

BAB II

TINJAUAN PUSTAKA

A. Tinjauan Umum STIKOM Uyelindo Kupang

1. Profil Perusahaan

Sekolah Tinggi Manajemen Informatika Komputer merupakan sebuah lembaga pendidikan yang didirikan pada tahun 2000. Rekomendasi Gubernur Provinsi NTT No 420/10/2000, tanggal 26 Oktober 2000 dimana pada prinsipnya Pemerintah Daerah Tingkat I NTT menyetujui dan mendukung berdirinya Sekolah Tinggi Manajemen Informatika Komputer (STIKOM) Uyelindo Kupang.

Pada tahun 2001 STIKOM Uyelindo Kupang mendapat ijin pendiriannya oleh Kementerian Pendidikan Nasional Republik Indonesia dengan nomor 77/D/O/2001. Sampai saat ini STIKOM Uyelindo Kupang memiliki tiga program studi yakni Sistem Informasi Strata Satu, Teknik Informatika Strata Satu, dan Teknik Informatika Diploma Tiga. Masing-masing program studi telah terakreditasi oleh badan akreditasi nasional.

Tiap tahun jumlah mahasiswa baru selalu mengalami peningkatan, namun jumlah lulusan tiap tahun tidak mengalami peningkatan sejalan dengan peningkatan mahasiswa baru. Pihak manajemen kesulitan untuk menelusuri sebaran lulusan tiap tahun terhadap sebaran angkatan (tahun masuk) dalam rangka melakukan evaluasi kesesuaian antara jumlah masuk dan jumlah lulusan.

Dukungan pihak manajemen terhadap sarana dan prasarana untuk menunjang akademik sangat baik. Termasuk sarana dan prasarana pengembangan

sistem akademik. Namun sumber daya manusia belum bekerja secara optimal. Perlu adanya pendidikan dan latihan khususnya di bidang sistem informasi/teknologi yang secara berkelanjutan.

Dalam mengembangkan sistem akademik STIKOM Uyelindo secara keseluruhan memang tidak mudah. Selain itu membutuhkan waktu dan perencanaan yang matang serta proses pelaksanaan secara bertahap dari waktu ke waktu agar semua sistem yang direncanakan dapat berjalan dengan baik. Namun tentu saja evaluasi dari kegiatan sesuai dengan tahapan pelaksanaan masing-masing.

Dalam menunjang proses kegiatan akademik, Stikom Uyelindo memiliki sebuah sistem akademik untuk memenuhi berbagai kebutuhan operasional kegiatan belajar mengajar. Sistem ini berjalan secara *stand alone* pada Biro Administrasi Akademik Kemahasiswaan (BAAK). Sementara unit-unit yang terlibat seperti bagian keuangan (BAUK), program studi, maupun para pembantu ketua dan ketua tidak dapat mengakses data secara *real time*. Demikian halnya dengan penerimaan mahasiswa baru, pengolahan data skripsi belum memiliki sebuah sistem untuk berinteraksi dengan sistem akademik yang ada.

2. Visi, Misi dan Tujuan

a. Visi

Adapun visi yang dimiliki oleh STIKOM Uyelindo Kupang yakni :

“Perguruan Tinggi Komputer Yang Unggul, Terdepan, Terbaik dan Terpercaya”.

b. Misi

Adapun misi yang dimiliki oleh STIKOM Uyelindo Kupang yakni :

1. Menyelenggarakan Tri Dharma Perguruan Tinggi berbasis teknologi informasi yang dapat beradaptasi terhadap perubahan global.
2. Mengembangkan sistem informasi atau sistem komputerisasi yang berdaya saing global.

c. Tujuan

Adapun tujuan yang dimiliki oleh STIKOM Uyelindo Kupang yakni :

1. Menghasilkan lulusan yang kreatif dan inovatif terhadap perkembangan teknologi informasi serta berbudi pekerti luhur.
2. Menghasilkan lulusan yang mampu bersaing dalam dunia kerja.
3. Mengembangkan STIKOM Uyelindo Kupang sebagai pusat penelitian dan pengembangan IT di kawasan Timur Indonesia.

