

BAB II

LANDASAN TEORI

2.1. Tinjauan Pustaka

Pada dasarnya *K-means clustering* merupakan salah satu metode data *clustering* non-hirarki yang mengelompokkan data berdasarkan bentuk satu atau lebih *cluster*. Penelitian ini dilakukan dengan merujuk dari beberapa hasil penelitian sebelumnya yang menggunakan Algoritma *K-means clustering* diantaranya adalah Analisis Data E-Absensi untuk Menganalisis Perbandingan Pola Disiplin Kerja menggunakan Algoritma *Clustering K-Means* (Atmojo, Reksa Suhud Tri dkk., 2019). Penelitian ini mengolah data yang ada pada sistem e-absensi menggunakan algoritma *K-Means* $K = 3$ menggunakan perangkat lunak *Orange*. Hasil penelitian menghasilkan hasil analisis pola perilaku kedisiplinan kerja dengan melihat presentase telat dan jumlah kehadiran yang diambil dari data E-Absensi SD Negeri X Taman Fajar Kecamatan Purbolinggo dan Puskesmas Y yang berada di Desa Negara Nabung, Kabupaten Lampung Timur sehingga dapat menjadi pertimbangan bagi instansi terkait peningkatan tingkat kedisiplinan kerja.

Penelitian selanjutnya adalah Implementasi *K-Means Clustering* pada Rapidminer untuk Analisis Daerah Rawan Kecelakaan (Rahmat C.T.I., Brilian dkk., 2017). Penelitian ini menganalisis tentang data kecelakaan menggunakan aplikasi RapidMiner dapat mengekstraksi beberapa informasi yang dibutuhkan untuk mengelompokkan. Kemudian data kecelakaan dibagi menjadi 3 buah kelompok/*cluster* dari 500 contoh data kecelakaan. Data dianalisis menggunakan algoritma *K-Means Clustering* dengan bantuan aplikasi RapidMiner Studio. Hasil analisis menunjukkan frekuensi tingkat kecelakaan di tiap lokasi beserta waktu-waktu rawan yang berpotensi terjadi kasus kecelakaan.

Penelitian berikutnya yaitu Mengetahui Profil Pegawai BPJT Dilihat dari Tingkat Kehadiran Dengan Metode *Clustering (K-Means)* (Hapsari dan Sadikin, 2018). Analisis dilakukan terhadap rekap absensi pegawai tahun 2017 di lingkungan Badan Pengatur Jalan Tol (BPJT). *Dataset* diambil dari rekap absensi

