BAB II
TINJAUAN PUSTAKA

2.1 Kajian Literatur

Penelitian-penelitian terkait *clustering* sudah banyak dilakukan, berikut beberapa penelitian terkait *clustering* dengan menggunakan algoritma *K-Means* dan *K-Medoids* sebagai berikut:

Pada tahun 2021, ada penelitian yang berjudul *Clustering* Kanker Serviks Berdasarkan Perbandingan Euclidean dan Manhattan Menggunakan Metode *K-Means*[5]. Pada penelitian tersebut, data yang digunakan sebanyak 401 data dengan 15 atribut yang diperoleh dari data pasien gynekologi RSPAD Gatot Soebroto Jakarta. Sebanyak 205 pasien yang terdeteksi kanker serviks, sedangkan 196 pasien lainnya tidak terkena kanker serviks. Pada penelitian ini, proses pemodelan menggunakan aplikasi WEKA dengan memanfaatkan tool classification Via *Clustering*. Nilai akurasi yang didapatkan adalah 79,30% dengan kurva ROC 79,17% pada *K-Means* Euclidean Metric, sedangkan *K-Means* Manhattan Metric sebesar 67,83% dengan kurva ROC 65,94%.

Pada tahun 2021, ada sebuah penelitian yang dilakukan oleh Tanty, Budi Serasi Ginting dan Magdalena Simanjuntak dengan judul Pengelompokan Penyakit Pada Pasien Berdasarkan Usia Dengan Metode *K-Means* Clustering (Studi Kasus : Puskesmas Bahorok)[6]. Pada penelitian ini menggunakan 20 data dengan 4 atribut. Data yang digunakan tidak memerlukan tahap pre-processing dan tidak menggunakan metode validasi. Hasil yang didapatkan dalam penelitian ini adalah Dari 20 data diproleh 3 grup, maka dapat disimpulkan sebagai berikut: Grup 1 Centroid 1 :(2,2 1,4 2,2) terdapat 5 data. Berdasarkan perhitungan diatas dapat diketahui bahwa pada claster 1 kelompok 1 merupakan pasiennya adalah Anak- Anak. Grup 2 Centroid 2 : (

1 1,6 1,6) terdapat 3 data. Berdasarkan perhitungan diatas dapat disimpulkan pada claster 2 kelompok 2 merupakan pasiennya adalah orang Dewasa. Grup 3 Centroid 3 : (1,75 1,58 6) terdapat 12 data. Berdasarkan perhitungan diatas dapat disimpulkan pada claster 3 kelompok 3 merupakan pasiennya adalah orang Lansia. Kesimpulan dari penelitian ini adalah Pengelompokan dengan metode *K-Means* dapat menghasilkan jumlah cluster yang sama dengan jumlah data yang berbeda – beda tanpa harus memiliki data yang sama, kemudian dengan dibangunnya sistem ini untuk mempermudahkan user dalam mengelompokan penyakit pada pasien berdasarkan usia secara efektif dan efisien khususnya untuk Staff Pegawai dan Administrasi, dan dengan metode *K-Means* sangatlah mempermudahkan user dalam mengelompokan suatu data hanya dengan memiliki karakteristik yang sama.

Penelitian selanjutnya adalah Perbandingan Algoritma *K-Means*, X-Means Dan *K-Medoids* Untuk Klasterisasi Awak Kabin Lion Air[7] yang dilakukan oleh Ahmad Jurnaidi Wahidin, Dana Indra Sensuse pada tahun 2021. Pada penelitian ini, data yang digunakan berasal dari Divisi Flight Operation Support (FOS) sebanyak 100 data dengan 7 atribut. Tujuan dari penelitian ini adalah membandingkan tiga algoritma dengan menghitung nilai Davies-Bouldin Index (DBI), pada tahapan pengolahan data dengan menghilangkan missing value dan menentukan atribut maka menghasilkan 100 data, pada tahapan pemodelan hasil paling optimum yang didapat dengan menggunakan algoritma *K-Means* adalah 4 klaster dan 6 atribut ditunjukan dengan nilai DBI sebesar 0.792, sedangkan nilai DBI algoritma x-means sebesar 0.812 dan algoritma *K-Medoids* sebesar 1,700 sehinga *K-Means* menjadi algoritma terbaik pada penelitian ini.

