BAB 2
PENELITIAN TERKAIT

Bab ini membahas mengenai penelitian terkait yang diambil dari penelitian-penelitian sebelumnya yang berkaitan. Penelitian ini selanjutnya akan dibangun sebagai landasan dalam melakukan penelitian ini.

2.1 Penelitian Terkait

Beberapa penelitian terkait dengan menggunakan Clustering Kelayakan Siswa Penerima Bantuan Operasional Sekolah Daerah (Bosda) Algoritma K-Means Dan K-Medoids diantaranya:

Tabel 2. 1 Penelitian Terkait

No	**Judul, penulis, tahun**	**Dataset**	**Tujuan**	**Metode**	**Hasil**
1	Implementasi Algoritma K-Means Classifier Sebagai Pendukung Keputusan Penerima Dana Bantuan Siswa Miskin (Studi Kasus : SMKN Sukoharjo) Aviv Fitria Yulia1 Handoyo Widi Nugroho2 Institut Informatika dan Bisnis Darmajaya, 27 Agustus 2022 ISSN: 2598-0256, E-ISSN: 2598-0238	Dalam peneliti ini menggunakan dataset dari sekolah. dengan dataset 1044 dan 4 atribut	Sebagai pendukung keputusan	Menggunakan Algoritma K-Means	Hasil pengujian mendapatkan nilai devies bouldin indeks sebesar 0,262 yang memiliki arti kesamaan antar anggota cluster

Penelitian Terkait Lanjutan

No	Judul, penulis, tahun	Dataset	Tujuan	Metode	Hasil
2	CLUSTERING DAERAH MISKIN DI PROVINSI RIAU MENGGUNAKAN METODE K-MEANS Fatimah Isyarah1, Mhd, Arief Hasan2, Fana Wiza3 Vol 1.No.1 2020	Peneliti menggunakan data set 12 dataset, 6 atribut.	mengelompokkan daerah miskin menjadi tiga klaster, yaitu Mampu, Menengah, dan Kurang mampu	Metode K-Means	Pada penelitian tersebut menunjukkan bahwa hasil pengujian bernilai sama. Yang artinya hasil pengujian bernilai sangat baik. 12 Kabupaten di Provinsi Riau, dapat dibentuk tiga buah Cluster. Cluster pertama termasuk dalam wilayah menengah, yang terdiri dari lima Kabupaten/kota, yakni : Rokan Hilir, Bengkalis, Rokan Hulu, Kampar, dan Indragiri Hilir. Sedangkan untuk Cluster kedua termasuk dalam wilayah kurang mampu, yang terdiri dari enam Kabupaten/kota, yakni : Siak, Pelalawan, Indragiri Hulu, Kuantan Singingi, Kep.Meranti, dan Kota Dumai. Lalu untuk Cluster Tiga termasuk dalam wilayah mampu, dan terdiri dari satu anggota saja, yakni : Kota Pekanbaru
3	Data Mining Menentukan Cluster Penerima Program Bantuan dengan Metode K-Means 1*Ririn Restu Aria, 2 Susi Susilowati, 3 Indra Riyana Rahadjeng Volume 7, Nomor 1, Januari 2023	Datset 34 dan 2 atribut	bisa memberikan gambaran kepada pemerintah daerah tentang provinsi yang menerima bantuan berdasarkan jenisnya yang ada sehingga bantuan sosial yang diberikan bisa tepat sasaran kepada penerimanya.	K-Means	mengelompokan data menjadi 2 cluster yaitu cluster 0 termasuk kategori tinggi memiliki 19 provinsi dan cluster 1 termasuk kategori rendah memiliki 15 provinsi.

