

BAB 1

PENGENALAN DATA MINING

1.1 Apa Itu Data Mining?

Dalam definisi sederhana, data mining merupakan ekstraksi informasi ataupun pola yang penting dan menarik dari data yang terdapat dalam sebuah database dalam jumlah besar. Data mining juga dapat diartikan sebagai suatu istilah yang digunakan untuk menguraikan pengetahuan di dalam database. Secara umum data mining terbagi atas 2(dua) kata yaitu:

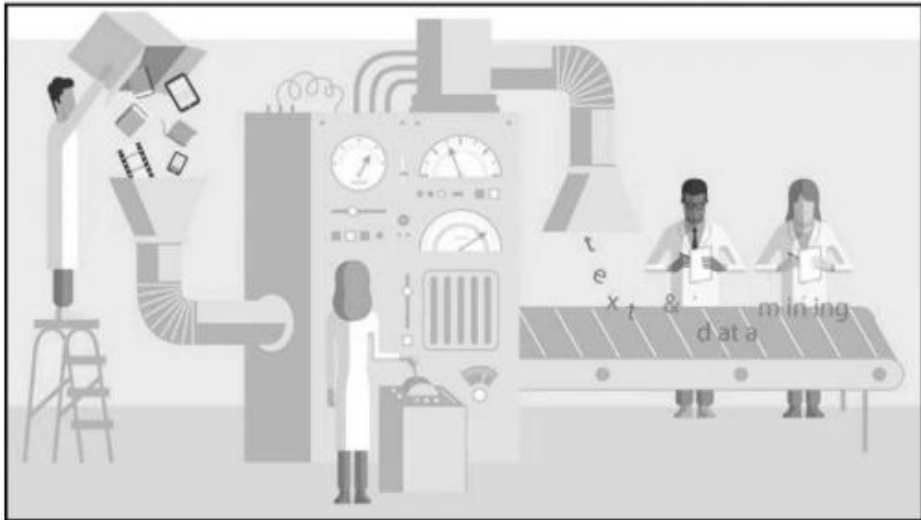
1. Data yaitu kumpulan fakta yang terekam atau sebuah entitas yang tidak memiliki arti dan selama ini terabaikan.
2. Mining yaitu proses penambangan.

BAB 1

Sehingga data mining dapat diartikan sebagai proses penambangan data yang menghasilkan sebuah output berupa pengetahuan. Berikut merupakan definisi data mining yang dikutip dari beberapa sumber, yaitu:

1. Data mining adalah analisis otomatis dari data yang berjumlah besar atau kompleks dengan tujuan untuk menemukan pola atau kecenderungan yang penting yang biasanya tidak disadari keberadaannya. (Pramudiono, 2006).
2. Data mining merupakan analisis dari peninjauan kumpulan data untuk menemukan hubungan yang tidak diduga dan meringkas data dengan cara berbeda dengan cara yang berbeda dengan sebelumnya, yang dapat dipahami dan bermanfaat bagi pemilik data. (Larose, 2005)
3. Data mining merupakan bidang dari beberapa bidang keilmuan yang menyatukan teknik dari pembelajaran mesin, pengenalan pola, statistik, database, dan visualisasi untuk penanganan permasalahan pengambilan informasi dari database yang besar (Larose, 2005)
4. Data mining adalah proses ekstraksi suatu data (sebelumnya tidak diketahui, bersifat implisit, dan dianggap tidak berguna) menjadi informasi atau pengetahuan atau pola dari data yang jumlahnya besar (Written, Ian H. Frank, 2011)
5. Menurut Gartner Group data mining adalah suatu proses menemukan hubungan yang berarti pola dan kecenderungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistik dan matematika (Larose, 2005).

BAB 1



Gambar 1.1 Ilustrasi Data Mining

Dari definisi yang telah disampaikan dapat disimpulkan bahwa data mining merupakan proses penambangan data dalam jumlah besar untuk memperoleh pengetahuan dengan menyatukan teknik dari berbagai bidang yaitu pembelajaran mesin, pengenalan pola, statistic, database, dan visualisasi informasi.

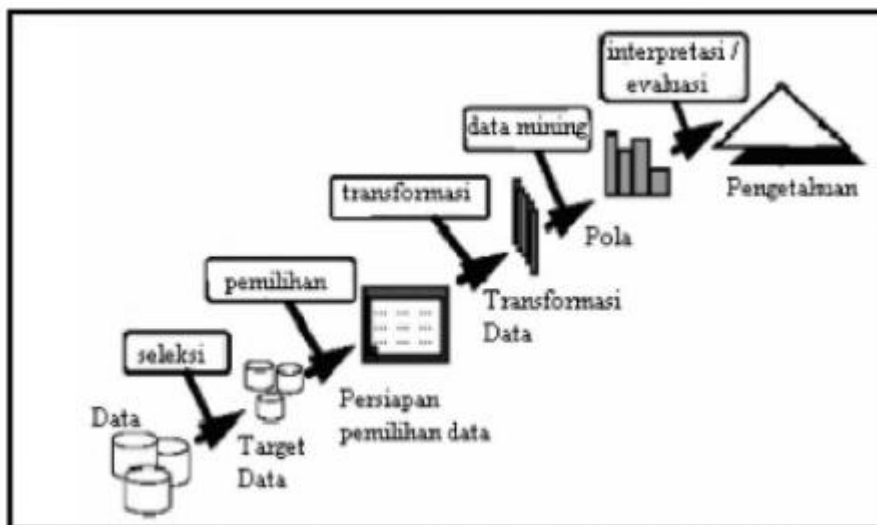
Data mining tidak hanya digunakan untuk melakukan penambangan data pada data transaksi saja. Penelitian di bidng data mining saat ini sudah merambah ke sistem database lanjut seperti *object oriented database*, *image/spatia database*, *time-series data/temporal database*, *text* (dikenal dengan nama *text mining*) dan *multimedia database*. Kemudian data mining juga memiliki karakteristik sebagai berikut:

- a. Data mining berhubungan dengan penemuan sesuatu yang tersembunyi dan pola data tertentu yang tidak diketahui sebelumnya.

BAB 1

- b. Data mining biasa menggunakan data yang sangat besar. Biasanya data tersebut digunakan untuk membuat hasil lebih dapat dipercaya.
- c. Data mining berguna untuk membuat keputusan kritis.

Dalam jurnal ilmiah data mining dikenal dengan istilah Knowledge Discovery in Database (KDD). Proses KDD secara garis besar dapat dijelaskan sebagai berikut (Fayyad, 1996).



Gambar 1.2 Proses Knowledge Discovery Database

1. Data Selection

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalan informasi saat KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses data mining disimpan dalam suatu berkas yang terpisah dari database operasional.

BAB 1

2. *Pre-processing/Cleaning*

Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi focus KDD. Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak. Selain itu juga dilakukan proses enrichment, yaitu proses memperkaya data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.

3. *Transformation*

Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses *coding* dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

4. *Data Mining*

Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

5. *Interpretation/Evaluation*

Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

BAB 1

Perluakah kita mempelajari data mining? Perlu, karena manusia banyak sekali menghasilkan data dalam jumlah besar dari berbagai bidang baik dalam bidang Bisnis, Kedokteran, Cuaca, Olahraga, Politik dan sebagainya. Dari bidang olahraga kita mengetahui berapa perolehan gol dari setiap tim dalam satu musim. Pada bidang bisnis khususnya saham, kita memperoleh data dari Bursa Efek Jakarta, kapan harga saham naik ataupun turun. Kemudian pada bidang cuaca kita memiliki data curah hujan, tingkat kelembaban, dan lain sebagainya yang diperoleh dari BMKG.

1.2 Sejarah Data Mining

Tahun 90-an merupakan awal kemunculan data mining. Data mining memang termasuk ke dalam salah satu cabang ilmu computer yang relative baru. Dan hingga saat ini orang-orang masih memperdebatkan data mining untuk ditempatkan di dalam bidang ilmu mana, karena data mining menyangkut berbagai aspek seperti database, kecerdasan buatan (*artificial intelligence*), statistic, dan sebagainya. Beberapa pihak ada yang berpendapat bahwa data mining tidak lebih dari machine learning atau analisa statistic yang berjalan di atas database saja. Namun beberapa pihak yang lain berpendapat bahawa database memiliki peranan penting dalam data mining karena data mining mengakses data yang ukurannya besar dan disini peran penting database terlihat terutama dalam proses optimisasi querynya. Data mining hadir dengan dilatarbelakangi terjadinya ledakan jumlah data yang dialami saat ini dimana banyak organisasi baik itu milik pemerintah maupun non-pemerintah yang telah mengumpulkan banyak data sekian tahun lamanya. Data tersebut hampir semuanya dimasukkan ke dalam komputer yang digunakan untuk menangani transaksi sehari-hari.

BAB 1

Bisa dibayangkan berapa transaksi perbankan yang terjadi dari sebuah bank dalam seharinya dan betapa besarnya data yang dimiliki oleh bank tersebut jika aktivitas tersebut telah berjalan beberapa tahun. Dari hal tersebut timbul sebuah pertanyaan, apakah data tersebut akan tersimpan begitu saja lalu pada akhirnya dibuang, ataukah kita dapat menggali informasi yang berguna untuk organisasi dari sekian banyak data tersebut.

Data mining melakukan eksplorasi pada basis data untuk melakukan pola-pola yang tersembunyi, mencari informasi yang mungkin saja terlupakan oleh pelaku bisnis karena hal tersebut diluar ekspektasi mereka. Sementara itu pelaku bisnis tersebut memiliki kebutuhan untuk memanfaatkan kumpulan data yang sudah dimiliki, melihat peluang tersebut para peneliti menciptakan sebuah teknologi baru yang dapat menjawab kebutuhan tersebut, yaitu data mining. Teknologi tersebut sekarang sudah banyak digunakan oleh berbagai perusahaan untuk memecahkan berbagai persoalan bisnis. Meningkatnya kebutuhan dunia bisnis yang selalu ingin memperoleh nilai tambah dari data yang telah dikumpulkan telah mendorong penerapan teknik analisis data yang berasal dari berbagai bidang seperti statistika, kecerdasan buatan, dan lain sebagainya untuk mengolah data berskala besar tersebut.

Berawal dari penerapan dalam dunia bisnis, kini data mining juga diterapkan pada bidang lain yang membutuhkan analisis data dalam skala besar seperti bioinformasi dan pertahanan negara.

1.3 Tujuan Data Mining

Penggunaan data mining tidak hanya sekedar hanya menggabungkan beberapa bidang untuk memperoleh suatu informasi saja, lebih dari itu

BAB 1

memiliki tujuan utama yaitu sebagaimana dijelaskan berikut ini. (Hoffer, Presscott, dan McFadden, 2007)

a. *Expalanatory*

Untuk menjelaskan beberapa kondisi yang terdapat di dalam penelitian, seperti mengapa penjualan handphone meningkat di suatu negara.

b. *Confirmatory*

Untuk mempertegas hipotesis, seperti halnya suatu promosi yang dilakukan pada social media lebih menarik perhatian banyak orang daripada promosi yang dilakukan pada media cetak.

c. *Exploratory*

Untuk menganalisis data yang memiliki hubungan yang baru. Misalnya, pola apa yang cocok diterapkan untuk strategi promosi penjualan.

1.4 Fungsi Data Mining

Teknik data mining tidak hanya digunakan untuk menemukan pola namun juga dapat digunakan untuk meperediksi tren masa kini. Selain itu data mining juga memiliki keuntungan yang kompetitif termasuk di dalamnya peningkatan pendapatan, berkurangnya pengeluaran, dan meningkatnya kemampuan pasar. Data mining dibagi menjadi dua kategori utama (Han dan Kamber, 2006:21-29) yaitu:

1. 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,

BAB 1

sedangkan atribut-atribut yang digunakan untuk membuat prediksi dikenal dengan istilah explanatory atau variable bebas.

2. Deskriptif

Tujuan dari tugas deskriptif adalah untuk menurunkan pola-pola (korelasi, trend, cluster, teritori, dan anomali) yang meringkas hubungan pokok dalam data. Tugas data mining deskriptif sering berupa penyelidikan dan seringkali memerlukan teknik *post-processing* untuk validasi dan penjelasan hasil.

Selain itu data mining juga memiliki beberapa fungsionalitas yaitu *Concept/Class Description: Characterization and Discrimination, Mining Frequent Patterns, Associations, and Correlations, Clasifiaction and Prediction, Cluster Analysis, Outlier analysis and Evolution analysis* (Han dan Kamber, 2006: 21-27)

Berikut ini merupakan penjelasan dari masing-masing fungsi yang telah disebutkan.

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

Data characterization adalah ringkasan dari semua karakteristik atau fitur dari data yang telah diperoleh dari target kelas 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. 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.

BAB 1

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

Frequent patterns adalah pola yang sering terjadi di dalam data. Frequent pattern memiliki jenis yang beragam 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 dalam sebuah kumpulan data transaksional.

Association analysis adalah pencarian aturan-aturan asosiasi yang menunjukkan kondisi-kondisi nilai atribut yang sering terjadi bersama dalam sekumpulan data. Association analysis 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. Model yang diturunkan tersebut dapat direpresentasikan dalam berbagai bentuk seperti if-then klasifikasi, decision tree, 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. Classification dapat direpresentasikan dalam bentuk pohon keputusan. Setiap node dalam pohon keputusan menyatakan suatu tes terhadap atribut dataset, sedangkan setiap cabang menyatakan hasil dari test tersebut. Pohon keputusan yang terbentuk dapat diterjemahkan menjadi sekumpulan aturan dalam bentuk IF condition THEN outcome. (Mewati Ayub, 2007: 7)

BAB 1

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. Sednagkan, clustering atau analisis cluster adalah proses pengelompokan satu set benda-benda fisik atau abstrak ke dalam kelas objek yang sama. Tujuannya adalah untuk menghasilkan pengelompokkan objek yang mirip satu sama lain dalam suatu kelompok. Semakin besar kemiripan objek dalam suatu cluster serta semakin besar pula perbedaan dalam suatu cluster maka kualitas analisis cluster tersebut semakin baik.

5. *Outlier Analysis*

Outlier merupakan objek data yang tidak mengikuti perilaku umum dari data. Outlier dianggap sebagai noise atau pengecualian. Analisis data outliers dapat dianggap sebagai noise atau pengecualian. Analisis data outlier dinamakan outlier mining. Teknik ini biasanya digunakan 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 krakterisasi, diskriminasi, asosiasi, klarifikasi, atau clustering dari data yang berkaitan dengan waktu

1.5 Tipe Data Pada Data Mining

Secara garis besar terdapat 2 (dua) tipe data ayang harus dipahami dalam data mining yaitu:

1. Numeric merupakan tipe data yang bisa di kalkulasi


BAB 1

2. Nominal merupakan tipe data yang tidak bisa di kalkulasi baik tambah, kurang, kali, maupun bagi.

Contoh pemanfaatan tipe data dapat terlihat pada tabel 1.1 berikut.

Tabel 1.1 Tipe Data Dalam Data Mining

No	NAMA	V1	V2	V3	Ket
1	Dini	0.25	73.6	79.3	Gagal
2	Dino	3.75	98.9	87	Lulus
3	Dina	3.85	99	85	Lulus
4	Dani	0.56	60.3	65	Gagal
5	Dana	3.15	95.7	84.3	Lulus
6	Danu	0.35	52.6	56	Gagal
7	Doni	1.72	68.3	73	Gagal
8	Dono	0.75	79.4	80	Gagal



1.6 Perkembangan Data Mining

Perkembangan awal dari data mining dimulai pada tahun 1763 ketika Thomas Bayes mempublikasikan Teorema Bayes. Teori ini sangat penting dalam data mining, karena memungkinkan estimasi suatu kejadian berdasarkan kejadian yang telah berlangsung. Pada tahun 1805 mulai berkembang teori regresi yang mempelajari hubungan antar variable. Regresi menjadi salah satu alat penting dalam data mining.

Penggunaan computer untuk mengolah data dalam jumlah besar dimulai ketika alan turing memperkenalkan ide mesin pengolah data yang bersifat universal pada tahun 1936. Tahun 1943 Warren McCulloch dan Walters Pitts menciptakan konsep dasar jaringan syaraf tiruan. Konsep

BAB 1

jaringan syaraf tiruan berusaha meniru cara kerja otak manusia dalam mengingat pola. Pengembangan system basis data yang pesat mulai tahun 1970 memungkinkan manusia untuk menyimpan dan mengelola data berukuran besar. Perkembangan itu diikuti pula oleh perkembangan berbagai algoritma untuk pengolahan data, misalnya algoritma genetika pada tahun 1975 dan *Support Vector Machines* (SVM) pada tahun 1992.

Perkembangan pesat data mining, baik dari segi perangkat keras maupun algoritma, memungkinkan implementasi data mining dalam berbagai bidang.

Kemajuan luar biasa yang terus berlanjut pada data mining yang didorong oleh beberapa faktor, antara lain (Larose,2005) :

1. Mempunyai pertumbuhan yang cepat dalam kumpulan data.
2. Penyimpanan data, sehingga seluruh perusahaan memiliki akses ke dalam database yang andal.
3. Peningkatan akses data melalui navigasi web dan intranet
4. Adanya tekanan kompetisi bisnis untuk meingkatkan penguasaan pasar dalam globalisasi ekonomi.
5. Perkembangan teknologi perangkat lunak untuk data mining (ketersediaan teknologi)
6. Perkembangan yang hebat dalam kemampuan komputasi dan pengembangan kapasitas media penyimpanan.

Hal penting yang terkait dengan data mining adalah :

1. Data mining adalah suatu proses otomatis terhadap data yang sudah ada.
2. Data yang akan diproses berupa data yang sangat besar
3. Tujuan data mining adalah mendapatkan hubungan atau pola yang mungkin memberikan indikasi yang bermanfaat.