Pada tahun 2021 ada penelitian dengan judul Perbandingan Tingkat Kepuasan Siswa Terhadap Pelayanan Sekolah Menggunakan Algoritma *K-Means* Dan *K-Medoids*[8]. Penelitian dilakukan oleh Maulana Abdur Rofik, Amril Mutoi

Siregar, Dwi Sulistya Kusumaningrum. Data diperoleh dari data pribadi SMK TI Muhammadiyah Cikampek dengan 509 data dan 5 atribut. Hasil pengelompokan memakai algoritma *K-Means* menghasilkan klaster puas sebanyak 276 siswa, klaster cukup puas sebanyak 216 siswa dan klaster kurang puas sebanyak 17 siswa. Lalu pada algoritma *K-Medoids* klaster puas sebanyak 324 siswa, klaster cukup puas sebanyak 11 siswa dan klaster kurang puas sebanyak 174 siswa. Perbedaan jumlah cluster pada kinerja tiap algoritma memiliki pola perhitungan yang berbeda pada masing-masing iterasi tergantung pada dataset yang digunakan serta titik centroid yang dijadikan perhitungan pada algoritma.

Penelitian lainnya adalah penelitian dengan judul Perbandingan Algoritma K Means dan K Medoids Untuk Clustering Kelas Siswa Tunagrahita[9] oleh Fitriana Harahap pada tahun 2021. Dengan menggunakan 36 data dengan 6 atribut yang berasal dari data pribadi Sekolah Luar Biasa C Muzdalifah Medan. Penelitian ini bertujuan untuk mengetahui perbandingan hasil pengklasteran kelas siswa tunagrahita menggunakan metode *K-Means* dan *K-Medoids* Clustering. Cluster yang dihasilkan kedua metode adalah 3. Dengan metode *K-Means* Clustering terdapat 8 siswa tunagrahita ringan, 14 siswa tunagrahita sedang, dan 14 siswa tuna grahita berat. Sedangkan dengan metode *K-Medoids* Clustering dapat diketahui bahwa terdapat 7 siswa tunagrahita ringan, 19 siswa tunagrahita sedang, dan 10 siswa tunagrahita berat. Nilai DBI untuk validasi *K-Means* adalah 0,161 dan nilai DBI untuk validasi *K-Medoids* adalah 0,281. Dengan demikian, pengklasteran menggunakan metode *K-Means* Clustering memiliki hasil yang lebih baik dibandingkan dengan metode *K-Medoids* Clustering, karena menghasilkan nilai DBI yang lebih kecil yaitu 0,16.

Pada tahun 2017 ada sebuah penelitian yang berjudul *K-Means* Dan Fuzzy C-Means Pada Analisis Data Polusi Udara Di Kota X[10] oleh Sandi Fajar

Rodiyansyah. Dengan menggunakan 9358 data dan 10 atribut yang diperoleh dari Penelitian De Vito. Sebelum diuji, data melalui proses Data Cleaning dan Data Transformation. Metode yang digunakan pada validasi adalah Standar Deviasi. Rata-rata standar deviasi pada hasil clustering Fuzzy C-Means lebih kecil dari pada rata-rata standar deviasi pada hasil clustering *K-Means*. Dengan sebaran anggota cluster 0 sebanyak 2272 data, anggota cluster 1 sebanyak 973 data, anggota cluster 2 sebanyak 873 data, anggota cluster 3 sebanyak 1639 data dan anggota cluster 4 sebanyak 3600 data. Terlihat pada hasil proses clustering dengan fuzzy c-means bahwa parameter yang berpengaruh pada proses clustering ini adalah parameter nitrogen oksida (NOx), Non Metanic Hydro Carbon (NMHC) dan natrium dioksida (NO2). Sementara itu, parameter yang tidak berpengaruh pada proses cluster adalah absolute humidity (AH), relative humidity (RH) dan benzene (C6H6).

