Data Mining Berbasis K-Means

dalam Analisis Penerima Bantuan Sosial

DATA MINING
BERBASIS K-MEANS
DALAM ANALISIS PENERIMA BANTUAN SOSIAL

Sanksi Pelanggaran Pasal 113
Undang-Undang No. 28 Tahun 2014 Tentang Hak Cipta

(i) Setiap Orang yang dengan tanpa hak melakukan pelanggaran hak ekonomi sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf i untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 1 (satu) tahun dan/atau pidana denda paling banyak Rp 100.000.000 (seratus juta rupiah).

(ii) Setiap Orang yang dengan tanpa hak dan/atau tanpa izin Pencipta atau pemegang Hak Cipta melakukan pelanggaran hak ekonomi Pencipta sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf c, huruf d, huruf f, dan/atau huruf h untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 3 (tiga) tahun dan/atau pidana denda paling banyak Rp 500.000.000,00 (lima ratus juta rupiah).

(iii) Setiap Orang yang dengan tanpa hak dan/atau tanpa izin Pencipta atau pemegang Hak Cipta melakukan pelanggaran hak ekonomi Pencipta sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf a, huruf b, huruf e, dan/atau huruf g untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 4 (empat) tahun dan/atau pidana denda paling banyak Rp 1.000.000.000,00 (satu miliar rupiah).

(iv) Setiap Orang yang memenuhi unsur sebagaimana dimaksud pada ayat (3) yang dilakukan dalam bentuk pembajakan, dipidana dengan pidana penjara paling lama 10 (sepuluh) tahun dan/atau pidana denda paling banyak Rp 4.000.000.000,00 (empat miliar rupiah)

DATA MINING
BERBASIS K-MEANS
DALAM ANALISIS PENERIMA BANTUAN SOSIAL

Dela Anggraini, Masrizal, Sudi Suryadi, Syaiful Zuhri Harahap

PUSTAKA iyadz

DATA MINING BERBASIS K-MEANS DALAM ANALISIS PENERIMA BANTUAN SOSIAL

Pemerhati Aksara — Tim Penerbit Pustaka Riyadz
Perancang Sampul — Tim Penerbit Pustaka Riyadz
Penata Letak— Tim Penerbit Pustaka Riyadz

Cetakan Pertama, Juli 2026
iv + 203 hal 15,5 × 23 cm
ISBN — 978-634-7078-93-3

Diterbitkan oleh

Penerbit Pustaka Riyadz
Desa Sawahan Jaya, Semurup, Kec. Air Hangat,
Kerinci - Jambi
pustakariyadz@gmail.com
Instagram @pustakariyadz
HP 0852-6988-7472
Anggota IKAPI: No. 08/JBI/2023

KATA PENGANTAR

Puji syukur ke hadirat Tuhan Yang Maha Esa atas segala rahmat dan karunia-Nya sehingga buku Data Mining Berbasis K-Means dalam Analisis Penerima Bantuan Sosial dapat diselesaikan dengan baik. Buku ini disusun sebagai salah satu bentuk kontribusi dalam pengembangan ilmu pengetahuan di bidang data mining serta penerapannya dalam mendukung pengambilan keputusan pada program bantuan sosial yang lebih tepat sasaran, transparan, dan berbasis data.

Peningkatan jumlah data yang dikelola oleh instansi pemerintah menuntut adanya metode analisis yang mampu menghasilkan informasi secara cepat, akurat, dan objektif. Dalam penyelenggaraan program bantuan sosial, proses identifikasi calon penerima yang sesuai dengan kriteria menjadi tantangan yang memerlukan pendekatan analitis. Salah satu metode yang dapat digunakan adalah algoritma K-Means Clustering, yang mampu mengelompokkan data berdasarkan karakteristik tertentu sehingga membantu proses analisis dan pengambilan keputusan secara lebih sistematis.

Buku ini membahas berbagai konsep dasar data mining serta implementasi algoritma K-Means dalam analisis penerima bantuan sosial. Pembahasan meliputi pengenalan data mining, tahapan Knowledge Discovery in Databases (KDD), konsep dan mekanisme kerja algoritma K-Means, proses prapemrosesan data, pembentukan klaster, evaluasi hasil pengelompokan, hingga implementasi analisis menggunakan studi kasus yang relevan.

Materi disajikan secara sistematis agar dapat memberikan pemahaman yang komprehensif, baik dari sisi teori maupun penerapannya dalam menyelesaikan permasalahan nyata.

Di era transformasi digital, pemanfaatan teknologi analisis data menjadi salah satu faktor penting dalam meningkatkan kualitas tata kelola pemerintahan dan pelayanan publik. Penggunaan metode data mining tidak hanya mendukung efisiensi dalam pengolahan data, tetapi juga membantu menghasilkan rekomendasi yang lebih objektif sehingga proses penyaluran bantuan sosial dapat dilakukan secara lebih efektif, adil, dan akuntabel.

Semoga buku Data Mining Berbasis K-Means dalam Analisis Penerima Bantuan Sosial dapat menjadi sumber referensi yang bermanfaat bagi mahasiswa, dosen, peneliti, praktisi teknologi informasi, analis data, aparatur pemerintah, serta seluruh pihak yang memiliki perhatian terhadap penerapan data mining dalam sektor pelayanan publik. Diharapkan kehadiran buku ini dapat memberikan kontribusi dalam pengembangan ilmu pengetahuan serta mendorong pemanfaatan teknologi analitik untuk mendukung pengambilan keputusan yang lebih berkualitas.

DAFTAR ISI

KATA PENGANTAR i

DAFTAR ISI iii

BAB 01

KONSEP DASAR DATA MINING 1
- 1.1 Pengertian Data Mining 1
- 1.2 Perkembangan Data Mining 2
- 1.3 Tujuan dan Manfaat Data Mining 6
- 1.4 Fungsi Data Mining dalam Pengolahan Data 10
- 1.5 Jenis-Jenis Teknik Data Mining 14
- 1.6 Peran Data Mining dalam Pengambilan Keputusan 21
- 1.7 Contoh Penerapan Data Mining dalam Berbagai Bidang 27

BAB 02

KNOWLEDGE DISCOVERY IN DATABASE DAN PENGOLAHAN DATA. 37
- 2.1 Pengertian Knowledge Discovery in Database 37
- 2.2 Tahapan Knowledge Discovery in Database 41
- 2.3 Seleksi Data 46
- 2.4 Pembersihan Data 53
- 2.5 Transformasi Data 59
- 2.6 Evaluasi dan Interpretasi Hasil 68
- 2.7 Pentingnya Kualitas Data dalam Data Mining 77

BAB 03

CLUSTERING DALAM DATA MINING 85
- 3.1 Pengertian Clustering 85
- 3.2 Tujuan Clustering 89
- 3.3 Karakteristik Metode Clustering 93
- 3.4 Perbedaan Clustering dan Classification 99

3.5 Jenis-Jenis Metode Clustering 105
3.6 Kelebihan dan Kekurangan Clustering 113

BAB 04

ALGORITMA K-MEANS DALAM *DATA MINING* 123
4.1 Pengertian Algoritma K-Means 123
4.2 Konsep Dasar K-Means 128
4.3 Prinsip Kerja Algoritma K-Means 135
4.4 Penentuan Jumlah Cluster 142
4.5 Penentuan Centroid Awal 150
4.6 Perhitungan Jarak Data ke Centroid 156
4.7 Pembentukan Cluster 157
4.8 Pembaruan Nilai Centroid 158
4.9 Iterasi dalam Algoritma K-Means 159
4.10 Kelebihan dan Kekurangan Algoritma K-Means 161

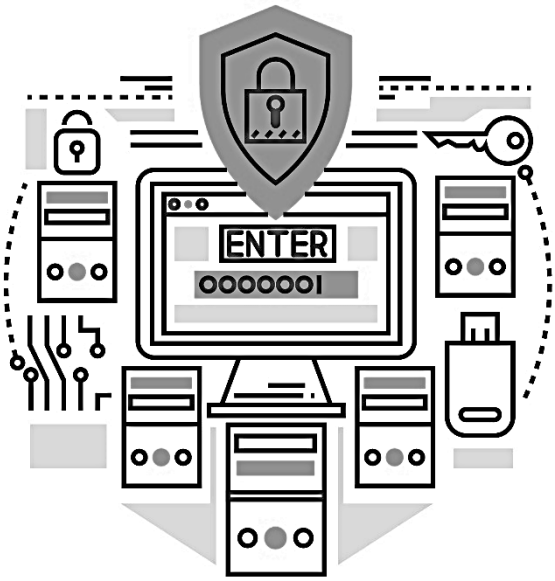
BAB 05

IMPLEMENTASI K-MEANS DALAM ANALISIS DATA 165
5.1 Gambaran Umum Studi Kasus 165
5.2 Pengumpulan Data 166
5.3. Metode Analisis Data 167
5.4 Penentuan Variabel dan Atribut Data 169
5.5 Tahapan Praproses Data 171
5.6 Transformasi Data untuk K-Means 172
5.7 Penerapan Algoritma K-Means 173
5.8 Analisis Hasil Cluster 176
5.9 Interpretasi Setiap Cluster 177
5.10 Evaluasi Hasil Pengelompokan Data 192
5.11 Kesimpulan Implementasi dan Rekomendasi 193

DAFTAR PUSTAKA 197

GLOSARIUM 199

PROFIL PENULIS 203



Bab 01

KONSEP DASAR DATA MINING

1.1 Pengertian Data Mining

Data mining merupakan proses analisis data yang bertujuan untuk menemukan pola, hubungan, kecenderungan, atau informasi bermakna dari sekumpulan data yang berukuran besar. Dalam kajian ilmu komputer dan sistem informasi, data mining dipahami sebagai salah satu pendekatan penting untuk mengubah data mentah menjadi pengetahuan yang dapat digunakan dalam pengambilan keputusan. Data yang tersimpan dalam basis data pada dasarnya belum memiliki makna strategis apabila hanya dikumpulkan tanpa dianalisis secara sistematis. Oleh karena itu, data mining berperan sebagai metode untuk menelusuri nilai tersembunyi di balik data melalui teknik statistik, pembelajaran mesin, pengenalan pola, dan kecerdasan buatan.

Menurut Fayyad, Piatetsky-Shapiro, dan Smyth (1996), data mining merupakan tahap dalam proses Knowledge Discovery in Database

(KDD) yang berfungsi untuk menerapkan algoritma tertentu dalam menemukan pola dari data. Definisi ini menunjukkan bahwa data mining tidak dapat dipandang hanya sebagai kegiatan teknis menjalankan algoritma, tetapi harus ditempatkan dalam kerangka kerja penemuan pengetahuan yang lebih luas. Dalam proses KDD, data mining berada setelah tahapan pemilihan data, pembersihan data, dan transformasi data. Hasil dari proses data mining kemudian masih perlu dievaluasi dan diinterpretasikan agar pola yang ditemukan benar-benar memiliki nilai pengetahuan.

Secara sederhana, data mining dapat diartikan sebagai proses menggali informasi penting dari data dalam jumlah besar dengan menggunakan metode tertentu. Istilah "mining" atau "penambangan" digunakan sebagai analogi karena proses ini menyerupai kegiatan menambang sumber daya berharga dari suatu tempat yang luas dan kompleks. Dalam konteks data, sumber daya yang dimaksud bukan berupa benda fisik, melainkan pola, model, aturan, atau pengetahuan yang sebelumnya tidak terlihat secara

1.2 Perkembangan Data Mining

Perkembangan data mining tidak dapat dilepaskan dari meningkatnya kebutuhan manusia dan organisasi dalam mengelola data secara lebih sistematis. Pada masa awal pemanfaatan komputer, data lebih banyak digunakan untuk kepentingan pencatatan, penyimpanan, dan pelaporan. Data diperlakukan sebagai arsip digital yang berfungsi mendokumentasikan aktivitas organisasi, seperti transaksi, administrasi, keuangan, pendidikan, dan layanan publik. Pada tahap ini, pengolahan data masih berorientasi pada penyajian informasi sederhana, misalnya laporan jumlah, rekapitulasi, atau ringkasan statistik dasar. Namun, seiring bertambahnya volume data, kebutuhan tidak lagi berhenti pada pertanyaan "apa yang terjadi", tetapi berkembang menjadi "pola apa yang dapat ditemukan" dan "keputusan apa yang dapat diambil berdasarkan pola tersebut".

Perubahan tersebut menjadi dasar lahirnya data mining sebagai bidang kajian yang menghubungkan sistem basis data, statistika, kecerdasan buatan, pembelajaran mesin, dan visualisasi data. Fayyad, Piatetsky-Shapiro, dan Smyth (1996) menjelaskan bahwa pada pertengahan 1990-an, istilah *knowledge discovery in databases* mulai digunakan untuk menggambarkan proses penemuan pengetahuan dari data, sedangkan data mining ditempatkan sebagai salah satu tahap utama dalam proses tersebut. Pandangan ini penting karena menunjukkan bahwa perkembangan data mining tidak hanya berkaitan dengan kecanggihan algoritma, tetapi juga dengan kebutuhan metodologis untuk menghasilkan pengetahuan yang valid, dapat dipahami, dan berguna dalam pengambilan keputusan.

Pada perkembangan berikutnya, data mining mulai digunakan secara lebih luas dalam berbagai sektor. Organisasi tidak hanya menyimpan data, tetapi juga berupaya memanfaatkannya untuk menemukan kecenderungan perilaku, membangun segmentasi, melakukan prediksi, serta mendukung proses perencanaan. Dalam bidang bisnis, data mining digunakan untuk memahami perilaku pelanggan, menganalisis pola pembelian, dan menentukan strategi pemasaran. Dalam bidang pendidikan, data mining dapat digunakan untuk menganalisis prestasi belajar, mengelompokkan karakteristik peserta didik, dan mendukung pengambilan keputusan akademik. Sementara itu, dalam bidang kesehatan, data mining dapat membantu menemukan pola pada data pasien, riwayat penyakit, atau efektivitas layanan kesehatan.

Perkembangan data mining juga ditandai dengan munculnya kebutuhan terhadap proses kerja yang lebih terstruktur. Salah satu model yang banyak digunakan adalah CRISP-DM atau *Cross-Industry Standard Process for Data Mining*. Model ini memperkenalkan tahapan kerja yang sistematis, meliputi pemahaman masalah, pemahaman data, persiapan data, pemodelan, evaluasi, dan penerapan hasil. Kehadiran CRISP-DM menunjukkan bahwa data mining tidak cukup dilakukan hanya dengan menjalankan perangkat lunak atau algoritma tertentu. Data

mining perlu diawali dengan pemahaman terhadap masalah, kesesuaian data, dan tujuan analisis agar hasil yang diperoleh tidak menyimpang dari kebutuhan pengguna atau peneliti.

Pada masa berikutnya, perkembangan data mining semakin dipengaruhi oleh kemajuan teknologi penyimpanan data dan kemampuan komputasi. Data yang sebelumnya hanya tersimpan dalam bentuk tabel terstruktur kemudian berkembang menjadi data dalam berbagai bentuk, seperti teks, gambar, log aktivitas, data transaksi digital, data sensor, dan data media sosial. Perubahan ini membuat data mining tidak lagi terbatas pada data kecil dan sederhana, tetapi juga harus mampu menangani data yang besar, beragam, dan terus bertambah. Dalam konteks ini, data mining beririsan dengan bidang *big data analytics* dan *data science*, karena ketiganya sama-sama berupaya mengekstraksi pengetahuan dari data untuk mendukung tindakan dan keputusan yang lebih tepat. Dhar (2013) menekankan bahwa data science berkaitan dengan ekstraksi pengetahuan yang dapat digeneralisasi dari data, dengan perhatian penting pada kemampuan prediktif dalam pengambilan keputusan.

Dari sisi metode, perkembangan data mining terlihat dari semakin beragamnya teknik yang digunakan. Pada awalnya, analisis data banyak bertumpu pada teknik statistik konvensional. Selanjutnya, data mining berkembang dengan memanfaatkan algoritma pembelajaran mesin, seperti *decision tree*, *neural network*, *support vector machine*, *association rule*, dan berbagai metode clustering. Teknik-teknik tersebut digunakan sesuai dengan tujuan analisis yang ingin dicapai. Jika tujuan analisis adalah mengelompokkan data berdasarkan kemiripan karakteristik, maka metode clustering dapat digunakan. Salah satu algoritma clustering yang paling banyak dikenal adalah K-Means karena memiliki konsep yang relatif sederhana, efisien, dan mudah diterapkan pada berbagai kasus pengelompokan data.

Perkembangan data mining juga memperlihatkan perubahan orientasi dari sekadar analisis deskriptif menuju analisis prediktif

dan preskriptif. Analisis deskriptif digunakan untuk menggambarkan keadaan data yang telah terjadi. Analisis prediktif digunakan untuk memperkirakan kemungkinan kejadian pada masa mendatang berdasarkan pola data sebelumnya. Sementara itu, analisis preskriptif berupaya memberikan alternatif tindakan berdasarkan hasil analisis. Perubahan orientasi ini memperluas fungsi data mining, dari alat bantu analisis menjadi bagian penting dalam sistem pendukung keputusan. Dengan demikian, data mining tidak hanya menghasilkan informasi, tetapi juga membantu organisasi atau peneliti menyusun strategi yang lebih tepat.

Dalam konteks penelitian ilmiah, perkembangan data mining mendorong peneliti untuk lebih hati-hati dalam memperlakukan data. Ketersediaan perangkat lunak analisis tidak serta-merta menjamin kualitas hasil penelitian. Kesalahan dalam memilih atribut, menentukan metode, membersihkan data, atau menafsirkan hasil dapat menyebabkan temuan yang tidak valid. Oleh karena itu, perkembangan data mining harus dipahami tidak hanya dari aspek teknologi, tetapi juga dari aspek metodologi penelitian. Seorang peneliti perlu memastikan bahwa data yang digunakan relevan dengan tujuan penelitian, metode yang dipilih sesuai dengan karakteristik data, dan hasil analisis diinterpretasikan secara objektif.

Dalam buku ini, perkembangan data mining menjadi landasan penting untuk memahami posisi algoritma K-Means. K-Means tidak muncul sebagai metode yang berdiri sendiri, melainkan sebagai bagian dari perkembangan teknik data mining, khususnya pada cabang clustering. Algoritma ini berkembang karena kebutuhan untuk mengelompokkan data dalam jumlah besar secara cepat dan sistematis. Dengan memahami perkembangan data mining, pembaca dapat melihat bahwa K-Means bukan hanya prosedur matematis untuk membagi data ke dalam beberapa kelompok, tetapi juga bagian dari tradisi ilmiah yang lebih luas dalam menemukan pola, struktur, dan pengetahuan dari data.

Berdasarkan uraian tersebut, perkembangan data mining dapat dipahami sebagai proses evolusi dari pengolahan data sederhana menuju penemuan pengetahuan berbasis data. Perkembangan ini dipengaruhi oleh kemajuan basis data, statistika, kecerdasan buatan, pembelajaran mesin, teknologi komputasi, dan kebutuhan pengambilan keputusan yang semakin kompleks. Data mining terus berkembang karena data semakin banyak, bentuk data semakin beragam, dan kebutuhan terhadap informasi yang akurat semakin tinggi. Oleh sebab itu, pemahaman mengenai perkembangan data mining menjadi penting sebelum membahas lebih jauh penerapan algoritma K-Means dalam proses pengelompokan data.

1.3 Tujuan dan Manfaat Data Mining

Data mining memiliki tujuan utama untuk membantu peneliti, lembaga, maupun organisasi dalam memahami data secara lebih mendalam. Data yang tersimpan dalam jumlah besar sering kali belum memberikan nilai pengetahuan apabila hanya dilihat sebagai kumpulan angka, catatan, atau atribut. Melalui data mining, data tersebut dapat diolah menjadi pola, informasi, dan pengetahuan yang lebih terarah. Dengan demikian, data mining tidak hanya berfungsi sebagai alat pengolahan data, tetapi juga sebagai pendekatan analitis untuk menjawab masalah, membaca kecenderungan, dan mendukung proses pengambilan keputusan secara lebih objektif.

Salah satu tujuan penting data mining adalah menemukan pola tersembunyi yang tidak mudah dikenali melalui pengamatan biasa. Pola tersebut dapat berupa kesamaan karakteristik, hubungan antarvariabel, kecenderungan tertentu, atau struktur kelompok dalam data. Dalam proses penemuan pengetahuan, pola yang dihasilkan tidak cukup hanya menarik secara teknis, tetapi juga harus memiliki nilai validitas, kebaruan, kegunaan, dan dapat dipahami oleh pengguna data. Fayyad, Piatetsky-Shapiro, dan Smyth (1996) menegaskan bahwa proses penemuan pengetahuan dari data diarahkan untuk mengidentifikasi pola yang valid, baru,

bermanfaat, dan dapat dimengerti. Dengan prinsip tersebut, data mining tidak boleh dipahami sebagai proses mencari hasil secara sembarangan, melainkan sebagai proses ilmiah yang perlu memperhatikan kualitas data, ketepatan metode, dan makna hasil analisis.

Tujuan berikutnya adalah membantu proses pengelompokan, pengklasifikasian, prediksi, estimasi, dan penemuan hubungan dalam data. Setiap tujuan tersebut memiliki arah analisis yang berbeda. Pengelompokan digunakan untuk membagi data ke dalam beberapa kelompok berdasarkan kemiripan karakteristik. Pengklasifikasian digunakan untuk menempatkan data ke dalam kategori yang telah ditentukan. Prediksi digunakan untuk memperkirakan kemungkinan yang akan terjadi berdasarkan pola data sebelumnya. Sementara itu, analisis asosiasi digunakan untuk menemukan keterkaitan antaritem atau antarvariabel. Han, Kamber, dan Pei (2011) menjelaskan bahwa data mining mencakup sejumlah fungsi penting, seperti deskripsi karakteristik data, klasifikasi, prediksi, analisis asosiasi, dan clustering. Dalam konteks buku ini, fungsi clustering menjadi perhatian utama karena berkaitan langsung dengan algoritma K-Means.

Data mining juga bertujuan untuk meningkatkan kualitas pengambilan keputusan. Keputusan yang hanya didasarkan pada intuisi atau dugaan sering kali memiliki risiko subjektivitas yang tinggi. Dengan bantuan data mining, keputusan dapat disusun berdasarkan pola yang diperoleh dari data. Hal ini tidak berarti bahwa data mining menggantikan peran manusia dalam mengambil keputusan, tetapi memberikan dasar analisis yang lebih kuat agar keputusan yang diambil lebih rasional, terukur, dan dapat dipertanggungjawabkan. Dalam penelitian ilmiah, tujuan ini sangat penting karena hasil analisis harus mampu menjelaskan hubungan antara data, metode, dan simpulan secara logis.

Selain itu, data mining bertujuan untuk menyederhanakan data yang kompleks menjadi informasi yang lebih mudah dipahami. Dalam banyak kasus, data memiliki jumlah atribut yang banyak,

ukuran yang besar, dan hubungan yang tidak sederhana. Kondisi tersebut dapat menyulitkan peneliti dalam membaca kecenderungan data secara langsung. Data mining membantu menyusun struktur pemahaman terhadap data, misalnya dengan membentuk kelompok, membuat model klasifikasi, atau menemukan aturan tertentu. Dengan cara ini, data yang semula sulit dibaca dapat diubah menjadi informasi yang lebih sistematis.

Manfaat data mining dapat dilihat dari kemampuannya dalam menghasilkan informasi yang bernilai bagi berbagai bidang. Dalam bidang pendidikan, data mining dapat dimanfaatkan untuk menganalisis prestasi akademik, mengelompokkan karakteristik peserta didik, mengidentifikasi kecenderungan hasil belajar, atau membantu lembaga pendidikan dalam menyusun strategi pembinaan. Dalam bidang bisnis, data mining bermanfaat untuk memahami perilaku pelanggan, menentukan segmentasi pasar, menganalisis transaksi, dan menyusun strategi pemasaran. Dalam bidang kesehatan, data mining dapat membantu membaca pola layanan, riwayat pasien, atau kecenderungan penyakit berdasarkan data yang tersedia.

Manfaat lain dari data mining adalah membantu menemukan informasi yang sebelumnya tidak terlihat secara langsung. Data dalam jumlah besar sering kali menyimpan pola yang tidak mudah dikenali melalui analisis manual. Dengan teknik data mining, peneliti dapat menemukan kecenderungan yang mungkin terlewat apabila hanya menggunakan pengamatan biasa. Informasi seperti kelompok data yang memiliki karakteristik serupa, atribut yang dominan, atau pola hubungan tertentu dapat menjadi dasar untuk memahami masalah secara lebih luas. Dalam konteks ini, data mining memberikan nilai tambah karena mampu mengubah data menjadi pengetahuan yang berguna.

Data mining juga bermanfaat dalam meningkatkan efisiensi analisis. Tanpa bantuan metode data mining, proses membaca dan menafsirkan data dalam jumlah besar dapat memerlukan waktu yang panjang dan berisiko menimbulkan kesalahan. Data mining

memungkinkan proses analisis dilakukan secara lebih cepat, sistematis, dan konsisten. Namun, efisiensi tersebut tetap harus diimbangi dengan ketelitian peneliti dalam mempersiapkan data dan menafsirkan hasil. Witten, Frank, Hall, dan Pal (2017) menekankan bahwa data mining berkaitan erat dengan teknik pembelajaran mesin yang digunakan untuk membangun model dari data, sehingga hasil yang diperoleh sangat dipengaruhi oleh pemilihan teknik dan kualitas data yang digunakan.

Dalam pelaksanaan penelitian, manfaat data mining juga terlihat pada kemampuannya mendukung proses evaluasi dan perencanaan. Hasil analisis data mining dapat digunakan untuk menilai kondisi yang sedang terjadi, mengidentifikasi kelompok yang membutuhkan perhatian khusus, serta menyusun rekomendasi berdasarkan hasil temuan. Model CRISP-DM menempatkan pemahaman masalah, pemahaman data, persiapan data, pemodelan, evaluasi, dan penerapan sebagai tahapan yang saling berhubungan dalam proyek data mining. Hal ini menunjukkan bahwa manfaat data mining tidak hanya terletak pada hasil akhir, tetapi juga pada proses sistematis yang membantu peneliti memahami masalah sejak awal hingga tahap interpretasi hasil.

Dalam kaitannya dengan algoritma K-Means, tujuan dan manfaat data mining menjadi dasar penting untuk memahami mengapa proses pengelompokan data perlu dilakukan. K-Means digunakan ketika peneliti ingin mengetahui struktur kelompok dalam data berdasarkan kemiripan karakteristik. Melalui algoritma ini, data dapat dibagi ke dalam beberapa cluster sehingga peneliti dapat melihat kelompok mana yang memiliki ciri tertentu. Oleh karena itu, manfaat K-Means tidak hanya terletak pada pembentukan cluster, tetapi juga pada kemampuan membantu peneliti menafsirkan pola kelompok yang muncul dari data.

Berdasarkan uraian tersebut, tujuan data mining dapat dipahami sebagai upaya untuk menemukan pola, membangun pemahaman terhadap data, mendukung pengambilan keputusan, serta

menghasilkan pengetahuan yang bermanfaat. Sementara itu, manfaat data mining terletak pada kemampuannya dalam mengolah data besar menjadi informasi yang lebih bermakna, meningkatkan efisiensi analisis, mengurangi subjektivitas keputusan, dan membantu peneliti dalam menyusun rekomendasi berbasis data. Dengan pemahaman ini, data mining menjadi fondasi penting sebelum membahas penerapan algoritma K-Means secara lebih teknis pada bab-bab berikutnya.

1.4 Fungsi Data Mining dalam Pengolahan Data

Data mining memiliki fungsi penting dalam pengolahan data karena mampu membantu peneliti memahami data secara lebih terarah, sistematis, dan bermakna. Dalam penelitian ilmiah, data tidak cukup hanya dikumpulkan, disimpan, dan disajikan dalam bentuk tabel. Data perlu diolah agar dapat memberikan gambaran yang jelas mengenai pola, kecenderungan, hubungan, atau struktur tertentu yang terdapat di dalamnya. Pada tahap inilah data mining berperan sebagai pendekatan analitis yang menghubungkan data mentah dengan informasi yang dapat ditafsirkan secara ilmiah.

Dalam pengolahan data, data mining berfungsi untuk menemukan pola yang tidak selalu terlihat melalui pengamatan langsung. Pola tersebut dapat berupa kemiripan antarobjek, perbedaan karakteristik antar kelompok, hubungan antarvariabel, kecenderungan nilai tertentu, atau kemungkinan munculnya suatu kejadian berdasarkan data sebelumnya. Fayyad, Piatetsky-Shapiro, dan Smyth (1996) menempatkan data mining sebagai tahap penerapan algoritma untuk menemukan pola dalam proses penemuan pengetahuan dari basis data. Dengan demikian, fungsi data mining tidak berdiri sendiri, tetapi menjadi bagian dari rangkaian pengolahan data yang mencakup pemilihan data, pembersihan data, transformasi data, penemuan pola, dan interpretasi hasil.

Secara umum, fungsi data mining dalam pengolahan data dapat dikelompokkan sebagai berikut.

- Membantu memahami karakteristik data
  Data mining berfungsi untuk membantu peneliti mengenali ciri-ciri utama dari data yang dianalisis. Melalui proses ini, peneliti dapat mengetahui sebaran data, pola umum, kecenderungan atribut, serta kemungkinan hubungan antarvariabel. Pemahaman terhadap karakteristik data menjadi dasar penting sebelum peneliti memilih metode analisis yang sesuai. Tanpa pemahaman yang baik terhadap data, proses analisis berisiko menghasilkan kesimpulan yang kurang tepat.
- Mendukung proses prapengolahan data
  Meskipun data mining bukanlah tahap pembersihan data, penerapannya sangat berkaitan dengan kualitas data yang digunakan. Data yang tidak lengkap, tidak konsisten, atau memiliki nilai yang tidak wajar dapat memengaruhi hasil analisis. Oleh karena itu, sebelum proses data mining dilakukan, data perlu dipersiapkan melalui tahapan seleksi, pembersihan, integrasi, dan transformasi. Dalam model CRISP-DM, persiapan data merupakan salah satu tahap penting sebelum pemodelan dilakukan, sehingga proses analisis tidak hanya bergantung pada algoritma, tetapi juga pada kesiapan data yang digunakan.
- Menemukan pola dan hubungan dalam data
  Salah satu fungsi utama data mining adalah menemukan pola dan hubungan yang tersembunyi dalam data. Pola tersebut dapat menjelaskan kecenderungan tertentu yang terjadi pada objek penelitian. Misalnya, dalam data pendidikan, data mining dapat membantu melihat hubungan antara nilai akademik, kehadiran, latar belakang peserta didik, atau kategori prestasi. Dalam data bisnis, data mining dapat digunakan untuk menemukan hubungan antara pola pembelian, perilaku pelanggan, dan segmentasi pasar.
- Mengelompokkan data berdasarkan kemiripan karakteristik
  Data mining berfungsi untuk mengelompokkan objek data ke dalam beberapa kelompok berdasarkan tingkat kemiripan tertentu. Fungsi ini dikenal sebagai clustering. Clustering

berguna ketika peneliti belum memiliki kategori awal, tetapi ingin menemukan struktur kelompok yang terbentuk secara alami dari data. Dalam konteks buku ini, fungsi clustering menjadi sangat penting karena berkaitan langsung dengan algoritma K-Means. K-Means digunakan untuk membagi data ke dalam sejumlah cluster berdasarkan kedekatan data terhadap titik pusat cluster atau centroid.

- Mengklasifikasikan data ke dalam kategori tertentu
  Selain clustering, data mining juga berfungsi untuk melakukan klasifikasi. Klasifikasi digunakan apabila kategori atau kelas data telah diketahui sebelumnya. Fungsi ini membantu peneliti membangun model yang dapat menempatkan data baru ke dalam kategori yang sesuai. Berbeda dengan clustering yang bersifat tidak terawasi, klasifikasi termasuk teknik terawasi karena menggunakan data berlabel sebagai dasar pembentukan model. Han, Kamber, dan Pei (2011) menjelaskan bahwa fungsi data mining mencakup beberapa bentuk analisis, antara lain deskripsi data, asosiasi, klasifikasi, prediksi, clustering, dan deteksi data menyimpang.
- Mendukung prediksi berdasarkan pola data
  Data mining juga berfungsi untuk memperkirakan kemungkinan yang akan terjadi berdasarkan pola data sebelumnya. Fungsi prediksi banyak digunakan dalam penelitian yang membutuhkan proyeksi, seperti memperkirakan keberhasilan, risiko, permintaan, atau kecenderungan perilaku. Dalam pengolahan data, prediksi tidak dilakukan berdasarkan dugaan semata, tetapi dibangun melalui model yang mempelajari pola dari data historis. Oleh karena itu, kualitas data masa lalu sangat menentukan ketepatan hasil prediksi.
- Mendeteksi data yang tidak wajar atau menyimpang
  Fungsi lain dari data mining adalah membantu mengenali data yang berbeda secara signifikan dari pola umum. Data semacam ini sering disebut sebagai outlier atau anomali. Dalam beberapa penelitian, data menyimpang dapat dianggap sebagai gangguan yang perlu diperiksa kembali. Namun, dalam konteks tertentu,

data menyimpang justru dapat menjadi temuan penting. Misalnya, anomali dalam data transaksi dapat menunjukkan kesalahan pencatatan atau indikasi perilaku yang tidak biasa. Oleh karena itu, deteksi data menyimpang perlu ditafsirkan secara hati-hati sesuai dengan konteks penelitian.

- Menyederhanakan data yang kompleks
  Data yang besar dan memiliki banyak atribut sering kali sulit dipahami secara langsung. Data mining membantu menyederhanakan kompleksitas tersebut melalui pembentukan pola, model, cluster, atau aturan tertentu. Dengan adanya hasil data mining, peneliti dapat melihat struktur data secara lebih ringkas tanpa kehilangan makna penting dari data tersebut. Fungsi ini sangat bermanfaat dalam penelitian yang melibatkan banyak variabel atau jumlah data yang cukup besar.
- Mendukung interpretasi hasil pengolahan data
  Data mining tidak hanya berfungsi pada tahap perhitungan, tetapi juga mendukung interpretasi hasil. Hasil analisis berupa cluster, model klasifikasi, aturan asosiasi, atau nilai prediksi perlu dijelaskan dalam bahasa ilmiah agar dapat dipahami oleh pembaca. Interpretasi menjadi tahap penting karena angka atau keluaran perangkat lunak tidak otomatis menjadi pengetahuan. Peneliti tetap perlu menjelaskan makna hasil, relevansinya dengan masalah penelitian, serta implikasinya terhadap pengambilan keputusan.

Dalam pengolahan data berbasis penelitian, fungsi data mining perlu ditempatkan secara proporsional. Data mining bukan alat yang secara otomatis menghasilkan kebenaran, melainkan metode analisis yang hasilnya sangat dipengaruhi oleh kualitas data, ketepatan pemilihan atribut, kesesuaian algoritma, dan kemampuan peneliti dalam menafsirkan hasil. Oleh karena itu, penggunaan data mining harus dilakukan secara kritis dan metodologis. Peneliti perlu memastikan bahwa data yang digunakan relevan dengan tujuan penelitian, metode yang dipilih

sesuai dengan karakteristik data, serta hasil analisis tidak ditafsirkan melebihi batas kemampuan data.

Dalam kaitannya dengan algoritma K-Means, fungsi data mining yang paling menonjol adalah fungsi pengelompokan data. K-Means membantu peneliti membagi objek data ke dalam beberapa kelompok berdasarkan kemiripan karakteristik. Hasil pengelompokan tersebut dapat digunakan untuk memahami perbedaan antarcluster, mengenali kelompok dominan, serta menyusun interpretasi terhadap struktur data. Misalnya, dalam bidang pendidikan, K-Means dapat digunakan untuk mengelompokkan peserta didik berdasarkan nilai, kemampuan awal, atau karakteristik akademik tertentu. Dalam bidang bisnis, K-Means dapat digunakan untuk mengelompokkan pelanggan berdasarkan perilaku transaksi atau preferensi pembelian.

Namun, perlu dipahami bahwa hasil clustering dengan K-Means tidak secara otomatis menunjukkan kategori yang benar secara mutlak. Cluster yang terbentuk merupakan hasil perhitungan berdasarkan atribut yang digunakan dan jumlah cluster yang ditentukan. Oleh karena itu, interpretasi hasil K-Means harus disesuaikan dengan konteks penelitian. Peneliti perlu menjelaskan alasan pemilihan atribut, jumlah cluster, dan makna setiap kelompok yang terbentuk. Dengan cara tersebut, fungsi data mining dalam pengolahan data tidak hanya bersifat teknis, tetapi juga memiliki nilai ilmiah yang dapat dipertanggungjawabkan.

1.5 Jenis-Jenis Teknik Data Mining

Teknik data mining merupakan cara atau metode yang digunakan untuk menemukan pola, hubungan, struktur, atau model tertentu dari data. Setiap teknik memiliki tujuan analisis yang berbeda sehingga pemilihannya harus disesuaikan dengan karakteristik data, bentuk masalah, dan hasil yang ingin dicapai. Dalam penelitian ilmiah, pemilihan teknik data mining tidak boleh dilakukan hanya karena suatu algoritma populer atau mudah

digunakan, tetapi harus didasarkan pada kebutuhan analisis dan kesesuaian metode terhadap data yang tersedia.

Secara umum, teknik data mining dapat dibedakan berdasarkan tujuan analisisnya. Ada teknik yang digunakan untuk menjelaskan pola yang telah ada dalam data, ada pula teknik yang digunakan untuk memperkirakan kemungkinan yang akan terjadi. Han, Kamber, dan Pei (2011) menjelaskan bahwa fungsi data mining mencakup beberapa bentuk analisis, seperti deskripsi data, penambangan pola berulang, asosiasi, klasifikasi, prediksi, clustering, dan analisis data menyimpang. Pembagian ini menunjukkan bahwa data mining memiliki cakupan yang luas dan tidak hanya terbatas pada satu jenis metode tertentu.

Dalam penerapannya, teknik data mining dapat dikelompokkan ke dalam beberapa jenis utama, yaitu classification, prediction, clustering, association rule mining, sequential pattern mining, dan outlier analysis. Setiap teknik tersebut memiliki peran yang berbeda dalam proses pengolahan data. Berikut uraian jenis-jenis teknik data mining yang umum digunakan dalam penelitian dan praktik analisis data.

1. Classification

Classification atau klasifikasi merupakan teknik data mining yang digunakan untuk menempatkan data ke dalam kelas atau kategori tertentu. Teknik ini digunakan apabila kategori data telah diketahui sebelumnya. Oleh karena itu, classification termasuk ke dalam teknik *supervised learning* karena proses pembentukan model dilakukan dengan menggunakan data yang telah memiliki label atau kelas.

Dalam classification, algoritma mempelajari pola dari data yang sudah dikategorikan, kemudian pola tersebut digunakan untuk menentukan kategori data baru. Sebagai contoh, dalam bidang pendidikan, data peserta didik dapat diklasifikasikan ke dalam kategori prestasi tinggi, sedang, atau rendah berdasarkan nilai akademik, kehadiran, atau atribut lain yang relevan. Dalam bidang

kesehatan, data pasien dapat diklasifikasikan berdasarkan tingkat risiko penyakit tertentu. Sementara itu, dalam bidang bisnis, pelanggan dapat diklasifikasikan berdasarkan kemungkinan membeli produk, berhenti berlangganan, atau merespons promosi.

Beberapa algoritma yang umum digunakan dalam teknik classification antara lain:

- Decision Tree, yaitu algoritma klasifikasi berbentuk struktur pohon keputusan.
- Naïve Bayes, yaitu algoritma berbasis probabilitas.
- Support Vector Machine, yaitu algoritma yang mencari batas pemisah terbaik antar kelas.
- K-Nearest Neighbor, yaitu algoritma yang menentukan kelas berdasarkan kedekatan data dengan data lain.
- Artificial Neural Network, yaitu algoritma yang meniru pola kerja jaringan saraf untuk mengenali pola kompleks.

Teknik classification sangat berguna ketika peneliti ingin membangun model pengambilan keputusan berdasarkan data historis. Namun, penggunaan teknik ini membutuhkan data berlabel yang baik. Apabila label pada data tidak akurat, maka model yang dibentuk juga berpotensi menghasilkan klasifikasi yang keliru.

2. Prediction

Prediction atau prediksi merupakan teknik data mining yang digunakan untuk memperkirakan nilai atau kejadian pada masa mendatang berdasarkan pola data sebelumnya. Teknik ini memiliki kedekatan dengan classification, tetapi keduanya tidak sepenuhnya sama. Classification biasanya digunakan untuk memprediksi kategori, sedangkan prediction dapat digunakan untuk memperkirakan nilai numerik atau kecenderungan tertentu.

Dalam penelitian, teknik prediksi dapat digunakan untuk memperkirakan hasil belajar mahasiswa, potensi kelulusan, permintaan produk, penjualan, risiko kredit, atau kecenderungan perilaku pengguna. Witten, Frank, Hall, dan Pal (2017)

menjelaskan bahwa data mining berkaitan erat dengan pembentukan model dari data, termasuk model yang digunakan untuk melakukan prediksi dan evaluasi hasil. Dengan demikian, prediksi dalam data mining bukan sekadar dugaan, melainkan hasil dari proses pemodelan berbasis data.

Teknik prediction dapat menggunakan beberapa pendekatan, antara lain regresi, decision tree, neural network, dan metode pembelajaran mesin lainnya. Dalam penggunaannya, peneliti perlu memperhatikan kualitas data masa lalu karena prediksi yang dihasilkan sangat bergantung pada pola historis yang dipelajari oleh model. Apabila data historis tidak representatif, maka hasil prediksi dapat menjadi kurang akurat.

3. Clustering

Clustering merupakan teknik data mining yang digunakan untuk mengelompokkan data berdasarkan kemiripan karakteristik. Berbeda dengan classification, clustering tidak memerlukan label kelas sebelumnya. Oleh karena itu, clustering termasuk ke dalam teknik *unsupervised learning*. Teknik ini digunakan ketika peneliti belum mengetahui kelompok data secara pasti, tetapi ingin menemukan struktur kelompok yang terbentuk secara alami dari data.

Clustering banyak digunakan untuk segmentasi data. Dalam bidang pendidikan, clustering dapat digunakan untuk mengelompokkan peserta didik berdasarkan kemampuan akademik, pola belajar, atau hasil evaluasi. Dalam bidang bisnis, clustering dapat digunakan untuk membagi pelanggan ke dalam beberapa segmen berdasarkan perilaku pembelian. Dalam bidang pemerintahan, clustering dapat digunakan untuk mengelompokkan wilayah berdasarkan karakteristik sosial, ekonomi, atau demografis.

Beberapa algoritma clustering yang umum digunakan antara lain:

- K-Means, yaitu algoritma clustering berbasis centroid.
- K-Medoids, yaitu metode clustering yang menggunakan objek nyata sebagai pusat kelompok.

- Hierarchical Clustering, yaitu metode yang membentuk struktur kelompok bertingkat.
- DBSCAN, yaitu metode clustering berbasis kepadatan data.
- Fuzzy C-Means, yaitu metode clustering yang memungkinkan satu data memiliki derajat keanggotaan pada lebih dari satu cluster.

Dalam buku ini, pembahasan difokuskan pada algoritma K-Means. Algoritma ini banyak digunakan karena memiliki konsep yang sederhana, proses perhitungan yang relatif mudah dipahami, dan dapat diterapkan pada berbagai jenis data numerik. Namun, K-Means juga memiliki keterbatasan, terutama dalam penentuan jumlah cluster, pemilihan centroid awal, dan sensitivitas terhadap data pencilan. Oleh sebab itu, penggunaan K-Means tetap memerlukan pemahaman metodologis agar hasil pengelompokan dapat ditafsirkan secara tepat.

4. Association Rule Mining

Association rule mining merupakan teknik data mining yang digunakan untuk menemukan hubungan atau keterkaitan antaritem dalam data. Teknik ini sering digunakan untuk mengetahui pola kemunculan bersama antara satu item dengan item lainnya. Salah satu contoh paling umum adalah analisis keranjang belanja atau *market basket analysis*, yaitu analisis yang digunakan untuk mengetahui produk apa saja yang sering dibeli secara bersamaan oleh pelanggan.

Dalam association rule mining, hasil analisis biasanya dinyatakan dalam bentuk aturan, misalnya "jika item A muncul, maka item B cenderung muncul". Aturan tersebut dinilai menggunakan ukuran tertentu, seperti support, confidence, dan lift. Support menunjukkan seberapa sering suatu kombinasi item muncul dalam data. Confidence menunjukkan tingkat kepercayaan terhadap hubungan antaritem. Lift menunjukkan kekuatan hubungan antara item dibandingkan dengan kemunculan secara acak.

Contoh sederhana aturan asosiasi dalam konteks bisnis adalah sebagai berikut:

Jika pelanggan membeli printer, maka pelanggan juga cenderung membeli tinta.

Dalam konteks pendidikan, association rule mining dapat digunakan untuk melihat pola hubungan antara mata pelajaran, aktivitas belajar, atau capaian tertentu. Namun, peneliti perlu berhati-hati dalam menafsirkan hasil asosiasi. Hubungan yang ditemukan dalam data tidak selalu menunjukkan hubungan sebab-akibat. Association rule mining hanya menunjukkan pola keterkaitan, bukan membuktikan bahwa satu variabel menyebabkan variabel lainnya.

5. Sequential Pattern Mining

Sequential pattern mining merupakan teknik data mining yang digunakan untuk menemukan pola urutan kejadian dalam data. Teknik ini berbeda dengan association rule mining karena memperhatikan urutan waktu atau urutan peristiwa. Dengan kata lain, sequential pattern mining tidak hanya melihat item apa yang muncul bersama, tetapi juga melihat pola kemunculan item atau kejadian dari waktu ke waktu.

Teknik ini banyak digunakan dalam analisis perilaku pengguna, riwayat transaksi, aktivitas pembelajaran digital, dan alur layanan. Misalnya, dalam sistem pembelajaran daring, sequential pattern mining dapat digunakan untuk mengetahui urutan aktivitas mahasiswa sebelum memperoleh hasil belajar tertentu. Dalam bisnis digital, teknik ini dapat membantu melihat urutan tindakan pelanggan sebelum melakukan pembelian.

Contoh pola sekuensial sederhana adalah sebagai berikut:

Pengguna membuka halaman produk, membaca ulasan, memasukkan produk ke keranjang, lalu melakukan pembelian.

