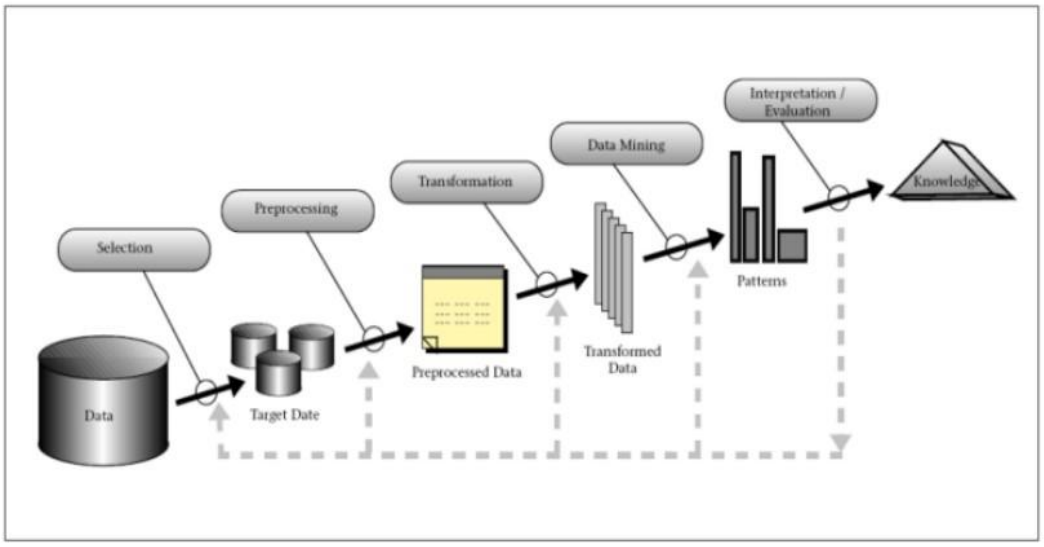
BAB II

TINJAUAN PUSTAKA

1.6 Knowledge Discover in Database (KDD)

Knowledge Discovery in Database (KDD) merupakan proses analisa terstruktur untuk memperoleh informasi yang benar, baru, bermanfaat dan menemukan pola dari data yang besar dan komplek. Data mining (DM) menjadi inti dari proses KDD, yakni dengan menggunakan algoritma tertentu untuk mengeksplorasi data, membangun model dan menemukan pola yang belum di ketahui (Zanuardi & Suprayitno, 2018).



Gambar 2. 1 Tahapan *Knowledge Discovery in Database* (KDD)
Sumber: (Budiman, Muliadi, & Ramadina, 2015)

Proses KDD secara garis besar dapat dijelaskan sebagai berikut (Budiman, Muliadi, & Ramadina, 2015) :

1. Data Selection adalah Pemilihan (seleksi) data dari sekumpulan data yang akan digunakan dalam proses data mining. Perlu dilakukan sebelum tahap penggalian informasi dalam *Knowledge Discovery in Database* (KDD) dimulai. Data hasil seleksi yang akan digunakan untuk proses data mining, disimpan dalam suatu berkas terpisah dari basis data operasional.
2. *Pre-processing / Cleaning* Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi fokus *Knowledge Discovery in Database* (KDD). Proses cleaning mencakup antara lain pembersihan data sepeerti menangani data yang tidak lengkap, menghilangkan gangguan atau outliner, membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak. Tahapan *Pre-Processing* dan *Cleaning Data* perlu dilakukan agar data terhindar dari duplikasi data, memperbaiki kesalahan data, atau dilakukan penambahan data untuk menunjang sistem yang dibuat.
3. *Transformation* adalah proses transformasi pada data yang telah dipilih, Setelah melakukan tahapan Pre-Processing dan Cleaning Datakemudian dilanjutkan dengantahapan Transformation, yaitu tahap untuk mengubah data dari bentuk asalnya menjadi data yang siap untuk ditambang. Karena hal tersebut dapat memudahkan dalam prosespenggalian data untuk menemukansuatupengetahuan baru.
4. *Data mining* adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik-teknik, metode-metode, atau algoritma dalam data mining sangat bervariasi.

Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses *Knowledge Discovery in Database* (KDD) secara keseluruhan.