B. Tinjauan Pustaka

Untuk mengetahui lebih jauh mengenai pemanfaatan analisis dan implementasi sistem pengolahan data yang besar dan pendukung keputusan untuk meningkatkan kesuksesan studi mahasiswa pada STIKOM Uyelindo, maka perlu ditinjau dahulu mengenai penelitian-penelitian sebelumnya, yang berkaitan dengan masalah-masalah profil akademik alumni sebagai bahan perbandingan dan kajian. Adapun hasil-hasil penelitian yang dijadikan perbandingan tidak terlepas dari topik penelitian yaitu tentang profil akademik alumni dengan menggunakan metode klasterisasi k-means.

Dalam melakukan penyusunan makalah ini, penulis telah menelaah beberapa jurnal yang ada untuk menjadi referensi yang diantaranya adalah Analisa Profil Data Mahasiswa Baru Terhadap Program Studi Yang Dipilih di Perguruan Tinggi Swasta Jawa Tengah Dengan Menggunakan Teknik Data Mining. Dalam jurnal ini membahas tentang teknik *data mining* klusterisasi yang diharapkan trend mahasiswa baru dapat diketahui berdasarkan bidang ilmu dan program studi yang dipilihnya, dengan demikian orientasi penerimaan mahasiswa baru dapat diprediksi dimasa yang akan datang, khususnya pada program studi yang banyak peminat dan program studi yang kurang peminat (Wahyudi dkk, 2011).

Penggunaan klusterisasi juga digunakan dalam berbagai bidang dalam penyampaian informasi diantara dalam bidang akademik dengan sebuah artikel Analisis Komparasi Untuk Prediksi Mahasiswa Non Aktif disini mengungkapkan bagaimana mengetahui faktor-faktor penyebab mahasiswa memiliki status non aktif. Berdasarkan komparasi algoritma klasifikasi data mining yaitu *logistic regression*, *decision tree*, *narve bayes*, dan *neural network* untuk prediksi mahasiswa non aktif dengan menggunakan 3861 data set mahasiswa studi kasus pada Universitas Dian Nuswantoro. Hasil dari proses klasifikasi dievaluasi dengan menggunakan *cross validation*, *confusion matrix*, *ROC Curve* dan *T-test* untuk mengetahui algoritma klasifikasi *data mining* yang paling akurat untuk prediksi mahasiswa non aktif. Berdasarkan analisis dapat disimpulkan bahwa dalam uji T-Test, algoritma *logistic regression* menunjukkan hasil bahwa algoritma tersebut paling dominan terhadap algoritma yang lain meskipun nilai *accuracy* tidak menunjukkan hasil yang tertinggi yaitu 81,64% dan nilai AUC 0,933. Algoritma

decision tree meskipun memiliki *accuracy* tertinggi yaitu 95,29% dan nilai AUC mencapai 0,963 ternyata dalam uji T-Test tidak dominan terhadap algoritma yang lain. Algoritma *naïve bayes* mencapai *accuracy* 93,47%, nilai AUC 0,976 dan dalam pengujian T-Test bersifat dominan terhadap algoritma yang lain. Algoritma *neural network* memiliki nilai *accuracy* yang cukup tinggi yaitu 94,56%, dalam uji AUC mencapai hasil maksimal yaitu 0,976 dan bersifat dominan terhadap algoritma yang lain (Hastuti, 2012).

Selain itu klasterisasi digunakan pada Pengembangan Perangkat Lunak prediktor Nilai Mahasiswa. Dalam jurnal ini membahas tentang pengelompokan data dengan menggunakan *spectral clustering* yang mendukung Bootstrap Aggregating Regresi Linier memiliki performa yang lebih baik jika dibandingkan dengan perangkat lunak yang menggunakan algoritma K-Means Clustering (Yusuf dkk, 2012). Rismawan dan Kusumadewi (2008) dalam jurnal Aplikasi K-means Untuk Pengelompokan Mahasiswa Berdasarkan Nilai Body Mass Index (BMI) & Ukuran Kerangka menyatakan bahwa algoritma klasifikasi K-means dapat digunakan berdasarkan status gizi dan ukuran kerangka.