Pola seperti ini dapat memberikan informasi penting bagi pengambil keputusan. Namun, seperti teknik data mining lainnya,

hasil sequential pattern mining tetap harus ditafsirkan sesuai konteks. Urutan kejadian yang sering muncul tidak selalu berarti bahwa urutan tersebut berlaku untuk semua kasus.

6. Outlier Analysis

Outlier analysis atau analisis data pencilan merupakan teknik data mining yang digunakan untuk menemukan data yang memiliki karakteristik berbeda secara signifikan dari pola umum. Data pencilan dapat muncul karena kesalahan input, kondisi khusus, atau kejadian yang memang tidak biasa. Dalam beberapa kasus, outlier dianggap sebagai gangguan yang perlu ditinjau kembali. Namun, dalam kasus lain, outlier justru menjadi informasi penting yang perlu dianalisis lebih lanjut.

Dalam bidang keuangan, outlier dapat menunjukkan transaksi yang tidak biasa. Dalam bidang pendidikan, outlier dapat menunjukkan peserta didik dengan capaian yang sangat berbeda dari kelompoknya. Dalam bidang kesehatan, outlier dapat menunjukkan kondisi pasien yang memerlukan perhatian khusus. Oleh karena itu, outlier tidak boleh langsung dihapus tanpa alasan yang jelas. Peneliti perlu memeriksa apakah data tersebut merupakan kesalahan, variasi alami, atau temuan penting.

Han et al. (2011) menempatkan analisis outlier sebagai salah satu bagian penting dalam data mining karena data yang menyimpang dari pola umum dapat mengandung informasi bernilai. Dalam konteks penelitian, analisis outlier membantu peneliti menjaga kualitas data sekaligus membuka kemungkinan ditemukannya fenomena yang tidak terlihat melalui pola umum.

7. Estimation

Estimation atau estimasi merupakan teknik yang digunakan untuk memperkirakan nilai tertentu berdasarkan atribut yang tersedia dalam data. Teknik ini memiliki kedekatan dengan prediction, tetapi estimation biasanya lebih berfokus pada nilai numerik. Misalnya, memperkirakan nilai akhir mahasiswa berdasarkan nilai tugas dan kehadiran, memperkirakan pendapatan pelanggan

berdasarkan riwayat transaksi, atau memperkirakan waktu penyelesaian layanan berdasarkan data sebelumnya.

Estimasi banyak menggunakan pendekatan statistik dan pembelajaran mesin, seperti regresi linear, regresi berganda, neural network, atau metode lainnya. Dalam penelitian kuantitatif, teknik estimasi dapat membantu peneliti memahami hubungan antara variabel prediktor dan variabel hasil. Namun, estimasi yang baik membutuhkan data yang relevan, bebas dari kesalahan besar, dan memiliki hubungan yang cukup jelas antara variabel yang dianalisis.

8. Summarization

Summarization merupakan teknik data mining yang digunakan untuk menyajikan ringkasan atau gambaran umum dari data. Teknik ini membantu peneliti memahami karakteristik utama data sebelum masuk ke analisis yang lebih kompleks. Summarization dapat berupa ringkasan statistik, distribusi data, visualisasi, atau deskripsi pola umum yang ditemukan dalam dataset.

Meskipun terlihat sederhana, summarization memiliki peran penting dalam penelitian berbasis data. Melalui ringkasan data, peneliti dapat melihat kecenderungan awal, mendeteksi ketidakwajaran data, dan menentukan apakah data layak dianalisis lebih lanjut. Dalam praktik data mining yang sistematis, tahap pemahaman data dan persiapan data menjadi bagian penting sebelum pemodelan dilakukan, sebagaimana dijelaskan dalam kerangka CRISP-DM.

1.6 Peran Data Mining dalam Pengambilan Keputusan

Pengambilan keputusan merupakan proses menentukan pilihan terbaik dari beberapa alternatif yang tersedia. Dalam konteks organisasi, lembaga pendidikan, bisnis, pemerintahan, maupun penelitian ilmiah, keputusan yang baik seharusnya tidak hanya didasarkan pada intuisi, pengalaman, atau kebiasaan, tetapi juga

didukung oleh data yang relevan. Data mining berperan penting dalam proses tersebut karena mampu membantu mengubah data menjadi informasi yang lebih terstruktur, sehingga pengambil keputusan dapat memahami kondisi, membaca pola, dan menyusun langkah yang lebih tepat.

Peran data mining dalam pengambilan keputusan berkaitan erat dengan konsep *data-driven decision making*, yaitu pengambilan keputusan yang didasarkan pada hasil analisis data. Provost dan Fawcett (2013) menjelaskan bahwa pengambilan keputusan berbasis data tidak sepenuhnya meniadakan intuisi manusia, tetapi menempatkan analisis data sebagai dasar yang lebih kuat dalam menilai suatu keadaan. Dengan demikian, data mining berfungsi sebagai alat bantu yang memperkuat proses berpikir rasional, bukan sebagai pengganti peran manusia dalam mengambil keputusan.

Dalam penelitian ilmiah, data mining membantu peneliti mengambil keputusan berdasarkan pola yang diperoleh dari data. Misalnya, seorang peneliti tidak hanya menyatakan bahwa suatu kelompok data memiliki karakteristik tertentu berdasarkan dugaan, tetapi dapat menunjukkan hasil analisis yang mendukung pernyataan tersebut. Melalui teknik seperti clustering, classification, prediction, dan association rule mining, data mining memberikan dasar empiris dalam menyusun interpretasi. Dasar empiris ini penting agar simpulan penelitian tidak bersifat subjektif dan dapat dipertanggungjawabkan secara akademik.

Secara umum, peran data mining dalam pengambilan keputusan dapat dijelaskan melalui beberapa aspek berikut.

1. Menyediakan Dasar Keputusan Berbasis Data

Data mining membantu menyediakan dasar keputusan yang bersumber dari data aktual. Dalam banyak kasus, pengambil keputusan berhadapan dengan data yang besar, kompleks, dan sulit dibaca secara langsung. Tanpa metode analisis yang tepat, data tersebut hanya menjadi kumpulan informasi yang belum

memberikan arah keputusan. Data mining membantu menelusuri pola dari data tersebut sehingga keputusan dapat disusun berdasarkan informasi yang lebih jelas.

Sebagai contoh, dalam bidang pendidikan, lembaga dapat menggunakan data mining untuk mengetahui kelompok peserta didik berdasarkan kemampuan akademik. Hasil pengelompokan tersebut dapat digunakan sebagai dasar untuk merancang strategi pembinaan, pendampingan, atau evaluasi pembelajaran. Dalam bidang bisnis, data mining dapat membantu perusahaan memahami segmentasi pelanggan sehingga strategi pemasaran dapat diarahkan pada kelompok pelanggan yang tepat.

2. Mengurangi Subjektivitas dalam Pengambilan Keputusan

Keputusan yang hanya didasarkan pada pendapat pribadi berisiko mengandung bias. Bias tersebut dapat muncul karena keterbatasan pengalaman, asumsi yang tidak diuji, atau kecenderungan melihat masalah dari satu sudut pandang saja. Data mining membantu mengurangi subjektivitas tersebut dengan menyajikan temuan berdasarkan proses analisis data.

Namun, perlu ditegaskan bahwa data mining tidak membuat keputusan menjadi sepenuhnya bebas dari kesalahan. Kesalahan tetap dapat terjadi apabila data yang digunakan tidak berkualitas, atribut yang dipilih tidak relevan, atau hasil analisis ditafsirkan secara berlebihan. Oleh karena itu, data mining sebaiknya dipahami sebagai alat bantu yang memperkuat keputusan, bukan sebagai alat yang secara otomatis menghasilkan kebenaran mutlak.

3. Membantu Mengidentifikasi Pola dan Kecenderungan

Salah satu peran penting data mining adalah membantu pengambil keputusan mengenali pola dan kecenderungan yang terdapat dalam data. Pola tersebut dapat berupa kelompok data yang memiliki karakteristik serupa, hubungan antarvariabel, kecenderungan perilaku, atau kemungkinan munculnya suatu keadaan tertentu.

Fayyad, Piatetsky-Shapiro, dan Smyth (1996) menekankan bahwa proses penemuan pengetahuan dari data diarahkan untuk menemukan pola yang valid, baru, bermanfaat, dan dapat dipahami. Dengan prinsip tersebut, pola yang ditemukan melalui data mining harus ditafsirkan secara hati-hati agar benar-benar memiliki nilai bagi pengambilan keputusan. Pola yang menarik secara statistik belum tentu bermanfaat apabila tidak sesuai dengan konteks masalah yang sedang dikaji.

4. Mendukung Penentuan Prioritas

Dalam praktik pengambilan keputusan, tidak semua masalah dapat diselesaikan secara bersamaan. Pengambil keputusan sering kali perlu menentukan prioritas berdasarkan kondisi data. Data mining dapat membantu menentukan kelompok mana yang membutuhkan perhatian lebih besar, faktor apa yang paling dominan, atau area mana yang perlu segera ditangani.

Contohnya, dalam konteks pendidikan, hasil clustering dapat menunjukkan adanya kelompok mahasiswa dengan capaian akademik rendah, sedang, dan tinggi. Informasi ini dapat membantu lembaga pendidikan menentukan program pendampingan yang lebih tepat. Kelompok dengan capaian rendah dapat diberikan pembinaan tambahan, sedangkan kelompok dengan capaian tinggi dapat diarahkan pada program pengembangan prestasi. Dengan demikian, keputusan yang diambil menjadi lebih terarah karena didasarkan pada hasil pemetaan data.

5. Membantu Menyusun Alternatif Strategi

Data mining tidak hanya membantu menjelaskan keadaan, tetapi juga dapat digunakan untuk menyusun alternatif strategi. Hasil analisis data dapat memberikan gambaran mengenai kemungkinan tindakan yang dapat dilakukan. Dalam kerangka CRISP-DM, hasil data mining tidak berhenti pada tahap pemodelan, tetapi dilanjutkan dengan evaluasi dan penerapan hasil. Hal ini menunjukkan bahwa data mining memiliki hubungan langsung

dengan kebutuhan praktis dalam menyelesaikan masalah dan mendukung keputusan.

Sebagai contoh, apabila hasil analisis menunjukkan bahwa pelanggan terbagi menjadi beberapa segmen berdasarkan pola pembelian, maka perusahaan dapat menyusun strategi pemasaran yang berbeda untuk setiap segmen. Jika hasil analisis menunjukkan bahwa peserta didik terbagi ke dalam kelompok kemampuan akademik tertentu, maka lembaga pendidikan dapat merancang strategi pembelajaran yang lebih sesuai dengan kebutuhan masing-masing kelompok.

6. Mendukung Evaluasi Keputusan

Data mining juga berperan dalam mengevaluasi keputusan yang telah diambil. Setelah suatu kebijakan, program, atau strategi diterapkan, data baru dapat dianalisis kembali untuk melihat apakah terjadi perubahan pola. Dengan cara ini, data mining tidak hanya digunakan sebelum keputusan dibuat, tetapi juga setelah keputusan dilaksanakan.

Evaluasi berbasis data penting untuk mengetahui apakah keputusan yang diambil sudah memberikan hasil yang sesuai dengan tujuan. Apabila hasil evaluasi menunjukkan bahwa strategi belum efektif, maka pengambil keputusan dapat melakukan perbaikan. Proses ini menjadikan pengambilan keputusan sebagai kegiatan yang berkelanjutan, bukan tindakan yang berhenti pada satu keputusan saja.

7. Membantu Mengantisipasi Risiko

Dalam beberapa kasus, data mining dapat membantu pengambil keputusan mengenali potensi risiko lebih awal. Teknik prediksi, klasifikasi, dan deteksi anomali dapat digunakan untuk melihat kemungkinan munculnya kondisi tertentu. Misalnya, dalam bidang keuangan, data mining dapat membantu mengenali transaksi yang tidak wajar. Dalam bidang pendidikan, data mining dapat digunakan untuk mengidentifikasi peserta didik yang berisiko mengalami penurunan prestasi. Dalam bidang kesehatan, data

mining dapat membantu mengenali pola risiko berdasarkan data pasien.

Meskipun demikian, hasil data mining yang bersifat prediktif tetap harus dipahami sebagai kemungkinan, bukan kepastian. Model prediksi bekerja berdasarkan pola data sebelumnya, sehingga hasilnya sangat bergantung pada kualitas dan relevansi data yang digunakan. Oleh sebab itu, keputusan akhir tetap perlu mempertimbangkan konteks, kebijakan, dan pertimbangan ahli.

8. Meningkatkan Efektivitas dan Efisiensi Keputusan

Data mining dapat meningkatkan efektivitas keputusan karena membantu pengambil keputusan memahami masalah dengan lebih akurat. Keputusan yang efektif adalah keputusan yang sesuai dengan masalah, sasaran, dan kondisi nyata. Selain itu, data mining juga dapat meningkatkan efisiensi karena proses analisis terhadap data yang besar dapat dilakukan secara lebih sistematis.

Dalam organisasi modern, kemampuan menggunakan data sebagai dasar keputusan menjadi salah satu faktor penting dalam meningkatkan kinerja. Davenport dan Harris (2007) menjelaskan bahwa organisasi yang mampu memanfaatkan analitik secara baik dapat memperoleh keunggulan karena keputusan tidak hanya didasarkan pada pengalaman, tetapi juga pada analisis kuantitatif dan pengelolaan data yang kuat.

Dalam kaitannya dengan algoritma K-Means, peran data mining dalam pengambilan keputusan terutama terlihat pada proses segmentasi atau pengelompokan data. K-Means membantu membagi data ke dalam beberapa cluster berdasarkan kemiripan karakteristik. Hasil cluster tersebut dapat digunakan sebagai dasar untuk memahami perbedaan antar kelompok dan menentukan tindakan yang sesuai untuk setiap kelompok.

Misalnya, apabila K-Means digunakan untuk mengelompokkan data peserta didik berdasarkan nilai akademik, maka hasilnya dapat menunjukkan kelompok peserta didik dengan kemampuan tinggi, sedang, dan rendah. Dari hasil tersebut, pengambil

keputusan dapat menyusun kebijakan pembelajaran yang lebih tepat. Kelompok dengan kemampuan rendah dapat diberikan pendampingan khusus, kelompok sedang dapat diarahkan pada penguatan materi, sedangkan kelompok tinggi dapat diberikan tantangan akademik yang lebih lanjut. Dengan demikian, K-Means tidak hanya menghasilkan pembagian kelompok, tetapi juga memberikan dasar bagi tindakan yang lebih terencana.

Namun, keputusan yang diambil berdasarkan hasil K-Means tetap memerlukan kehati-hatian. Cluster yang terbentuk bergantung pada atribut yang digunakan, jumlah cluster yang ditentukan, serta kualitas data yang dianalisis. Jika atribut yang digunakan tidak relevan, maka hasil cluster dapat menyesatkan. Jika jumlah cluster tidak dipertimbangkan dengan baik, maka interpretasi kelompok dapat menjadi kurang tepat. Oleh karena itu, hasil K-Means perlu dikaji bersama konteks masalah, tujuan penelitian, dan pertimbangan teoritis.

1.7 Contoh Penerapan Data Mining dalam Berbagai Bidang

Data mining dapat diterapkan dalam berbagai bidang karena hampir setiap sektor menghasilkan dan menyimpan data. Data tersebut dapat berasal dari transaksi, aktivitas pengguna, catatan akademik, rekam medis, layanan publik, media digital, maupun sistem administrasi. Apabila data hanya disimpan tanpa dianalisis, maka nilainya menjadi terbatas. Sebaliknya, ketika data diolah menggunakan teknik data mining, data dapat memberikan informasi yang membantu proses pemahaman masalah, penyusunan strategi, dan pengambilan keputusan.

Penerapan data mining tidak selalu memiliki bentuk yang sama pada setiap bidang. Dalam bidang tertentu, data mining digunakan untuk melakukan prediksi. Pada bidang lain, data mining digunakan untuk klasifikasi, segmentasi, deteksi anomali, atau pencarian pola hubungan. Oleh karena itu, penerapan data mining perlu disesuaikan dengan jenis data, tujuan analisis, dan kebutuhan

pengguna hasil analisis. Han, Kamber, dan Pei (2011) menjelaskan bahwa data mining dapat digunakan untuk menemukan pola dari data dalam berbagai bentuk analisis, seperti klasifikasi, prediksi, asosiasi, clustering, dan deteksi data menyimpang.

1. Penerapan Data Mining dalam Bidang Pendidikan

Bidang pendidikan merupakan salah satu bidang yang banyak memanfaatkan data mining. Data pendidikan dapat berupa nilai siswa, kehadiran, latar belakang peserta didik, aktivitas pembelajaran, hasil ujian, data pendaftaran, maupun data kelulusan. Melalui data mining, data tersebut dapat dianalisis untuk menemukan pola yang berkaitan dengan prestasi belajar, keberhasilan akademik, atau kebutuhan pembinaan peserta didik.

Dalam pendidikan, data mining sering dikaitkan dengan istilah *educational data mining*. Romero dan Ventura (2020) menjelaskan bahwa educational data mining berfokus pada pengembangan metode untuk mengeksplorasi data yang berasal dari lingkungan pendidikan. Tujuannya adalah memahami peserta didik, proses pembelajaran, dan lingkungan belajar secara lebih baik. Dengan pendekatan ini, lembaga pendidikan dapat menggunakan data untuk mendukung perencanaan akademik dan peningkatan kualitas pembelajaran.

Contoh penerapan data mining dalam bidang pendidikan antara lain:

- mengelompokkan siswa berdasarkan tingkat prestasi akademik;
- memprediksi potensi kelulusan mahasiswa;
- mengidentifikasi peserta didik yang membutuhkan pendampingan belajar;
- menganalisis hubungan antara kehadiran dan hasil belajar;
- mengelompokkan calon siswa berdasarkan nilai akademik awal;
- mengevaluasi efektivitas program pembelajaran.

Dalam konteks buku ini, penerapan data mining dalam bidang pendidikan dapat dilakukan menggunakan algoritma K-Means. Misalnya, data nilai siswa dapat dikelompokkan ke dalam beberapa cluster, seperti kelompok prestasi tinggi, sedang, dan rendah. Hasil pengelompokan tersebut dapat membantu pihak sekolah atau lembaga pendidikan dalam menyusun strategi pembinaan yang lebih sesuai dengan karakteristik setiap kelompok.

2. Penerapan Data Mining dalam Bidang Bisnis dan Pemasaran

Dalam bidang bisnis, data mining banyak digunakan untuk memahami perilaku pelanggan dan mendukung strategi pemasaran. Perusahaan biasanya memiliki data transaksi, data pelanggan, data pembelian, data kunjungan, dan data respons terhadap promosi. Data tersebut dapat dianalisis untuk mengetahui pola pembelian, segmentasi pelanggan, produk yang sering dibeli bersama, serta kecenderungan pelanggan terhadap produk tertentu.

Provost dan Fawcett (2013) menjelaskan bahwa data mining dan pemikiran analitis berbasis data memiliki peran penting dalam dunia bisnis karena dapat membantu organisasi mengambil keputusan berdasarkan bukti dari data. Dalam konteks ini, data mining tidak hanya digunakan untuk melihat apa yang telah terjadi, tetapi juga membantu perusahaan memperkirakan peluang dan risiko pada masa mendatang.

Contoh penerapan data mining dalam bidang bisnis dan pemasaran antara lain:

- melakukan segmentasi pelanggan berdasarkan perilaku pembelian;
- menganalisis produk yang sering dibeli secara bersamaan;
- memprediksi pelanggan yang berpotensi berhenti berlangganan;
- menentukan target promosi berdasarkan karakteristik pelanggan;

- mengidentifikasi pola penjualan berdasarkan waktu atau wilayah;
- menyusun rekomendasi produk sesuai preferensi pelanggan.

Algoritma K-Means dapat digunakan dalam bidang bisnis untuk membagi pelanggan ke dalam beberapa segmen. Misalnya, pelanggan dapat dikelompokkan berdasarkan frekuensi pembelian, nilai transaksi, dan jenis produk yang dibeli. Hasil cluster tersebut dapat membantu perusahaan menentukan strategi yang berbeda untuk setiap kelompok pelanggan.

3. Penerapan Data Mining dalam Bidang Kesehatan

Bidang kesehatan menghasilkan data yang sangat beragam, seperti data pasien, riwayat penyakit, hasil pemeriksaan, data pengobatan, data kunjungan, dan data layanan kesehatan. Data mining dapat membantu tenaga kesehatan, pengelola rumah sakit, maupun peneliti dalam memahami pola layanan dan kondisi pasien berdasarkan data yang tersedia.

Raghupathi dan Raghupathi (2014) menyatakan bahwa analisis data dalam bidang kesehatan memiliki potensi besar untuk mendukung peningkatan kualitas layanan, efisiensi operasional, dan pengambilan keputusan klinis maupun manajerial. Meskipun demikian, penggunaan data kesehatan harus dilakukan secara hati-hati karena berkaitan dengan kerahasiaan, etika, dan ketepatan interpretasi.

Contoh penerapan data mining dalam bidang kesehatan antara lain:

- mengelompokkan pasien berdasarkan karakteristik penyakit;
- memprediksi risiko penyakit berdasarkan data riwayat kesehatan;
- menganalisis pola kunjungan pasien;
- mengidentifikasi faktor yang berkaitan dengan lama perawatan;
- mendeteksi data tidak wajar dalam klaim layanan kesehatan;
- membantu evaluasi efektivitas layanan rumah sakit.

Dalam penerapan clustering, K-Means dapat digunakan untuk mengelompokkan pasien berdasarkan usia, hasil pemeriksaan, gejala, atau riwayat penyakit tertentu. Namun, hasil pengelompokan tersebut tidak boleh langsung dijadikan dasar diagnosis tanpa pertimbangan tenaga ahli. Data mining dalam kesehatan sebaiknya digunakan sebagai alat bantu analisis, bukan sebagai pengganti keputusan profesional medis.

4. Penerapan Data Mining dalam Bidang Keuangan dan Perbankan

Bidang keuangan dan perbankan memiliki data yang sangat besar, terutama data transaksi, data nasabah, data kredit, data pembayaran, dan data risiko. Data mining dapat digunakan untuk membantu lembaga keuangan dalam mengenali pola transaksi, mengelola risiko, mendeteksi kemungkinan kecurangan, dan memahami karakteristik nasabah.

Salah satu penerapan penting data mining dalam bidang keuangan adalah deteksi anomali. Transaksi yang berbeda secara signifikan dari pola umum dapat diperiksa lebih lanjut untuk mengetahui apakah transaksi tersebut merupakan kesalahan, aktivitas tidak biasa, atau indikasi kecurangan. Bolton dan Hand (2002) menjelaskan bahwa deteksi kecurangan secara statistik banyak berkaitan dengan upaya mengenali pola yang tidak biasa dari data.

Contoh penerapan data mining dalam bidang keuangan dan perbankan antara lain:

- menganalisis kelayakan kredit nasabah;
- memprediksi risiko gagal bayar;
- mendeteksi transaksi yang tidak wajar;
- mengelompokkan nasabah berdasarkan profil keuangan;
- menganalisis pola penggunaan kartu kredit;
- menyusun strategi layanan berdasarkan segmen nasabah.

Dalam konteks clustering, K-Means dapat digunakan untuk mengelompokkan nasabah berdasarkan jumlah transaksi, saldo rata-rata, jenis layanan yang digunakan, atau pola pembayaran.

Hasil pengelompokan tersebut dapat membantu lembaga keuangan memahami karakteristik nasabah dan menyusun layanan yang lebih sesuai.

5. Penerapan Data Mining dalam Bidang Pemerintahan dan Layanan Publik

Pemerintahan dan layanan publik juga dapat memanfaatkan data mining untuk meningkatkan kualitas pelayanan. Data yang dimiliki pemerintah dapat berupa data kependudukan, data sosial ekonomi, data pendidikan, data kesehatan, data bantuan sosial, data pajak, dan data administrasi pelayanan publik. Dengan data mining, pemerintah dapat memahami kondisi masyarakat secara lebih terukur dan menyusun kebijakan yang lebih tepat sasaran.

Contoh penerapan data mining dalam bidang pemerintahan dan layanan publik antara lain:

- mengelompokkan wilayah berdasarkan tingkat kesejahteraan masyarakat;
- menganalisis data penerima bantuan sosial;
- memetakan kebutuhan layanan publik berdasarkan wilayah;
- mengidentifikasi kelompok masyarakat yang membutuhkan program tertentu;
- menganalisis pola pengaduan masyarakat;
- mengevaluasi efektivitas kebijakan berbasis data.

Algoritma K-Means dapat digunakan untuk mengelompokkan masyarakat atau wilayah berdasarkan indikator tertentu, seperti pendapatan, jumlah tanggungan, pendidikan, pekerjaan, atau kondisi tempat tinggal. Hasil clustering dapat membantu pemerintah dalam menyusun kebijakan yang lebih terarah. Namun, penggunaan data mining dalam layanan publik harus memperhatikan keakuratan data dan keadilan interpretasi, karena keputusan yang diambil dapat berdampak langsung pada masyarakat.

6. Penerapan Data Mining dalam Bidang Industri dan Manufaktur

Dalam bidang industri dan manufaktur, data mining digunakan untuk menganalisis data produksi, kualitas produk, penggunaan mesin, persediaan bahan baku, dan distribusi. Tujuannya adalah meningkatkan efisiensi produksi, mengurangi kesalahan, serta membantu perusahaan memahami faktor-faktor yang memengaruhi kualitas dan produktivitas.

Contoh penerapan data mining dalam bidang industri dan manufaktur antara lain:

- memprediksi kerusakan mesin berdasarkan data operasional;
- menganalisis penyebab cacat produk;
- mengelompokkan jenis produk berdasarkan karakteristik produksi;
- mengoptimalkan persediaan bahan baku;
- menganalisis pola permintaan barang;
- mengevaluasi kinerja proses produksi.

Dalam hal ini, data mining dapat membantu perusahaan mengambil keputusan yang lebih cepat dan tepat. Misalnya, apabila data menunjukkan pola kerusakan mesin tertentu, perusahaan dapat melakukan perawatan sebelum kerusakan menjadi lebih serius. Dengan demikian, data mining dapat mendukung efisiensi biaya dan menjaga kelancaran proses produksi.

7. Penerapan Data Mining dalam Bidang Perpustakaan dan Informasi

Bidang perpustakaan dan informasi juga dapat memanfaatkan data mining, terutama untuk memahami perilaku pengguna layanan. Perpustakaan memiliki data peminjaman buku, kunjungan pengguna, kategori koleksi, riwayat pencarian, dan data keanggotaan. Data tersebut dapat dianalisis untuk mengetahui kebutuhan pembaca dan meningkatkan kualitas layanan.

Contoh penerapan data mining dalam bidang perpustakaan dan informasi antara lain:

- menganalisis buku yang paling sering dipinjam;
- mengelompokkan pengguna berdasarkan minat bacaan;
- mengetahui kategori koleksi yang paling dibutuhkan;
- memprediksi kebutuhan pengadaan buku;
- menganalisis pola kunjungan perpustakaan;
- menyusun rekomendasi bacaan bagi pengguna.

K-Means dapat digunakan untuk mengelompokkan pengguna perpustakaan berdasarkan jenis buku yang dipinjam, frekuensi kunjungan, atau kategori bacaan yang diminati. Hasil tersebut dapat membantu perpustakaan mengembangkan layanan yang lebih sesuai dengan kebutuhan pengguna.

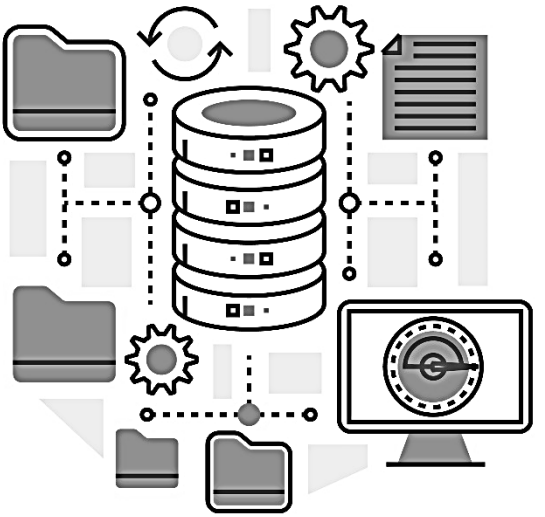
8. Penerapan Data Mining dalam Bidang Media Digital

Perkembangan media digital menghasilkan data yang sangat besar, seperti data kunjungan situs web, interaksi media sosial, komentar pengguna, durasi akses, dan pola konsumsi konten. Data mining dapat digunakan untuk memahami perilaku pengguna media digital dan meningkatkan kualitas konten maupun layanan.

Contoh penerapan data mining dalam bidang media digital antara lain:

- menganalisis pola interaksi pengguna media sosial;
- mengelompokkan pengguna berdasarkan minat konten;
- memprediksi jenis konten yang berpotensi menarik perhatian pembaca;
- menganalisis sentimen komentar pengguna;
- menyusun rekomendasi konten;
- mengevaluasi efektivitas kampanye digital.

Dalam konteks ini, data mining dapat membantu pengelola media memahami jenis konten yang paling diminati oleh pengguna. Namun, analisis data digital harus tetap memperhatikan etika penggunaan data, privasi pengguna, dan keakuratan interpretasi terhadap perilaku digital.



Bab 02

KNOWLEDGE DISCOVERY IN DATABASE DAN PENGOLAHAN DATA

2.1 Pengertian Knowledge Discovery in Database

Knowledge Discovery in Database atau KDD merupakan proses penemuan pengetahuan dari sekumpulan data melalui tahapan yang sistematis. KDD tidak hanya berkaitan dengan penggunaan algoritma, tetapi mencakup seluruh rangkaian kegiatan mulai dari memahami data, memilih data yang relevan, membersihkan data, mentransformasikan data, menerapkan metode data mining, hingga mengevaluasi dan menafsirkan pola yang ditemukan. Dengan demikian, KDD dapat dipahami sebagai kerangka kerja yang lebih luas daripada data mining.

Fayyad, Piatetsky-Shapiro, dan Smyth (1996) mendefinisikan KDD sebagai proses nontrivial untuk mengidentifikasi pola dalam data yang bersifat valid, baru, berpotensi bermanfaat, dan dapat

dipahami. Definisi ini menunjukkan bahwa pengetahuan yang dihasilkan melalui KDD tidak boleh hanya berupa pola yang muncul secara kebetulan, tetapi harus memiliki dasar yang dapat dipertanggungjawabkan. Pola yang ditemukan perlu diuji, ditafsirkan, dan dikaitkan dengan kebutuhan analisis agar benar-benar dapat disebut sebagai pengetahuan.

Dalam konteks tersebut, istilah "pengetahuan" tidak sama dengan data mentah. Data mentah merupakan kumpulan fakta atau catatan yang belum diolah secara mendalam. Setelah data diproses, data dapat berubah menjadi informasi. Informasi kemudian dapat menjadi pengetahuan apabila telah ditafsirkan, dipahami, dan digunakan untuk menjawab masalah tertentu. Oleh karena itu, KDD berfungsi sebagai jembatan antara data yang masih bersifat mentah dengan pengetahuan yang dapat digunakan dalam pengambilan keputusan.

Perlu ditegaskan bahwa KDD dan data mining bukan istilah yang sepenuhnya sama. Data mining merupakan salah satu tahap penting dalam KDD, yaitu tahap ketika algoritma diterapkan untuk menemukan pola dari data. Sementara itu, KDD mencakup proses yang lebih lengkap, termasuk pemilihan data, prapengolahan, transformasi, evaluasi, dan interpretasi hasil. Kekeliruan yang sering terjadi adalah menganggap data mining sebagai keseluruhan proses penemuan pengetahuan. Padahal, secara konseptual, data mining hanya menjadi bagian inti dari proses KDD.

Secara umum, KDD memiliki beberapa ciri utama sebagai berikut.

- Bersifat sistematis
  KDD dilakukan melalui tahapan yang teratur, mulai dari pemahaman data sampai interpretasi hasil. Proses ini tidak dilakukan secara acak, tetapi mengikuti langkah ilmiah agar hasil yang diperoleh dapat dipertanggungjawabkan.
- Berorientasi pada penemuan pengetahuan
  Tujuan akhir KDD bukan hanya menghasilkan angka, grafik, atau keluaran perangkat lunak, tetapi menemukan pengetahuan yang memiliki makna bagi pemecahan masalah.

- Melibatkan data dalam jumlah tertentu
  KDD umumnya digunakan pada kumpulan data yang cukup besar atau kompleks, sehingga pola di dalamnya sulit ditemukan melalui pengamatan manual.
- Menggunakan metode analitis
  Dalam proses KDD, digunakan metode statistik, data mining, pembelajaran mesin, atau teknik komputasi lain yang sesuai dengan karakteristik data.
- Memerlukan interpretasi manusia
  Hasil KDD tidak dapat langsung dianggap sebagai simpulan akhir. Peneliti tetap perlu menafsirkan hasil berdasarkan teori, konteks masalah, dan tujuan penelitian.

Dalam penerapannya, KDD biasanya dimulai dari pemilihan data yang sesuai dengan kebutuhan analisis. Tidak semua data yang tersedia harus digunakan. Peneliti perlu menentukan data mana yang relevan dengan tujuan penelitian. Setelah itu, data diperiksa untuk menemukan kemungkinan kesalahan, data kosong, duplikasi, atau ketidakkonsistenan. Tahap ini penting karena kualitas data sangat memengaruhi hasil analisis. Data yang buruk dapat menghasilkan pola yang keliru, meskipun algoritma yang digunakan sudah tepat.

Setelah data dibersihkan, data kemudian ditransformasikan ke dalam bentuk yang sesuai untuk dianalisis. Transformasi dapat berupa pengubahan format data, normalisasi nilai, pengkodean variabel, atau pemilihan atribut tertentu. Dalam penelitian yang menggunakan algoritma K-Means, tahap transformasi menjadi sangat penting karena K-Means bekerja berdasarkan perhitungan jarak. Oleh sebab itu, atribut yang digunakan perlu disiapkan secara hati-hati agar hasil pengelompokan tidak bias terhadap skala data tertentu.

Tahap berikutnya adalah data mining, yaitu penerapan metode atau algoritma untuk menemukan pola dari data. Pada tahap ini, peneliti dapat menggunakan teknik yang sesuai dengan tujuan analisis, seperti classification, clustering, association, prediction,

atau outlier analysis. Dalam buku ini, teknik yang menjadi perhatian utama adalah clustering dengan algoritma K-Means. K-Means digunakan untuk menemukan kelompok data berdasarkan kemiripan karakteristik antarobjek.

Namun, proses KDD tidak berhenti setelah algoritma dijalankan. Hasil yang diperoleh masih perlu dievaluasi dan diinterpretasikan. Evaluasi dilakukan untuk melihat apakah pola yang ditemukan benar-benar bermakna, relevan, dan sesuai dengan tujuan penelitian. Interpretasi dilakukan untuk menjelaskan arti dari pola tersebut dalam konteks masalah yang sedang dikaji. Tanpa evaluasi dan interpretasi, hasil data mining hanya menjadi keluaran teknis yang belum tentu memiliki nilai ilmiah.

Dalam penelitian berbasis K-Means, KDD dapat digunakan sebagai kerangka kerja agar analisis tidak hanya berpusat pada proses perhitungan cluster. Peneliti perlu memahami data yang digunakan, menentukan atribut yang relevan, membersihkan data dari kesalahan, menyiapkan data dalam format numerik, menjalankan algoritma K-Means, kemudian menafsirkan hasil cluster secara ilmiah. Dengan demikian, KDD membantu memastikan bahwa penerapan K-Means dilakukan secara terarah dan tidak hanya bersifat mekanis.

Sebagai contoh, apabila seorang peneliti ingin mengelompokkan data siswa berdasarkan nilai akademik, maka proses KDD dapat dimulai dari pemilihan data nilai yang relevan. Data tersebut kemudian diperiksa kelengkapannya, dibersihkan dari kesalahan input, dan ditransformasikan agar siap dianalisis. Setelah itu, algoritma K-Means digunakan untuk membentuk cluster. Hasil cluster kemudian dievaluasi dan ditafsirkan, misalnya menjadi kelompok siswa dengan kemampuan tinggi, sedang, dan rendah. Interpretasi inilah yang menjadikan hasil analisis memiliki nilai pengetahuan.

2.2 Tahapan Knowledge Discovery in Database

Tahapan *Knowledge Discovery in Database* atau KDD merupakan rangkaian proses yang digunakan untuk mengubah data menjadi pengetahuan yang dapat dipahami dan dimanfaatkan. KDD tidak dilakukan secara langsung dengan menjalankan algoritma, tetapi melalui beberapa tahap yang saling berhubungan. Setiap tahap memiliki fungsi tertentu, mulai dari menentukan data yang relevan, memperbaiki kualitas data, menyiapkan bentuk data, menerapkan metode data mining, hingga mengevaluasi pola yang ditemukan. Fayyad, Piatetsky-Shapiro, dan Smyth menjelaskan bahwa proses KDD mencakup penggunaan basis data melalui seleksi, praproses, subsampling, transformasi, penerapan metode data mining, serta evaluasi hasil untuk menentukan pola yang layak dianggap sebagai pengetahuan.

Dalam buku ini, tahapan KDD disajikan dalam lima tahap utama, yaitu data selection, preprocessing, transformation, data mining, serta interpretation/evaluation. Pembagian ini digunakan karena lebih sederhana, sistematis, dan mudah diterapkan dalam penelitian berbasis algoritma K-Means. Meskipun beberapa literatur menambahkan tahapan lain, seperti pemahaman domain, integrasi data, reduksi data, atau presentasi pengetahuan, inti proses KDD tetap mengarah pada alur pengolahan data dari data mentah menuju pengetahuan yang bermakna.

1. Data Selection

Tahap pertama dalam KDD adalah data selection atau seleksi data. Tahap ini dilakukan untuk memilih data yang relevan dengan tujuan analisis. Dalam penelitian, tidak semua data yang tersedia harus digunakan. Data yang tidak berhubungan dengan tujuan penelitian dapat mengganggu proses analisis dan menyebabkan hasil yang kurang tepat. Oleh karena itu, peneliti perlu menentukan data mana yang akan digunakan sebagai dasar proses data mining.

Seleksi data bertujuan untuk membentuk *target data*, yaitu kumpulan data yang telah dipilih sesuai dengan kebutuhan analisis.

Misalnya, apabila penelitian bertujuan mengelompokkan siswa berdasarkan kemampuan akademik, maka data yang dipilih dapat berupa nilai mata pelajaran, nilai ujian, atau indikator akademik lain yang relevan. Data seperti alamat lengkap, nomor induk, atau informasi administratif tertentu tidak selalu diperlukan apabila tidak memiliki hubungan langsung dengan tujuan clustering.

Pada tahap ini, peneliti perlu memperhatikan beberapa hal berikut.

- Menentukan tujuan analisis secara jelas.
- Memilih atribut yang sesuai dengan tujuan penelitian.
- Menghindari penggunaan data yang tidak relevan.
- Memastikan data yang dipilih tersedia dan dapat diolah.
- Menyesuaikan data dengan metode yang akan digunakan.

Dalam konteks algoritma K-Means, tahap seleksi data sangat penting karena algoritma ini bekerja berdasarkan kemiripan karakteristik antarobjek. Atribut yang dipilih akan memengaruhi hasil cluster. Jika atribut yang digunakan tidak relevan, maka cluster yang terbentuk dapat menghasilkan interpretasi yang keliru.

2. Preprocessing

Tahap kedua adalah preprocessing atau praproses data. Tahap ini bertujuan untuk memperbaiki kualitas data sebelum dilakukan analisis lebih lanjut. Data yang diperoleh dari sumber asli sering kali belum siap digunakan. Data dapat mengandung nilai kosong, data ganda, kesalahan penulisan, format yang tidak seragam, atau nilai yang tidak wajar. Apabila kondisi tersebut dibiarkan, maka hasil data mining dapat menjadi tidak akurat.

Preprocessing dilakukan untuk memastikan bahwa data berada dalam kondisi layak analisis. Tahap ini menjadi salah satu bagian penting dalam KDD karena kualitas hasil data mining sangat bergantung pada kualitas data yang digunakan. Data yang tidak bersih dapat menyebabkan algoritma membaca pola yang sebenarnya berasal dari kesalahan data, bukan dari fenomena yang sesungguhnya.

Beberapa kegiatan yang umum dilakukan pada tahap preprocessing antara lain:

- Membersihkan data kosong, yaitu menangani data yang tidak memiliki nilai pada atribut tertentu.
- Menghapus data duplikat, yaitu menghilangkan data yang tercatat lebih dari satu kali.
- Memperbaiki kesalahan input, seperti salah ketik, perbedaan format, atau kesalahan penulisan angka.
- Menangani data pencilan, yaitu memeriksa data yang nilainya jauh berbeda dari pola umum.
- Menyamakan format data, misalnya format tanggal, kode kategori, atau satuan nilai.

Dalam penelitian berbasis K-Means, preprocessing harus dilakukan secara cermat. Misalnya, jika terdapat nilai yang kosong pada atribut akademik, peneliti perlu menentukan apakah data tersebut akan dihapus, diganti dengan nilai tertentu, atau diperbaiki berdasarkan sumber data asli. Keputusan tersebut harus dijelaskan secara metodologis agar proses analisis dapat dipertanggungjawabkan.

3. Transformation

Tahap ketiga adalah transformation atau transformasi data. Transformasi data merupakan proses mengubah data ke dalam bentuk yang sesuai untuk digunakan dalam proses data mining. Data yang sudah dipilih dan dibersihkan belum tentu dapat langsung dianalisis. Beberapa metode data mining membutuhkan bentuk data tertentu agar algoritma dapat bekerja dengan baik.

Dalam transformasi data, peneliti dapat melakukan pengubahan skala, pengkodean, pengelompokan nilai, normalisasi, atau pembentukan atribut baru. Tahap ini sangat penting dalam algoritma K-Means karena K-Means menggunakan perhitungan jarak. Apabila terdapat perbedaan skala yang terlalu besar antaratribut, maka atribut dengan nilai besar dapat mendominasi proses pengelompokan.

Sebagai contoh, apabila satu atribut memiliki rentang nilai 0–100, sedangkan atribut lain memiliki rentang 0–10.000, maka perhitungan jarak dapat lebih banyak dipengaruhi oleh atribut dengan rentang nilai lebih besar. Untuk menghindari hal tersebut, peneliti dapat melakukan normalisasi atau standardisasi data. Dengan cara ini, setiap atribut memiliki peluang yang lebih seimbang dalam membentuk cluster.

Bentuk transformasi data yang sering digunakan antara lain:

- Normalisasi data, yaitu mengubah nilai ke dalam rentang tertentu.
- Standardisasi data, yaitu menyesuaikan nilai berdasarkan ukuran statistik tertentu.
- Pengkodean data kategorik, yaitu mengubah kategori menjadi bentuk numerik.
- Agregasi data, yaitu menggabungkan beberapa data menjadi ringkasan tertentu.
- Reduksi atribut, yaitu mengurangi atribut yang tidak diperlukan.

Dalam konteks buku ini, transformasi data perlu mendapat perhatian khusus karena K-Means lebih sesuai digunakan pada data numerik. Apabila data yang tersedia berbentuk kategorik, maka data tersebut perlu dikodekan terlebih dahulu secara tepat. Namun, pengkodean tidak boleh dilakukan sembarangan karena dapat memengaruhi makna data dan hasil perhitungan jarak.

4. Data Mining

Tahap keempat adalah data mining, yaitu proses penerapan metode atau algoritma untuk menemukan pola dari data. Pada tahap ini, data yang telah melalui seleksi, preprocessing, dan transformasi dianalisis menggunakan teknik tertentu. Teknik yang digunakan bergantung pada tujuan penelitian, misalnya classification, prediction, association, clustering, atau outlier analysis.

Dalam buku ini, tahap data mining diarahkan pada teknik clustering menggunakan algoritma K-Means. K-Means digunakan untuk membagi data ke dalam beberapa kelompok berdasarkan kemiripan karakteristik. Proses ini dilakukan dengan menentukan jumlah cluster, memilih centroid awal, menghitung jarak setiap data terhadap centroid, membentuk kelompok, memperbarui centroid, dan mengulangi proses sampai hasil cluster stabil.

Tahap data mining perlu dilakukan secara hati-hati karena hasil algoritma sangat dipengaruhi oleh keputusan peneliti pada tahap sebelumnya. Data yang tidak relevan, tidak bersih, atau tidak ditransformasikan dengan baik dapat menghasilkan pola yang kurang tepat. Oleh karena itu, data mining tidak dapat dipisahkan dari tahapan KDD yang mendahuluinya.

Dalam penerapan K-Means, beberapa hal yang perlu diperhatikan pada tahap data mining adalah:

- menentukan jumlah cluster yang akan dibentuk;
- memastikan atribut yang digunakan sesuai dengan tujuan analisis;
- menggunakan data numerik atau data yang telah ditransformasikan;
- memperhatikan pengaruh centroid awal terhadap hasil cluster;
- mengevaluasi apakah hasil pengelompokan dapat ditafsirkan secara logis.

Data mining menghasilkan keluaran berupa pola, model, aturan, nilai prediksi, atau cluster. Namun, keluaran tersebut belum otomatis menjadi pengetahuan. Hasil data mining masih perlu dievaluasi dan ditafsirkan agar memiliki makna ilmiah.

5. Interpretation dan Evaluation

Tahap terakhir dalam KDD adalah interpretation dan evaluation. Tahap ini dilakukan untuk menilai apakah pola yang ditemukan benar-benar bermakna, relevan, dan sesuai dengan tujuan penelitian. Evaluasi diperlukan untuk memastikan bahwa hasil

data mining tidak hanya benar secara teknis, tetapi juga dapat dipahami dan dimanfaatkan.

Interpretasi merupakan proses menjelaskan makna hasil data mining. Dalam clustering menggunakan K-Means, interpretasi dilakukan dengan membaca karakteristik setiap cluster. Misalnya, apabila data siswa dikelompokkan menjadi tiga cluster berdasarkan nilai akademik, maka peneliti perlu menjelaskan perbedaan antara cluster pertama, kedua, dan ketiga. Peneliti tidak cukup hanya menyebutkan jumlah anggota cluster, tetapi juga perlu menjelaskan karakteristik masing-masing kelompok.

Evaluasi hasil KDD dapat mencakup beberapa pertanyaan berikut.

- Apakah pola yang ditemukan sesuai dengan tujuan penelitian?
- Apakah hasil analisis dapat dijelaskan secara logis?
- Apakah cluster atau model yang terbentuk memiliki makna praktis?
- Apakah hasil analisis didukung oleh kualitas data yang memadai?
- Apakah hasil tersebut dapat digunakan sebagai dasar pengambilan keputusan?