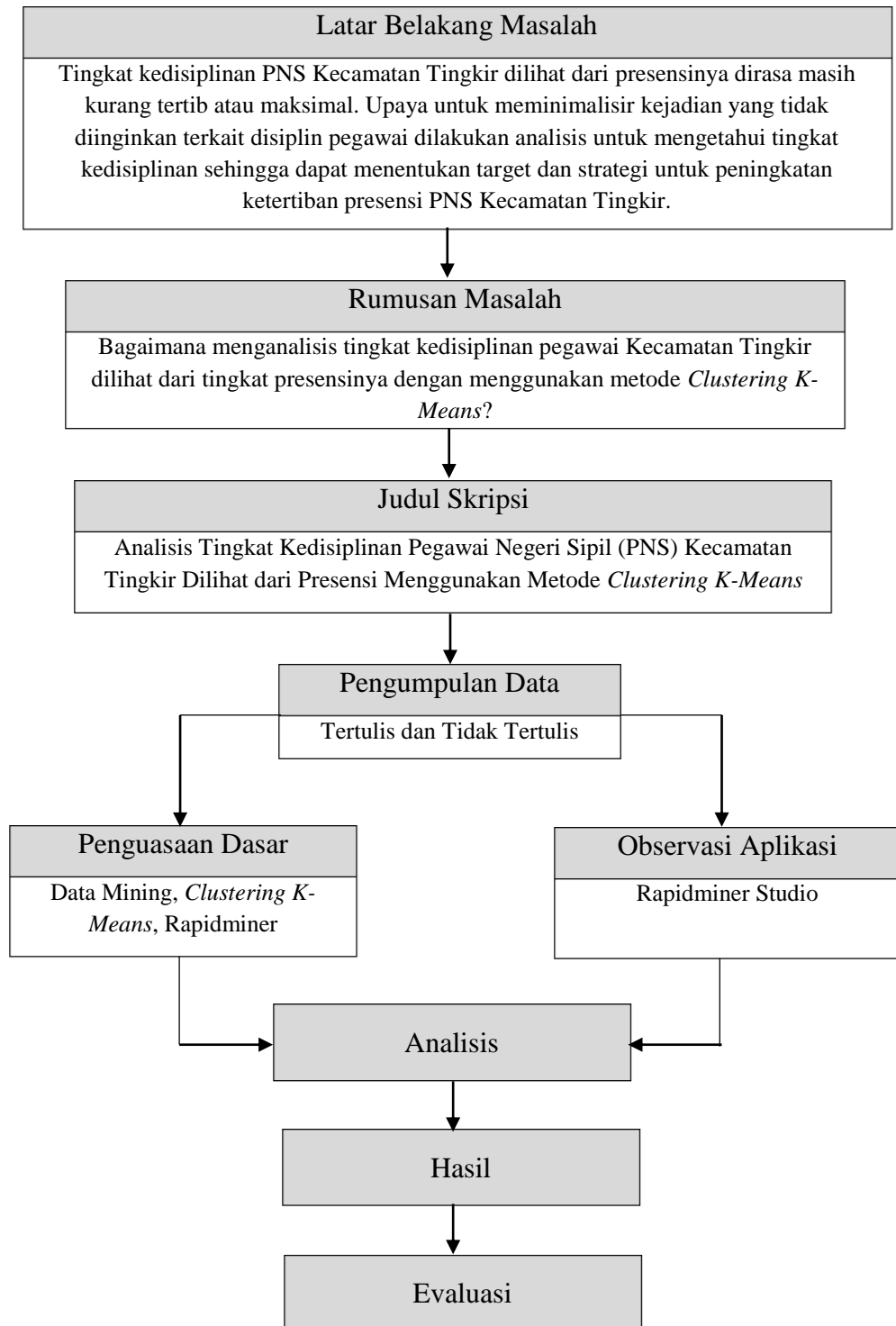
pegawai digunakan untuk mengetahui profil pegawai BPJT berdasarkan tingkat kehadirannya. Pengujian *dataset* dilakukan dengan metode *clustering* dan menggunakan *algoritma k-means* yang terdapat pada aplikasi Rapidminer Studio 8.1. Agar memperoleh hasil terbaik maka dilakukan pengujian dalam pembagian kelompoknya, yaitu dengan mengganti nilai *k* dari 2 sampai 10. Berdasarkan beberapa pengujian dengan menggunakan parameter *k* diperoleh hasil yang paling baik adalah dengan *k=6*. Prinsipnya adalah adanya variasi hasil pada kolom persentase kehadiran (*Per_kehadiran*), terutama untuk mengetahui pegawai yang nilai kehadirannya paling rendah. Hasilnya diperoleh 6 kelompok dengan masing-masing karakteristik yang hampir sama, diantaranya 5 kelompok pegawai yang sudah baik tingkat disiplinnya dan 1 kelompok pegawai yang perlu ditingkatkan.

Penelitian selanjutnya oleh Hardika Khusnuliawati pada tahun 2018 tentang Algoritma Pengelompokan Menggunakan *Self-Organizing Map* dan *K-Means* Pada Data Sumber Daya Manusia Provinsi Indonesia (Khusnuliawati, 2018). Penelitian ini menggunakan metode pengelompokan data yang merupakan gabungan dari SOM dan *K-Means* dapat diimplementasikan pada permasalahan dunia nyata untuk mengelompokkan kondisi suatu wilayah berdasarkan informasi demografisnya. Algoritma SOM merupakan algoritma yang diimplementasikan pada tahap pertama untuk memperoleh visualisasi dari hasil pengelompokan provinsi di Indonesia. Sedangkan tahap kedua diujikan algoritma *K-Means* untuk memperjelas hasil pengelompokan data yang kemudian dievaluasi menggunakan algoritma *Silhouette*. Hasil uji coba yang dilakukan menunjukkan pengelompokan terbaik dari 33 provinsi di Indonesia berdasarkan informasi sumber daya manusia yaitu sejumlah dua hingga tiga kelompok dengan rata-rata nilai *silhouette value* yang dihasilkan yaitu 0.6238 dan 0.6116.

2.2. Kerangka Pemikiran

Kerangka pemikiran merupakan garis besar dari langkah – langkah penelitian yang sedang dilakukan, kerangka pemikiran dijadikan acuan untuk melakukan tahap – tahap yang sedang dilakukan dalam penelitian.

Kerangka pemikiran dalam penyusunan skripsi dapat dilihat pada Gambar 2.1.