Meinanda (2009) menyatakan bahwa variabel Indeks Prestasi Kumulatif, jumlah mata kuliah yang diambil, jumlah mata kuliah mengulang, dan jumlah pengambilan mata kuliah tertentu dapat mempengaruhi masa studi. Dalam melakukan prediksi masa studi, model regresi akan menghasilkan prediksi masa studi yang bias karena asumsi *Homoscedasticity Error* tidak dapat terpenuhi. Alfina dkk, 2012 juga menerapkan gabungan metode *single linkage clustering* dan

k-means yang memberikan hasil cluster yang lebih baik dengan parameter uji *cluster variance* dan metode *silhouette coefisien*.

Penggunaan klasterisasi ini juga digunakan dalam bidang bisnis dalam sebuah artikel Penerapan Data Mining Algoritma Asosiasi Untuk Meningkatkan penjualan. Dalam jurnal ini membahas penggunaan data mining algoritma asosiasi untuk menyusun sebuah sistem yang memiliki kemampuan melihat pola penjualan barang yang selanjutnya dapat digunakan untuk menyujusun strategi penjualan baru (Luthfi, 2009).

Pada tabel 1. terdapat perbandingan tabel antara penelitian yang dibuat dengan penelitian - penelitian sebelumnya, disini terlihat jelas bahwa penggunaan sebuah klasterisasi k-means dapat membantu dalam pengambilan keputusan di STIKOM Uyelindo Kupang untuk meningkatkan kesuksesan studi mahasiswa.

Tabel 1. Tabel Perbandingan

Fitur	Emha Taufiq Luthfi	Eko Nur Wahyudi, Jananto, Narwati	Arief Arief	Tahta Alfina, Budi Barakbah	Penulis
Judul	Penerapan data mining algoritma asosiasi untuk meningkatkan penjualan	Analisa profil data mahasiswa baru terhadap program studi yang di pilih di PT swasta jawa tengah dengan menggunakan teknik data mining		Analisa perbandingan hierachical clustering, k-means, dan gabungan keduanya dalam cluster data	Analisa profil akademik alumni dengan menggunakan metode klasterisasi k-means
Tahun	2009	2011	2012	2013	
Metode	Data mining asosiasi	Data Mining, Klastering, K-means	<i>K-means, Hierarchica, Clustering, Cluster Variance, Metode Silhouette Coeficient</i>		Data mining, klasterisasi, k-means
Subjek	Pelanggan	Mahasiswa		Mahasiswa	Alumni
Konten	Penjualan	Program Studi		Hirarchical clustering dan k-means	Akademik
Web base	-	-		-	√
SKPL & DPPL	-	-		-	√

C. Landasan Teori

1. *Knowledge Discovery In Database (KDD)*

Knowledge Discovery In Database sendiri merupakan sekumpulan proses untuk menemukan pengetahuan yang bermanfaat dari data (Abriansyah, 2010).

KDD terdiri dari serangkaian langkah perubahan, termasuk data *preprocessing* dan juga *post processing*. *Data preprocessing* merupakan langkah untuk mengubah data mentah menjadi format yang sesuai untuk tahap analisis berikutnya. Selain itu data *preprocessing* juga digunakan untuk membantu dalam pengenalan atribut dan data segmen yang relevan dengan task *data mining* (Abriansyah 2010).