Pada tahun 2021, ada penelitian berjudul Penerapan Algoritma *K-Medoids* untuk Pengelompokan Penyakit di Pekanbaru Riau[11] oleh Tri Juninda, Mustakim, dan Elvia Andri. Data yang digunakan adalah Kuesioner online yang diisi oleh masyarakat Pekanbaru dengan 2811 data dan 4 atribut. ALgoritma yang digunakan adalah *K-Medoids*. Tahap pre-processing dalam penelitian ini meliputi Data Cleaning, Data Transformasi, dan Data Normalisasi. Berdasarkan penelitian yang dilakukan terhadap data penyakit yang sering diderita di Pekanbaru Riau didapatkan hasil 4 cluster sebagai pengelompokan terbaik dengan nilai Devies Bouildien Indeks sebesar 0,043 . Pada cluster 1 didapatkan 420 record dengan penyakit dominan adalah Maag sebesar 44,39%, cluster 2 didapatkan 349 record dengan penyakit dominan adalah Diare dan Sakit Perut sebesar 16,98%, pada cluster 3 didapatkan 794 record dengan penyakit dominan adalah Batuk dan Pilek sebesar 65,21% dan pada cluster 4 didapatkan 1248 record dengan penyakit dominan adalah Batuk dan Pilek sebesar 54,10%. Hasil dari penelitian ini menunjukkan bahawa

algoritma *K-Medoids* mampu melakukan pengelompokan terhadap penyakit di Pekanbaru Riau.

Penelitian yang berjudul Perbandingan Algoritma *K-Means* Dan *K-Medoids* Dalam Klasterisasi Produk Asuransi Perusahaan Nasional[12] pada tahun 2019 oleh Frenda Farahdinna, Irfan Nurdiansyah, Apriati Suryani dan Arief Wibowo. Data pada penelitian ini bersumber dari perusahaan asuransi berskala nasional sejak tahun 1996 hingga 2019. Data tersebut melalui tahap pra-pemrosesan data sebelum digunakan untuk melakukan proses klasterisasi pada beberapa atribut. Penelitian ini bertujuan untuk membandingkan kelompok mana yang paling optimal diantara metode *K-Means* dan *K-Medoids* dalam pengelompokan data produk asuransi perusahaan nasional pada cluster 1 sampai cluster 10 dengan menggunakan metode *K-Means* pada tabel 3 dan metode *K-Medoids*. Setelah melakukan analisis dan pengolahan data dengan membandingkan metode *K-Means* dan *K-Medoids* dalam klasterisasi produk asuransi, maka dapat disimpulkan bahwa nilai DBI yang diperoleh dari metode *K-Means* dan K-Medoid dengan eksperimen pembentukan cluster sebanyak sembilan menghasilkan nilai DBI terkecil pada metode *K-Means* adalah pada nilai k=5 yaitu 0,018. Sebagai pembanding kinerja, pembentukan cluster dengan metode *K-Medoids* memiliki nilai DBI terkecil pada k=2 yaitu sebesar 0,027. Dengan demikian maka pembentukan cluster yang paling optimal dalam klasterisasi produk asuransi perusahaan nasional adalah menggunakan metode *K-Means*.

Penelitian selanjutnya adalah Perbandingan Algoritma *K-Means* Dengan Algoritma Fuzzy C-Means Untuk Clustering Tingkat Kedisiplinan Kinerja Karyawan[13] oleh Nova Agustina, Prihandoko pada tahun 2018. Algoritma yang digunakan dalam penelitian ini adalah *K-Means* dan Fuzzy C-Means. Metode analisis data yang dipakai adalah Penelitian Tindakan. Tahap awal adalah melakukan pengumpulan data melalui observasi, wawancara dan studi

literatur. Selanjutnya adalah membandingkan Algoritma *K-Means* dengan Fuzzy C-Means. Validasi dari algoritma *K-Means* menggunakan Silhoutte Index dan untuk algoritma Fuzzy C-Means menggunakan Partition Coefficient Index (PCI). Berdasarkan hasil penelitian yang telah diperoleh, dapat disimpulkan bahwa hasil cluster dari data presensi karyawan menggunakan metode *K-Means* dan Fuzzy C-Means berbeda. Hal ini dapat dilihat dari jumlah cluster yang diperoleh dari kedua metode tersebut. Dilihat dari hasil validasi, Fuzzy C-Means dominan menghasilkan metode yang lebih baik, dengan nilai validasinya adalah 0,758 dikarenakan nilai validasinya lebih mendekati nilai 1, dibandingkan dengan metode *K-Means* dengan nilai validasinya adalah 0,528.