Dalam penelitian ilmiah, tahap evaluasi dan interpretasi menjadi sangat penting karena hasil data mining harus dapat dipertanggungjawabkan. Pola yang muncul dari algoritma tidak boleh langsung diterima tanpa pemeriksaan. Peneliti harus menilai apakah pola tersebut relevan dengan teori, konteks data, dan tujuan penelitian.

2.3 Seleksi Data

Seleksi data merupakan tahap awal yang penting dalam proses *Knowledge Discovery in Database* karena menentukan data mana yang akan digunakan dalam proses analisis. Pada tahap ini, peneliti memilih data yang relevan dengan tujuan penelitian dan mengabaikan data yang tidak berhubungan langsung dengan kebutuhan analisis. Seleksi data tidak boleh dilakukan secara

sembarangan, sebab data yang dipilih akan memengaruhi seluruh proses berikutnya, mulai dari praproses data, transformasi data, penerapan algoritma, hingga interpretasi hasil.

Dalam kerangka KDD, seleksi data menjadi bagian dari proses untuk membentuk *target data*, yaitu data terpilih yang akan dianalisis lebih lanjut. Fayyad, Piatetsky-Shapiro, dan Smyth menjelaskan bahwa proses KDD melibatkan pemilihan, praproses, transformasi, penerapan metode data mining, dan interpretasi pola untuk menemukan pengetahuan dari data. Artinya, sebelum algoritma dijalankan, peneliti perlu memastikan bahwa data yang masuk ke tahap analisis benar-benar sesuai dengan masalah yang dikaji.

Seleksi data berfungsi untuk membatasi ruang lingkup analisis agar proses data mining lebih terarah. Dalam penelitian, data yang tersedia sering kali sangat banyak dan tidak seluruhnya relevan. Apabila semua data digunakan tanpa penyaringan, hasil analisis dapat menjadi bias, tidak fokus, atau sulit ditafsirkan. Oleh karena itu, seleksi data dilakukan untuk memilih objek, atribut, dan periode data yang benar-benar memiliki hubungan dengan tujuan penelitian.

Pada penelitian berbasis algoritma K-Means, seleksi data memiliki peran yang sangat penting karena K-Means bekerja dengan cara menghitung kedekatan antarobjek berdasarkan atribut yang digunakan. Apabila atribut yang dipilih tidak tepat, maka hasil cluster yang terbentuk juga dapat menjadi kurang bermakna. Misalnya, jika penelitian bertujuan mengelompokkan siswa berdasarkan prestasi akademik, maka atribut yang relevan dapat berupa nilai mata pelajaran, nilai ujian, atau rata-rata nilai. Sebaliknya, atribut seperti nomor induk, nama lengkap, atau alamat lengkap tidak perlu digunakan sebagai dasar clustering karena tidak menggambarkan kemiripan akademik.

Secara umum, seleksi data dilakukan untuk mencapai beberapa tujuan berikut.

- Menentukan data yang relevan dengan masalah penelitian
  Data yang dipilih harus memiliki hubungan langsung dengan rumusan masalah dan tujuan penelitian. Relevansi data menjadi dasar agar hasil analisis dapat menjawab persoalan yang dikaji.
- Mengurangi data yang tidak diperlukan
  Tidak semua data yang tersedia harus digunakan. Data yang tidak mendukung tujuan analisis dapat dihilangkan agar proses pengolahan lebih fokus dan efisien.
- Menentukan atribut yang akan dianalisis
  Atribut merupakan variabel atau karakteristik yang melekat pada objek data. Dalam K-Means, atribut yang dipilih akan menjadi dasar perhitungan jarak antarobjek.
- Membentuk dataset yang siap diproses
  Seleksi data menghasilkan kumpulan data awal yang akan diproses pada tahap berikutnya, yaitu pembersihan, transformasi, dan data mining.
- Mencegah kesalahan interpretasi hasil
  Pemilihan data yang tidak tepat dapat menghasilkan pola yang tampak benar secara teknis, tetapi keliru secara makna. Oleh karena itu, seleksi data membantu menjaga kualitas interpretasi hasil penelitian.

Dalam pelaksanaannya, seleksi data perlu memperhatikan objek penelitian. Objek penelitian adalah unit yang akan dianalisis, misalnya siswa, mahasiswa, pelanggan, pasien, wilayah, transaksi, atau penerima bantuan. Penentuan objek harus jelas karena objek inilah yang akan dikelompokkan atau dianalisis. Pada algoritma K-Means, setiap objek biasanya direpresentasikan dalam satu baris data, sedangkan setiap atribut ditampilkan dalam kolom data.

Selain objek penelitian, peneliti juga perlu menentukan atribut yang digunakan. Atribut yang dipilih harus mencerminkan karakteristik yang ingin dianalisis. Dalam penelitian pendidikan,

atribut dapat berupa nilai akademik, kehadiran, jumlah tugas, atau hasil ujian. Dalam bidang bisnis, atribut dapat berupa jumlah transaksi, frekuensi pembelian, total belanja, atau jenis produk. Dalam bidang sosial, atribut dapat berupa penghasilan, jumlah tanggungan, kondisi tempat tinggal, atau tingkat pendidikan.

Pemilihan atribut harus dilakukan secara hati-hati. Atribut yang terlalu sedikit dapat membuat hasil analisis kurang menggambarkan keadaan yang sebenarnya. Sebaliknya, atribut yang terlalu banyak tetapi tidak relevan dapat mengganggu proses clustering dan menyulitkan interpretasi. Dalam hal ini, peneliti perlu menyeimbangkan antara kelengkapan data dan relevansi data. Data yang baik bukan hanya banyak, tetapi juga sesuai dengan tujuan analisis.

Dalam kaitannya dengan K-Means, kriteria "sesuai metode" perlu mendapat perhatian khusus. K-Means lebih tepat digunakan pada data numerik karena algoritma ini menghitung jarak antara data dan centroid. Apabila data yang digunakan berbentuk kategorik, peneliti perlu melakukan transformasi terlebih dahulu. Namun, transformasi data kategorik harus dilakukan secara hati-hati agar tidak menimbulkan makna angka yang keliru. Misalnya, pengkodean kategori pekerjaan menjadi angka 1, 2, dan 3 tidak boleh langsung dianggap memiliki jarak matematis yang sama, kecuali peneliti memiliki dasar yang jelas untuk memperlakukannya demikian.

Seleksi data juga berkaitan dengan penentuan periode data. Dalam beberapa penelitian, data yang digunakan berasal dari periode tertentu, misalnya tahun akademik tertentu, bulan tertentu, atau periode transaksi tertentu. Penentuan periode harus dijelaskan agar pembaca memahami batasan data yang dianalisis. Apabila penelitian menggunakan data siswa tahun ajaran tertentu, maka hasil analisis tidak boleh digeneralisasikan secara berlebihan untuk semua tahun ajaran tanpa dasar yang memadai.

Selain periode, sumber data juga harus dijelaskan dengan jelas. Data dapat berasal dari dokumen akademik, database sekolah,

laporan transaksi, rekam administrasi, kuesioner, sistem informasi, atau sumber lain yang relevan. Kejelasan sumber data penting untuk menunjukkan bahwa data yang digunakan dapat ditelusuri dan dipertanggungjawabkan. Dalam penelitian ilmiah, data yang tidak jelas sumbernya dapat melemahkan validitas analisis.

Tahap seleksi data dapat dilakukan melalui langkah-langkah berikut.

- Menentukan tujuan analisis
  Peneliti terlebih dahulu menetapkan tujuan analisis. Misalnya, tujuan penelitian adalah mengelompokkan siswa berdasarkan prestasi akademik menggunakan K-Means.
- Menentukan objek data
  Peneliti menetapkan unit data yang akan dianalisis. Objek data dapat berupa siswa, mahasiswa, pelanggan, wilayah, atau objek lain sesuai konteks penelitian.
- Menentukan atribut yang relevan
  Peneliti memilih atribut yang sesuai dengan tujuan analisis. Untuk pengelompokan prestasi akademik, atribut yang dipilih dapat berupa nilai mata pelajaran atau rata-rata nilai.
- Menentukan sumber data
  Peneliti menjelaskan asal data yang digunakan. Sumber data dapat berupa data primer atau data sekunder, bergantung pada desain penelitian.
- Menentukan batasan data
  Peneliti menetapkan batasan data, seperti periode waktu, jumlah objek, wilayah, atau kategori tertentu yang digunakan dalam penelitian.
- Menghapus data yang tidak relevan
  Data atau atribut yang tidak mendukung tujuan analisis dikeluarkan dari dataset agar proses analisis lebih fokus.
- Menyusun dataset awal
  Data yang telah dipilih disusun menjadi dataset awal untuk diproses pada tahap berikutnya, yaitu preprocessing.

Sebagai contoh, apabila penelitian membahas pengelompokan siswa menggunakan algoritma K-Means berdasarkan nilai akademik, maka peneliti dapat memilih atribut berupa nilai Matematika, Bahasa Indonesia, Bahasa Inggris, dan Ilmu Pengetahuan Alam. Atribut seperti nama siswa tetap dapat digunakan sebagai identitas dalam tabel, tetapi tidak digunakan dalam perhitungan clustering. Hal ini penting karena nama tidak memiliki nilai numerik yang menggambarkan kemiripan akademik.

Dalam praktik data mining, seleksi data juga harus mempertimbangkan kemungkinan adanya atribut yang saling berulang atau terlalu mirip. Misalnya, jika data sudah memiliki nilai rata-rata, sementara semua nilai mata pelajaran juga digunakan, peneliti perlu mempertimbangkan apakah rata-rata tersebut perlu dimasukkan atau cukup digunakan sebagai informasi tambahan. Penggunaan atribut yang terlalu mirip dapat membuat bobot informasi tertentu menjadi terlalu dominan dalam analisis.

Model CRISP-DM juga menempatkan pemahaman dan persiapan data sebagai bagian penting sebelum pemodelan. Dalam kerangka tersebut, data perlu dipahami, dipilih, dibersihkan, dibentuk, dan disiapkan sebelum masuk ke tahap pemodelan. Hal ini sejalan dengan prinsip KDD bahwa kualitas hasil data mining sangat bergantung pada kesiapan data yang digunakan.

Kesalahan dalam seleksi data dapat menimbulkan dampak serius terhadap hasil penelitian. Beberapa kesalahan yang perlu dihindari antara lain:

- memilih atribut yang tidak sesuai dengan tujuan penelitian;
- menggunakan data identitas sebagai atribut analisis;
- memasukkan data yang tidak memiliki hubungan dengan masalah penelitian;
- tidak menjelaskan sumber dan periode data;
- menggunakan data yang terlalu sedikit tanpa alasan metodologis;
- mengabaikan karakteristik algoritma yang digunakan;

- tidak membedakan antara data untuk identifikasi dan data untuk perhitungan.

Dalam algoritma K-Means, kesalahan pemilihan atribut dapat menyebabkan cluster yang terbentuk tidak mencerminkan kelompok yang sebenarnya diinginkan. Misalnya, jika tujuan penelitian adalah mengelompokkan kemampuan akademik, tetapi atribut yang digunakan mencampurkan nilai akademik dengan data administratif yang tidak relevan, maka hasil cluster dapat menjadi tidak valid secara konseptual. Oleh karena itu, peneliti harus mampu menjelaskan mengapa atribut tertentu digunakan dan mengapa atribut lain tidak digunakan.

Seleksi data juga perlu memperhatikan aspek etika. Data yang digunakan dalam penelitian harus diperlakukan secara bertanggung jawab, terutama apabila berkaitan dengan identitas individu. Informasi pribadi seperti nama lengkap, alamat, nomor identitas, atau data sensitif lainnya sebaiknya tidak digunakan dalam perhitungan jika tidak diperlukan. Dalam pelaporan hasil penelitian, identitas individu dapat disamarkan atau diganti dengan kode agar kerahasiaan data tetap terjaga.

Dalam penelitian ilmiah, penjelasan tentang seleksi data biasanya dimasukkan dalam bagian metode penelitian. Peneliti perlu menjelaskan jenis data, sumber data, jumlah data, atribut yang digunakan, serta alasan pemilihan atribut. Penjelasan ini penting agar pembaca dapat menilai apakah data yang digunakan sudah sesuai dengan tujuan penelitian. Semakin jelas proses seleksi data dijelaskan, semakin mudah pembaca memahami dasar analisis yang dilakukan.

Dengan demikian, seleksi data merupakan tahap penting dalam KDD karena menentukan kualitas dan arah analisis data mining. Pada tahap ini, peneliti memilih objek dan atribut yang relevan, menentukan batasan data, memastikan sumber data, serta menyiapkan dataset awal untuk diproses lebih lanjut. Dalam penerapan K-Means, seleksi data menjadi sangat penting karena atribut yang dipilih akan menentukan hasil pengelompokan. Oleh

sebab itu, seleksi data harus dilakukan secara cermat, logis, dan dapat dipertanggungjawabkan secara ilmiah.

2.4 Pembersihan Data

Pembersihan data merupakan tahap penting dalam proses *Knowledge Discovery in Database* yang bertujuan untuk memastikan bahwa data yang akan dianalisis berada dalam kondisi layak, konsisten, dan dapat dipertanggungjawabkan. Data yang diperoleh dari sumber asli sering kali belum siap digunakan karena masih mengandung kesalahan, kekosongan nilai, duplikasi, ketidaksesuaian format, atau data yang tidak wajar. Apabila data seperti ini langsung digunakan dalam proses data mining, maka pola yang dihasilkan dapat menjadi keliru dan berpotensi menyesatkan interpretasi penelitian.

Dalam proses KDD, pembersihan data berada setelah tahap seleksi data. Setelah peneliti menentukan data dan atribut yang relevan, langkah berikutnya adalah memeriksa kualitas data tersebut. Fayyad, Piatetsky-Shapiro, dan Smyth menempatkan tahap praproses sebagai bagian penting dalam KDD sebelum proses data mining dilakukan. Hal ini menunjukkan bahwa kualitas hasil data mining sangat dipengaruhi oleh kualitas data yang masuk ke dalam proses analisis.

Pembersihan data tidak hanya berarti menghapus data yang salah, tetapi juga mencakup proses memeriksa, memperbaiki, menyesuaikan, dan mendokumentasikan kondisi data. Rahm dan Do menjelaskan bahwa masalah kualitas data dapat muncul dalam berbagai bentuk, seperti kesalahan pada sumber data tunggal, ketidaksesuaian antar sumber data, duplikasi, dan perbedaan struktur data. Oleh karena itu, pembersihan data harus dilakukan secara teliti, terutama apabila data berasal dari beberapa sumber yang berbeda.

Dalam penelitian berbasis algoritma K-Means, pembersihan data memiliki kedudukan yang sangat penting karena algoritma ini bekerja berdasarkan perhitungan jarak antar data. Kesalahan kecil

pada nilai numerik dapat memengaruhi posisi data terhadap centroid dan akhirnya memengaruhi pembentukan cluster. Dengan demikian, data yang tidak bersih dapat menyebabkan hasil pengelompokan tidak mencerminkan kondisi yang sebenarnya.

Beberapa masalah data yang umum ditemukan dalam tahap pembersihan data antara lain sebagai berikut.

- Data kosong atau *missing value*
  Data kosong terjadi ketika suatu atribut tidak memiliki nilai. Misalnya, terdapat siswa yang tidak memiliki nilai pada salah satu mata pelajaran, pelanggan yang tidak memiliki data usia, atau responden yang tidak mengisi salah satu item kuesioner. Data kosong perlu ditangani dengan hati-hati karena dapat mengganggu proses perhitungan.
- Data duplikat
  Data duplikat terjadi ketika satu objek tercatat lebih dari satu kali. Dalam penelitian, duplikasi dapat menyebabkan satu objek memiliki pengaruh berlebihan terhadap hasil analisis. Pada K-Means, data yang berulang dapat memengaruhi posisi centroid karena data tersebut dihitung lebih dari satu kali.
- Kesalahan penulisan atau input data
  Kesalahan input dapat berupa salah ketik, kesalahan angka, perbedaan format penulisan, atau penggunaan satuan yang tidak konsisten. Misalnya, nilai 85 ditulis menjadi 850, tanggal ditulis dengan format berbeda, atau kategori yang sama ditulis dengan ejaan yang tidak seragam.
- Ketidakkonsistenan format data
  Ketidakkonsistenan format terjadi ketika data yang seharusnya memiliki bentuk seragam justru ditulis dalam format berbeda. Contohnya adalah penulisan jenis kelamin sebagai “L”, “Laki-laki”, “Pria”, atau “Male”. Ketidakkonsistenan seperti ini perlu diseragamkan sebelum data digunakan.
- Data pencilan atau *outlier*
  Data pencilan adalah data yang nilainya jauh berbeda dari pola umum. Dalam beberapa kasus, outlier dapat terjadi karena kesalahan input. Namun, dalam kasus lain, outlier dapat

menjadi informasi penting yang memang menggambarkan kondisi khusus. Oleh karena itu, outlier tidak boleh langsung dihapus tanpa pemeriksaan.

- Data tidak relevan
  Data tidak relevan merupakan data yang tidak mendukung tujuan analisis. Meskipun data tersebut tidak selalu salah, keberadaannya dapat mengganggu fokus analisis apabila dimasukkan ke dalam proses data mining.

Tahap pembersihan data perlu dilakukan dengan prinsip kehati-hatian. Peneliti tidak boleh mengubah data hanya agar hasil analisis terlihat lebih baik. Setiap perubahan, penghapusan, atau penyesuaian data harus memiliki alasan metodologis yang jelas. Dalam naskah ilmiah, proses pembersihan data sebaiknya dijelaskan agar pembaca mengetahui bagaimana data disiapkan sebelum dianalisis.

Langkah-langkah pembersihan data dapat dilakukan sebagai berikut.

1. Memeriksa Kelengkapan Data

Langkah pertama adalah memeriksa apakah setiap atribut memiliki nilai yang lengkap. Pemeriksaan ini dapat dilakukan dengan melihat jumlah data kosong pada setiap kolom. Apabila ditemukan nilai kosong, peneliti perlu menentukan tindakan yang sesuai.

Beberapa cara menangani data kosong antara lain:

- menghapus baris data apabila jumlah data kosong sedikit dan tidak memengaruhi keseluruhan dataset;
- mengisi nilai kosong menggunakan rata-rata, median, atau modus;
- mengisi nilai berdasarkan sumber data asli;
- menggunakan estimasi tertentu apabila memiliki dasar yang kuat;
- tidak menggunakan atribut tersebut apabila terlalu banyak nilai yang kosong.

Dalam penelitian K-Means, pengisian nilai kosong harus dilakukan dengan hati-hati. Misalnya, penggantian nilai kosong dengan rata-rata dapat mengubah karakteristik data. Oleh sebab itu, peneliti perlu menjelaskan alasan pemilihan metode penanganan data kosong.

2. Menghapus atau Menggabungkan Data Duplikat

Data duplikat perlu diperiksa karena dapat menyebabkan hasil analisis menjadi tidak seimbang. Apabila satu objek tercatat dua kali, maka objek tersebut akan memiliki bobot lebih besar dalam proses perhitungan. Pada algoritma K-Means, kondisi ini dapat memengaruhi pembentukan centroid dan hasil cluster.

Namun, tidak semua data yang tampak sama dapat langsung dianggap duplikat. Peneliti perlu memeriksa identitas data secara menyeluruh. Misalnya, dua siswa dapat memiliki nama yang sama, tetapi nomor induk berbeda. Dalam kasus seperti ini, data tersebut bukan duplikat. Oleh karena itu, pemeriksaan duplikasi sebaiknya dilakukan berdasarkan identitas unik, seperti nomor induk, kode responden, atau kode transaksi.

3. Menyeragamkan Format Data

Format data yang tidak seragam dapat menghambat proses analisis. Penyeragaman format perlu dilakukan agar data dapat dibaca dan diproses secara konsisten. Dalam data numerik, peneliti perlu memastikan bahwa satuan dan skala nilai telah sesuai. Dalam data kategorik, peneliti perlu menyeragamkan penulisan kategori.

Penyeragaman format penting agar data tidak terbaca sebagai kategori yang berbeda padahal memiliki makna yang sama. Dalam data mining, perbedaan kecil dalam penulisan dapat memengaruhi hasil analisis, terutama apabila data digunakan dalam proses transformasi atau pengkodean.

4. Memeriksa Kesalahan Nilai

Kesalahan nilai dapat terjadi karena kekeliruan input atau kesalahan pengolahan sebelumnya. Misalnya, nilai ujian yang

seharusnya berada pada rentang 0–100 tercatat sebagai 120. Data seperti ini harus diperiksa kembali karena berada di luar batas yang wajar.

Pemeriksaan kesalahan nilai dapat dilakukan dengan menetapkan batas logis untuk setiap atribut. Contohnya:

- nilai akademik berada pada rentang 0–100;
- usia siswa berada pada rentang yang sesuai dengan jenjang pendidikan;
- jumlah kehadiran tidak boleh melebihi jumlah pertemuan;
- total transaksi tidak boleh bernilai negatif;
- jumlah tanggungan keluarga tidak boleh kurang dari nol.

Apabila ditemukan nilai yang tidak wajar, peneliti perlu memeriksa sumber data asli. Jika nilai tersebut merupakan kesalahan input, maka data dapat diperbaiki. Namun, jika nilai tersebut benar-benar terjadi, maka peneliti perlu mempertimbangkan apakah data tersebut termasuk outlier yang perlu dianalisis lebih lanjut.

5. Menangani Data Pencilan

Data pencilan atau outlier perlu ditangani secara hati-hati. Dalam algoritma K-Means, outlier dapat memengaruhi posisi centroid karena centroid dihitung berdasarkan rata-rata nilai anggota cluster. Jika terdapat data dengan nilai yang terlalu ekstrem, maka centroid dapat bergeser dan hasil clustering menjadi kurang representatif.

Penanganan outlier dapat dilakukan dengan beberapa cara, antara lain:

- memeriksa kembali apakah outlier berasal dari kesalahan input;
- mempertahankan outlier apabila data tersebut valid dan penting;
- menghapus outlier apabila terbukti sebagai kesalahan data;
- melakukan transformasi data apabila diperlukan;
- menjelaskan keberadaan outlier dalam interpretasi hasil.

Keputusan terhadap outlier tidak boleh dilakukan hanya karena data tersebut berbeda dari data lain. Dalam penelitian ilmiah, outlier dapat menjadi temuan penting. Oleh karena itu, peneliti perlu membedakan antara outlier sebagai kesalahan dan outlier sebagai fenomena nyata.

6. Mendokumentasikan Proses Pembersihan Data

Dokumentasi pembersihan data merupakan bagian penting dalam penelitian. Setiap tindakan yang dilakukan terhadap data harus dicatat, termasuk jumlah data awal, jumlah data yang dihapus, atribut yang diperbaiki, serta alasan perubahan. Dokumentasi ini membantu menjaga transparansi proses penelitian.

Model CRISP-DM menempatkan persiapan data sebagai tahap penting sebelum pemodelan, termasuk kegiatan memilih, membersihkan, membangun, mengintegrasikan, dan memformat data. Dengan demikian, pembersihan data bukan pekerjaan tambahan, tetapi bagian utama dari proses analisis data mining yang sistematis.

Dalam konteks penelitian berbasis K-Means, pembersihan data sangat menentukan kualitas cluster yang dihasilkan. K-Means mengelompokkan data berdasarkan jarak antara objek dan centroid. Oleh sebab itu, data yang kosong, duplikat, salah format, atau memiliki nilai ekstrem tanpa penjelasan dapat memengaruhi hasil perhitungan. Jika data tidak dibersihkan dengan baik, maka cluster yang terbentuk mungkin tidak menggambarkan kemiripan karakteristik yang sebenarnya.

Sebagai contoh, dalam pengelompokan siswa berdasarkan nilai akademik, satu nilai yang salah input dapat mengubah posisi siswa dalam cluster. Apabila nilai Matematika seorang siswa seharusnya 80 tetapi tercatat 800, maka siswa tersebut akan tampak sangat berbeda dari siswa lain. Akibatnya, perhitungan jarak menjadi tidak wajar dan dapat memengaruhi pembentukan cluster. Contoh ini menunjukkan bahwa pembersihan data bukan sekadar tahap administratif, tetapi bagian penting dari validitas hasil analisis.

Pembersihan data juga berkaitan dengan etika penelitian. Peneliti tidak boleh menghapus atau mengubah data hanya karena hasilnya tidak sesuai dengan harapan. Data hanya boleh diperbaiki apabila terdapat alasan yang jelas, misalnya kesalahan input, duplikasi, atau ketidaksesuaian format. Jika data dihapus, peneliti perlu menjelaskan dasar penghapusannya. Sikap ini penting agar penelitian tetap objektif dan dapat dipercaya.

2.5 Transformasi Data

Transformasi data merupakan tahap dalam proses Knowledge Discovery in Database yang bertujuan mengubah data ke dalam bentuk yang sesuai untuk dianalisis menggunakan metode data mining. Setelah data dipilih dan dibersihkan, data belum tentu dapat langsung digunakan oleh algoritma. Beberapa data masih perlu disesuaikan formatnya, skala nilainya, satuannya, atau bentuk penyajiannya agar proses analisis dapat berjalan dengan baik. Oleh karena itu, transformasi data menjadi tahap penting yang menghubungkan data bersih dengan proses pemodelan atau penerapan algoritma.

Dalam konteks KDD, transformasi data dilakukan sebelum tahap data mining. Fayyad, Piatetsky-Shapiro, dan Smyth (1996) menempatkan transformasi sebagai bagian dari proses penyiapan data agar pola yang ditemukan pada tahap data mining memiliki kualitas yang lebih baik. Transformasi data bukan hanya kegiatan teknis mengubah bentuk angka, tetapi juga proses metodologis untuk memastikan bahwa data benar-benar sesuai dengan tujuan analisis dan karakteristik algoritma yang digunakan.

Pada penelitian berbasis algoritma K-Means, transformasi data memiliki peran yang sangat penting. K-Means bekerja berdasarkan perhitungan jarak antara objek data dengan pusat cluster atau centroid. Oleh karena itu, bentuk dan skala data sangat memengaruhi hasil pengelompokan. Apabila atribut yang digunakan memiliki rentang nilai yang berbeda jauh, maka atribut dengan rentang nilai lebih besar dapat mendominasi proses

perhitungan jarak. Akibatnya, cluster yang terbentuk tidak sepenuhnya mencerminkan kemiripan data secara seimbang.

Sebagai contoh, apabila data siswa dianalisis berdasarkan dua atribut, yaitu nilai ujian dengan rentang 0-100 dan jumlah pendapatan orang tua dengan rentang jutaan rupiah, maka atribut pendapatan dapat memiliki pengaruh yang jauh lebih besar terhadap hasil perhitungan jarak. Kondisi ini dapat menyebabkan hasil clustering lebih banyak dipengaruhi oleh pendapatan daripada nilai akademik. Untuk menghindari ketidakseimbangan tersebut, data perlu ditransformasikan melalui proses normalisasi atau standardisasi.

Secara umum, transformasi data memiliki beberapa tujuan berikut.

- Menyesuaikan data dengan kebutuhan algoritma. Setiap algoritma memiliki karakteristik tertentu. K-Means membutuhkan data numerik karena bekerja berdasarkan perhitungan jarak.
- Menyamakan skala antaratribut. Atribut dengan skala berbeda perlu disesuaikan agar tidak ada atribut yang terlalu dominan dalam proses analisis.
- Mengubah data kategorik menjadi bentuk numerik. Data kategorik perlu dikodekan apabila akan digunakan dalam proses analisis berbasis perhitungan numerik.
- Meningkatkan kualitas hasil data mining. Transformasi yang tepat dapat membantu algoritma menghasilkan pola yang lebih stabil dan lebih mudah diinterpretasikan.
- Menyederhanakan data yang kompleks. Data yang terlalu rinci dapat diringkas atau dibentuk ulang agar lebih sesuai dengan tujuan analisis.

Han, Kamber, dan Pei (2011) menjelaskan bahwa transformasi data dapat mencakup beberapa kegiatan, seperti normalisasi, agregasi, generalisasi, konstruksi atribut, dan reduksi data. Dalam praktik penelitian, bentuk transformasi yang digunakan perlu disesuaikan dengan jenis data dan metode analisis yang dipilih.

Bentuk-Bentuk Transformasi Data

Transformasi data dapat dilakukan melalui beberapa teknik. Tidak semua teknik harus digunakan dalam satu penelitian. Peneliti perlu memilih teknik yang paling sesuai dengan karakteristik data dan tujuan analisis.

1. Normalisasi Data

Normalisasi data merupakan proses mengubah nilai data ke dalam rentang tertentu. Tujuan normalisasi adalah menyamakan skala antaratribut agar tidak terjadi dominasi atribut tertentu dalam perhitungan jarak. Dalam algoritma K-Means, normalisasi sering digunakan karena algoritma ini sensitif terhadap skala nilai.

Salah satu teknik normalisasi yang umum digunakan adalah min-max normalization, yaitu mengubah nilai data ke dalam rentang tertentu, misalnya 0 sampai 1. Rumus sederhana normalisasi min-max adalah sebagai berikut.

$$X' = (X - Xmin) / (Xmax - Xmin)$$

Simbol	Keterangan
X'	Nilai hasil normalisasi
X	Nilai asli data
Xmin	Nilai terkecil pada atribut
Xmax	Nilai terbesar pada atribut

Sebagai contoh, jika nilai tertinggi suatu atribut adalah 100 dan nilai terendahnya adalah 50, maka nilai 75 dapat dinormalisasi sebagai berikut.

$$X' = (75 - 50) / (100 - 50) = 25 / 50 = 0,5$$

Hasil tersebut menunjukkan bahwa nilai 75 berada pada posisi tengah dalam rentang nilai atribut tersebut. Dengan normalisasi, setiap atribut dapat berada pada skala yang lebih seimbang sehingga proses perhitungan jarak dalam K-Means menjadi lebih proporsional.

2. Standardisasi Data

Standardisasi data merupakan proses mengubah nilai data berdasarkan rata-rata dan simpangan baku. Teknik ini sering digunakan apabila data memiliki distribusi yang berbeda atau satuan pengukuran yang tidak seragam. Standardisasi menghasilkan nilai yang menunjukkan posisi suatu data terhadap rata-rata kelompoknya.

Rumus standardisasi yang umum digunakan adalah sebagai berikut.

$$Z = (X - \mu) / \sigma$$

Simbol	Keterangan
Z	Nilai hasil standardisasi
X	Nilai asli data
μ	Rata-rata atribut
σ	Simpangan baku atribut

Nilai hasil standardisasi dapat bernilai positif, negatif, atau nol. Nilai positif menunjukkan bahwa data berada di atas rata-rata, nilai negatif menunjukkan bahwa data berada di bawah rata-rata, sedangkan nilai nol menunjukkan bahwa data sama dengan rata-rata. Dalam penelitian berbasis K-Means, standardisasi dapat digunakan apabila atribut memiliki skala dan sebaran nilai yang berbeda.

3. Pengkodean Data Kategorik

Data kategorik adalah data yang berbentuk kategori, seperti jenis kelamin, status pekerjaan, wilayah, kategori prestasi, atau jenis produk. Karena K-Means bekerja dengan perhitungan numerik, data kategorik tidak dapat langsung digunakan tanpa proses pengkodean. Pengkodean dilakukan untuk mengubah kategori menjadi bentuk numerik.

Namun, pengkodean data kategorik harus dilakukan dengan hati-hati. Tidak semua kategori memiliki hubungan urutan atau jarak yang sama. Misalnya, apabila pekerjaan dikodekan sebagai 1 =

petani, 2 = pedagang, dan 3 = pegawai, angka tersebut hanya berfungsi sebagai kode, bukan menunjukkan bahwa pegawai lebih tinggi daripada pedagang atau petani. Kesalahan memahami kode sebagai nilai matematis dapat menyebabkan interpretasi yang keliru.

Beberapa bentuk pengkodean data kategorik antara lain:

- Label encoding, yaitu memberi angka pada setiap kategori.
- One-hot encoding, yaitu mengubah setiap kategori menjadi kolom tersendiri.
- Ordinal encoding, yaitu memberi angka berdasarkan urutan kategori yang memang memiliki tingkatan.
- Binary encoding, yaitu mengubah kategori menjadi representasi biner.

Dalam penerapan K-Means, data kategorik sebaiknya hanya digunakan apabila telah ditransformasikan secara tepat dan memiliki dasar metodologis yang jelas. Jika sebagian besar data berbentuk kategorik, peneliti perlu mempertimbangkan apakah K-Means merupakan metode yang paling sesuai atau perlu menggunakan metode clustering lain yang lebih cocok untuk data kategorik.

4. Agregasi Data

Agregasi data merupakan proses menggabungkan beberapa nilai menjadi satu nilai ringkasan. Teknik ini digunakan ketika data terlalu rinci dan perlu disederhanakan agar sesuai dengan tujuan analisis. Contohnya adalah menghitung rata-rata nilai beberapa mata pelajaran, total transaksi pelanggan, jumlah kehadiran selama satu semester, atau rata-rata pembelian per bulan.

Dalam penelitian pendidikan, data nilai beberapa ujian dapat diagregasi menjadi nilai rata-rata akademik. Dalam bidang bisnis, data transaksi harian dapat diagregasi menjadi total transaksi bulanan. Agregasi membantu menyederhanakan data, tetapi peneliti perlu memastikan bahwa ringkasan tersebut tidak menghilangkan informasi penting.

Data Awal	Bentuk Agregasi	Tujuan
Nilai tugas 1, tugas 2, tugas 3	Rata-rata nilai tugas	Menyederhanakan indikator akademik
Transaksi harian pelanggan	Total transaksi bulanan	Melihat tingkat aktivitas pelanggan
Kehadiran per pertemuan	Persentase kehadiran	Mengukur konsistensi kehadiran
Nilai beberapa mata pelajaran	Rata-rata nilai akademik	Menentukan gambaran prestasi umum

Agregasi data bermanfaat apabila peneliti ingin mengurangi kompleksitas data. Namun, agregasi harus dilakukan sesuai tujuan penelitian. Apabila detail data diperlukan untuk melihat variasi tertentu, maka agregasi yang terlalu sederhana dapat mengurangi ketajaman analisis.

5. Pembentukan Atribut Baru

Pembentukan atribut baru atau attribute construction dilakukan dengan membuat variabel baru dari atribut yang sudah ada. Tujuannya adalah menghasilkan atribut yang lebih informatif dan lebih sesuai dengan kebutuhan analisis. Misalnya, dari data nilai beberapa mata pelajaran dapat dibuat atribut rata-rata nilai. Dari data jumlah hadir dan jumlah pertemuan dapat dibuat atribut persentase kehadiran.

Atribut Asal	Atribut Baru	Rumus atau Cara Pembentukan
Jumlah hadir dan jumlah pertemuan	Persentase kehadiran	Jumlah hadir / jumlah pertemuan x 100
Nilai Matematika, IPA, Bahasa Indonesia	Rata-rata akademik	Jumlah nilai / jumlah mata pelajaran
Total belanja dan jumlah transaksi	Rata-rata belanja	Total belanja / jumlah transaksi
Pendapatan dan jumlah tanggungan	Rasio beban ekonomi	Jumlah tanggungan / pendapatan

Dalam algoritma K-Means, pembentukan atribut baru dapat membantu memperjelas karakteristik data. Namun, atribut baru harus dibuat berdasarkan pertimbangan logis dan teoritis. Peneliti tidak boleh membuat atribut baru hanya untuk menyesuaikan hasil, tetapi harus menjelaskan alasan pembentukannya.

6. Reduksi Data

Reduksi data merupakan proses mengurangi jumlah data atau atribut tanpa menghilangkan informasi penting. Reduksi dilakukan apabila dataset terlalu besar, memiliki atribut yang terlalu banyak, atau terdapat atribut yang saling berulang. Tujuan reduksi adalah membuat proses analisis lebih efisien dan hasilnya lebih mudah ditafsirkan.

Reduksi data dapat dilakukan dengan beberapa cara, antara lain:

- menghapus atribut yang tidak relevan;
- menggabungkan atribut yang memiliki makna serupa;
- memilih atribut utama berdasarkan tujuan penelitian;
- menggunakan teknik statistik untuk mereduksi dimensi;
- mengambil sampel data apabila jumlah data sangat besar.

Dalam penelitian K-Means, reduksi atribut perlu dilakukan secara hati-hati. Atribut yang tidak relevan dapat mengganggu proses clustering, tetapi pengurangan atribut yang terlalu banyak dapat membuat hasil cluster kurang mencerminkan keadaan data. Oleh sebab itu, reduksi harus dilakukan berdasarkan pertimbangan ilmiah, bukan semata-mata untuk menyederhanakan perhitungan.

Relevansi Transformasi Data dengan Algoritma K-Means

Transformasi data memiliki hubungan langsung dengan kualitas hasil algoritma K-Means. Hal ini disebabkan oleh karakteristik K-Means yang menggunakan jarak sebagai dasar pengelompokan. Jain (2010) menjelaskan bahwa K-Means merupakan salah satu algoritma clustering yang sederhana dan populer, tetapi memiliki sensitivitas terhadap pemilihan parameter, pusat awal, dan

karakteristik data. Oleh karena itu, data yang digunakan perlu disiapkan dengan tepat sebelum algoritma dijalankan.

Beberapa alasan transformasi data penting dalam K-Means adalah sebagai berikut.

- K-Means menggunakan perhitungan jarak. Setiap data akan dikelompokkan berdasarkan jaraknya terhadap centroid. Skala data yang tidak seimbang dapat memengaruhi hasil jarak.
- K-Means lebih sesuai untuk data numerik. Data kategorik perlu diubah terlebih dahulu sebelum digunakan, atau peneliti perlu memilih metode lain yang lebih sesuai.
- Centroid dihitung berdasarkan rata-rata. Karena centroid merupakan nilai rata-rata dari anggota cluster, nilai ekstrem atau data dengan skala besar dapat memengaruhi posisi centroid.
- Hasil cluster bergantung pada atribut yang digunakan. Atribut yang telah ditransformasikan akan menentukan bagaimana kemiripan antarobjek dihitung.
- Interpretasi cluster membutuhkan data yang bermakna. Transformasi yang keliru dapat menghasilkan cluster yang sulit dijelaskan secara ilmiah.

Dalam penerapan praktis, peneliti perlu menjelaskan transformasi apa yang dilakukan sebelum menjalankan K-Means. Penjelasan ini penting agar pembaca dapat memahami bagaimana data disiapkan dan mengapa bentuk data tersebut dianggap layak untuk dianalisis. Misalnya, jika peneliti melakukan normalisasi min-max, maka peneliti perlu menjelaskan bahwa normalisasi dilakukan untuk menyamakan skala antaratribut agar tidak ada atribut yang mendominasi proses perhitungan jarak.

Contoh Transformasi Data untuk K-Means

Misalnya seorang peneliti ingin mengelompokkan siswa berdasarkan nilai akademik. Data awal yang tersedia adalah nilai Matematika, Bahasa Indonesia, dan IPA. Karena ketiga nilai tersebut berada dalam rentang yang sama, yaitu 0-100, data dapat

digunakan secara langsung atau tetap dinormalisasi untuk menjaga keseragaman proses. Namun, jika terdapat atribut lain seperti jumlah penghasilan orang tua atau jumlah absensi, maka normalisasi menjadi lebih penting.

Contoh sederhana data sebelum transformasi adalah sebagai berikut.

Siswa	Nilai Matematika	Nilai Bahasa Indonesia	Nilai IPA
S1	80	85	82
S2	65	70	68
S3	90	88	92
S4	55	60	58

Apabila data tersebut dinormalisasi menggunakan min-max normalization, setiap nilai akan diubah ke dalam rentang 0 sampai 1. Dengan demikian, data menjadi lebih seragam dan siap digunakan untuk proses perhitungan jarak dalam K-Means. Transformasi ini tidak mengubah urutan kemampuan siswa, tetapi hanya mengubah skala angka agar lebih proporsional dalam proses analisis.

Contoh hasil transformasi dapat disajikan sebagai berikut.

Siswa	Matematika Normalisasi	Bahasa Indonesia Normalisasi	IPA Normalisasi
S1	0,71	0,89	0,71
S2	0,29	0,36	0,29
S3	1,00	1,00	1,00
S4	0,00	0,00	0,00

Tabel tersebut menunjukkan bahwa data telah berada pada rentang yang sama. Dengan kondisi ini, perhitungan jarak dalam K-Means menjadi lebih seimbang karena setiap atribut memiliki skala nilai yang sama.

Kesalahan yang Perlu Dihindari dalam Transformasi Data

Transformasi data perlu dilakukan dengan cermat karena kesalahan pada tahap ini dapat memengaruhi hasil akhir penelitian. Beberapa kesalahan yang perlu dihindari antara lain:

- melakukan normalisasi tanpa menjelaskan alasannya;
- mengubah data kategorik menjadi angka tanpa memahami maknanya;
- menyamakan semua data tanpa mempertimbangkan jenis atribut;
- membentuk atribut baru tanpa dasar teoritis;
- menghilangkan atribut penting karena dianggap memperumit analisis;
- menggunakan data dengan skala berbeda tanpa transformasi;
- menafsirkan hasil kode kategorik sebagai nilai matematis yang memiliki jarak.

Dalam penelitian ilmiah, transformasi data harus dijelaskan secara terbuka. Peneliti perlu menyebutkan jenis transformasi yang digunakan, alasan penggunaannya, dan dampaknya terhadap proses analisis. Penjelasan ini penting agar proses penelitian dapat ditelusuri dan dipertanggungjawabkan.

2.6 Evaluasi dan Interpretasi Hasil

Evaluasi dan interpretasi hasil merupakan tahap akhir dalam proses Knowledge Discovery in Database yang bertujuan menilai apakah pola yang ditemukan melalui data mining benar-benar bermakna, relevan, dan dapat digunakan sebagai dasar pengetahuan. Tahap ini sangat penting karena keluaran algoritma tidak secara otomatis dapat dianggap sebagai temuan ilmiah. Hasil perhitungan masih perlu diperiksa, dijelaskan, dan dikaitkan dengan tujuan penelitian agar memiliki makna yang dapat dipertanggungjawabkan.

Dalam proses KDD, evaluasi dan interpretasi menjadi penghubung antara hasil teknis dengan pengetahuan yang dapat dimanfaatkan.

Fayyad, Piatetsky-Shapiro, dan Smyth (1996) menegaskan bahwa pola yang ditemukan dalam proses KDD harus bersifat valid, baru, bermanfaat, dan dapat dipahami. Dengan demikian, hasil data mining tidak cukup hanya menunjukkan angka, model, atau cluster, tetapi harus mampu menjawab persoalan yang dianalisis secara logis dan ilmiah.

Evaluasi dan interpretasi memiliki fokus yang berbeda, meskipun keduanya saling berkaitan. Evaluasi berfokus pada penilaian kualitas hasil analisis, sedangkan interpretasi berfokus pada pemberian makna terhadap hasil tersebut. Dalam konteks algoritma K-Means, evaluasi dilakukan untuk menilai apakah cluster yang terbentuk cukup baik, stabil, dan sesuai dengan tujuan analisis. Sementara itu, interpretasi dilakukan untuk menjelaskan karakteristik setiap cluster dan implikasinya terhadap masalah penelitian.

Aspek	Evaluasi Hasil	Interpretasi Hasil
Fokus utama	Menilai kualitas hasil data mining	Menjelaskan makna hasil data mining
Pertanyaan utama	Apakah hasilnya baik dan layak digunakan?	Apa arti hasil tersebut dalam konteks penelitian?
Bentuk kegiatan	Mengukur kualitas model, cluster, atau pola	Membaca karakteristik dan memberikan penjelasan
Contoh pada K-Means	Menilai nilai SSE, silhouette, atau struktur cluster	Menjelaskan cluster tinggi, sedang, dan rendah
Tujuan akhir	Memastikan hasil tidak menyesatkan	Mengubah hasil teknis menjadi pengetahuan

Dalam penerapan K-Means, evaluasi hasil dilakukan untuk melihat kualitas pengelompokan yang terbentuk. K-Means menghasilkan cluster berdasarkan kedekatan data terhadap centroid. Namun, terbentuknya cluster tidak selalu berarti bahwa hasil tersebut sudah tepat. Peneliti perlu memeriksa apakah data dalam satu

cluster memiliki kemiripan yang cukup tinggi dan apakah antarcluster memiliki perbedaan yang cukup jelas. Jika anggota dalam satu cluster masih sangat beragam atau antarcluster sulit dibedakan, maka hasil clustering perlu dievaluasi kembali.

Salah satu ukuran yang sering digunakan dalam evaluasi K-Means adalah Sum of Squared Error atau SSE. SSE menggambarkan jumlah kuadrat jarak antara setiap data dengan centroid pada cluster masing-masing. Semakin kecil nilai SSE, semakin dekat data terhadap centroid dalam cluster. Namun, nilai SSE tidak boleh ditafsirkan secara terpisah tanpa mempertimbangkan jumlah cluster. Semakin banyak jumlah cluster, nilai SSE cenderung menurun. Oleh karena itu, SSE sering digunakan bersama metode elbow untuk membantu menentukan jumlah cluster yang lebih sesuai.

$$SSE = \Sigma_{i=1}^{k} \Sigma_{x \in C_i} dist(x, c_i)^2$$

Simbol	Keterangan
SSE	Jumlah kuadrat kesalahan dalam cluster
k	Jumlah cluster
C_i	Cluster ke-i
x	Data atau objek dalam cluster
c_i	Centroid cluster ke-i
$dist(x, c_i)$	Jarak data terhadap centroid

Selain SSE, evaluasi cluster juga dapat dilakukan menggunakan silhouette coefficient. Ukuran ini digunakan untuk melihat seberapa baik suatu data berada dalam cluster-nya dibandingkan dengan cluster lain. Nilai silhouette berada pada rentang -1 sampai 1. Nilai yang mendekati 1 menunjukkan bahwa data sesuai dengan cluster-nya. Nilai mendekati 0 menunjukkan bahwa data berada di antara dua cluster. Sementara itu, nilai negatif menunjukkan kemungkinan data lebih sesuai ditempatkan pada cluster lain. Rousseeuw (1987) memperkenalkan silhouette sebagai alat untuk menilai struktur pengelompokan berdasarkan kedekatan dalam cluster dan pemisahan antarcluster.

Dalam penelitian yang lebih teknis, evaluasi K-Means juga dapat menggunakan ukuran lain, seperti Davies-Bouldin Index dan Calinski-Harabasz Index. Davies-Bouldin Index digunakan untuk menilai rasio antara kedekatan dalam cluster dan pemisahan antarcluster. Nilai Davies-Bouldin yang lebih kecil umumnya menunjukkan cluster yang lebih baik. Sementara itu, Calinski-Harabasz Index menilai perbandingan antara variasi antarcluster dan variasi dalam cluster. Nilai yang lebih besar umumnya menunjukkan hasil clustering yang lebih baik. Meskipun demikian, penggunaan ukuran-ukuran tersebut perlu disesuaikan dengan tingkat kebutuhan penelitian dan kemampuan pembaca sasaran buku.