Penelitian Terkait Lanjutan

No	Judul, penulis, tahun	Dataset	Tujuan	Metode	Hasil
4	Penerapan Algoritma K-Means untuk Klasterisasi Penduduk Miskin pada Kota Pagar Alam Febriansyah Febriansyah (1)* , Siti Muntari (2) Vol. 8, No. 1, JANUARI, 2023	Dataset 471	Tujuan dari penelitian ini adalah untuk mendapatkan klaster data kemiskinan di Kota Pagar Alam	K_Means	Dari hasil penelitian terdiri dari tiga cluster. Cluster dimulai dari cluster_0, cluster_1, dan cluster_2. Data yang berada pada cluster_0 berjumlah 156 data, cluster_1 berjumlah 82 data, dan cluster_2 berjumlah 233 data.
5	Perbandingan K-Means dan K-Medoids Pada Pengelompokan Data Miskin di Indonesia Nanda Try Luchia1 , Hani Handayani2 , Fathan Surya Hamdi3 , Dwi Erlangga4 , Sania Fitri Octavia5 Vol. 2 Iss. 2 October 2022	Dataset 34	Membantu pemerintah untuk bantuan tepat sasaran	K-Means dan K-Medoids	KMeans dan K-Medoids, didapatkan cluster terbaik dengan nilai k=8 pada algortitma K-Means. Sedangkan cluster terbaik pada K-Medoids dengan nilai k=2. Hasil klasterisasi yang dilakukan sudah divalidasi dengan Davies Bouldin Index (DBI). Maka pada penelitian ini, algoritma K-Means lebih unggul dibanding K-Medoids pada pengelompokan data miskin berdasarkan provinsi dengan nilai DBI terbaik yaitu 0.041 dengan nilai k=8. Percobaan K-Means dengan nilai k=8 membagi data menjadi 8 cluster dengan 4 anggota dari cluster 1, 6 anggota dari cluster 2, 1 anggota dari cluster 3 dan 4, 7 anggota dari cluster 5, dan 4 anggota dari cluster 6.

Penelitian Terkait Lanjutan

No	Judul, penulis, tahun	Dataset	Tujuan	Metode	Hasil
6	Perbandingan Tingkat Kepuasan Siswa Terhadap Pelayanan Sekolah Menggunakan Algoritma K-Means Dan K-Medoids Maulana Abdur Rofik, Amril Mutoi Siregar, Dwi Sulistya Kusumaningrum Vol. II No: 1, Januari 2021	Dataset 509	Membantu sekolah dalam tingkat kepuasan pelayanan di sekolah	K-Means Dan K-Medoids	hasil yang di peroleh k-means setelah 3 iterasi siswa merasa puas dengan pelayanan sekolah berjumlah 276 siswa, cukup puas berjumlah 216 siswa dan kurang puas dengan pelayanan sekolah berjumlah 17 siswa. Sedangkan k-medoids lebih baik karena hanya 2 iterasi dengan hasil siswa merasa puas dengan pelayanan sekolah berjumlah 324 siswa, cukup puas berjumlah 11 siswa dan kurang puas dengan pelayanan sekolah berjumlah 174 siswa, sehingga algoritma k-medoids lebih cocok digunakan untuk klastering kepuasan siswa disekolah
7	PENGELOMPOKAN MENGGUNAKAN ALGORITMA KMEDOID UNTUK EVALUASI PERFORMA SISWA Yoga Religia1 , Rifki Tia Bayu Jaya2 Vol. 15 2020	Dataset 909	Membantu sekolah untuk evaluasi performa siswa dengan metode kmedoids	Algoritma Kmedoid	Dari penelitian ini diperoleh cluster status performa siswa sebagai berikut: • Cluster Status performa siswa kurang sebanyak 426 siswa • Cluster Status performa siswa ratarata sebanyak 191 siswa • Cluster Status performa siswa bagus sebanyak 292 siswa

Penelitian Terkait Lanjutan

No	Judul, penulis, tahun	Dataset	Tujuan	Metode	Hasil
8	The comparison of k-means and k-medoids algorithms for clustering the spread of the covid-19 outbreak in Indonesia Wargijono Utomo ILKOM Jurnal Ilmiah Vol. 13, No. 1, April 2021	Terdapat 34 provinsi dan 4 atribut	untuk mengelompokkan penyebaran wabah Covid-19 di Indonesia	algoritma k-means dan k-medoids	perbandingan antara metode K-Means dan K-Medoids untuk mengelompokkan penyebaran virus corona di Indonesia, kesimpulan diperoleh. Dengan nilai indeks Davies Boulden dari nilai K2 hingga K9, ternyata metode K-Means mendapatkan hasil yang lebih baik. nilai terkecil pada K-5 sebesar 0,064, sedangkan K-Medoids pada nilai k-2 sebesar 0,411. Jadi dari dua cara yang digunakan bisa saja menyimpulkan bahwa metode terbaik untuk mengelompokkan penyebaran wabah virus corona di Indonesia adalah metode K-Means.