Kumpulan proses dalam KDD meliputi : pembersihan data (*data cleaning*), integrasi data (*data integration*), pemilihan data (*data selection*), transformasi data (*data transformation*), penambangan data (*data mining*), evaluasi pola (*pattern evaluation*), dan presentasi pengetahuan (*knowledge presentation*). Berdasarkan definisi ini terlihat bahwa *data mining* hanya merupakan salah satu proses dari keseluruhan proses yang ada pada KDD, tetapi merupakan proses yang sangat penting dalam usaha menemukan pola-pola yang berguna dari sejumlah data yang besar (data tersebut bisa disimpan dalam basisdata, *data warehouse*, atau media penyimpanan informasi lainnya) (Abriansyah, 2010).

2. *Business Intelligence*

Business Intelligence (BI) merupakan sebuah proses untuk melakukan ekstraksi data-data operasional perusahaan dan mengumpulkannya dalam sebuah *data warehouse*. Selanjutnya data di *data warehouse* diproses menggunakan

berbagai analisis statistik dalam proses data mining, sehingga didapat berbagai kecenderungan atau pattern dari data. Hasil penyederhanaan dan peringkasan ini disajikan kepada *end user* yang biasanya merupakan pengambil keputusan bisnis. Dengan demikian manajemen dapat mengambil keputusan berdasarkan fakta-fakta aktual, dan tidak hanya mengandalkan intuisi dan pengalaman kuantitatif saja (Subaridargombe, 2006).

Pada prakteknya, BI akan berfungsi sebagai analisis, penghitung *scorecard*, sekaligus memberikan rekomendasi pada user terhadap tindakan yang sebaiknya diambil. Dengan menjalankan fungsi *dashboard*, user BI akan mengenali potensi ketidakberesan pada perusahaan sekaligus dengan penyebabnya sebelum hal tersebut berkembang menjadi masalah yang besar. BI akan berfungsi memberikan advance alarm, memberikan informasi trend dan melakukan benchmark. BI juga berfungsi untuk membantu dalam membuat keputusan perusahaan atau bisnis secara cepat dan akurat (Subaridargombe, 2006).

Didalam BI juga terdapat komponen *data mining*/OLAP yaitu proses dalam bisnis intelligent yang tidak hanya berhenti sampai dalam tahap pembuatan data warehouse, tetapi yang lebih penting ialah kemampuan untuk menganalisis data tersebut menjadi salah satu bahan pertimbangan dalam mengambil keputusan. OLAP/Data Mining dapat membantu untuk menganalisis data data dalam *warehouse* (Subaridargombe, 2006).

3. *Data Mining*

Data mining adalah istilah yang digunakan untuk menggambarkan proses penggalian nilai dari database (Nitu&Ashish, 2012). Ada beberapa peneliti

berpendapat bahwa data mining merupakan salah satu solusi untuk menjelaskan proses penggalian informasi dalam suatu basis data yang berskala besar dan proses klasifikasi otomatisasi kasus berdasarkan pola data yang diperoleh dari data set (Kusnawi, 2007;Varun &Nisha, 2011).

Data mining juga didefinisikan sebagai satu set teknik yang digunakan secara otomatis untuk mengeksplorasi secara menyeluruh dan membawa ke permukaan relasi-relasi yang kompleks pada set data yang sangat besar. Set data yang dimaksud disini adalah set data yang berbentuk tabulasi, seperti yang banyak diimplementasikan dalam teknologi manajemen basis data rasional. Akan tetapi, teknik-teknik *data mining* dapat juga diaplikasikan seperti domain data spasial, berbasis text, dan multimedia (Huda, 2010).