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

1.7 Cross-Industry Standard Process for Data Mining (CRISP-DM)

CRISP-DM adalah kerangka kerja yang terstruktur dan terdiri dari enam tahapan yang membantu memastikan proses data mining dilakukan secara sistematis dan komprehensif. Tahapan ini meliputi Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment (Mahanani, Utsalina, & Akhriza, 2024). Setiap tahapan saling berkaitan dan dapat dilakukan secara iteratif, sehingga memungkinkan peneliti untuk kembali ke tahap sebelumnya apabila ditemukan ketidaksesuaian atau diperlukan penyesuaian terhadap proses yang sedang berlangsung. Penerapan tahapan CRISP-DM ini bertujuan untuk menghasilkan analisis data yang valid, akurat, serta mendukung pengambilan keputusan yang lebih tepat. Dengan demikian, pendekatan ini dipilih dalam penelitian sebagai acuan metodologis untuk mengarahkan proses pengolahan dan analisis data secara komprehensif.

Gambar 2.2 Framework *Cross-Industry Standart Process for Data Mining (CRISP-DM)*
Sumber: https://outsource-philippines.com/what-is-crisp-dm/

1. Business Understanding, Tahap ini bertujuan untuk memahami secara mendalam tujuan bisnis yang ingin dicapai dari proyek data mining. Aktivitas utama meliputi identifikasi masalah bisnis, penetapan tujuan analisis data, serta pembuatan rencana awal proyek. Dengan memahami konteks bisnis, analis dapat merancang solusi data mining yang relevan dan bernilai guna bagi organisasi. Misalnya, pada kasus lain seperti segmentasi pelanggan toko online dengan tahap proses mencakup:
    a. Masalah bisnis: Toko mengalami kesulitan dalam melakukan promosi yang tepat sasaran, dan banyak pelanggan tidak kembali melakukan pembelian.
    b. Tujuan analitik: Melakukan segmentasi pelanggan berdasarkan perilaku belanja untuk mendukung strategi pemasaran yang lebih efektif.
    c. Rencana proyek:

1) Mengumpulkan data Riwayat pembelian pelanggan.
2) Menggunakan algoritma *clustering* seperti K-Means untuk mengelompokkan pelanggan berdasarkan frekuensi pembelian, jumlah transaksi, dan nilai belanja.
3) Menganalisis hasil segmen untuk menyusun strategi promosi khusus pada tiap kelompok pelanggan (misal: pelanggan loyal, pelanggan baru, pelanggan yang hampir hilang).

2. Data Understanding, Setelah tujuan bisnis dipahami, tahap ini berfokus pada pengumpulan dan eksplorasi data awal yang akan digunakan. Analisis awal dilakukan untuk memahami struktur, kualitas, dan karakteristik data. Proses ini membantu mengidentifikasi potensi masalah dalam data seperti nilai hilang, outlier, atau data tidak konsisten, serta memberikan wawasan awal yang berguna untuk proses selanjutnya. Misalnya, pada kasus Prediksi kepuasan pelanggan restoran tahap proses mencakup:

   1) Pengumpulan data awal

      Data diambil dari survei pelanggan dan sistem POS (Point of Sales), mencakup: lama pelayanan, kualitas makanan, kebersihan restoran, metode pembayaran, dan rating kepuasan (1–5).

   2) Pemeriksaan struktur data

      Terdapat 1.000 baris data dan 8 atribut, termasuk atribut target "kepuasan_pelanggan". Ditemukan beberapa atribut numerik (lama pelayanan, rating), dan kategorikal (metode pembayaran).

   3) Eksplorasi awal data

Ditemukan bahwa sebagian besar pelanggan memberikan rating 4 dan 5, namun rating rendah (1–2) lebih sering terjadi jika lama pelayanan >20 menit.

4) Identifikasi masalah kualitas data

Ada 5% data yang tidak memiliki rating, dan beberapa data metode pembayaran berisi entri tidak valid seperti “xxxx” atau kosong.