Kemudian, pada tahun 2019 pada penelitian yang berjudul Perbandingan Pengelompokan *K-Means* dan *K-Medoids* Pada Data Potensi Kebakaran Hutan/Lahan Berdasarkan Persebaran Titik Panas[14] oleh Athifaturrofifah, Rito Goejantoro, dan Desi Yuniarti, menguji 42 data dan 7 atribut dengan menggunakan algoritma *K-Means* dan *K-Medoids* yang kemudian hasilnya divalidasi dengan menggunakan Silhouette Coefficient. Berdasarkan hasil penelitian dan pembahasan, maka kesimpulan yang dapat diambil adalah hasil pengelompokan yang terbentuk dengan menggunakan algoritma *K-Means* yaitu untuk cluster 1 beranggotakan 37 wilayah, sedangkan untuk cluster 2 sebanyak 5 wilayah. Kemudian hasil pengelompokan yang terbentuk dengan menggunakan algoritma *K-Medoids* yaitu untuk cluster 1 beranggotakan 36 wilayah, sedangkan untuk cluster 2 sebanyak 6 wilayah. Dan selanjutnya untuk nilai Silhouette Coefficient dengan metode *K-Means* adalah sebesar 0,558. Sedangkan Nilai Silhouette Coefficient dengan metode *K-Medoids* adalah sebesar 0,529 yang menyatakan bahwa metode *K-Means* menghasilkan nilai SC lebih besar dari pada *K-Medoids*, sehingga *K-Means* dapat memberikan hasil pengelompokan yang lebih baik.

Tabel berikut adalah rangkuman dari penelitian-penelitian yang telah diuraikan diatas:

Judul, Penulis, Tahun	**Jumlah & Atribut**	**Algoritma**	***Preproc essing***	**Validasi**	**Open Source dataset**	**Akurasi**
Clustering Kanker Serviks Berdasarkan Perbandingan Euclidean dan Manhattan Menggunakan Metode *K-Means*. Slamet Widodo, et al. (2021)	401 data, 15 atribut	*K-Means*	-	Confussion Matrix dan kurva ROC.	Data pasien RSPAD Gatot Soebroto Jakarta	*K-Means* Eulidean 79,30%, *K-Means* Manhattan 67,83%
Pengelompokan Penyakit Pada Pasien Berdasarkan Usia Dengan Metode *K-Means* Clustering (Studi Kasus : Puskesmas Bahorok) Tanty, Budi Serasi Ginting, Magdalena Simanjuntak (2021)	20 data, 4 atribut	*K-Means*	-	-	Data pasien dari Puskesmas Bahorok	Ideal 3 klaster
Perbandingan Algoritma *K-Means*, X-Means Dan *K-Medoids* Untuk Klasterisasi Awak Kabin Lion Air. Ahmad Jurnaidi Wahidin, Dana Indra Sensuse (2021)	100 data, 7 atribut.	*K-Means*, X-Means Dan *K-Medoids*	-	Davies Bouldin Indeks	Divisi Flight Operation Support (FOS)	*K-Means* 0.792 X-Means 0.812 *K-Medoids* 1.700

Perbandingan Tingkat Kepuasan Siswa Terhadap Pelayanan Sekolah Menggunakan Algoritma *K-Means* Dan *K-Medoids*. Maulana Abdur Rofik, Amril Mutoi Siregar, Dwi Sulistya Kusumaningrum (2021)	509 data dan 5 atribut	*K-Means* Dan *K-Medoids*	-	-	Data pribadi SMK TI Muhammadiyah Cikampek	*K-Means*: Puas 276, Cukup Puas 216, Kurang Puas 17 *K-Medoids*:Puas 324, Cukup Puas 11, Kurang Puas 174 Algoritma Ideal: *K-Medoids*
Perbandingan Algoritma K Means dan K Medoids Untuk Clustering Kelas Siswa Tunagrahita. Fitriana Harahap (2021)	36 data dengan 6 atrib	*K-Means* dan *K-Medoids*		Davies Bouldin Indeks	data pribadi Sekolah Luar Biasa C Muzdalifah Medan	DBI *K-Means* 0,161, DBI *K-Medoids* 0,281
K-Means Dan Fuzzy C-Means Pada Analisis Data Polusi Udara Di Kota X Sandi Fajar Rodiyansyah (2017)	9358 data, 10 atribut		*Data Cleaning, Data Transformation*	Standar Deviasi	Penelitian De Vito	
Penerapan Algoritma *K-Medoids* untuk Pengelompokan Penyakit di Pekanbaru Riau Tri Juninda, Mustakim, Elvia Andri (2019)	2811 data, 4 atribut	*K-Medoids*	Data Cleaning, Data Transformasi, dan Data Normalisasi		Kuesioner online yang diisi oleh masyarakat Pekanbaru	Klaster 1 : 420 Klaster 1 : 349 Klaster 1 : 794 Klaster 1 : 1248