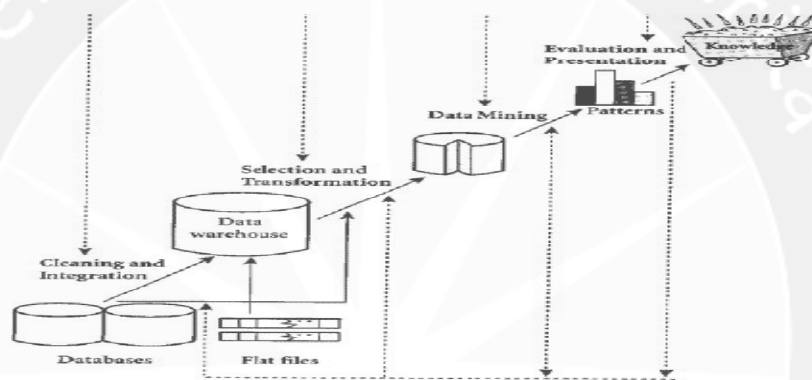
Data mining menggunakan pendekatan *discovery-based* dimana pencocokan pola dan algoritma yang lain digunakan untuk menentukan relasi relasi kunci didalam data yang dieksplorasi. *Data mining* merupakan komponen baru pada arsitektur sistem pendukung keputusan (DSS) (Huda, 2010).

Komponen *data mining* pada proses KDD seringkali merupakan aplikasi iteratif yang berulang dari metodologi data mining tertentu. Tujuan utama dari data mining adalah untuk membuat prediksi dan deskripsi. Prediksi menggunakan beberapa variabel atau field-field basis data untuk memprediksi nilai-nilai variabel masa mendatang yang diperlukan, yang belum diketahui saat ini (Huda, 2010).

Salah satu tuntutan dari *data mining* ketika diterapkan pada data berskala besar adalah diperlukan metodologi sistematis tidak hanya ketika melakukan analisa saja tetapi juga ketika mempersiapkan data dan juga melakukan

interpretasi dari hasilnya sehingga dapat menjadi aksi ataupun keputusan yang bermanfaat (Kusnawi, 2007).

Data mining seharusnya dipahami sebagai suatu proses, yang memiliki tahapan-tahapan tertentu dan juga ada umpan balik dari setiap tahapan sebelumnya. Pada umumnya proses data mining berjalan interaktif karena tidak jarang hasil *data mining* pada awalnya tidak sesuai dengan hasil analisisnya, sehingga perlu dilakukan desain ulang prosesnya (Kusnawi, 2007).



Gambar 1. Tahap – Tahap Data Mining (Kusnawi, 2007).

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 tidak relevan.

2. Integrasi data (*data integration*)

Integrasi data merupakan penggabungan data dari berbagai database ke dalam satu database baru.

3. Seleksi data

Data yang ada pada database sering kali tidak semuanya dipakai. Oleh karena itu, hanya data yang sesuai untuk dianalisis yang akan diambil dari database.

4. Transformasi data

Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam data mining.

5. Proses mining

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

6. Evaluasi pola

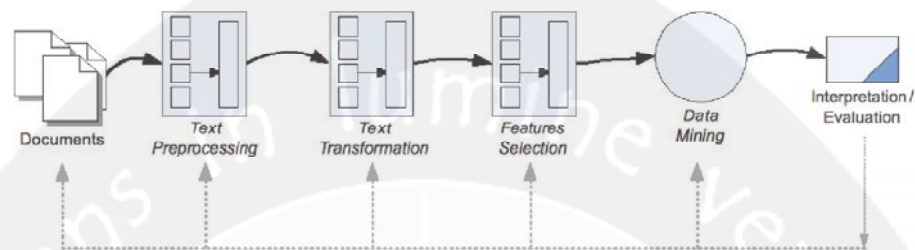
Untuk mengidentifikasi pola-pola menarik ke dalam knowledge base yang ditemukan. Dalam tahap ini hasil dari teknik data mining berupa pola-pola yang khas maupun model prediksi dievaluasi untuk menilai apakah hipotesa yang ada memang tercapai.

4. *Text Mining*

Text mining merupakan penerapan konsep dan teknik data mining untuk mencari pola dalam teks. Proses penganalisisan teks guna menyarikan informasi yang bermanfaat untuk tujuan tertentu. Proses data mining untuk data dokumen atau teks memerlukan lebih banyak tahapan, mengingat teks memiliki karakteristik yang lebih kompleks daripada data biasa (Santoso, 2010).