Berdasarkan tabel penelitian diatas, maka dapat disimpulkan bahwa penerapan algoritma K-means membuat cluster lebih akurat namun ada juga yang dari Sebagian peneliti diatas algoritma K-medoids sama akuratnya. Dari peneliti diatas pada judul "Perbandingan K-Means dan K-Medoids Pada Pengelompokan Data Miskin di Indonesia" KMeans dan K-Medoids, didapatkan cluster terbaik dengan nilai k=8 pada algortitma K-Means. Sedangkan cluster terbaik pada K-Medoids dengan nilai k=2. Hasil klasterisasi yang dilakukan sudah divalidasi dengan Davies Bouldin Index (DBI). Maka pada penelitian ini, algoritma K-Means lebih unggul dibanding K-Medoids pada pengelompokan data miskin berdasarkan provinsi dengan nilai DBI terbaik yaitu 0.041 dengan nilai k=8. Percobaan K-Means dengan nilai k=8 membagi data menjadi 8 cluster dengan 4 anggota dari cluster 1, 6 anggota dari cluster 2, 1 anggota dari cluster 3 dan 4, 7 anggota dari cluster 5, dan 4 anggota dari cluster 6. Algortima K-means lebih akurat dibandingkan dengan K-medoids.

Sedangkan pada peneliti dengan judul " Perbandingan Tingkat Kepuasan Siswa Terhadap Pelayanan Sekolah Menggunakan Algoritma K-Means Dan K-Medoids" hasil yang di peroleh k-means setelah 3 iterasi siswa merasa puas dengan pelayanan sekolah

berjumlah 276 siswa, cukup puas berjumlah 216 siswa dan kurang puas dengan pelayanan sekolah berjumlah 17 siswa. Sedangkan k-medoids lebih baik karena hanya 2 iterasi dengan hasil siswa merasa puas dengan pelayanan sekolah berjumlah 324 siswa, cukup puas berjumlah 11 siswa dan kurang puas dengan pelayanan sekolah berjumlah 174 siswa, sehingga algoritma k-medoids lebih cocok digunakan untuk klastering kepuasan siswa disekolah. Algoritma K-medoids lebih akurat dibandingkan K-means.

2.2 SMKN 1 Katibung

SMK Negeri 1 Katibung merupakan sebuah Sekolah Menengah Kejuruan (SMK) yang terletak di Kalianda, Lampung Selatan, Provinsi Lampung. Sebagai institusi pendidikan menengah kejuruan negeri, SMK Negeri 1 Katibung menyediakan berbagai program keahlian dan pelatihan keterampilan kepada siswa-siswi yang ingin memperoleh pendidikan dalam bidang-bidang tertentu yang relevan dengan kebutuhan industri dan pasar kerja. Sebagai SMK negeri, lembaga ini dioperasikan dan dikelola oleh pemerintah daerah setempat dengan tujuan menyediakan pendidikan berkualitas bagi masyarakat di sekitarnya. .

2.3 BOSDA

Bantuan Operasional Sekolah Daerah (BOSDA) merupakan suatu program pemerintah yang bertujuan untuk memberikan dukungan finansial kepada sekolah-sekolah di daerah dengan populasi peserta didik yang mayoritas berasal dari keluarga kurang mampu. Dalam konteks ini, "keluarga tidak mampu" merujuk kepada keluarga yang berada dalam kondisi ekonomi yang terbatas sehingga sulit untuk memenuhi kebutuhan pendidikan anak-anak mereka.Program BOSDA berfokus pada penyediaan dana bantuan yang dihitung berdasarkan dua faktor utama. Pertama, jumlah siswa yang memenuhi kriteria tidak mampu di setiap sekolah, yang mencakup siswa-siswa dari keluarga kurang mampu. Kedua, program ini menggunakan satuan biaya yang telah ditetapkan, yang dikenal sebagai Unit Cost, sebagai dasar perhitungan untuk menentukan besaran dana bantuan yang diberikan kepada setiap sekolah. Dengan menggunakan pendekatan ini, dana bantuan dari program BOSDA dapat disesuaikan dengan kebutuhan dan skala sekolah,

sehingga sekolah-sekolah dengan jumlah peserta didik yang lebih banyak dan tingkat ketersediaan sumber daya yang lebih rendah dapat menerima bantuan yang sesuai untuk mendukung kegiatan operasional dan penyelenggaraan pendidikan mereka. Program ini diharapkan dapat membantu mengurangi disparitas pendidikan antara daerah yang lebih mampu dan kurang mampu dengan memastikan bahwa sekolah-sekolah di daerah yang membutuhkan mendapatkan dukungan finansial yang memadai[10].