Ukuran Evaluasi	Tujuan Penggunaan	Arah Interpretasi Umum
SSE atau WCSS	Menilai kedekatan data dengan centroid dalam cluster	Lebih kecil lebih baik, tetapi dipengaruhi jumlah cluster
Elbow Method	Membantu menentukan jumlah cluster	Titik siku menunjukkan jumlah cluster yang relatif sesuai
Silhouette Coefficient	Menilai kepaduan dan pemisahan cluster	Mendekati 1 menunjukkan hasil lebih baik
Davies-Bouldin Index	Menilai kemiripan antarcluster	Lebih kecil umumnya lebih baik
Calinski-Harabasz Index	Menilai pemisahan antarcluster	Lebih besar umumnya lebih baik
Analisis domain	Menilai kesesuaian hasil dengan konteks penelitian	Hasil harus logis dan dapat dijelaskan

Meskipun ukuran evaluasi kuantitatif penting, peneliti tidak boleh hanya bergantung pada angka. Evaluasi hasil data mining juga harus mempertimbangkan konteks penelitian. Dalam penelitian pendidikan, misalnya, cluster yang terbentuk harus dapat dijelaskan berdasarkan kondisi akademik peserta didik. Jika cluster yang dihasilkan secara matematis baik tetapi sulit dijelaskan secara substantif, maka peneliti perlu meninjau kembali atribut yang digunakan, jumlah cluster, atau proses transformasi data.

Setelah evaluasi dilakukan, tahap berikutnya adalah interpretasi hasil. Interpretasi bertujuan menjelaskan makna pola atau cluster yang ditemukan. Dalam K-Means, interpretasi biasanya dilakukan dengan membaca nilai centroid, rata-rata atribut pada setiap cluster, jumlah anggota cluster, dan karakteristik dominan dari masing-masing kelompok. Centroid dapat membantu peneliti memahami gambaran umum setiap cluster karena centroid mewakili pusat karakteristik kelompok.

Langkah-langkah interpretasi hasil K-Means dapat dilakukan sebagai berikut.

- Membaca jumlah anggota setiap cluster. Peneliti perlu melihat berapa banyak data yang masuk ke dalam setiap cluster. Jumlah anggota cluster dapat menunjukkan kelompok yang dominan atau kelompok yang relatif kecil.
- Menganalisis nilai centroid. Nilai centroid menunjukkan pusat dari setiap cluster. Dengan membaca centroid, peneliti dapat mengetahui karakteristik umum masing-masing kelompok.
- Membandingkan atribut antarcluster. Peneliti perlu membandingkan nilai atribut pada setiap cluster untuk melihat perbedaan utama antar kelompok.
- Memberikan label pada cluster. Setelah karakteristik cluster dipahami, peneliti dapat memberikan label yang sesuai, misalnya tinggi, sedang, dan rendah. Label harus diberikan berdasarkan data, bukan berdasarkan dugaan.

- Mengaitkan hasil dengan tujuan penelitian. Interpretasi harus kembali pada tujuan awal penelitian. Hasil cluster perlu dijelaskan sebagai jawaban terhadap masalah yang dikaji.
- Menyusun implikasi atau rekomendasi. Apabila diperlukan, hasil interpretasi dapat digunakan untuk menyusun rekomendasi yang relevan dengan konteks penelitian.

Contoh interpretasi sederhana dapat dilihat pada kasus pengelompokan siswa berdasarkan nilai akademik. Apabila K-Means menghasilkan tiga cluster, peneliti dapat membaca rata-rata nilai setiap cluster. Jika cluster pertama memiliki nilai rata-rata tinggi pada seluruh mata pelajaran, maka cluster tersebut dapat diberi label prestasi tinggi. Jika cluster kedua memiliki nilai rata-rata sedang, maka dapat diberi label prestasi sedang. Jika cluster ketiga memiliki nilai rata-rata rendah, maka dapat diberi label prestasi rendah.

Cluster	Karakteristik Umum	Label Interpretatif	Makna
Cluster 1	Nilai akademik tinggi pada sebagian besar atribut	Prestasi tinggi	Kelompok siswa dengan kemampuan akademik kuat
Cluster 2	Nilai akademik berada pada tingkat menengah	Prestasi sedang	Kelompok siswa dengan kemampuan cukup dan masih perlu penguatan
Cluster 3	Nilai akademik rendah pada sebagian besar atribut	Prestasi rendah	Kelompok siswa yang membutuhkan pendampingan lebih intensif

Pemberian label seperti tinggi, sedang, dan rendah harus dilakukan secara hati-hati. Label tersebut bukan bawaan dari algoritma, melainkan hasil interpretasi peneliti berdasarkan karakteristik

data. K-Means hanya menghasilkan pembagian kelompok berdasarkan perhitungan jarak. Makna setiap kelompok ditentukan melalui analisis peneliti terhadap nilai atribut, konteks data, dan tujuan penelitian.

Interpretasi hasil juga perlu menghindari generalisasi yang berlebihan. Cluster yang terbentuk hanya berlaku untuk data, atribut, dan periode yang digunakan dalam penelitian. Jika penelitian menggunakan data siswa pada satu tahun ajaran tertentu, maka hasilnya tidak otomatis berlaku untuk seluruh siswa pada tahun ajaran lain. Peneliti perlu menjelaskan batasan ini agar pembaca memahami ruang lingkup temuan.

Dalam konteks KDD, hasil akhir yang baik bukan hanya hasil yang akurat secara teknis, tetapi juga hasil yang dapat dipahami dan bermanfaat. Han, Kamber, dan Pei (2011) menjelaskan bahwa evaluasi pola diperlukan untuk mengidentifikasi pola yang benar-benar menarik dan berguna. Oleh karena itu, hasil data mining harus dinilai dari dua sisi, yaitu kualitas perhitungan dan kegunaan makna. Pola yang baik seharusnya dapat menjelaskan kondisi data dan membantu pengguna dalam memahami masalah.

Beberapa pertanyaan yang dapat digunakan peneliti dalam mengevaluasi dan menginterpretasikan hasil adalah sebagai berikut.

- Apakah hasil analisis sesuai dengan tujuan penelitian?
- Apakah cluster yang terbentuk memiliki perbedaan karakteristik yang jelas?
- Apakah atribut yang digunakan cukup menjelaskan pembentukan cluster?
- Apakah jumlah cluster yang dipilih dapat dipertanggung-jawabkan?
- Apakah hasil analisis dapat dijelaskan secara logis?
- Apakah hasil analisis memiliki manfaat praktis atau ilmiah?
- Apakah ada data pencilan yang memengaruhi hasil cluster?

- Apakah interpretasi hasil tidak melampaui batas data yang digunakan?

Dalam penelitian berbasis K-Means, evaluasi juga perlu memperhatikan kemungkinan kelemahan algoritma. K-Means sensitif terhadap penentuan jumlah cluster dan pemilihan centroid awal. Selain itu, K-Means juga sensitif terhadap skala data dan outlier. Jain (2010) menjelaskan bahwa K-Means merupakan algoritma yang sederhana dan banyak digunakan, tetapi hasilnya dapat dipengaruhi oleh karakteristik data dan pemilihan parameter. Oleh sebab itu, evaluasi perlu dilakukan untuk memastikan bahwa hasil clustering tidak hanya terbentuk secara matematis, tetapi juga masuk akal secara substantif.

Interpretasi hasil sebaiknya disajikan dalam bentuk narasi ilmiah yang jelas. Peneliti dapat memadukan tabel, grafik, dan uraian tertulis agar pembaca lebih mudah memahami hasil analisis. Tabel digunakan untuk menyajikan nilai centroid, jumlah anggota cluster, atau rata-rata atribut. Grafik dapat digunakan untuk memperlihatkan pola atau perbedaan antarcluster. Narasi digunakan untuk menjelaskan makna hasil secara lebih mendalam.

Contoh narasi interpretasi dapat ditulis sebagai berikut.

Berdasarkan hasil clustering menggunakan algoritma K-Means, data terbagi ke dalam tiga cluster. Cluster pertama menunjukkan karakteristik nilai akademik yang relatif tinggi pada sebagian besar atribut, sehingga dapat diinterpretasikan sebagai kelompok prestasi tinggi. Cluster kedua memiliki nilai rata-rata sedang dan dapat diinterpretasikan sebagai kelompok prestasi sedang. Sementara itu, cluster ketiga menunjukkan nilai rata-rata yang lebih rendah dibandingkan cluster lainnya, sehingga dapat dikategorikan sebagai kelompok yang memerlukan perhatian akademik lebih lanjut.

Narasi tersebut menunjukkan bahwa interpretasi tidak hanya menyebutkan hasil cluster, tetapi juga menjelaskan makna dari setiap kelompok. Dalam buku ilmiah, gaya penulisan seperti ini penting agar hasil analisis tidak tampak sebagai keluaran

perangkat lunak semata, melainkan sebagai hasil pemikiran akademik yang terarah.

Secara umum, kesalahan yang perlu dihindari dalam evaluasi dan interpretasi hasil adalah sebagai berikut.

- menerima hasil algoritma tanpa evaluasi;
- memberi label cluster tanpa membaca karakteristik data;
- menyimpulkan hubungan sebab-akibat dari hasil clustering;
- mengabaikan pengaruh atribut dan skala data;
- menggeneralisasikan hasil di luar batas dataset;
- hanya menampilkan angka tanpa penjelasan;
- mengabaikan data pencilan yang memengaruhi hasil;
- tidak menjelaskan dasar pemilihan jumlah cluster.

Kesalahan paling umum dalam interpretasi K-Means adalah menganggap bahwa cluster yang terbentuk merupakan kategori mutlak. Padahal, cluster merupakan hasil pengelompokan berdasarkan atribut dan jumlah cluster yang telah ditentukan. Jika atribut berubah, jumlah cluster berubah, atau data diperbarui, maka hasil cluster juga dapat berubah. Oleh karena itu, interpretasi hasil K-Means harus selalu dikaitkan dengan batasan data dan metode yang digunakan.

Dengan demikian, evaluasi dan interpretasi hasil merupakan tahap penting dalam KDD karena menentukan apakah hasil data mining dapat menjadi pengetahuan yang bermakna. Evaluasi berfungsi menilai kualitas pola atau cluster yang terbentuk, sedangkan interpretasi berfungsi menjelaskan arti hasil dalam konteks penelitian. Dalam penerapan algoritma K-Means, tahap ini dilakukan dengan menilai kualitas cluster, membaca centroid, membandingkan atribut antarcluster, memberi label kelompok, dan mengaitkan hasil dengan tujuan penelitian. Melalui evaluasi dan interpretasi yang tepat, hasil K-Means tidak hanya menjadi keluaran teknis, tetapi juga menjadi dasar pemahaman dan pengambilan keputusan yang lebih ilmiah.

2.7 Pentingnya Kualitas Data dalam Data Mining

Kualitas data merupakan salah satu faktor utama yang menentukan keberhasilan proses data mining. Data mining tidak hanya bergantung pada kecanggihan algoritma, tetapi juga pada kondisi data yang digunakan sebagai dasar analisis. Algoritma yang baik tetap dapat menghasilkan kesimpulan yang keliru apabila data yang digunakan tidak lengkap, tidak konsisten, tidak relevan, atau mengandung banyak kesalahan. Oleh karena itu, kualitas data harus menjadi perhatian utama sejak tahap awal proses *Knowledge Discovery in Database*.

Dalam penelitian berbasis data, kualitas data berkaitan dengan tingkat kelayakan data untuk digunakan dalam menjawab tujuan analisis. Data yang berkualitas bukan hanya data yang banyak, tetapi data yang sesuai, akurat, lengkap, konsisten, dan dapat ditafsirkan secara tepat. Wang dan Strong (1996) menjelaskan bahwa kualitas data tidak hanya dapat dilihat dari aspek akurasi, tetapi juga dari dimensi lain yang penting bagi pengguna data, seperti relevansi, kelengkapan, ketepatan waktu, dan kemudahan pemahaman. Pandangan ini menunjukkan bahwa data yang benar secara angka belum tentu cukup apabila data tersebut tidak relevan dengan kebutuhan analisis.

Dalam data mining, kualitas data menjadi penting karena setiap proses analisis bertumpu pada data yang tersedia. Apabila data yang digunakan memiliki kesalahan, maka pola yang ditemukan juga dapat menjadi salah. Kesalahan tersebut dapat berupa nilai yang hilang, data duplikat, nilai ekstrem yang tidak diperiksa, perbedaan format, atau atribut yang tidak sesuai dengan tujuan penelitian. Rahm dan Do (2000) menjelaskan bahwa persoalan kualitas data sering muncul dalam proses pembersihan data, terutama ketika data berasal dari sumber yang berbeda, memiliki duplikasi, atau mengalami ketidaksesuaian struktur.

Dalam konteks KDD, kualitas data menjadi landasan sebelum data masuk ke tahap data mining. Tahapan seperti seleksi data,

pembersihan data, dan transformasi data sebenarnya bertujuan memastikan bahwa data yang digunakan benar-benar layak dianalisis. Model CRISP-DM juga menempatkan pemahaman data dan persiapan data sebagai tahap penting sebelum pemodelan dilakukan. Hal ini menunjukkan bahwa proses data mining tidak seharusnya langsung dimulai dari algoritma, tetapi harus diawali dengan pemeriksaan dan penyiapan data secara cermat.

Pada algoritma K-Means, kualitas data memiliki pengaruh yang sangat besar terhadap hasil clustering. K-Means bekerja dengan menghitung jarak antara data dan centroid. Oleh karena itu, kesalahan nilai, perbedaan skala atribut, data kosong, atau data pencilan dapat mengubah hasil perhitungan jarak. Jika hasil perhitungan jarak berubah, maka posisi data terhadap cluster juga dapat berubah. Dengan demikian, kualitas data secara langsung memengaruhi pembentukan cluster dan interpretasi hasil pengelompokan.

Secara umum, kualitas data dalam data mining dapat dilihat dari beberapa aspek berikut.

1. Akurasi Data

Akurasi data menunjukkan sejauh mana data menggambarkan kondisi yang sebenarnya. Data yang akurat adalah data yang nilainya benar dan sesuai dengan sumber aslinya. Misalnya, jika nilai ujian seorang siswa adalah 85, maka data yang tercatat dalam dataset juga harus 85. Apabila terjadi kesalahan input menjadi 58 atau 850, maka data tersebut tidak akurat dan dapat memengaruhi hasil analisis.

Dalam algoritma K-Means, akurasi data sangat penting karena setiap nilai digunakan dalam perhitungan jarak. Satu nilai yang salah dapat menyebabkan objek data masuk ke cluster yang tidak sesuai. Oleh karena itu, pemeriksaan akurasi data perlu dilakukan sebelum data dianalisis. Pemeriksaan dapat dilakukan dengan membandingkan data dalam dataset dengan dokumen sumber,

sistem informasi, atau catatan resmi yang digunakan sebagai acuan.

2. Kelengkapan Data

Kelengkapan data berkaitan dengan ada atau tidaknya nilai pada setiap atribut yang dibutuhkan. Data yang tidak lengkap dapat menyebabkan proses analisis terganggu, terutama apabila nilai yang kosong terdapat pada atribut penting. Misalnya, dalam penelitian pengelompokan prestasi siswa, data nilai Matematika, Bahasa Indonesia, dan IPA diperlukan sebagai dasar clustering. Jika salah satu nilai tidak tersedia, maka proses pengelompokan dapat menjadi kurang akurat.

Data kosong dapat ditangani dengan beberapa cara, seperti memperbaiki berdasarkan sumber asli, menghapus data yang tidak lengkap, atau mengisi nilai kosong menggunakan metode tertentu. Namun, setiap tindakan harus memiliki alasan yang jelas. Pengisian nilai kosong secara sembarangan dapat mengubah karakteristik data dan memengaruhi hasil clustering.

3. Konsistensi Data

Konsistensi data menunjukkan keseragaman format, satuan, dan penulisan dalam dataset. Data yang tidak konsisten dapat menyebabkan kesalahan pembacaan oleh sistem atau peneliti. Misalnya, kategori “Laki-laki” ditulis dengan beberapa bentuk, seperti “L”, “Pria”, dan “Male”. Meskipun memiliki makna yang sama, variasi penulisan tersebut dapat terbaca sebagai kategori yang berbeda.

Dalam data numerik, ketidakkonsistenan juga dapat terjadi pada satuan nilai. Misalnya, sebagian data pendapatan ditulis dalam rupiah penuh, sedangkan sebagian lain ditulis dalam jutaan rupiah. Jika data seperti ini langsung digunakan, hasil analisis dapat menjadi tidak valid. Oleh karena itu, penyeragaman format dan satuan perlu dilakukan sebelum proses data mining.

4. Relevansi Data

Relevansi data menunjukkan kesesuaian data dengan tujuan penelitian. Data yang relevan adalah data yang memiliki hubungan langsung dengan masalah yang dianalisis. Dalam data mining, penggunaan data yang tidak relevan dapat mengganggu hasil analisis karena algoritma akan tetap memproses atribut yang diberikan, meskipun atribut tersebut tidak memiliki makna terhadap tujuan penelitian.

Pada K-Means, relevansi atribut menjadi sangat penting. Jika tujuan penelitian adalah mengelompokkan siswa berdasarkan prestasi akademik, maka atribut yang relevan adalah nilai akademik atau indikator belajar lainnya. Atribut seperti nama, nomor induk, dan alamat lengkap dapat digunakan sebagai identitas, tetapi tidak layak digunakan sebagai dasar perhitungan cluster. Penggunaan atribut yang tidak relevan dapat menghasilkan cluster yang sulit dijelaskan secara ilmiah.

5. Ketepatan Skala Data

Ketepatan skala data berkaitan dengan kesesuaian rentang nilai antaratribut yang digunakan. Dalam K-Means, atribut dengan skala nilai yang lebih besar dapat mendominasi perhitungan jarak. Misalnya, jika satu atribut memiliki rentang 0–100 dan atribut lain memiliki rentang 0–10.000.000, maka atribut dengan rentang nilai lebih besar akan lebih memengaruhi hasil cluster.

Untuk mengatasi persoalan ini, data dapat ditransformasikan melalui normalisasi atau standardisasi. Transformasi tersebut bertujuan menyamakan skala antaratribut agar perhitungan jarak menjadi lebih seimbang. Han, Kamber, dan Pei (2011) menjelaskan bahwa praproses data, termasuk pembersihan dan transformasi, merupakan bagian penting dalam data mining karena data yang tersedia sering kali masih tidak lengkap, mengandung gangguan, dan tidak konsisten.

6. Keandalan Sumber Data

Keandalan sumber data menunjukkan bahwa data berasal dari sumber yang dapat dipercaya dan dapat ditelusuri. Dalam penelitian ilmiah, data sebaiknya berasal dari dokumen resmi, sistem informasi, catatan administrasi, hasil observasi, instrumen penelitian, atau sumber lain yang jelas. Data yang tidak diketahui asalnya dapat menurunkan kredibilitas penelitian karena sulit diverifikasi.

Keandalan sumber data juga penting untuk menjaga validitas penelitian. Jika data diperoleh dari sumber yang tidak jelas, maka hasil analisis sulit dipertanggungjawabkan. Oleh karena itu, peneliti perlu menjelaskan sumber data, periode pengambilan data, jumlah data, serta atribut yang digunakan dalam analisis.

7. Ketepatan Waktu Data

Ketepatan waktu data berkaitan dengan kesesuaian data terhadap periode analisis. Data yang terlalu lama atau tidak sesuai dengan periode penelitian dapat menghasilkan interpretasi yang kurang tepat. Misalnya, apabila penelitian bertujuan menganalisis kondisi siswa tahun ajaran tertentu, maka data yang digunakan harus berasal dari tahun ajaran tersebut.

Dalam konteks pengambilan keputusan, data yang mutakhir lebih berguna untuk menggambarkan kondisi yang sedang terjadi. Namun, dalam penelitian historis, data lama tetap dapat digunakan selama sesuai dengan tujuan penelitian. Dengan demikian, ketepatan waktu data harus dipahami berdasarkan konteks analisis.

Kualitas data yang baik memberikan beberapa manfaat dalam proses data mining. Pertama, kualitas data membantu meningkatkan ketepatan hasil analisis. Data yang akurat dan lengkap akan memberikan dasar yang lebih kuat bagi algoritma untuk menemukan pola. Kedua, kualitas data membantu memperjelas interpretasi hasil. Hasil clustering, klasifikasi, atau prediksi akan lebih mudah dijelaskan apabila data yang digunakan

memiliki struktur yang baik. Ketiga, kualitas data membantu meningkatkan kepercayaan terhadap hasil penelitian. Pembaca akan lebih mudah menerima hasil analisis apabila proses penyiapan data dijelaskan secara transparan.

Dalam penelitian berbasis K-Means, kualitas data juga berpengaruh terhadap stabilitas cluster. Jika data mengandung banyak kesalahan atau outlier yang tidak diperiksa, centroid dapat bergeser dari posisi yang seharusnya. Akibatnya, anggota cluster dapat berubah dan interpretasi kelompok menjadi kurang tepat. Oleh karena itu, sebelum menerapkan K-Means, peneliti perlu memastikan bahwa data telah melalui proses seleksi, pembersihan, dan transformasi yang memadai.

Selain memengaruhi hasil teknis, kualitas data juga berpengaruh terhadap etika penelitian. Data yang digunakan harus dikelola secara bertanggung jawab, terutama jika berkaitan dengan identitas individu. Peneliti perlu membedakan antara data yang diperlukan untuk analisis dan data yang hanya berfungsi sebagai identitas. Informasi pribadi yang tidak relevan sebaiknya tidak digunakan dalam proses data mining. Jika data individu harus ditampilkan, peneliti dapat menggunakan kode agar kerahasiaan tetap terjaga.

Dalam praktik penelitian, kualitas data perlu dijaga melalui beberapa langkah. Peneliti dapat melakukan pemeriksaan awal terhadap dataset, mengecek data kosong, menghapus duplikasi, menyeragamkan format, memeriksa nilai ekstrem, dan memastikan bahwa atribut yang digunakan sesuai dengan tujuan analisis. Langkah-langkah ini bukan sekadar tahapan teknis, tetapi bagian dari upaya menjaga validitas penelitian.

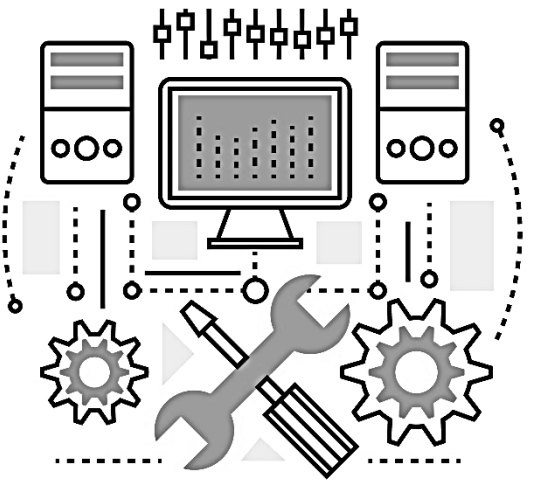
Beberapa langkah praktis untuk menjaga kualitas data adalah sebagai berikut.

- Menentukan sumber data yang jelas dan dapat dipercaya.
- Memilih atribut yang sesuai dengan tujuan penelitian.
- Memeriksa kelengkapan nilai pada setiap atribut.

- Menghapus atau memperbaiki data duplikat.
- Menyeragamkan format penulisan dan satuan data.
- Memeriksa nilai yang tidak wajar atau berada di luar rentang logis.
- Melakukan normalisasi atau standardisasi apabila skala atribut berbeda.
- Mendokumentasikan setiap proses perbaikan data.
- Menjaga kerahasiaan data yang bersifat pribadi.
- Menjelaskan batasan data dalam laporan penelitian.

Dokumentasi proses pengelolaan kualitas data sangat penting dalam naskah ilmiah. Peneliti perlu menjelaskan jumlah data awal, jumlah data setelah pembersihan, atribut yang digunakan, perlakuan terhadap data kosong, serta bentuk transformasi yang dilakukan. Penjelasan tersebut menunjukkan bahwa penelitian dilakukan secara sistematis dan dapat ditelusuri. Tanpa dokumentasi, pembaca sulit menilai apakah hasil data mining benar-benar berasal dari data yang layak.

Dalam buku ini, kualitas data ditempatkan sebagai fondasi utama sebelum pembahasan algoritma K-Means dilakukan secara lebih teknis. Hal ini penting karena K-Means bukan sekadar proses menghitung jarak dan membentuk cluster, tetapi bagian dari proses penemuan pengetahuan yang membutuhkan data berkualitas. Jika data yang digunakan tidak layak, maka hasil cluster tidak dapat dijadikan dasar interpretasi yang kuat.



Bab 03

CLUSTERING DALAM DATA MINING

3.1 Pengertian Clustering

Clustering merupakan salah satu teknik dalam data mining yang digunakan untuk mengelompokkan objek data ke dalam beberapa kelompok berdasarkan kemiripan karakteristik. Objek-objek yang berada dalam satu kelompok atau *cluster* diharapkan memiliki tingkat kemiripan yang tinggi, sedangkan objek yang berada pada kelompok berbeda diharapkan memiliki tingkat perbedaan yang lebih besar. Dengan demikian, clustering bertujuan menemukan struktur alami dalam data tanpa harus menentukan kategori terlebih dahulu.

Dalam data mining, clustering termasuk ke dalam teknik *unsupervised learning* atau pembelajaran tidak terawasi. Hal ini berarti proses pengelompokan dilakukan tanpa menggunakan label kelas yang sudah ditentukan sebelumnya. Berbeda dengan klasifikasi yang membutuhkan data berlabel, clustering digunakan ketika peneliti belum mengetahui kelompok data secara pasti dan

ingin menemukan pola pengelompokan berdasarkan karakteristik data yang tersedia. Jain (2010) menjelaskan bahwa clustering berupaya menemukan kelompok alami dari data tidak berlabel berdasarkan ukuran kemiripan, sehingga objek dalam kelompok yang sama memiliki kemiripan tinggi dan objek pada kelompok berbeda memiliki kemiripan rendah.

Clustering dapat dipahami sebagai proses eksploratif karena digunakan untuk memahami struktur data. Dalam penelitian, teknik ini berguna ketika peneliti ingin mengetahui apakah data memiliki kecenderungan membentuk kelompok tertentu. Misalnya, dalam data pendidikan, clustering dapat digunakan untuk mengelompokkan siswa berdasarkan nilai akademik. Dalam data bisnis, clustering dapat digunakan untuk mengelompokkan pelanggan berdasarkan perilaku pembelian. Dalam data sosial, clustering dapat digunakan untuk mengelompokkan masyarakat berdasarkan indikator ekonomi atau demografis.

Han, Kamber, dan Pei menempatkan clustering sebagai salah satu konsep penting dalam data mining, bersama dengan teknik lain seperti klasifikasi, pola asosiasi, dan deteksi outlier. Buku *Data Mining: Concepts and Techniques* juga membahas clustering sebagai bagian dari metode dasar dalam penambangan data, khususnya untuk menemukan struktur kelompok dari sekumpulan data. Dengan posisi tersebut, clustering menjadi teknik yang penting untuk memahami data ketika kategori atau kelas belum tersedia.

Secara sederhana, prinsip clustering dapat dijelaskan melalui dua konsep utama, yaitu kemiripan dalam cluster dan perbedaan antarcluster. Kemiripan dalam cluster berarti objek yang berada pada kelompok yang sama memiliki karakteristik yang relatif serupa. Perbedaan antarcluster berarti objek yang berada pada kelompok berbeda memiliki karakteristik yang relatif berbeda. Semakin tinggi kemiripan dalam cluster dan semakin jelas perbedaan antarcluster, maka hasil clustering dapat dianggap semakin baik.

Ciri utama clustering dapat dijelaskan sebagai berikut.

- Tidak memerlukan label awal
- Clustering tidak membutuhkan kategori yang sudah ditentukan sebelumnya. Kelompok terbentuk berdasarkan pola atau kemiripan data.
- Bertujuan menemukan struktur data
- Clustering digunakan untuk melihat apakah data memiliki susunan kelompok tertentu yang dapat ditafsirkan.
- Menggunakan ukuran kemiripan atau jarak
- Proses pengelompokan biasanya didasarkan pada ukuran jarak atau kesamaan antarobjek.
- Bersifat eksploratif
- Hasil clustering membantu peneliti memahami data, tetapi tetap memerlukan interpretasi sesuai konteks penelitian.
- Dipengaruhi oleh atribut yang digunakan
- Cluster yang terbentuk bergantung pada variabel atau atribut yang dimasukkan ke dalam analisis.

Dalam praktiknya, clustering dapat diterapkan pada berbagai jenis data. Data yang digunakan dapat berupa nilai akademik, data transaksi, data pelanggan, data wilayah, data kesehatan, maupun data sosial ekonomi. Namun, setiap metode clustering memiliki karakteristik tersendiri. Ada metode yang lebih sesuai untuk data numerik, ada pula metode yang dapat menangani data kategorik, data berdimensi tinggi, atau data dengan bentuk cluster yang tidak beraturan.

Salah satu hal penting dalam clustering adalah ukuran kemiripan. Ukuran kemiripan digunakan untuk menentukan seberapa dekat hubungan antara satu objek data dengan objek lainnya. Dalam metode berbasis jarak, dua objek dianggap mirip apabila jaraknya kecil. Sebaliknya, dua objek dianggap berbeda apabila jaraknya besar. Pada algoritma K-Means, ukuran yang sering digunakan adalah jarak Euclidean. Oleh karena itu, data yang digunakan dalam K-Means umumnya berbentuk numerik atau telah ditransformasi-kan ke dalam bentuk numerik.

Contoh sederhana konsep clustering dapat dilihat pada data siswa berdasarkan nilai akademik. Jika terdapat sejumlah siswa dengan nilai tinggi, sedang, dan rendah, maka clustering dapat membantu membentuk kelompok berdasarkan kemiripan nilai tersebut. Siswa yang memiliki nilai relatif tinggi akan cenderung masuk ke dalam kelompok yang sama. Siswa dengan nilai sedang akan membentuk kelompok lain, sedangkan siswa dengan nilai rendah akan berada pada kelompok yang berbeda. Namun, label seperti "tinggi", "sedang", dan "rendah" bukan diberikan oleh algoritma secara otomatis, melainkan merupakan hasil interpretasi peneliti setelah melihat karakteristik setiap cluster.

Dalam konteks ini, clustering berbeda dengan klasifikasi. Pada klasifikasi, kategori data sudah diketahui sejak awal. Misalnya, data siswa sudah memiliki label "lulus" dan "tidak lulus", kemudian algoritma digunakan untuk mempelajari pola dari label tersebut. Sementara itu, pada clustering, kategori belum ditentukan. Algoritma berusaha menemukan kelompok berdasarkan kemiripan data. Perbedaan ini penting agar peneliti tidak keliru dalam memilih metode analisis.

Clustering memiliki manfaat besar dalam proses analisis data karena mampu membantu peneliti memahami pola kelompok yang tidak tampak secara langsung. Dengan clustering, data yang besar dan kompleks dapat disusun menjadi kelompok-kelompok yang lebih mudah dipahami. Hasil clustering kemudian dapat digunakan sebagai dasar untuk interpretasi, pengambilan keputusan, atau penyusunan strategi.

Dalam bidang pendidikan, clustering dapat digunakan untuk mengelompokkan peserta didik berdasarkan prestasi akademik, pola belajar, atau kemampuan awal. Dalam bidang bisnis, clustering dapat digunakan untuk segmentasi pelanggan. Dalam bidang kesehatan, clustering dapat digunakan untuk mengelompokkan pasien berdasarkan karakteristik tertentu. Dalam bidang pemerintahan, clustering dapat digunakan untuk

mengelompokkan wilayah atau masyarakat berdasarkan indikator sosial ekonomi.

Meskipun demikian, clustering tidak boleh dipahami sebagai proses yang selalu menghasilkan kelompok yang benar secara mutlak. Jain (2010) menekankan bahwa definisi cluster tidak selalu sederhana karena cluster dapat berbeda dalam bentuk, ukuran, dan kepadatan. Selain itu, keberadaan noise atau data pencilan dapat membuat proses clustering menjadi lebih sulit. Oleh karena itu, hasil clustering harus dievaluasi dan ditafsirkan dengan mempertimbangkan karakteristik data serta konteks penelitian.

Dalam penelitian berbasis K-Means, pengertian clustering menjadi dasar yang sangat penting. K-Means merupakan salah satu algoritma clustering yang bekerja dengan membagi data ke dalam sejumlah cluster berdasarkan kedekatan data terhadap centroid. Setiap data akan ditempatkan pada cluster dengan centroid terdekat. Proses ini dilakukan secara berulang sampai posisi cluster relatif stabil. Dengan demikian, K-Means merupakan bentuk penerapan clustering yang menggunakan pendekatan partisi dan perhitungan jarak.

Namun, sebelum menggunakan K-Means, peneliti perlu memahami bahwa hasil clustering sangat dipengaruhi oleh pemilihan atribut, jumlah cluster, skala data, dan kualitas data. Jika atribut yang digunakan tidak relevan, maka cluster yang terbentuk dapat sulit dijelaskan. Jika skala data tidak seimbang, maka hasil pengelompokan dapat didominasi oleh atribut tertentu. Oleh sebab itu, clustering harus ditempatkan dalam proses analisis yang sistematis, mulai dari seleksi data, pembersihan data, transformasi data, penerapan algoritma, hingga interpretasi hasil.

3.2 Tujuan Clustering

Tujuan utama clustering adalah mengelompokkan objek data ke dalam beberapa kelompok berdasarkan tingkat kemiripan karakteristik. Melalui proses ini, data yang semula berjumlah banyak dan sulit dipahami dapat disusun menjadi kelompok-

kelompok yang lebih terstruktur. Setiap kelompok atau *cluster* berisi objek yang memiliki karakteristik relatif serupa, sedangkan objek pada cluster yang berbeda memiliki karakteristik yang lebih berbeda. Dengan demikian, clustering membantu peneliti memahami susunan data secara lebih sederhana dan bermakna.

Dalam data mining, clustering digunakan ketika peneliti belum memiliki kategori awal terhadap data yang dianalisis. Oleh karena itu, clustering bertujuan menemukan pola kelompok yang terbentuk secara alami dari data. Jain (2010) menjelaskan bahwa clustering berhubungan dengan proses menemukan struktur kelompok dalam data tidak berlabel berdasarkan ukuran kemiripan tertentu. Artinya, tujuan clustering bukan untuk memasukkan data ke dalam kelas yang sudah ditentukan, melainkan untuk menemukan sendiri pola kelompok berdasarkan karakteristik data.

Salah satu tujuan penting clustering adalah menyederhanakan data yang kompleks. Data dalam jumlah besar sering kali sulit dibaca secara langsung karena terdiri atas banyak objek dan atribut. Melalui clustering, data dapat dipetakan ke dalam beberapa kelompok sehingga peneliti lebih mudah melihat pola umum. Misalnya, data siswa yang terdiri atas puluhan atau ratusan nilai dapat dikelompokkan menjadi beberapa cluster berdasarkan kemiripan capaian akademik. Dengan cara ini, data yang besar dapat dibaca dalam bentuk kelompok yang lebih mudah dipahami.

Tujuan clustering juga berkaitan dengan segmentasi data. Segmentasi merupakan proses membagi data ke dalam kelompok-kelompok tertentu agar setiap kelompok dapat dianalisis secara lebih spesifik. Dalam bidang pendidikan, segmentasi dapat digunakan untuk mengelompokkan siswa berdasarkan prestasi akademik, pola belajar, atau kemampuan awal. Dalam bidang bisnis, segmentasi dapat digunakan untuk mengelompokkan pelanggan berdasarkan perilaku pembelian. Dalam bidang sosial, segmentasi dapat digunakan untuk mengelompokkan masyarakat berdasarkan indikator ekonomi atau demografis.

Secara umum, tujuan clustering dapat dijelaskan sebagai berikut.

- Menemukan struktur kelompok dalam data
  Clustering bertujuan menemukan susunan kelompok yang belum diketahui sebelumnya. Struktur kelompok ini terbentuk berdasarkan kemiripan data, bukan berdasarkan label yang sudah ditentukan.
- Mengelompokkan objek berdasarkan kemiripan karakteristik
  Objek yang memiliki karakteristik serupa ditempatkan dalam cluster yang sama. Sebaliknya, objek dengan karakteristik berbeda ditempatkan pada cluster yang berbeda.
- Menyederhanakan data yang besar dan kompleks
  Clustering membantu menyusun data yang banyak menjadi beberapa kelompok sehingga lebih mudah dianalisis dan ditafsirkan.
- Membantu proses segmentasi
  Clustering dapat digunakan untuk membagi data menjadi segmen-segmen tertentu, seperti kelompok siswa, pelanggan, pasien, wilayah, atau penerima layanan.
- Mendukung pengambilan keputusan
  Hasil clustering dapat menjadi dasar untuk menyusun strategi, kebijakan, atau rekomendasi yang berbeda pada setiap kelompok.
- Mengenali pola yang tidak terlihat secara langsung
  Clustering membantu menemukan pola kelompok yang mungkin tidak mudah dikenali melalui pengamatan manual.
- Menjadi dasar analisis lanjutan
  Hasil clustering dapat digunakan sebagai tahap awal sebelum dilakukan analisis lain, seperti evaluasi kelompok, pemetaan risiko, atau penyusunan strategi intervensi.

Dalam konteks penelitian, tujuan clustering tidak hanya menghasilkan pembagian data ke dalam beberapa kelompok. Hasil clustering harus dapat membantu peneliti memahami perbedaan karakteristik antar kelompok. Misalnya, apabila data siswa dikelompokkan menjadi tiga cluster, peneliti tidak cukup hanya menyebutkan jumlah siswa pada setiap cluster. Peneliti perlu

menjelaskan karakteristik masing-masing kelompok, seperti kelompok dengan nilai tinggi, kelompok dengan nilai sedang, dan kelompok dengan nilai rendah. Dengan demikian, clustering memiliki tujuan analitis, bukan sekadar teknis.

Pada algoritma K-Means, tujuan clustering diwujudkan melalui pembentukan sejumlah cluster berdasarkan kedekatan data terhadap centroid. Setiap objek data ditempatkan pada cluster yang memiliki centroid paling dekat. Proses ini dilakukan secara berulang sampai posisi cluster relatif stabil. Dengan mekanisme tersebut, K-Means bertujuan membentuk kelompok yang memiliki jarak antaranggota relatif kecil di dalam cluster dan jarak yang lebih besar antarcluster.

Tujuan clustering juga dapat dikaitkan dengan efisiensi analisis. Tanpa clustering, peneliti harus membaca setiap data secara satu per satu. Cara tersebut dapat dilakukan pada data yang kecil, tetapi menjadi tidak efisien apabila data berjumlah besar. Clustering membantu peneliti melihat gambaran umum data melalui kelompok-kelompok yang terbentuk. Dengan demikian, clustering dapat mempercepat proses pemahaman terhadap data tanpa mengabaikan karakteristik penting yang ada di dalamnya.

Selain itu, clustering bertujuan membantu peneliti mengenali kelompok yang memerlukan perhatian khusus. Dalam bidang pendidikan, hasil clustering dapat menunjukkan kelompok siswa dengan kemampuan akademik rendah yang membutuhkan pendampingan lebih intensif. Dalam bidang kesehatan, clustering dapat membantu mengenali kelompok pasien dengan karakteristik risiko tertentu. Dalam bidang pemerintahan, clustering dapat membantu mengelompokkan masyarakat atau wilayah berdasarkan tingkat kebutuhan layanan. Dengan cara ini, clustering dapat mendukung keputusan yang lebih terarah.

Namun, tujuan clustering harus dipahami secara proporsional. Clustering tidak bertujuan membuktikan hubungan sebab-akibat antarvariabel. Hasil clustering hanya menunjukkan pengelompokan berdasarkan kemiripan atribut yang digunakan.

Oleh karena itu, peneliti tidak boleh menyimpulkan bahwa satu atribut menyebabkan terbentuknya cluster tertentu tanpa analisis tambahan. Clustering lebih tepat digunakan untuk menemukan struktur, pola, atau segmentasi awal dalam data.

Kualitas tujuan clustering juga bergantung pada pemilihan atribut. Jika atribut yang digunakan relevan dengan tujuan penelitian, maka hasil cluster lebih mudah ditafsirkan. Sebaliknya, jika atribut yang digunakan tidak sesuai, maka hasil cluster dapat menjadi tidak bermakna. Sebagai contoh, apabila tujuan penelitian adalah mengelompokkan siswa berdasarkan prestasi akademik, maka atribut yang digunakan sebaiknya berupa nilai akademik atau indikator belajar. Atribut administratif seperti nomor induk atau nama tidak layak digunakan dalam perhitungan clustering karena tidak menggambarkan kemiripan akademik.

3.3 Karakteristik Metode Clustering

Metode clustering memiliki karakteristik khusus yang membedakannya dari teknik data mining lainnya. Karakteristik tersebut berkaitan dengan cara clustering membentuk kelompok, jenis data yang digunakan, ukuran kemiripan, serta cara hasil pengelompokan ditafsirkan. Pemahaman terhadap karakteristik clustering penting agar peneliti tidak hanya mampu menjalankan algoritma, tetapi juga dapat memahami batasan dan makna dari hasil cluster yang terbentuk.

Karakteristik utama clustering adalah bekerja pada data yang belum memiliki label kelas. Oleh karena itu, clustering termasuk ke dalam pendekatan *unsupervised learning*. Dalam pendekatan ini, algoritma tidak diberi informasi awal mengenai kategori data. Algoritma akan mengelompokkan data berdasarkan pola kemiripan yang ditemukan dari atribut yang tersedia. Han, Kamber, dan Pei (2011) menjelaskan bahwa clustering digunakan untuk menganalisis objek data tanpa mengetahui label kelas sebelumnya, sehingga tujuan utamanya adalah menemukan kelompok atau struktur yang tersembunyi dalam data.

Karakteristik berikutnya adalah clustering sangat bergantung pada ukuran kemiripan atau jarak. Objek data dianggap memiliki hubungan yang dekat apabila memiliki karakteristik yang serupa. Sebaliknya, objek yang memiliki perbedaan besar akan ditempatkan pada cluster yang berbeda. Dalam banyak metode clustering, ukuran jarak menjadi dasar utama dalam menentukan kedekatan antarobjek. Pada algoritma K-Means, ukuran jarak yang umum digunakan adalah jarak Euclidean, yaitu jarak antara dua titik dalam ruang multidimensi.

Secara umum, karakteristik metode clustering dapat dijelaskan sebagai berikut.

1. Tidak Menggunakan Label Kelas Awal

Clustering tidak memerlukan label atau kategori yang sudah ditentukan sebelumnya. Data dikelompokkan berdasarkan kemiripan yang ditemukan oleh algoritma. Hal ini berbeda dengan klasifikasi, yang membutuhkan data berlabel sebagai dasar pembentukan model. Karena tidak menggunakan label awal, clustering lebih bersifat eksploratif. Peneliti menggunakan clustering untuk menemukan kemungkinan struktur kelompok yang terdapat dalam data.

Sebagai contoh, dalam penelitian pendidikan, data siswa dapat dikelompokkan berdasarkan nilai akademik tanpa perlu memberi label awal seperti “tinggi”, “sedang”, atau “rendah”. Label tersebut baru diberikan setelah hasil clustering dianalisis. Dengan demikian, proses clustering menghasilkan kelompok terlebih dahulu, kemudian peneliti menafsirkan makna dari kelompok tersebut berdasarkan karakteristik data.

2. Berbasis Kemiripan atau Kedekatan Data

Metode clustering bekerja berdasarkan prinsip bahwa data yang mirip seharusnya berada dalam kelompok yang sama. Kemiripan ini dapat dihitung menggunakan ukuran tertentu, seperti jarak Euclidean, jarak Manhattan, cosine similarity, atau ukuran

kemiripan lainnya. Pemilihan ukuran kemiripan harus disesuaikan dengan jenis data dan metode clustering yang digunakan.

Dalam K-Means, setiap data akan dihitung jaraknya terhadap centroid. Data kemudian dimasukkan ke dalam cluster dengan centroid terdekat. Artinya, hasil cluster sangat dipengaruhi oleh kedekatan nilai antarobjek. Oleh karena itu, data yang digunakan dalam K-Means sebaiknya berbentuk numerik atau telah melalui proses transformasi yang tepat.

3. Bertujuan Membentuk Kelompok yang Homogen

Salah satu karakteristik penting clustering adalah berusaha membentuk kelompok yang homogen. Homogen berarti anggota dalam satu cluster memiliki kemiripan karakteristik yang tinggi. Semakin mirip anggota dalam satu cluster, semakin baik kepaduan cluster tersebut. Sebaliknya, cluster yang berisi data dengan karakteristik terlalu beragam akan sulit ditafsirkan.

Dalam penelitian berbasis K-Means, homogenitas cluster dapat dilihat dari kedekatan data terhadap centroid. Jika data dalam satu cluster memiliki jarak yang kecil terhadap centroid, maka cluster tersebut dapat dianggap lebih kompak. Namun, kekompakan cluster tetap perlu ditafsirkan bersama konteks penelitian agar tidak hanya menjadi ukuran matematis.

4. Memiliki Perbedaan Antarcluster

Selain membentuk kelompok yang homogen, clustering juga bertujuan menghasilkan perbedaan yang jelas antarcluster. Objek pada cluster yang berbeda seharusnya memiliki karakteristik yang berbeda pula. Jika antarcluster tidak memiliki perbedaan yang jelas, maka hasil pengelompokan menjadi sulit dimaknai.

Dalam konteks analisis akademik, misalnya, jika tiga cluster yang terbentuk memiliki rata-rata nilai yang hampir sama, maka interpretasi kelompok menjadi lemah. Sebaliknya, jika satu cluster menunjukkan nilai akademik tinggi, cluster lain sedang, dan cluster lainnya rendah, maka hasil clustering lebih mudah dijelaskan. Dengan demikian, hasil clustering yang baik tidak hanya

memperhatikan kemiripan dalam cluster, tetapi juga perbedaan antarcluster.