BAB 1

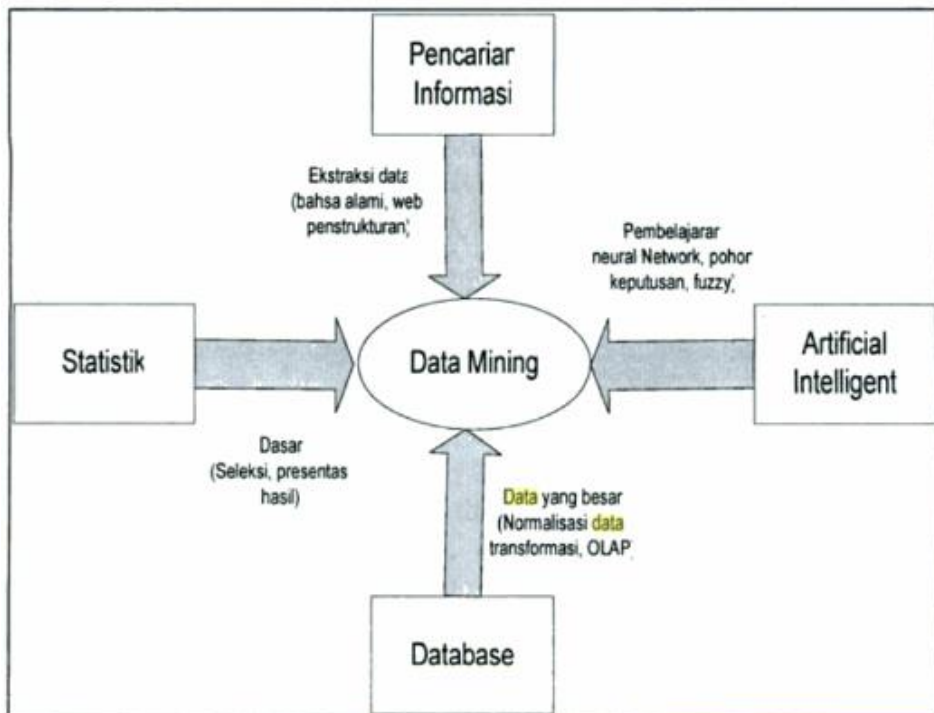
Hubungan yang dicari dalam data mining dapat berupa hubungan antara dua atau lebih dalam satu dimensi. Misalnya dalam dimensi produk dapat melihat keterkaitan pembelian suatu produk dengan produk yang lain. Selain itu, hubungan juga dapat dilihat antara dua atau lebih atribut dan dua atau lebih objek (Ponniah, 2001). Beberapa definisi awal dari data mining menyertakan fokus pada proses otomatisasi. Berry dan Linoff dalam buku *Data Mining Technique for Marketing, Sales, and Customers Support* mendefinisikan data mining sebagai suatu proses eksplorasi dan analisis secara otomatis maupun semiotomatis terhadap data dalam jumlah besar dengan tujuan menemukan pola atau aturan yang berarti (Larose, 2005).

Tiga tahun kemudian, dalam buku *mastering Data mining* memberikan definisi ulang terhadap pengertian data mining dan memberikan pernyataan bahwa jika ada yang kami sesalkan adalah frasa “secara otomatis maupun semiotomatis” karena kami merasa hal tersebut memberikan fokus lebih pada teknik otomatis dan kurang pada eksplorasi dan analisis. Hal tersebut memberikan pemahaman yang salah bahwa data mining merupakan produk yang dapat dibeli dibandingkan keilmuan yang harus dikuasai (Larose, 2005).

Dari pernyataan tersebut menegaskan bahwa dalam data mining otomatisasi tidak menggantikan campur tangan manusia. Manusia harus ikut aktif dalam setiap fase proses data mining. Kehebatan data mining yang terdapat saat ini memungkinkan terjadinya kesalahan penggunaan yang berakibat fatal. Pengguna mungkin menrapkan analisis yang tidak tepat terhadap kumpulan data dengan menggunakan pendekatan yang berbeda. Oleh karena itu, dibutuhkan pemahaman tentang statistik dan

BAB 1

struktur model matematika yang mendasari kerja perangkat lunak (Larose, 2005).



Gambar 1.3 Bidang Ilmu Data Mining

Salah satu kesulitan untuk mendefinisikan data mining adalah kenyataan bahwa data mining mewarisi banyak aspek dan teknik dari bidang-bidang ilmu yang sudah mapan terlebih dahulu. Gambar 1.3 menunjukkan bahwa data mining memiliki akar yang panjang dari bidang ilmu seperti kecerdasan buatan (artificial intelligent), machine learning, statistik, data base, dan juga information retrieval (Pramudiono, 2006).

BAB 1

1.7 Implementasi Data Mining

Data mining memiliki penerapan yang sangat luas di berbagai bidang, hal tersebut dikarenakan aktivitas yang dilakukan banyak menghasilkan data. Berikut ini merupakan contoh penerapan data mining dalam berbagai bidang.

1. Kesehatan

Pada bidang kesehatan data mining memiliki potensi yang besar untuk memperbaiki kesehatan. Analisis terhadap data praktik terbaik untuk meningkatkan perawatan serta meminimalkan biaya. Selain itu juga data mining diimplementasikan untuk diagnosis sebuah penyakit dan digunakan untuk memprediksi volume pasien dari setiap kategori. Selain itu data mining juga digunakan oleh perusahaan asuransi untuk menghindari kecurangan.

2. Analisis pasar

Analisis pasar digunakan untuk melakukan pemodelan dengan menarik kesimpulan bahwa jika seseorang membeli item tertentu, maka cenderung akan membeli item lainnya. Teknik ini memungkinkan seorang penjual memahami perilaku dari pelanggannya sehingga penataan item pada toko akan disesuaikan dengan hasil analisis tersebut. Selain itu analisis tersebut juga untuk perbandingan antara pelanggan dalam suatu kelompok demografis yang berbeda.

3. Pendidikan

Terdapat bidang baru yang muncul dalam Pendidikan, yaitu Educational Data Mining (EDM). Bidang tersebut berkaitan dengan metode pengembangan dengan menemukan pengetahuan baru yang berasal dari lingkungan Pendidikan. Tujuan dari EDM dapat

BAB 1

diidentifikasi sebagai perilaku belajar siswa di masa yang akan datang, mempelajari dampak dari dukungan pendidikan serta untuk memajukan pengetahuan ilmiah tentang pembelajaran.

4. Rekayasa Manufaktur

Dalam bidang manufaktur data mining dapat digunakan untuk menemukan proses manufaktu yang kompleks. Data mining dapat juga digunakan untuk membuat suatu rancangan system yang digunakan untuk mengekstrak hubungan antara asitektur produk, portofolio produk, data kebutuhan pelanggan, serta perkembangan produk.

5. CRM

Customer Relationship Management merupakan bagian yang memiliki tugas untuk mengakuisisi dan mempertahankan pelanggan, serta meningkatkan loyalitas pelanggan dan menerapkan strategi yang berfokus pada pelanggan. Untuk menjaga hubungan yang benara dengan pelanggan maka pelaku bisnis harus mengumpulkan data dan menganalisis informasi. Dengan menggunakan teknologi data mining, data yang telah dikumpulkan dapat digunakan untuk analisis guna menghasilkan informasi yang bermanfaat.

6. Fraud Detection (Deteksi Penipuan/Kecurangan)

Metode tradisional yang digunakan untuk melakukan deteksi terhadap kecurangan memakan waktu dan kompleks dalam pelaksanaannya.

Data mining berperan membantu dalam menemukan pola yang berarti serta mengubah data menjadi sebuah informasi, dimana setiap informasi yang valid dan berguna merupakan pengetahuan. Sebuah system dikatakan sempurna adalah sistem yang dapat melindungi semua

BAB 1

informasi pengguna. Metode yang mendapat pengawasan yaitu pengumpulan catatan sampel, karena catatan ini tergolong curang atau tidak palsu. Kemudian dibangun sebuah model dengan menggunakan data ini dan algoritma dibuat untuk mengidentifikasi apakah rekamannya itu salah atau tidak.

7. Intrusion detection

Gangguan dapat didefinisikan sebagai setiap tindakan yang membahayakan integritas dan kerahasiaan sumber daya. Langkah yang dilakukan untuk menghindari gangguan tersebut meliputi otentikasi pengguna, meminimalkan kesalahan pemrograman, dan perlindungan terhadap informasi. Data mining dapat membantu memperbaiki deteksi intrusi dengan menambah tingkat fokus terhadap deteksi anomaly. Hal tersebut membantu analisis untuk membedakan aktivitas yang terjadi sehari-hari pada jaringan. Data mining juga dapat berguna untuk membantu mengekstrak data yang lebih relevan dengan masalah yang ada.

8. Deteksi Kebohongan

Penegakan hukum bisa dilakukan dengan menggunakan teknik data mining untuk menyelidiki kejahatan, memantau komunikasi tersangka yang dianggap sebagai teroris. Hal ini termasuk ke dalam teks mining. Proses ini berjalan untuk menemukan pola yang berarti dalam data yang biasanya berupa teks yang tidak terstruktur. Hasil pengumpulan sampel data dari penelitian sebelumnya akan dibandingkan dengan model untuk melakukan deteksi terhadap kebohongan yang dilakukan dimana model yang dibuat diciptakan sesuai dengan kebutuhan.

BAB 1

9. Segmentasi Pelanggan

Penelitian terhadap suatu pasar dapat membantu untuk melakukan segmentasi pelanggan, namun dengan penggunaan data mining hal tersebut dapat dilakukan lebih mendalam serta dapat meningkatkan efektivitas pasar. Selain itu data mining juga dapat digunakan untuk menyelaraskan pelanggan menjadi segmen yang berbeda dan dapat melakukan penentuan kebutuhan berdasarkan pelanggan. Dalam buku ini akan dilakukan simulasi dalam implementasi data mining yaitu segmentasi data pelanggan pada sebuah perusahaan yang akan dijelaskan pada bab selanjutnya.

10. Perbankan/Keuangan

Komputerisasi data dalam jumlah besar pada perbankan pelanggan mempengaruhi bertambahnya data yang diperoleh dari setiap transaksi terutama transaksi baru. Dalam hal ini data mining memiliki kontribusi untuk memecahkan masalah bisnis di bidang perbankan dan keuangan dengan menemukan pola, sebab-akibat, serta korelasi dalam informasi bisnis dan harga pasar yang tidak jelas terlihat oleh manajer dikarenakan volume data yang terlalu besar atau dihasilkan terlalu cepat. Para ahli melakukan penyaringan dan pengolahan terhadap data sehingga para manajer dapat memperoleh informasi untuk segmentasi, penargetan, perolehan, penahanan, serta pemeliharaan pelanggan yang lebih baik

11. Pengawasan Perusahaan

Pengawasan perusahaan merupakan pemantauan terhadap perilaku seseorang atau kelompok oleh perusahaan. Data yang diperoleh biasanya sering digunakan untuk tujuan pemasaran atau dijual ke perusahaan lain, namun dibagi secara reguler dengan instansi pemerintah terkait. Hal

BAB 1

tersebut dapat digunakan oleh para pelaku bisnis untuk menyesuaikan produk mereka yang diharapkan oleh pelanggan. Data tersebut juga dapat digunakan untuk tujuan pemasaran langsung, seperti iklan bertarget pada mesin pencari Google, dimana iklan ditargetkan pada pengguna dengan menganalisis riwayat pencarian mereka.

12. Analisis Riset

Data mining memiliki peran yang sangat membantu dalam pembersihan data, pra-pengolahan data serta integrasi database. Penemuan yang didapat oleh peneliti dari data yang serupa dalam database dapat membawa perubahan dalam penelitian. Identifikasi dan korelasi antar aktivitas apapun dapat diketahui. Kemudian visualisasi data yang dilakukan oleh data mining dapat memberi gambaran yang jelas tentang data.

13. Investigasi Kriminal

Proses yang bertujuan untuk mengidentifikasi karakteristik kejahatan disebut dengan kriminologi. Analisis kejahatan meliputi peninjauan dan deteksi kejahatan serta hubungannya dengan penjahat. Volume dataset kejahatan yang tinggi juga kompleksitas hubungan antara data semacam ini menjadikan kriminologi sebagai bidang yang tepat untuk menerapkan teknik data mining. Hasil laporan kejahatan berbasis teks dapat diubah menjadi data pengolahan kata dimana informasi yang diperoleh bisa digunakan untuk melakukan proses pencocokan kejahatan.

14. Bioinformatika

Pendekatan data mining dalam bioinformatika membantu untuk mengekstrak pengetahuan yang berguna dari kumpulan data yang dikumpulkan dalam biologi serta bidang ilmu kehidupan lainnya yang

BAB 1

terkait seperti kedokteran dan ilmu saraf. Pengaplikasian data mining untuk bioinformatika meliputi penemuan gen, inferensi fungsi protein, diagnosis terhadap penyakit, optimasi pengobatan penyakit, rekonstruksi jaringan interkasi protein dan gen serta prediksi lokasi sub-seluler protein.

1.8 Pengelompokan Data Mining

Berdasarkan tugas yang dapat dilakukan, data mining dibagi ke dalam beberapa kelompok, yaitu (Larose, 2005):

1. Deskripsi

Peneliti dan analis terkadang secara sederhana ingin mencoba menemukan cara untuk menggambarkan pola serta kecenderungan yang terdapat dalam data. Misal, petugas pengumpulan suara mungkin tidak dapat menemukan keterangan atau fakta bahwa siapa yang tidak cukup professional akan sedikit didukung dalam pemilihan kepala daerah. Deskripsi dari pola dan kecenderungan sering memeberikan kemungkinan penjelasan terhadap suatu pola atau kecenderungan.

2. Estimasi

Estimasi memiliki kesamaan dengan klasifikasi, kecuali pada variable target estimasi lebih ke arah numerik daripada ke arah kategori. Sebuah model dibangun dengan menggunakan record lengkap yang menyediakan nilai dari variable target sebagai nilai prediksi. Selanjutnya pada tahap peninjauan berikutnya estimasi nilai dari variable target dibuat berdasarkan nilai dari variable prediksi. Misal, akan dilakukan estimasi tekanan dara sistolik pada seorang pasien di sebuah rumah sakit berdasarkan umur pasien, jenis kelamin, indeks berat badan, dan level sodium darah. Hubungan antara tekanan darah sitolik dan nilai variable

BAB 1

prediksi dalam proses pembelajaran akan menghasilkan model estimasi. Model estimasi yang dihasilkan dapat digunakan untuk kasus baru lainnya. Contoh lain yaitu estimasi nilai indeks prestasi kumulatif mahasiswa program pascasarjana berdasarkan nilai indeks prestasi mahasiswa tersebut pada saat mengikuti program sarjana.

3. Prediksi

Prediksi memiliki karakteristik yang hamper sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa mendatang.

Contoh prediksi dalam dunia bisnis dan penelitian adalah:

- Prediksi harga beras dalam tiga bulan yang akan datang.
- Prediksi persentase kenaikan kecelakaan lalu lintas tahun depan jika batas bawah kecepatan dinaikan.

Beberapa metode dan teknik yang digunakan dalam klasifikasi dan estimasi bisa juga digunakan (dalam keadaan yang tepat) untuk prediksi.

4. Klasifikasi

Pada klasifikasi, terdapat target variable kategori. Misal, penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang dan pendapatan rendah.

Contoh lain klasifikasi dalam dunia bisnis maupun penelitian adalah:

- Menentukan apakah suatu transaksi pada sebuah kartu kredit merupakan transaksi yang curang atau bukan.
- Memperkirakan apakah suatu pengajuan hipotek oleh nasabah merupakan suatu kredit yang baik atau buruk
- Mendiagnosis penyakit seorang pasien untuk mendapatkan informasi penyakit tersebut termasuk dalam kategori apa.

BAB 1

5. Pengklusteran

Pengklusteran bisa juga disebut dengan pengelompokan *record*, pengamatan, atau memperhatikan dan memebentuk kelas objek-objek yang memiliki kemiripan. Kluster adalah kumpulan *record* yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan dengan *record* dalam kluster lain.

Pengklusteran berbeda dengan klasifikasi yaitu tidak adanya variable target dalam pengklusteran. Dalam pengklusteran tidak dilakukan proses klasifikasi, estimasi, atau memprediksi nilai dari variable target. Namun demikian, algoritma pengklusteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan (homogen), yang mana kemiripan *record* dalam suatu kelompok akan bernilai maksimal, sedangkan kemiripan dengan record dalam kelompok lain akan bernilai minimal

Contoh pengkluteraan dalam dunia bisnis maupun penelitian adalah:

- Mendapatkan kelompok-kelompok pelanggan untuk target pemasaran dari suatu produk bagi perusahaan yang tidak memiliki dana pemasaran yang besar.
- Untuk tujuan audit akuntansi, yaitu melakukan pemisahan terhadap perilaku finansial dalam baik dan mencurigakan.
- Melakukan pengklusteran terhadap ekspresi dari gen, untuk mendapatkan kemiripan perilaku dari gen dalam jumlah besar.

6. Asosiasi

Dalam data mining, asosiasi memiliki tugas menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis leih umum disebut dengan *Market Basket Analysis* atau analisis keranjang belanja.

BAB 1

Contoh asosiasi dalam dunia bisnis maupun penelitian adalah:

- Meneliti jumlah pelanggan dari perusahaan telekomunikasi seluler yang diharapkan untuk memberikan respons positif terhadap penawaran peningkatan layanan yang diberikan.
- Menemukan barang dalam supermarket yang dibeli secara bersamaan dan barang yang tidak pernah dibeli secara bersamaan.

1.9 Cara Kerja Data Mining

Cara kerja data mining yaitu menggali hal-hal penting yang belum diketahui sebelumnya atau memprediksi apa yang akan terjadi. Teknik yang digunakan untuk melaksanakan tugas ini disebut pemodelan. Pemodelan adalah sebuah kegiatan untuk membangun sebuah model pada situasi yang telah diketahui “jawabannya” dan kemudian menerapkannya pada situasi lain yang akan dicari jawabannya.

Data Mining untuk menentukan pola-pola dalam data. Pola yang ditemukan harus penuh arti dan pola tersebut memberikan keuntungan. Karakteristik data mining sebagai berikut :

- a. Data mining berhubungan dengan penemuan sesuatu yang tersembunyi dan pola data tertentu yang tidak diketahui sebelumnya.
- b. Data mining biasa menggunakan data yang sangat besar. Biasanya data yang besar digunakan untuk membuat hasil lebih mudah dipercaya
- c. Data mining berguna untuk membuat keputusan yang kritis, terutama strategi (Davies,2004), juga dapat digunakan untuk pengambilan keputusan dimasa depan berdasarkan informasi yang diperoleh dari data masa lalu. Tergantung pada aplikasinya, data bisa

BAB 1

berupa data mahasiswa, pasien, dll. Banyak Kasus dalam sehari-hari yang dapat diselesaikan dengan data mining diantaranya :

- 1) Memprediksi berapa jumlah mahasiswa baru di perguruan tinggi berdasarkan data pendaftar pada tahun-tahun sebelumnya.
- 2) Memprediksi nilai IPK berdasarkan nilai IP setiap semester sebelumnya.
- 3) Produk apa yang akan dibeli pelanggan secara bersamaan jika membeli produk di swalayan

Tentu masih banyak lagi contoh-contoh dalam bidang lain atau kasus lain yang kaitannya dengan panggalian data sehingga bias menghasilkan pengetahuan baru dan informasi baru yang dapat menjadi strategi dalam mengembangkan suatu bidang usaha.

1.10 Algoritma Data Mining

Proses pemecahan masalah yang dilakukan saat pengolahan data dalam data mining yang bertujuan untuk menemukan sebuah pola yang tersembunyi dalam data tidak lepas dari sebuah algoritma. Penggunaan algoritma itu sendiri disesuaikan dengan melihat informasi apa yang ingin didapat serta data yang akan diolah menggunakan data mining. Berikut ini merupakan algoritma yang populer dan sering digunakan dalam data mining.

