, 2006).

يَٰٓأَيُّهَا ٱلنَّاسُ إِنَّا خَلَقۡنَٰكُم مِّن ذَكَرٖ وَأُنثَىٰ وَجَعَلۡنَٰكُمۡ شُعُوبٗا وَقَبَآئِلَ لِتَعَارَفُوٓاْۚ إِنَّ أَكۡرَمَكُمۡ عِندَ ٱللَّهِ أَتۡقَىٰكُمۡۚ إِنَّ ٱللَّهَ عَلِيمٌ خَبِيرٞ ١٣

"Hai manusia, Sesungguhnya Kami menciptakan kamu dari seorang laki-laki dan seorang perempuan dan menjadikan kamu berbangsa - bangsa dan bersuku-suku supaya kamu saling kenal-mengenal. Sesungguhnya orang yang paling mulia diantara kamu disisi Allah ialah orang yang paling taqwa diantara kamu. Sesungguhnya Allah Maha mengetahui lagi Maha Mengenal" (Qs. 49(Al-hujrat): 13).

Mengawali pembahasan analisis klaster, ayat Al-Quran di atas sengaja dikutipkan. Terdapat banyak karakteristik dalam diri manusia. Kita berbeda dalam hal bahasa, warna kulit, warna bola mata, bentuk rambut, postur tubuh dan masih

banyak lagi perbedaan lainnya. Untuk memudahkan identifikasi, manusia kita kelompok-kelompokkan menjadi bagian-bagian kecil. Manusia penghuni dunia bisa kita kelompokkan menurut bangsanya. Di dalam satu bangsa bisa dikelompokkan lagi menurut suku-suku dalam satu bangsa dan seterusnya. Dalam analisis multivariat, untuk pengelompokkan objek digunakan analisis kelompok atau lebih dikenal dengan analisis klaster (*cluster analysis*).

Salah satu teknik yang dikenal dalam data *mining* yaitu *clustering*. Pengertian *clustering* dalam data *mining* adalah pengelompokan sejumlah data atau objek ke dalam *cluster* (*group*) sehingga setiap dalam *cluster* tersebut akan berisi data yang semirip mungkin dan berbeda dengan objek dalam *cluster* yang lainnya (Santosa B., 2007).

Sampai saat ini, para ilmuwan masih terus melakukan berbagai usaha untuk melakukan perbaikan model *cluster* dan menghitung jumlah *cluster* yang optimal sehingga dapat dihasilkan *cluster* yang paling baik. Ada beberapa metode *clustering* yang kita kenal, yaitu *hierarchical, K-means, self organizing maps (SOM) clustering* (Alfina, 2012).

Metode K-*means* merupakan metode *clustering* yang paling sederhana dan umum. Hal ini dikarenakan K-*means* mempunyai kemampuan mengelompokkan data dalam jumlah yang cukup besar dengan waktu komputasi yang relatif cepat dan efisien. Namun, K-*means* mempunyai mempunyai kelemahan yang diakibatkan oleh penentuan pusat awal *cluster*. Hasil *cluster* yang terbentuk dari metode K-*means* ini sangatlah tergantung pada inisiasi nilai pusat awal *cluster*

yang diberikan. Hal ini menyebabkan hasil klasternya berupa solusi yang sifatnya *local optimal* (K. Arai, 2007).

Metode hierarki dapat dibedakan menjadi dua bagian, yaitu metode penggabungan (*agglomerative*) dan metode pemecahan (*devisive*). Pembentukan kelompok dalam metode hierarki, menggunakan beberapa cara, antara lain pautan tunggal (*single linkage*), pautan lengkap (*complete linkage*), dan pautan rata-rata (*average linkage*). Metode ini bisa terjadi masalah untuk set data yang mengandung *noise*, dan data berdimensi tinggi. Biasanya, untuk masalah ini dibantu dengan metode lain secara parsial, seperti k-*means* (Prasetyo, Eko. 2012).