Secara khusus, program Bantuan Operasional Sekolah Daerah (BOSDA) untuk Sekolah Menengah Atas (SMA) dan Sekolah Menengah Kejuruan (SMK) memiliki beberapa tujuan yang dijabarkan sebagai berikut:

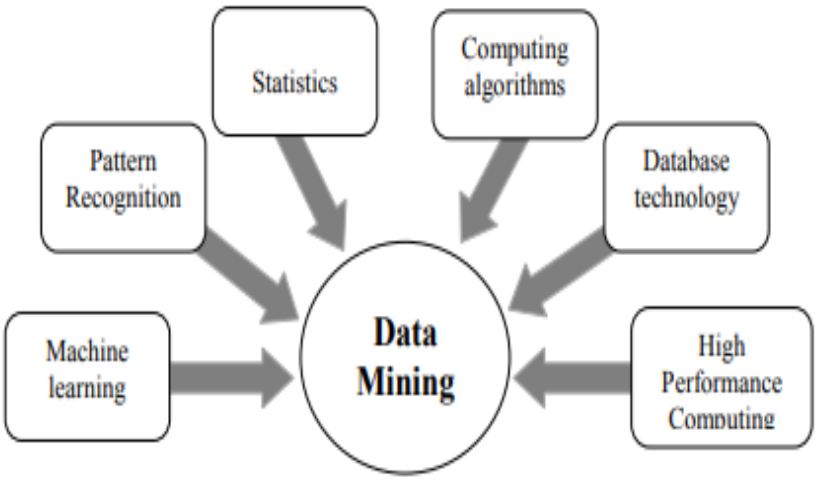
a. Membantu biaya operasional sekolah: Salah satu tujuan utama dari program BOSDA untuk SMA/SMK adalah untuk memberikan dukungan finansial yang dapat digunakan oleh sekolah dalam menutupi biaya operasional sehari-hari. Biaya operasional ini mencakup berbagai aspek, seperti pembayaran gaji guru dan pegawai, pemeliharaan fasilitas sekolah, pembelian perlengkapan pendidikan, dan berbagai keperluan lain yang diperlukan untuk menjalankan kegiatan belajar mengajar secara efektif.
b. Meningkatkan akses dan kualitas pendidikan negeri/swasta: Program BOSDA SMA/SMK bertujuan untuk meningkatkan aksesibilitas pendidikan bagi siswa di tingkat menengah atas, baik untuk sekolah negeri maupun swasta. Dengan memberikan bantuan operasional kepada sekolah-sekolah ini, diharapkan dapat memperluas kesempatan bagi siswa untuk mengakses pendidikan berkualitas tanpa terkendala oleh faktor finansial. Selain itu, dukungan ini juga diharapkan dapat meningkatkan kualitas pendidikan yang diberikan oleh kedua jenis sekolah tersebut, sehingga memberikan dampak positif bagi proses pembelajaran dan pencapaian siswa.
c. Memberikan kesempatan yang setara bagi peserta didik untuk mendapatkan layanan pendidikan yang terjangkau dan bermutu: Program BOSDA SMA/SMK bertujuan untuk menciptakan kesempatan yang setara bagi semua peserta didik, tanpa memandang latar belakang ekonomi mereka. Dengan memberikan bantuan operasional kepada sekolah-sekolah di daerah yang membutuhkan, program ini berupaya untuk memastikan bahwa semua siswa, terutama yang berasal dari

keluarga kurang mampu, memiliki akses yang sama terhadap pendidikan yang terjangkau dan berkualitas.

2.4 Data Minning

Istilah *data mining* memiliki hakikat sebagai disiplin ilmu yang tujuan utamanya adalah untuk menemukan, menggali, atau menambang pengetahuan dari data atau informasi yangkita miliki. Data mining, sering juga disebut sebagai Konowledge Discovery Database (KDD)[11] Data Mining adalah proses menemukan pola wawasan, menarik, dan baru, sebagai serta model deskriptif, dapat dipahami, dan prediktif dari data berskala besar [12].