5. Dipengaruhi oleh Atribut yang Digunakan

Hasil clustering sangat bergantung pada atribut atau variabel yang digunakan dalam analisis. Atribut yang dipilih menjadi dasar bagi algoritma dalam menghitung kemiripan antarobjek. Jika atribut yang digunakan relevan, maka cluster yang terbentuk cenderung lebih bermakna. Sebaliknya, jika atribut yang digunakan tidak sesuai dengan tujuan penelitian, maka hasil cluster dapat menyesatkan.

Sebagai contoh, apabila tujuan penelitian adalah mengelompokkan siswa berdasarkan prestasi akademik, maka atribut yang tepat adalah nilai mata pelajaran, nilai ujian, atau indikator akademik lainnya. Atribut seperti nama, nomor induk, atau alamat lengkap tidak tepat digunakan sebagai dasar perhitungan clustering karena tidak menggambarkan kemiripan akademik. Oleh karena itu, pemilihan atribut merupakan bagian penting dalam keberhasilan metode clustering.

6. Sensitif terhadap Skala Data

Beberapa metode clustering, terutama K-Means, sangat sensitif terhadap skala data. Atribut dengan rentang nilai yang besar dapat mendominasi perhitungan jarak. Misalnya, apabila satu atribut memiliki rentang 0–100 dan atribut lain memiliki rentang 0–10.000, maka atribut kedua dapat memberikan pengaruh yang lebih besar terhadap hasil cluster.

Untuk mengatasi masalah tersebut, data perlu melalui tahap transformasi, seperti normalisasi atau standardisasi. Transformasi ini bertujuan agar setiap atribut berada pada skala yang lebih seimbang. Jain (2010) menjelaskan bahwa keberhasilan metode clustering sangat dipengaruhi oleh representasi data, pemilihan ukuran kedekatan, dan karakteristik algoritma yang digunakan. Oleh karena itu, tahap penyiapan data tidak dapat diabaikan dalam analisis clustering.

7. Memerlukan Penentuan Jumlah Cluster

Pada beberapa metode clustering, jumlah cluster perlu ditentukan sebelum proses analisis dilakukan. K-Means merupakan salah satu metode yang memerlukan penentuan jumlah cluster di awal. Peneliti harus menentukan nilai *k*, yaitu jumlah cluster yang akan dibentuk. Penentuan jumlah cluster ini sangat penting karena akan memengaruhi hasil akhir pengelompokan.

Jika jumlah cluster terlalu sedikit, data yang sebenarnya berbeda dapat tergabung dalam kelompok yang sama. Sebaliknya, jika jumlah cluster terlalu banyak, kelompok yang terbentuk dapat menjadi terlalu rinci dan sulit ditafsirkan. Oleh karena itu, penentuan jumlah cluster perlu mempertimbangkan tujuan penelitian, karakteristik data, dan metode evaluasi tertentu, seperti *elbow method* atau silhouette coefficient.

8. Bersifat Eksploratif

Clustering memiliki sifat eksploratif karena digunakan untuk menemukan pola kelompok yang belum diketahui sebelumnya. Hasil clustering dapat membuka pemahaman awal mengenai struktur data. Namun, karena bersifat eksploratif, hasil clustering perlu ditafsirkan secara hati-hati. Peneliti tidak boleh langsung menganggap cluster sebagai kategori final tanpa evaluasi dan pembahasan lebih lanjut.

Dalam penelitian ilmiah, sifat eksploratif ini justru menjadi keunggulan clustering. Peneliti dapat menemukan pola yang sebelumnya tidak tampak melalui pengamatan biasa. Namun, hasil tersebut tetap perlu dikaji berdasarkan teori, konteks data, dan tujuan penelitian agar memiliki nilai ilmiah.

9. Memerlukan Evaluasi Hasil

Hasil clustering perlu dievaluasi untuk mengetahui apakah kelompok yang terbentuk memiliki kualitas yang baik. Evaluasi dapat dilakukan dengan melihat kepadatan dalam cluster, pemisahan antarcluster, nilai jarak, atau ukuran validasi cluster.

Dalam K-Means, evaluasi dapat menggunakan SSE, *elbow method*, silhouette coefficient, atau ukuran lain yang sesuai.

Evaluasi penting karena clustering akan selalu membentuk kelompok selama algoritma dijalankan. Namun, tidak semua kelompok yang terbentuk memiliki makna yang kuat. Tanpa evaluasi, peneliti berisiko menerima hasil algoritma secara mentah. Oleh karena itu, evaluasi menjadi bagian penting untuk memastikan bahwa hasil clustering layak digunakan dan dapat dijelaskan secara ilmiah.

10. Membutuhkan Interpretasi Peneliti

Clustering tidak memberikan makna kelompok secara otomatis. Algoritma hanya membagi data ke dalam cluster berdasarkan perhitungan tertentu. Makna dari setiap cluster ditentukan melalui interpretasi peneliti. Peneliti perlu membaca nilai centroid, rata-rata atribut, jumlah anggota cluster, dan karakteristik dominan dari setiap kelompok.

Sebagai contoh, K-Means dapat menghasilkan Cluster 1, Cluster 2, dan Cluster 3. Nama cluster tersebut belum memiliki makna substantif. Setelah peneliti membaca karakteristik setiap cluster, barulah cluster dapat diberi label seperti “prestasi tinggi”, “prestasi sedang”, dan “prestasi rendah”. Label ini harus didasarkan pada data, bukan pada asumsi pribadi.

Dalam penerapannya, karakteristik clustering menunjukkan bahwa metode ini tidak dapat digunakan secara sembarangan. Peneliti perlu memahami jenis data yang digunakan, tujuan analisis, atribut yang dipilih, teknik transformasi data, serta cara mengevaluasi hasil. Pada algoritma K-Means, perhatian terhadap karakteristik clustering menjadi semakin penting karena K-Means sangat dipengaruhi oleh nilai numerik, skala data, centroid awal, dan jumlah cluster.

Xu dan Wunsch (2005) menjelaskan bahwa clustering merupakan proses penting dalam analisis data karena dapat digunakan untuk menemukan struktur dalam data, tetapi hasilnya sangat

bergantung pada representasi data, ukuran kemiripan, dan algoritma yang digunakan. Dengan demikian, clustering bukan hanya proses membagi data, tetapi proses analitis yang membutuhkan pertimbangan metodologis.

3.4 Perbedaan Clustering dan Classification

Clustering dan *classification* merupakan dua teknik penting dalam data mining yang sama-sama digunakan untuk mengelompokkan data. Meskipun keduanya terlihat mirip karena menghasilkan pembagian data ke dalam kelompok tertentu, secara konsep keduanya memiliki perbedaan mendasar. Perbedaan tersebut terutama terletak pada keberadaan label kelas, tujuan analisis, cara kerja algoritma, serta bentuk hasil yang dihasilkan. Pemahaman terhadap perbedaan ini penting agar peneliti tidak keliru dalam memilih metode analisis.

Clustering merupakan teknik pengelompokan data yang dilakukan tanpa menggunakan label kelas awal. Artinya, data yang dianalisis belum memiliki kategori tertentu sebelum proses algoritma dijalankan. Algoritma clustering akan menemukan kelompok berdasarkan kemiripan karakteristik antarobjek. Sebaliknya, classification merupakan teknik yang digunakan untuk menempatkan data ke dalam kelas atau kategori yang telah diketahui sebelumnya. Dalam classification, data pelatihan sudah memiliki label sehingga algoritma dapat mempelajari pola dari label tersebut.

Han, Kamber, dan Pei (2011) menjelaskan bahwa clustering termasuk ke dalam teknik *unsupervised learning*, sedangkan classification termasuk ke dalam teknik *supervised learning*. Perbedaan ini menjadi dasar utama dalam memahami kedua metode tersebut. Pada clustering, algoritma bekerja tanpa arahan kelas. Pada classification, algoritma belajar dari data yang sudah diberi kelas untuk membangun model yang dapat digunakan pada data baru.

Dalam clustering, peneliti biasanya ingin mengetahui struktur kelompok yang terdapat dalam data. Misalnya, seorang peneliti memiliki data nilai siswa dan ingin mengetahui apakah siswa dapat dikelompokkan ke dalam beberapa kelompok berdasarkan kemiripan nilai. Dalam kasus ini, peneliti belum menentukan sejak awal siapa yang termasuk kelompok tinggi, sedang, atau rendah. Kelompok tersebut akan muncul setelah proses clustering dilakukan. Label seperti "tinggi", "sedang", dan "rendah" diberikan setelah peneliti membaca karakteristik setiap cluster.

Berbeda dengan itu, classification digunakan apabila kategori data sudah diketahui. Misalnya, peneliti memiliki data siswa yang sudah diberi label "lulus" dan "tidak lulus". Algoritma classification kemudian digunakan untuk mempelajari pola dari data tersebut, misalnya berdasarkan nilai ujian, kehadiran, atau tugas. Setelah model terbentuk, model tersebut dapat digunakan untuk memprediksi apakah siswa baru termasuk kategori "lulus" atau "tidak lulus".

Dengan demikian, clustering lebih tepat digunakan untuk menemukan kelompok, sedangkan classification lebih tepat digunakan untuk menentukan kelas. Clustering bersifat eksploratif karena bertujuan mencari pola kelompok yang belum diketahui. Classification bersifat prediktif atau penentuan kategori karena bertujuan membangun model berdasarkan kelas yang telah tersedia.

Perbedaan clustering dan classification dapat dilihat dari beberapa aspek berikut.

1. Berdasarkan Keberadaan Label Data

Perbedaan paling mendasar antara clustering dan classification terletak pada keberadaan label data. Clustering tidak membutuhkan label data sebelum analisis dilakukan. Data dikelompokkan berdasarkan kemiripan atribut yang tersedia. Sementara itu, classification membutuhkan label kelas pada data

pelatihan. Label tersebut menjadi dasar bagi algoritma untuk mempelajari pola.

Dalam penelitian berbasis K-Means, data yang digunakan tidak perlu memiliki kelas awal. Misalnya, data siswa cukup berisi nilai Matematika, Bahasa Indonesia, IPA, atau atribut akademik lain. Algoritma K-Means kemudian mengelompokkan siswa berdasarkan kemiripan nilai. Sebaliknya, apabila peneliti menggunakan classification, maka data harus memiliki label seperti "prestasi tinggi", "prestasi sedang", atau "prestasi rendah" sejak awal.

2. Berdasarkan Tujuan Analisis

Clustering bertujuan menemukan struktur atau pola kelompok dalam data. Teknik ini digunakan ketika peneliti ingin mengetahui apakah data memiliki kecenderungan membentuk kelompok tertentu. Clustering membantu peneliti memahami karakteristik data secara eksploratif.

Classification bertujuan membangun model yang dapat menentukan kelas suatu data. Teknik ini digunakan ketika kelas data telah diketahui dan peneliti ingin mengklasifikasikan data baru berdasarkan pola yang dipelajari dari data sebelumnya. Dengan demikian, classification lebih banyak digunakan untuk prediksi kategori.

Contohnya, apabila tujuan penelitian adalah mengelompokkan pelanggan berdasarkan perilaku pembelian tanpa kategori awal, maka clustering lebih sesuai. Namun, apabila tujuan penelitian adalah menentukan apakah pelanggan termasuk kategori "loyal" atau "tidak loyal" berdasarkan data yang telah memiliki label, maka classification lebih sesuai.

3. Berdasarkan Cara Kerja Algoritma

Pada clustering, algoritma bekerja dengan menghitung kemiripan atau jarak antarobjek. Objek yang memiliki karakteristik mirip akan ditempatkan dalam cluster yang sama. Pada K-Means, proses

ini dilakukan dengan menghitung jarak antara data dan centroid. Data akan masuk ke cluster dengan centroid terdekat.

Pada classification, algoritma bekerja dengan mempelajari hubungan antara atribut dan label kelas. Setelah pola hubungan tersebut dipelajari, algoritma membentuk model klasifikasi. Model ini kemudian digunakan untuk menentukan kelas data baru. Contoh algoritma classification antara lain Decision Tree, Naïve Bayes, Support Vector Machine, K-Nearest Neighbor, dan Artificial Neural Network.

4. Berdasarkan Jenis Pembelajaran

Clustering termasuk dalam *unsupervised learning* karena tidak menggunakan data berlabel. Algoritma hanya menerima atribut data, kemudian mencari pola kelompok berdasarkan kesamaan karakteristik. Classification termasuk dalam *supervised learning* karena memerlukan data pelatihan yang telah memiliki label kelas.

Perbedaan jenis pembelajaran ini berdampak pada proses penelitian. Pada clustering, peneliti perlu lebih kuat dalam menafsirkan hasil karena algoritma tidak memberikan nama kelompok secara otomatis. Pada classification, peneliti perlu memastikan bahwa label kelas pada data pelatihan benar dan konsisten, sebab kesalahan label dapat menyebabkan model klasifikasi menjadi tidak akurat.

5. Berdasarkan Bentuk Hasil

Hasil clustering berupa cluster atau kelompok data. Cluster yang dihasilkan biasanya diberi nama seperti Cluster 1, Cluster 2, dan Cluster 3. Nama tersebut belum memiliki makna substantif sampai peneliti melakukan interpretasi. Setelah karakteristik setiap cluster dibaca, peneliti dapat memberikan label interpretatif seperti “kelompok prestasi tinggi”, “kelompok prestasi sedang”, atau “kelompok prestasi rendah”.

Hasil classification berupa kelas atau kategori yang telah diketahui sebelumnya. Misalnya, hasil klasifikasi dapat berupa “lulus” atau “tidak lulus”, “layak” atau “tidak layak”, “rendah”, “sedang”, atau

"tinggi". Karena kelas sudah ditentukan sejak awal, hasil classification lebih langsung mengarah pada penentuan kategori data.

6. Berdasarkan Contoh Algoritma

Clustering dan classification memiliki algoritma yang berbeda. Algoritma clustering dirancang untuk menemukan kelompok berdasarkan kemiripan data. Algoritma classification dirancang untuk mempelajari pola dari data berlabel dan menentukan kelas data baru.

Contoh algoritma clustering antara lain:

- K-Means;
- K-Medoids;
- Hierarchical Clustering;
- DBSCAN;
- Fuzzy C-Means.

Contoh algoritma classification antara lain:

- Decision Tree;
- Naïve Bayes;
- Support Vector Machine;
- K-Nearest Neighbor;
- Random Forest;
- Artificial Neural Network.

K-Means berada pada kelompok clustering, bukan classification. Oleh karena itu, penggunaan K-Means tidak tepat apabila tujuan penelitian adalah memprediksi kelas yang telah diketahui berdasarkan data berlabel. K-Means lebih sesuai digunakan apabila tujuan penelitian adalah menemukan kelompok berdasarkan kemiripan data.

7. Berdasarkan Kebutuhan Data

Clustering membutuhkan data yang memiliki atribut untuk dihitung kemiripannya. Pada K-Means, data yang digunakan

sebaiknya berbentuk numerik atau telah ditransformasikan ke dalam bentuk numerik. Clustering tidak membutuhkan label kelas, tetapi tetap membutuhkan atribut yang relevan agar hasil pengelompokan bermakna.

Classification membutuhkan data yang terdiri atas atribut dan label kelas. Atribut digunakan sebagai variabel prediktor, sedangkan label digunakan sebagai target. Tanpa label kelas, proses classification tidak dapat dilakukan secara tepat karena algoritma tidak memiliki dasar untuk mempelajari kategori.

8. Berdasarkan Cara Evaluasi

Evaluasi pada clustering dilakukan dengan menilai kualitas cluster. Beberapa ukuran yang dapat digunakan antara lain SSE, *elbow method*, silhouette coefficient, Davies-Bouldin Index, dan interpretasi berbasis konteks. Evaluasi clustering menilai apakah kelompok yang terbentuk cukup kompak dan memiliki perbedaan yang jelas antarcluster.

Evaluasi pada classification dilakukan dengan menilai ketepatan model dalam menentukan kelas. Ukuran evaluasi yang umum digunakan antara lain akurasi, presisi, recall, F1-score, confusion matrix, dan ROC curve. Evaluasi classification lebih menekankan seberapa baik model memprediksi kelas yang benar.

Dalam konteks penelitian pendidikan, perbedaan keduanya dapat dilihat melalui contoh berikut. Apabila peneliti memiliki data nilai siswa dan ingin mengetahui kelompok siswa berdasarkan kemiripan nilai tanpa kategori awal, maka metode yang digunakan adalah clustering. K-Means dapat digunakan untuk membagi siswa ke dalam beberapa kelompok. Setelah cluster terbentuk, peneliti menafsirkan kelompok tersebut berdasarkan nilai rata-rata atau centroid.

Namun, apabila peneliti telah memiliki data siswa dengan label "berprestasi" dan "tidak berprestasi", kemudian ingin membangun model untuk memprediksi kategori siswa baru, maka metode yang lebih tepat adalah classification. Dalam kondisi ini, algoritma

seperti Decision Tree, Naïve Bayes, atau Support Vector Machine dapat digunakan.

Perbedaan ini menunjukkan bahwa pemilihan metode tidak boleh hanya didasarkan pada popularitas algoritma. Peneliti harus terlebih dahulu memahami tujuan analisis dan bentuk data yang dimiliki. Jika data belum memiliki label dan tujuan penelitian adalah menemukan kelompok, maka clustering lebih tepat. Jika data sudah memiliki label dan tujuan penelitian adalah menentukan kategori, maka classification lebih sesuai.

3.5 Jenis-Jenis Metode Clustering

Jenis-Jenis Metode Clustering

Metode clustering memiliki berbagai jenis pendekatan yang dapat digunakan sesuai dengan karakteristik data dan tujuan analisis. Setiap metode memiliki cara kerja, kelebihan, dan keterbatasan masing-masing. Oleh karena itu, pemilihan metode clustering tidak boleh dilakukan secara sembarangan, tetapi harus mempertimbangkan bentuk data, jumlah data, ukuran kemiripan, kebutuhan interpretasi, serta hasil yang ingin diperoleh.

Dalam literatur data mining, metode clustering umumnya dikelompokkan ke dalam beberapa jenis, antara lain partitioning method, hierarchical method, density-based method, grid-based method, model-based method, dan fuzzy clustering. Pengelompokan ini menunjukkan bahwa clustering tidak hanya terbatas pada K-Means. K-Means merupakan salah satu metode clustering yang termasuk ke dalam kelompok partitioning method atau metode partisi. Han, Kamber, dan Pei (2011) menjelaskan bahwa metode clustering dapat dibedakan berdasarkan cara pembentukan cluster, seperti metode partisi, hierarki, kepadatan, grid, dan model.

1. Partitioning Method

Partitioning method atau metode partisi merupakan metode clustering yang membagi data ke dalam sejumlah cluster tertentu.

Pada metode ini, jumlah cluster biasanya ditentukan terlebih dahulu sebelum proses clustering dilakukan. Setiap objek data akan ditempatkan ke dalam salah satu cluster berdasarkan ukuran kemiripan atau jarak tertentu. Tujuan utama metode partisi adalah membentuk cluster yang memiliki anggota relatif mirip di dalam kelompoknya dan berbeda dengan kelompok lainnya.

Metode partisi banyak digunakan karena konsepnya sederhana dan mudah dipahami. Salah satu algoritma paling dikenal dalam kelompok ini adalah K-Means. K-Means bekerja dengan cara menentukan jumlah cluster, memilih centroid awal, menghitung jarak setiap data terhadap centroid, menempatkan data pada cluster terdekat, memperbarui centroid, dan mengulangi proses tersebut sampai posisi cluster stabil.

Selain K-Means, metode partisi lain yang sering digunakan adalah K-Medoids. Perbedaan utama K-Means dan K-Medoids terletak pada pusat cluster. K-Means menggunakan centroid berupa rata-rata nilai anggota cluster, sedangkan K-Medoids menggunakan salah satu objek nyata sebagai pusat cluster atau *medoid*. Karena menggunakan objek nyata sebagai pusat cluster, K-Medoids umumnya lebih tahan terhadap data pencilan dibandingkan K-Means.

Ciri utama partitioning method adalah sebagai berikut.

- Jumlah cluster biasanya ditentukan di awal.
- Setiap objek data masuk ke salah satu cluster.
- Proses pengelompokan didasarkan pada ukuran jarak atau kemiripan.
- Cocok untuk data numerik, terutama pada algoritma K-Means.
- Relatif mudah diterapkan pada dataset berukuran sedang hingga besar.

Kelebihan metode partisi adalah prosesnya relatif cepat dan hasilnya mudah dipahami. Namun, metode ini juga memiliki keterbatasan. Pada K-Means, peneliti harus menentukan jumlah cluster terlebih dahulu. Selain itu, hasil K-Means dapat dipengaruhi

oleh centroid awal, skala data, dan keberadaan outlier. Jain (2010) menjelaskan bahwa K-Means merupakan metode clustering yang populer karena sederhana dan efisien, tetapi tetap memiliki tantangan, seperti pemilihan jumlah cluster, ukuran kemiripan, dan validasi hasil cluster.

2. Hierarchical Method

Hierarchical method atau metode hierarki merupakan metode clustering yang membentuk cluster secara bertingkat. Hasil metode ini biasanya ditampilkan dalam bentuk struktur pohon yang disebut dendrogram. Melalui dendrogram, peneliti dapat melihat bagaimana objek data bergabung atau terpisah pada berbagai tingkat kemiripan.

Metode hierarki terbagi menjadi dua pendekatan utama, yaitu agglomerative dan divisive. Pendekatan agglomerative dimulai dari setiap objek sebagai cluster tersendiri, kemudian cluster yang paling mirip digabungkan secara bertahap hingga terbentuk cluster yang lebih besar. Sebaliknya, pendekatan divisive dimulai dari semua objek dalam satu cluster besar, kemudian cluster tersebut dipecah secara bertahap menjadi kelompok yang lebih kecil.

Ciri utama hierarchical method adalah sebagai berikut.

- Membentuk struktur cluster bertingkat.
- Dapat ditampilkan melalui dendrogram.
- Tidak selalu membutuhkan penentuan jumlah cluster di awal.
- Berguna untuk melihat hubungan antarobjek secara bertahap.
- Cocok untuk analisis eksploratif terhadap struktur data.

Metode hierarki memiliki kelebihan karena mampu menunjukkan hubungan antarcluster secara lebih rinci. Peneliti dapat melihat pada tingkat kemiripan tertentu objek-objek mulai bergabung atau terpisah. Namun, metode ini umumnya membutuhkan waktu komputasi lebih besar dibandingkan K-Means, terutama jika jumlah data sangat besar. Selain itu, keputusan penggabungan atau pemisahan pada tahap awal dapat memengaruhi hasil akhir.

Dalam konteks penelitian ilmiah, metode hierarki berguna apabila peneliti ingin melihat struktur hubungan data secara bertingkat. Namun, untuk data yang besar dan membutuhkan proses yang lebih cepat, K-Means sering lebih praktis digunakan. Oleh karena itu, pemilihan antara hierarchical clustering dan K-Means harus disesuaikan dengan kebutuhan analisis.

3. Density-Based Method

Density-based method atau metode berbasis kepadatan merupakan metode clustering yang membentuk cluster berdasarkan kepadatan data pada suatu area. Data yang berada pada wilayah padat akan dikelompokkan sebagai cluster, sedangkan data yang berada pada wilayah jarang dapat dianggap sebagai noise atau data pencilan. Metode ini sangat berguna untuk menemukan cluster dengan bentuk tidak beraturan.

Salah satu algoritma terkenal dalam metode berbasis kepadatan adalah DBSCAN atau *Density-Based Spatial Clustering of Applications with Noise*. DBSCAN dirancang untuk menemukan cluster berdasarkan daerah yang memiliki kepadatan tinggi dan dapat mengenali data pencilan sebagai noise. Ester, Kriegel, Sander, dan Xu (1996) memperkenalkan DBSCAN sebagai algoritma berbasis kepadatan yang dapat menemukan cluster dengan bentuk arbitrer dan menangani noise pada basis data spasial berukuran besar.

Ciri utama density-based method adalah sebagai berikut.

- Membentuk cluster berdasarkan kepadatan titik data.
- Mampu menemukan cluster dengan bentuk tidak beraturan.
- Dapat mengenali noise atau data pencilan.
- Tidak selalu membutuhkan jumlah cluster di awal.
- Cocok untuk data spasial atau data dengan pola kepadatan tertentu.

Kelebihan metode berbasis kepadatan adalah kemampuannya menemukan bentuk cluster yang tidak selalu bulat atau teratur. Hal ini berbeda dengan K-Means yang lebih cocok untuk cluster

berbentuk relatif bulat dan seimbang. Namun, metode berbasis kepadatan juga memiliki keterbatasan, terutama dalam menentukan parameter kepadatan. Jika parameter tidak tepat, hasil cluster dapat menjadi terlalu banyak, terlalu sedikit, atau tidak stabil.

Dalam penelitian berbasis K-Means, pemahaman terhadap metode berbasis kepadatan penting agar peneliti menyadari bahwa K-Means bukan satu-satunya pilihan clustering. Jika data memiliki banyak noise atau bentuk cluster yang tidak beraturan, metode seperti DBSCAN dapat menjadi alternatif yang lebih sesuai.

4. Grid-Based Method

Grid-based method atau metode berbasis grid merupakan metode clustering yang membagi ruang data ke dalam sejumlah sel atau grid. Proses clustering kemudian dilakukan pada struktur grid tersebut, bukan langsung pada setiap objek data. Dengan cara ini, proses analisis dapat menjadi lebih cepat, terutama pada data berukuran besar.

Pada metode berbasis grid, ruang data dibagi menjadi beberapa bagian. Setiap bagian atau sel grid dianalisis berdasarkan jumlah objek atau kepadatan data yang berada di dalamnya. Sel yang memiliki kepadatan tinggi dapat dianggap sebagai bagian dari cluster, sedangkan sel yang memiliki kepadatan rendah dapat diabaikan atau dianggap tidak signifikan.

Ciri utama grid-based method adalah sebagai berikut.

- Membagi ruang data ke dalam bentuk grid atau sel.
- Proses clustering dilakukan pada sel, bukan langsung pada objek individual.
- Cocok untuk dataset besar karena lebih efisien.
- Bergantung pada ukuran dan pembagian grid.
- Sering digunakan pada data spasial atau data multidimensi tertentu.

Kelebihan metode berbasis grid adalah efisiensi komputasi. Karena analisis dilakukan pada struktur grid, proses clustering dapat berlangsung lebih cepat daripada metode yang membandingkan seluruh objek secara langsung. Namun, kualitas hasil sangat dipengaruhi oleh ukuran grid. Jika grid terlalu besar, detail pola data dapat hilang. Jika grid terlalu kecil, proses analisis dapat menjadi lebih kompleks.

5. Model-Based Method

Model-based method merupakan metode clustering yang menganggap data berasal dari model atau distribusi tertentu. Dalam metode ini, cluster dibentuk berdasarkan asumsi bahwa data dihasilkan oleh beberapa model probabilistik. Salah satu pendekatan yang sering digunakan adalah mixture model, misalnya *Gaussian Mixture Model*.

Berbeda dengan K-Means yang menempatkan data ke dalam cluster berdasarkan jarak terdekat terhadap centroid, model-based clustering memperkirakan kemungkinan suatu data menjadi bagian dari cluster tertentu berdasarkan model statistik. Dengan demikian, metode ini dapat memberikan informasi probabilitas keanggotaan data terhadap cluster.

Ciri utama model-based method adalah sebagai berikut.

- Menggunakan asumsi model statistik atau distribusi data.
- Cluster dibentuk berdasarkan probabilitas.
- Dapat memberikan peluang keanggotaan data pada cluster.
- Cocok untuk data yang mengikuti pola distribusi tertentu.
- Membutuhkan pemahaman statistik yang lebih kuat.

Kelebihan metode berbasis model adalah kemampuannya memberikan dasar probabilistik terhadap pembentukan cluster. Namun, metode ini juga memiliki keterbatasan karena hasilnya sangat bergantung pada kesesuaian asumsi model dengan data. Jika asumsi distribusi tidak sesuai, hasil cluster dapat menjadi kurang tepat.

6. Fuzzy Clustering

Fuzzy clustering merupakan metode clustering yang memungkinkan satu objek data memiliki derajat keanggotaan pada lebih dari satu cluster. Hal ini berbeda dengan K-Means yang bersifat tegas atau *hard clustering*, yaitu setiap data hanya masuk ke satu cluster. Dalam fuzzy clustering, satu data dapat memiliki tingkat keanggotaan berbeda pada beberapa cluster.

Algoritma yang paling dikenal dalam kelompok ini adalah Fuzzy C-Means. Pada Fuzzy C-Means, setiap data memiliki nilai keanggotaan terhadap setiap cluster. Misalnya, satu data dapat memiliki keanggotaan 0,70 pada Cluster 1 dan 0,30 pada Cluster 2. Hal ini berguna apabila batas antarcluster tidak terlalu jelas.

Ciri utama fuzzy clustering adalah sebagai berikut.

- Data dapat memiliki derajat keanggotaan pada lebih dari satu cluster.
- Cocok untuk data dengan batas kelompok yang tidak tegas.
- Memberikan informasi keanggotaan secara lebih fleksibel.
- Membutuhkan interpretasi yang lebih hati-hati.
- Berbeda dengan K-Means yang menempatkan data secara tegas pada satu cluster.

Fuzzy clustering bermanfaat ketika objek data tidak dapat dipisahkan secara jelas. Dalam beberapa kasus, data memang berada di antara dua kelompok. Dengan fuzzy clustering, kondisi tersebut dapat direpresentasikan melalui derajat keanggotaan. Namun, metode ini membutuhkan penafsiran yang lebih kompleks dibandingkan K-Means.

7. Constraint-Based Clustering

Constraint-based clustering merupakan metode clustering yang menggunakan batasan atau aturan tertentu dalam proses pembentukan cluster. Batasan tersebut dapat berasal dari kebutuhan pengguna, karakteristik domain, atau aturan penelitian. Misalnya, peneliti dapat menentukan bahwa setiap cluster harus

memiliki jumlah anggota tertentu, jarak maksimum tertentu, atau atribut tertentu yang harus dipertimbangkan.

Metode ini berguna apabila proses clustering tidak hanya bergantung pada kemiripan data, tetapi juga harus memenuhi persyaratan tertentu. Dalam konteks layanan publik, misalnya, pengelompokan wilayah mungkin tidak hanya mempertimbangkan kemiripan indikator sosial, tetapi juga batas geografis atau kapasitas layanan. Dengan demikian, constraint-based clustering memungkinkan proses clustering lebih sesuai dengan kebutuhan praktis.

Ciri utama constraint-based clustering adalah sebagai berikut.

- Menggunakan batasan tertentu dalam pembentukan cluster.
- Batasan dapat berasal dari pengguna atau konteks penelitian.
- Membantu menghasilkan cluster yang sesuai kebutuhan praktis.
- Cocok untuk kasus yang memiliki aturan khusus.
- Membutuhkan definisi batasan yang jelas sejak awal.

Metode ini lebih kompleks karena tidak hanya mempertimbangkan jarak atau kemiripan, tetapi juga aturan tambahan. Oleh karena itu, peneliti perlu menjelaskan alasan penggunaan batasan dan dampaknya terhadap hasil clustering.

Dari berbagai jenis metode tersebut, K-Means termasuk dalam partitioning method. K-Means banyak digunakan karena memiliki alur kerja yang relatif sederhana dan mudah diterapkan dalam berbagai bidang. Namun, K-Means bukan metode yang selalu paling tepat untuk semua jenis data. K-Means lebih sesuai digunakan apabila data berbentuk numerik, jumlah cluster dapat ditentukan, dan bentuk cluster relatif kompak. Jika data memiliki banyak noise, bentuk cluster tidak beraturan, atau batas antarcluster tidak jelas, maka metode lain dapat dipertimbangkan.

Pemilihan metode clustering dapat dipandu oleh beberapa pertanyaan berikut.

- Apakah data berbentuk numerik atau kategorik?
- Apakah jumlah cluster sudah dapat ditentukan?
- Apakah data memiliki banyak outlier?
- Apakah bentuk cluster diperkirakan teratur atau tidak beraturan?
- Apakah peneliti membutuhkan hasil yang mudah dijelaskan?
- Apakah data berukuran sangat besar?
- Apakah ada batasan khusus dalam pembentukan cluster?

Dalam penelitian ilmiah, jawaban atas pertanyaan tersebut penting untuk menentukan metode yang tepat. Jika tujuan penelitian adalah mengelompokkan data numerik secara sederhana dan interpretatif, K-Means dapat menjadi pilihan yang sesuai. Namun, jika data memiliki karakteristik khusus, peneliti perlu mempertimbangkan metode lain agar hasil clustering lebih tepat.

Dengan demikian, jenis-jenis metode clustering menunjukkan bahwa proses pengelompokan data dapat dilakukan melalui berbagai pendekatan. Setiap metode memiliki prinsip kerja dan kegunaan yang berbeda. Partitioning method, hierarchical method, density-based method, grid-based method, model-based method, fuzzy clustering, dan constraint-based clustering dapat digunakan sesuai dengan kebutuhan analisis. Dalam buku ini, pembahasan difokuskan pada K-Means sebagai salah satu metode partisi yang banyak digunakan dalam data mining. Pemahaman terhadap metode clustering lainnya tetap diperlukan agar pembaca dapat melihat posisi K-Means secara tepat dan memahami bahwa pemilihan metode harus disesuaikan dengan karakteristik data serta tujuan penelitian.

3.6 Kelebihan dan Kekurangan Clustering

Saya lanjutkan dengan subbab Kelebihan dan Kekurangan Clustering secara seimbang: tidak hanya memuji manfaatnya,

tetapi juga menjelaskan batasan metodologis yang perlu diperhatikan dalam penelitian K-Means.

Thought for 7s

Kelebihan dan Kekurangan Clustering

Clustering merupakan salah satu teknik data mining yang banyak digunakan karena mampu membantu peneliti menemukan struktur kelompok dalam data tanpa memerlukan label kelas sebelumnya. Teknik ini sangat berguna ketika peneliti belum mengetahui kategori data secara pasti, tetapi ingin melihat apakah data memiliki pola kemiripan tertentu. Meskipun demikian, clustering juga memiliki keterbatasan yang perlu dipahami agar hasil analisis tidak ditafsirkan secara berlebihan. Dalam penelitian ilmiah, pemahaman terhadap kelebihan dan kekurangan clustering penting agar metode ini digunakan secara tepat sesuai karakteristik data dan tujuan penelitian.

1. Kelebihan Clustering

Salah satu kelebihan utama clustering adalah kemampuannya menemukan kelompok data secara eksploratif. Clustering dapat digunakan ketika data belum memiliki kategori awal. Dalam kondisi tersebut, peneliti tidak perlu menentukan label terlebih dahulu, karena algoritma akan membentuk kelompok berdasarkan kemiripan antarobjek. Hal ini menjadikan clustering sangat berguna untuk penelitian yang bertujuan menemukan pola awal dalam data.

Kelebihan clustering dapat dijelaskan sebagai berikut.

a. Tidak Memerlukan Label Data Awal

Clustering termasuk teknik *unsupervised learning*, sehingga tidak membutuhkan data yang telah diberi label. Hal ini berbeda dengan classification yang memerlukan kelas atau kategori pada data pelatihan. Dalam clustering, data dikelompokkan berdasarkan kemiripan karakteristik yang terdapat pada atribut yang digunakan. Menurut Han, Kamber, dan Pei (2011), clustering

digunakan untuk mengelompokkan objek data tanpa mengetahui label kelas sebelumnya, sehingga metode ini sesuai untuk menemukan struktur tersembunyi dalam data.

Keunggulan ini sangat membantu ketika peneliti memiliki data mentah yang belum dikategorikan. Misalnya, dalam bidang pendidikan, data nilai siswa dapat langsung dikelompokkan berdasarkan kemiripan nilai tanpa harus terlebih dahulu menentukan siswa mana yang termasuk kategori tinggi, sedang, atau rendah. Label tersebut baru diberikan setelah hasil cluster dianalisis dan ditafsirkan.

b. Mampu Menemukan Pola Tersembunyi dalam Data

Clustering dapat membantu menemukan pola yang tidak mudah terlihat melalui pengamatan manual. Pada dataset yang besar, peneliti sering kali sulit membaca kecenderungan data secara langsung. Clustering membantu menyusun data menjadi beberapa kelompok sehingga pola kemiripan dan perbedaan antarobjek lebih mudah dipahami.

Sebagai contoh, data pelanggan yang terdiri atas jumlah transaksi, frekuensi pembelian, dan nilai belanja dapat dikelompokkan menjadi beberapa segmen. Dari hasil tersebut, perusahaan dapat mengetahui kelompok pelanggan aktif, pelanggan sedang, dan pelanggan pasif. Dalam konteks pendidikan, clustering dapat membantu menemukan kelompok siswa berdasarkan capaian akademik atau kemampuan awal.

c. Membantu Menyederhanakan Data yang Kompleks

Data yang besar dan memiliki banyak atribut sering kali sulit ditafsirkan secara langsung. Clustering membantu menyederhanakan kompleksitas tersebut dengan membagi data ke dalam kelompok-kelompok yang lebih kecil. Dengan demikian, peneliti dapat memahami data melalui karakteristik setiap cluster, bukan membaca setiap objek satu per satu.

Penyederhanaan ini sangat bermanfaat dalam penelitian yang melibatkan banyak objek data. Misalnya, ratusan data siswa dapat diringkas menjadi tiga kelompok berdasarkan kemiripan nilai akademik. Hasil ini memudahkan peneliti atau pengambil keputusan untuk memahami kondisi umum data secara lebih terstruktur.

d. Berguna untuk Segmentasi Data

Clustering sangat kuat dalam proses segmentasi. Segmentasi merupakan proses membagi objek data ke dalam kelompok tertentu berdasarkan karakteristik yang serupa. Dalam berbagai bidang, segmentasi menjadi dasar penting untuk menyusun strategi atau rekomendasi yang lebih tepat.

Contoh penerapan segmentasi melalui clustering antara lain:

- mengelompokkan siswa berdasarkan prestasi akademik;
- mengelompokkan pelanggan berdasarkan pola pembelian;
- mengelompokkan pasien berdasarkan karakteristik kesehatan;
- mengelompokkan wilayah berdasarkan indikator sosial ekonomi;
- mengelompokkan penerima bantuan berdasarkan tingkat kebutuhan.

Dalam buku ini, segmentasi menjadi salah satu alasan utama penggunaan algoritma K-Means. K-Means dapat membantu membagi data ke dalam beberapa cluster sehingga setiap kelompok dapat dianalisis berdasarkan karakteristiknya.

e. Dapat Digunakan pada Berbagai Bidang

Clustering memiliki cakupan penerapan yang luas. Teknik ini dapat digunakan dalam bidang pendidikan, bisnis, kesehatan, pemerintahan, industri, perpustakaan, hingga media digital. Keluasan penerapan ini terjadi karena banyak bidang memiliki kebutuhan untuk memahami kelompok atau segmen dari data yang tersedia.

Dalam pendidikan, clustering dapat digunakan untuk mengelompokkan peserta didik. Dalam bisnis, clustering dapat digunakan untuk segmentasi pelanggan. Dalam pemerintahan, clustering dapat digunakan untuk mengelompokkan wilayah atau masyarakat berdasarkan indikator tertentu. Hal ini menunjukkan bahwa clustering bersifat fleksibel dan dapat disesuaikan dengan berbagai kebutuhan analisis.

f. Mendukung Pengambilan Keputusan Berbasis Data

Hasil clustering dapat menjadi dasar dalam proses pengambilan keputusan. Setelah data dikelompokkan, pengambil keputusan dapat menyusun strategi yang berbeda untuk setiap kelompok. Keputusan yang diambil tidak lagi hanya berdasarkan intuisi, tetapi didukung oleh pola yang ditemukan dari data.

Misalnya, jika hasil clustering menunjukkan adanya kelompok siswa dengan capaian akademik rendah, sekolah dapat merancang program pendampingan khusus. Jika terdapat kelompok pelanggan dengan nilai transaksi tinggi, perusahaan dapat menyusun strategi loyalitas pelanggan. Dengan demikian, clustering membantu mengarahkan keputusan agar lebih sesuai dengan karakteristik kelompok.

g. Membantu Analisis Lanjutan

Hasil clustering dapat digunakan sebagai dasar untuk analisis berikutnya. Cluster yang terbentuk dapat dianalisis lebih lanjut untuk mengetahui karakteristik dominan, faktor pembeda antar kelompok, atau kebutuhan setiap kelompok. Dalam beberapa penelitian, hasil clustering juga dapat digunakan sebagai variabel tambahan untuk analisis berikutnya.

Sebagai contoh, setelah siswa dikelompokkan menggunakan K-Means, peneliti dapat membandingkan karakteristik setiap cluster berdasarkan nilai rata-rata, kehadiran, atau faktor lain. Dengan demikian, clustering dapat menjadi tahap awal untuk memahami data secara lebih mendalam.

2. Kekurangan Clustering

Meskipun memiliki banyak kelebihan, clustering juga memiliki keterbatasan. Kekurangan ini perlu dipahami agar peneliti tidak menganggap hasil clustering sebagai kebenaran mutlak. Clustering menghasilkan kelompok berdasarkan atribut, ukuran kemiripan, dan algoritma yang digunakan. Jika salah satu unsur tersebut kurang tepat, hasil cluster dapat menjadi kurang bermakna.

a. Hasil Sangat Bergantung pada Atribut yang Digunakan

Clustering sangat dipengaruhi oleh atribut yang dipilih. Atribut menjadi dasar bagi algoritma untuk menghitung kemiripan antarobjek. Jika atribut yang digunakan tidak relevan dengan tujuan penelitian, maka cluster yang terbentuk dapat sulit ditafsirkan.

Sebagai contoh, apabila tujuan penelitian adalah mengelompokkan siswa berdasarkan prestasi akademik, atribut yang digunakan seharusnya berkaitan dengan nilai atau capaian belajar. Jika atribut administratif seperti nomor induk, alamat, atau nama dimasukkan ke dalam perhitungan, maka hasil cluster tidak menggambarkan prestasi akademik secara tepat.

b. Sensitif terhadap Skala Data

Banyak metode clustering berbasis jarak, termasuk K-Means, sangat sensitif terhadap skala data. Atribut dengan rentang nilai yang besar dapat mendominasi perhitungan jarak. Akibatnya, hasil clustering dapat lebih dipengaruhi oleh atribut tertentu, sementara atribut lain menjadi kurang berpengaruh.

Misalnya, jika atribut nilai akademik memiliki rentang 0–100, sedangkan atribut pendapatan memiliki rentang jutaan rupiah, maka atribut pendapatan dapat mendominasi hasil perhitungan jarak. Oleh karena itu, normalisasi atau standardisasi sering diperlukan sebelum clustering dilakukan. Jain (2010) menegaskan bahwa representasi data dan ukuran kedekatan sangat menentukan kualitas hasil clustering.

c. Memerlukan Penentuan Jumlah Cluster pada Metode Tertentu

Beberapa metode clustering, seperti K-Means, memerlukan penentuan jumlah cluster sejak awal. Peneliti harus menentukan nilai *k* sebelum algoritma dijalankan. Penentuan jumlah cluster ini tidak selalu mudah karena jumlah kelompok yang paling tepat belum tentu diketahui sebelumnya.

Jika jumlah cluster terlalu sedikit, data yang sebenarnya berbeda dapat tergabung dalam kelompok yang sama. Sebaliknya, jika jumlah cluster terlalu banyak, hasil pengelompokan dapat menjadi terlalu terpecah dan sulit diinterpretasikan. Oleh karena itu, penentuan jumlah cluster perlu didukung oleh pertimbangan teoritis, tujuan penelitian, dan metode evaluasi seperti *elbow method* atau silhouette coefficient.

d. Sensitif terhadap Data Pencilan

Clustering, terutama K-Means, dapat dipengaruhi oleh data pencilan atau outlier. Outlier merupakan data yang nilainya jauh berbeda dari pola umum. Karena K-Means menggunakan centroid yang dihitung berdasarkan rata-rata, keberadaan outlier dapat menggeser posisi centroid dan memengaruhi hasil cluster.

Sebagai contoh, jika terdapat satu data siswa dengan nilai yang sangat ekstrem akibat kesalahan input, data tersebut dapat memengaruhi pembentukan cluster. Oleh sebab itu, data pencilan harus diperiksa sebelum proses clustering dilakukan. Outlier tidak selalu harus dihapus, tetapi perlu dianalisis apakah data tersebut merupakan kesalahan atau fenomena yang memang valid.

e. Hasil Cluster Tidak Selalu Mudah Ditafsirkan

Algoritma clustering hanya menghasilkan kelompok, tetapi tidak memberikan makna kelompok secara otomatis. Peneliti harus menafsirkan hasil cluster berdasarkan karakteristik data. Jika perbedaan antarcluster tidak jelas, maka hasil pengelompokan menjadi sulit dijelaskan.

Misalnya, K-Means dapat menghasilkan Cluster 1, Cluster 2, dan Cluster 3. Namun, nama cluster tersebut belum memiliki makna substantif. Peneliti perlu membaca nilai centroid, rata-rata atribut, dan karakteristik anggota cluster sebelum memberi label seperti "tinggi", "sedang", atau "rendah". Jika label diberikan tanpa dasar data, interpretasi hasil dapat menjadi keliru.

f. Tidak Menunjukkan Hubungan Sebab-Akibat

Clustering hanya menunjukkan pengelompokan berdasarkan kemiripan data. Teknik ini tidak dapat digunakan untuk membuktikan hubungan sebab-akibat antarvariabel. Kesalahan umum dalam penelitian adalah menganggap bahwa terbentuknya suatu cluster disebabkan oleh atribut tertentu, padahal clustering hanya menunjukkan kedekatan karakteristik berdasarkan data yang dianalisis.

Sebagai contoh, apabila siswa dengan nilai rendah berada dalam satu cluster, peneliti tidak dapat langsung menyimpulkan penyebab rendahnya nilai tanpa analisis tambahan. Clustering hanya menunjukkan bahwa siswa dalam kelompok tersebut memiliki karakteristik nilai yang mirip. Untuk mengetahui penyebabnya, diperlukan metode analisis lain atau kajian tambahan.

g. Kualitas Hasil Bergantung pada Kualitas Data

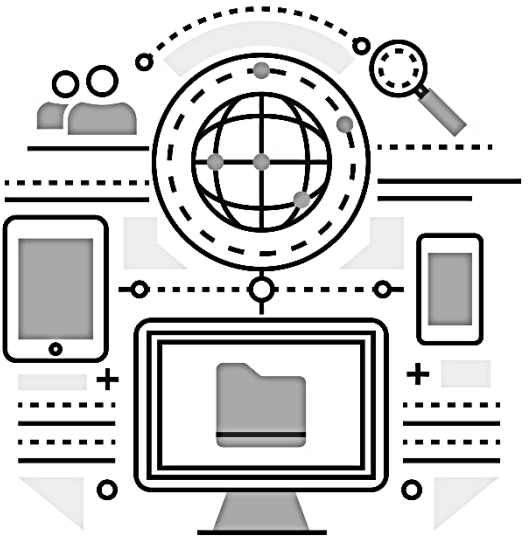
Clustering sangat bergantung pada kualitas data. Data yang tidak lengkap, tidak konsisten, mengandung kesalahan, atau tidak relevan dapat menghasilkan cluster yang tidak valid. Oleh karena itu, tahapan seleksi data, pembersihan data, dan transformasi data sangat penting sebelum clustering dilakukan.