3. Data Preparation, Tahap ini melibatkan semua proses untuk mempersiapkan data agar layak digunakan dalam pemodelan. Kegiatan umum mencakup seleksi data yang relevan, pembersihan data, transformasi variabel, dan penggabungan data dari berbagai sumber. Tujuannya adalah menghasilkan dataset akhir yang bersih, konsisten, dan sesuai format yang dibutuhkan oleh algoritma pemodelan.
4. Modeling, Dalam tahap modeling, berbagai teknik data mining diterapkan pada data yang telah dipersiapkan untuk membangun model prediktif atau deskriptif. Pemilihan algoritma dilakukan berdasarkan jenis data dan tujuan analisis, seperti klasifikasi, regresi, atau *clustering*. Parameter dari model juga diuji dan disesuaikan agar menghasilkan performa terbaik.
5. Evaluation, Tahap evaluasi dilakukan untuk mengukur kualitas dan keakuratan model yang telah dibangun. Selain menilai performa model dari sisi teknis (seperti *cluster*, akurasi, presisi, recall), evaluasi juga mempertimbangkan apakah model tersebut benar-benar menjawab kebutuhan dan tujuan bisnis. Jika model belum memenuhi harapan, perbaikan dapat dilakukan sebelum diterapkan.

6. Deployment, Tahap terakhir adalah penerapan hasil analisis ke dalam proses bisnis nyata. Model yang telah dievaluasi dapat diintegrasikan ke dalam sistem operasional, digunakan untuk membuat laporan, atau dijadikan dasar pengambilan keputusan. Deployment juga mencakup dokumentasi, pelatihan pengguna, serta pemantauan performa model secara berkala untuk memastikan hasilnya tetap relevan dan akurat.

1.8 Data Mining

Data mining merupakan proses teknik matematika, teknik statistic, machine learning dan juga kecerdasan buatan yang dapat menguraikan pengetahuan baru yang ada didalam basis data serta mengidentifikasikan informasi bermanfaat yang memiliki keterkaitan dengan basis data yang besar. Data Mining adalah teknik yang memanfaatkan data dalam jumlah yang besar untuk memperoleh informasi berharga yang sebelumnya tidak diketahui dan dapat dimanfaatkan untuk pengambilan keputusan penting (Aldisa, 2023). Salah satu tugas utama dan metode yang sering diterapkan dalam data mining adalah *clustering*, yaitu teknik untuk mengelompokkan data yang serupa ke dalam *cluster* yang berbeda. Kegiatan ini menghasilkan keluaran yang dapat digunakan untuk membuat keputusan di masa mendatang (Helia, Mulyawan, Rohmat, & Fathurrohman, 2024).

Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu (Samsi, 2021) :

1) *Description* (Deskripsi) Terkadang peneliti dan analis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data. Sebagai contoh, petugas

pengumpulan suara mungkin tidak dapat menemukan keterangan atau fakta bahwa siapa yang tidak cukup profesional akan sedikit didukung dalam pemilihan presiden. Deskripsi dari pola dan kecenderungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecenderungan.

2) *Estimation* (Estimasi) adalah hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah numerik daripada ke arah kategori. Model dibangun menggunakan record lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi. Sebagai contoh yaitu estimasi nilai indeks prestasi kumulatif mahasiswa program pasca sarjana dengan melihat nilaiindeks prestasi mahasiswa tersebut pada saat mengikuti program sarjana.
3) *Prediction* (Prediksi) hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada dimasa mendatang. Contoh prediksi dalam bisnis dan penelitian adalah:
    a. Prediksi harga beras dalam tiga bulan yang akan datang.
    b. Prediksi tingkat penganguran lima tahun akan datang.
    c. Prediskси persentase kanaikan kecelakaan lalu lintas tahun depan jika batas bawah kecepatan dinaikan.

    Beberapa metode dan teknik yang digunakan dalam klasifikasi dan estimasi dapat pula digunakan (untuk keadaan yang tepat) untuk prediksi.
4) *Classification* (Klasifikasi) Dalam klasifikasi, terdapat target variabel kategori. Sebagai contoh, penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, dan

pendapatan rendah. Contoh lain klasifikasi dalam bisnis dan penelitian adalah:

a. Menentukan apakah suatu transaksi kartu kredit merupakan transaksi yang curang atau bukan.
b. Memperkirakan apakah suatu pengajuan hipotek oleh nasabah merupakan suatu kredit yang baik atau buruk.
c. Mendiagnosis penyakit seorang pasien untuk mendapatkan termasuk penyakit apa.