1. Klasifikasi.

Pada data mining proses klasifikasi bertujuan untuk mengelompokkan data menjadi beberapa kelompok. Proses pengelompokan data mengacu pada data yang telah diketahui terlebih

BAB 1

dahulu kelompok atau kelasnya. Data yang belum memiliki kelompok ditentukan kelompoknya melalui proses perbandingan dengan data yang sudah diketahui kelompoknya. Berikut merupakan algoritma klasifikasi populer.

- a. *Decision tree*
- b. *Naïve bayes*
- c. *K-nearest neighbor (KNN)*
- d. Jaringan syaraf tiruan
- e. Algoritma genetika
- f. *Support vector machine (SVM)*

2. *Clustering.*

Clustering memiliki tujuan yang sama dengan klasifikasi yaitu mengelompokkan data. Namun, proses pengelompokan data pada clustering tidak menggunakan data lain yang sudah diketahui kelompoknya sebagai perbandingan. Pengelompokan clustering berlangsung otonom dengan cara membandingkan semua data yang belum memiliki kelas dan membaginya kedalam beberapa kelas berdasarkan kemiripan anatara data. Beberapa algoritma clustering yang banyak digunakan adalah.

- a. *K-means*
- b. *Density-based spatial clustering of applications with noise (DBSCAN)*
- c. *Expectation-Maximization (EM)*
- d. *Fuzzy C-Means*
- e. *Hierarchical clustering*
- f. *Gaussian mixtures*

BAB 1

3. Regresi

Regresi berbeda dengan klasifikasi dan clustering yang bertujuan dalam mengelompokkan data. Regresi bertujuan untuk melakukan prediksi atau peramalan. Konsep dasar regresi pada data mining diturunkan dari teori statistika. Pada dasarnya, regresi berusaha mengidentifikasi relasi antar beberapa variable terikat dengan variable bebas. Selanjutnya model matematika yang telah dihasilkan dapat digunakan untuk memperkirakan nilai dari suatu variable terikat berdasarkan nilai variable bebasnya. Berikut ini beberapa algoritma regresi yang banyak digunakan adalah.

- a. Regresi linear sederhana
- b. Regresi linear berganda

4. *Association Rule*.

Association rule merupakan metode pencarian pola relasi antar data dalam sebuah kumpulan data. Berdasarkan pola tersebut, kemunculan suatu data dapat diprediksi berdasarkan kemunculan data lainnya. Algoritma *association rule* yang populer diantaranya adalah sebagai berikut.

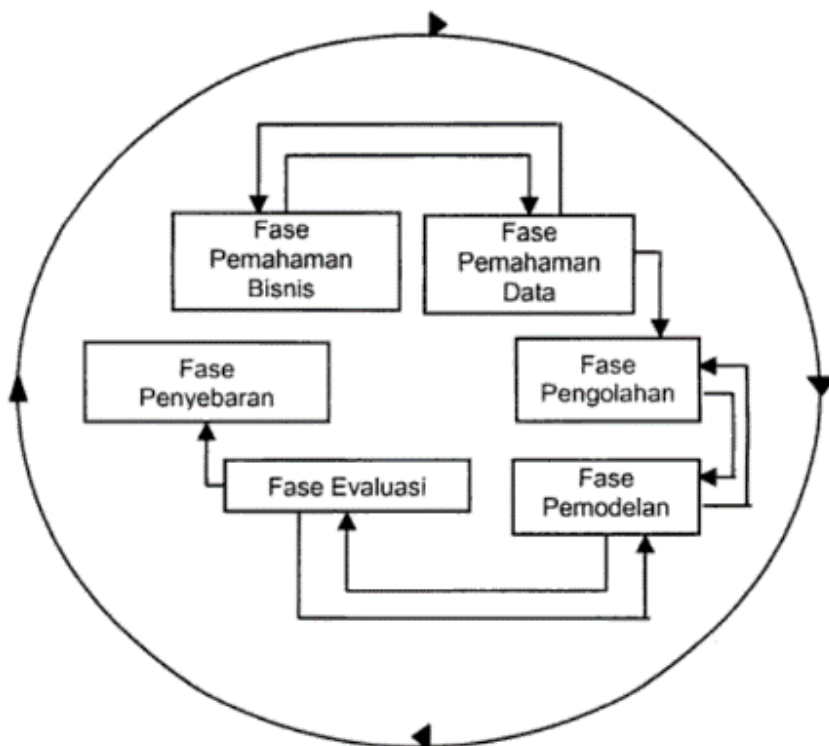
- a. Apriori
- b. Eclat
- c. *Frequent-pattern growth (FP-growth)*

1.11 CRISP-DM

Pada tahun 1996 analis dari beberapa industri seperti DaimlerChrysler, SPSS dan NCR mengembangkan Cross-Industry Standard Process for Data Mining (CRISP-DM). CRISP DM menyediakan standar proses data mining sebagai strategi pemecahan masalah secara umum dari bisnis atau unit penelitian. Dalam CRISP-DM, sebuah proyek

BAB 1

data mining memiliki siklus hidup yang terbagi dalam enam fase. Keseluruhan fase berurutan yang ada tersebut bersifat adaptif. Fase berikutnya dalam urutan bergantung kepada keluaran dari fase sebelumnya. Hubungan penting antarfase yang digambarkan oleh panah. Misalkan, jika proses berada fase modeling. Berdasar pada perilaku dan karakteristik model, proses mungkin harus kembali kepada fase data preparation untuk perbaikan lebih lanjut terhadap data atau berpindah maju kepada fase evaluation.



Gambar 1.4 Proses Data Mining menurut CRISP-DM

BAB 1

Terdapat enam fase CRISP-DM (Larose, 2005) :

1. Fase Pemahaman Bisnis (Business Understanding Phase)
 - a. Penentuan tujuan proyek dan kebutuhan secara detail dalam lingkup bisnis atau unit penelitian secara keseluruhan.
 - b. Menerjemahkan tujuan dan batasan menjadi formula dari permasalahan data mining.
 - c. Menyiapkan strategi awal untuk mencapai tujuan
2. Fase Pemahaman Data (Data Understanding Phase)
 - a. Mengumpulkan data
 - b. Menggunakan analisis penyelidikan data untuk mengenali lebih lanjut data dan pencarian pengetahuan awal
 - c. Mengevaluasi kualitas data
 - d. Jika diinginkan, pilih sebagian kecil group data yang mungkin mengandung pola dari permasalahan
3. Fase Pengolahan Data (Data Preparation Phase)
 - a. Siapkan data awal, kumpulan data yang akan digunakan untuk keseluruhan fase berikutnya. Fase ini merupakan pekerjaan perlu dilaksanakan secara intensif
 - b. Pilih kasus dan variabel yang akan dianalisis dan sesuai analisis yang akan dilakukan.
 - c. Lakukan perubahan pada beberapa variabel jika dibutuhkan.
 - d. Siapkan data awal sehingga siap untuk ke perangkat pemodelan.
4. Fase Pemodelan (Modeling Phase)
 - a. Pilih dan aplikasikan teknik pemodelan yang sesuai
 - b. Kalibrasi aturan model untuk mengoptimalkan hasil

BAB 1

- c. Perlu diperhatikan bahwa beberapa teknik mungkin untuk digunakan pada permasalahan data mining yang sama.
 - d. Jika diperlukan, proses dapat kembali ke fase pengolahan data untuk menjadikan data ke dalam bentuk yang sesuai dengan spesifikasi kebutuhan teknik data mining tertentu.
5. Fase Evaluasi (Evaluation Phase)
- a. Mengevaluasi satu atau lebih model yang digunakan dalam fase pemodelan untuk mendapatkan kualitas dan efektivitas sebelum disebarkan untuk digunakan.
 - b. Menetapkan apakah terdapat model yang memenuhi tujuan pada fase awal
 - c. Menentukan apakah terdapat permasalahan penting dari bisnis atau penelitian yang tidak tertangani dengan baik
 - d. Mengambil keputusan berkaitan dengan penggunaan hasil dari data mining.
6. Fase Penyebaran (Deployment Phase)
- a. Menggunakan model yang dihasilkan. Terbentuknya model tidak menandakan telah terselesaikannya proyek.
 - b. Pembuatan Laporan. Penerapan proses data mining secara paralel pada departemen lain

1.12 Tools Data Mining

Tools data mining digunakan untuk menunjang proses kerja dari data mining. Tools tersebut dibuat untuk mendefinisikan dan mencapai berbagai tujuan serta untuk membantu mendapatkan informasi yang lebih

BAB 1

terperinci. Berikut ini beberapa tools atau software yang digunakan dalam data mining.

1. Rapidminer



Gambar 1.5 Logo Rapidminer

Rapidminer merupakan software yang bersifat open source. Rapidminer merupakan salahsatu solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. Rapidminer menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik. Rapidminer merupakan software yang berdiri sendiri untuk analisis data dan sebagai mesin data mining yang dapat diintegrasikan pada produknya sendiri. Rapidminer ditulis dengan menggunakan bahasa pemrograman java sehingga dapat bekerja pada semua sistem operasi. Rapidminer memiliki beberapa sifat sebagai berikut.

1. Ditulis dengan bahasa pemrograman java sehingga dapat dijalankan di berbagai sistem operasi.
2. Proses penemuan pengetahuan dimodelkan sebagai operator trees.

BAB 1

3. Representasi XML internal untuk memastikan format standar pertukaran data.
4. Bahasa scripting memungkinkan untuk eksperimen skala besar dan otomatisasi eksperimen
5. Konsep multi-layer untuk menjamin tampilan data yang efisien dan menjamin penanganan data.
6. Memiliki GUI, command line mode, dan java API yang dapat dipanggil dari program lain.

Selain itu Rapidminer juga memiliki beberapa fitur diantaranya adalah sebagai berikut.

1. Banyaknya algoritma data mining, seperti decision tree dan self-organization map.
2. Bentuk grafis yang grafis , seperti tumpeng tindih diagram histogram, tree chart dan 3D Scatter plots.
3. Banyaknya variasi plugin, seperti text plugin untuk melakukan analisis teks.
4. Menyediakan prosedur data mining dan machine learning termasuk ETL (extraction, transformation, loading), data preprocessing, visualisasi, modelling dan evaluasi.
5. Proses data mining tersusun atas operator-operator yang nestable, dideskripsikan dengan XML, dan dibuat dengan GUI.
6. Mengintegrasikan proyek data mining Weka dan statistika R.

BAB 1

2. Weka



Gambar 1.6 Logo Weka

Weka merupakan software terintegrasi yang berisi implementasi dari metode-metode data mining. Weka dikembangkan oleh Universitas Waikato, Selandia Baru menggunakan bahasa pemrograman java. Oleh karena itu, Weka merupakan singkatan dari *Waikato Environment for Knowledge Analysis*. Dengan mengadopsi konsep open source software, menjadikan Weka dapat digunakan dan dimodifikasi siapapun secara gratis.

Weka memiliki keunggulan jika dibandingkan dengan perangkat lunak data mining lainnya. Penggunaan Weka murah karena aplikasi tersebut berlisensi *GNU General Public License*, yang artinya dapat

BAB 1

digunakan secara gratis. Penggunaan bahasa pemrograman java dalam pengembangan Weka menyebabkan Weka dapat diinstal pada hampir semua sistem operasi, sepanjang sistem operasi tersebut mendukung *Java Virtual Machine*. Berbagai macam algoritma data mining, mulai dari pemrosesan awal sampai dengan pemodelan data, telah disertakan dalam Weka sehingga memudahkan pengguna dalam menganalisis data. Apabila algoritma yang akan digunakan tidak tersedia pada Weka, pengguna dapat menambahkan algoritma tersebut melalui bahasa pemrograman java. Penggunaan Weka pun tergolong mudah karena telah dibekali dengan antarmuka grafis (*Graphical User Interface*) sehingga pengguna dapat menggunakannya tanpa perlu menulis satu baris kode pun.

3. R



Gambar 1.7 Logo R

BAB 1

R adalah nama bahas pemrograman computer yang ditujukan secara khusus untuk menangani komputasi statistic dan memudahkan penyajian grafik. Bahasa ini diciptakan oleh Ross Ihaka dan Robert Gentleman di Universitas Auckland, Selandia Baru. Versi pertama dirilis pada tahun 1995. Adapun nama R disematkan berdasarkan nama depan kedua penciptanya.

Bahasa R tergolong sebagai bahasa skrip, yakni bahasa yang memungkinkan perintah-perintah yang digunakan ditulis dalam skrip, berekstensi .r, yang disimpan dalam bentuk berkas teks. Selain itudengan menggunakan konsol, dimungkinkan untuk memberikan perintah secara interaktif. Dalam hal ini, begitu perintah diberikan dan tombol enter ditekan, perintah tersebut akan dieksekusi dan hasilnya akan segera terlihat. R menggunakan lisensi *GNU General Public License* sehingga dapat digunakan atau bebas dipakai oleh siapa saja selain itu bahasa R juga dapat berjalan hampir di semua sistem operasi baik itu Windows, Linux, maupun Mac OS X.

Bahasa R sangat cocok digunakan untuk menangani operasi yang melibatkan vektor dan matriks. Dengan bahasa R, operasi vektor dan matriks dapat dikerjakan dengan perintah yang sangat singkat. Dengan demikian, R dapat digunakan sebagai alternative terhadap bahasa MATLAB maupun Octave. Aplikasi utama R adalah untuk menangani komputasi statistic dan memudahkan dalam penyajian grafik. Namun , dengan paket-paket tambahan yang juga bersifat gratis, R dapat digunakan untuk menangani pengolahan citra (*image processing*), pembelajaran mesin (*machine learning*), amupun data besar (*big data*).

BAB 2

SEGMENTASI DAN *PROFILING* PELANGGAN

3.1 Segmentasi Pelanggan

Segmentasi terus menjadi konsep pemasaran yang penting juga dalam konteks relationship marketing. Meningkatkan hubungan dengan pelanggan menjadi lebih menarik dan akan menghasilkan pemahaman yang lebih baik tentang kebutuhan pelanggan. Segmentasi adalah proses membagi pelanggan menjadi beberapa klaster dengan kategori loyalitas pelanggan untuk membangun strategi pemasaran. Segmentasi pelanggan adalah salah satu langkah awal dalam membuat model bisnis.

BAB 2

Don Pepper dan Martha Roger dalam bukunya “*Managing Customer Relationship : A Strategic Framework*”. Melakukan kategori pelanggan sebagai berikut:

1. *Most Valuble Customer* (MVC), yaitu pelanggan dengan nilai paling tinggi bagi perusahaan. Merupakan pelanggan yang memberikan keuntungan terbesar bagi perusahaan.
2. *Most Groable Customer*, yaitu pelanggan yang tanpa disadari memiliki potensi besar.
3. *Below Zeros*, yaitu pelanggan yang memberikan keuntungan lebih sedikit daripada biaya untuk memberikan pelayanan.
4. *Migrators*, yaitu pelanggan yang berada pada posisi diantara *below zeros* dan *most growable customer* dan perlu dianalisis agar dapat diketahui kategori asalnya.

Segmentasi pelanggan dapat didefinisikan sebagai pembagian basis pelanggan menjadi kelompok yang berbeda dan konsisten secara internal dengan karakteristik serupa dimana memungkinkan perusahaan untuk mengembangkan strategi pemasaran yang berbeda sesuai dengan karakteristik pelanggan. Pemahaman seperti itu akan membantu perusahaan dalam mempertahankan pelanggan dan menciptakan nilai tambah bagi pelanggan itu melalui pengembangan hubungan pelanggan. Akibat dari manajemen hubungan pelanggan (Customer Relationship Management - CRM).

Segmentasi pelanggan juga mempresentasikan elemen kunci dalam identifikasi pelanggan dalam customer relationship management. (Ngai dkk, 2009). Customer relationship management berguna untuk meningkatkan hubungan dengan pelanggan, memfokuskan dalam hal

BAB 2

mengintegrasikan nilai, harapan, dan perilaku pelanggan dengan melakukan analisa data dari transaksi pelanggan (Peppard, 2000). Untuk mencapai tujuan customer relationship management , maka biasanya perusahaan memanfaatkan teknologi informasi untuk membantu perusahaan dalam mengatur hubungan pelanggan dengan suatu cara yang sistematis untuk meningkatkan loyalitas pelanggan dan meningkatkan keuntungan bisnis secara menyeluruh (Kalakota & Robinson, 1999).

Karakteristik pelanggan dapat direpresentasikan oleh beberapa kategori variabel yang terkait dengan pengelompokan, seperti berikut ini:

- Demographics: Umur, jenis kelamin, besarnya keluarga, besarnya kediaman, siklus kehidupan keluarga, pemasukan, pekerjaan atau profesi, pendidikan, kepemilikan rumah, status sosial ekonomi, agama, kewarganegaraan.
- Psychographics: kepribadian, gaya hidup, nilai-nilai, sikap.
- Behaviour: manfaat yang dicari, status pembelian, tingkat penggunaan produk, frekuensi pembelian.
- Geographic: negara, provinsi, kota, kode pos, iklim.

Skema segmentasi yang berbeda dapat dikembangkan menurut tujuan bisnis yang spesifik dari organisasi. Segmentasi umumnya digunakan melalui riset data pasar untuk mendapatkan wawasan tentang sikap pelanggan, keinginan, pandangan, preferensi, dan opini tentang perusahaan dan kompetisi. Segmentasi pelanggan berdasarkan pada riset pasar dan demografi seringkali membutuhkan pemahaman karakteristik semua pelanggan agar lebih efektif mengetahui segmen apa yang menjadi menarik pelanggan. Penggalan data dapat mengembangkan segmentasi pelanggan yang juga mengidentifikasi segmentasi pada perilaku

BAB 2

pelanggan. Selain data penelitian eksternal atau pasar, data transaksi dan pembayaran pelanggan juga dapat digunakan untuk mendapatkan wawasan tentang perilaku pelanggan. Segmentasi dengan cara tersebut, dapat mengalokasikan pelanggan untuk membentuk kelompok berdasarkan jumlah pengeluaran mereka. Hal ini dapat digunakan untuk mengidentifikasi pelanggan yang bernilai tinggi dan memprioritaskan pelayanan. Karakter dari pelanggan dijelaskan pada tabel 2.1.

Tabel 2.1 Karakteristik Pelanggan

Kelas Pelanggan	Karakteristik
Superstar	<ul style="list-style-type: none">a. Pelanggan dengan loyalti yang tinggi.b. Mempunyai nilai monetary yang tinggi.c. Mempunyai frekuensi yang tinggi.d. Mempunyai transaksi paling tinggi.
Golden Customer	<ul style="list-style-type: none">a. Mempunyai nilai monetary tertinggi yang ke dua.b. Frequency yang tinggi.c. Mempunyai rata-rata transaksi.
Typical Customer	Mempunyai rata-rata nilai monetary dan rata-rata transaksi.