Perbandingan Algoritma *K-Means* Dan *K-Medoids* Dalam Klasterisasi Produk Asuransi Perusahaan Nasional Frenda Farahdinna, Irfan Nurdiansyah, Apriati Suryani, Arief Wibowo (2019)	102 data, 7 atribut	*K-Means* dan *K-Medoids*	Cleansing data	Davies Boldin Index	Data pribadi perusahaan asuransi	*K-Means*, k=5, 0,018
Perbandingan Algoritma *K-Means* dengan Algoritma Fuzzy C-Means untuk Clustering Tingkat Kedisiplinan Kinerja Karyawan Nova Agustina, Prihandoko (2018)		*K-Means* dan Fuzzy C-Means	-	*K-Means*:Silhoutte Index Fuzzy C-Means:artition Coefficient Index (PCI)	STT-Bandung	*K-Means*:Silhoutte Index Fuzzy C-Means:0,758
Perbandingan Pengelompokan *K-Means* dan *K-Medoids* Pada Data Potensi Kebakaran Hutan/Lahan Berdasarkan Persebaran Titik Panas Athifaturrofifah, Rito Goejantoro, dan Desi Yuniarti (2019)	42 data, 7 atribut	*K-Means* dan *K-Medoids*	-	Silhouette Coefficient	Data Titik Panas Di Indonesia	*K-Means*:0,558 *K-Medoids*:0,529
Perbandingan Algoritma *K-Means* Dan *K-Medoids* Untuk Klasterisasasi Penyakit Kanker Serviks. Lilis Setya Rini (2022)	858 Data dan 36 atribut	*K-Means* Dan *K-Medoids*	Data Cleaning	Davies Bouldin Index	UCI Machine Learning Repository	Davies Bouldin *K-Means*: 0.30 *K-Medoids*: 0.

Tabel 2. 1 Kajian Literatur

Dari review beberapa jurnal diatas terkait dengan metode *clustering*, belum menerapkan klasterisasi dengan algoritma *K-Means* dan *K-Medoids* untuk penyakit kanker serviks menggunakan *tool RapidMiner*, sehingga peneliti akan menerapkan metode *K-Means dan K-Medoids* untuk klasterisasi penyakit kanker serviks dan dibantu dengan *tools RapidMiner Studio.*

2.2 Kanker Serviks

Kanker seviks adalah tumor ganas primer yang berasal dari sel epitel skuamosa. Sebelum terjadinya kanker, akan didahului oleh keadaan yang disebut lesi prakanker atau neoplasia intraepitel serviks (NIS)[5]. Kanker ini adalah kanker yang terjadi pada area leher rahim yaitu bagian rahim yang menghubungkan rahim bagian atas dengan vagina. Kanker serviks disebabkan infeksi virus HPV (Human Papilloma Virus) atau virus papiloma manusia. HPV menimbulkan kutil pada pria maupun wanita, termasuk kutil pada kelamin yang disebut kondiloma akuminatum. Hanya beberapa saja dari ratusan varian HPV yang dapat menyebabkan kanker. Kanker serviks atau kanker leher rahim bisa terjadi jika terjadi infeksi yang tidak sembuh-sembuh untuk waktu lama. Sebaliknya, kebanyakan infeksi HPV akan hilang sendiri, teratasi oleh sistem kekebalan tubuh. Pada tahap awal, penyakit ini tidak menimbulkan gejala yang mudah diamati. Gejala fisik serangan penyakit ini pada umumnya hanya dirasakan oleh penderita kanker stadium lanjut. Apabila kanker sudah menyebar ke panggul, maka pasien akan menderita keluhan nyeri punggung, hambatan dalam berkemih, serta pembesaran ginjal.

Akan tetapi apabila ditangani lebih cepat maka kemungkinan penyembuhan terhadap kanker bisa diatasi tergantung dari stadium dari kanker tersebut[2].

2.3 *Data Mining*

Data mining adalah metode pengolahan data yang difokuskan untuk mencari pola tersembunyi dalam data tersebut[15] dan menurut Turban dkk. (2005) data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan mesin learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait berbagai database besar.