5) *Clustering* (Pengklusteran) merupakan pengelompokan record, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. Kluster adalah kumpulan record yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan dengan record-record dalam kluster lain. Pengklusteran berbeda dengan klasifikasi yaitu tidak adanya variabel target dalam pengklusteran. Pengklusteran tidak mencoba untuk melakukan klasifikasi, mengestimasi, atau memprediksi nilai dari variabel target. Akan tetapi, algoritma pengklusteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi *cluster-cluster* yang memiliki kemiripan (homogen), yang mana kemiripan record dalam satu kelompok akan bernilai maksimal, sedangkan kemiripan dengan record dalam kelompok lain akan bernilai minimal.
6) *Association* (Asosiasi) Tugas asosiasi dalam data mining adalah menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja. Contoh asosiasi dalam bisnis dan penelitian adalah:

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

1.9 Metode Clustering

Clustering merupakan metode yang digunakan untuk mengumpulkan data menjadi beberapa kelompok berdasarkan kriteria yang sama. Suatu data akan memiliki karakteristik yang sama dengan data lain yang berada dalam satu kelompok, sedangkan karakteristiknya akan berbeda dengan data lain yang berada di kelompok yang berbeda (Kumarahadi, Pratiwi, & Subianti, 2023). Sebuah claster terdiri dari Kumpulan objek yang mirip satu sama lain dan berbeda dari objek dalam *cluster* lainnya. Algoritma clastering terdiri dari dua jenis, yaitu hirarkis dan partitional. Algoritma hirarkis menemukan claster secara berurutan dengan claster yang ditetapkan sebelumnya, sedangkan algoritma partitional menentukan semua kelompok pada satu waktu. Seperti contoh dalam dunia sosial media, di mana platform mungkin menggunakan *clustering* untuk mengelompokkan pengguna dengan minat atau perilaku yang serupa. Ini dapat membantu dalam menyajikan konten yang lebih relevan atau rekomendasi teman yang lebih akurat kepada pengguna. Pada pembahasan ini, objek atau data dikelompokkan berdasarkan tingkat kemiripan antar mereka sehingga objek dalam satu kelompok akan lebih mirip satu sama lain daripada dengan objek dalam kelompok lain. Melalui pendekatan ini, kita dapat membantu mengidentifikasi serta menemukan kelompok-kelompok alami yang terdapat dalam sebuah dataset. *Clustering* merupakan salah

satu teknik dalam data mining yang termasuk dalam kategori Unsupervised, karena tidak ada atribut tunggal yang digunakan sebagai panduan dalam proses pembelajaran. Sebagai gantinya, semua atribut input diperlakukan dengan cara yang sama. Sebagian besar Algoritma *clustering* membangun model dengan mengikuti serangkaian iterasi dan berhenti ketika model tersebut telah mencapai pusat atau kumpulan titik-titik (batasan dari segmen ini telah mencapai kestabilan) (Faran & Aldisa, 2023).

1.10 Metode Hierarchical Clustering

Hierarchical Clustering merupakan salah satu metode *clustering* yang didasarkan pada struktur seperti dendogram, yaitu membagi atau menggabungkan data seperti cabang-cabang pohon dengan mengelompokkan dua atau lebih data yang memiliki kesamaan terdekat. *Hierarchical Clustering* dapat dilakukan dengan menggunakan dua metode, yaitu *agglomerative nesting* dan *divisive analysis* (Kumarahadi, Pratiwi, & Subianti, 2023).

Hierarchical Clustering merupakan metode *clustering* yang dapat melakukan pengelompokan objek pada data ke dalam sebuah hierarki. Terdapat dua teknik pengelompokan objek pada hierarchical *clustering* yaitu secara agglomerative *(bottom-up)* dan *divisive (top-down). Hierarchical agglomerative Clustering* menggabungkan setiap objek hingga menjadi satu kelompok, sedangkan *Hierarchical Divisive Clustering* memisahkan semua objek pada sebuah kelompok besar menjadi kelompok yang hanya memiliki satu objek. Ketika *agglomerative* telah menggabungkan dua kelompok maka mereka tidak dapat dipisahkan kembali, dan ketika divisive telah memisahkan dua objek maka mereka tidak dapat digabungkan kembali (Abdurrahman, 2019).