BAB 2

Occasional customer	<ul style="list-style-type: none">a. Nilai monetary terendah kedua setelah dormant customerb. Nilai recency paling rendahc. Transaksi paling tinggi
Everyday shopper	<ul style="list-style-type: none">a. Memiliki peningkatan transaksib. Transaksi yang rendahc. Mempunyai nilai monetary sedang sampai dengan rendah.
Dormant customer	<ul style="list-style-type: none">a. Mempunyai frequency dan monetary yang paling rendahb. Nilai recency yang paling rendah

3.2 *Profiling Pelanggan*

Customer Profiling merupakan langkah yang dilakukan untuk memetakan dan mendalami profil pelanggan atau pelanggan dengan lebih baik. Pemetaan profil pelanggan dapat dilakukan dengan kombinasi data eksplisit (informasi mengenai pelanggan yang didapatkan dari proses pendaftaran dan survei) dan data implisit (informasi perilaku pelanggan yang didapatkan dengan pengamatan langsung).

Istilah pelanggan diartikan sebagai dua jenis pelanggan, yaitu: pelanggan individu dan pelanggan organisasi. Pelanggan individu membeli barang dan jasa untuk digunakan sendiri, maupun oleh anggota keluarga yang lain. Pelanggan individu sering disebut pelanggan akhir karena langsung digunakan oleh individunya. Sedangkan pelanggan

BAB 2

organisasi meliputi organisasi bisnis, yayasan, kantor, dan lembaga lainnya. Jenis pelanggan organisasi membeli produk dan jasa untuk menjalankan kegiatan organisasinya (tidak dikonsumsi sendiri). Perilaku pelanggan adalah tindakan yang langsung terlibat dalam mendapatkan, mengonsumsi, dan menghabiskan produk atau jasa, termasuk proses keputusan yang mendahului dan menyusuli tindakan ini.

Kebutuhan dan keinginan pelanggan selalu menjadi perhatian utama bagi pemilik usaha, yaitu dengan selalu memperhatikan perilaku pelanggannya. Oleh sebab itu, suatu perusahaan dituntut untuk selalu memperhatikan perilaku pelanggan dan menyesuaikan pengenalan produknya kepada pelanggan dengan mengadakan penyempurnaan dan perbaikan terhadap produknya serta menyesuaikan kembali kebutuhan mereka untuk saat ini maupun kebutuhan masa depan. Berikut ini merupakan definisi perilaku pelanggan menurut beberapa ahli.

1. Perilaku Pelanggan (consumer behavior) didefinisikan sebagai studi tentang unit pembelian (Buying units) dan proses pertukaran yang melibatkan perolehan, konsumsi, dan pembuangan barang, jasa, pengalaman, serta ide-ide. Proses pertukaran merupakan unsur mendasar dari perilaku pelanggan. Pertukaran terjadi antara pelanggan dengan perusahaan, disamping itu juga terjadi di antara perusahaan pada situasi pembelian industrial. Akhirnya, pertukaran juga terjadi diantara pelanggan sendiri, seperti pada saat tetangga meminjam secangkir gula atau mesin pemotong rumput (John C. Mowen and Minor, 2002).
2. Perilaku pelanggan sebagai tindakan yang langsung terlibat dalam mendapatkan, mengonsumsi, dan menghabiskan produk dan jasa,

BAB 2

termasuk proses keputusan yang mendahului dan menyusuli tindakan ini. (Engel, et al 2008).

3. Perilaku pelanggan menurut Shiffman adalah perilaku yang ditunjukkan dalam mencari, membeli, menggunakan, menilai dan menentukan produk jasa dan gagasan. Sedangkan menurut Philip perilaku pelanggan adalah Bidang ilmu perilaku pelanggan mempelajari bagaimana individu, kelompok dan organisasi memilih, memakai serta memanfaatkan barang, jasa, gagasan atau pengalaman dalam rangka memuaskan kebutuhan dan Hasrat mereka.
4. Menurut Carl McDaniel perilaku pelanggan menggambarkan bagaimana pelanggan membuat keputusan pembelian dan bagaimana mereka menggunakan serta mengatur pembelian barang atau jasa. Dari beberapa pengertian diatas disimpulkan bahwa setiap pelanggan dalam membeli produk mempunyai perilaku yang berbeda antara satu dengan yang lain.
5. Perilaku pelanggan adalah perilaku dari pelanggan akhir, individu dan rumah tangga, yang membeli barang dan jasa untuk konsumsi pribadi. Faktor-faktor yang mempengaruhi perilaku pelanggan adalah kebudayaan, sosial, pribadi, psikologis (Kotler dan Keller, 2007).

Berdasarkan pendapat para ahli tersebut, maka dapat disimpulkan bahwa perilaku pelanggan adalah tindakan yang dilakukan oleh individu, kelompok, atau organisasi yang secara langsung terlibat atau berhubungan dengan proses pengambilan keputusan yang meliputi tindakan mengevaluasi, mendapatkan, dan mengkonsumsi produk, baik barang maupun jasa yang dapat dipengaruhi lingkungan, termasuk sebelum dan sesudah proses pengambilan keputusan pembelian. Perilaku pelanggan

BAB 2

adalah dinamis, berarti bahwa perilaku seorang pelanggan, group pelanggan, ataupun masyarakat luas selalu berubah dan bergerak sepanjang waktu. Perilaku pelanggan melibatkan pertukaran. Itu merupakan hal terakhir yang ditekankan dalam definisi perilaku pelanggan, yaitu pertukaran di antara individu. Hal tersebut membuat definisi dari perilaku pelanggan tetap konsisten dengan definisi pemasaran yang sejauh ini juga menekankan pertukaran. Namun pada kenyataannya, peran pemasaran adalah untuk menciptakan pertukaran dengan pelanggan melalui formulasi dan penerapan strategi pemasaran.

3.3 Perlunya Mempelajari Perilaku Pelanggan

Kajian atau studi tentang perilaku pelanggan yang dilakukan oleh para ahli menyimpulkan bahwa mempelajari perilaku dari pelanggan itu harus dilakukan dikarenakan akan memiliki dampak yang dapat membantu para pelaku bisnis untuk melakukan hal berikut ini.

1. Merancang bauran pemasaran
2. Menetapkan segmentasi
3. Merumuskan positioning dan pembedaan produk
4. Memformulasikan analisis lingkungan bisnisnya
5. Mengembangkan riset pemasaran

Selain itu, analisis perilaku pelanggan juga memiliki peranan penting dalam merancang kebijakan publik. Bagi orang yang memiliki peranan penting pada suatu negara, kajian ini diperlukan untuk merumuskan kebijakannya dalam kerangka perlindungan pelanggan. Dengan mengetahui perilaku pelanggan mungkin dapat dimanfaatkan untuk

BAB 2

kepentingan pengembangan kemampuan seorang pelaku bisnis dalam menjalankan tugasnya.

3.4 Jenis-Jenis Pelanggan

Keputusan pelanggan untuk pembelian dan mengonsumsi suatu produk sangat dipengaruhi oleh berbagai faktor. Sebagai seorang individu, konsumsi suatu produk akan dipengaruhi oleh persepsi, proses pembelajaran dan memori, motivasi dan nilai, konsep diri, sikap, kepribadian dan gaya hidup. Sebagai pengambil keputusan, hal ini akan tergantung dari tipe keputusan (rutin atau jarang), situasi pembelian yang dihadapi, kelompok atau orang yang mempengaruhi dan menjadi acuan. Selanjutnya, kebudayaan dan subbudaya juga memiliki pengaruh kepada perilaku pelanggan. Pembahasan lengkap dari topik-topik di atas akan Saudara temukan pada modul-modul berikutnya beserta contoh-contoh untuk memudahkan saudara memahaminya.

Kata pelanggan (consumer) lebih umum menjelaskan setiap orang yang terlibat dengan suatu kegiatan, seperti yang tercantum pada definisi perilaku pelanggan di atas, yaitu mengevaluasi, memperoleh, menggunakan, dan membuang barang atau jasa. Dengan demikian, pelanggan terkait dengan hubungannya dengan perusahaan tertentu, sedangkan pelanggan tidak.

Pelanggan umum merupakan seseorang yang memiliki kebutuhan atau dorongan, melakukan pembelian, selanjutnya membuang produk dalam 3 tahap proses konsumsi (Solomon, 2002). Pelanggan memiliki beberapa peran dalam ketiga proses tersebut, yaitu berikut ini.

1. Pencetus ide (initiator).

BAB 2

2. Pembeli (Purchaser/Buyer).
3. Membayar (Payer).
4. Pengguna/pemakai (User).
5. Pemberi pengaruh (Influencer).
6. Pengambil keputusan (decision maker).
7. Pelanggan organisasi atau kelompok, di mana satu orang atau sekelompok orang akan membuat keputusan untuk organisasi

3.5 Macam-Macam Model Perilaku Pelanggan

Menurut model perilaku pelanggan yang dikemukakan oleh Henry Assael (1998) terdapat beberapa faktor yang mempengaruhi perilaku pelanggan. Dengan model perilaku pelanggan yang sederhana Henry Assael menunjukkan bahwa interaksi antara pemasar dengan pelanggan perlu dilakukan karena dapat menimbulkan adanya proses untuk merasakan dan mengevaluasi informasi merek produk, mempertimbangkan berbagai alternatif merek dapat memenuhi kebutuhan pelanggan dan pada akhirnya memutuskan merek apa yang akan dibeli pelanggan. Model perilaku pelanggan adalah suatu gambar atau kerangka yang mencerminkan atau menjelaskan tahap demi tahap yang akan dilakukan oleh pelanggan dalam memutuskan untuk melakukan keputusan pembelian.

3.6 Faktor-Faktor yang Mempengaruhi Perilaku Pelanggan

Terdapat banyak faktor yang mampu mempengaruhi perilaku pelanggan dalam melakukan sebuah pembelian. Faktor-faktor tersebut

BAB 2

berawal dari dalam diri pelanggan serta dari luar pelanggan. Keputusan pembelian dari pembeli sangat dipengaruhi oleh faktor kebudayaan, sosial, pribadi dan psikologi dari pembeli. Sebagian besar adalah faktor-faktor yang tidak dapat dikendalikan oleh pemasar, tetapi harus benar-benar diperhitungkan. Faktor-Faktor yang Mempengaruhi Perilaku Pelanggan

1. Faktor Budaya

Menurut Sumarwan (2004) budaya adalah segala nilai, pemikiran, simbol yang mempengaruhi perilaku, sikap, kepercayaan dan kebiasaan seseorang dan masyarakat. Adapun unsur-unsur budaya antara lain budaya, subbudaya dan kelas sosial.

2. Subbudaya

Setiap kebudayaan terdiri dari subbudaya-subbudaya yang lebih kecil yang memberikan identifikasi dan sosialisasi yang lebih spesifik untuk para anggotanya. Subbudaya dapat dibedakan menjadi empat jenis: kelompok nasionalisme, kelompok keagamaan, kelompok, ras, dan area geografis.

3. Kelas social

Kelas-kelas social adalah kelompok yang relative homogen dan bertahan lama dalam suatu masyarakat, yang tersusun secara hierarki dan yang keanggotaannya mempunyai nilai, minat, dan perilaku yang serupa

4. Faktor Sosial

- Kelompok Referensi

Kelompok referensi seseorang terdiri dari seluruh kelompok yang mempunyai pengaruh langsung maupun tidak langsung terhadap sikap atau perilaku seseorang. Beberapa diantaranya adalah

BAB 2

kelompok primer yaitu kelompok yang memiliki interaksi yang cukup berkesinambungan seperti, keluarga, teman, tetangga, dan teman sejawat. Kelompok sekunder merupakan kelompok yang cenderung lebih resmi dan yang mana interaksi yang terjadi kurang berkesinambungan. Kelompok yang seseorang ingin menjadi anggotanya disebut kelompok aspirasi. Kemudian kelompok diasosiatif merupakan sebuah kelompok yang nilai atau perilakunya tidak disukai oleh individu.

- Keluarga

Pengaruh Keluarga yaitu keluarga memberikan pengaruh yang besar dalam perilaku pembelian. Para pelaku pasar telah memeriksa peran dan pengaruh suami, istri, dan anak dalam pembelian produk yang berbeda. Anak-anak sebagai contoh, memberikan pengaruh yang besar dalam keputusan yang melibatkan restoran fast food. Faktor sosial terdiri dari kelompok acuan, keluarga, peran dan status (Setiadi, 2003). Dalam kehidupan pembeli keluarga dibedakan menjadi dua yaitu.

- 1) Keluarga orientasi

Merupakan orangtua seseorang. Dari orangtualah seseorang mendapatkan pandangan tentang agama, politik, ekonomi, dan merasakan ambisi pribadi nilai atau harga diri dan cinta.

- 2) Keluarga prokreasi

Merupakan pasangan hidup serta anak-anak dari seseorang, keluarga ini merupakan organisasi pembeli dari seorang pelanggan yang paling penting dalam suatu masyarakat dan telah diteliti secara intensif.

BAB 2

- Peran dan Status

Setiap orang umumnya berpartisipasi dalam kelompok selama hidupnya, baik itu keluarga, organisasi ataupun klub. Posisi seseorang dalam setiap kelompok dapat diidentifikasi dalam peran dan status.

5. Faktor Pribadi

Keputusan pembelian juga dipengaruhi oleh karakteristik pribadi, antara lain sebagai berikut

- 1) Umur dan tahapan dalam siklus hidup

Konsumsi yang dilakukan oleh seseorang juga dipengaruhi oleh tahapan siklus hidup keluarga. Beberapa penelitian juga pernah melakukan identifikasi terhadap tahapan-tahapan dalam siklus hidup psikologis. Orang-orang dewasa biasanya mengalami perubahan atau transformasi tertentu pada saat mereka menjalani hidupnya.

- 2) Pekerjaan

Para pelaku bisnis mengidentifikasi kelompok-kelompok pekerja yang memiliki minat di atas rata-rata terhadap produk dan jasa tertentu.

- 3) Kepribadian dan konsep diri

Dalam hal ini yang dimaksud kepribadian adalah karakteristik psikologis yang berbeda dan setiap orang yang memandang responsnya terhadap lingkungan yang relative konsisten. Kepribadian merupakan suatu unsur yang sangat berguna dalam hal analisis perilaku pelanggan. Bila keragaman dari kepribadian dapat diklasifikasikan dan

BAB 2

memiliki korelasi yang kuat antara jenis-jenis kepribadian tersebut dan berbagai pilihan [rpduk atau merek.

4) Situasi ekonomi

Keadaan atau situasi ekonomi yang dimaksud dari seseorang terdiri dari pendapatan yang dapat dibelanjakan, tabungan dan hartanya, kemampuan untuk meminjam dan sikap terhadap mengeluarkan lawan menabung.

5) Gaya hidup.

Merupakan pola hidup seseorang di dunia yang diekspresikan dengan kegiatan, minat dan pendapat seseorang. Gaya hidup menggambarkan seseorang secara utuh yang berinteraksi dengan lingkungannya. Gaya hidup juga dapat mencerminkan sesuatu di balik kelas sosial seseorang

6. Faktor Psikologis

Pilihan pembelian seseorang dipengaruhi oleh empat faktor psikologis utama yaitu

1) Motivasi

Manusia memiliki beberapa kebutuhan diantaranya bersifat biogenic, kebutuhan biogenic yaitu kebutuhan yang timbul dari suatu keadaan biologis tertentu, seperti rasa lapar, haus, resah tidak nyaman. Adapun kebutuhan lain bersifat psikogenik, yaitu kebutuhan yang timbul dari keadaan psikologis tertentu, seperti kebutuhan untuk diakui oleh orang lain ataupun sebagainya. Untuk mencukupi kebutuhan

BAB 2

tersebut diperlukan lah sebuah motivasi dimana hal tersebut dapat mendorong seseorang untuk melakukan sesuatu.

2) Persepsi

Persepsi dapat diartikan sebagai proses dimana seseorang memilih, mengorganisasikan, mengartikan masukan informasi untuk menciptakan suatu gambaran yang berarti dari dunia ini.

3) Pembelajaran

Pembelajaran biasa juga disebut dengan proses belajar dapat menjelaskan perubahan dalam perilaku seseorang yang timbul diakibatkan dari pengalaman.

4) Keyakinan dan sikap.

Keyakinan dan sikap merupakan salahsatu faktor yang penting, karena keyakinan merupakan sebuah gagasan deskriptif yang dimiliki seseorang terhadap sesuatu.

Keputusan pembelian terhadap sebuah produk yang dilakukan oleh seseorang memiliki keterkaitan yang cukup rumit dan dipengaruhi oleh berbagai faktor yang telah disebutkan sebelumnya. Namun faktor-faktor tersebut sangat berguna dalam hal mengidentifikasi pembeli atau pelanggan yang mungkin memiliki minat terhadap suatu produk.

BAB 3

PENGENALAN ALGORITMA

3.1 Apa itu Algoritma?

Dalam pekerjaan sehari-hari, sangat diperlukan untuk memiliki pengetahuan serta pemahaman terlebih dahulu terkait dengan pekerjaan yang akan dilakukan agar tujuan, proses serta hasil dari pekerjaan tersebut dapat terlaksana. Demikian pula dalam pemrograman komputer, seorang programmer harus memiliki pemahaman yang cukup tentang logika serta konsep dari pemrograman, agar dalam pelaksanaannya program yang dibuat dapat memenuhi kebutuhan yang ada. Sebuah program tidak akan terlepas dari kinerja sebuah komputer, dimana dalam hal tersebut terdapat logika dasar dari setiap siklus yang ada dalam komputer seperti *input*,

BAB 3

process, hingga *output*. Algoritma merupakan sebuah pemahaman dasar yang harus dimiliki, karena dengan algoritma ini dapat membantu menata tahapan-tahapan dalam penyelesaian masalah menggunakan komputer, dalam hal ini membuat program.

Algoritma berasal dari nama seorang penulis dan juga seorang ahli matematika yang bernama Abu Ja'far Muhammad Ibnu Musa al-Khawarizmi yang mana oleh orang barat kata al-Khawarizmi dibaca *algorism* yang kemudian lambat laun berubah menjadi *algorithm* atau diterjemahkan dalam bahasa Indonesia menjadi algoritma. Salahsatu karya terkenal dari al-Khawarizmi yaitu *al-Kitab al-mukhtasar fi hisab al-jabr wa'l-muqabala (rules of restoration and reduction* atau buku rangkuman untuk kalkulasi dengan melengkapi dan menyeimbangkan).

Algoritma dapat didefinisikan sebagai langkah-langkah atau urutan untuk memecahkan suatu masalah berdasarkan tahapan logis secara sistematis dalam periode tertentu. Dalam literatur yang lain, definisi dari algoritma adalah langkah-langkah perhitungan dasar untuk mengubah suatu inputan menjadi keluaran.

3.2 Kriteria Algoritma

Menurut Donald E. Kurth sebuah algoritma yang baik harus memiliki kriteria sebagai berikut.