Gambar 2.1. Diagram Kerangka Pemikiran

Berikut ini adalah penjelasan Gambar 2.1 yang merupakan kerangka pemikiran yang dijalankan dalam penelitian ini.

1) Latar Belakang Masalah

Munculnya masalah yang dihadapi Kecamatan Tingkir Kota Salatiga menjadikan dasar analisis ini dibuat.

2) Rumusan Masalah

Latar belakang masalah yang ada menjadikan rumusan masalah yang akan dijadikan sebagai bahan analisis.

3) Judul Skripsi

Judul yang sesuai untuk menangani masalah yang dihadapi oleh Kecamatan Tingkir Kota Salatiga.

4) Pengumpulan Data (Tertulis dan Tidak Tertulis)

Pengumpulan data dilakukan baik dengan tanya-jawab (*interview*), observasi, maupun studi literatur di perpustakaan.

5) Penguasaan Dasar (Data Mining, *Clustering K-Means*, Rapidminer)

Tahap untuk mempelajari dasar-dasar *Data Mining*, *Clustering K-Means*, Rapidminer agar lebih menguasai materi dan aplikasi yang akan digunakan untuk melakukan analisis.

6) Observasi Aplikasi (Rapidminer Studio)

Merupakan tahap pengamatan sampel aplikasi yang telah ada, jurnal, buku, maupun karya ilmiah untuk kajian yang dapat dijadikan referensi untuk melakukan analisis.

7) Analisis

Penjelasan tentang prosedur analisis tingkat kedisiplinan PNS Kecamatan Tingkir dilihat dari presensi menggunakan metode *Clustering K-Means* yang berkaitan dengan landasan teori yang mendukung dalam melakukan analisis.

8) Hasil

Menghasilkan sebuah dokumen analisis tingkat kedisiplinan PNS Kecamatan Tingkir dilihat dari presensi menggunakan metode *clustering k-*

means yang mampu mengelompokkan data tingkat kedisiplinan PNS pada Kecamatan Tingkir.

9) Evaluasi

Hasil analisis tingkat kedisiplinan PNS Kecamatan Tingkir dilihat dari presensi menggunakan metode *clustering k-means* nanti dapat digunakan dalam menentukan target dan strategi untuk peningkatan ketertiban presensi pegawai Kecamatan Tingkir.

2.3. Landasan Teori

2.3.1. Pengertian Analisis

Menurut Spradley (Sugiyono, 2015) mengatakan bahwa analisis adalah sebuah kegiatan untuk mencari suatu pola selain itu analisis merupakan cara berpikir yang berkaitan dengan pengujian secara sistematis terhadap sesuatu untuk menentukan bagian, hubungan antar bagian dan hubungannya dengan keseluruhan. Analisis adalah suatu usaha untuk mengurai suatu masalah atau fokus kajian menjadi bagian-bagian (*decomposition*) sehingga susunan/tatanan bentuk sesuatu yang diurai itu tampak dengan jelas dan karenanya bisa secara lebih terang ditangkap maknanya atau lebih jernih dimengerti duduk perkaranya (Satori dan Komariyah, 2014).

Nasution dalam (Sugiyono, 2015) melakukan analisis adalah pekerjaan sulit, memerlukan kerja keras. Tidak ada cara tertentu yang dapat diikuti untuk mengadakan analisis, sehingga setiap peneliti harus mencari sendiri metode yang dirasakan cocok dengan sifat penelitiannya. Bahan yang sama bisa diklasifikasikan berbeda.





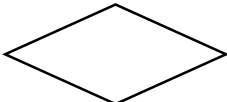

Jadi dapat ditarik kesimpulan bahwa analisis merupakan penguraian suatu pokok secara sistematis dalam menentukan bagian, hubungan antar bagian serta hubungannya secara menyeluruh untuk memperoleh pengertian dan pemahaman yang tepat.

2.3.2. Diagram Alir (*Flowchart*)

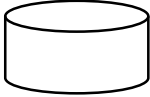

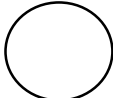
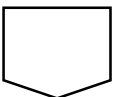
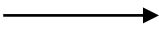
Flowchart atau diagram alir merupakan sebuah diagram dengan simbol-simbol grafis yang menyatakan aliran algoritma atau proses berjalannya program. *Flowchart* adalah suatu diagram yang berupa simbol-simbol dan dapat menunjukkan alur data serta operasi yang terjadi pada suatu sistem. Bagan alur digunakan sebagai alat bantu komunikasi dan dokumentasi (Arianto, 2016).

Bagan alur sistem digambarkan dengan menggunakan simbol-simbol yang tampak pada Tabel 2.1.