Pengelompokan data menggunakan metode *Hierarchical Agglomerative Clustering* diawali dengan merepresentasikan setiap objek pada data sebagai satu kelompok, kemudian dilakukan perhitungan jarak (*distance measure*) antar kelompok tersebut. Setelah itu dua kelompok yang memiliki jarak terdekat digabungkan menjadi sebuah kelompok baru. Kemudian jarak antara kelompok yang baru dengan kelompok lain dihitung menggunakan salah satu metode perhitungan jarak antar kelompok (*single linkage, complete linkage, Average Linkage*, dll). Pengertian lain dari *Agglomerative* (metode penggabungan) adalah strategi pengelompokan hirarki yang dimulai dengan setiap objek dalam satu *cluster* yang terpisah kemudian membentuk *cluster* yang semakin membesar. Jadi, banyaknya *cluster* awal adalah sama dengan banyaknya objek. Sedangkan *divisive* (metode pembagian) adalah strategi pengelompokan hirarki yang dimulai dari semua objek dikelompokkan menjadi *cluster* tunggal kemudian dipisah sampai setiap objek berada dalam *cluster* yang terpisah (Samsi, 2021).

1.11 Algoritma Agglomerative Hierarchical Clustering

Strategi pengelompokan hierarchical *hierarchical* dibagi menjadi dua, yakni *divisive (top to down)* dan *agglomerative (down to top)* (Abdurrahman, 2019). Namun untuk penelitian ini menggunakan pendekatan Agglomerative Hierarchical *Clustering*.

Agglomerative Hierarchical *Clustering* (AHC) adalah metode *clustering* bersifat bottom-up yaitu menggabungkan *cluster-cluster* menjadi satu *cluster* tunggal. Agglomerative Hierarchical *Clustering* (AHC) dengan menggunakan buttom –up , dimulai dari masing -masing data sebagai sebuah *cluster*, kemudian secara rekrusif mencari kelompok terdekat sebagai pasangan yang kemudian akan

digabungkan menjadi kelompok yang lebih besar. Proses tersebut diulang terus sehingga tampak bergerak keatas membentuk hirarki (Suhirman & Wintolo, 2019). Di dalam algoritma *agglomerative* terdapat empat macam metode, yaitu: *single linkage*, *centroid linkage*, *complete linkage*, dan *Average Linkage*. Salah satu algoritma *agglomerative yakni Average Linkage*, yang selanjutnya dipilih sebagai algoritma di dalam penelitian ini, disebut sebagai algoritmaA*gglomerative Hierarchical Clustering (*AHC*) Average Linkage.* Algoritma ini dipilih karena merupakan algoritma terbaik di antara algoritma *hierarchical* yang lain walaupun dengan waktu komputasi tertinggi di antara algoritma *hierarchical* yang lain (Abdurrahman, 2019).

Terdapat tiga Teknik kedekatan dalam *Hierarchical Clustering*, yaitu; *single linkage* (jarak terdekat) atau tautan tunggal, *Average Linkage* (jarak rata-rata) atau tautan rata-rata, dan *complete linkage* (jarak terjauh) atau tautan lengkap.

1.11.1 Single Linkage

Single Linkage (MIN) menentukan kedekatan diantara dua kelompok terdekat (terkecil) antar dua data dari *cluster* yang berbeda.

Formulasi untuk *Single linkage* adalah:

$$\text{l}\boldsymbol{duv} = \boldsymbol{min}\{\boldsymbol{duv}\} \qquad (1)$$

Keterangan:

$\{duv\}$ adalah jarak antara data U dan V dari masing-masing *cluster* U dan V.

1.11.2 Average Linkage

Average Linkage (AVERAGE) menentukan kedekatan diantara dua kelompok dari jarak rata-rata antar dua data dari *cluster* yang berbeda.

Formulasi untuk *Average Linkage* adalah:

$$d_{uv} = \frac{1}{|u*|v|} \sum u€D \sum v€D^{d}{}_{uv} \qquad (2)$$

Keterangan:

|U| dan |V| adalah jumlah data yang ada dalam *cluster* U dan V.

1.11.3 Complete Linkage

Complete linkage (MAX) menentukan kedekatan diantara dua kelompok dari jarak terjauh (terbesar) antara dua data dari *cluster* yang berbeda.

Formulasi untuk *complete linkage* adalah:

$$\boldsymbol{duv} = \mathbf{max}\,\{\boldsymbol{duv}\} \qquad (3)$$

Keterangan:

$\{duv\}$ adalah jarak antara data U dan V dari masing-masing *cluster* U dan V.

1.12 Evaluasi Cluster Davies-Bouldin Index (DBI)

DBI diperkenalkan oleh David L. Davies dan Donald W. Bouldin pada 1979, untuk mengukur kualitas *clustering* dengan menghitung rasio antara jarak *cluster*. Nilai yang lebih kecil menunjukkan *clustering* yang lebih baik (Ginting, Harianja, & Sipayung, 2024).