Self Organizing Maps (SOM) merupakan suatu tipe *Artificial Neural Networks* yang di-*training* secara *unsupervised*. SOM menghasilkan map yang terdiri dari *output* dalam dimensi yang rendah (2 atau 3 dimensi). Map ini berusaha mencari *property* dari *input* data. Komposisi *input* dan *output* dalam SOM mirip dengan komposisi dari proses *feature scaling (multidimensional scaling).* Walaupun proses *learning* yang dilakukan mirip dengan *Artificial Neural Networks*, tetapi proses untuk meng-*assign input* data ke map, lebih mirip dengan K-*Means* dan kNN *Algorithm* (Agusta, 2007)

Untuk itu, metode K-*means, hierarchical* dan *self organizing maps* akan dibandingkan untuk mendapatkan hasil *cluster* yang sesuai. Dari proses pengelompokan ini nantinya diharapkan akan diketahui kemiripan atau kedekatan antar data sehingga dapat dikelompokkan ke dalam beberapa *cluster*, dimana antar anggota *cluster* memiliki tingkat kemiripan yang tinggi. Maka berdasarkan hal itu

judul skripsi ini adalah "**Studi Komparatif Penerapan Metode *Hierarchical*, K-*Means* dan *Self Organizing Maps* (SOM) *Clustering* Pada Basis Data**".

1.2 Rumusan Masalah

Berdasarkan penjelasan di atas, maka dapat diidentifikasi masalah-masalah yang dihadapi, yaitu:

1. Bagaimana membandingkan metode *cluster* yang sesuai dengan data yang akan dikelompokan.
2. Bagaimana cara menentukan jumlah *cluster* yang ideal.
3. Bagaimana cara mendapatkan anggota *cluster* yang memiliki tingkat kemiripan yang tinggi.

1.3 Maksud dan Tujuan

Adapun maksud dari penelitian ini adalah menganalisis beberapa metode yang diterapkan pada proses *clustering* sehingga menghasilkan *cluster (group)* yang sesuai.

Adapun tujuan yang akan dicapai dalam penelitian ini adalah :

1. Membandingkan hasil *cluster* metode *hierarchial*, k-*means* dan *self organizing maps (SOM)*.
2. Menentukan jumlah *cluster* yang ideal untuk masing-masing metode tersebut.
3. Mengetahui kualitas kemiripan hasil pengelompokan data menggunakan metode *hierarchial*, k-*means* dan *self organizing maps (SOM)*.

1.4 Batasan Masalah

Agar penelitian ini tidak meluas dari lungkup permasalahan dan supaya lebih terfokus dan terarah maka akan diberikan batasan terhadap penelitian yang akan dibahas, yaitu:

1. Pengelompokan data yang digunakan menggunakan metode *hierarchial*, k-*means* dan *self organizing maps (SOM)*.
2. Sumber data uji merupakan sampel data yang telah dikumpulkan sebelumnya atau bisa didownload pada situs penyedia data set uji untuk klaster seperti *http://cml.ics.uci.edu/* dan *http://kdd.ics.uci.edu/*.
3. Jenis data uji merupakan file dengan *extension .txt (*berupa *tab-detimited)* atau *.xls*
4. Penggunaan metode perancangan perangkat lunak RAD (*Rapid application development*).
5. Tidak terdapat hak akses untuk menggunakan aplikasi.

1.5 State of the Art

Banyak penelitian yang sebelumnya dilakukan mengenai perbandingan metode-metode *clustering*. Dalam upaya mengembangkan dan menyempurnakan maka perlu dilakukan studi literatur sebagai salah satu dari penerapan metode penelitian yang akan dilakukan. Adapun manfaat dari studi literatur ini antara lain

1. Menghindari membuat ulang sehingga banyak menghemat waktu dan juga menghindari kesalahan-kesalahan yang dilakukan orang lain.
2. Mengidentifikasi metode yang pernah dilakukan dan relevan terhadap penelitian ini.

3. Meneruskan penelitian sebelumnya yang telah dicapai orang lain. Sehingga, dengan adanya studi literatur, penelitian yang akan dilakukan dapat membangun di atas *platform* atau ide yang sudah ada.

Berikut ini adalah penelitian yang telah dilakukan dan memiliki korelasi yang searah dengan penelitian yang dibahas, antara lain :

Penelitian Tahta Alfina (2012) membahas tentang analisa perbandingan metode k-*means, hierarchical clustering* yang menghasilkan suatu kesimpulan bahwa dalam studi kasus Problem Kerja Praktek jurusan Teknik Industri ITS, dari kombinasi *hierarchical clustering* dan K-*means* yang ada, kombinasi *single linkage clustering* dan K-*means* menghasilkan pengelompokan data yang terbaik dibandingkan dengan metode *hierarki* yang lainnya.