Data mining, secara sederhana merupakan suatu langkah ekstraksi untuk mendapatkan informasi penting yang sifatnyaimplisit dan belum diketahui. Data miningmempunyai hubungan dengan berbagai bidang seperti statistic, machine learning, *computing algorithms, database technology.* Gambar 2.1 merupakan diagram hubungan *data mining* :



Gambar 2. 1 Diagram Hubungan Data Minning

Secara sistematis, langkah utama untuk melakukan *data mining* terdiri dari tahap, yaitu sebagai berikut :

1) Ekspolasi Atau Pemrosesan Awal Data

Eksplorasi atau pemrosesan awal data terdiri dari pembersihan data, normalisasi data, transformasi data, penanganan missing value, reduksi dimensi, pemilihan subset fitur, dan sebagainya.

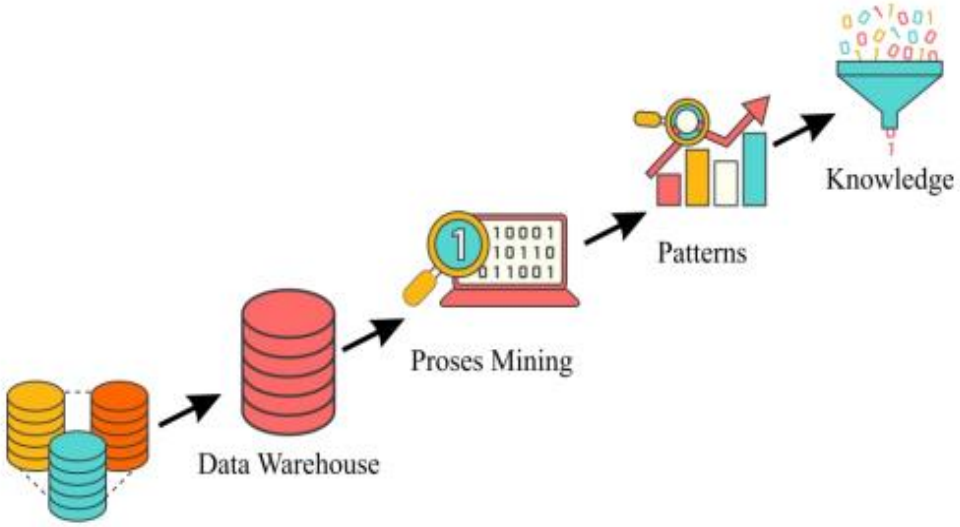
2) Membangun Model Dan Validasi

Membangun model dan validasi, merupakan melakukan analisis dari berbagai model dan memlih model sehingga menghasilkan kinerja yang terbaik. Pembangunan model dilakukan menggunakan metode-metode seperti klasifikasi, regresi,analisis cluster, dan asosiasi.

3) Penerapan

Penerapan dilakukan dengan menerapkan model yang dipilih pada data baru untuk menghasilkan kinerja yang baik pada msalah yang diinvestigasi.

Tahapan proses data mining ada beberapa yang sesuai dengan proses KDD (*Knowledge Discovery in Database*) . Gambar 2.2 merupakan proses KDD (*Knowledge Discovery in Database*) :



Gambar 2. 2 Proses KDD (*Knowledge Discovery in Database*)

1. *Cleaning And Integration.*
   a. *Data Cleaning* (Pembersih data)

      Data cleaning (Pembersihan data) adalah proses yang dilakukan untuk menghilangkan noise pada data yang tidak konsisten atau bisa disebut tidak

relevan. Data yang diperoleh dari database suatu perusahaan maupun hasil eksperimen yang sudah ada, tidak semuanya memiliki isian yang sempurna misalnya data yang hilang, data yang tidak valid, atau bisa juga hanya sekedar salah ketik. Data yang tidak relevan itu dapat ditangani dengan cara dibuang atau sering disebut dengan proses cleaning. Proses cleaning dapat berpengaruh terhadap performa dari teknik *data mining*.

b. *Data Integration* (Integrasi Data)

Integrasi data merupakan proses penggabungan data dari berbagai database sehingga menjadi satu database baru. Data yang perlukan pada proses *data mining* tidak hanya berasal dari beberapa database.