1. *Input*

Sebuah algoritma harus memiliki sebuah baik itu masukan dari pengguna ataupun data yang diinisialisasikan atau dibangkitkan dalam suatu algoritma.

BAB 3

2. *Output*

Dalam suatu algoritma harus memiliki satu atau lebih *output*. Algoritma yang tidak memiliki output merupakan sebuah algoritma yang sia-sia untuk dilakukan. Karena tujuan dibuatnya algoritma adalah untuk mendapatkan sebuah *output*.

3. *Finiteness*

Algoritma yang dijalankan harus memiliki sebuah jaminan untuk berhenti setelah melakukan suatu proses atau setelah output yang diinginkan tercapai.

4. *Definiteness*

Setiap pernyataan yang terdapat dalam suatu algoritma harus pasti dan tidak memiliki makna ganda sehingga tidak membingungkan pembaca dari algoritma tersebut, sehingga diharapkan mampu untuk memberikan hasil atau output sesuai dengan yang diharapkan.

5. *Effectiveness*

Sebuah algoritma sebisa mungkin harus dilaksanakan dalam jangka waktu yang wajar serta efektif, sehingga segala aktivitas yang dilakukan merupakan aktivitas yang memiliki dampak dan tidak ada aktivitas yang sia-sia dalam proses pengerjaannya.

3.3 Jenis-Jenis Proses Algoritma

Secara umum, algoritma dibedakan menjadi beberapa jenis yaitu sebagai berikut.

BAB 3

1. Algoritma Sekuensial (*Sequential*)

Algoritma sekuensial adalah langkah-langkah pemecahan suatu masalah yang dilakukan sesuai dengan penulisannya. Jika salahsatu langkah atau urutan nya dirubah maka kemungkinan akan mempengaruhi output yang dihasilkan.

2. Algoritma Percabangan (*Branching*)

Dalam suatu algoritma tertentu, sebuah aksi terkadang akan dilakukan jika terdapat kondisi yang terpenuhi ataupun tidak dilakukan tergantung dari situasi yang terjadi. Pada algoritma percabangan ini, hanya aka nada satu aksi yang dijalankan dari sejumlah pilihan aksi yang diberikan.

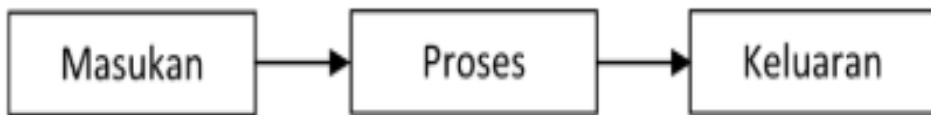
3. Algoritma Perulangan (*Looping*)

Dalam algoritma dikenal istilah perulangan. Perulangan yang dimaksud adalah satu atau beberapa aksi yang dijalankan secara berulang sesuai dengan kondisi dan kebutuhan.

3.4 Prinsip Kerja Algoritma

Berdasarkan definisi, algoritma merupakan sebuah deskripsi dari pelaksanaan dalam suatu proses, dimana proses yang dikerjakan akan sesuai dengan algoritma yang ditulis atau dibuat. Prinsip kerja dari sebuah algoritma yaitu adanya suatu masukan (*input*) yang kemudian diproses hingga menghasilkan sebuah keluaran (*output*) seperti yang digambarkan pada Gambar 3.1.

BAB 3



Gambar 3.1 Prinsip Kerja Algoritma

3.5 *Clustering*

Clustering merupakan salahsatu teknik yang digunakan untuk mengelompokan data. *Clustering* adalah sebuah proses yang digunakan untuk mengelompokan data menjadi beberapa kelompok atau *cluster* sehingga memiliki tingkat kemiripan yang tinggi terhadap anggota yang lain dalam satu kelompok serta memiliki kemiripan yang minimum dengan cluster yang lain (Tan, 2006). Menurut Rui Xu dan Donald (2009), *clustering* merupakan peroses membagi kumpulan data menjadi beberapa kelompok dimana anggota dari masing-masing kelompok memiliki kesamaan. Gagasan terkait dengan penglompokan data atau *clustering*, memiliki sifat yang sederhana dan cukup dekat dengan cara berpikir manusia. Selain itu, sebagian besar dari data yang diperoleh dalam jumlah besar terkadang memiliki banyak masalah yang dapat dilihat dari beberapa sifat yang melekat serta mengalami proses pengelompokan-pengelompokan secara natural (Hammuda dan Karay, 2003).

Clustering dapat digunakan sebagai langkah awal atau pendahuluan dalam sebuah proses pengumpuland ata. Dengan adanya *cluster-cluster* yang dihasilkan tersebut dapat digunakan untuk masukan lebih lanjut dalam suatu teknik yang berbeda, seperti natural yang disebutkan sebelumnya dapat diperoleh sebagai jarak dari pembaharuan formula Lance-Williams (Lance dan Williams, 1967).

BAB 3

Analisis kluster atau *clustering* merupakan proses membagi data pada suatu himpunan ke dalam beberapa kelompok atau *cluster* yang memiliki kesamaan data dalam suatu kelompok yang memiliki kesamaan yang lebih besar daripada kesamaan data tersebut dengan data yang berada pada kelompok yang lain (Jang, Sun, dan Mizutani, 2004). Analisis kluster juga dapat dikatakan sebagai teknik multivariat dengan tujuan utama yaitu untuk mengelompokkan objek berdasarkan karakteristik yang dimiliki dari masing-masing objek. Analisis kluster mengklasifikasi objek sehingga setiap objek yang memiliki kesamaan paling dekat dengan objek lainnya yang berada pada cluster yang sama. Solusi yang diperoleh dari hasil analisis kluster bersifat tidak unik, anggota dari masing-masing *cluster* dari masing-masing penyelesaian bergantung pada beberapa elemen prosedur serta beberapa solusi yang berbeda dapat diperoleh dengan mengubah salahsatu elemen atau lebih. Secara keseluruhan, solusi untuk analisis kluster bergantung pada variable-variabel yang dijadikan sebagai dasar untuk menilai kesamaan tersebut. Pengurangan atau penambahan dari masing-masing variable yang relevan dapat mempengaruhi hasil dari *clustering*.

Clustering merupakan suatu proses membagi kumpulan objek data ke dalam suatu himpunan yang biasa disebut dengan cluster. Objek yang berada di dalam cluster tersebut memiliki kesamaan antara satu dengan yang lainnya dan memiliki perbedaan dengan anggota kelompok dari *cluster* lain. Proses pembagian data dilakukan dengan menggunakan sebuah algoritma *clustering*. Maka dari itu, hal tersebut sangat berguna serta dapat menemukan kelompok atau *cluster* yang belum dikenal dalam data. *Clustering* banyak diimplementasikan pada berbagai aplikasi seperti

BAB 3

business intelligence, pengenalan pola citra, *web search*, bidang ilmu biologi, serta untuk keamanan (*security*).

Dalam *business intelligence*, *clustering* dapat digunakan untuk membagi *customer* atau pelanggan ke dalam banyak kelompok sesuai dengan karakteristik yang diharapkan dari masing-masing kelompok. Selain itu, *clustering* dikenal sebagai data segmentasi dikarenakan *clustering* dapat melakukan pembagian data terhadap banyak data set ke dalam banyak kelompok atau *cluster* berdasarkan kemiripan dari masing-masing data. Kemudian *clustering* juga dikenal sebagai *outlier detection*.

3.6 Manfaat *Clustering*

Teknik *clustering* memiliki beberapa manfaat dalam penggunaannya. Manfaat tersebut diantaranya adalah sebagai berikut.

1. *Clustering* adalah metode segmentasi data yang sangat berguna dalam melakukan prediksi dan analisis terhadap masalah bisnis tertentu (Berson dan Smith, 2001). Contohnya segmentasi pasar dan pelanggan, *marketing* dan pemetaan pada zonasi wilayah.
2. *Clustering* juga berguna untuk melakukan identifikasi obyek dalam berbagai bidang seperti *computer vision* dan *image processing*.

3.7 Konsep Dasar *Clustering*

Hasil yang baik dari sebuah *clustering* akan memiliki tingkat kesamaan yang tinggi dalam suatu kelompok atau *cluster* dan memiliki tingkat kesamaan yang rendah dengan anggota *cluster* lain. Tingkat kesamaan yang dimaksud tersebut merupakan hasil pengukuran secara numerik terhadap dua objek. Hasil kesamaan antar dua objek akan

BAB 3

memiliki nilai yang semakin tinggi jika kedua objek tersebut dibandingkan. Begitupun sebaliknya. Kualitas yang dimiliki dari hasil *clustering* bergantung pada metode yang digunakan. Terdapat 4 tipe data yang dikenal dalam *clustering*. Adapun keempat tipe data tersebut adalah sebagai berikut.

1. Variable berskala interval
2. Variable biner
3. Variable nominal, ordinal, dan rasio
4. Variable dengan tipe lainnya.

3.8 Syarat *Clustering*

Algoritma *clustering* memiliki syarat sekaligus tantangan yang harus dipenuhi. Syarat tersebut adalah sebagai berikut (Han dan Kamber, 2012).

1. Skalabilitas

Sebuah algoritma clustering harus memiliki kemampuan untuk mengolah data dalam jumlah yang besar. Penggunaan data dalam jumlah besar pada saat ini sudah sangat biasa digunakan di berbagai bidang contohnya saja database. Database ini tidak hanya berisi ratusan objek, melainkan berisi lebih dari jutaan objek.

2. Kemampuan analisa beragam bentuk data

Selain memiliki kemampuan skalabilitas yang tinggi, algoritma *clustering* juga harus mampu untuk dapat diaplikasikan dalam berbagai bentuk data, seperti pada tipe data yang telah disebutkan sebelumnya yaitu nominal, ordinal maupun gabungannya.

BAB 3

3. Menemukan *cluster* dengan bentuk yang tidak terduga

Pada umumnya algoritma *clustering* yang menggunakan metode *Euclidian* atau *Manhattan* memiliki hasil berbentuk bulat. Namun sebenarnya, hasil clustering dapat memiliki bentuk yang aneh dan tidak sama antara satu dengan yang lain. Oleh karena itu, dibutuhkan kemampuan untuk menganalisa *cluster* dalam berbagai bentuk pada suatu algoritma *clustering*.

4. Kemampuan untuk dapat menangani noise

Data yang diperoleh dari suatu pengumpulan data tidak selalu dalam keadaan yang baik. Terkadang terdapat beberapa data dalam kondisi rusak, sulit dimengerti ataupun hilang. Dengan adanya sistem inilah, suatu algoritma *clustering* dituntut untuk dapat menangani data yang rusak tersebut.

5. Sensitifitas terhadap perubahan input

Perubahan serta penambahan data yang terjadi pada sebuah masukan dapat mengakibatkan terjadinya perubahan pada cluster yang telah ada bahkan dapat menyebabkan perubahan yang sangat mencolok jika menggunakan algoritma *clustering* yang memiliki tingkat sensitifitas yang rendah.

6. Mampu melakukan *clustering* untuk data dimensi tinggi

Dimensi atau atribut yang dimiliki dari sekumpulan data dapat berbeda-beda. Oleh karena itu sangat diperlukan sebuah algoritma *clustering* yang dapat menangani data dengan perbedaan jumlah dimensi yang tidak sedikit.

BAB 3

7. Interpretasi dan kegunaan

Hasil dari proses pengolahan data menggunakan algoritma *clustering* harus dapat diinterpretasikan dan memiliki informasi yang berguna untuk penelitian.

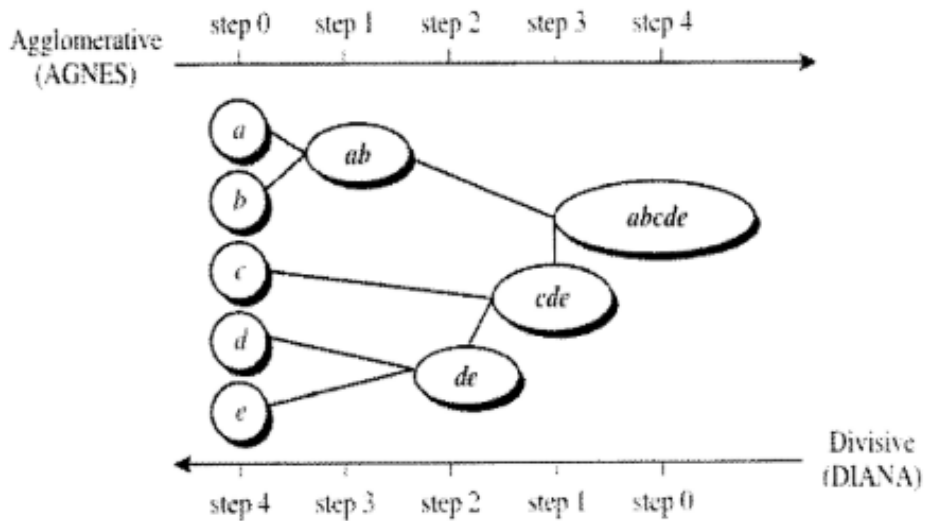
3.9 Metode *Clustering*

Secara umum metode *clustering* dapat dibagi menjadi dua yaitu *hierarchical clustering* dan *partitional clustering* (Tan, 2011). Kemudian sebagai tambahan terdapat juga metode Density-Based dan Grid-Based yang sering juga diterapkan dalam pengimplemetasian *clustering*.

3.9.1 *Hierarchical Clustering*

Dalam metode *hierarchical clustering* pengelompokan data dilakukan menggunakan suatu bagan yang berbentuk hirarki. Pada bagan tersebut terdapat penggabungan dua kelompok yang terdekat pada setiap iterasinya ataupun pembagian dari seluruh data set ke dalam suatu *cluster* seperti yang terlihat pada Gambar 3.2.

BAB 3



Gambar 3.2 *Hierarchical Clustering*

Berikut ini tahapan-tahapan dalam melakukan *Hierarchical Clustering*.

1. Identifikasi *item* yang memiliki jarak terdekat.
2. Gabungkan *item* tersebut kedalam suatu *cluster*.
3. Hitung jarak antar *cluster*.
4. Ulangi tahapan tersebut dari awal hingga semua terhubung.

Contoh metode *hierarchical clustering* diantaranya yaitu sebagai berikut.

1. *Single Linkage*.
2. *Complete Linkage*.
3. *Average Linkage*.
4. *Average Group Linkage*.

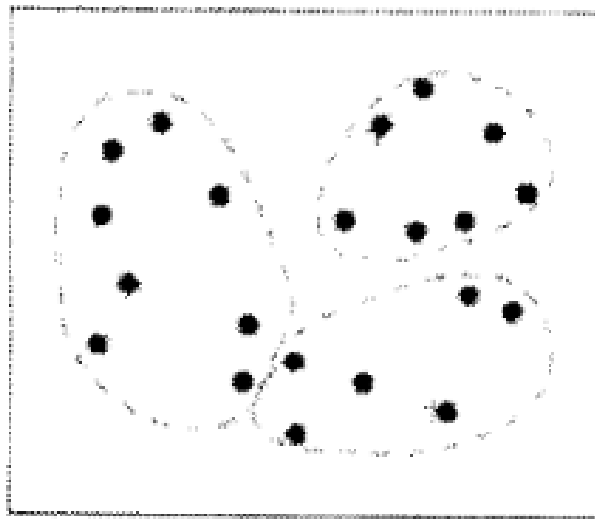
BAB 3

3.9.2 *Partitional Clustering*

Partitional clustering merupakan pengelompokan data kedalam sejumlah *cluster* tanpa adanya struktur hirarki antara satu dengan yang lainnya. Dalam metode ini setiap *cluster* memiliki titik pusat cluster atau biasa disebut dengan *centroid*. Kemudian secara umum metode ini memiliki sebuah fungsi dan tujuan yaitu meminimumkan jarak atau *dissimilarity* dari keseluruhan data terhadap pusat *cluster* masing-masing.

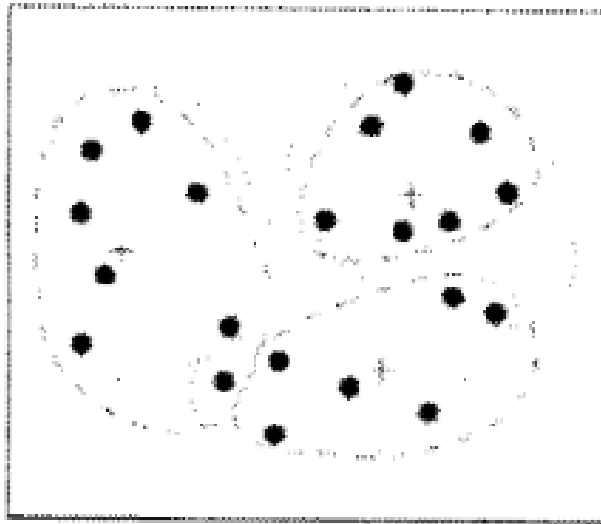
Contoh metode *partitional clustering* diantaranya yaitu sebagai berikut.

1. *K-Means*.
2. *Fuzzy K-Means*.
3. *Mixture Modelling*.

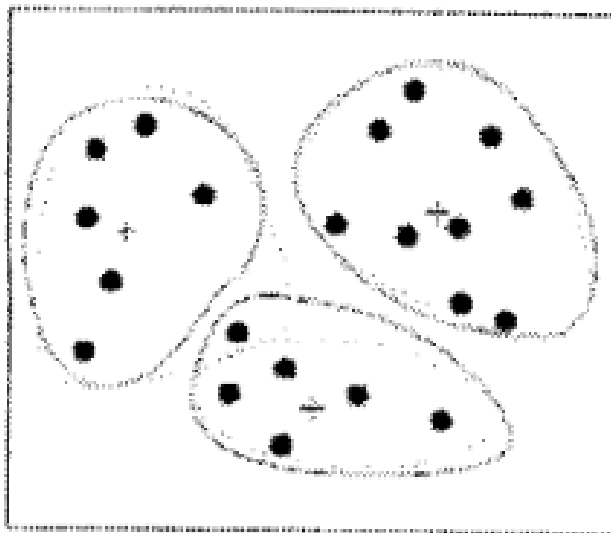


Gambar 3.3 *Cluster Awal*

BAB 3



Gambar 3.4 Proses Iterasi



Gambar 3.5 *Cluster Akhir*

BAB 3

3.10 Algoritma K-Means

Algoritma K-Means merupakan algoritma clustering yang paling sederhana dan umum digunakan. Hal tersebut dikarenakan kemampuan yang dimiliki K-Means yaitu dapat mengelompokkan data dalam jumlah yang besar dengan waktu komputasi yang dibutuhkan relatif cepat dan efisien. K-means merupakan salahsatu algoritma *clustering* dengan metode partisi atau *partitioning method* yang memisahkan data ke dalam kelompok yang berbeda. Dengan *partitioning* secara iteratif, K-Means mampu meminimalkan rata-rata jarak setiap data ke *cluster* nya. Algoritma K-Means pertama kali diusulkan oleh MacQueen pada tahun 1967 kemudian dikembangkan oleh Hartigan dan Wong pada tahun 1975 dengan tujuan algoritma K-Means dapat digunakan dapat membagi M *data point* dalam N dimensi kedalam sejumlah k *cluster* dimana proses *clustering* tersebut dilakukan dengan meminimalkan jarak *sum squares* antara data dengan masing-masing pusat *cluster* atau *centroid-based*.