Berdasarkan ketidakaturan struktur data teks, maka proses *text mining* memerlukan tahapan awal yang pada intinya adalah mempersiapkan agar teks

dapat diubah menjadi lebih terstruktur. Bentuk perubahan yang dilakukan adalah dalam spreadsheet, kolom menunjuk dokumen dan baris menunjuk kata, sedangkan selnya menunjuk frekuensi dalam dokumen (Santoso, 2010).



Gambar 2. Proses Text Mining (Santoso, 2010).

5. Clustering

Salah satu metode yang diterapkan dalam KDD adalah *Clustering*. *Clustering* adalah membagi data ke dalam grup-grup yang mempunyai objek yang karakteristiknya sama. *Clustering* memegang peranan penting dalam aplikasi data mining, misalnya eksplorasi data ilmu pengetahuan, pengaksesan informasi dan *text mining*, aplikasi basis data spesial dan analisis web. *Clustering* diterapkan dalam mesin pencari di internet. Web mesin pencari akan mencari ratusan dokumen yang cocok dengan kata kunci yang dimasukkan (Andayani, 2007).

Clustering bermanfaat untuk melakukan analisis pola-pola yang ada, mengelompokkan, membuat keputusan dan *machine learning*, termasuk data mining, dokumen retrieval, segmentasi citra dan klasifikasi pola. Metodologi clustering lebih cocok digunakan untuk eksplorasi hubungan antar data untuk membuat suatu penilaian terhadap strukturnya.

Dengan menggunakan klusterisasi, kita dapat mengidentifikasi daerah yang padat, menentukan pola-pola distribusi secara keseluruhan, dan menemukan pola-

pola distribusi secara keseluruhan, dan menemukan keterkaitan yang menarik antara atribut-tribut data. Dalam data mining, usaha difokuskan pada metode-metode penemuan untuk klaster pada basisdata berukuran besar secara efektif dan efisien.

Secara garis besar, terdapat beberapa metode klasterisasi data pemilihan metode klasterisasi bergantung pada tipe data dan tujuan klasterisasi itu sendiri. Metode-metode beserta algoritma yang termasuk didalamnya meliputi :

1. *Portitional methods* : membangun berbagai partisi dan kemudian mengevaluasi partisi tersebut dengan beberapa kriteria. Yang termasuk dalam metode ini meliputi algoritma *K-Means*, *K-medoid*, PROCLUS, CLARA, CLARANS, dan PAM.
2. *Hierarchical methods* : membuat suatu penguraian secara hierarikal dari himpunan data dengan menggunakan beberapa kriteria. Metode ini teridiri atas dua macam, yaitu *Agglomerative* yang menggunakan strategi *top-down*. Metode ini meliputi algoritma BIRCH, AGNES, DIANA, CURE, dan CHAMALEON.
3. *Density-based methods* : suatu model dihoptesakan untuk masing-masing klaster dan ide untuk mencari *best fit* dari model tersebut untuk masing-masing yang lain. Metode klasterisasi ini meliputi pendekatan statistik, yaitu algoritma COBWEB dan jaringan syaraf tiruan, yaitu SOM.

6. Algoritma K-Means

Algoritma K-Means diperkenalkan oleh J.B. MacQueen pada Tahun 1976. K-means adalah salah satu algoritma *unsupervised* yang paling sederhana yang

dikenal dapat menyelesaikan permasalahan *clustering* dengan baik. Data di dalam *cluster* mempunyai ciri-ciri (fitur, karakteristik, atribut, properti) serupa dan tidak serupa dengan data pada cluster lain (Agusta, 2007).

K-means merupakan algoritma *clustering* yang berulang-ulang. Algoritma K-means dimulai dengan pemilihan secara acak K, K disini merupakan banyaknya cluster yang ingin dibentuk. Kemudian tetapkan nilai K secara random, untuk sementara nilai tersebut menjadi pusat dari *cluster* atau biasa disebut dengan centroid, mean atau “*means*”. Hitung jarak setiap data yang ada terhadap masing-masing centroid menggunakan rumus *Euclidean* hingga ditemukan jarak yang paling dekat dari setiap data dengan centroid. Klasifikasikan setiap data berdasarkan kedekatannya dengan centroid (Agusta, 2007).