Dalam proses KDD, clustering tidak boleh langsung dilakukan tanpa persiapan data yang baik. Data yang buruk akan menghasilkan pola yang buruk pula. Prinsip ini sering disebut dengan ungkapan *garbage in, garbage out*, yaitu apabila data yang dimasukkan tidak berkualitas, maka hasil yang keluar juga tidak berkualitas.

h. Tidak Semua Bentuk Data Cocok untuk Semua Metode Clustering

Setiap metode clustering memiliki karakteristik tertentu. K-Means lebih cocok untuk data numerik dan cluster yang relatif kompak. Metode ini kurang sesuai untuk data yang memiliki banyak noise, bentuk cluster tidak beraturan, atau data yang sebagian besar bersifat kategorik. Dalam kondisi seperti itu, metode lain seperti DBSCAN, hierarchical clustering, K-Medoids, atau Fuzzy C-Means mungkin lebih sesuai.

Oleh karena itu, peneliti perlu memahami karakteristik metode sebelum menggunakannya. Pemilihan metode yang tidak sesuai dapat menyebabkan hasil pengelompokan tidak bermakna atau sulit dipertanggungjawabkan.



Bab 04

ALGORITMA K-MEANS DALAM *DATA MINING*

4.1 Pengertian Algoritma K-Means

Algoritma K-Means merupakan salah satu metode clustering yang digunakan untuk mengelompokkan data ke dalam sejumlah kelompok atau cluster berdasarkan kemiripan karakteristik. Algoritma ini bekerja dengan cara membagi data ke dalam k cluster, kemudian menempatkan setiap objek data pada cluster yang memiliki pusat kelompok atau centroid terdekat. Dalam konteks data mining, K-Means termasuk ke dalam metode unsupervised learning karena proses pengelompokan dilakukan tanpa menggunakan label kelas awal.

Secara historis, istilah K-Means dikenal melalui karya MacQueen pada tahun 1967 yang membahas proses pemartisian populasi berdimensi banyak ke dalam sejumlah kelompok berdasarkan sampel data. Gagasan ini kemudian berkembang luas dalam kajian

clustering karena memiliki konsep yang sederhana, efisien, dan mudah diterapkan pada berbagai bentuk data numerik. Selain itu, algoritma yang berkaitan erat dengan prosedur K-Means juga dijelaskan dalam karya Lloyd yang dipublikasikan pada tahun 1982 melalui kajian mengenai least squares quantization. Kedua rujukan tersebut menjadi dasar penting dalam perkembangan algoritma K-Means sebagai salah satu metode clustering yang banyak digunakan (Lloyd, 1982; MacQueen, 1967).

K-Means disebut sebagai metode clustering berbasis partisi karena algoritma ini membagi data ke dalam beberapa kelompok yang telah ditentukan jumlahnya. Nilai k menunjukkan jumlah cluster yang ingin dibentuk. Jika peneliti menentukan nilai k = 3, maka data akan dibagi ke dalam tiga cluster. Jumlah cluster ini tidak ditentukan secara otomatis oleh algoritma, tetapi harus ditentukan oleh peneliti berdasarkan tujuan analisis, karakteristik data, atau metode evaluasi tertentu seperti elbow method dan silhouette coefficient.

Prinsip dasar K-Means adalah menempatkan objek data ke dalam cluster yang memiliki jarak paling dekat dengan centroid. Centroid merupakan titik pusat dari suatu cluster yang dihitung berdasarkan nilai rata-rata atribut dari anggota cluster tersebut. Setelah objek data dikelompokkan, centroid diperbarui kembali berdasarkan anggota cluster yang baru. Proses ini dilakukan secara berulang sampai tidak terjadi perubahan berarti pada anggota cluster atau posisi centroid. Lloyd (1982) menjelaskan pendekatan kuantisasi berbasis kuadrat terkecil yang menjadi dasar prosedur iteratif seperti K-Means, yaitu memperbarui pusat representasi berdasarkan pembagian wilayah data.

Secara matematis, tujuan K-Means adalah meminimalkan jumlah kuadrat jarak antara setiap data dan centroid cluster-nya. Ukuran ini sering disebut sebagai within-cluster sum of squares atau SSE. Semakin kecil jarak antara data dan centroid dalam satu cluster, semakin kompak cluster tersebut.

$$J = \Sigma_{i=1}^{k} \Sigma_{x \in C_i} ||x - \mu_i||^2$$

Keterangan:

Simbol	Keterangan
J	Nilai fungsi objektif atau jumlah kuadrat jarak dalam cluster
k	Jumlah cluster
C_i	Cluster ke-i
x	Objek data dalam cluster
μ_i	Centroid atau pusat cluster ke-i
$\|\|x - \mu_i\|\|^2$	Kuadrat jarak antara data dan centroid

Dari rumus tersebut dapat dipahami bahwa K-Means berusaha membentuk cluster yang anggotanya memiliki kedekatan tinggi terhadap centroid masing-masing. Dengan kata lain, algoritma ini berupaya membuat objek dalam satu cluster menjadi sehomogen mungkin berdasarkan atribut yang digunakan. Jain (2010) menjelaskan bahwa K-Means merupakan salah satu algoritma clustering yang paling dikenal karena kesederhanaan dan efisiensinya, meskipun kualitas hasilnya tetap dipengaruhi oleh pemilihan jumlah cluster, pusat awal, ukuran jarak, dan karakteristik data.

Dalam penerapannya, K-Means umumnya menggunakan ukuran jarak Euclidean untuk menghitung kedekatan antara objek data dan centroid. Jarak Euclidean digunakan untuk mengukur seberapa jauh posisi suatu data dari pusat cluster dalam ruang berdimensi banyak. Semakin kecil jarak suatu data terhadap centroid, semakin besar kemungkinan data tersebut menjadi anggota cluster tersebut. Oleh karena itu, K-Means lebih sesuai digunakan pada data numerik atau data yang telah ditransformasikan ke dalam bentuk numerik.

K-Means termasuk ke dalam metode hard clustering. Artinya, setiap objek data hanya menjadi anggota dari satu cluster. Jika suatu data telah ditempatkan pada Cluster 1, maka data tersebut tidak menjadi anggota Cluster 2 atau Cluster 3. Hal ini berbeda dengan fuzzy clustering yang memungkinkan satu objek memiliki derajat keanggotaan pada lebih dari satu cluster. Sifat hard clustering membuat hasil K-Means lebih mudah dipahami, tetapi juga menuntut kehati-hatian dalam interpretasi, terutama apabila terdapat data yang posisinya berada di antara dua cluster.

Secara sederhana, cara kerja K-Means dapat dipahami melalui tahapan berikut.

- Peneliti menentukan jumlah cluster atau nilai k.
- Algoritma menentukan centroid awal.
- Setiap data dihitung jaraknya terhadap masing-masing centroid.
- Data ditempatkan pada cluster dengan centroid terdekat.
- Centroid diperbarui berdasarkan rata-rata anggota cluster.
- Proses pengelompokan dan pembaruan centroid diulang sampai cluster stabil.

Tahapan tersebut menunjukkan bahwa K-Means bekerja secara iteratif. Proses tidak berhenti pada pembentukan cluster pertama, tetapi terus diperbarui sampai posisi cluster tidak banyak berubah. Dengan mekanisme ini, K-Means berusaha menemukan pembagian kelompok yang paling sesuai berdasarkan kedekatan data terhadap centroid.

Dalam penelitian data mining, K-Means banyak digunakan karena memiliki alur yang mudah dipahami dan dapat diterapkan pada berbagai bidang. Dalam bidang pendidikan, K-Means dapat digunakan untuk mengelompokkan siswa berdasarkan nilai akademik. Dalam bidang bisnis, K-Means dapat digunakan untuk segmentasi pelanggan. Dalam bidang pemerintahan, K-Means dapat digunakan untuk mengelompokkan masyarakat atau wilayah berdasarkan indikator sosial ekonomi. Dalam bidang kesehatan, K-

Means dapat digunakan untuk mengelompokkan pasien berdasarkan karakteristik tertentu.

Namun, penggunaan K-Means harus dilakukan dengan pemahaman metodologis yang tepat. K-Means bukan algoritma untuk memprediksi kelas yang telah diketahui, melainkan algoritma untuk menemukan kelompok berdasarkan kemiripan data. Oleh karena itu, hasil K-Means tidak boleh diperlakukan sebagai kategori mutlak sebelum dievaluasi dan ditafsirkan. Cluster yang terbentuk merupakan hasil perhitungan berdasarkan atribut, jumlah cluster, skala data, dan centroid yang digunakan.

Sebagai contoh, apabila K-Means digunakan untuk mengelompokkan siswa berdasarkan nilai Matematika, Bahasa Indonesia, dan IPA, maka algoritma akan membentuk kelompok siswa berdasarkan kemiripan nilai pada ketiga atribut tersebut. Hasilnya dapat berupa Cluster 1, Cluster 2, dan Cluster 3. Setelah itu, peneliti membaca nilai centroid atau rata-rata setiap cluster untuk memberi makna, misalnya kelompok prestasi tinggi, sedang, dan rendah. Label tersebut bukan hasil otomatis dari algoritma, melainkan hasil interpretasi peneliti berdasarkan karakteristik cluster.

Komponen Utama K-Means	Penjelasan
Data	Objek yang akan dikelompokkan
Atribut	Variabel yang menjadi dasar perhitungan kemiripan
Nilai k	Jumlah cluster yang ditentukan di awal
Centroid	Titik pusat setiap cluster
Jarak	Ukuran kedekatan data terhadap centroid
Iterasi	Proses pengulangan sampai hasil cluster stabil
Cluster	Kelompok data yang terbentuk berdasarkan kemiripan

Kelebihan utama K-Means terletak pada kesederhanaan konsep dan efisiensi prosesnya. Algoritma ini relatif mudah dipahami oleh peneliti pemula dan dapat dijalankan menggunakan berbagai perangkat lunak data mining, statistik, maupun bahasa pemrograman. Meskipun demikian, K-Means juga memiliki keterbatasan. Algoritma ini sensitif terhadap penentuan jumlah cluster, centroid awal, data pencilan, dan perbedaan skala antaratribut. Karena itu, sebelum K-Means diterapkan, data perlu melalui proses seleksi, pembersihan, dan transformasi yang baik.

4.2 Konsep Dasar K-Means

Konsep dasar K-Means berangkat dari gagasan bahwa sekumpulan data dapat dikelompokkan ke dalam beberapa kelompok berdasarkan tingkat kemiripan antarobjek. Objek yang memiliki karakteristik relatif sama akan ditempatkan dalam kelompok yang sama, sedangkan objek yang memiliki karakteristik berbeda akan ditempatkan pada kelompok lain. Dalam konteks data mining, K-Means digunakan untuk menemukan pola pengelompokan data tanpa memerlukan label kelas sebelumnya.

K-Means termasuk dalam metode clustering berbasis partisi. Artinya, data dibagi ke dalam sejumlah cluster yang telah ditentukan sebelumnya. Jumlah cluster tersebut dinyatakan dengan simbol k. Jika nilai k ditentukan sebanyak 3, maka algoritma akan membagi data ke dalam tiga cluster. Setiap cluster memiliki titik pusat yang disebut centroid. Centroid berfungsi sebagai pusat representasi dari setiap cluster dan menjadi dasar dalam menentukan kedekatan suatu data terhadap kelompok tertentu.

Konsep utama dalam K-Means dapat dipahami melalui tiga unsur penting, yaitu jumlah cluster, centroid, dan jarak data terhadap centroid. Ketiga unsur ini menjadi dasar kerja algoritma K-Means dalam membentuk kelompok data.

1. Jumlah Cluster

Jumlah cluster merupakan banyaknya kelompok yang ingin dibentuk dalam proses K-Means. Jumlah cluster dinyatakan dengan simbol k. Penentuan nilai k dilakukan sebelum algoritma dijalankan. Hal ini menjadi salah satu ciri penting K-Means karena algoritma tidak secara otomatis menentukan jumlah cluster terbaik.

Sebagai contoh, jika peneliti ingin mengelompokkan data siswa ke dalam tiga kelompok prestasi, maka nilai k dapat ditentukan sebesar 3. Hasil pengelompokan dapat berupa Cluster 1, Cluster 2, dan Cluster 3. Setelah proses clustering selesai, peneliti dapat menafsirkan karakteristik masing-masing cluster, misalnya sebagai kelompok prestasi tinggi, sedang, dan rendah.

Penentuan jumlah cluster perlu dilakukan secara hati-hati. Jika jumlah cluster terlalu sedikit, data yang sebenarnya berbeda dapat tergabung dalam satu kelompok. Sebaliknya, jika jumlah cluster terlalu banyak, hasil pengelompokan dapat menjadi terlalu terpecah dan sulit ditafsirkan. Oleh sebab itu, nilai k sebaiknya ditentukan berdasarkan tujuan penelitian, karakteristik data, atau metode evaluasi tertentu seperti elbow method dan silhouette coefficient.

2. Centroid sebagai Pusat Cluster

Centroid merupakan titik pusat dari suatu cluster. Dalam K-Means, centroid dihitung berdasarkan nilai rata-rata dari seluruh anggota cluster. Centroid bukan selalu merupakan salah satu data asli, melainkan titik rata-rata yang mewakili posisi umum anggota cluster. Oleh karena itu, centroid dapat berubah selama proses iterasi berlangsung.

Pada tahap awal, centroid dapat dipilih secara acak atau menggunakan metode tertentu. Setelah data dikelompokkan berdasarkan centroid terdekat, nilai centroid diperbarui kembali berdasarkan anggota cluster yang baru terbentuk. Proses

pembaruan ini dilakukan berulang sampai posisi centroid relatif stabil.

Secara sederhana, centroid dapat dipahami sebagai pusat kecenderungan dari suatu kelompok. Jika suatu cluster berisi siswa dengan nilai akademik tinggi, maka centroid cluster tersebut akan menunjukkan nilai rata-rata yang juga relatif tinggi. Sebaliknya, jika suatu cluster berisi siswa dengan nilai rendah, maka centroid-nya akan berada pada posisi nilai yang lebih rendah.

3. Jarak Data terhadap Centroid

K-Means mengelompokkan data berdasarkan jarak terdekat antara objek data dan centroid. Data akan dimasukkan ke dalam cluster yang memiliki centroid paling dekat. Ukuran jarak yang umum digunakan dalam K-Means adalah Euclidean distance.

Rumus jarak Euclidean antara dua titik dapat dituliskan sebagai berikut.

$$d(x,y) = \sqrt{(\Sigma_{i=1}^{n} (x_i - y_i)^2)}$$

Simbol	Keterangan
d(x,y)	Jarak antara data x dan y
x_i	Nilai atribut ke-i pada data x
y_i	Nilai atribut ke-i pada data y
n	Jumlah atribut yang digunakan

Dalam K-Means, rumus ini digunakan untuk menghitung jarak antara data dan centroid. Data yang memiliki jarak paling kecil terhadap suatu centroid akan menjadi anggota cluster tersebut. Semakin kecil jarak data terhadap centroid, semakin tinggi tingkat kemiripan data tersebut dengan karakteristik cluster.

4. Mekanisme Dasar K-Means

Secara umum, mekanisme dasar K-Means terdiri atas beberapa tahap. Tahap pertama adalah menentukan jumlah cluster atau nilai k. Tahap kedua adalah menentukan centroid awal. Tahap ketiga adalah menghitung jarak setiap data terhadap seluruh centroid. Tahap keempat adalah menempatkan data ke dalam cluster dengan centroid terdekat. Tahap kelima adalah memperbarui centroid berdasarkan anggota cluster yang baru. Tahap keenam adalah mengulangi proses tersebut sampai tidak terjadi perubahan cluster atau perubahan centroid sudah sangat kecil.

Tahapan dasar K-Means dapat diringkas sebagai berikut.

Tahap	Proses	Penjelasan
1	Menentukan nilai k	Menentukan jumlah cluster yang akan dibentuk
2	Menentukan centroid awal	Memilih titik pusat awal untuk setiap cluster
3	Menghitung jarak	Mengukur jarak setiap data terhadap centroid
4	Membentuk cluster	Menempatkan data pada centroid terdekat
5	Memperbarui centroid	Menghitung ulang pusat cluster berdasarkan anggota baru
6	Melakukan iterasi	Mengulangi proses sampai cluster stabil

Proses tersebut menunjukkan bahwa K-Means bekerja secara berulang atau iteratif. Iterasi dilakukan karena posisi centroid dapat berubah setelah anggota cluster diperbarui. Ketika centroid

berubah, jarak data terhadap centroid juga dapat berubah. Akibatnya, beberapa data mungkin berpindah dari satu cluster ke cluster lain. Proses ini terus berlangsung sampai hasil pengelompokan tidak berubah secara signifikan.

5. Fungsi Objektif K-Means

Konsep dasar K-Means juga berkaitan dengan fungsi objektif yang ingin dicapai oleh algoritma. K-Means bertujuan meminimalkan jumlah jarak kuadrat antara data dan centroid dalam setiap cluster. Dengan kata lain, algoritma berusaha membentuk cluster yang anggotanya berada sedekat mungkin dengan pusat cluster.

Fungsi objektif K-Means dapat dituliskan sebagai berikut.

$$J = \Sigma_{i=1}^{k} \Sigma_{x \in C_i} ||x - \mu_i||^2$$

Simbol	Keterangan
J	Nilai fungsi objektif
k	Jumlah cluster
C_i	Cluster ke-i
x	Data yang menjadi anggota cluster
μ_i	Centroid cluster ke-i
$\|\|x - \mu_i\|\|^2$	Kuadrat jarak data terhadap centroid

Fungsi objektif tersebut menunjukkan bahwa K-Means berupaya menghasilkan cluster yang kompak. Cluster dikatakan kompak apabila jarak antara anggota cluster dan centroid relatif kecil. Semakin kecil nilai fungsi objektif, semakin dekat anggota cluster terhadap centroid-nya. Namun, nilai ini perlu ditafsirkan secara hati-hati karena penambahan jumlah cluster biasanya akan menurunkan nilai jarak dalam cluster.

6. Contoh Konseptual K-Means

Misalnya terdapat data siswa berdasarkan nilai Matematika dan IPA. Peneliti ingin mengelompokkan siswa ke dalam tiga cluster. Pada tahap awal, algoritma menentukan tiga centroid. Setiap siswa

kemudian dihitung jaraknya terhadap ketiga centroid tersebut. Siswa akan masuk ke cluster dengan centroid terdekat.

Apabila siswa A memiliki nilai Matematika 90 dan IPA 88, maka siswa tersebut kemungkinan akan dekat dengan centroid kelompok nilai tinggi. Jika siswa B memiliki nilai Matematika 65 dan IPA 68, maka siswa tersebut kemungkinan akan berada pada kelompok nilai sedang. Jika siswa C memiliki nilai Matematika 45 dan IPA 50, maka siswa tersebut kemungkinan akan dekat dengan kelompok nilai rendah. Setelah semua data dikelompokkan, centroid diperbarui berdasarkan rata-rata nilai anggota setiap cluster.

Contoh ini menunjukkan bahwa K-Means tidak memberi label "tinggi", "sedang", atau "rendah" secara otomatis. Algoritma hanya membentuk cluster berdasarkan perhitungan jarak. Label tersebut diberikan oleh peneliti setelah membaca karakteristik setiap cluster, terutama melalui nilai centroid dan rata-rata atribut anggota cluster.

7. Karakter Data yang Sesuai untuk K-Means

K-Means lebih sesuai digunakan pada data numerik karena prosesnya bergantung pada perhitungan jarak. Data kategorik tidak dapat langsung digunakan kecuali telah ditransformasikan secara tepat. Selain itu, K-Means lebih baik digunakan apabila data memiliki cluster yang relatif kompak dan tidak terlalu banyak mengandung outlier.

Beberapa karakter data yang sesuai untuk K-Means antara lain:

- data berbentuk numerik;
- atribut memiliki skala yang seimbang;
- jumlah cluster dapat ditentukan secara logis;
- data tidak terlalu banyak mengandung pencilan ekstrem;
- atribut yang digunakan relevan dengan tujuan pengelompokan;
- setiap objek data dapat dibandingkan berdasarkan ukuran jarak.

Jika data memiliki skala berbeda, peneliti dapat melakukan normalisasi atau standardisasi sebelum proses K-Means. Hal ini penting karena atribut dengan rentang nilai besar dapat mendominasi hasil perhitungan jarak. Sebagai contoh, atribut pendapatan dengan rentang jutaan rupiah dapat lebih dominan daripada atribut nilai akademik dengan rentang 0–100 jika keduanya digunakan tanpa transformasi.

8. Posisi K-Means dalam Data Mining

Dalam data mining, K-Means berada pada kelompok teknik clustering. K-Means tidak digunakan untuk memprediksi kelas yang sudah diketahui, melainkan untuk menemukan kelompok berdasarkan kemiripan data. Oleh karena itu, K-Means tepat digunakan ketika peneliti ingin melakukan segmentasi atau pengelompokan awal terhadap data.

Dalam bidang pendidikan, K-Means dapat digunakan untuk mengelompokkan siswa berdasarkan nilai akademik. Dalam bidang bisnis, K-Means dapat digunakan untuk segmentasi pelanggan. Dalam bidang pemerintahan, K-Means dapat digunakan untuk mengelompokkan masyarakat berdasarkan indikator sosial ekonomi. Dalam bidang kesehatan, K-Means dapat digunakan untuk mengelompokkan pasien berdasarkan karakteristik tertentu. Penerapan ini menunjukkan bahwa konsep dasar K-Means dapat digunakan secara luas selama data yang dianalisis sesuai dengan karakteristik algoritma.

9. Hal yang Perlu Diperhatikan dalam Konsep K-Means

Meskipun K-Means memiliki konsep yang sederhana, terdapat beberapa hal yang perlu diperhatikan. Pertama, penentuan nilai k sangat memengaruhi hasil clustering. Kedua, centroid awal dapat memengaruhi hasil akhir karena proses K-Means dimulai dari pusat awal tertentu. Ketiga, skala data harus diperhatikan agar tidak terjadi dominasi atribut tertentu. Keempat, outlier dapat memengaruhi centroid karena centroid dihitung menggunakan rata-rata.

Hal-hal tersebut menunjukkan bahwa K-Means tidak boleh digunakan hanya sebagai prosedur teknis. Peneliti perlu memahami konsep dasarnya agar dapat menyiapkan data, menentukan jumlah cluster, menjalankan algoritma, dan menafsirkan hasil secara tepat.

4.3 Prinsip Kerja Algoritma K-Means

Prinsip kerja algoritma K-Means didasarkan pada proses pengelompokan data ke dalam sejumlah cluster berdasarkan kedekatan data terhadap titik pusat cluster atau centroid. Algoritma ini bekerja secara iteratif, yaitu melalui proses pengulangan sampai susunan cluster mencapai kondisi yang relatif stabil. Pada setiap iterasi, data akan dihitung jaraknya terhadap centroid, kemudian ditempatkan pada cluster dengan jarak terdekat. Setelah seluruh data dikelompokkan, centroid diperbarui berdasarkan rata-rata anggota cluster yang baru terbentuk.

Secara umum, K-Means bekerja dengan tujuan membentuk cluster yang anggotanya memiliki kemiripan tinggi di dalam kelompoknya dan memiliki perbedaan yang cukup jelas dengan kelompok lain. Prinsip ini sejalan dengan tujuan clustering, yaitu memaksimalkan kemiripan dalam cluster dan meminimalkan kemiripan antarcluster. MacQueen (1967) memperkenalkan K-Means sebagai metode untuk mengelompokkan data multivariat ke dalam beberapa kelompok berdasarkan karakteristik data. Sementara itu, Lloyd (1982) menjelaskan prosedur iteratif yang menjadi dasar pembaruan pusat kelompok dalam pendekatan kuantisasi kuadrat terkecil.

Dalam penerapannya, K-Means dimulai dengan menentukan jumlah cluster atau nilai k. Nilai k menunjukkan banyaknya kelompok yang akan dibentuk. Penentuan nilai k merupakan langkah penting karena algoritma K-Means tidak menentukan jumlah cluster secara otomatis. Apabila nilai k ditentukan sebesar 3, maka algoritma akan membagi data ke dalam tiga cluster. Penentuan jumlah cluster dapat didasarkan pada tujuan penelitian,

kebutuhan analisis, atau metode evaluasi seperti elbow method dan silhouette coefficient.

Setelah jumlah cluster ditentukan, algoritma menetapkan centroid awal. Centroid awal dapat dipilih secara acak atau menggunakan metode tertentu. Pemilihan centroid awal berpengaruh terhadap proses iterasi dan kemungkinan hasil akhir. Pada beberapa kasus, centroid awal yang berbeda dapat menghasilkan susunan cluster yang berbeda. Oleh karena itu, dalam penelitian ilmiah, peneliti perlu memahami bahwa hasil K-Means dapat dipengaruhi oleh tahap awal ini.

Setelah centroid awal tersedia, setiap data dihitung jaraknya terhadap seluruh centroid. Ukuran jarak yang umum digunakan dalam K-Means adalah Euclidean distance. Jarak ini digunakan untuk mengetahui seberapa dekat suatu data dengan pusat cluster. Data akan ditempatkan ke dalam cluster yang memiliki jarak paling kecil terhadap centroid. Dengan prinsip ini, setiap data hanya menjadi anggota dari satu cluster.

Rumus Euclidean distance dapat dituliskan sebagai berikut.

$$d(x,c) = sqrt(\Sigma_{i=1}^{n} (x_i - c_i)^2)$$

Simbol	Keterangan
d(x,c)	Jarak antara data dan centroid
x_i	Nilai atribut ke-i pada objek data
c_i	Nilai atribut ke-i pada centroid
n	Jumlah atribut yang digunakan

Setelah seluruh data ditempatkan ke cluster terdekat, centroid setiap cluster diperbarui. Pembaruan centroid dilakukan dengan menghitung rata-rata nilai seluruh anggota cluster pada setiap atribut. Centroid baru ini kemudian digunakan kembali untuk menghitung jarak pada iterasi berikutnya. Proses tersebut terus dilakukan sampai tidak ada lagi perubahan anggota cluster atau perubahan centroid sudah sangat kecil.

Rumus pembaruan centroid dapat dituliskan sebagai berikut.

$$\mu_j = (1/N_j) \Sigma_x \in C_j \, x$$

Simbol	Keterangan
μ_j	Centroid cluster ke-j
N_j	Jumlah anggota pada cluster ke-j
C_j	Cluster ke-j
x	Data yang menjadi anggota cluster

Berdasarkan rumus tersebut, centroid merupakan nilai rata-rata dari seluruh data yang berada dalam satu cluster. Karena centroid dihitung dari rata-rata, maka keberadaan data yang sangat ekstrem dapat memengaruhi posisi centroid. Hal ini menjadi salah satu alasan mengapa data perlu dibersihkan dan diperiksa sebelum proses K-Means dilakukan.

Secara sistematis, prinsip kerja algoritma K-Means dapat dijelaskan melalui tahapan berikut.

1. Menentukan Jumlah Cluster

Tahap pertama adalah menentukan jumlah cluster yang akan dibentuk. Jumlah cluster dinyatakan dengan nilai k. Nilai ini harus ditentukan sebelum algoritma dijalankan. Penentuan nilai k dapat disesuaikan dengan kebutuhan penelitian. Misalnya, dalam pengelompokan prestasi akademik, peneliti dapat menentukan tiga cluster, yaitu kelompok tinggi, sedang, dan rendah. Namun, label tersebut tidak diberikan di awal oleh algoritma, melainkan diberikan setelah hasil cluster dianalisis.

Pemilihan nilai k tidak boleh dilakukan secara sembarangan. Jika nilai k terlalu kecil, beberapa data yang memiliki karakteristik berbeda dapat tergabung dalam satu cluster. Sebaliknya, jika nilai k terlalu besar, cluster yang terbentuk dapat terlalu banyak dan sulit ditafsirkan. Oleh karena itu, penentuan nilai k perlu mempertimbangkan tujuan penelitian, karakteristik data, dan evaluasi hasil clustering.

2. Menentukan Centroid Awal

Setelah nilai k ditentukan, langkah berikutnya adalah menentukan centroid awal. Jumlah centroid awal sama dengan jumlah cluster yang telah ditentukan. Jika k = 3, maka diperlukan tiga centroid awal. Centroid awal dapat dipilih secara acak dari data atau ditentukan dengan pendekatan tertentu.

Pemilihan centroid awal memiliki pengaruh terhadap hasil clustering karena proses perhitungan jarak dimulai dari centroid tersebut. Jika centroid awal dipilih kurang representatif, proses iterasi dapat menghasilkan cluster yang kurang optimal. Untuk mengurangi pengaruh pemilihan awal, beberapa perangkat lunak data mining menggunakan pengulangan proses dengan inisialisasi yang berbeda atau menggunakan metode seperti K-Means++.

3. Menghitung Jarak Setiap Data terhadap Centroid

Tahap berikutnya adalah menghitung jarak setiap data terhadap masing-masing centroid. Perhitungan ini dilakukan untuk menentukan centroid mana yang paling dekat dengan setiap data. Pada K-Means, jarak yang paling sering digunakan adalah jarak Euclidean karena mudah dihitung dan sesuai untuk data numerik.

Sebagai contoh, apabila terdapat tiga centroid, maka setiap data akan dihitung jaraknya terhadap ketiga centroid tersebut. Data kemudian akan dibandingkan berdasarkan nilai jarak yang diperoleh. Centroid dengan jarak paling kecil menunjukkan cluster yang paling dekat dengan data tersebut.

4. Mengelompokkan Data ke Cluster Terdekat

Setelah jarak dihitung, setiap data ditempatkan ke dalam cluster dengan centroid terdekat. Tahap ini disebut proses alokasi atau penugasan data ke cluster. Setiap data hanya masuk ke satu cluster karena K-Means merupakan metode hard clustering. Artinya, tidak ada data yang menjadi anggota lebih dari satu cluster pada waktu yang sama.

Misalnya, jika data siswa memiliki jarak paling dekat dengan centroid Cluster 1, maka siswa tersebut dimasukkan ke Cluster 1. Jika pada iterasi berikutnya centroid berubah dan jarak terdekat siswa tersebut berpindah ke Cluster 2, maka siswa tersebut dapat berpindah cluster. Perpindahan ini merupakan bagian dari proses iteratif K-Means.

5. Memperbarui Nilai Centroid

Setelah semua data dikelompokkan, centroid setiap cluster diperbarui. Pembaruan dilakukan dengan menghitung rata-rata nilai atribut dari seluruh anggota cluster. Centroid baru ini menggantikan centroid lama dan menjadi dasar perhitungan pada iterasi berikutnya.

Tahap pembaruan centroid penting karena pusat cluster harus menyesuaikan dengan anggota cluster terbaru. Jika anggota cluster berubah, maka centroid juga dapat berubah. Perubahan centroid inilah yang menyebabkan proses K-Means perlu dilakukan berulang sampai mencapai kondisi stabil.

6. Mengulangi Proses sampai Cluster Stabil

Setelah centroid diperbarui, proses perhitungan jarak dan pengelompokan data dilakukan kembali. Iterasi terus berlangsung sampai memenuhi kondisi berhenti. Kondisi berhenti dapat berupa tidak adanya perubahan anggota cluster, perubahan centroid yang sangat kecil, atau jumlah iterasi maksimum telah tercapai.

Kondisi stabil dalam K-Means menunjukkan bahwa proses clustering telah mencapai hasil akhir berdasarkan data, jumlah cluster, dan centroid yang digunakan. Namun, stabil secara algoritmik tidak selalu berarti hasil tersebut paling bermakna secara substantif. Oleh karena itu, hasil akhir tetap perlu dievaluasi dan diinterpretasikan.

Tahap	Kegiatan	Tujuan
1	Menentukan nilai k	Menentukan jumlah cluster yang akan dibentuk
2	Menentukan centroid awal	Menetapkan pusat awal setiap cluster
3	Menghitung jarak data ke centroid	Mengetahui kedekatan data terhadap setiap cluster
4	Menempatkan data pada cluster terdekat	Membentuk susunan cluster sementara
5	Memperbarui centroid	Menyesuaikan pusat cluster berdasarkan anggota terbaru
6	Melakukan iterasi	Mengulangi proses sampai cluster stabil
7	Mengevaluasi dan menafsirkan hasil	Memberikan makna pada cluster yang terbentuk

Prinsip kerja K-Means dapat dipahami melalui contoh sederhana. Misalnya, peneliti memiliki data siswa berdasarkan nilai Matematika dan IPA. Peneliti menentukan nilai k = 3, sehingga data akan dikelompokkan ke dalam tiga cluster. Pada tahap awal, algoritma menentukan tiga centroid. Setiap siswa dihitung jaraknya terhadap ketiga centroid tersebut. Siswa kemudian ditempatkan pada cluster dengan centroid terdekat. Setelah semua siswa masuk ke cluster, centroid diperbarui berdasarkan rata-rata nilai siswa dalam setiap cluster. Proses ini diulang sampai posisi cluster tidak berubah.

Contoh tersebut menunjukkan bahwa K-Means tidak langsung memberi nama pada kelompok yang terbentuk. Algoritma hanya menghasilkan cluster berdasarkan perhitungan jarak. Setelah hasil diperoleh, peneliti perlu membaca nilai centroid dan karakteristik anggota cluster. Jika centroid cluster pertama menunjukkan nilai akademik tinggi, maka cluster tersebut dapat ditafsirkan sebagai kelompok prestasi tinggi. Jika cluster kedua menunjukkan nilai sedang, maka dapat ditafsirkan sebagai kelompok prestasi sedang. Jika cluster ketiga menunjukkan nilai rendah, maka dapat ditafsirkan sebagai kelompok prestasi rendah.

Dalam praktik penelitian, prinsip kerja K-Means juga perlu dikaitkan dengan kualitas data. Data yang digunakan harus bersih, relevan, dan memiliki skala yang sesuai. K-Means sangat dipengaruhi oleh perbedaan skala atribut karena algoritma ini menggunakan perhitungan jarak. Atribut dengan rentang nilai besar dapat mendominasi hasil clustering. Oleh karena itu, normalisasi atau standardisasi sering diperlukan sebelum algoritma dijalankan.

Selain itu, K-Means juga sensitif terhadap outlier. Karena centroid dihitung berdasarkan rata-rata, data yang nilainya sangat ekstrem dapat menggeser posisi centroid. Jika outlier berasal dari kesalahan input, maka data perlu diperbaiki. Namun, jika outlier merupakan fenomena yang valid, peneliti perlu mempertimbangkan pengaruhnya terhadap hasil clustering. Jain (2010) menegaskan bahwa meskipun K-Means sederhana dan populer, hasilnya dapat dipengaruhi oleh representasi data, pemilihan ukuran jarak, dan kondisi awal algoritma.

Prinsip kerja K-Means juga menunjukkan bahwa algoritma ini cocok untuk data numerik. Data kategorik tidak dapat langsung digunakan dalam perhitungan jarak Euclidean. Jika data kategorik tetap ingin digunakan, data tersebut perlu ditransformasikan terlebih dahulu. Namun, proses transformasi harus dilakukan secara hati-hati agar angka yang diberikan tidak menimbulkan makna jarak yang keliru.

Dalam konteks fungsi objektif, K-Means berusaha meminimalkan jumlah kuadrat jarak dalam cluster. Tujuan ini dapat dituliskan dalam bentuk berikut.

$$J = \Sigma_{j=1}^{k} \Sigma_{x \in C_j} ||x - \mu_j||^2$$

Fungsi tersebut menunjukkan bahwa algoritma berusaha membuat setiap data berada sedekat mungkin dengan centroid cluster-nya. Semakin kecil nilai jarak dalam cluster, semakin kompak cluster tersebut. Akan tetapi, nilai fungsi objektif tidak dapat digunakan secara tunggal untuk menentukan kualitas cluster

karena nilai tersebut cenderung menurun ketika jumlah cluster ditambah. Oleh sebab itu, evaluasi K-Means perlu mempertimbangkan ukuran lain dan konteks penelitian.

Beberapa hal penting yang perlu diperhatikan dalam prinsip kerja K-Means adalah sebagai berikut.

- Nilai k harus ditentukan sebelum algoritma dijalankan.
- Centroid awal dapat memengaruhi hasil akhir clustering.
- Jarak data terhadap centroid menjadi dasar pengelompokan.
- Centroid diperbarui berdasarkan rata-rata anggota cluster.
- Proses dilakukan secara iteratif sampai cluster stabil.
- Data numerik lebih sesuai digunakan dalam K-Means.
- Perbedaan skala atribut dapat memengaruhi hasil perhitungan jarak.
- Outlier dapat menggeser posisi centroid.
- Hasil cluster memerlukan evaluasi dan interpretasi.
- Label cluster ditentukan oleh peneliti berdasarkan karakteristik data.

Dengan memahami prinsip kerja tersebut, peneliti dapat menggunakan K-Means secara lebih tepat. K-Means bukan sekadar algoritma yang membagi data menjadi beberapa kelompok, tetapi metode yang membutuhkan penentuan parameter, kesiapan data, proses iteratif, dan interpretasi hasil. Jika setiap tahap dilakukan dengan baik, K-Means dapat membantu peneliti menemukan struktur kelompok yang relevan dan bermanfaat dalam analisis data.

4.4 Penentuan Jumlah Cluster

Penentuan jumlah cluster merupakan salah satu tahap penting dalam penerapan algoritma K-Means. Jumlah cluster dinyatakan dengan simbol k, yaitu banyaknya kelompok yang akan dibentuk dari data yang dianalisis. Berbeda dengan beberapa metode clustering lain yang dapat membentuk struktur kelompok secara bertingkat atau berdasarkan kepadatan, K-Means mengharuskan

peneliti menentukan nilai k sebelum algoritma dijalankan. Dengan demikian, kualitas hasil K-Means sangat dipengaruhi oleh ketepatan dalam menentukan jumlah cluster.

Penentuan jumlah cluster tidak boleh dilakukan secara sembarangan. Jika jumlah cluster terlalu sedikit, data yang sebenarnya memiliki karakteristik berbeda dapat tergabung dalam kelompok yang sama. Sebaliknya, jika jumlah cluster terlalu banyak, data dapat terbagi ke dalam kelompok yang terlalu kecil sehingga sulit ditafsirkan secara substantif. Oleh karena itu, penentuan nilai k perlu mempertimbangkan tujuan penelitian, karakteristik data, kebutuhan interpretasi, serta ukuran evaluasi clustering yang sesuai.

Dalam penelitian ilmiah, jumlah cluster yang baik bukan hanya jumlah yang menghasilkan nilai perhitungan paling kecil, tetapi juga jumlah yang dapat dijelaskan secara logis sesuai konteks penelitian. Pada algoritma K-Means, penambahan jumlah cluster biasanya akan menurunkan nilai jarak dalam cluster karena data dibagi ke dalam kelompok yang lebih kecil. Namun, penurunan nilai tersebut tidak selalu berarti bahwa hasil clustering menjadi lebih baik secara makna. Oleh sebab itu, peneliti perlu menyeimbangkan antara hasil matematis dan interpretasi substantif.

Secara umum, penentuan jumlah cluster dalam K-Means dapat dilakukan melalui beberapa pendekatan berikut.

1. Berdasarkan Tujuan Penelitian

Pendekatan pertama dalam menentukan jumlah cluster adalah berdasarkan tujuan penelitian. Peneliti perlu memahami terlebih dahulu mengapa data dikelompokkan dan hasil seperti apa yang ingin diperoleh. Jika tujuan penelitian adalah membagi data ke dalam kategori umum, maka jumlah cluster dapat disesuaikan dengan kebutuhan analisis.

Sebagai contoh, dalam penelitian pendidikan, peneliti dapat menentukan tiga cluster apabila ingin mengelompokkan siswa ke dalam kelompok kemampuan tinggi, sedang, dan rendah. Dalam

bidang bisnis, peneliti dapat menentukan beberapa cluster pelanggan berdasarkan perilaku transaksi, misalnya pelanggan aktif, pelanggan potensial, dan pelanggan pasif. Namun, jumlah cluster tersebut tetap perlu diperiksa kembali melalui karakteristik data agar tidak hanya berdasarkan asumsi awal.

Penentuan jumlah cluster berdasarkan tujuan penelitian sangat berguna ketika peneliti memiliki dasar konseptual atau kebutuhan praktis yang jelas. Meskipun demikian, pendekatan ini sebaiknya tidak berdiri sendiri. Peneliti tetap perlu mengevaluasi apakah jumlah cluster yang dipilih memang sesuai dengan pola data.

2. Berdasarkan Karakteristik Data

Jumlah cluster juga dapat ditentukan dengan memperhatikan karakteristik data. Peneliti perlu melihat apakah data memiliki kecenderungan membentuk kelompok tertentu. Hal ini dapat dilakukan melalui pemeriksaan deskriptif, visualisasi data, atau eksplorasi awal terhadap sebaran atribut.

Pada data yang memiliki perbedaan karakteristik cukup jelas, jumlah cluster biasanya lebih mudah ditentukan. Misalnya, jika data nilai siswa menunjukkan tiga kelompok nilai yang relatif berbeda, maka nilai k = 3 dapat dipertimbangkan. Namun, apabila data tersebar secara merata tanpa batas kelompok yang jelas, penentuan jumlah cluster menjadi lebih sulit dan memerlukan metode evaluasi tambahan.

Karakteristik data yang perlu diperhatikan dalam penentuan jumlah cluster antara lain:

- jumlah data yang dianalisis;
- jumlah atribut yang digunakan;
- sebaran nilai setiap atribut;
- keberadaan data pencilan;
- perbedaan skala antaratribut;
- pola kemiripan antarobjek;
- kemudahan interpretasi hasil cluster.

Dalam algoritma K-Means, karakteristik data sangat penting karena proses clustering didasarkan pada perhitungan jarak. Jika data tidak memiliki struktur kelompok yang cukup jelas, K-Means tetap akan membentuk cluster, tetapi hasilnya belum tentu bermakna. Oleh karena itu, peneliti perlu mengevaluasi hasil cluster sebelum menarik kesimpulan.

3. Menggunakan Elbow Method

Salah satu metode yang umum digunakan untuk menentukan jumlah cluster adalah elbow method. Metode ini dilakukan dengan menjalankan K-Means pada beberapa nilai k, kemudian menghitung nilai Sum of Squared Error atau Within-Cluster Sum of Squares pada setiap nilai k. Nilai tersebut menunjukkan jumlah kuadrat jarak antara data dengan centroid pada cluster masing-masing.

Rumus umum SSE dapat dituliskan sebagai berikut.

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$$

Simbol	Keterangan
SSE	Jumlah kuadrat jarak data terhadap centroid
k	Jumlah cluster
C_i	Cluster ke-i
x	Objek data dalam cluster
μ_i	Centroid cluster ke-i
$\|\|x - \mu_i\|\|^2$	Kuadrat jarak data terhadap centroid

Pada elbow method, nilai SSE biasanya akan menurun ketika jumlah cluster bertambah. Namun, penurunan tersebut tidak selalu terjadi secara signifikan pada setiap penambahan cluster. Titik yang menunjukkan perubahan penurunan mulai melambat disebut sebagai titik siku atau elbow. Titik inilah yang sering dijadikan pertimbangan untuk memilih jumlah cluster.

Sebagai contoh, jika nilai SSE turun tajam dari k = 1 ke k = 2 dan dari k = 2 ke k = 3, tetapi penurunannya mulai kecil setelah k = 3, maka nilai k = 3 dapat dipertimbangkan sebagai jumlah cluster yang sesuai. Namun, elbow method bersifat heuristik, sehingga hasilnya tetap perlu dikaitkan dengan konteks penelitian dan interpretasi cluster.

Kelebihan elbow method adalah mudah dipahami dan mudah diterapkan. Akan tetapi, kelemahannya adalah titik siku tidak selalu terlihat jelas. Pada beberapa dataset, grafik SSE dapat menurun secara bertahap tanpa membentuk siku yang tegas. Dalam kondisi seperti itu, peneliti perlu menggunakan ukuran evaluasi lain sebagai pembanding.

4. Menggunakan Silhouette Coefficient

Metode lain yang dapat digunakan untuk membantu menentukan jumlah cluster adalah silhouette coefficient. Ukuran ini digunakan untuk menilai seberapa baik suatu data berada dalam cluster-nya dibandingkan dengan cluster lain. Nilai silhouette berada pada rentang -1 sampai 1. Nilai yang mendekati 1 menunjukkan bahwa data sesuai dengan cluster-nya. Nilai mendekati 0 menunjukkan bahwa data berada di antara dua cluster. Sementara itu, nilai negatif menunjukkan kemungkinan data lebih sesuai berada pada cluster lain. Rousseeuw (1987) memperkenalkan silhouette sebagai alat grafis untuk membantu interpretasi dan validasi hasil cluster.

Secara umum, interpretasi nilai silhouette dapat dijelaskan sebagai berikut.

Nilai Silhouette	Interpretasi Umum
Mendekati 1	Data sesuai dengan cluster-nya
Mendekati 0	Data berada di batas antara dua cluster
Bernilai negatif	Data mungkin lebih tepat berada pada cluster lain

Dalam penentuan jumlah cluster, K-Means dapat dijalankan pada beberapa nilai k, kemudian nilai rata-rata silhouette dihitung untuk setiap k. Nilai k dengan rata-rata silhouette yang lebih tinggi dapat dipertimbangkan sebagai jumlah cluster yang lebih baik. Namun, sama seperti metode lain, silhouette tidak boleh digunakan sebagai satu-satunya dasar. Peneliti tetap perlu melihat apakah hasil cluster dapat dijelaskan secara substantif.

Silhouette coefficient memiliki kelebihan karena menilai dua aspek penting dalam clustering, yaitu kedekatan data dalam cluster dan keterpisahan data dari cluster lain. Dengan demikian, ukuran ini tidak hanya melihat kekompakan cluster, tetapi juga memperhatikan perbedaan antarcluster. Kaufman dan Rousseeuw (1990) juga membahas penggunaan silhouette dalam analisis cluster sebagai bagian dari upaya memahami kualitas pengelompokan data.