Tabel 2.1. Simbol dan Keterangan *Flowchart*

Simbol	Keterangan
 Terminal	Menunjukkan awal atau akhir aliran proses.
 Proses	Melambangkan proses yang dilakukan oleh komputer.
 Proses	Melambangkan proses atau operasi yang dilakukan secara manual.
 Proses	Melambangkan proses yang dilakukan oleh manusia dan komputer seperti memasukkan data ke dalam komputer (<i>input</i>).
 Decision	Melambangkan pengambilan keputusan bagaimana alur dalam <i>flowchart</i> berjalan selanjutnya berdasarkan kriteria atau pernyataan tertentu.
 Stored Data	Melambangkan informasi yang disimpan ke dalam media penyimpanan umum.

Lanjutan Tabel 2.1

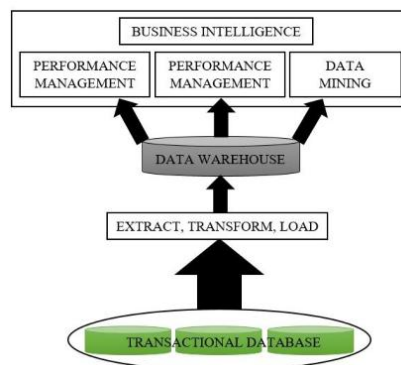
Simbol	Keterangan
 Database	Melambangkan basis data atau <i>database</i> .
 Predefined Process	Melambangkan proses yang telah kita jelaskan lebih rinci di dalam <i>flowchart</i> tersendiri.
 Koneksi	Melambangkan koneksi yang digunakan pada satu halaman, sebagai pengganti garis penghubung.
 Koneksi	Melambangkan koneksi yang digunakan pada halaman lain, sebagai pengganti garis penghubung.
 Garis	Melambangkan garis penghubung aliran algoritma.

2.3.3. Data Mining

Data mining memiliki pengertian lain yaitu *knowledge discovery* ataupun *pattern recognition* merupakan suatu istilah yang digunakan untuk mendapatkan pengetahuan yang tersembunyi dari kumpulan data yang berukuran sangat besar. Tujuan utama *data mining* adalah untuk menemukan, menggali, atau menambang pengetahuan dari data atau informasi yang kita miliki.

Data mining merupakan 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. Kegunaan *data mining* adalah untuk menspesifikasikan pola yang harus ditemukan dalam tugas *data mining*. Salah satu teknik dalam *data mining* yaitu untuk membangun sebuah model dalam penelusuran data.

Data mining merupakan suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam *database*. Menurut Turban, dkk dalam bukunya Kusri data mining merupakan suatu proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait berbagai *database* besar (Daud, 2017).



Gambar 2.2. Posisi *data mining* dalam *business Intelligence*

(Sumber : Daud, 2017:7)

Gambar 2.2 mengilustrasikan posisi antara *data mining* dan *data warehouse*. *Data mining* adalah suatu bidang yang sepenuhnya menggunakan apa yang dihasilkan oleh *data warehouse* untuk menangani masalah pelaporan dan manajemen data. Sedangkan *data warehouse* bertugas untuk menarik/meng-*query* data dari *database* mentah untuk menghasilkan data yang nantinya akan digunakan oleh bidang yang menangani manajemen, pelaporan, dan *data mining*.

Secara umum *data mining* memiliki empat tugas utama :

a. Klasifikasi (*Classification*)

Klasifikasi bertujuan untuk mengklasifikasikan item data menjadi satu dari beberapa kelas standar. Beberapa algoritma yang digunakan dalam mengklasifikasi data diantaranya pohon keputusan, *nearest neighbor*, *naive bayes*, *neural networks* dan *support vector machines*.

b. Regresi (*Regression*)

Suatu pemodelan dan investigasi hubungan dua atau lebih variabel. Dalam analisis regresi ada satu lebih variabel *independent* / prediktor yang

biasa diwakili dengan notasi x dan satu variabel *respons* yang biasa diwakili dengan notasi.

c. Pengelompokan (*Clustering*)