Agusta, 2007 menyatakan beberapa *distance space* telah diimplementasikan dalam menghitung jarak (*distance*) antara data dan centroid termasuk diantaranya L_1 (*Manhattan/City Block distance space*), L_2 (*Euclidean distance space*), dan L_p (*Minkowski distance space*). Jarak antara dua titik x_1 dan x_2 pada *Manhattan/City Block Distance Space* dihitung dengan menggunakan rumus sebagai berikut

$$D_{L_1}(x_2, x_1) = \|x_2 - x_1\|_1 = \sum_{j=1}^p |x_{2j} - x_{1j}| \dots\dots\dots (1)$$

Dimana :

P : Dimensi data

$|\cdot|$: Nilai absolut

Sedangkan untuk L_2 (*Euclidean distance space*), jarak antara dua titik dihitung menggunakan rumus sebagai berikut :

$$D_{L_2}(x_2, x_1) = \|x_2 - x_1\|_2 = \sqrt{\sum_{j=1}^p (x_{2j} - x_{1j})^2} \dots\dots\dots(2)$$

Dimana :

P : dimensi data

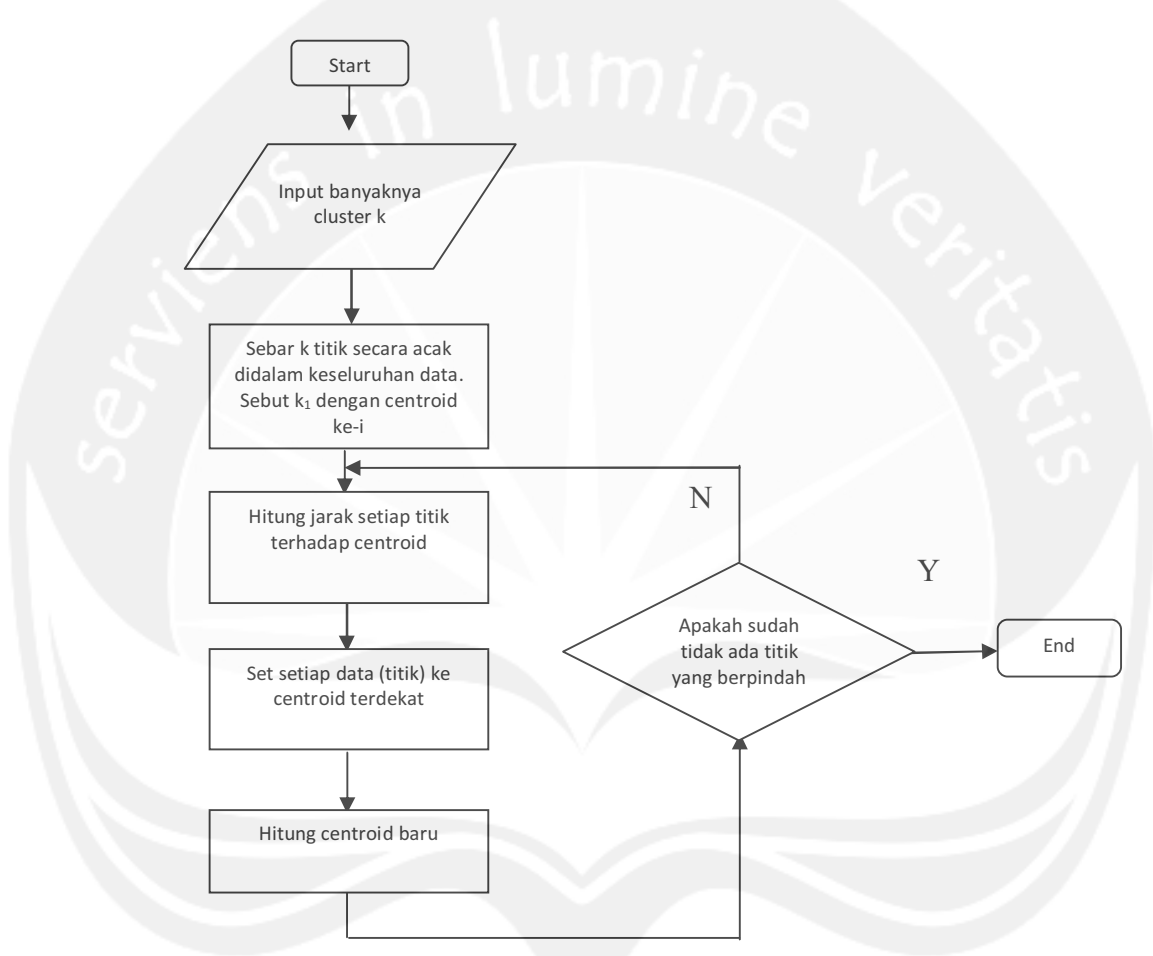
L_p (*Minkowski distance space*) yang merupakan generalisasi dari beberapa *distance space* yang ada seperti L_1 (*Manhattan/City Block*) dan L_2 (*Euclidean*), juga telah diimplementasikan. Tetapi secara umum *distance space* yang sering digunakan adalah *manhattan* dan *Euclidean*. *Euclidean* sering digunakan karena perhitungan jarak dalam *distance space* ini merupakan jarak terpendek yang bisa didapatkan antara dua titik yang diperhitungkan, sedangkan *Manhattan* sering digunakan karena kemampuannya dalam mendeteksi keadaan khusus seperti keberadaan outliers dengan lebih baik.