Agar DBI menghasilkan nilai terbaik maka sebisa mungkin S_i harus bernilai sekecil mungkin dan M_{ij} harus bernilai sebesar mungkin. Dengan menggunakan DBI suatu *cluster* akan dianggap memiliki skema *clustering* yang optimal apabila memiliki

nilai DBI yang kecil. Tahapan dalam melakukan perhitungan *Davies-Bouldin Index* (DBI) adalah sebagai berikut (Samsi, 2021) :

1. *Sum of Square Within Cluster* (SSW)

   Untuk mengetahui kohesi dalam sebuah *cluster* ke-*i* adalah dengan menghitung nilai dari *Sum of Square Within Cluster* (SSW). Kohesi maksudnya yaitu jumlah dari kedekatan data terhadap titik pusat *Cluster* dari sebuah *cluster* yang terbentuk.

   $$SSW_i = \frac{1}{m_i}\sum_{j=i}^{m_i} d(x_{j,}\, c_i) \qquad (5)$$

2. *Sum of Square Between Cluster* (SSB)

   Perhitungan *Sum of Square Between Cluster* (SSB) bertujuan untuk mengetahui jarak antar *cluster*.

   $$SSB_{ij} = d(c_{i,}c_j) \qquad (6)$$

3. *Ratio* (Rasio)

   Bertujuan untuk mengetahui nilai perbandingan antara *cluster* ke-i dan *cluster* ke-j.

   $$R_{ij} = \frac{SSW_I + SSW_j}{SSB_{ij}} \qquad (7)$$

4. *Davies Bouldin Index*

   Nilai rasio yang diperoleh dari persamaan Rasio digunakan untuk mencari nilai *davies bouldin index* (DBI).

   $$DBI = \frac{1}{K}\sum_{i=1}^{k} max_{i \neq j}(R_{ij}) \qquad (8)$$

1.13 Dendogram

Dendrogram adalah representasi visual berbentuk pohon yang digunakan dalam metode klasterisasi hierarkis, khususnya Agglomerative Hierarchical *Clustering* (AHC). Diagram ini menggambarkan proses penggabungan antar objek atau *cluster* secara bertahap berdasarkan tingkat kemiripan atau jarak antar data. Setiap percabangan dalam dendrogram menunjukkan penggabungan dua *cluster*, dimulai dari objek individual hingga membentuk satu *cluster* besar yang menaungi seluruh data. Salah satu keunggulan dendrogram adalah kemampuannya dalam membantu menentukan jumlah *cluster* optimal secara visual. Dengan memotong dendrogram pada titik tertentu (threshold), peneliti dapat mengidentifikasi jumlah *cluster* yang paling representatif sesuai dengan struktur data. Sebagai contoh, dalam penelitian yang menggunakan AHC untuk menganalisis topik skripsi mahasiswa, pemotongan dendrogram pada level 0.25 menghasilkan *cluster* yang lebih spesifik, memudahkan analisis topik dominan di antara judul-judul skripsi yang dikaji (Februariyanti & Santoso, 2017).

Selain itu, dendrogram juga memudahkan dalam menginterpretasikan hubungan hierarkis antar data. Dalam studi lain yang membandingkan metode *K-Means* dan *Hierarchical Clustering* untuk pengelompokan data penduduk miskin di Kabupaten Cianjur, visualisasi dendrogram membantu mengungkap pola hubungan antar data yang kompleks, yang mungkin tidak terdeteksi dengan metode *clustering* lainnya (Putra & Fadhillah, 2025).

Dengan demikian, penggunaan dendrogram dalam visualisasi hasil *clustering* tidak hanya memberikan gambaran struktur data yang lebih mendalam,

tetapi juga mendukung pengambilan keputusan dalam menentukan jumlah *cluster* yang optimal dan memahami hubungan antar *cluster* dalam dataset.

1.14 Penerapan Algoritma AHC dalam *cluster* data perceraian

Penelitian yang dilakukan (Dani, Wahyuningsih, & Rizki, 2019) melakukan pengukuran pengelompokkan Kabupaten/Kota Provinsi Kaltim berdasarkan jumlah penduduk dengan data runtun waktu. Pada penelitian ini melakukan perbandingan antara *single linkage, complete linkage, dan Average Linkage*. Berdasarkan hasil analisis, diperoleh jarak pengukuran kemiripan terbaik dalam mengelompokkan Kabupaten/Kota di Provinsi Kalimantan Timur adalah jarak *autocorrelation based distance* (ACF) dengan nilai koefisien korelasi *cophenetic* sebesar 0,99. Algoritma pengelompokkan yang optimal adalah algoritma *Average Linkage*, dikarenakan memiliki nilai koefisien korelasi *cophenetic* yang terbesar yaitu 0,99 diantara algoritma pengelompokkan lainnya, dengan jumlah *cluster* yang representatif berdasarkan koefisien *silhouette* adalah 2 *cluster*.