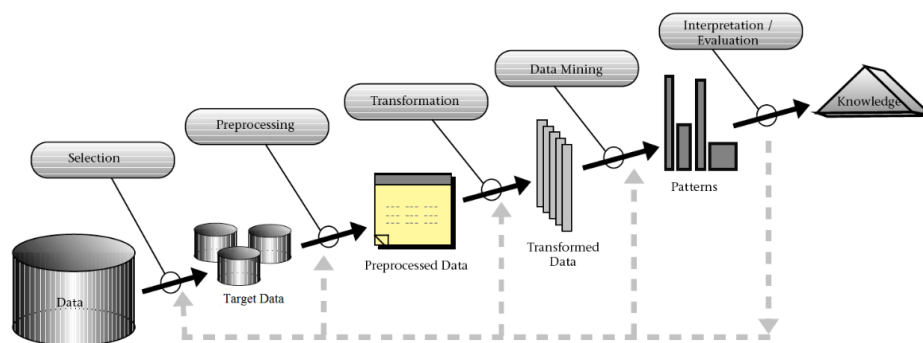
Suatu metode pengelompokan data ke dalam *cluster* sehingga dalam setiap *cluster* berisi data yang semirip mungkin.

d. Pembelajaran Aturan Asosiasi (*Association Rule Learning*)

Suatu tugas untuk menemukan atribut-atribut yang “terjadi” bersamaan. Tugas asosiasi mencoba untuk menemukan aturan untuk mengkuantifikasi hubungan antara dua atau lebih atribut. Aturan asosiasi berbentuk “*If antecedent, then consequent*”, bersama-sama dengan ukuran *support* dan *confidence* yang berhubungan dengan aturan.

2.3.3.1 Tahapan *Data Mining*

Suatu rangkaian proses *data mining* dapat dibagi menjadi beberapa tahap yang bersifat interaktif. Tahap-tahap dapat diilustrasikan pada Gambar 2.3.



Gambar 2.3. Tahapan *Data Mining*

(Sumber : Daud, 2017:9)

Tahap-tahapan tersebut yaitu :

1) Pembersihan data

Suatu proses menghilangkan *noise* dan data dari *database* maupun hasil dari eksperimen yang tidak konsisten. Pada tahap ini dilakukan penghapusan terhadap data-data yang tidak memiliki kelengkapan atribut yang dibutuhkan

karena keberadaannya nantinya dapat mempengaruhi akurasi hasil dari *data mining* nantinya.

2) Integrasi data

Merupakan suatu gabungan data dari berbagai macam *database* ke dalam satu *database* baru. Integrasi data dapat dilakukan pada atribut-atribut yang memiliki entitas-entitas yang unik seperti atribut nama, nomor pelanggan, jenis, produk dan lainnya. Integrasi data perlu dilakukan secara cermat karena jika tidak bisa menghasilkan hasil yang menyimpang dan bahkan menyesatkan dalam pengambilan aksi nantinya.

3) Seleksi data

Seleksi data merupakan suatu proses dimana data dari *database* yang berkaitan diambil untuk di analisis. Tidak semuanya data yang ada pada *database* dipakai, hanya data yang sesuai untuk dianalisis itulah yang akan diambil.

4) Transformasi data

Tahap ini data diubah dan dikonsolidasikan ke dalam bentuk yang sesuai untuk diproses dalam *data mining* dengan melakukan ringkasan atau penggabungan operasi. Sebagai contoh beberapa metode standar seperti analisis asosiasi dan *clustering* hanya bisa menerima *input* data kategorikal. Karenanya data berupa angka numerik yang berlanjut perlu dibagi-bagi menjadi beberapa interval.

5) Proses *mining*

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

6) Evaluasi pola

Tahap ini merupakan suatu indentifikasi kebenaran pola yang pada dasarnya pengetahuan guna untuk menilai apakah hipotesa yang ada memang tercapai. Jika hasil yang diperoleh belum tercapai atau ternyata tidak sesuai maka dapat memperbaiki proses *data mining* tersebut sampai hasil yang diinginkan tercapai, mencoba menggunakan metode lain yang lebih sesuai atau hanya

menerima hasil ini sebagai satuan hasil yang diluar dugaan yang mungkin bermanfaat.

7) Presentasi pengetahuan

Merupakan tahap terakhir dimana penemuan penyajian pengetahuan dapat direpresentasikan secara visualisasi oleh pengguna untuk memperoleh pengetahuan yang didapat.

2.3.3.2 Teknik – Teknik *Data Mining*

Teknik – teknik dalam *data mining* di bedakan berdasarkan tugas yang dilakukan. Teknik - teknik tersebut antara lain sebagai berikut (Han dan Kamber, 2012):

1) Deskripsi

Secara umum digunakan oleh para peneliti dan analisis ingin mencoba untuk menggambarkan pola yang terdapat dalam data dan cara mendeskripsikannya.

2) Estimasi

Terdapat kesamaan antara teknik estimasi dengan teknik klasifikasi, yang menjadi pembeda kedua teknik tersebut pada variabelnya. Teknik estimasi variabel yang digunakan yaitu numerik, sedangkan dalam teknik klasifikasi variabel yang digunakan kategori.