Widyawati (2010) menyatakan bahwa proses algoritma K-Means sebagai berikut :

- a. Pilih secara acak objek sebanyak k, objek-objek tersebut akan dipresentasikan sebagai *mean* pada cluster.
- b. Untuk setiap objek dimasukkan ke dalam cluster yang tingkat kemiripan objek terhadap *cluster* tersebut tinggi. Tingkat kemiripan ditentukan dengan jarak objek terhadap *mean* atau *centroid cluster* tersebut. Semakin dekat jarak objek tersebut dengan *mean* semakin mirip pula karakteristik objek dengan *mean*.
- c. Hitung nilai *centroid* yang baru pada masing-masing *cluster* yang terbentuk.

- d. Proses tersebut diulang hingga anggota pada kumpulan *cluster* tersebut tidak berubah.

Kurniawan dkk (2010) menyatakan proses algoritma k-means seperti dijabarkan dalam flowchart berikut :



Gambar 3. Cara kerja algoritma K-Means (Kurniawan dkk, 2010).

Saepulloh (2010) menyatakan penentuan *centroid* awal dilakukan secara *random*/acak dari data/objek yang tersedia sebanyak jumlah kluster k , kemudian untuk menghitung *centroid cluster* berikutnya ke i , v_i digunakan rumus sebagai berikut :

$$V_k = \frac{\sum_{i=1}^{N_k} X_i}{N_k} \dots\dots\dots(3)$$

V_k : *centroid* pada *cluster* ke k

X_i : Data ke i

N_k : Banyaknya objek/jumlah data yang menjadi anggota *cluster* ke k

Susanto dan Suryadi (2010) menyatakan bahwa range tidak dapat dijadikan pijakan yang kokoh untuk menilai keberagaman data. Oleh karena itu, ukuran varians yang menggunakan prinsip pencarian jarak antara setiap data dengan pusatnya (rata-rata) seringkali digunakan. Secara sistematis, hal tersebut dirumuskan menjadi :

$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \dots\dots\dots(4)$$

Dimana :

x = data ke- n

\bar{x} = x rata-rata = nilai rata-rata sampel

n = banyaknya data.