Dalam penerapannya algoritma K-Means memerlukan tiga parameter yang keseluruhan nya ditentukan pengguna yaitu jumlah *cluster* k , inisialisasi *cluster* dan jarak sistem. K-Means biasanya dijalankan secara independent dengan inisialisasi yang berbeda sehingga menghasilkan *cluster* akhir yang berbeda karena algoritma ini pada prinsipnya hanya mengelompokkan data menuju *local minimal*. Salah satu cara yang dapat dilakukan untuk mengatasi *local minimal* adalah dengan menerapkan algoritma K-Means, untuk jumlah K yang diberikan dengan memberikan beberapa nilai *initial partition* yang berbeda kemudians elanjutnya dipilih partisi dengan nilai kesalahan kuadrat terkecil (Jain, 2009).

BAB 3

Algoritma K-Means merupakan model *centroid*. Model *centroid* adalah model yang menggunakan centroid untuk membuat *cluster*. *Centroid* adalah titik tengah suatu *cluster*. *Centroid* berupa nilai. *Centroid* digunakan untuk menghitung jarak suatu objek data terhadap *centroid*. Suatu objek data termasuk dalam *cluster* jika memiliki jarak terpendek terhadap centroid *cluster* tersebut. Selain itu algoritma K-means memiliki aturan dalam proses *clustering* yaitu sebagai berikut.

1. Berapa jumlah *cluster* yang perlu dimasukkan.
2. Hanya memiliki atribut bertipe numerik.

K-Means mempunyai kemampuan mengelompokkan data dalam jumlah yang cukup besar dengan waktu komputasi yang relatif cepat dan efisien. Namun, K-Means 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.

Secara detail teknik ini menggunakan ukuran ketidakmiripan untuk mengelompokkan obyek. Ketidakmiripan dapat diterjemahkan dalam konsep jarak. Dua obyek dikatakan mirip jika jarak dua objek tersebut dekat. Semakin tinggi nilai jarak, semakin tinggi nilai ketidakmiripannya. Tahapan awal yang dilakukan pada proses pengelompokan data dengan menggunakan algoritma K-Means adalah pembentukan titik awal centroid c_j . Pada umumnya pembentukan titik awal centroid dibangkitkan secara acak. Jumlah centroid c_j yang dibangkitkan sesuai dengan jumlah klaster yang ditentukan di awal. Setelah k centroid terbentuk kemudian dihitung jarak tiap data x_i dengan centroid ke- j sampai k , dinotasikan dengan $d(x_i, c_j)$. Terdapat beberapa ukuran jarak yang digunakan sebagai ukuran

BAB 3

kemiripan suatu instance data, salah satunya adalah jarak Euclidean. Perhitungan jarak Euclidean seperti pada Persamaan 3.1.

$$d_{ij} = \sqrt{\sum_{k=1}^p (X_{ik} - C_{jk})^2} \dots\dots\dots(3.1)$$

Duran dan Odell (1974) menyatakan jika semakin kecil, kesamaan antara dua $d(X_i, C_j)$ unit pengamatan semakin dekat. Syarat menggunakan jarak Euclid adalah jika semua fitur dalam dataset tidak saling berkorelasi. Jika terdapat fitur yang berkorelasi maka menggunakan konsep jarak Mahalanobis. Agusta (2007) menyatakan kelanjutan dari jarak tersebut dicari yang terdekat sehingga data akan mengelompok berdasarkan centroid yang paling dekat. Tahap berikutnya adalah update titik centroid dengan menghitung rata-rata jarak seluruh data terhadap centroid. Selanjutnya akan kembali lagi ke proses awal. Iterasi ini akan diulangi terus sampai didapatkan centroid yang konstan artinya titik centroid sudah tidak berubah lagi. Atau iterasi dihentikan berdasarkan jumlah iterasi maksimal yang ditentukan.

Algoritma K-Means secara *iterative* dapat meningkatkan variasi dari nilai dari masing-masing cluster dimana obyek selanjutnya ditempatkan dalam *cluster* terdekat, dihitung dari titik tengah *cluster*. Titik tengah baru tersebut dapat ditentukan apabila semua data telah ditempatkan dalam *cluster* terdekat. Proses penentuan *centroid* dan penempatan data dalam *cluster* diulangi sampai nilai tengah dari semua *cluster* yang terbentuk tidak berubah lagi (Han dkk, 2012).

BAB 3

Sebelum melakukan proses *clustering* dilakukan tahap persiapan data, salahsatu langkah yang dilakukan adalah menganalisis atribut serta nilai dari atribut yang akan digunakan untuk melakukan *clustering*. Apabila terdapat nilai yang berbeda atau memiliki rentang yang berbeda maka diperlukan adanya proses normalisasi terlebih dahulu. Normalisasi adalah proses transformasi untuk merubah nilai data. Normalisasi data yang dilakukan dengan menggunakan metode Min-Max merupakan metode normalisasi yang dapat menghasilkan transformasi linier dari data asal dimana normalisai menggunakan metod Min-Max ini dapat memetakan sebuah nilai v dari A menjadi v' dalam *range* nilai minimal dan maksimal yang baru. Normalisasi juga dapat digunakan untuk menyamakan skala atribut data kedalam *range* yang lebih spesifik yang lebih kecil seperti -1 sampai 1 atau 0 – 1. Untuk melakukan proses normalisasi dapat dilakukan dengan menggunakan Persamaan 3.2.

$$\text{Nilai normalisasi} = \frac{(\text{nilai awal}-\text{nilai minimal})}{(\text{nilai maksimal}-\text{nilai minimal})} \dots\dots\dots(3.2)$$

3.11 Implementasi K-Means

Algoritma K-Means telah mendapatkan perluasan atau *extension* terhadap kemampuannya hingga saat ini. Kumar dan Wasan, 2010 mencatat ada tiga varian dari algoritma K-Means hasil modifikasi yaitu algoritma global K-Means (Likas dkk, 20013), algoritma *efficient* K-Means (Zhang dkk, 2003) dan algoritma X-Means (Pelleg dan Moore, 2000).

BAB 3

Peningkatan akan kemampuan tersebut antara lain dengan diusulkannya *fast adaptive K-Means clustering* algoritma (Darken dan Moody, 1990), *intelligent K-Means* (Mirkin, 2005), algoritma *improved genetic K-Means* (Guo dkk, 2006), *constrained intelligent K-Means* (Amorim, 2008) serta usulan terkait dengan *shift-based initialization* pada K-Means (Cabria dan Gondra, 2012).

Penggunaan algoritma K-Means dengan menggunakan data spasial hingga saat ini telah diimplementasikan pada berbagai aplikasi diantaranya adalah *clustering* daerah resiko kebakaran di wilayah perkotaan (Lizhi dan Aizhu, 2008), identifikasi *cluster* pepohonan dengan menggunakan data dari citra satelit (Fan dkk, 2010) serta perencanaan sistem transportasi yang memiliki keterkaitan dengan penentuan jumlah lokasi yang sesuai untuk digunakan sebagai pusat layanan *cassava* (Tangkitjaroenongkol, 2011).

3.12 Kelebihan dan Kekurangan Algoritma K-Means

Setiap algoritma dapat memiliki kelebihan dan kekurangan dalam pengopersian serta pengimplementasiannya, hal tersebut tidak menutup kemungkinan bahwa pada algoritma K-Means pun demikian. Berikut ini merupakan kelebihan serta kekurangan dari algoritma K-Means.

Kelebihan

1. Algoritma K-Means memiliki kemudahan dalam pengimplementasian serta pengopeasiannya.
2. Waktu yang dibutuhkan untuk proses yang dilakukan oleh algoritma K-Means untuk melakukan proses pembelajaran relatif cepat.
3. Memiliki tingkat fleksibel yang tinggi serta adaptasi dalam penggunaan algoritma ini dapat dilakukan dengan mudah.

BAB 3

4. Penggunaan algoritma K-Means sangat umum, sehingga ketika terdapat *error* saat pengimplementasian akan banyak sekali dokumentasi yang diperlukan untuk melakukan penyelesaiannya.
5. Prinsip yang digunakan oleh algoritma K-Means yang sederhana sehingga dapat dijelaskan secara umum dan non statistic.

Kekurangan

1. Saat algoritma K-Means pertama kali dijalankan, nilai K dinisialisasikan secara acak sehingga dalam hal penglompokan data hasil yang berbeda-beda pada setiap percobaan yang dilakukan. Namun, jika nilai yang didapat secara acak tersebut digunakan untuk inisialisasi hasilnya kurang baik, maka hasil yang didapat dari pengelompokan menjadi tidak optimal.
2. Apabila terjebak dalam sebuah permasalahan yang biasanya dinamakan *Curse of Dimensionality*. Hal tersebut akan terjadi apabila salahsatu data yang digunakan sebagai data latih memiliki dimensi yang sangat banyak. Misalnya, bila terdapat data dengan terdiri dari 2 atribut saja maka dimensinya hanya 2 juga. Namun, situasinya akan berbeda jika terdapat 20 atribut maka akan ada 20 dimensi yang dimiliki pula. Salahsatu cara kerja algoritma *clustering* ini adalah untuk memperoleh jarak terdekat diantara dari masing-masing k titik dengan titik lainnya. Pencarian jarak terdekat tersebut mudah dilakukan apabila atribut yang dimiliki hanya 2 atau dua dimensi saja, namun bila lebih dari itu hal tersebut akan menjadi lebih sulit untuk dilakukan proses perhitungan dari jarak terdekat.

BAB 3

3.13 Permasalahan Terkait Algoritma K-Means

Dalam penggunaannya beberapa permasalahan sering ditemukan dalam algoritma K-Means. Beberapa permasalahan yang sering muncul tersebut diantaranya adalah sebagai berikut.

1. Sering ditemukannya beberapa model *clustering* yang berbeda.

Penyebab dari permasalahan ini biasanya disebabkan oleh adanya perbedaan pada proses inisialisasi dari masing-masing anggota *cluster*. Dalam proses inisialisasi biasanya digunakan proses inisialisasi secara acak. Dalam suatu kajian tentang perbandingan, proses inisialisasi yang dilakukan secara acak memiliki kecenderungan untuk memberikan hasil yang lebih baik serta independen, namun memiliki kekurangan yaitu lambatnya kecepatan untuk konvergen.

2. Penentuan nilai K, untuk jumlah *cluster* yang paling tepat.

Permasalahan ini merupakan permasalahan utama dalam algoritma K-Means. Oleh karena itu diperlukan pendekatan untuk penentuan nilai K yang optimal agar memperoleh hasil *clustering* optimal.

3. Terjadi kegagalan dalam penentuan kriteria untuk memperoleh hasil yang konvergen.

Kegagalan untuk konvergen, dapat dimungkinkan terjadi dalam metode Hard K-Means ataupun Fuzzy K-Means. Kemungkinan dari hal tersebut akan semakin besar terjadi dalam metode Hard K-Means, dikarenakan setiap data yang berada dalam dataset dialokasikan secara tegas (*hard*) untuk menjadi bagian dari suatu cluster tertentu. Perpindahan data yang terjadi dari suatu dataset ke dalam suatu *cluster*, dapat mempengaruhi berubahnya karakteristik dari suatu model clustering sehingga menyebabkan data yang telah berpindah tersebut lebih cocok

BAB 3

untuk berada pada cluster semula sebelum data tersebut dipindahkan, begitu pulan sebaliknya.

Pada Fuzzy K-Means walaupun terjadi permasalahan ini, kemungkinan terjadinya adalah sangat kecil, dikarenakan setiap data dilengkapi dengan *membership function* untuk menjadi anggota dari sebuah *cluster*.

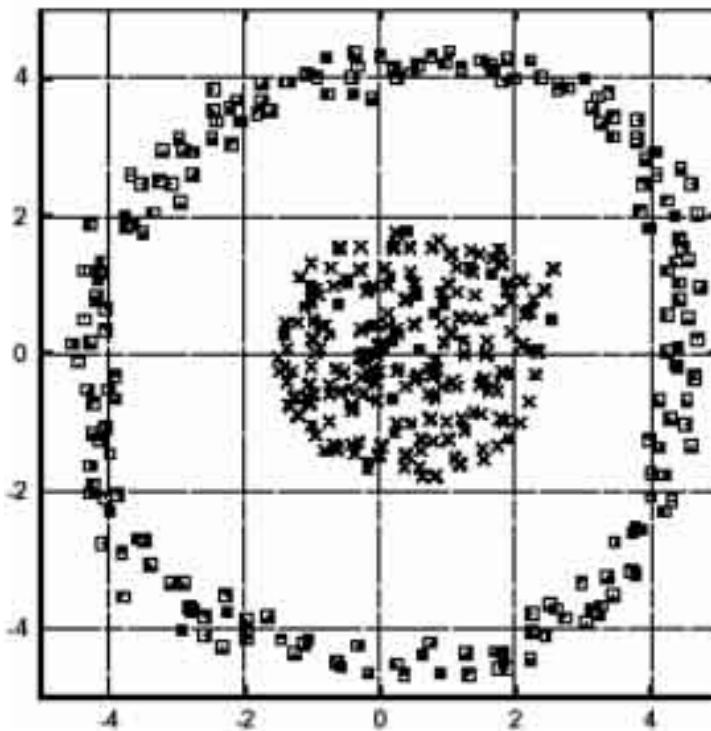
4. *Outliers*.

Permasalahan ini hampir terjadi pada setiap metode yang digunakan untuk memodelkan data. Dalam metode K-Means hal ini dapat menjadi sebuah permasalahan yang cukup serius dan menentukan terhadap hasil dari *clustering*. Hal yang perlu diperhatikan dalam mendeteksi *outliers* di dalam proses peneglompokan data termasuk bagaimana menentukan apakah suatu data merupakan *outliers* dari suatu *cluster* tertentu serta apakah data dalam jumlah kecil yang memebentuk suatu *cluster* dapat dianggap sebagai sebuah *outliers*. Proses tersebut memerlukan sebuah pendekatan yang berbeda dengan proses pendeteksian *outliers* tersebut dalam suatu dataset yang hanya memiliki satu populasi yang sama.

5. Bentuk dari *cluster*.

Bentuk dari suatu *cluster* yang diperoleh di dalam algoritma K-Means merupakan hal yang perlu dicermati. Berbeda dengan metode *clustering* lainnya. K-Means secara umum tidak mememprhatikan bentuk dari tiap-tiap cluster yang menjadi dasar terbentuknya suatu model, walaupun biasanya secara natural sebuah *cluster* berbentuk bulat. Diperlukan beberapa pendekatan untuk dataset yang memiliki bentuk yang tidak biasa.

BAB 3



Gambar 3.6 Contoh Dataset yang Memiliki Bentuk Khusus

6. *Overlapping.*

Permasalahan yang terkait dengan *overlapping* ini dapat dikatakan sebagai permasalahan yang sering sekali diabaikan dikarenakan biasanya permasalahan ini sulit untuk dideteksi. Hal tersebut bias terjadi pada metode Hard K-Means maupun metode Fuzzy K-Means, dikarenakan pada dasarnya, metode tersebut tidak dilengkapi dengan fungsi untuk pendeteksian apakah terdapat *cluster* tersembunyi didalam suatu *cluster*

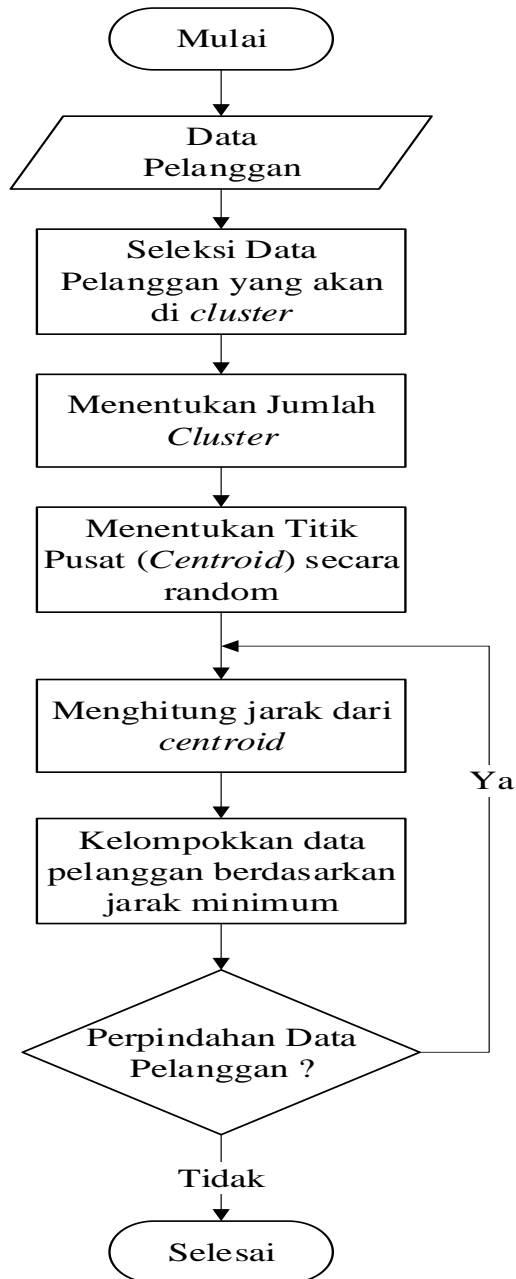
BAB 3

Permasalahan tersebut harus sangat diperhatikan dalam proses *clustering* atau pengelompokan data menggunakan algoritma K-Means sehingga dapat dilakukan langkah antisipasi agar beberapa permasalahan tersebut tidak terjadi.

3.14 Contoh Perhitungan *Clustering* dengan Algoritma K-Means

Pada pembahasan ini akan dijelaskan bagaimana contoh implementasi pengolahan data menggunakan algoritma K-Means. Dalam hal ini algoritma K-Means akan diaplikasikan untuk *clustering* data pelanggan. Sebelum melakukan pengolahan data, terlebih dahulu harus dipahami langkah-langkah yang dilakukan dalam proses *clustering* agar prosesnya berjalan dengan baik. Berikut ini merupakan diagram alir rancangan *clustering* yang diimplementasikan dalam segmentasi pelanggan.

BAB 3



Gambar 3.7 Diagram Alir Rancangan Penentuan *Cluster* Atau Segmentasi Data Pelanggan

BAB 3

Pada gambar 3.6 merupakan langkah-langkah yang dilakukan untuk melakukan *clustering* menggunakan algoritma K-Means .