3) Prediksi

Tahapan prediksi mempunyai kesamaan dengan estimasi dan klasifikasi, prediksi nilai dari hasil akan ada di masa yang akan datang.

4) Klasifikasi

Menerapkan suatu teknik yang bisa mengklasifikasikan satu objek berdasar atribut-atributnya. Kelas target sudah tersedia dalam data sebelumnya, sehingga data yang ada agar klasifikator bisa mengklasifikasikan sendiri atau dapat membentuk kelas yang tidak diketahui labelnya. Ada beberapa algoritma yang digunakan dalam klasifikasi antara lain adalah *CART*, *K-Nearest Neighbor*, *C.45*, dan *Naive Bayes*.

5) Pengklusteran (*Clustering*)

Merupakan suatu proses pengelompokan *record* yang akan membentuk kelas yang terdiri dari objek – objek yang memiliki kemiripan. Kumpulan *cluster* yang membentuk kelas memiliki kemiripan antara *record* satu dengan yang lain atau bernilai maksimal dalam satu *cluster* dan tidak memiliki kemiripan dengan *record-record* atau bernilai minimal dalam *cluster* lain. Beberapa algoritma yang digunakan dalam *clustering* adalah *K-Means* dan *Fuzzy C-Means*.

6) Asosiasi

Tahapan asosiasi yaitu menentukan atribut berdasarkan hasil analisa dalam satu waktu. Asosiasi sering disebut dengan *market basket* analisis dalam dunia bisnis. Algoritma yang digunakan dalam asosiasi diantaranya yaitu FP-Growth dan Apriori.

2.3.4. *Clustering*

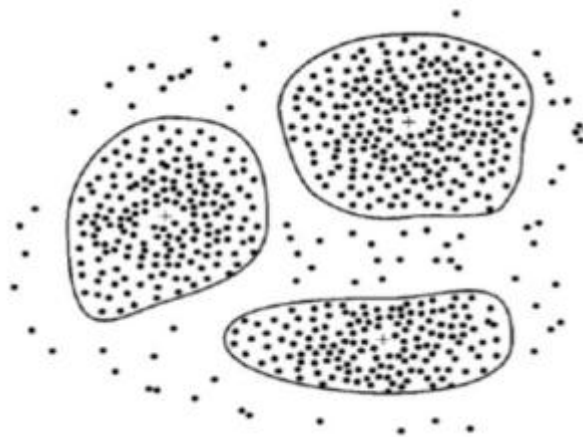
Pengelompokan atau *Clustering* merupakan suatu teknik *data mining* yang digunakan untuk menganalisis data untuk memecahkan permasalahan dalam pengelompokan data atau lebih tepatnya mempartisi dari *dataset* ke dalam *subset*. Teknik *clustering* targetnya adalah untuk kasus pendistribusian (objek, orang, peristiwa dan lainnya) ke dalam satu kelompok, hingga derajat tingkat keterhubungan antar anggota *cluster* yang sama adalah kuat dan lemah antara anggota *cluster* yang berbeda (Wardhani, 2014).

Baskoro menyatakan *Clustering* atau klasterisasi merupakan satu alat bantu pada *data mining* yang mempunyai tujuan untuk mengelompokkan objek ke dalam *cluster*. *Cluster* merupakan kelompok atau kumpulan objek – objek data yang mirip antara satu dengan yang lain dalam *cluster* yang sama dan sejenis terhadap objek– objek yang berbeda *cluster* (Rohmawati W, Nurul dkk., 2015).

Pada dasarnya *clustering* merupakan suatu proses pengelompokan dari sekian banyak data untuk menemukan kelompok atau identifikasi kelompok obyek yang hampir sama. Ketidakmiripan dan kesamaan dinilai dari nilai atribut yang menggambarkan objek dan sering melibatkan perlakuan jarak. *Clustering* berbeda

dengan *group* yang hanya mempunyai kondisi jika tidak ya pasti bukan kelompoknya. Tetapi kalau *cluster* tidak harus sama namun kedekatan atau kemiripan dari satu karakteristik populasi yang ada dengan menggunakan rumus jarak *ecludian*. Hal yang penting dalam proses pengklasteran adalah menyatakan sekumpulan pola ke kelompok yang sesuai yang berguna untuk menemukan kesamaan dan perbedaan sehingga dapat menghasilkan kesimpulan yang berharga.