2. *Selection and Transformation*

a. *Data Selection* (Seleksi Data)

Tidak semua data yang ada didatabase akan dipakai, karena hanya data yang sesuai saja yang akan dianalisis dan diambil dari database.Misalnya pada sebuah kasus market basket analysis yang akan meneliti faktor kecenderungan pelanggan, maka tidak perlu mengambil nama pelanggan, cukup dengan id pelanggan.

b. *Data Trnasformation* (Transformasi Data)

Transformasi data merupakan proses pengubahan data dan penggabungan data ke dalam format tertentu, *data mining* membutuhkan format data khusus sebelum diaplikasikan. Misalnya metode standar seperti analysis asosiasi dan clustering haya bias menerima inputan data yang bersifat katagorial. Karenanya data yang berupa angka numeric apabila mempunyai sifat kontinyu perlu dibagi menjdi beberapa interval. Proses ini sering disebut dengan transformasi data.

3. *Poses Mining*

Proses mining dapat disebut juga sebagai proses penambangan data. Proses mining merupakan proses utama yang menggunakan metode untuk menemukan pengetahuaan beharga yang tersembunyi dari data.

4. *Evaluation and Precentation*
   a. Evaluasi Pola *(Pattren Evaluation)*

      Evaluasi pola bertugas untuk mengidentifikasi pola-pola yang menarik ke dalam knowledge based yang ditemukan. Pada tahap ini dihasilkan polapola yang khas dari model klasifikasi yang dievaluasi untuk menilai apakah hipotesa yang ada memang tercapai. Bila ternyata hasil yang diperoleh tidak sesuai dengan hipotesa, terdapat beberapa alternative yang bias diambil seperti menjadikanya umpan baik untuk memprbaiki proses *data mining*, atau mencoba metode *data mining* lain yang lebih sesuai.

   b. Presentasi Pengetahuan *(Knowledge Presentation)*

      Knowledge presentation merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan atau informasi yang telah digali oleh pengguna. Tahap terakhir dari proses data mining adalah memformulasikan keputusan dari hasil analisis yang didapat.

2.4.1 Clustering

Clustering, dalam konteks analisis data, merujuk pada proses pengelompokan sekumpulan data menjadi beberapa kelompok atau kluster berdasarkan tingkat kemiripan di antara data-data tersebut. Konsep dasar dari clustering adalah untuk mengelompokkan objek-objek atau entitas-entitas data yang memiliki karakteristik atau fitur yang serupa atau mirip dalam suatu kluster yang sama. Dalam proses clustering, setiap objek atau entitas data direpresentasikan sebagai titik dalam ruang fitur. Titik-titik ini kemudian dikelompokkan bersama berdasarkan kesamaan atau kemiripan fitur-fitur yang dimiliki. Tujuan utama dari clustering adalah untuk memaksimalkan kesamaan di dalam kluster dan meminimalkan kesamaan antar kluster. Proses clustering dapat dilakukan dengan berbagai metode, termasuk K-Means, K-Medoids, Hierarchical Clustering, dan Density-Based Clustering, di antara lain. Setiap metode clustering memiliki pendekatan dan karakteristik yang berbeda, tetapi tujuannya adalah sama: untuk mengelompokkan data menjadi kelompok-kelompok yang serupa atau homogen[13].

Metode clustering merupakan suatu metode untuk mencari dan mengelompokkan data yang memiliki kemiripan karakteriktik (similarity) antara satu data dengan data yang

lain.[14]. *Clustering* membagi data menjadi kelompok-kelompok atau *cluster* berdasarkan karakteristik yang serupa. Pengelompokan sejumlah objek atau data ke dalam kelompok (kelompok) adalah proses yang dikenal sebagai "*clustering*". Tujuan dari *clustering* adalah untuk memastikan bahwa setiap kelompok berisi data yang seminimal mungkin dan objek dalam setiap kelompok akan berbeda satu sama lain. *Clustering* adalah salah satu metode pengelompokan dalam *data mining*. Dalam ilmu *data mining*, pengelompokan berarti mengelompokkan sejumlah objek atau data ke dalam kelompok atau kelompok, sehingga setiap kelompok berisi data yang seminimal mungkin mirip dan berbeda dari kelompok lainnya [15].