1. Data pelanggan : pengumpulan dan penyiapan data pelanggan yang akan di *cluster*.
2. Seleksi data pelanggan : dilakukan untuk memilih atribut yang akan digunakan untuk proses *clustering*.
3. Penentuan jumlah *cluster* : menentukan jumlah *cluster* atau nilai k yang diinginkan.
4. Penentuan titik pusat : dilakukan untuk menentukan nilai *centroid* pada setiap *cluster*.
5. Menghitung jarak dari centroid : proses untuk menghitung jarak minimum antara data dengan titik pusat atau *centroid*, dapat dilakukan dengan menggunakan Persamaan 3.1.

$$d_{ij} = \sqrt{\sum_{k=1}^p (X_{ik} - C_{jk})^2} \dots \dots \dots (3.1)$$

Keterangan :

d_{ij} : *Euclidian distance*, Jarak antara data pelanggan i dan j

p : Dimensi data yang digunakan

X_{ik} : Data pelanggan ke-i

C_{jk} : Centroid ke-j

BAB 3

6. Pengelompokkan data : pada tahap pengelompokkan ini dilakukan berdasarkan hasil perhitungan *euclidian distance* setiap data pelanggan. Suatu data pelanggan akan menjadi anggota dari *cluster* ke-k apabila jarak data tersebut ke pusat *cluster* bernilai paling kecil jika dibandingkan dengan jarak ke pusat *cluster* lainnya.
7. Hitung nilai centroid baru berdasarkan data yang mengikuti *cluster* masing-masing. Untuk memperoleh titik *centroid* baru dilakukan dengan cara menghitung nilai rata-rata dari data yang ada pada *cluster* yang sama .
8. Apabila tidak ada data yang berpindah ke *cluster* lain maka proses *clustering* selesai. Ulangi dari langkah ke tiga hingga langkah ke lima apabila terdapat perpindahan data ke *cluster* lain.

Berikut ini merupakan penerapan proses segmentasi data pelanggan sesuai dengan diagram alir pada Gambar 3.6.

• Penjelasan Data

Pada pembahasan ini akan dijelaskan terkait dengan data yang akan digunakan dalam proses clustering. Data yang akan digunakan merupakan data pelanggan dimana hasil akhir dari *clustering* ini adalah pengelompokan data ke dalam masing-masing *cluster* sesuai dengan kemiripan dari masing-masing data.

1. Data Pelanggan Indihome

Data pelanggan indihome merupakan data pelanggan yang berlangganan layanan jaringan internet.

BAB 3

Tabel 3.1 Data Pelanggan Indihome

NCLI	ND_INTERNET	ND	CITEM_SPEEDY	KECEPATAN
39684298	131167143669	2287274505	INETF10M	10240
39713960	131167143707	2287272529	INETFL10M	10240
39716635	131167142784	2287274029	INETFL20M	20480
68586	131167113058	227207516	INETFL10M	10240
78255	131167124056	227274422	INETFL10M	10240
39783299	131167143737		INETF10M	10240
110900	131167139046	227207489	INETF10M	10240
39771851	131167143754	2271520899	INET10Q053	10240
64148	131167134218	227104625	INETF10M	10240

Lanjutan Tabel 3.1 Data Pelanggan Indihome

DESKRIPSI	TGL_REG	TGL_ETAT	NAMA
CS18 - Indihome Gamer USEE	31-Dec-18	01-jan-2019 14:01:11	IDA SUSANTI
CS18 - Sensasi Akhir Tahun 2018 UseeTV	02-Jan-19	02-jan-2019 20:01:08	SRI MULYANI
CS18 - Sensasi Akhir Tahun 2018 UseeTV	05-Jan-19	05-jan-2019 19:01:10	Hesti Handayani.drg
CS18 - Sensasi Akhir Tahun 2018 UseeTV	07-Jan-19	07-jan-2019 13:01:06	BPK. SYARIF
CS18 - Sensasi Akhir Tahun 2018 UseeTV	07-Jan-19	07-jan-2019 20:01:05	ABDUL MANAP
CS17 - New Indihome Netizen II USEETV	08-Jan-19	08-jan-2019 19:01:09	yosep
CS17 - New Indihome Lower Value USEETV	08-Jan-19	09-jan-2019 12:01:04	SANDRA WISNU WENDHARI
CS17 - Paket Indihome Per ODP UseeTV New Entry	09-Jan-19	09-jan-2019 20:01:07	Andreas Asmara
CS17 - New Indihome Lower Value USEETV	09-Jan-19	10-jan-2019 11:01:02	SURYANA AFFANDI AKT

2. Data Pelanggan *Add On*

Data pelanggan *Add On* merupakan data pelanggan yang berlangganan layanan tambahan produk digital yang disediakan oleh perusahaan untuk melengkapi layanan internet indihome.

BAB 3

Tabel 3.2 Data Pelanggan *Add On*

WITEL	NCLI	NDOS	NDEM	NO_INET	ITEM	PRICE	TGL_VA
BANDUNG	39681602	1	661506709	131183111875	OTTSTUDY1	5000	22-Jan-19
BANDUNG	39687618	1	661504659	131183156226	OTTSTUDY1	5000	22-Jan-19
BANDUNG	39644960	1	661483159	131165155637	OTTSTUDY1	5000	22-Jan-19
BANDUNG	554326	3	661470469	131161124294	OTTSTUDY1	5000	22-Jan-19
BANDUNG	39742477	1	661423809	131165154465	OTTSTUDY1	5000	22-Jan-19
BANDUNG	39766152	1	661378249	131183112749	OTTSTUDY1	5000	22-Jan-19
BANDUNG	39251881	1	662014819	131184123706	OTTSTUDY1	5000	23-Jan-19
BANDUNG	39654108	1	661704709	131183156168	OTTSTUDY1	5000	23-Jan-19
BANDUNG	39626404	1	663181239	131165155613	OTTSTUDY1	5000	25-Jan-19

Lanjutan Tabel 3.2 Data Pelanggan *Add On*

TGL_VA	TGL_PS	KCONTACT
22-Jan-19	22-Jan-19	SC15370830;Upgrade Entry to Essential 3bln s.d 31 Maret 2019/By.CC
22-Jan-19	22-Jan-19	SC15370616;Upgrade Entry to Essential 3bln s.d 31 Maret 2019/By.CC
22-Jan-19	22-Jan-19	SC15369250;Upgrade Entry to Essential 3bln s.d 31 Maret 2019/By.CC
22-Jan-19	22-Jan-19	SC15368130;Upgrade Entry to Essential 3bln s.d 31 Maret 2019/By.CC
22-Jan-19	22-Jan-19	SC15365469;Upgrade Entry to Essential 3bln s.d 31 Maret 2019/By.CC
22-Jan-19	22-Jan-19	SC15362013;Upgrade Entry to Essential 3bln s.d 31 Maret 2019 (40Kper1A
23-Jan-19	23-Jan-19	SC15394066;Upgrade Entry to Essential 3bln s.d 31 Maret 2019/By.CC
23-Jan-19	23-Jan-19	SC15376505;Upgrade Entry to Essential 3bln s.d 31 Maret 2019/By.CC
25-Jan-19	25-Jan-19	SC15449227;CCW;Upgrade Entry to Essential 3bln s.d 31 Maret 2019

3. Data *Churn* pelanggan

Data *Churn* pelanggan merupakan data pelanggan yang berhenti berlangganan layanan.

BAB 3

Tabel 3.3 Data *Churn* Pelanggan

KAWASAN	WITEL	DATTEL	NCLI	ND_INTERNET	DESKRIPSI
DIVRE 3	BANDUNG	BANDUNG	33670266	131165127698	CS16 - New USEETV indiHOME Essential
DIVRE 3	BANDUNG	BANDUNG	35943467	131165137843	CS16 - USeeTV indiHOME Solution
DIVRE 3	BANDUNG	BANDUNG	34403798	131165132052	CS17 - New Indihome Lower Value USEETV
DIVRE 3	BANDUNG	BANDUNG	37168158	131165127120	CS16 - Paket IndiHome Dinamic Price Premium (UseeTV)
DIVRE 3	BANDUNG	BANDUNG	36809423	131165141515	CS18 - Indihome Khusus Imlek USEETV
DIVRE 3	BANDUNG	BANDUNG	37506312	131165146424	CS18 - New UseeTV-OTT IFLIX Indihome Penuh Berkah
DIVRE 3	BANDUNG	BANDUNG	21740	131165112733	Program Indihome Add On UseeTV 49Rb
DIVRE 3	BANDUNG	BANDUNG	21204	131165110393	New Indihome Pemenangan UseeTV Winning
DIVRE 3	BANDUNG	BANDUNG	35876141	131165136321	CS17 - IndiHome Promo NaRu 2017 (USEE)
DIVRE 3	BANDUNG	BANDUNG	35920444	131165136573	CS17 - IndiHome Promo NaRu 2017 (USEE)

Lanjutan Tabel 3.3 Data *Churn* Pelanggan

TGL_REG	TGL_ETAT	STATUS_ORDER
31-Jan-19	31-jan-2019 16:01:40	CHURN OUT
31-Jan-19	31-jan-2019 17:01:49	CHURN OUT
31-Jan-19	31-jan-2019 17:01:32	CHURN OUT
31-Jan-19	31-jan-2019 20:01:57	CHURN OUT
10-Jan-19	21-jan-2019 10:01:14	CHURN OUT
11-Jan-19	21-jan-2019 10:01:11	CHURN OUT
31-Jan-19	31-jan-2019 16:01:28	CHURN OUT
31-Jan-19	31-jan-2019 17:01:54	CHURN OUT
31-Jan-19	31-jan-2019 17:01:44	CHURN OUT
31-Jan-19	31-jan-2019 18:01:35	CHURN OUT

• Seleksi Data Pelanggan

Proses ini dilakukan untuk memilih atribut yang akan digunakan untuk proses *clustering*. Atribut yang digunakan untuk mewakili setiap pelanggan diantaranya adalah nomor pelanggan atau NCLI, lama berlangganan, jumlah layanan yang digunakan serta total tagihan namun dalam proses *clustering*, atribut yang digunakan hanya 3 atribut kecuali nomor pelanggan atau NCLI.

BAB 3

Tabel 3.4 Atribut Yang Digunakan Untuk *Clustering*

NCLI	LAMA_LANGGANAN	JUMLAH_LAYANAN	JUMLAH_TAGIHAN_HARUS_DIBAYAR
39684298	11	1	418000
39713960	11	2	511500
39716635	11	2	698500
39818227	11	2	291500
68228	11	2	517000
39937697	11	1	352000
40011078	11	2	847000
39817407	11	1	621500
39864058	11	2	1100000
39952623	11	1	902000
39891161	11	1	352000
40033755	11	2	291500
39786171	11	2	2090000
39995005	11	2	621500
32745610	11	2	814000
380513	1	2	1358500
39765624	11	1	286000
39600770	11	1	275000
39999904	11	1	275000
39892083	11	1	495000

Pada pembahasan seleksi data pelanggan dilakukan juga proses normalisasi. Hal ini dilakukan untuk menyamakan *range* nilai dari masing-masing atribut yang dipilih. Proses normalisasi dilakukan dengan menggunakan Persamaan 3.2. Berikut ini merupakan contoh perhitungan normalisasi.

1. Lama Langganan

Lama Langganan	
Max	11
Min	1

$$\begin{aligned}
 X_{11} &= (X_{\text{Lama langganan NCLI 1}} - X_{\min}) / (X_{\max} - X_{\min}) \\
 &= (11 - 1) / (11 - 1) \\
 &= 1
 \end{aligned}$$

BAB 3

2. Jumlah Layanan

Jumlah Layanan	
Max	2
Min	1

$$\begin{aligned}
 X_{21} &= (X_{\text{Jumlah Layanan NCLI 1}} - X_{\min}) / (X_{\max} - X_{\min}) \\
 &= (1 - 1) / (2 - 1) \\
 &= 0
 \end{aligned}$$

3. Total Tagihan

Tagihan	
Max	2090000
Min	275000

$$\begin{aligned}
 X_{31} &= (X_{\text{Total Tagihan NCLI 1}} - X_{\min}) / (X_{\max} - X_{\min}) \\
 &= (418000 - 275000) / (2090000 - 275000) \\
 &= 0.08
 \end{aligned}$$

Perhitungan dilakukan sampai semua nilai atribut dinormalisasi. Hasil dari proses normalisasi dapat dilihat pada Tabel 3.5.

Tabel 3.5 Hasil Normalisasi

NCLI	LAMA_LANGGANAN	JUMLAH_LAYANAN	TAGIHAN
39684298	1.00	0.00	0.08
39713960	1.00	0.25	0.13
39716635	1.00	0.25	0.23
39818227	1.00	0.25	0.01
39813816	1.00	0.00	0.32
30389436	1.00	0.25	0.32

BAB 3

Lanjutan Tabel 3.5 Data *Churn* Pelanggan

NCLI	LAMA_LANGGANAN	JUMLAH_LAYANAN	TAGIHAN
39950721	1.00	0.00	0.48
39872630	1.00	0.25	0.48
39880310	1.00	0.00	0.03
616655	1.00	0.50	0.20
39792742	0.90	0.25	0.19
39802417	1.00	0.00	0.08
40012860	1.00	0.25	0.24
110281	1.00	0.50	0.24
39739034	1.00	0.00	0.23
39742383	1.00	0.25	0.23
40007439	1.00	0.00	0.35
39884232	1.00	0.25	0.35
39884288	1.00	0.25	0.35
639184	1.00	0.75	0.87

- **Penentuan Jumlah *Cluster***

Pada tahap ini dilakukan penentuan jumlah *cluster* atau nilai K. Adapun nilai K yang digunakan pada penelitian ini berjumlah 3.

- **Penentuan Titik Pusat**

Pada tahap ini dilakukan penentuan nilai centroid pada setiap *cluster*, nilai dari *centroid* itu sendiri ditentukan secara acak.

Tabel 3.6 *Centroid* awal

K	NCLI	LAMA_LANGGANAN	JUMLAH_LAYANAN	TAGIHAN
0	30389436	1.00	0.25	0.32
1	110281	1.00	0.50	0.24
2	639184	1.00	0.75	0.87

BAB 3

- **Perhitungan Jarak Data Centroid**

Inisialisasi

Pada tahap inisialisasi merupakan perhitungan jarak antara data dengan nilai *centroid* awal (*euclidian distance*) dengan menggunakan persamaan 3.1.

1. Perhitungan jarak minimum data ke-1 (1.00, 0, 0.08)

- ❖ DC0

$$\begin{aligned} & \sqrt{(1 - 1)^2 + (0 - 0.25)^2 + (0.08 - 0.32)^2} \\ & = 0.346554469 \end{aligned}$$

- ❖ DC1

$$\begin{aligned} & \sqrt{(1 - 1)^2 + (0 - 0.5)^2 + (0.08 - 0.24)^2} \\ & = 0.52497619 \end{aligned}$$

- ❖ DC2

$$\begin{aligned} & \sqrt{(1 - 1)^2 + (0 - 0.75)^2 + (0.08 - 0.87)^2} \\ & = 1.089311709 \end{aligned}$$

2. Perhitungan jarak minimum data ke-2 (1.00, 0.25, 0.13)

- ❖ DC0

$$\begin{aligned} & \sqrt{(1 - 1)^2 + (0.25 - 0.25)^2 + (0.13 - 0.32)^2} \\ & = 0.19 \end{aligned}$$

- ❖ DC1

$$\begin{aligned} & \sqrt{(1 - 1)^2 + (0.25 - 0.5)^2 + (0.13 - 0.24)^2} \\ & = 0.273130006 \end{aligned}$$

BAB 3

❖ DC2

$$\sqrt{(1 - 1)^2 + (0.25 - 0.75)^2 + (0.13 - 0.87)^2}$$

$$= 0.893084542$$

Lakukan perhitungan yang sama pada data yang lain sehingga menghasilkan nilai jarak data terhadap *centroid* sebagai berikut.

Tabel 3.7 Hasil Perhitungan Jarak Data Dengan *Centroid* Pada Inisialisasi

NO	NCLI	DC0	DC1	DC2
1	39684298	0.346554	0.524976	1.089312
2	39713960	0.19	0.27313	0.893085
3	39716635	0.09	0.2502	0.812158
4	39818227	0.31	0.339706	0.994786
5	39813816	0.25	0.50636	0.930054
6	30389436	0	0.262488	0.743303
7	39950721	0.296816	0.554617	0.84534
8	39872630	0.16	0.346554	0.634114
9	39880310	0.382884	0.54231	1.126099
10	616655	0.277308	0.04	0.715122
11	39792742	0.164012	0.273861	0.849941
12	39802417	0.346554	0.524976	1.089312
13	40012860	0.08	0.25	0.804301
14	110281	0.262488	0	0.677791
15	39739034	0.265707	0.5001	0.985951
16	39742383	0.09	0.2502	0.812158
17	40007439	0.251794	0.511957	0.912634
18	39884232	0.03	0.27313	0.721388
19	39884288	0.03	0.27313	0.721388
20	639184	0.743303	0.677791	0

BAB 3

Penentuan data berada pada *cluster* tertentu didasarkan pada jarak minimum perhitungan *euclidian distance* sehingga didapat hasil sebagai berikut.

Tabel 3.8 Hasil Pemetaan *Centroid* Awal

NO	NCLI	DC0	DC1	DC2	HASIL
1	39684298	*			0
2	39713960	*			0
3	39716635	*			0
4	39818227	*			0
5	39813816	*			0
6	30389436	*			0
7	39950721	*			0
8	39872630	*			0
9	39880310	*			0
10	616655		*		1
11	39792742	*			0
12	39802417	*			0
13	40012860	*			0
14	110281		*		1
15	39739034	*			0
16	39742383	*			0
17	40007439	*			0
18	39884232	*			0
19	39884288	*			0
20	639184			*	2

BAB 3

- Iterasi Pertama

Setelah data terbagi ke dalam *cluster* pada tahap inisialisasi, untuk melanjutkan tahap iterasi pertama diperlukan adanya penentuan centroid baru untuk menghitung jarak menggunakan euclidian *distance*. Untuk menentukan centroid baru dapat dilakukan dengan menggunakan Persamaan 3.3 .

$$Ci = \frac{1}{M} \sum_{j=1}^M X_j \dots\dots\dots (3.3)$$

Lokasi centroid setiap kelompok diambil dari rata-rata (mean) semua nilai data pada setiap fiturnya. Dimana rata-rata semua nilai dihitung dengan membagi jumlah data dengan jumlah atribut untuk pembagi pada perhitungan atribut baru.

- ❖ C0

$$\left(\frac{(1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 0.9 + 1 + 1 + 1 + 1 + 1 + 1 + 1)}{17} \right)$$

$$= 0.994117647$$

$$\left(\frac{(0 + 0.25 + 0.25 + 0.25 + 0 + 0.25 + 0 + 0.25 + 0 + 0.25 + 0 + 0.25 + 0 + 0.25 + 0 + 0.25 + 0 + 0.25)}{17} \right)$$

$$= 0.147058824$$

$$\left(\frac{(0.08 + 0.13 + 0.23 + 0.01 + 0.32 + 0.32 + 0.48 + 0.48 + 0.03 + 0.19 + 0.08 + 0.24 + 0.23 + 0.23 + 0.35 + 0.35 + 0.35)}{17} \right)$$

$$= 0.241176471$$

BAB 3

Sehingga *centroid* C0 yang baru bernilai (0.994117647, 0.147058824, 0.241176471).