Tujuan dari *clustering* data dapat dibedakan menjadi dua, pertama pengelompokan untuk pemahaman terbentuk harus menangkap struktur alami data yang bertujuan hanya sebagai proses awal untuk kemudian dilanjutkan dengan pekerjaan inti seperti peringkasan atau *summarization* (rata-rata , standar deviasi), pemberian label kelas pada setiap kelompok bertujuan untuk digunakan sebagai data latih klasifikasi. Kedua pengelompokan untuk penggunaan bertujuan untuk mencari *prototipe* kelompok yang paling representatif terhadap data memberikan abstraksi dari setiap objek data dalam kelompok dimana sebuah data terletak di dalamnya.



Gambar 2.4. *Clustering*

(Sumber : Daud, 2017:16)

Ilustrasi *clustering* pada gambar di atas menjelaskan dari pelanggan suatu toko dapat dikelompokkan menjadi beberapa *cluster* dengan pusat *cluster* ditunjukkan oleh tanda positif (+).

2.3.5. Algoritma *K-Means*

K-Means merupakan algoritma *clustering* yang pertama kali diperkenalkan oleh James B MacQueen pada tahun 1976. Metode ini merupakan suatu metode *clustering non-heirarchial* yang umum digunakan yang relatif sederhana untuk mengelompokkan data dalam jumlah besar.

K-Means merupakan metode klasterisasi yang sering digunakan diberbagai bidang karena penggunaannya sederhana, mudah untuk diimplementasikan, mampu untuk mengklaster data yang besar. Algoritma *K-Means* merupakan metode berbasis jarak yang membagi data kedalam sejumlah *cluster* dan dalam setiap tahapan tertentu setiap objek harus masuk dalam kelompok, pada tahap selajutnya objek dapat berpindah ke kelompok lain. Algoritma ini pada dasarnya melakukan proses *clustering* tetapi tergantung dari data yang didapat dan konklusi yang dicapai. Maka dari itu algoritma *K-Means* mempunyai aturan yaitu ada jumlah *cluster* yang akan diinputkan dan hanya dapat memiliki atribut yang bertipe numerik.

Pada awalnya dalam algoritma *K-Means* melakukan pengambilan sebagian dari banyaknya populasi untuk dijadikan *cluster* awal. Ada banyak cara dalam memberi nilai awal misalnya dengan pengambilan data sampel awal dari objek. Pusat *cluster* dipilih secara acak yang berada dari beberapa populasi data. Setelah mendapatkan pusat *cluster* awal, algoritma *K-Means* melakukan pengujian masing-masing komponen ke salah satu pusat *cluster* yang telah didefinisikan jarak minimumnya antar komponen dengan tiap-tiap pusat *cluster*. Posisi pusat *cluster* akan melakukan perhitungan kembali sampai semua komponen data dapat digolongkan ke dalam setiap *cluster* dan akan membentuk posisi *cluster* baru.

Algoritma *K-Means* terdapat 3 komponen didalamnya, yaitu :

- 1) Jumlah *Cluster* K

Metode ini jumlah k harus ditentukan dulu, setelah jumlah k didapatkan dengan melalui pendekatan metode hirarki dapat melakukan pengambilan *cluster* awal. Aturan khusus dalam menentukan jumlah *cluster* k bahkan tidak ada, namun ada juga jumlah *cluster* yang diinginkan sesuai dengan kebutuhan subjektif seseorang.

2) *Cluster* Awal

Ada banyak pendapat saat melakukan pengambilan *cluster* awal untuk metode *K-Means* misalnya pemilihan terhadap interval dari jumlah setiap observasi, melalui pendekatan salah satu metode hirarki dan ada juga dengan melalui pemilihan *cluster* secara acak dari sekumpulan observasi. Dengan adanya beberapa cara pengambilan *cluster* awal tersebut dapat memungkinkan solusi terbaik yang dihasilkan.

3) Ukuran Jarak

Tahapan ini ukuran jarak juga penting dalam menempatkan observasi ke dalam *cluster* berdasarkan nilai *centroid* terdekat. *Euclidian Distance* adalah jarak yang digunakan untuk mengukur jarak dalam metode *K-Means*.