Penelitian Lathifaturrahman (2010) membahas tentang perbandingan hasil penggrombolan metode k-*means*, *fuzzy* k-*means* dan *two step cluster* Jumlah gerombol ideal yang dihasilkan oleh masing-masing metode tersebut adalah 2 gerombol karena memiliki nilai *rasio* yang lebih kecil antara nilai rata-rata jumlah kuadrat dalam gerombol dengan antar gerombol. Hasil dari masing-masing gerombol metode k-*means* dan *fuzzy* k-*means* lebih mirip pada penggerombolan 2 gerombol, sedangkan metode *two step cluster* dari awal penggerombolan jumlah anggota gerombol yang agak jauh berbeda dengan kedua metode lainnya.

Penelitian Nursinta Adi Wahanani (2012) yang membahas tentang optimasi *clustering* K-*means* dengan algoritma genetika multiobyektif yang menghasilkan sebuah kesimpulan bahwa Perbaikan kinerja K-*Means* bisa dilakukan dengan menggunakan metode algoritma genetika multiobyektif dengan pendekatan *pareto*

rangking. Hasil yang didapat berupa *pareto front* yang merupakan himpunan solusi yang memenuhi tujuan meminimalkan *varian* dalam *cluster* dan memaksimalkan *varian* antar *cluster*.

Penelitian Edward (2006) yang membahas tentang *clustering* menggunakan *self organizing maps* dengan studi kasus Panitia Penerimaan Mahasiswa Baru Institut Pertanian Bogor (PPMB IPB) yang menghasilkan sebuah kesimpulan bahwa penelitian tersebut belum difokuskan untuk optimasi kombinasi nilai-nilai parameter algoritma SOM untuk memperoleh hasil yang optimal.

Penelitian Liesca Levy Shandy (2008) yang membahas tentang Perbandingan Metode *Diskretisasi* Data *Partisi Intuitif* dan K-*Means Clustering* Terhadap Pembuatan Pohon Keputusan yang menghasilkan sebuah kesimpulan bahwa hasil penelitian dapat dinyatakan bahwa diskretisasi atribut dengan menggunakan algoritma K-*Means clustering* dengan 4 *cluster* memberikan akurasi yang paling tinggi sebesar 87,40 %, diikuti metode *Partisi Intuitif* yang mempunyai akurasi pohon keputusan sebesar 84,54% dan terakhir oleh algoritma K-*Means clustering* dengan 5 *cluster* sebesar 76,87% .

Dari hasil studi literatur yang telah diambil dari beberapa sumber dapat dilihat pada tabel 1.1 berikut:

Tabel 1.1 Perbandingan Studi Literatur

No.	Peneliti	Judul	Metode	Keterangan
1	Tahta Alfina	analisa perbandingan metode k-*means*, *hierarchical clustering* dan gabungan keduanya	K-*Means*, *hierarchial* dan gabungan keduanya	Pada perbandingan tersebut terfokus terhadap salah satu metode.
2	Lathifaturrahman	perbandingan hasil penggrombolan metode k-*means*, *fuzzy* k-*means* dan *two step cluster*	metode k-*means*, *fuzzy* k-*means* dan *two step cluster*	*Clustering* yang dihasilkan hanya terdapat 2 *cluster*.
3	Nursinta Adi Wahanani	optimasi *clustering* K-*means* dengan algoritma genetika multiobyektif	K-*means*	Optimasi yang dihasilkan belum cukup optimal
4	Edward	*clustering* menggunakan *self organizing maps*	*self organizing maps (SOM)*	Befokus pada nilai-nilai paramete
5	Liesca Levy Shandy	Perbandingan Metode *Diskretisasi* Data *Partisi Intuitif* dan K-*Means Clustering* Terhadap Pembuatan Pohon Keputusan	*Diskretisasi* Data *Partisi Intuitif* dan K-*Means*	Akurasi yang dihasilkan belum mendapat hasil yang maksimal
6	Ijang Badruzaman	Studi Komparatif Penerapan Metode *Hierarchical*, K-*Means dan Self Organizing Maps* (SOM) *Clustering* Pada Basis Data	*K-means, Hierarchical, SOM*	Menganalisis hasil dari klaster dan menentukan tingkat akurasi dari hasil metode tersebut.