5. Menggunakan Gap Statistic

Gap statistic merupakan metode yang digunakan untuk memperkirakan jumlah cluster dengan membandingkan variasi dalam cluster pada data asli terhadap variasi yang diharapkan dari data acak. Metode ini dikembangkan oleh Tibshirani, Walther, dan Hastie (2001) sebagai pendekatan untuk memperkirakan jumlah kelompok dalam dataset. Metode ini dapat digunakan pada berbagai algoritma clustering, termasuk K-Means dan hierarchical clustering.

Secara sederhana, gap statistic membandingkan apakah struktur cluster pada data asli lebih kuat dibandingkan dengan struktur yang mungkin muncul secara acak. Jika suatu nilai k menghasilkan perbedaan yang besar antara data asli dan data acak, maka nilai tersebut dapat dipertimbangkan sebagai jumlah cluster yang sesuai.

Kelebihan gap statistic adalah memiliki dasar statistik yang lebih kuat dibandingkan penentuan jumlah cluster secara visual. Namun, metode ini lebih kompleks dibandingkan elbow method dan

membutuhkan proses komputasi yang lebih panjang. Oleh karena itu, penggunaannya perlu disesuaikan dengan kebutuhan penelitian, perangkat analisis yang tersedia, dan tingkat kedalaman analisis yang diinginkan.

6. Berdasarkan Kemudahan Interpretasi

Selain ukuran kuantitatif, penentuan jumlah cluster juga perlu memperhatikan kemudahan interpretasi. Dalam penelitian ilmiah, hasil clustering harus dapat dijelaskan dengan jelas. Jumlah cluster yang terlalu banyak dapat membuat interpretasi menjadi rumit. Sebaliknya, jumlah cluster yang terlalu sedikit dapat membuat perbedaan karakteristik antar kelompok menjadi kurang terlihat.

Sebagai contoh, dalam penelitian pengelompokan prestasi akademik siswa, tiga cluster sering lebih mudah ditafsirkan karena dapat dikaitkan dengan kelompok prestasi tinggi, sedang, dan rendah. Namun, jika data menunjukkan variasi yang lebih kompleks, empat atau lima cluster mungkin lebih sesuai. Pemilihan jumlah cluster harus mempertimbangkan apakah setiap cluster memiliki karakteristik yang berbeda dan dapat diberi makna berdasarkan data.

Kemudahan interpretasi penting karena tujuan akhir clustering bukan hanya membagi data, tetapi menghasilkan pemahaman. Cluster yang baik seharusnya dapat dijelaskan melalui nilai centroid, rata-rata atribut, jumlah anggota cluster, dan karakteristik dominan dari setiap kelompok.

7. Berdasarkan Keseimbangan Jumlah Anggota Cluster

Penentuan jumlah cluster juga dapat mempertimbangkan keseimbangan jumlah anggota pada setiap cluster. Cluster yang terlalu kecil perlu diperiksa lebih lanjut. Cluster kecil tidak selalu salah, tetapi dapat menunjukkan adanya kelompok khusus, data pencilan, atau hasil pengelompokan yang kurang stabil.

Sebagai contoh, jika K-Means dengan k = 4 menghasilkan tiga cluster besar dan satu cluster yang hanya berisi satu atau dua data, peneliti perlu meninjau kembali hasil tersebut. Cluster kecil dapat

bermakna jika data tersebut memang memiliki karakteristik khusus. Namun, jika cluster kecil muncul karena outlier atau kesalahan data, maka peneliti perlu melakukan pemeriksaan ulang terhadap dataset.

Keseimbangan jumlah anggota cluster bukan berarti setiap cluster harus memiliki jumlah anggota yang sama. Dalam data nyata, ukuran cluster dapat berbeda. Namun, peneliti perlu memastikan bahwa perbedaan tersebut dapat dijelaskan secara logis.

Langkah-Langkah Menentukan Jumlah Cluster

Dalam penelitian berbasis K-Means, penentuan jumlah cluster dapat dilakukan secara sistematis melalui langkah-langkah berikut.

- Menentukan tujuan pengelompokan. Peneliti terlebih dahulu menetapkan tujuan analisis, misalnya mengelompokkan siswa, pelanggan, wilayah, atau objek lain berdasarkan atribut tertentu.
- Memahami karakteristik data. Peneliti memeriksa jumlah data, atribut yang digunakan, sebaran nilai, data pencilan, dan skala data.
- Menyiapkan beberapa alternatif nilai k. Peneliti menjalankan K-Means pada beberapa nilai k, misalnya k = 2 sampai k = 6.
- Menghitung ukuran evaluasi. Peneliti menghitung nilai SSE, silhouette coefficient, atau ukuran lain yang relevan.
- Membandingkan hasil setiap nilai k. Peneliti membandingkan hasil pengelompokan berdasarkan ukuran evaluasi dan karakteristik cluster.
- Membaca karakteristik cluster. Peneliti melihat nilai centroid, jumlah anggota, dan perbedaan atribut antarcluster.
- Memilih nilai k yang paling masuk akal. Nilai k dipilih berdasarkan keseimbangan antara hasil evaluasi dan kemudahan interpretasi.
- Menjelaskan dasar pemilihan nilai k. Peneliti perlu menjelaskan alasan pemilihan jumlah cluster dalam laporan penelitian agar dapat dipertanggungjawabkan.

Proses ini menunjukkan bahwa penentuan jumlah cluster bukan hanya kegiatan teknis, tetapi bagian dari pertimbangan metodologis. Peneliti harus mampu menjelaskan mengapa jumlah cluster tertentu dipilih dan bagaimana jumlah tersebut mendukung tujuan penelitian.

4.5 Penentuan Centroid Awal

Penentuan centroid awal merupakan salah satu tahap penting dalam algoritma K-Means karena centroid menjadi titik pusat awal yang digunakan untuk membentuk cluster. Pada tahap ini, algoritma menentukan sejumlah pusat cluster sesuai dengan nilai **k** yang telah ditetapkan sebelumnya. Jika nilai k adalah 3, maka diperlukan tiga centroid awal. Setiap data kemudian dihitung jaraknya terhadap centroid tersebut untuk menentukan cluster terdekat.

Centroid awal memiliki pengaruh besar terhadap proses dan hasil clustering. Hal ini terjadi karena K-Means bekerja secara iteratif dengan memulai proses pengelompokan dari titik pusat awal. Apabila centroid awal dipilih dengan baik, proses clustering dapat menghasilkan kelompok yang lebih stabil dan mudah ditafsirkan. Sebaliknya, apabila centroid awal kurang representatif, hasil clustering dapat kurang optimal, membutuhkan iterasi lebih banyak, atau menghasilkan cluster yang sulit dijelaskan. Jain (2010) menjelaskan bahwa salah satu persoalan penting dalam K-Means adalah pengaruh kondisi awal terhadap kualitas hasil clustering.

Dalam K-Means, centroid merupakan titik pusat dari suatu cluster. Pada awal proses, centroid dapat dipilih secara acak dari data atau ditentukan dengan metode tertentu. Setelah data dikelompokkan berdasarkan centroid terdekat, centroid diperbarui menggunakan rata-rata nilai anggota cluster. Proses tersebut diulang sampai tidak terjadi perubahan yang berarti pada anggota cluster atau posisi centroid. Dengan demikian, centroid awal berfungsi sebagai titik awal yang menentukan arah proses clustering selanjutnya.

Peran Centroid Awal dalam K-Means

Centroid awal berperan sebagai pusat sementara dari setiap cluster pada tahap awal algoritma. Seluruh data akan dibandingkan dengan centroid tersebut berdasarkan ukuran jarak, biasanya menggunakan jarak Euclidean. Data kemudian dimasukkan ke dalam cluster dengan centroid terdekat. Setelah seluruh data masuk ke cluster masing-masing, centroid diperbarui berdasarkan rata-rata anggota cluster.

Peran centroid awal dapat dijelaskan melalui beberapa fungsi berikut.

- Menjadi titik pusat awal cluster
  Centroid awal menjadi acuan pertama dalam proses pengelompokan data.
- Menentukan pembagian data pada iterasi pertama
  Data akan ditempatkan ke cluster berdasarkan jarak terdekat terhadap centroid awal.
- Memengaruhi posisi centroid berikutnya
  Karena centroid baru dihitung dari anggota cluster, pemilihan centroid awal dapat memengaruhi pembaruan centroid pada iterasi berikutnya.
- Memengaruhi jumlah iterasi
  Centroid awal yang baik dapat mempercepat proses menuju kondisi stabil.
- Memengaruhi kualitas hasil cluster
  Centroid awal yang kurang tepat dapat menyebabkan hasil clustering terjebak pada solusi lokal yang kurang optimal.

Dalam algoritma K-Means, solusi yang diperoleh tidak selalu merupakan solusi terbaik secara global. Hal ini terjadi karena K-Means berusaha meminimalkan jarak dalam cluster berdasarkan proses iteratif yang dimulai dari centroid awal. Arthur dan Vassilvitskii (2007) menjelaskan bahwa K-Means banyak digunakan karena sederhana dan cepat, tetapi pemilihan pusat awal yang kurang baik dapat memengaruhi kualitas hasil clustering.

Pengaruh Centroid Awal terhadap Hasil Clustering

Pengaruh centroid awal dapat terlihat dari kemungkinan perbedaan hasil ketika algoritma dijalankan lebih dari satu kali. Jika centroid awal dipilih secara acak, maka hasil clustering dapat berbeda antara satu percobaan dan percobaan lain. Perbedaan tersebut dapat terjadi karena data pada iterasi pertama masuk ke cluster yang berbeda, sehingga pembaruan centroid berikutnya juga berbeda.

Sebagai contoh, dalam pengelompokan siswa berdasarkan nilai akademik, pemilihan centroid awal yang mewakili nilai tinggi, sedang, dan rendah dapat membantu algoritma membentuk cluster yang lebih jelas. Namun, jika ketiga centroid awal berada terlalu dekat pada kelompok nilai yang sama, maka proses clustering dapat menjadi kurang seimbang. Beberapa cluster mungkin memiliki anggota yang sangat sedikit, sedangkan cluster lain terlalu banyak. Kondisi seperti ini dapat menyulitkan interpretasi hasil.

Secara metodologis, pengaruh centroid awal perlu diperhatikan karena penelitian ilmiah membutuhkan hasil yang dapat dijelaskan dan dipertanggungjawabkan. Peneliti sebaiknya tidak hanya menerima hasil clustering pertama tanpa mengevaluasi kestabilannya. Jika algoritma dijalankan dengan centroid awal yang berbeda dan menghasilkan cluster yang sangat berbeda, maka peneliti perlu memeriksa kembali data, jumlah cluster, metode inisialisasi, atau kualitas atribut yang digunakan.

Metode Penentuan Centroid Awal

Penentuan centroid awal dapat dilakukan melalui beberapa pendekatan. Setiap pendekatan memiliki kelebihan dan keterbatasan. Pemilihan pendekatan perlu disesuaikan dengan karakteristik data, tujuan penelitian, serta perangkat analisis yang digunakan.

1. Pemilihan Centroid Secara Acak

Metode paling sederhana dalam menentukan centroid awal adalah memilih centroid secara acak dari data yang tersedia. Pada metode ini, algoritma memilih sejumlah data sebagai pusat awal sesuai dengan nilai **k**. Jika k = 3, maka tiga data dipilih secara acak sebagai centroid awal.

Kelebihan metode acak adalah mudah diterapkan dan tidak membutuhkan proses perhitungan tambahan. Namun, kelemahannya adalah hasil dapat berbeda pada setiap percobaan. Centroid yang dipilih secara acak belum tentu mewakili sebaran data dengan baik. Jika centroid awal berada pada posisi yang kurang tepat, hasil clustering dapat kurang stabil.

Dalam penelitian, penggunaan centroid acak masih dapat dilakukan, tetapi sebaiknya disertai pengulangan proses beberapa kali. Hasil terbaik dapat dipilih berdasarkan nilai evaluasi tertentu, seperti SSE atau silhouette coefficient. Dengan cara ini, peneliti dapat mengurangi risiko hasil yang terlalu bergantung pada satu pemilihan centroid awal.

2. Pemilihan Centroid Berdasarkan Data yang Menyebar

Pendekatan lain adalah memilih centroid awal dari data yang memiliki posisi cukup menyebar. Tujuannya adalah agar centroid awal tidak berada terlalu dekat satu sama lain. Jika centroid awal tersebar dengan baik, maka proses awal pengelompokan dapat lebih representatif terhadap variasi data.

Secara sederhana, peneliti dapat memilih data yang mewakili nilai rendah, sedang, dan tinggi apabila tujuan penelitian adalah mengelompokkan data menjadi tiga cluster. Misalnya, dalam data nilai akademik siswa, centroid awal dapat dipilih dari siswa dengan nilai rendah, nilai sedang, dan nilai tinggi. Namun, pendekatan ini harus dilakukan secara hati-hati agar tidak terlalu subjektif.

Metode ini dapat membantu proses interpretasi karena centroid awal dipilih berdasarkan pemahaman terhadap data. Akan tetapi, peneliti tetap perlu memastikan bahwa pemilihan tersebut

memiliki dasar yang jelas dan tidak dilakukan hanya untuk menghasilkan cluster yang sesuai dengan dugaan awal.

3. Menggunakan Beberapa Percobaan Inisialisasi

Salah satu cara untuk mengurangi pengaruh centroid awal adalah menjalankan K-Means beberapa kali dengan centroid awal yang berbeda. Dari beberapa hasil tersebut, peneliti dapat memilih hasil clustering yang memiliki kualitas terbaik berdasarkan ukuran evaluasi tertentu. Pendekatan ini umum digunakan dalam perangkat lunak data mining dan statistik.

Metode ini penting karena satu kali proses K-Means belum tentu menghasilkan susunan cluster terbaik. Dengan menjalankan beberapa inisialisasi, peluang memperoleh hasil yang lebih baik menjadi lebih besar. Bradley dan Fayyad (1998) membahas pentingnya penyempurnaan titik awal dalam K-Means karena titik awal dapat memengaruhi kualitas solusi clustering.

Dalam praktik penelitian, peneliti dapat menjelaskan bahwa algoritma dijalankan beberapa kali dan hasil terbaik dipilih berdasarkan nilai SSE terkecil atau ukuran validasi cluster lainnya. Penjelasan ini membantu meningkatkan transparansi metode penelitian.

4. Menggunakan Metode K-Means++

K-Means++ merupakan metode inisialisasi centroid yang dirancang untuk memilih centroid awal secara lebih hati-hati. Metode ini tidak memilih seluruh centroid secara acak murni, tetapi memilih centroid berikutnya dengan mempertimbangkan jarak terhadap centroid yang telah dipilih sebelumnya. Dengan cara ini, centroid awal cenderung lebih menyebar dan lebih representatif terhadap struktur data.

Arthur dan Vassilvitskii (2007) memperkenalkan K-Means++ sebagai teknik *careful seeding* yang bertujuan memperbaiki pemilihan pusat awal pada K-Means. Penelitian tersebut menunjukkan bahwa inisialisasi yang lebih baik dapat membantu

meningkatkan kualitas hasil clustering dibandingkan pemilihan pusat awal secara acak.

Secara umum, langkah K-Means++ dapat dijelaskan sebagai berikut.

- Pilih satu centroid awal secara acak dari data.
- Hitung jarak setiap data terhadap centroid terdekat yang telah dipilih.
- Pilih centroid berikutnya dengan peluang lebih besar pada data yang jauh dari centroid yang sudah ada.
- Ulangi proses sampai jumlah centroid sama dengan nilai **k**.
- Jalankan algoritma K-Means seperti biasa.

Kelebihan K-Means++ adalah dapat menghasilkan centroid awal yang lebih baik dan mengurangi kemungkinan cluster yang buruk akibat pemilihan awal yang tidak representatif. Namun, metode ini membutuhkan proses tambahan sebelum iterasi K-Means dimulai. Dalam banyak kasus, tambahan proses ini sebanding dengan peningkatan stabilitas hasil clustering.

5. Penentuan Centroid Berdasarkan Pengetahuan Domain

Dalam beberapa penelitian, peneliti dapat menentukan centroid awal berdasarkan pengetahuan domain atau pemahaman terhadap data. Misalnya, dalam pengelompokan prestasi akademik, peneliti dapat menggunakan nilai yang mewakili kategori rendah, sedang, dan tinggi sebagai pusat awal. Pendekatan ini dapat digunakan apabila peneliti memiliki alasan teoritis atau praktis yang kuat.

Namun, pendekatan berbasis pengetahuan domain perlu digunakan secara hati-hati. Peneliti tidak boleh menentukan centroid awal hanya untuk mengarahkan hasil agar sesuai dengan harapan tertentu. Jika digunakan, dasar penentuannya harus dijelaskan secara terbuka. Pendekatan ini lebih tepat apabila konteks penelitian memang memiliki batasan atau kategori yang dapat dipertanggungjawabkan.

4.6 Perhitungan Jarak Data ke Centroid

Perhitungan jarak data ke centroid merupakan bagian inti dalam algoritma K-Means. Setelah jumlah cluster dan centroid awal ditentukan, setiap data dihitung jaraknya terhadap seluruh centroid. Data kemudian ditempatkan pada cluster yang memiliki jarak paling dekat. Dengan demikian, jarak menjadi dasar utama dalam menentukan keanggotaan suatu data pada cluster tertentu.

Dalam K-Means, centroid berfungsi sebagai titik pusat cluster. Semakin kecil jarak suatu data terhadap centroid, semakin besar kemiripan data tersebut dengan karakteristik cluster. Sebaliknya, semakin besar jarak data terhadap centroid, semakin kecil kemungkinan data tersebut menjadi anggota cluster tersebut. Prinsip ini menunjukkan bahwa K-Means bekerja berdasarkan kedekatan numerik antarobjek data.

Ukuran jarak yang paling umum digunakan dalam K-Means adalah Euclidean distance. Ukuran ini sesuai untuk data numerik karena menghitung jarak lurus antara dua titik dalam ruang berdimensi banyak. Pada data dua atribut, jarak dapat dipahami sebagai jarak antar titik pada bidang koordinat. Pada data dengan lebih banyak atribut, prinsip perhitungannya tetap sama, yaitu menghitung selisih nilai setiap atribut antara data dan centroid.

$$\mathbf{d(x,c) = \sqrt{\Sigma_{i=1}^{n} (x_i - c_i)^2}}$$

Keterangan rumus tersebut dapat dijelaskan sebagai berikut.

Simbol	Keterangan
d(x,c)	Jarak antara data x dan centroid c
x_i	Nilai atribut ke-i pada data
c_i	Nilai atribut ke-i pada centroid
n	Jumlah atribut yang digunakan
Σ	Penjumlahan seluruh atribut

Langkah-Langkah Perhitungan

- Menentukan atribut yang digunakan sebagai dasar clustering, misalnya nilai Matematika dan IPA.
- Menentukan centroid yang akan menjadi pusat pembanding setiap cluster.
- Menghitung selisih nilai setiap atribut antara data dan centroid.
- Mengkuadratkan setiap selisih nilai agar seluruh hasil menjadi positif.
- Menjumlahkan seluruh hasil kuadrat dari setiap atribut.
- Mengambil akar dari jumlah kuadrat tersebut untuk memperoleh jarak Euclidean.
- Membandingkan jarak data terhadap semua centroid dan memilih jarak terkecil sebagai cluster terdekat.

4.7 Pembentukan Cluster

Pembentukan cluster merupakan proses pengelompokan data ke dalam beberapa kelompok berdasarkan tingkat kemiripan karakteristik yang dimiliki oleh setiap data. Dalam data mining, cluster digunakan untuk menemukan pola tersembunyi dari kumpulan data yang belum memiliki label atau kategori tertentu. Data yang memiliki karakteristik serupa akan dikelompokkan ke dalam cluster yang sama, sedangkan data yang memiliki perbedaan karakteristik akan ditempatkan pada cluster yang berbeda.

Pada algoritma K-Means, pembentukan cluster diawali dengan menentukan jumlah cluster atau nilai K. Nilai K menunjukkan banyaknya kelompok yang ingin dibentuk dalam proses analisis. Penentuan jumlah cluster dapat disesuaikan dengan kebutuhan penelitian, tujuan analisis, atau karakteristik data yang digunakan. Pemilihan jumlah cluster yang tepat sangat penting karena akan memengaruhi hasil akhir pengelompokan data.

Setelah jumlah cluster ditentukan, langkah berikutnya adalah menentukan titik pusat awal atau centroid. Centroid berfungsi sebagai pusat dari masing-masing cluster. Setiap data kemudian

dihitung jaraknya terhadap setiap centroid. Data akan dimasukkan ke dalam cluster yang memiliki jarak paling dekat dengan centroid tersebut. Proses ini menunjukkan bahwa kedekatan jarak menjadi dasar utama dalam pembentukan cluster pada algoritma K-Means.

Setelah seluruh data masuk ke dalam cluster tertentu, centroid akan diperbarui berdasarkan rata-rata nilai data yang terdapat dalam masing-masing cluster. Selanjutnya, proses perhitungan jarak dan pembaruan centroid dilakukan secara berulang sampai tidak ada lagi perubahan posisi data dalam cluster atau sampai hasil pengelompokan mencapai kondisi stabil.

Hasil dari pembentukan cluster berupa kelompok-kelompok data yang memiliki karakteristik relatif sama di dalam satu cluster. Pembentukan cluster ini dapat membantu peneliti atau pengambil keputusan dalam memahami pola data, membandingkan karakteristik antar kelompok, serta menentukan langkah atau kebijakan berdasarkan hasil pengelompokan tersebut.

Dengan demikian, pembentukan cluster merupakan tahap penting dalam proses clustering karena menjadi dasar dalam menemukan struktur atau pola pada data. Melalui proses ini, data yang awalnya sulit dipahami secara keseluruhan dapat disusun menjadi kelompok-kelompok yang lebih teratur, jelas, dan mudah dianalisis.

4.8 Pembaruan Nilai Centroid

Pembaruan nilai centroid merupakan salah satu tahap penting dalam algoritma K-Means Clustering. Centroid adalah titik pusat dari suatu cluster yang digunakan sebagai acuan dalam proses pengelompokan data. Setelah data dikelompokkan ke dalam cluster berdasarkan jarak terdekat terhadap centroid awal, nilai centroid perlu diperbarui agar posisi pusat cluster menjadi lebih sesuai dengan anggota data yang berada di dalam cluster tersebut.

Proses pembaruan centroid dilakukan dengan cara menghitung nilai rata-rata dari seluruh data yang menjadi anggota dalam setiap

cluster. Nilai rata-rata tersebut kemudian digunakan sebagai centroid baru. Dengan demikian, centroid tidak bersifat tetap, melainkan dapat berubah mengikuti karakteristik data yang tergabung dalam cluster.

Pembaruan centroid bertujuan untuk meningkatkan ketepatan hasil pengelompokan. Apabila centroid baru telah terbentuk, maka setiap data akan kembali dihitung jaraknya terhadap centroid tersebut. Data kemudian dapat tetap berada pada cluster sebelumnya atau berpindah ke cluster lain jika jaraknya lebih dekat dengan centroid cluster yang berbeda.

Tahapan ini dilakukan secara berulang sampai tidak ada lagi perubahan anggota cluster atau sampai posisi centroid dianggap stabil. Kondisi stabil menunjukkan bahwa proses clustering telah mencapai hasil akhir, karena data telah berada pada kelompok yang paling sesuai berdasarkan kemiripan karakteristiknya.

Dalam penerapan K-Means, pembaruan nilai centroid sangat berpengaruh terhadap kualitas hasil cluster. Semakin tepat posisi centroid, semakin baik pula proses pengelompokan data yang dihasilkan. Oleh karena itu, tahap pembaruan centroid menjadi bagian penting dalam memastikan bahwa setiap cluster benar-benar mewakili karakteristik data yang berada di dalamnya.

Dengan adanya pembaruan nilai centroid, algoritma K-Means dapat menyesuaikan pusat cluster secara bertahap hingga menghasilkan kelompok data yang lebih akurat, teratur, dan mudah dianalisis. Proses ini menjadikan K-Means sebagai metode clustering yang sederhana namun efektif dalam menemukan pola atau struktur kelompok pada data.

4.9 Iterasi dalam Algoritma K-Means

Iterasi dalam algoritma K-Means merupakan proses pengulangan tahapan pengelompokan data sampai diperoleh hasil cluster yang stabil. Proses ini menjadi bagian penting karena pada awal pengelompokan, posisi centroid atau titik pusat cluster belum

tentu berada pada posisi yang paling tepat. Oleh karena itu, algoritma K-Means melakukan perhitungan secara berulang agar setiap data dapat ditempatkan pada cluster yang paling sesuai.

Pada tahap awal, algoritma K-Means menentukan jumlah cluster yang akan dibentuk dan memilih centroid awal. Setelah itu, setiap data dihitung jaraknya terhadap masing-masing centroid. Data kemudian ditempatkan pada cluster dengan jarak centroid terdekat. Hasil pengelompokan pertama ini belum tentu menjadi hasil akhir, sehingga diperlukan proses iterasi untuk memperbaiki posisi cluster.

Setelah seluruh data masuk ke dalam cluster tertentu, centroid akan diperbarui dengan menghitung nilai rata-rata dari seluruh data yang berada dalam masing-masing cluster. Centroid baru tersebut kemudian digunakan kembali sebagai acuan untuk menghitung jarak setiap data. Jika terdapat data yang lebih dekat dengan centroid cluster lain, maka data tersebut dapat berpindah ke cluster yang lebih sesuai.

Proses perhitungan jarak, penempatan data ke cluster, dan pembaruan centroid dilakukan secara berulang. Iterasi akan terus berlangsung sampai tidak ada lagi perubahan anggota cluster atau sampai perubahan nilai centroid sangat kecil. Kondisi ini menunjukkan bahwa algoritma telah mencapai titik konvergen, yaitu keadaan ketika hasil pengelompokan dianggap stabil.

Jumlah iterasi yang dibutuhkan dalam algoritma K-Means dapat berbeda-beda tergantung pada jumlah data, jumlah cluster, pemilihan centroid awal, dan karakteristik data yang digunakan. Semakin kompleks data yang dianalisis, semakin besar kemungkinan proses iterasi membutuhkan waktu lebih lama untuk mencapai hasil yang stabil.

Dengan adanya proses iterasi, algoritma K-Means mampu memperbaiki hasil pengelompokan secara bertahap. Iterasi membantu memastikan bahwa setiap data ditempatkan pada cluster yang memiliki karakteristik paling mirip. Oleh karena itu,

iterasi menjadi salah satu tahapan penting dalam menghasilkan cluster yang lebih akurat, konsisten, dan mudah dianalisis.

4.10 Kelebihan dan Kekurangan Algoritma K-Means

Algoritma K-Means merupakan salah satu metode clustering yang banyak digunakan dalam data mining karena memiliki konsep yang sederhana dan mudah diterapkan. Algoritma ini bekerja dengan cara mengelompokkan data ke dalam sejumlah cluster berdasarkan kedekatan jarak antara data dengan titik pusat cluster atau centroid. Meskipun demikian, seperti metode analisis data lainnya, K-Means memiliki kelebihan dan kekurangan yang perlu dipahami sebelum digunakan dalam suatu penelitian.

1. Kelebihan Algoritma K-Means

Salah satu kelebihan utama algoritma K-Means adalah mudah dipahami dan diimplementasikan. Proses kerja K-Means relatif sederhana, yaitu menentukan jumlah cluster, memilih centroid awal, menghitung jarak data ke centroid, mengelompokkan data, kemudian memperbarui centroid hingga hasil cluster stabil. Kesederhanaan ini membuat K-Means cocok digunakan oleh peneliti, mahasiswa, maupun praktisi yang baru mempelajari teknik clustering.

Kelebihan berikutnya adalah proses komputasinya relatif cepat. K-Means mampu mengolah data dalam jumlah besar dengan waktu yang cukup efisien, terutama jika dibandingkan dengan beberapa metode clustering lain yang lebih kompleks. Hal ini menjadikan K-Means banyak digunakan dalam berbagai bidang, seperti pendidikan, pemasaran, kesehatan, sosial, dan bisnis.

Selain itu, K-Means juga memiliki keunggulan dalam menghasilkan kelompok data yang jelas dan mudah diinterpretasikan. Setiap data akan masuk ke dalam satu cluster tertentu berdasarkan kedekatannya dengan centroid. Hasil pengelompokan ini dapat membantu peneliti memahami pola atau karakteristik data,

misalnya kelompok siswa dengan prestasi tinggi, sedang, dan rendah.

K-Means juga cocok digunakan untuk data numerik. Apabila data yang dianalisis berbentuk angka, seperti nilai akademik, jumlah transaksi, usia, pendapatan, atau skor tertentu, maka K-Means dapat bekerja dengan baik. Oleh karena itu, algoritma ini sering digunakan dalam penelitian yang melibatkan data kuantitatif.

Secara umum, kelebihan algoritma K-Means dapat dirangkum sebagai berikut:

- Mudah dipahami dan diterapkan.
- Proses perhitungan relatif cepat.
- Cocok untuk data berjumlah besar.
- Hasil cluster mudah dianalisis.
- Efektif untuk data numerik.
- Dapat digunakan dalam berbagai bidang penelitian.
- Membantu menemukan pola tersembunyi dalam data.

2. Kekurangan Algoritma K-Means

Meskipun memiliki banyak kelebihan, algoritma K-Means juga memiliki beberapa kekurangan. Salah satu kelemahan utamanya adalah jumlah cluster harus ditentukan sejak awal. Pengguna harus menentukan nilai K sebelum proses clustering dilakukan. Jika jumlah cluster yang dipilih tidak tepat, maka hasil pengelompokan dapat menjadi kurang akurat atau kurang sesuai dengan kondisi data sebenarnya.

Kelemahan lain dari K-Means adalah sensitif terhadap pemilihan centroid awal. Centroid awal yang berbeda dapat menghasilkan hasil clustering yang berbeda pula. Jika centroid awal dipilih secara kurang tepat, maka proses clustering dapat menghasilkan kelompok yang tidak optimal.

K-Means juga sensitif terhadap data pencilan atau outlier. Data yang memiliki nilai sangat jauh dari data lainnya dapat memengaruhi posisi centroid. Akibatnya, hasil cluster dapat

bergeser dan tidak mencerminkan pola data secara keseluruhan. Oleh karena itu, sebelum menggunakan K-Means, data perlu melalui tahap pembersihan agar outlier yang mengganggu dapat diminimalkan.

Selain itu, K-Means lebih cocok digunakan untuk data yang memiliki bentuk cluster relatif bulat atau seimbang. Jika data memiliki bentuk cluster yang tidak beraturan, tumpang tindih, atau memiliki ukuran cluster yang sangat berbeda, maka K-Means dapat mengalami kesulitan dalam menghasilkan pengelompokan yang tepat.

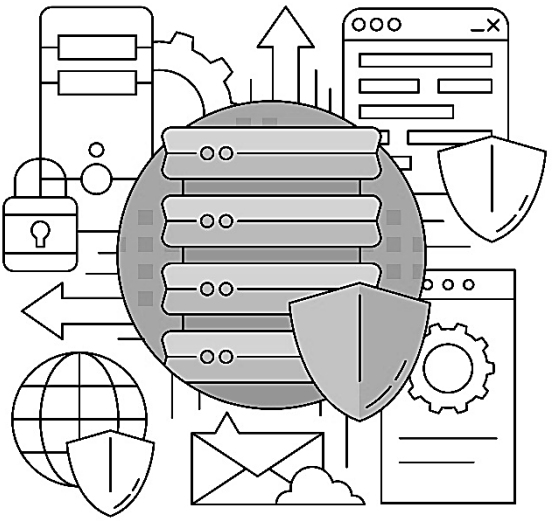
Kelemahan lainnya adalah K-Means hanya dapat bekerja secara optimal pada data numerik. Jika data yang digunakan berbentuk kategorikal, maka data tersebut perlu ditransformasikan terlebih dahulu ke dalam bentuk angka. Proses transformasi ini harus dilakukan dengan hati-hati agar tidak mengubah makna data.

Secara ringkas, kekurangan algoritma K-Means adalah sebagai berikut:

- Jumlah cluster harus ditentukan di awal.
- Hasil clustering dipengaruhi oleh centroid awal.
- Sensitif terhadap outlier.
- Kurang cocok untuk data kategorikal.
- Kurang optimal untuk cluster dengan bentuk tidak beraturan.
- Dapat menghasilkan hasil berbeda pada percobaan yang berbeda.
- Memerlukan normalisasi data jika skala antaratribut berbeda.

Berdasarkan kelebihan dan kekurangannya, algoritma K-Means dapat dikatakan sebagai metode clustering yang sederhana, cepat, dan efektif untuk mengelompokkan data numerik. Algoritma ini sangat berguna untuk menemukan pola kelompok dalam data, terutama pada data yang memiliki struktur relatif jelas.

Namun, penggunaan K-Means tetap memerlukan kehati-hatian, khususnya dalam menentukan jumlah cluster, memilih centroid awal, menangani outlier, dan menyiapkan data sebelum dianalisis. Dengan proses persiapan data yang baik, K-Means dapat menghasilkan pengelompokan yang bermanfaat dalam mendukung proses analisis dan pengambilan keputusan.



Bab 05

IMPLEMENTASI K-MEANS DALAM ANALISIS DATA

5.1 Gambaran Umum Studi Kasus

Program Bantuan Langsung Tunai (BLT) pertama kali diperkenalkan pada tahun 2005 sebagai respons terhadap kenaikan harga bahan bakar minyak (BBM) yang berdampak pada daya beli masyarakat miskin. Program ini diberikan dalam bentuk bantuan tunai langsung kepada rumah tangga miskin untuk membantu memenuhi kebutuhan sehari-hari, menjaga daya beli, serta mengurangi tekanan ekonomi, khususnya pada masa krisis atau kondisi ekonomi yang tidak stabil.

Dalam pelaksanaannya, proses penentuan penerima BLT di tingkat lingkungan atau kelurahan masih sering dilakukan secara manual melalui pendataan sederhana dan musyawarah. Cara ini berpotensi menimbulkan berbagai permasalahan, seperti subjektivitas penilaian, ketidaktepatan sasaran, kurangnya transparansi,

kesalahan pencatatan, serta ketidakkonsistenan dalam pengambilan keputusan.

Permasalahan tersebut juga terlihat dari belum optimalnya pengolahan data sosial ekonomi masyarakat, seperti penghasilan, jumlah tanggungan keluarga, kondisi tempat tinggal, dan kepemilikan aset. Data yang tersedia belum dimanfaatkan secara sistematis untuk mengelompokkan masyarakat berdasarkan tingkat kesejahteraan, sehingga sulit membedakan calon penerima yang sangat layak, layak, dan kurang layak menerima bantuan.

Untuk mengatasi masalah tersebut, penelitian ini mengusulkan penggunaan metode K-Means Clustering dalam mengelompokkan data calon penerima BLT di N2 Aek Nabara. Metode ini digunakan untuk membagi masyarakat ke dalam beberapa cluster berdasarkan kemiripan karakteristik sosial ekonomi. Dengan penerapan K-Means, proses seleksi penerima BLT diharapkan menjadi lebih objektif, tepat sasaran, transparan, dan berbasis data.

5.2 Pengumpulan Data

Teknik pengumpulan data merupakan metode atau cara yang digunakan peneliti untuk memperoleh data yang dibutuhkan dalam penelitian. Pada penelitian ini, teknik pengumpulan data dilakukan untuk mendapatkan informasi mengenai calon penerima Bantuan Langsung Tunai (BLT) di N Dua Aek Nabara yang nantinya digunakan sebagai bahan analisis dan proses pengelompokan menggunakan metode K-Means Clustering. Adapun teknik pengumpulan data yang digunakan dalam penelitian ini adalah sebagai berikut:

1. Observasi

Observasi dilakukan dengan cara melakukan pengamatan secara langsung di N Dua Aek Nabara. Pada tahap ini peneliti mengamati kondisi sosial ekonomi masyarakat, proses pendataan calon penerima BLT, serta mekanisme penentuan penerima bantuan yang dilakukan oleh pihak terkait. Observasi bertujuan untuk

mengetahui kondisi nyata yang terjadi di lapangan sehingga peneliti dapat memahami permasalahan yang berkaitan dengan penentuan penerima BLT.

2. Wawancara

Wawancara dilakukan dengan pihak yang terlibat dalam proses pendataan dan penyaluran Bantuan Langsung Tunai (BLT), seperti perangkat desa atau pihak terkait lainnya. Melalui wawancara, peneliti memperoleh informasi mengenai kriteria penerima BLT, proses seleksi penerima bantuan, serta kendala yang dihadapi dalam menentukan masyarakat yang layak menerima bantuan. Hasil wawancara digunakan sebagai informasi pendukung dalam proses analisis penelitian.

3. Studi Pustaka

Studi pustaka dilakukan dengan mempelajari berbagai sumber referensi seperti buku, jurnal, artikel ilmiah, serta penelitian terdahulu yang berkaitan dengan data mining, Knowledge Discovery in Database (KDD), metode K-Means Clustering, sistem pendukung keputusan, dan program Bantuan Langsung Tunai (BLT). Studi pustaka bertujuan untuk memperkuat landasan teori dan mendukung penelitian agar sesuai dengan metode ilmiah yang digunakan.

5.3. Metode Analisis Data

Metode analisis data merupakan tahapan yang dilakukan untuk mengolah data sehingga menghasilkan informasi yang dapat digunakan dalam menjawab tujuan penelitian. Pada penelitian ini metode analisis data dilakukan menggunakan tahapan Knowledge Discovery in Database (KDD) dengan penerapan metode K-Means Clustering untuk mengelompokkan calon penerima Bantuan Langsung Tunai (BLT). Proses analisis dilakukan secara sistematis agar hasil yang diperoleh lebih objektif dan sesuai dengan kondisi masyarakat.

1. Data Selection

Tahap data selection merupakan proses pemilihan data yang relevan dengan tujuan penelitian. Pada tahap ini peneliti memilih data masyarakat yang berhubungan dengan penentuan penerima BLT. Data yang digunakan meliputi pendapatan, jumlah tanggungan keluarga, kondisi rumah, dan pekerjaan. Pemilihan data dilakukan agar proses analisis menjadi lebih fokus dan sesuai dengan kebutuhan penelitian.

2. Data Preprocessing

Data preprocessing merupakan tahap pembersihan dan persiapan data sebelum dilakukan proses analisis. Pada tahap ini peneliti memeriksa apakah terdapat data yang kosong, tidak lengkap, ganda, atau tidak konsisten. Jika ditemukan kesalahan maka data akan diperbaiki agar dataset menjadi lebih bersih dan siap digunakan dalam proses clustering. Tahap ini sangat penting karena kualitas data akan mempengaruhi hasil pengelompokan yang dihasilkan.

3. Transformation

Transformation atau transformasi data merupakan proses mengubah data ke dalam bentuk yang sesuai dengan kebutuhan metode K-Means Clustering. Pada tahap ini setiap kriteria diberikan bobot numerik berdasarkan tingkat kelayakan penerima BLT. Data kemudian diubah ke dalam bentuk numerik agar dapat dihitung menggunakan metode clustering.

4. Data Mining

Tahap data mining merupakan proses utama dalam penelitian. Pada tahap ini dilakukan penerapan metode K-Means Clustering untuk mengelompokkan masyarakat berdasarkan tingkat kemiripan karakteristik sosial ekonomi yang dimiliki. Pengelompokan dilakukan dengan menentukan jumlah cluster, memilih centroid awal, menghitung jarak menggunakan Euclidean Distance, dan melakukan iterasi hingga diperoleh cluster yang stabil.

5.4 Penentuan Variabel dan Atribut Data

Variabel data clustering merupakan komponen utama yang digunakan dalam proses pengelompokan data menggunakan metode K-Means Clustering. Variabel yang digunakan dalam penelitian ini berasal dari data calon penerima Bantuan Langsung Tunai (BLT) yang diperoleh dari hasil pendataan masyarakat di N Dua Aek Nabara.

Dataset yang digunakan terdiri dari variabel pendapatan (X1), jumlah tanggungan keluarga (X2), kondisi rumah (X3), dan pekerjaan (X4). Variabel-variabel tersebut digunakan sebagai dasar dalam proses pengelompokan masyarakat berdasarkan tingkat kelayakan penerimaan BLT.

1. Pendapatan (X1)

Pendapatan merupakan jumlah penghasilan yang diperoleh kepala keluarga setiap bulan. Variabel ini digunakan untuk mengukur tingkat kemampuan ekonomi masyarakat. Semakin rendah pendapatan yang dimiliki maka semakin tinggi tingkat prioritas untuk menerima BLT.

2. Jumlah Tanggungan Keluarga (X2)

Jumlah tanggungan keluarga menunjukkan banyaknya anggota keluarga yang menjadi tanggungan kepala keluarga. Semakin banyak jumlah tanggungan maka semakin besar kebutuhan ekonomi keluarga sehingga menjadi salah satu pertimbangan dalam menentukan kelayakan penerima BLT.

3. Kondisi Rumah (X3)

Kondisi rumah digunakan untuk menggambarkan tingkat kesejahteraan masyarakat berdasarkan kondisi tempat tinggal yang dimiliki. Rumah dengan kondisi kurang layak menunjukkan tingkat ekonomi yang lebih rendah dibandingkan rumah yang layak huni.

4. Pekerjaan (X4)

Pekerjaan menunjukkan jenis pekerjaan yang dimiliki kepala keluarga. Variabel ini digunakan untuk melihat tingkat kestabilan penghasilan masyarakat. Pekerjaan yang tidak tetap umumnya memiliki tingkat risiko ekonomi yang lebih tinggi dibandingkan pekerjaan tetap.

5. Hasil Cluster

Hasil akhir dari proses K-Means Clustering berupa pengelompokan masyarakat ke dalam beberapa cluster berdasarkan tingkat kemiripan data yang dimiliki. Dalam penelitian ini cluster dibagi menjadi tiga kategori yaitu:

- Cluster 1 (C1) : Sangat Layak menerima BLT.
- Cluster 2 (C2) : Layak menerima BLT.
- Cluster 3 (C3) : Kurang Layak menerima BLT.

Hasil cluster tersebut digunakan sebagai dasar rekomendasi dalam membantu proses penentuan penerima Bantuan Langsung Tunai (BLT) secara objektif, transparan, dan tepat sasaran.

Tabel 3.1 DataSet

No	Nama KK	Pendapatan	Tanggungan	Rumah	Pekerjaan
1	Suparman	Rp1.250.000	3	5	5
2	Hendra Wijaya	Rp950.000	7	5	5
3	Siti Rahayu	Rp2.850.000	5	3	3
...	...	...	...	...	...

X1 = Pendapatan (Rp)
X2 = Jumlah Tanggungan
X3 = Kondisi Rumah (1-5)
X4 = Pekerjaan (1-5)

Keterangan:

Kondisi Rumah: 1=Sangat Baik, 2=Baik, 3=Sedang, 4=Buruk, 5=Sangat Buruk

Pekerjaan: 1=Pegawai, 2=Wiraswasta, 3=Petani, 4=Buruh, 5=Tidak Tetap

Contoh Perhitungan Nilai Kriteria Hendra Wijaya

Pendapatan = Rp950.000

$$X_1 = 5$$

Jumlah Tanggungan = 7 orang

$$X_2 = 5$$

Kondisi Rumah = Sangat Buruk

$$X_3 = 5$$

Pekerjaan = Buruh Harian

$$X_4 = 5$$

Sehingga diperoleh nilai:

Nama KK	X1	X2	X3	X4
Hendra Wijaya	5	5	5	5

Atau jika yang dimaksud dosen adalah **centroid C1 = 3,9**, tulis:

$$C1 = \frac{\sum X}{n}$$

$$C1 = \frac{78}{20}$$

$$C1 = 3{,}9$$

5.5 Tahapan Praproses Data

Seleksi data merupakan tahapan dalam preprocessing yang dilakukan untuk memilih variabel atau atribut data yang relevan dengan tujuan penelitian. Pada tahap ini, dataset calon penerima Bantuan Langsung Tunai (BLT) di N Dua Aek Nabara diseleksi untuk menentukan variabel yang akan digunakan dalam proses pengelompokan menggunakan metode K-Means Clustering.

Variabel yang digunakan dalam penelitian ini meliputi pendapatan, jumlah tanggungan keluarga, kondisi rumah, dan pekerjaan**.** Keempat variabel tersebut dipilih karena dianggap mampu menggambarkan kondisi sosial ekonomi masyarakat yang menjadi dasar dalam menentukan tingkat kelayakan penerima BLT. Sementara itu, atribut Nama Kepala Keluarga (KK) tidak digunakan dalam proses perhitungan clustering karena hanya berfungsi sebagai identitas data. Proses seleksi data dilakukan agar data yang digunakan lebih terstruktur, relevan, dan sesuai untuk proses analisis menggunakan metode K-Means Clustering. Hasil dari tahap seleksi data adalah dataset yang hanya berisi atribut-atribut yang berpengaruh terhadap pengelompokan calon penerima BLT sehingga proses clustering dapat dilakukan dengan lebih efektif dan menghasilkan kelompok yang lebih akurat.

Tabel 3.2 Data Seleksi

No	Nama KK	Pendapatan	Tanggungan	Kondisi Rumah	Pekerjaan	Status
1	Suparman	Rp1.250.000	3	5	5	Digunakan
2	Hendra Wijaya	Rp950.000	7	5	5	Digunakan
3	Siti Rahayu	Rp2.850.000	5	3	3	Digunakan
…	…	…	…	…	…	…

5.6 Transformasi Data untuk K-Means

Transformasi data merupakan proses mengubah data mentah ke dalam bentuk numerik yang sesuai untuk proses perhitungan menggunakan metode K-Means Clustering. Pada tahap ini, data calon penerima Bantuan Langsung Tunai (BLT) yang terdiri dari pendapatan, jumlah tanggungan, kondisi rumah, dan pekerjaan diubah menjadi nilai bobot berdasarkan kriteria yang telah ditentukan.

Tujuan transformasi data adalah untuk menyamakan skala data sehingga seluruh variabel dapat diproses dalam perhitungan jarak (distance) pada metode K-Means Clustering.

Tabel 3.3 Transformasi Data

Nama KK	Pendapatan	Tanggungan	Kondisi Rumah	Pekerjaan	X 1	X 2	X 3	X 4
Suparman	Rp1.250.000	3	5	5	4	2	5	5
Hendra Wijaya	Rp950.000	7	5	5	5	5	5	5
Siti Rahayu	Rp2.850.000	5	3	3	3	4	3	3
...	...	...	...	...	...	...	...	...