❖ C1

$$\left(\frac{(1+1)}{2}\right)$$

$$= 1$$

$$\left(\frac{(0.5+0.5)}{2}\right)$$

$$= 0.5$$

$$\left(\frac{(0.2+0.24)}{2}\right)$$

$$= 0.22$$

Sehingga *centroid* C1 yang baru bernilai (1, 0.5, 0.22).

❖ C2

$$\left(\frac{(1)}{1}\right)$$

$$= 1$$

$$\left(\frac{(0.75)}{1}\right)$$

$$= 0.75$$

$$\left(\frac{(0.87)}{17}\right)$$

$$= 0.87$$

Sehingga *centroid* C2 yang baru bernilai (1, 0.75, 0.87)

BAB 3

Hitung kembali jarak euclidian *distance* antara data dengan centroid baru seperti pada tahap sebelumnya, sehingga didapat hasil sebagai berikut.

Tabel 3.9 Hasil Perhitungan Jarak Data Dengan Centroid Pada Iterasi Pertama

NO	NCLI	DC0	DC1	DC2
1	39684298	0.218263	0.51923	1.089312
2	39713960	0.15163	0.265707	0.893085
3	39716635	0.103713	0.2502	0.812158
4	39818227	0.253129	0.326497	0.994786
5	39813816	0.166955	0.509902	0.930054
6	30389436	0.129787	0.269258	0.743303
7	39950721	0.280531	0.56356	0.84534
8	39872630	0.260131	0.360694	0.634114
9	39880310	0.257403	0.534883	1.126099
10	616655	0.355384	0.02	0.715122
11	39792742	0.148573	0.270924	0.849941
12	39802417	0.218263	0.51923	1.089312
13	40012860	0.103116	0.250799	0.804301
14	110281	0.352992	0.02	0.677791
15	39739034	0.1476	0.5001	0.985951
16	39742383	0.103713	0.2502	0.812158
17	40007439	0.18304	0.516624	0.912634
18	39884232	0.149913	0.28178	0.721388
19	39884288	0.149913	0.28178	0.721388
20	639184	0.871201	0.696419	0

BAB 3

Tentukan kembali data berada pada *cluster* yang sesuai dengan jarak minimum perhitungan euclidian *distance* sehingga didapat hasil sebagai berikut.

Tabel 3.10 Hasil Pemetaan Centroid Iterasi Pertama

NO	NCLI	DC0	DC1	DC2	HASIL
1	39684298	*			0
2	39713960	*			0
3	39716635	*			0
4	39818227	*			0
5	39813816	*			0
6	30389436	*			0
7	39950721	*			0
8	39872630	*			0
9	39880310	*			0
10	616655		*		1
11	39792742	*			0
12	39802417	*			0
13	40012860	*			0
14	110281		*		1
15	39739034	*			0
16	39742383	*			0
17	40007439	*			0
18	39884232	*			0
19	39884288	*			0
20	639184			*	2

BAB 3

- Iterasi Kedua

Untuk melanjutkan perhitungan pada iterasi kedua, diperlukan centroid baru. Centroid baru tersebut dihitung menggunakan cara yang sama dengan yang dilakukan untuk mencari centroid pada iterasi kedua dengan persamaan 3.3.

Berikut ini merupakan perhitungan untuk memperoleh centroid baru pada iterasi kedua.

❖ C0

$$\left(\frac{(1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 0.9 + 1 + 1 + 1 + 1 + 1 + 1)}{17} \right)$$

$$= 0.994117647$$

$$\left(\frac{(0 + 0.25 + 0.25 + 0.25 + 0 + 0.25 + 0 + 0.25 + 0 + 0.25 +)}{0 + 0.25 + 0 + 0.25 + 0 + 0.25 + 0.25} \right)$$

$$= 0.147058824$$

$$\left(\frac{(0.08 + 0.13 + 0.23 + 0.01 + 0.32 + 0.32 + 0.48 + 0.48 + 0.03 + 0.19 +)}{0.08 + 0.24 + 0.23 + 0.23 + 0.35 + 0.35 + 0.35} \right)$$

$$= 0.241176471$$

Sehingga *centroid* C0 yang baru bernilai (0.994117647, 0.147058824, 0.241176471).

❖ C1

$$\left(\frac{(1 + 1)}{2} \right)$$

$$= 1$$

BAB 3

$$\left(\frac{(0.5 + 0.5)}{2}\right)$$
$$= 0.5$$

$$\left(\frac{(0.2 + 0.24)}{2}\right)$$
$$= 0.22$$

Sehingga *centroid* C1 yang baru bernilai (1, 0.5, 0.22).

❖ C2

$$\left(\frac{(1)}{1}\right)$$
$$= 1$$

$$\left(\frac{(0.75)}{1}\right)$$
$$= 0.75$$

$$\left(\frac{(0.87)}{17}\right)$$
$$= 0.87$$

Sehingga *centroid* C2 yang baru bernilai (1, 0.75, 0.87).

Setelah *centroid* baru didapat, hitung kembali jarak antara *centroid* baru dengan data, kemudian pilih jarak minimum untuk menentukan data tersebut berada pada *cluster* tertentu.

BAB 3

Tabel 3.11 Hasil Perhitungan Jarak Data Dengan Centroid Pada Iterasi Kedua

NO	NCLI	DC0	DC1	DC2
1	39684298	0.218263	0.51923	1.089312
2	39713960	0.15163	0.265707	0.893085
3	39716635	0.103713	0.2502	0.812158
4	39818227	0.253129	0.326497	0.994786
5	39813816	0.166955	0.509902	0.930054
6	30389436	0.129787	0.269258	0.743303
7	39950721	0.280531	0.56356	0.84534
8	39872630	0.260131	0.360694	0.634114
9	39880310	0.257403	0.534883	1.126099
10	616655	0.355384	0.02	0.715122
11	39792742	0.148573	0.270924	0.849941
12	39802417	0.218263	0.51923	1.089312
13	40012860	0.103116	0.250799	0.804301
14	110281	0.352992	0.02	0.677791
15	39739034	0.1476	0.5001	0.985951
16	39742383	0.103713	0.2502	0.812158
17	40007439	0.18304	0.516624	0.912634
18	39884232	0.149913	0.28178	0.721388
19	39884288	0.149913	0.28178	0.721388
20	639184	0.871201	0.696419	0

BAB 3

Tabel 3.12 Hasil Pemetaan *Centroid* Iterasi Kedua

NO	NCLI	DC0	DC1	DC2
1	39684298	*		
2	39713960	*		
3	39716635	*		
4	39818227	*		
5	39813816	*		
6	30389436	*		
7	39950721	*		
8	39872630	*		
9	39880310	*		
10	616655		*	
11	39792742	*		
12	39802417	*		
13	40012860	*		
14	110281		*	
15	39739034	*		
16	39742383	*		
17	40007439	*		
18	39884232	*		
19	39884288	*		
20	639184			*

Dikarenakan centroid pada iterasi kedua sama dengan centroid pada iterasi pertama berarti pengulangan perhitungan atau iterasi dihentikan pada iterasi kedua dikarenakan posisi data sudah tidak ada yang berpindah posisi dan sudah berada pada *cluster* nya.

BAB 3

3.15 Metode Penentuan K

Nilai K dalam suatu proses *clustering* merupakan sebuah elemen penting. Dikarenakan K ini adalah inisialisasi dari jumlah *cluster* atau kelompok untuk pembagian data yang dilakukan dalam *clustering*. Untuk memperoleh hasil pengelompokan data yang maksimal, penentuan nilai K ini tidak dapat ditentukan secara sembarang. Oleh karena itu digunakanlah sebuah metode yang dapat menentukan nilai K untuk mendapatkan hasil *clustering* yang optimal dan dinamakan dengan metode *Elbow*.

3.13.1 Metode *Elbow*

Metode *Elbow* merupakan suatu metode yang dapat digunakan untuk menghasilkan informasi dalam menentukan jumlah *cluster* terbaik dengan cara melihat persentase hasil perbandingan antara jumlah *cluster* yang akan membentuk siku pada suatu titik tertentu. Metode ini memberikan ide/gagasan dengan cara memilih nilai *cluster* dan kemudian menambah nilai *cluster* tersebut untuk dijadikan model data dalam penentuan *cluster* terbaik. Dan selain itu persentase perhitungan yang dihasilkan menjadi pembandingan antara jumlah *cluster* yang ditambah. Hasil persentase yang berbeda dari setiap nilai *cluster* dapat ditunjukkan dengan menggunakan grafik sebagai sumber informasinya. Jika nilai *cluster* pertama dengan nilai *cluster* kedua memberikan sudut dalam grafik atau nilainya mengalami penurunan paling besar maka nilai *cluster* tersebut yang terbaik.

Menurut Bholowalia dan Kumar (2014) tahapan metode Elbow dalam menentukan nilai K pada K-Means:

1. Menginisialisasi awal nilai k.

BAB 3

2. Menaikan nilai k .
3. Menghitung hasil *Sum of Square Error* dari tiap nilai k .
4. Analisa hasil *Sum of Square Error* dari nilai k yang mengalami penurunan secara drastis.
5. Cari dan tetapkan nilai k yang berbentuk siku.

Pada metode *Elbow* nilai *cluster* terbaik dilihat dengan membandingkan nilai yang akan diambil dari hasil perhitungan *Sum of Square Error* (SSE) dari masing-masing *cluster*, karena semakin besar jumlah *cluster* K maka nilai dari *Sum of Square Error* (SSE) tersebut akan semakin kecil. Untuk menghitung SSE dapat menggunakan Persamaan 3.4.

$$SSE = \sum_{K=1}^K \sum_{x_i \in S_k} \|X_i - C_k\|^2 \dots\dots\dots(3.4)$$

Dimana:

K = jumlah *cluster*

x_i = data ke – i

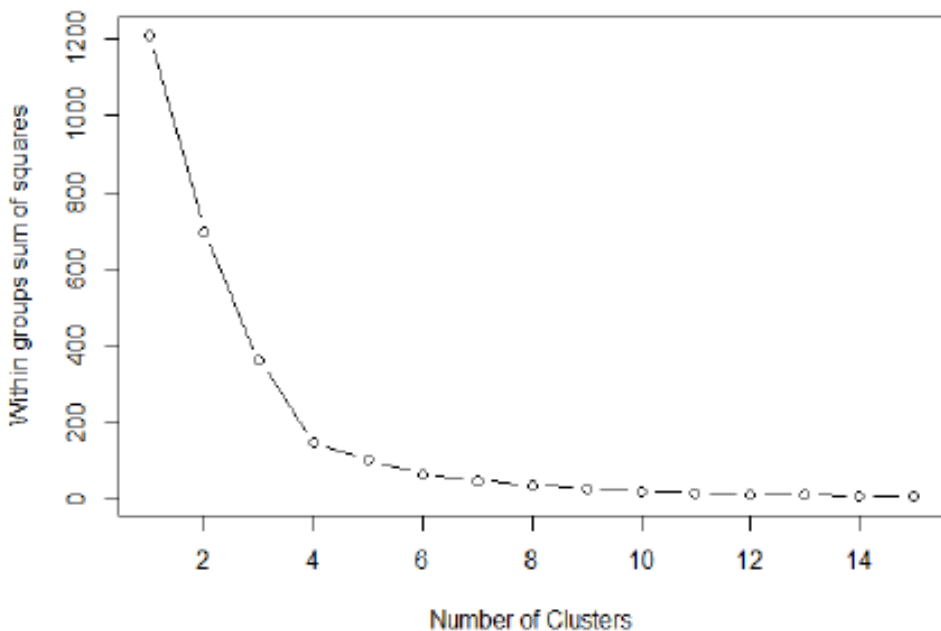
C_k = centroid *cluster*

Sum of Square Error (SSE) merupakan rumus yang digunakan untuk mengukur perbedaan antara data yang diperoleh dengan model perkiraan yang telah dilakukan sebelumnya. SSE sering digunakan sebagai acuan penelitian terkait dalam menentukan optimal *cluster*.

Setelah dilihat hasil dari SSE maka akan diperoleh beberapa nilai K yang mengalami penurunan yang paling besar dan signifikan dan selanjutnya hasil dari nilai K tersebut akan turun secara perlahan-lahan

BAB 3

sampai hasil nilai K tersebut stabil. Misal, nilai *cluster* dengan nilai $K = 2$ ke *cluster* yang nilai $K = 3$, kemudian dari *cluster* dengan nilai $K = 3$ ke *cluster* dengan nilai $K = 4$, dari perubahan jumlah cluster yang berbeda tersebut dapat dilihat penurunan yang signifikan yang nantinya akan membentuk siku pada titik *cluster* dengan nilai $K = 3$. Dengan demikian dapat disimpulkan bahwa hasil dari penggunaan metode *elbow* untuk menentukan nilai K yang terbaik dan diperoleh nilai K yang terbaik adalah 3 seperti ditunjukkan pada Gambar 3.7.



Gambar 3.8 Grafik Metode Elbow

BAB 3

3.16 Metode Evaluasi *Cluster*

Evaluasi dapat diartikan sebagai sebuah proses untuk memeriksa, menilai, membuat suatu keputusan ataupun menyediakan informasi yang berguna terhadap program atau proses yang telah dilaksanakan serta untuk mengukur sejauh mana ketercapaian dari sebuah proses tersebut. Dalam hal ini evaluasi clustering dilakukan untuk mneguji performa dari suatu cluster. Dengan adanya evaluasi *cluster* dapat dilihat dari berbagai aspek bahwa proses *clustering* yang dilakuakn memiliki hasil yang baik serta telah dilakukan dengan tahapan-tahapan yang tepat. Berikut ini merupakan beberapa metode yang digunakan untuk melakukan evaluasi *cluster*.

3.14.1 Silhouette index

Secara umum, indeks validitas Silhouette menghitung rata-rata nilai setiap titik pada himpunan data. Lebih spesifik, perhitungan nilai setiap titik merupakan selisih dari nilai *separation* dan *compactness* yang dibagi dengan maksimum antara keduanya. Jumlah *cluster* yang terbaik ditunjukkan dengan nilai Silhouette yang semakin mendekati 1 (Rosseeuw, 1987). Misalkan terdapat N buah titik pada suatu himpunan data, terdapat pula di dalamnya *cluster* p dan *cluster* q dengan x_i adalah titik pada *cluster* p dan y_j adalah titik pada *cluster* q , sehingga $a_{p,i}$ adalah rata-rata jarak titik x_i ke setiap titik pada *cluster* p , dan $d_{q,i}$ adalah rata-rata jarak titik x_i ke setiap titik pada *cluster* q . Maka rumus perhitungan indeks validitas Silhouette dapat dilihat pada Persamaan 3.5.

BAB 3

$$\begin{aligned}
 SIL &= \frac{1}{N} \sum_{i=0}^N s_{x_i}, \\
 s_{x_i} &= \frac{(b_{q,i} - a_{p,i})}{\max \{a_{p,i}, b_{q,i}\}}, p \neq q, \\
 &\dots\dots\dots(3.5) \\
 b_{q,i} &= \min d_{q,i} ; q = 1, \dots, k, \\
 d_{q,i} &= \frac{1}{n_q} \sum_{j=1}^{n_q} d(x_i, y_j), \\
 a_{p,i} &= \frac{1}{n_p} \sum_{k=1}^{n_p} d(x_i, x_k).
 \end{aligned}$$

3.14.2 Davies-Bouldin Index

Indeks validitas Davies-Bouldin (DB) menghitung rata-rata nilai setiap titik pada himpunan data. Perhitungan nilai setiap titik adalah jumlah nilai *compactness* yang dibagi dengan jarak antara kedua titik pusat *cluster* sebagai *separation*.

Jumlah *cluster* terbaik ditunjukkan dengan nilai DB yang semakin kecil (Davies & Bouldin, 1979). Misalkan terdapat suatu himpunan data dengan k buah *cluster*, terdapat n_p buah titik pada *cluster* p dan n_q buah titik pada *cluster* q dengan titik pusatnya masing-masing adalah c_p dan c_q , sehingga M_{pq} adalah jarak antara titik pusat *cluster* p dan *cluster* q , S_p dan S_q berturut-turut merupakan rata-rata jarak setiap titik pada *cluster* p dan

BAB 3

q ke titik pusatnya pada *cluster* yang terkait, yaitu c_p dan c_q , dengan perhitungan indeks validitas DB dapat dilihat pada Persamaan 3.6.

$$\begin{aligned}DB &= \frac{1}{k} \sum_{p=1}^k R_p, \\R_p &= \max R_{p,q}, \quad p \neq q, \\R_{p,q} &= \frac{(S_p + S_q)}{M_{pq}}, \quad \dots\dots\dots(3.6) \\S_p &= \frac{1}{n_p} \sum_{i=1}^{n_p} d(x_i, c_p), \\S_q &= \frac{1}{n_q} \sum_{j=1}^{n_q} d(y_j, c_q), \\M_{pq} &= d(c_p, c_q).\end{aligned}$$

3.14.3 Calinski Harabasz Index

Indeks validitas Calinski-Harabasz (CH) menghitung perbandingan antara nilai *Sum of Square between cluster* (SSB) sebagai *separation* dan nilai *Sum of Square within cluster* (SSW) sebagai *compactness* yang dikalikan dengan faktor normalisasi, yaitu selisih jumlah data dengan jumlah *cluster* dibagi dengan jumlah *cluster* dikurang satu. Jumlah *cluster* terbaik ditunjukkan dengan semakin besar nilai CH (Baarsch & Celebi, 2012). Misalkan terdapat suatu himpunan data dengan k buah *cluster* dan

BAB 3

N buah titik data, misal C_l adalah *cluster* ke - l dengan x_i adalah titik ke - i pada *cluster* ke - l , N_l adalah jumlah titik pada *cluster* ke - l , dan \bar{x}_l adalah titik pusat *cluster* ke - l , maka perhitungan indeks validitas CH dapat dilihat pada Persamaan 3.7.

$$CH = \frac{\text{trace}(SSB)}{\text{trace}(SSW)} \times \frac{N - k}{k - 1},$$

$$SSW = \sum_{l=1}^k \sum_{x_i \in C_l} (x_i - \bar{x}_l)(x_i - \bar{x}_l)^T, \dots\dots\dots(3.7)$$

$$SSB = \sum_{l=1}^k N_l(\bar{x}_l - \bar{x})(\bar{x}_l - \bar{x})^T,$$

$$\bar{x}_l = \frac{1}{N_l} \sum_{x_i \in C_l} x_i,$$

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i.$$