1.6 Metodologi Penelitian

Metodologi yang digunakan dalam penelitian ini terdiri dari tahap pengumpulan data dan metode pemgembangan sistem:

1. **Tahap Pengumpulan Data**

   a. Studi Lapangan

      1. Observasi.

         Teknik pengumpulan data dengan mengadakan penelitian dan peninjauan langsung terhadap permasalahan yang diambil.

      2. Wawancara.

         Teknik pengumpulan data dengan mengadakan tanya jawab secara langsung yang ada kaitannya dengan topik yang diambil.

   b. Studi Pustaka

      Dalam penyusunan laporan tugas akhir ini, penulis menggunakan beberapa buku sebagai bahan landasan teoritis untuk memperoleh suatu keterangan yang dapat menunjang penyusunan laporan tugas akhir ini.

2. **Metode Pengembangan Sistem**

Rapid application development (RAD) atau *rapid prototyping* adalah model proses pembangunan perangkat lunak yang tergolong dalam teknik *incremental* (bertingkat). RAD menekankan pada siklus pembangunan pendek, singkat, dan cepat. Waktu yang singkat adalah batasan yang penting untuk model ini. *Rapid application development* menggunakan metode *iteratif* (berulang) dalam mengembangkan sistem dimana *working model* (model bekerja) sistem

dikonstruksikan diawal tahap pengembangan dengan tujuan menetapkan kebutuhan (*requirement*) user dan selanjutnya disingkirkan. *Working model* digunakan kadang-kadang saja sebagai basis desain dan implementasi sistem final (Christanta Mega, 2011).

Metode RAD digunakan pada aplikasi sistem konstruksi, maka menekankan fase-fase sebagai berikut:

1. ***Bussiness Modelling***

   Pada tahap ini, aliran informasi (*information flow*) pada fungsi-fungsi bisnis dimodelkan untuk mengetahui informasi apa yang mengendalikan proses bisnis, informasi apa yang hasilkan, siapa yang membuat informasi itu, kemana saja informasi mengalir, dan siapa yang mengolahnya.

2. **Data *Modelling***

   Aliran informasi yang didefinisikan dari *business modeling*, disaring lagi agar bisa dijadikan bagianbagian dari objek data yang dibutuhkan untuk mendukung bisnis tersebut. Karakteristik setiap objek ditentukan beserta relasi antar objeknya.

3. ***Process Modelling***

   Aliran informasi pada fase data *modelling* ditransformasikan untuk mendapatkan aliran informasi yang diperlukan pada implementasi fungsi bisnis. Pemrosesan diciptakan untuk menambah, memodifikasi, menghapus, atau mendapatkan kembali objek data tertentu

4. ***Application Generation***

   Selain menggunakan bahasa pemrograman generasi ketiga, RAD juga memakai komponen program yang telah ada atau menciptakan komponen yang bisa dipakai lagi. Alat-alat bantu bisa dipakai untuk memfasilitasi konstruksi perangkat lunak.

5. ***Testing and Turnover***

   Karena menggunakan kembali komponen yang telah ada, maka akan mengurangi waktu pengujian. Tetapi komponen baru harus diuji dan semua *interface* harus dilatih secara penuh.

1.7 Sistematika Penulisan

Sistematika penulisan laporan ini disusun dalam beberapa bab yang masing-masing bab menguraikan beberapa pokok pembahasan. Adapun sistematika penulisan laporan ini adalah sebagai berikut :

BAB I PENDAHULUAN

Bab ini berisikan tentang latar belakang permasalahan, perumusan masalah yang dihadapi, batasan masalah, tujuan, metodologi, serta bagaimana penulisan sistematika penulisan.

BAB II LANDASAN TEORI

Bab ini menjelaskan tentang teori-teori apa saja yang berkaitan dengan topik yang telah dibuat berdasarkan hasil penelitian dan hal-hal yang berguna dalam proses penyusunan tugas akhir ini.

BAB III ANALISIS KEBUTUHAN

Memuat gambaran analisis yang dibutuhkan oleh sistem, diantaranya proses bisnis sistem, kebutuhan perangkat lunak dan skenario untuk pembuatan proses pembuatan aplikasi.

BAB IV IMPLEMENTASI

Menerangkan pengimplementasian dari sistem yang telah dibangun baik itu *software* yang diperlukan, *hardware* yang mendukung, implementasi *user interface* termasuk pengujian sistem yang telah dibangun.

BAB V PENUTUP

Bab ini berisikan tentang kesimpulan dan saran yang diperoleh dari hasil penulisan laporan tugas akhir.