5.7 Penerapan Algoritma K-Means

Setelah data calon penerima Bantuan Langsung Tunai (BLT) melalui tahap seleksi dan transformasi data, langkah selanjutnya adalah melakukan proses clustering menggunakan metode K-Means. Metode K-Means digunakan untuk mengelompokkan data berdasarkan tingkat kemiripan karakteristik yang dimiliki oleh setiap kepala keluarga.

Dalam penelitian ini digunakan 3 cluster (K=3), yaitu:

Tabel 3.4 Kategori Cluster

Cluster	Keterangan
C1	Sangat Layak Menerima BLT
C2	Layak Menerima BLT
C3	Tidak Layak Menerima BLT

Adapun tahapan K-Means yang digunakan dalam penelitian ini adalah sebagai berikut:

- Menentukan jumlah cluster (K)
- Menentukan centroid awal
- Menghitung jarak setiap data terhadap centroid
- Menentukan cluster berdasarkan jarak terdekat
- Menghitung centroid baru

- Melakukan iterasi hingga cluster tidak berubah

1. Menentukan Centroid Awal

Centroid awal dipilih secara manual untuk mewakili kelompok masyarakat dengan kondisi ekonomi rendah, sedang, dan tinggi.

Tabel 3.5 Centroid Awal

Cluster	X1	X2	X3	X4
C1	5	5	5	5
C2	3	2	2	3
C3	1	1	1	1

Keterangan:

X1 = Pendapatan
X2 = Jumlah Tanggungan
X3 = Kondisi Rumah
X4 = Pekerjaan

2. Menghitung Jarak Euclidean Distance

Setiap data dihitung jaraknya terhadap masing-masing centroid menggunakan rumus Euclidean Distance.

Rumus:

$$d(x,y)=\sqrt{[(x1-y1)^2+(x2-y2)^2+(x3-y3)^2+(x4-y4)^2]}$$

Keterangan:

d(x,y) = jarak data terhadap centroid
x = nilai data
y = nilai centroid

3. Menentukan Cluster

Data akan dimasukkan ke cluster yang memiliki nilai jarak paling kecil terhadap centroid.

Jarak ke C1

$$d_{C1} = \sqrt{(5-5)^2 + (5-5)^2 + (5-5)^2 + (5-5)^2}$$

$$= \sqrt{0+0+0+0}$$

$$= \sqrt{0}$$

$$= 0$$

Jarak ke C2

$$d_{C2} = \sqrt{(5-3)^2 + (5-2)^2 + (5-2)^2 + (5-3)^2}$$

$$= \sqrt{4+9+9+4}$$

$$= \sqrt{26}$$

$$= 5{,}099$$

$$\approx 5{,}10$$

Jarak ke C3

$$d_{C3} = \sqrt{(5-1)^2 + (5-1)^2 + (5-1)^2 + (5-1)^2}$$

$$= \sqrt{16+16+16+16}$$

$$= \sqrt{64}$$

$$= 8$$

Penentuan Cluster

Karena:

$$d_{C1} = 0 < d_{C2} = 5{,}10 < d_{C3} = 8$$

maka data Hendra Wijaya masuk ke Cluster C1.

4. Menghitung Centroid Baru

Setelah seluruh data mendapatkan cluster, dilakukan perhitungan centroid baru dengan mencari rata-rata setiap variabel dalam cluster.

Rumus:

Centroid Baru = ΣX / n

Keterangan:

ΣX = jumlah data dalam cluster
n = jumlah anggota cluster

5. Iterasi

Perhitungan jarak dan pembentukan centroid baru dilakukan secara berulang hingga tidak terjadi perubahan anggota cluster. Hasil akhir berupa kelompok calon penerima BLT yang terbagi menjadi Sangat Layak, Layak, dan Tidak Layak menerima BLT.

5.8 Analisis Hasil Cluster

Hasil clustering merupakan tahap akhir dari proses K-Means yang menunjukkan pengelompokan data calon penerima BLT berdasarkan karakteristik yang dimiliki. Setelah proses iterasi dilakukan hingga tidak terjadi perubahan anggota cluster, maka diperoleh cluster akhir yang stabil.

Data yang telah dikelompokkan kemudian dibagi ke dalam tiga kategori, yaitu:

Cluster	Keterangan
C1	Sangat Layak Menerima BLT
C2	Layak Menerima BLT
C3	Tidak Layak Menerima BLT

Hasil clustering digunakan sebagai dasar dalam menentukan calon penerima BLT yang sesuai dengan kriteria yang telah ditetapkan.

Tabel 3.6 Hasil Clustering

D(C1)	D(C2)	D(C3)	Jarak Min	Cluster
3,16227766	3,741657387	6,480740698	3,16227766	C1
0	5,099019514	8	0	C1
3,60555128	2,236067977	4,582575695	2,236067977	C2
...	...	...	...	...

5.9 Interpretasi Setiap Cluster

Tahap implementasi dilakukan menggunakan RapidMiner untuk mengolah data penelitian penerima Bantuan Langsung Tunai (BLT) di Desa N2 Aek Nabara. Dataset yang telah dikumpulkan dari data penduduk dan kondisi ekonomi masyarakat diimpor ke dalam RapidMiner melalui operator *Read Excel*. Selanjutnya, dilakukan proses pengolahan data dan penerapan algoritma yang digunakan untuk mengelompokkan masyarakat berdasarkan tingkat kelayakan penerimaan bantuan. Setelah model terbentuk, dilakukan pengujian untuk memastikan hasil pengelompokan berjalan dengan baik. Hasil yang diperoleh kemudian dianalisis sebagai dasar dalam menentukan calon penerima BLT yang lebih tepat sasaran di Desa N2 Aek Nabara.

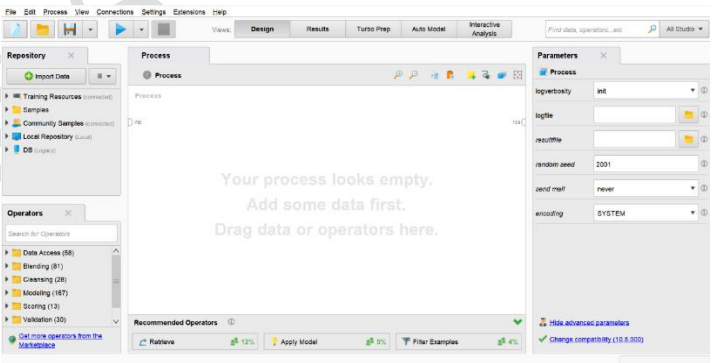
1. Import Data

Tahap import data dilakukan dengan memasukkan dataset penerima Bantuan Langsung Tunai (BLT) Desa N2 Aek Nabara ke dalam RapidMiner menggunakan operator *Read Excel*. Dataset

yang digunakan berisi data masyarakat seperti pendapatan, jumlah tanggungan, kondisi rumah, dan pekerjaan. Setelah data berhasil diimpor, dilakukan pemeriksaan untuk memastikan seluruh atribut dan data telah terbaca dengan benar sehingga siap digunakan pada proses pengolahan dan analisis selanjutnya.

Gambar 4.1 Proses Membuka Aplikasi Rapidminer

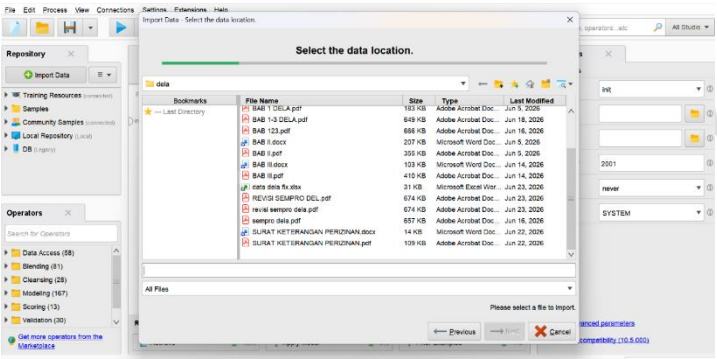
menunjukkan tampilan awal aplikasi RapidMiner versi 10.1 saat dijalankan. Pada tahap ini, aplikasi sedang melakukan proses pemuatan (*loading*) sistem dan ekstensi yang diperlukan sebelum pengguna dapat mengakses halaman utama. Setelah proses pemuatan selesai, pengguna dapat mulai mengimpor dataset dan melakukan pengolahan data untuk menerapkan algoritma yang digunakan dalam penelitian.



Gambar 4.2 Lembar Kerja Utama Aplikasi Rapidminer

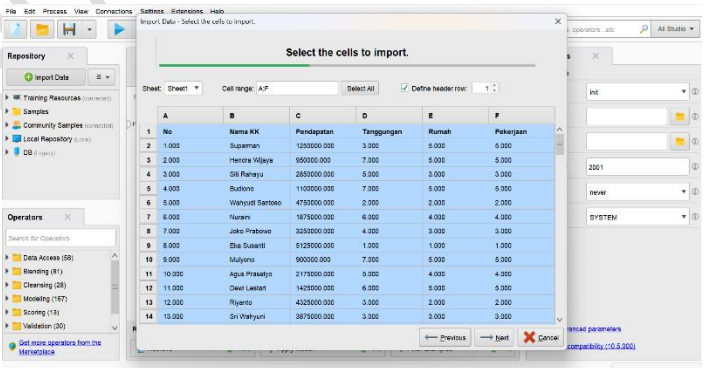
Gambar 4.2 menunjukkan tampilan lembar kerja utama RapidMiner yang digunakan untuk melakukan proses pengolahan

data. Pada tampilan ini terdapat beberapa bagian utama, yaitu *Repository* untuk menyimpan dataset, *Operators* yang berisi berbagai operator pengolahan data, area *Process* sebagai tempat merancang alur proses analisis, serta *Parameters* untuk mengatur konfigurasi operator yang digunakan. Melalui lembar kerja ini, pengguna dapat mengimpor data, membangun model, dan menjalankan proses analisis sesuai dengan kebutuhan penelitian.



Gambar 4.3 Direktori File Import Data

Gambar 4.3 menunjukkan proses pemilihan lokasi file yang akan diimpor ke dalam aplikasi RapidMiner. Pada tahap ini, peneliti memilih dataset yang tersimpan pada direktori komputer untuk digunakan dalam proses pengolahan data. Pemilihan file ini merupakan langkah awal sebelum data dimasukkan ke repository RapidMiner dan selanjutnya digunakan dalam proses clustering menggunakan metode K-Means.



Gambar 4.4 Select The Cells To Import

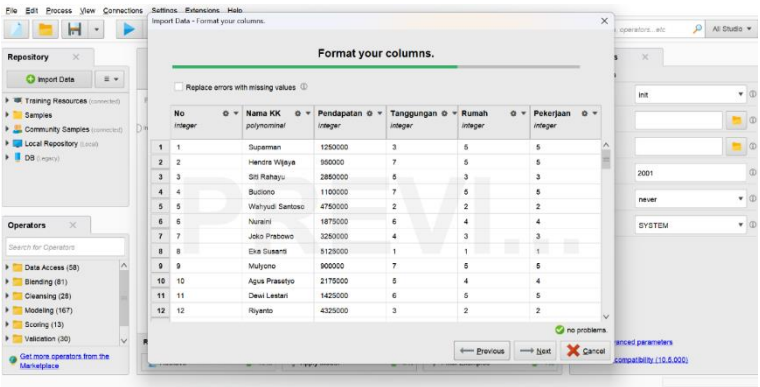
Gambar 4.4 menunjukkan tahap pemilihan sel data yang akan diimpor ke dalam RapidMiner. Pada tahap ini, peneliti menentukan rentang data yang akan digunakan, mulai dari kolom A sampai F, serta mengaktifkan opsi Define header row untuk menjadikan baris pertama sebagai nama atribut. Data yang dipilih terdiri dari atribut No, Nama KK, Pendapatan, Tanggungan, Rumah, dan Pekerjaan yang akan digunakan dalam proses clustering menggunakan metode K-Means.

Tabel 4.1 Keterangan Atribut Data

No	Nama Atribut	Tipe Data	Keterangan
1	No	Integer	Nomor urut data kepala keluarga.
2	Nama KK	String	Nama kepala keluarga yang menjadi objek penelitian.
3	Pendapatan	Numerik	Jumlah pendapatan kepala keluarga dalam satu bulan (Rp).
4	Tanggungan	Numerik	Jumlah anggota keluarga yang menjadi tanggungan kepala keluarga.
5	Rumah	Numerik	Kondisi atau status rumah yang dimiliki kepala keluarga berdasarkan kriteria yang telah ditentukan.
6	Pekerjaan	Numerik	Kategori pekerjaan kepala keluarga yang telah dikonversi ke dalam bentuk angka.

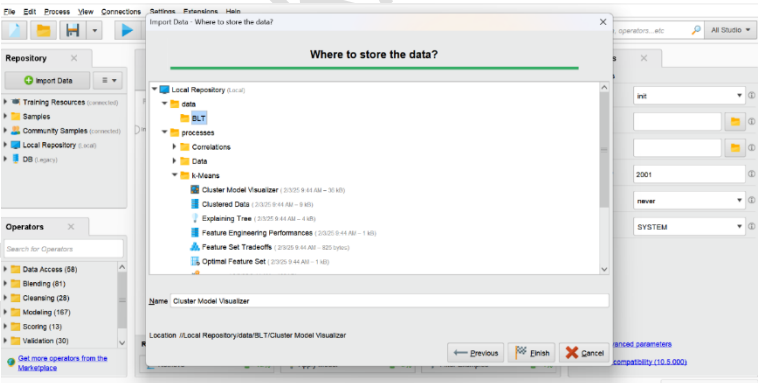
Penjelasan:

- Atribut Nama KK berfungsi sebagai identitas data dan tidak digunakan dalam perhitungan K-Means.
- Atribut Pendapatan, Tanggungan, Rumah, dan Pekerjaan digunakan sebagai variabel dalam proses clustering.
- Atribut No hanya digunakan sebagai penanda data dan tidak ikut dalam perhitungan cluster.



Gambar 4.5 Set Role Variabel

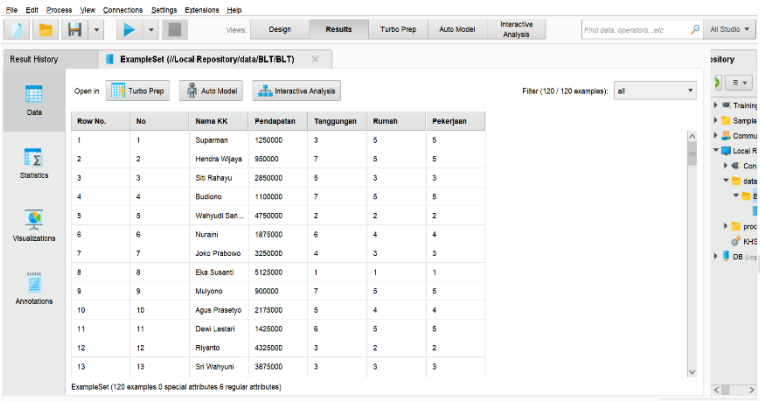
Gambar 4.5 menunjukkan tahap pengaturan format atau tipe data setiap atribut pada RapidMiner. Pada tahap ini atribut No, Pendapatan, Tanggungan, Rumah, dan Pekerjaan ditetapkan sebagai data bertipe integer, sedangkan atribut Nama KK ditetapkan sebagai polynominal karena berisi data kategori berupa nama. Pengaturan tipe data dilakukan agar proses pengolahan dan clustering menggunakan metode K-Means dapat berjalan dengan benar



Gambar 4.6 Direktori Penyimpanan File

Gambar 4.6 menunjukkan tahap penentuan lokasi penyimpanan dataset pada repository RapidMiner setelah proses impor data selesai dilakukan. Pada tahap ini, dataset disimpan ke dalam folder BLT yang berada pada Local Repository, sehingga data dapat digunakan kembali dalam proses pengolahan dan analisis

menggunakan metode K-Means Clustering. Penyimpanan data pada repository memudahkan pengelolaan dataset serta akses data pada tahap pemodelan berikutnya.



Gambar 4.7 Preview Dataset

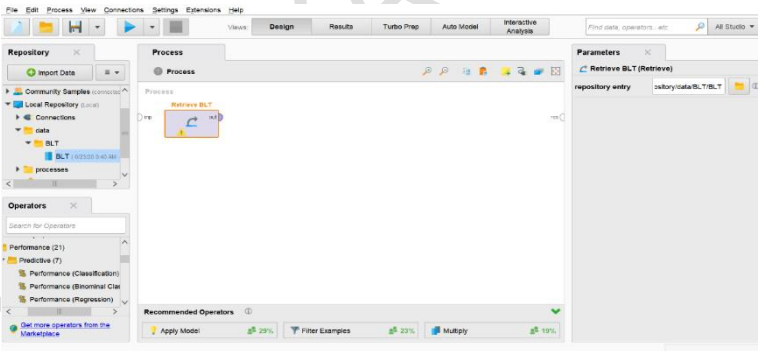
Gambar 4.7 menunjukkan tampilan preview dataset yang telah berhasil diimpor ke dalam RapidMiner. Pada tahap ini, seluruh data ditampilkan dalam bentuk tabel sehingga peneliti dapat memeriksa kesesuaian atribut dan nilai data sebelum dilakukan proses clustering. Dataset terdiri dari atribut No, Nama KK, Pendapatan, Tanggungan, Rumah, dan Pekerjaan yang akan digunakan sebagai dasar pengelompokan penerima Bantuan Langsung Tunai (BLT) menggunakan metode K-Means Clustering.

Melalui tampilan preview ini, peneliti dapat memastikan bahwa setiap atribut telah memiliki tipe data yang sesuai. Atribut Pendapatan, Tanggungan, Rumah, dan Pekerjaan telah terbaca sebagai data numerik sehingga dapat digunakan dalam perhitungan jarak Euclidean pada algoritma K-Means. Selain itu, tidak ditemukan data yang kosong (*missing value*) maupun kesalahan format yang dapat memengaruhi hasil proses clustering. Pemeriksaan ini dilakukan untuk meningkatkan kualitas data sehingga hasil pengelompokan yang dihasilkan lebih akurat dan dapat dipertanggungjawabkan

2. Pra-pemrosesan Data

Pra-pemrosesan data (*data preprocessing*) merupakan tahapan penting yang dilakukan sebelum proses clustering menggunakan metode K-Means. Tahap ini bertujuan untuk memastikan bahwa data yang digunakan telah sesuai dengan kebutuhan analisis sehingga dapat menghasilkan pengelompokan yang lebih akurat. Pada penelitian ini, proses pra-pemrosesan dilakukan menggunakan aplikasi RapidMiner terhadap data calon penerima Bantuan Langsung Tunai (BLT) yang terdiri dari atribut Pendapatan, Tanggungan, Rumah, dan Pekerjaan.

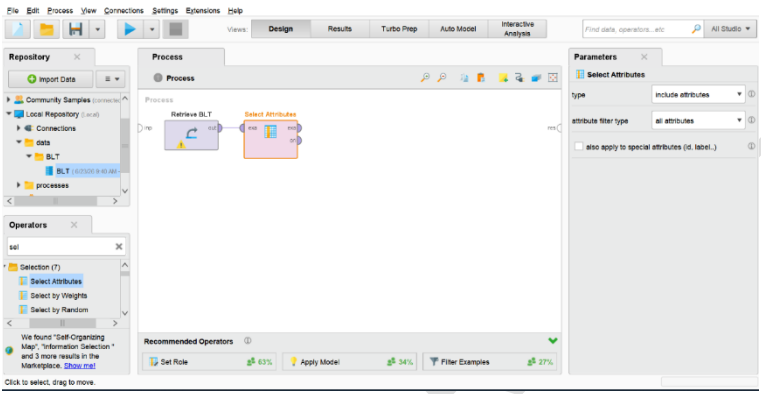
Tahap pertama yang dilakukan adalah pengumpulan dan impor data ke dalam RapidMiner. Data yang sebelumnya tersimpan dalam format Microsoft Excel (*.xlsx*) diimpor ke dalam repository RapidMiner agar dapat diolah pada tahap berikutnya. Setelah proses impor selesai, dilakukan pemeriksaan terhadap struktur data untuk memastikan seluruh atribut dan nilai data telah terbaca dengan benar.



Gambar 4.8 Retrieve Data

Gambar 4.8 menunjukkan proses pemanggilan dataset yang telah tersimpan pada repository RapidMiner menggunakan operator Retrieve. Operator ini berfungsi untuk mengambil data BLT dari folder repository sehingga dapat digunakan dalam proses pengolahan dan analisis data. Tahap retrieve merupakan langkah awal dalam penyusunan model clustering karena seluruh data yang akan diproses harus terlebih dahulu dipanggil ke area kerja

(process). Dataset yang digunakan berisi data kepala keluarga dengan atribut Pendapatan, Tanggungan, Rumah, dan Pekerjaan yang akan digunakan sebagai dasar pengelompokan menggunakan metode K-Means Clustering. Setelah data berhasil dipanggil melalui operator Retrieve, proses dapat dilanjutkan ke tahap seleksi atribut dan pembentukan cluster.



Gambar 4.9 Select Attributes

Pada Gambar 4.9 ditunjukkan penggunaan operator Select Attributes pada aplikasi RapidMiner. Operator ini digunakan untuk memilih atribut atau variabel yang akan digunakan dalam proses analisis K-Means Clustering. Tujuannya adalah agar hanya atribut yang relevan yang diproses, sehingga hasil clustering menjadi lebih optimal dan sesuai dengan kebutuhan penelitian.

Pada penelitian ini, atribut yang dipilih adalah Pendapatan, Tanggungan, Kondisi Rumah, dan Pekerjaan, sedangkan atribut lain yang tidak diperlukan, seperti nomor identitas atau kode data, dapat dikecualikan dari proses pengelompokan. Pengaturan dilakukan pada panel Parameters dengan memilih tipe *include attributes*, sehingga hanya atribut yang dipilih yang akan diteruskan ke tahap berikutnya.

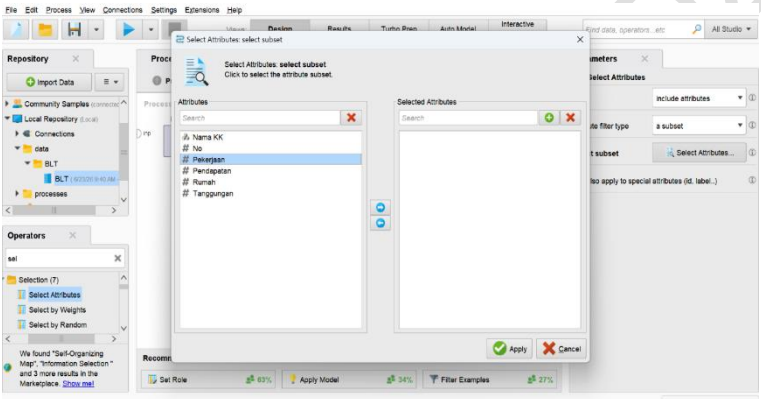
Penggunaan operator Select Attributes membantu mengurangi data yang tidak relevan (*irrelevant attributes*), mempercepat proses komputasi, serta meningkatkan kualitas hasil

pengelompokan data penerima Bantuan Langsung Tunai (BLT) menggunakan metode K-Means Clustering.

Keterangan Gambar 4.9:

Retrieve BLT : Mengambil data BLT dari repository RapidMiner. Select Attributes : Memilih atribut yang digunakan dalam proses clustering.

Output (exa) : Menghasilkan dataset yang telah diseleksi dan siap digunakan pada proses K-Means Clustering.



Gambar 4.10 Memilih Select Attributes

Pada Gambar 4.10 ditunjukkan proses pemilihan atribut yang akan digunakan dalam analisis menggunakan operator Select Attributes pada RapidMiner. Pada jendela *Select Attributes: select subset*, tersedia beberapa atribut yaitu Nama KK, No, Pekerjaan, Pendapatan, Rumah, dan Tanggungan.

Dalam penelitian ini, atribut yang dipilih untuk proses K-Means Clustering adalah Pendapatan, Tanggungan, Rumah, dan Pekerjaan karena atribut tersebut menjadi variabel utama dalam menentukan tingkat kelayakan penerima Bantuan Langsung Tunai (BLT). Sementara itu, atribut Nama KK dan No tidak digunakan karena hanya berfungsi sebagai identitas data dan tidak berpengaruh terhadap proses pengelompokan.

Pemilihan atribut dilakukan dengan memindahkan atribut yang diperlukan dari kolom Attributes ke kolom Selected Attributes menggunakan tombol panah yang tersedia. Setelah seluruh atribut yang dibutuhkan dipilih, pengguna menekan tombol Apply untuk menyimpan konfigurasi dan melanjutkan ke tahap berikutnya.

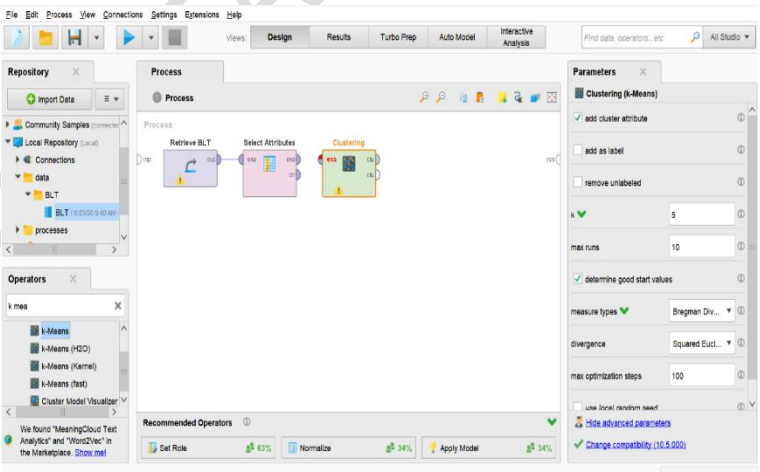
Melalui proses ini, dataset yang digunakan dalam analisis hanya berisi atribut-atribut yang relevan sehingga proses clustering dapat menghasilkan kelompok data yang lebih akurat dan sesuai dengan tujuan penelitian.

Atribut yang dipilih:

- Pendapatan = menggambarkan tingkat penghasilan keluarga.
- Tanggungan = menunjukkan jumlah anggota keluarga yang menjadi tanggungan.
- Rumah = menggambarkan kondisi tempat tinggal keluarga.
- Pekerjaan = menunjukkan jenis pekerjaan kepala keluarga.

Atribut yang tidak dipilih:

- Nama KK = hanya sebagai identitas kepala keluarga.
- No hanya = sebagai nomor urut data.



Gambar 4.11 K-Means

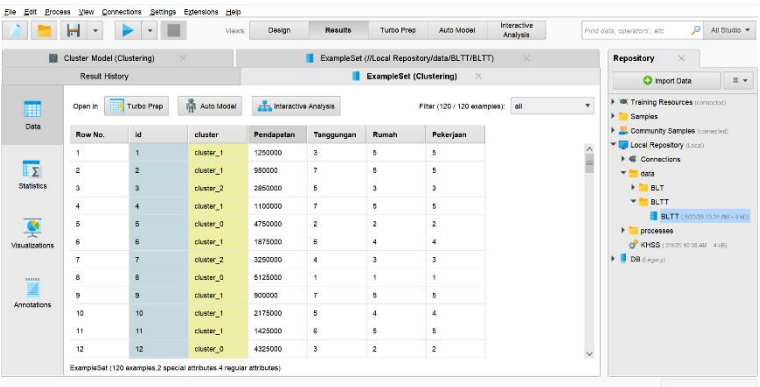
Pada Gambar 4.11 ditunjukkan proses penerapan operator K-Means Clustering pada RapidMiner.Operator K-Means digunakan untuk mengelompokkan data penerima Bantuan Langsung Tunai (BLT) berdasarkan kemiripan karakteristik yang dimiliki setiap Kepala Keluarga (KK). Data yang telah melalui tahap Select Attributes kemudian diproses menggunakan metode K-Means untuk menghasilkan beberapa kelompok (cluster).

Pada panel Parameters, nilai k = 3 ditentukan sebagai jumlah cluster yang akan dibentuk. Pemilihan tiga cluster disesuaikan dengan tujuan penelitian, yaitu mengelompokkan data ke dalam kategori:

- Cluster 1 (C1) : Sangat Layak Menerima BLT.
- Cluster 2 (C2) : Layak Menerima BLT.
- Cluster 3 (C3) : Tidak Layak Menerima BLT.

Metode K-Means bekerja dengan menentukan centroid awal, kemudian menghitung jarak setiap data terhadap centroid menggunakan rumus Euclidean Distance. Setiap data ditempatkan pada cluster dengan jarak terdekat. Proses iterasi dilakukan secara berulang hingga posisi centroid stabil dan tidak terjadi perpindahan anggota cluster secara signifikan.

Pada gambar terlihat bahwa output dari Select Attributes dihubungkan ke input operator Clustering (K-Means) melalui port exa. Hasil dari proses ini berupa model cluster dan data yang telah memiliki atribut cluster, sehingga setiap Kepala Keluarga dapat diketahui termasuk ke dalam kelompok mana.



Gambar 4.12 Hasil Clustering

Pada Gambar 4.12 ditampilkan hasil proses K-Means Clustering yang telah dijalankan menggunakan RapidMiner. Hasil clustering berupa pengelompokan data Kepala Keluarga (KK) ke dalam beberapa cluster berdasarkan atribut Pendapatan, Tanggungan, Rumah, dan Pekerjaan. Setiap data memperoleh label cluster yang menunjukkan kelompok tempat data tersebut berada.

Pada tabel hasil clustering terlihat adanya tiga kelompok, yaitu cluster_0, cluster_1, dan cluster_2. Masing-masing cluster memiliki karakteristik yang berbeda sesuai dengan tingkat kemiripan data yang dihitung oleh algoritma K-Means. Pengelompokan ini dilakukan berdasarkan jarak terdekat antara data dengan centroid menggunakan metode Euclidean Distance.

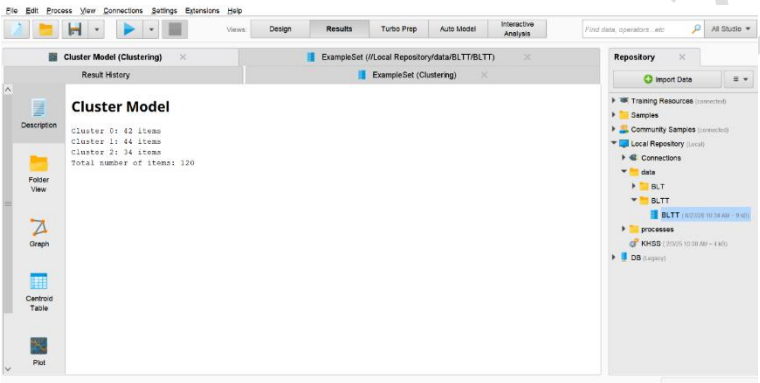
Kolom cluster menunjukkan hasil klasifikasi setiap Kepala Keluarga. Sebagai contoh, data dengan ID 1 masuk ke cluster_1, data dengan ID 3 masuk ke cluster_2, dan data dengan ID 5 masuk ke cluster_0. Hasil tersebut menunjukkan bahwa setiap data telah berhasil ditempatkan pada kelompok yang memiliki karakteristik paling mirip.

Dalam penelitian ini, hasil cluster kemudian diinterpretasikan sebagai berikut:

- Cluster 0 (C1) : Tidak Layak Menerima BLT.
- Cluster 1 (C2) : Layak Menerima BLT.

- Cluster 2 (C3) : Sangat Layak Menerima BLT.

Interpretasi cluster dilakukan dengan menganalisis nilai rata-rata atribut pada setiap kelompok. Cluster yang memiliki pendapatan lebih rendah, jumlah tanggungan lebih banyak, kondisi rumah kurang baik, dan pekerjaan tidak tetap dikategorikan sebagai Sangat Layak Menerima BLT. Sebaliknya, cluster dengan kondisi ekonomi lebih baik dikategorikan sebagai Tidak Layak Menerima BLT.



Gambar 4.13 Cluster Model

Pada Gambar 4.13 ditampilkan Cluster Model yang dihasilkan setelah proses K-Means Clustering selesai dijalankan pada RapidMiner. Tampilan ini menunjukkan jumlah anggota yang terdapat pada setiap cluster hasil pengelompokan data penerima Bantuan Langsung Tunai (BLT).

Berdasarkan hasil clustering, diperoleh tiga cluster dengan jumlah anggota sebagai berikut:

Cluster	Jumlah Data
Cluster 0	42 data
Cluster 1	44 data
Cluster 2	34 data
Total	120 data

Hasil tersebut menunjukkan bahwa seluruh 120 data Kepala Keluarga (KK) berhasil dikelompokkan ke dalam tiga cluster sesuai

dengan karakteristik yang dimiliki berdasarkan atribut Pendapatan, Tanggungan, Rumah, dan Pekerjaan. Setiap cluster memiliki jumlah anggota yang berbeda karena proses pengelompokan dilakukan berdasarkan tingkat kemiripan data terhadap centroid masing-masing cluster.

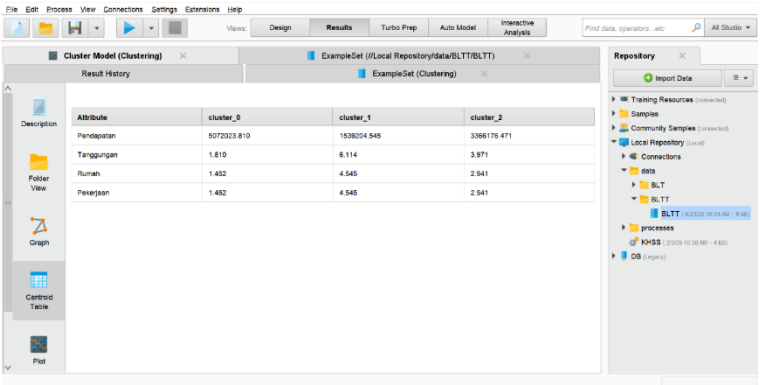
Distribusi anggota cluster menunjukkan bahwa Cluster 1 memiliki jumlah anggota terbanyak yaitu 44 data atau sekitar 36,67% dari total data. Selanjutnya Cluster 0 berjumlah 42 data atau sekitar 35,00%, sedangkan Cluster 2 memiliki 34 data atau sekitar 28,33% dari keseluruhan data.

Dalam penelitian ini, hasil cluster kemudian diinterpretasikan sebagai berikut:

- Cluster 0 : Tidak Layak Menerima BLT (42 KK)
- Cluster 1 : Layak Menerima BLT (44 KK)
- Cluster 2 : Sangat Layak Menerima BLT (34 KK)

Interpretasi tersebut diperoleh berdasarkan analisis karakteristik rata-rata pada setiap cluster. Cluster dengan kondisi ekonomi paling rendah, jumlah tanggungan lebih banyak, kondisi rumah kurang layak, dan pekerjaan tidak tetap dikategorikan sebagai Sangat Layak Menerima BLT, sedangkan cluster dengan kondisi ekonomi lebih baik dikategorikan sebagai Tidak Layak Menerima BLT.

Dengan adanya Cluster Model, peneliti dapat mengetahui jumlah anggota pada setiap kelompok sehingga memudahkan proses evaluasi dan penentuan sasaran penerima bantuan. Hasil ini menunjukkan bahwa metode K-Means Clustering mampu mengelompokkan data masyarakat secara efektif berdasarkan tingkat kemiripan karakteristik sosial dan ekonomi yang dimiliki oleh setiap Kepala Keluarga



Gambar 4.14 Ceteroid

Pada Gambar 4.14 ditampilkan Centroid Table yang merupakan pusat cluster (*centroid*) hasil proses K-Means Clustering. Centroid merupakan nilai rata-rata dari setiap atribut pada masing-masing cluster yang digunakan sebagai acuan dalam proses pengelompokan data. Nilai centroid menunjukkan karakteristik utama yang dimiliki oleh anggota dalam setiap cluster.

Berdasarkan hasil pada tabel centroid, diperoleh nilai sebagai berikut:

Atribut	Cluster 0	Cluster 1	Cluster 2
Pendapatan	507.023,810	1.539.204,545	3.366.176,471
Tanggungan	1,810	6,114	3,971
Rumah	1,452	4,545	2,941
Pekerjaan	1,452	4,545	2,941

Berdasarkan nilai centroid tersebut dapat diketahui karakteristik masing-masing cluster sebagai berikut:

Cluster 0

Cluster 0 memiliki rata-rata pendapatan sebesar Rp507.023,81, jumlah tanggungan 1,81 orang, nilai kondisi rumah 1,45, dan nilai pekerjaan 1,45. Pendapatan yang relatif rendah menunjukkan bahwa anggota cluster ini berada pada kelompok ekonomi menengah ke bawah.

Cluster

Cluster 1 memiliki rata-rata pendapatan sebesar Rp1.539.204,55, jumlah tanggungan 6,11 orang, nilai kondisi rumah 4,55, dan nilai pekerjaan 4,55. Cluster ini memiliki jumlah tanggungan paling tinggi dibandingkan cluster lainnya sehingga dapat menjadi kelompok yang membutuhkan perhatian dalam penyaluran bantuan.

Cluster 2

Cluster 2 memiliki rata-rata pendapatan sebesar Rp3.366.176,47, jumlah tanggungan 3,97 orang, nilai kondisi rumah 2,94, dan nilai pekerjaan 2,94. Pendapatan yang lebih tinggi dibandingkan cluster lainnya menunjukkan bahwa kelompok ini memiliki kondisi ekonomi yang relatif lebih baik.

Dari hasil analisis centroid, peneliti dapat memberikan interpretasi terhadap tingkat kelayakan penerima BLT sebagai berikut:

- Cluster 0 : Sangat Layak Menerima BLT
- Cluster 1 : Layak Menerima BLT
- Cluster 2 : Tidak Layak Menerima BLT

Analisis centroid sangat penting karena digunakan untuk memahami karakteristik setiap cluster yang terbentuk. Dengan melihat nilai rata-rata pada setiap atribut, peneliti dapat menentukan kelompok masyarakat yang paling berhak menerima bantuan secara objektif berdasarkan kondisi sosial dan ekonomi yang dimiliki. Hasil ini menunjukkan bahwa metode K-Means Clustering mampu memberikan pengelompokan data yang jelas dan dapat dijadikan dasar dalam proses pengambilan keputusan penentuan penerima Bantuan Langsung Tunai (BLT).

5.10 Evaluasi Hasil Pengelompokan Data

Evaluasi dilakukan untuk mengetahui kualitas hasil pengelompokan yang dihasilkan oleh metode K-Means. Pada penelitian ini evaluasi dilakukan dengan melihat tingkat kesamaan

karakteristik anggota pada masing-masing cluster dan perbedaan karakteristik antar cluster.

Selain itu, evaluasi dapat dilakukan menggunakan nilai Within Cluster Sum of Squares (WCSS) yang digunakan untuk mengukur tingkat kedekatan data terhadap centroid cluster.

Rumus WCSS:

$$WCSS = \sum(x_i - c_i)^2$$

Keterangan:

x_i = data ke-i
c_i = centroid cluster
WCSS = total variasi data dalam cluster

Semakin kecil nilai WCSS maka semakin baik kualitas cluster yang terbentuk.

5.11 Kesimpulan Implementasi dan Rekomendasi

1. Kesimpulan

Berdasarkan hasil pembahasan yang telah diuraikan, dapat disimpulkan bahwa metode K-Means Clustering dapat diterapkan sebagai salah satu pendekatan dalam proses pengelompokan calon penerima Bantuan Langsung Tunai (BLT) di Desa N2 Aek Nabara. Pengelompokan dilakukan berdasarkan beberapa atribut sosial ekonomi, yaitu pendapatan, jumlah tanggungan, kondisi rumah, dan pekerjaan. Atribut-atribut tersebut menjadi dasar penting dalam menilai tingkat kelayakan masyarakat sebagai penerima bantuan.

Hasil pengolahan data menggunakan RapidMiner terhadap 120 Kepala Keluarga (KK) menunjukkan bahwa data calon penerima BLT berhasil dikelompokkan ke dalam tiga cluster. Cluster 0 terdiri atas 42 KK, Cluster 1 terdiri atas 44 KK, dan Cluster 2 terdiri atas 34 KK. Berdasarkan analisis nilai centroid, Cluster 0 dikategorikan sebagai kelompok sangat layak menerima BLT, Cluster 1 sebagai

kelompok layak menerima BLT, dan Cluster 2 sebagai kelompok tidak layak menerima BLT.

Penerapan metode K-Means Clustering dalam penelitian ini menunjukkan bahwa proses penentuan penerima BLT dapat dilakukan secara lebih sistematis, objektif, dan berbasis data. Metode ini mampu membantu mengurangi subjektivitas dalam proses seleksi manual serta memberikan gambaran yang lebih jelas mengenai kondisi sosial ekonomi masyarakat. Dengan demikian, K-Means Clustering dapat menjadi alternatif pendukung pengambilan keputusan dalam menentukan penerima bantuan agar lebih tepat sasaran, transparan, dan efektif.

2. Rekomendasi

Berdasarkan hasil penelitian yang telah diperoleh, Pemerintah Desa N2 Aek Nabara disarankan untuk memanfaatkan metode K-Means Clustering sebagai salah satu alat pendukung dalam proses pengambilan keputusan penentuan penerima Bantuan Langsung Tunai (BLT). Penggunaan metode ini dapat membantu proses seleksi penerima bantuan menjadi lebih sistematis, objektif, efektif, dan tepat sasaran berdasarkan kondisi sosial ekonomi masyarakat.

Selain itu, penggunaan data yang akurat dan terbaru sangat diperlukan agar hasil pengelompokan calon penerima BLT dapat mencerminkan kondisi masyarakat yang sebenarnya. Oleh karena itu, pemerintah desa perlu melakukan pembaruan data secara berkala, terutama terkait pendapatan, jumlah tanggungan, kondisi rumah, pekerjaan, dan aspek sosial ekonomi lainnya.

Untuk penelitian selanjutnya, disarankan agar jumlah data yang digunakan lebih besar serta menambahkan variabel lain, seperti tingkat pendidikan, kondisi kesehatan, kepemilikan aset, dan status pekerjaan. Penambahan variabel tersebut diharapkan dapat meningkatkan kualitas hasil pengelompokan dan memberikan gambaran yang lebih lengkap mengenai tingkat kelayakan masyarakat sebagai penerima bantuan.

Penelitian berikutnya juga dapat mengembangkan analisis dengan membandingkan metode K-Means Clustering dengan metode data mining lainnya. Perbandingan tersebut penting dilakukan untuk mengetahui metode yang paling sesuai dan optimal dalam mendukung proses klasifikasi atau pengelompokan calon penerima bantuan sosial.

DAFTAR PUSTAKA

Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (pp. 1027–1035).

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science, 17*(3), 235–255. https://doi.org/10.1214/ss/1042727940

Bradley, P. S., & Fayyad, U. M. (1998). Refining initial points for K-means clustering. In *Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 91–99). Morgan Kaufmann.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM, 56*(12), 64–73.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (pp. 226–231). AAAI Press.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3), 37–54. https://doi.org/10.1609/aimag.v17i3.1230

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge discovery and data mining: Towards a unifying framework.

In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (pp. 82–88). AAAI Press.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*(2), 129–137.

Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: Promise and potential. *Health Information Science and Systems, 2*, Article 3. https://doi.org/10.1186/2047-2501-2-3

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Bulletin of the Technical Committee on Data Engineering, 23*(4), 3–13.

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery, 10*(3), Article e1355. https://doi.org/10.1002/widm.1355

Tan, P.-N., Steinbach, M., & Kumar, V. (2019). *Introduction to data mining* (2nd ed.). Pearson.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). *Data mining: Practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann.

Xu, R., & Wunsch, D. C., II. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks, 16*(3), 645–678. https://doi.org/10.1109/TNN.2005.845141

GLOSARIUM

Data Mining	Proses menemukan pola, hubungan, dan informasi penting dari kumpulan data yang besar.
KDD (Knowledge Discovery in Databases)	Proses sistematis untuk menemukan pengetahuan dari data melalui beberapa tahapan analisis.
Dataset	Kumpulan data yang digunakan sebagai objek analisis.
Basis Data (Database)	Tempat penyimpanan data yang terstruktur dan dapat diakses untuk berbagai kebutuhan analisis.
Clustering	Teknik pengelompokan data berdasarkan tingkat kemiripan karakteristik.
K-Means	Algoritma clustering yang membagi data ke dalam sejumlah kelompok berdasarkan centroid.
Cluster	Kelompok data yang memiliki karakteristik serupa.
Centroid	Titik pusat suatu cluster yang digunakan sebagai acuan pengelompokan data.

Iterasi	Proses pengulangan langkah-langkah algoritma hingga mencapai kondisi stabil.
Atribut	Karakteristik atau variabel yang dimiliki oleh setiap data.
Variabel	Faktor atau unsur yang diamati dan dianalisis dalam penelitian.
Preprocessing	Tahap persiapan data sebelum dianalisis, seperti pembersihan dan perbaikan data.
Data Cleaning	Proses membersihkan data dari kesalahan, duplikasi, dan nilai yang tidak valid.
Transformasi Data	Proses mengubah format atau struktur data agar sesuai untuk analisis.
Normalisasi Data	Teknik menyamakan rentang nilai data agar tidak terjadi dominasi atribut tertentu.
Seleksi Data	Proses memilih data yang relevan dengan tujuan analisis.
Euclidean Distance	Metode pengukuran jarak yang umum digunakan dalam algoritma K-Means.
Similaritas	Tingkat kemiripan antarobjek dalam suatu dataset.
Outlier	Data yang memiliki karakteristik jauh berbeda dari sebagian besar data lainnya.
Classification	Teknik data mining untuk mengelompokkan data ke dalam kategori yang telah ditentukan.
Prediction	Teknik untuk memperkirakan nilai atau kejadian berdasarkan pola data sebelumnya.
Association Rule	Teknik untuk menemukan hubungan atau keterkaitan antaritem dalam data.

CRISP-DM	Kerangka kerja standar dalam proyek data mining yang terdiri dari enam tahapan utama.
Big Data	Kumpulan data berukuran sangat besar yang memerlukan teknologi khusus untuk pengolahannya.
Machine Learning	Cabang kecerdasan buatan yang memungkinkan sistem belajar dari data secara otomatis.
Analisis Deskriptif	Analisis yang bertujuan menggambarkan kondisi atau karakteristik data.
Analisis Prediktif	Analisis yang digunakan untuk memperkirakan kondisi atau kejadian di masa mendatang.
Bantuan Sosial (Bansos)	Program pemerintah yang bertujuan membantu masyarakat yang membutuhkan.
Penerima Bantuan Sosial	Individu atau keluarga yang memenuhi kriteria untuk memperoleh bantuan sosial.
Sistem Pendukung Keputusan	Sistem yang membantu pengambil keputusan dengan menyediakan informasi hasil analisis data.