Partisi dan metode hirarki adalah metode clustering yang paling banyak dipelajari. Metode hirarki memanfaatkan pendekatan dengan membuat struktur berbasis pohon biner yang disebut dendogram, sedangkan metode partisi bertujuan untuk menemukan pengelompokan data dengan mengoptimalkan fungsi tujuan yang dapat meningkatkan kualitas partisi. *K-Means, SOM (Self Organizing Maps), Fuzzy C-Means, dan PAM (Partitioning Around Medoid)* adalah beberapa metode partisi [16]. Karena tidak dapat lepas dengan banyak data yang menghasilkan informasi untuk kebutuhan hidup, clustering dapat sangat penting dalam kehidupan sehari-hari. Mengklasifikasikan atau mengelompokkan data ke dalam seperangkat kategori atau *cluster* adalah salah satu metode yang paling penting dalam hubungan dengan data. Banyak aplikasi saat ini menggunakan clustering di berbagai bidang. Contoh pengelompokan data yang digunakan untuk analisis data statistik adalah pengelompokan untuk pembelajaran mesin, *data mining*, pengenalan pola, analisis gambar, dan bio informatika [17].

2.4.2 K-Means

K-Means merupakan metode klasterisasi yang paling terkenal dan banyak digunakan diberbagai bidang karena sederhana, mudah diimplementasikan, mempunyai kemampuan untuk mengklaster data yang sangat besar dan kompleksitas waktunya linear O(nKT) dengan n adalah jumlah dokumen, K adalah jumlah klaster, dan T adalah jumlah iterasi.[18]. *K-means* clustering adalah metode pengelompokan data *non-hirarki* yang mengelompokkan data dalam satu atau lebih *cluster* atau kelompok. Clustering *K-means* menggabungkan data yang memiliki karakteristik yang sama ke dalam satu *cluster* atau kelompok, dan kemudian mengelompokkan data dengan karakteristik yang berbeda

ke dalam *cluster* atau kelompok lain sehingga ada tingkat variasi yang kecil antara data yang tergabung dalam satu *cluster* atau kelompok. Tujuan pengelompokan data ini adalah untuk mengurangi fungsi objektif proses pengelompokan, yang biasanya bertujuan untuk mengurangi variasi dalam kelompok dan memaksimalkan variasi antar kelompok [19].

K-Means merupakan algoritma untuk cluster objek berdasarkan atribut menjadi k partisi, dimana k < n. Secara Umum KMeans Clustering merupakan salah satu metode data Clustering non-hirarki yang mengelompokan data dalam bentuk satu atau lebih cluster atau kelompok [20]. Metode ini mempartisi data ke dalam cluster sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu cluster yang sama dan data yang mempunyai karateristik yang berbeda di kelompokan ke dalam cluster yang lain. Istilah-istilah dalam K-Means :

1. N data : data set yang akan diolah sebanyak N data dimana N data tersebut terdiri dari atributatributnya
2. K centroid : Inisialisasi dari pusat cluster data adalah sebanyak K dimana pusat-pusat awal tersebut digunakan sebagai banyaknya kelas yang akan tercipta. Centroid didapatkan secara random dari N data set yang ada.
3. Euclidian Distance: merupakan jarak yang didapat dari perhitungan antara semua N data dengan K centroid dimana akan memperoleh tingkat kedekatan dengan kelas yang terdekat dengan populasi data tersebut.

Kelemahan K-Means :

1. Bila jumlah data tidak terlalu banyak, mudah untuk menentukan cluster awal.
2. Jumlah cluster, sebanyak K, harus ditentukan sebelum dilakukan perhitungan.
3. Tidak pernah mengetahui real cluster dengan menggunakan data yang sama, namun jika dimasukkan dengan cara yang berbeda mungkin dapat memproduksi cluster yang berbeda jika jumlah datanya sedikit.
4. Tidak tahu kontribusi dari atribut dalam proses pengelompokan karena dianggap bahwa setiap atribut memiliki bobot yang sama.

Langkah-langkah dalam Algoritma K-Means Clustering :

1. Menentukan jumlah cluster .
2. Menentukan nilai centroid. Dalam menentukan nilai centroid untuk awal iterasi, nilai awal centroid dilakukan secara acak. Sedangkan jika menentukan nilai

centroid yang merupakan tahap dari iterasi, maka digunakan rumus sebagai berikut :

a. Menghitung jarak antara titik centroid dengan titik tiap objek
b. Pengelompokan objek untuk menentukan anggota cluster adalah dengan memperhitungkan jarak minimum objek.
c. Kembali ke tahap 2, lakukan perulangan hingga nilai centroid yang dihasilkan tetap dan anggota cluster tidak berpindah ke cluster lain.