Berdasarkan beberapa pengertian di atas, dapat disimpulkan bahwa data mining adalah suatu proses analisis untuk menggali informasi yang tersembunyi dengan menggunakan statistik dan artificial intelligence di dalam suatu database dengan ukuran sangat besar, sehingga ditemukan suatu pola dari data yang sebelumnya tidak diketahui, dan pola tersebut direpresentasikan dengan grafik komputer agar mudah dimengerti. Proses penganalisaan dari data yang banyak dengan tujuan menemukan suatu jawaban untuk dijadikan informasi yang berguna dalam mengambil keputusan, pada prosesnya data mining memiliki banyak metode yang dapat digunakan. Banyak fungsi data mining yang dapat digunakan, dalam kasus tertentu fungsi data mining dapat digabungkan untuk memecahkan sebuah permasalahan. Fungsi dari data mining, secara umum dijelaskan oleh Han dkk. (2012) adalah deskripsi, estimasi, prediksi, klasifikasi, pengelompokan, asosiasi.

2.4 *Clustering*

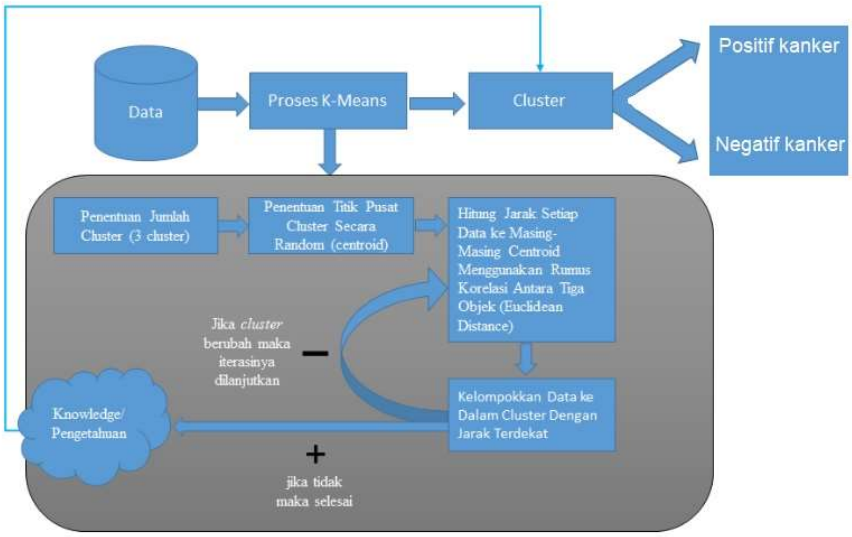
Clustering merupakan cara untuk menemukan kelompok objek yang memiliki kemiripan dan dapat menemukan pola penyebaran dan pola hubungan dalam kumpulan data yang besar[16]. Data-data yang memiliki

kemiripan karakter akan berkumpul disebuah *cluster* yang sama. Pada data mining terdapat dua pembagian jenis metode *clustering* untuk proses pengelompokkan data, yakni *hierarchical clustering* dan *non-hierarchical clustering*. Pada dasarnya *clustering* adalah metode untuk mengkategorikan atau pengelompokan sekelompok objek sesuai dengan atribut yang sama atau karakteristik dengan data-data lainnya. *Clustering* merupakan suatu metode pada data mining dimana proses kerja pada algoritma ini sifatnya tanpa arahan (*unsupervised*), artinya metode ini tidak lagi memerlukan lagi suatu training dan tanpa guru bahkan output tidak di perlukan.

2.5 *K-Means*

K-Means merupakan salah satu metode clustering non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih cluster[17]. Dengan demikian data yang memiliki kemiripan berada pada satu cluster yang sama dan data yang memiliki ketidaksamaan berada pada cluster yang lain.

Algoritma *K-Means*, mungkin yang pertama dari algoritma pengelompokan yang diusulkan, didasarkan pada ide yang sangat sederhana. Diberi serangkaian kumpulan awal, menetapkan setiap titik ke salah satunya, lalu setiap pusat gugus diganti dengan titik rata-rata pada klaster masing-masing. Dua langkah sederhana ini diulang hingga konvergensi. Suatu titik ditugaskan ke cluster yang dekat dalam jarak Euclidean ke titik[18].



Gambar 2. 1 Alur Kerja K-Means Clustering

Pada gambar 2.1 merupakan proses algoritma *K-Means*, dimana tahap pertama adalah mengumpulkan data dan *pre-processing data*. Tahapan selanjutnya adalah proses pengolahan *K-Means* dengan menentukan jumlah *cluster* secara random, titik pusat *cluster*, menghitung jarak setiap data dan mengelompokkan data pada *cluster* sehingga menghasilkan *knowledger* yaitu *clustering*.