Berikut ini adalah rumus *Euclidian Distance*:

$$D_{ik} = \sqrt{\sum_{j=1}^m (X_{ij} - C_{kj})^2}$$

Keterangan :

D_{ik} = titik, dokumen/jarak data ke-i

m = jumlah Variabel

X_{ij} = data yang akan dilakukan pengklasteran

C_{kj} = pusat dari *cluster*

Jarak yang terpendek antara *centroid* dengan dokumen menentukan posisi *cluster* suatu dokumen. Berikut merupakan rumus penentuan *centroid* baru :

$$V_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i} X_{kj}$$

Keterangan :

V_{ij} = *centroid* / rata-rata *cluster* ke-i untuk *variable* ke-j

N_i = jumlah data yang menjadi anggota *cluster* ke-i

i,k = indeks dari *cluster*

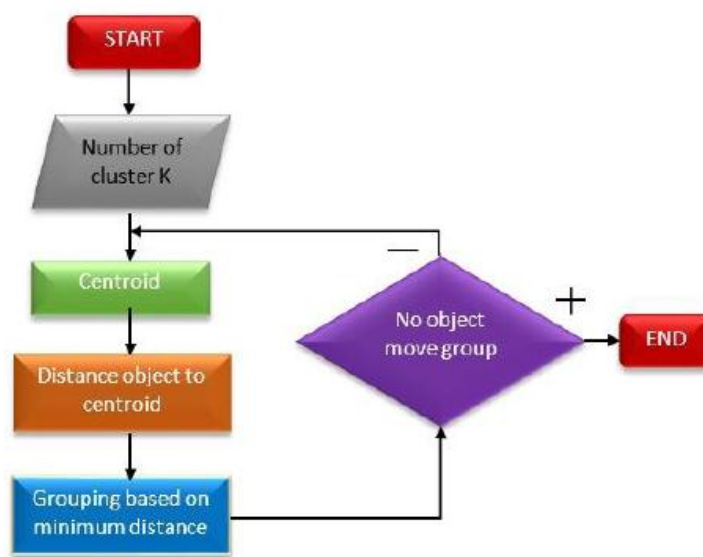
j = indeks dari variabel

X_{kj} = nilai data ke-k yang ada di dalam *cluster* tersebut untuk *variable* ke-j

Algoritma *K-Means* bersifat *partitional clustering* yaitu dengan cara membagi himpunan, objek data ke dalam sub himpunan (*cluster*) yang tidak *overlap*, sehingga hasil dari setiap objek data berada tepat dalam satu *cluster*.

2.3.5.1 Tahapan Algoritma *K-Means*

Menurut (Daud, 2017) menyatakan tahapan algoritma *K-Means* seperti pada Gambar 2.5.



Gambar 2.5. Tahapan Algoritma *K-Means*

(Sumber : Daud, 2017:19)

Berikut penjelasan dari Gambar 2.5 :

- Tentukan K sebagai jumlah *cluster* yang ingin dibentuk.
- Bangkitkan K *centroids* (titik pusat *cluster*) awal secara random.
- Hitung jarak setiap data ke masing-masing *centroids*.
- Setiap data memilih *centroids* yang terdekat.
- Tentukan posisi *centroids* baru dengan cara menghitung nilai rata-rata dari data-data yang terletak pada *centroids* yang sama.
- Kembali ke langkah 3 jika posisi *centroids* baru dengan *centroids* lama tidak sama.

2.3.6. RapidMiner

RapidMiner merupakan perangkat lunak yang bersifat terbuka (*open source*). RapidMiner adalah sebuah 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 memiliki kurang lebih 500 operator *data mining*, termasuk operator untuk *input*, *output*, *data preprocessing* dan visualisasi. 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 java sehingga dapat bekerja di semua sistem operasi.

RapidMiner sebelumnya bernama YALE (*Yet Another Learning Environment*), dimana versi awalnya mulai dikembangkan pada tahun 2001 oleh RalfKlinkenberg, Ingo Mierswa, dan Simon Fischer di *Artificial Intelligence Unit* dari *University of Dortmund*. RapidMiner didistribusikan di bawah lisensi AGPL (*GNU Affero General Public License*) versi 3. Hingga saat ini telah ribuan aplikasi yang dikembangkan menggunakan RapidMiner di lebih dari 40 negara. RapidMiner sebagai *software open source* untuk *data mining* tidak perlu diragukan lagi karena *software* ini sudah terkemuka di dunia. RapidMiner menempati peringkat pertama sebagai *software data mining* pada *polling* oleh KDnuggets, sebuah portal *data-mining* pada 2010-2011.

RapidMiner menyediakan GUI (*Graphic User Interface*) untuk merancang sebuah *pipeline* analitis. GUI ini akan menghasilkan file XML (*Extensible Markup Language*) yang mendefinisikan proses analisis keinginan pengguna untuk diterapkan ke data. File ini kemudian dibaca oleh RapidMiner untuk menjalankan analisis secara otomatis (Shella, 2015).