Metode k-means ini mengelompokkan data yang ada kedalam beberapa kelompok yang masing-masing memiliki karakteristik.[21] Pada algoritma K-Means perlu menghitung jarak data ke centroid terdekat dengan rumus [22]:

$$(x, y) = \sqrt{(xi - yi)2 + (xi - yi)^2} \quad (1)$$

Dimana : D(x,y) = Jarak data ke centroid
x= Record / Data
y= Centroid / Pusat cluster

2.4.3 K-medoids

Metode K-Medoids adalah salah satu metode dalam analisis cluster yang serupa dengan metode K-Means, tetapi menggunakan medoids sebagai representasi titik pusat dari setiap kluster. Medoids adalah titik data aktual dalam dataset yang secara optimal mewakili kluster. Metode ini mirip dengan K-Means namun lebih tahan terhadap outliers karena medoids diambil dari data aktual. Selain itu, metode K-Medoids juga berkaitan dengan metode MedoidShift yang fokus pada pergeseran titik medoid untuk memperbaiki kualitas clustering [23]. Algoritma PAM (Partioning Around Medoids) atau biasa juga disebut dengan algoritma K- Medoids, merupakan algoritma yang diwakili oleh cluster yaitu medoid. Perbedaan antara algoritma K- Medoids dengan algoritma K-Means yaitu algoritma K-Medoids menggunakan objek sebagai perwakilan (medoid) pusat cluster

untuk tiap cluster, sementara algoritma K-Means membutuhkan nilai rata-rata (mean) sebagai pusat cluster [24]. Langkahlangkah algoritma K-Medoids:

1. Inisialisasi pusat cluster sebanyak k (jumlah cluster)
2. Alokasikan setiap data (objek) ke cluster terdekat menggunakan persamaan ukuran jarak Euclidian Distance dengan persamaan:

   $d(x, y) = ||x - y||$

   $= \sum_{i=1}^{n} (\text{xi} - \text{yi})^2$; 1,2,3, … n (2)
3. Pilih secara acak objek pada masing-masing cluster sebagai kandidat medoid baru.
4. Hitung jarak setiap objek yang berada pada masing-masing cluster dengan kandidat medoid baru.
5. Hitung total simpangan (S) dengan menghitung nilai total distance baru – total distance lama. Jika S < 0, maka tukar objek dengan data cluster untuk membentuk sekumpulan k objek baru sebagai medoid.
6. Ulangi langkah 3 sampai 5 hingga tidak terjadi perubahan medoid, sehingga didapatkan cluster beserta anggota cluster masing-masing.

2.5 Rapid Minner

RapidMiner adalah platform perangkat lunak data ilmu pengetahuan yang dikembangkan oleh perusahaan dengan nama yang sama, yang menyediakan lingkungan terpadu untuk pembelajaran mesin (machine learning), pembelajaran mendalam (deep learning), penambangan teks (textmining), dan analisis prediktif (predictive analytics). Aplikasi ini digunakan untuk aplikasi bisnis dan komersial serta untuk penelitian, pendidikan, pelatihan, pembuatan prototype dengan cepat, dan pengembangan aplikasi serta mendukung semua Langkah proses pembelajaran mesin termasuk persiapan data, visualisasi hasil, validasi dan pengoptimalan. RapidMiner dikembangkan dengan model open core [25].

2.6 Davies-Bouldin Index (DBI)

Davies-Bouldin Index (DBI) adalah metrik untuk mengevaluasi algoritma pengelompokan. Ini adalah skema evaluasi internal, tempat validasi seberapa baik pengelompokan yang telah dilakukan dibuat menggunakan kuantitas dan fitur yang

melekat pada dataset. Sedangkan pemisahannya berdasarkan pada jarak antara titik pusat cluster ke clusternya. Sebagai ukuran, DBI dapat memaksimalkan jarak antar cluster Ci dan Cj sambil mencoba meminimalkan jarak antar titik dalam cluster. Ketika jarak antara clusternya maksimal, hal itulah yang membedakan setiap cluster secara signifikan. Jadi, ada sedikit perbedaan di antara keduanya cluster menjadi lebih jelas. Jika jarak antar cluster minimal, berarti setiap objek dalam cluster tersebut memiliki akarakteristik tingkat tinggi [26].