2.6 *K-Medoids*

K-Medoids tidak menentukan nilai rata-rata dari objek dalam cluster sebagai titik acuan, tapi menggunakan medoids (median), yang merupakan objek yang paling terletak dipusat sebuah cluster. Dengan demikian, metode partisi masih dapat dilakukan berdasarkan prinsip meminimalkan jumlah dari ketidak samaan antara setiap objek dan titik acuan yang sesuai (medoids).

Metode *K-Medoids* adalah bagian dari partitioning *clustering*. Metode *K-Medoids* cukup efisien dalam dataset yang kecil. Langkah awal *K-Medoids*

ialah mencari titik yang paling representatif (medoids) dalam dataset dengan menghitung jarak dari kelompok dalam semua kemungkinan kombinasi dari medoids sehingga jarak antar titik dalam suatu cluster kecil sedangkan jarak titik antar cluster besar[19].

2.7 *Davies-Bouldin Index (DBI)*

DBI adalah suatu metode yang digunakan untuk mengukur validitas *cluster* dalam metode *clustering*, kohesi didefinisikan sebagai jumlah dari kedekatan data terhadap titik pusat *cluster* dari *cluster* yang diikuti. Sedangkan separasi didasarkan pada jarak antar titik pusat *cluster* terhadap klasternya. Pengukuran dengan DBI dapat memaksimalkan jarak antara *cluster* Ci dan Cj dan pada saat yang sama mencoba untuk meminimalkan jarak antara titik-titik dalam *cluster*. Jika jarak antar *cluster* adalah maksimum, perbedaan signifikan pada setiap *cluster*, maka perbedaan kecil antara *cluster* lebih jelas. Jika jarak *intra-cluster* minimal, itu berarti setiap objek dalam *cluster* memiliki tingkat karakteristik penting yang tinggi[20].

Sum of square within cluster (SSW) merupakan persamaan yang digunakan untuk mengetahui matrik kohesi dalam sebuah cluster ke-i yang dirumuskan sebagai berikut:

$$SSW_i = \frac{1}{m_i} \sum_{j=i}^{m_i} d(x_{j|}, c_i)$$

Dari persamaan tersebut, mi merupakan jumlah data dalam cluster ke-i, ci adalah centroid cluster ke-i, dan d () merupakan jarak setiap data kecentroid yang dihitung menggunakan jarak euclidean.

Sum of square between cluster (SSB) merupakan persamaan yang digunakan untuk mengetahui separasi antar cluster yang dihitung menggunakan persamaan:

$$SSB_{i,j} = d(c_i, c_j)$$

Setelah nilai kohesi dan separasi diperoleh, kemudian dilakukan pengukuran rasio (Rij) untuk mengetahui nilai perbandingan antara cluster ke-i dan cluster ke-j. Cluster yang baik adalah cluster yang memiliki nilai kohesi sekecil mungkin dan separasi yang sebesar mungkin. Nilai rasio dihitung menggunakan persamaan sebagai berikut:

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}}$$

Nilai rasio yang diperoleh tersebut digunakan untuk mencari nilai *Davies-Bouldin Index (DBI)* dari persamaan berikut:

$$\text{DBI} = \frac{1}{k}\sum_{i=1}^{k} Ri, qt$$

Dari persamaan tersebut, *k* adalah banyaknya *cluster* yang digunakan. Jika nilai DBI yang diperoleh semakin kecil (non-negatif >=0), maka *clustering* yang diperoleh semakin baik[21].

2.8 *RapidMiner*

RapidMiner adalah perangkat lunak yang dibuat oleh Dr. Markus Hofmann dari *Institute of Technology Blanchardstown* dan Raif Klinkenberg dari *rapid-i.com* dengan tampilan *GUI (Graphical User Interface)* sehingga memudahkan pengguna dalam menggunakan perangkat lunak ini. Perangkat lunak ini bersifat *open source* dan dibuat dengan menggunakan bahasa java dibawah lisensi *GNU Public License* dan *Rapid Miner* dapat dijalankan disistem operasi manapun. Dengan menggunakan *Rapid Miner*, tidak dibutuhkan kemampuan koding khusus. *Rapid Miner* dikhususkan untuk penggunaan data mining.