Penelitian selanjutnya dilakukan (Fadliana & Rozi, 2015) menunjukkan bahwa berdasarkan hasil uji validitas *cluster*, diketahui bahwa metode *Average Linkage* memberikan solusi *cluster* yang lebih baik bila dibandingkan dengan metode *Agglomerative Hierarchical Clustering* lainnya (*Single Linkage*, *Complete Linkage*, dan *Ward*). Solusi *cluster* pada metode *Average Linkage* menghasilkan 4 *cluster* dengan karakteristik yang berbeda. *cluster* 1 memiliki karakteristik tingkat kualifikasi klinik KB dan tingkat kompetensi tenaga pelayanan KB “sangat rendah”. *Cluster* 2 memiliki karakteristik tingkat kualifikasi klinik KB “cukup baik”, dan Tingkat kompetensi tenaga pelayanan KB “rendah”. *Cluster* 3 memiliki

karakteristik tingkat kualifikasi klinik KB “rendah” dan tingkat kompetensi tenaga pelayanan KB “sedang”. *Cluster* 4 terdiri dari empat kabupaten dengan karakteristik tingkat kualifikasi klinik KB “sedang” dan tingkat kompetensi tenaga pelayanan KB“cukup baik”.

Pada penelitian (Widyawati, Saptomo, & Utami, 2020) dijelaskan bahwa metode AHC merupakan metode yang lebih baik daripada K-Means terutama metode AHC *Complete Linkage* dan *Average Linkage.* Hasil pengelompokan diperoleh *cluster* 1 terdapat 20 anggota, *cluster* 2 terdapat 43 anggota, *cluster* 3 terdapat 75 anggota, *cluster* 4 terdapat 158 anggota, *cluster* 5 terdapat 9 anggota, *cluster* 6 terdapat 2 anggota dan *cluster* 7 terdapat 1 anggota.

Pada penelitian (Pratikto & Damastuti, 2021) uji *cluster* optimal dengan *elbow method*, provinsi Jawa Timur terbagi menjadi 3 kelompok *cluster* daerah terdampak potensi banjir yaitu karakteristik rendah, sedang, tinggi. Hasil uji performa *cluster* menggunakan *cophenetic correlation coefficient* menunjukkan bahwa metode *Average Linkage* memberikan solusi *cluster* yang lebih baik dibandingkan dengan metode AHC lainnya yakni sebesar 0,92.

Penelitian (Paramadina, Sudarmin, & Aidid, 2019) membandingkan metode *Average Linkage* dan *Complete Linkage* dalam analisis *cluster*. Hasilnya menunjukkan bahwa metode *Average Linkage* lebih efektif dalam mengelompokkan objek atau variabel ke dalam *cluster* yang memiliki sifat berbeda antara satu dengan yang lainnya. Pada penelitian ini dibentuk beberapa pengelompokan awal yang selanjutnya akan ditentukan *cluster* yang baik digunakan dengan menggunakan Validasi *Cluster Index Dunn*. Diperoleh bahwa pengelompokan dengan menggunakan Metode *Average Linkage* menghasilkan

Index Dunn yang terbaik dengan nilai 0.55 dengan jumlah *cluster* yang terbentuk sebanyak 8 *cluster* dibandingkan dengan Metode *Ward* yang menghasilkan nilai *Index Dunn* sebesar 0.43.

Penelitian lainnya dengan menggunakan metode DBI yang menguji *cluster* yaitu (Suraya & Wijayanto, 2022) Indeks davies bouldin adalah suatu ukuran yang menunjukkan kesamaan *cluster* yang kerapatan datanya merupakan fungsi penurunan jarak dari karakteristik vektor *cluster*. Semakin kecil nilai indeks davies bouldin pada suatu *cluster*, semakin baik hasil *clustering*-nya. Dengan meminimalkan indeks ini terdapat perbedaan yang jelas antara satu *Cluster* dengan *cluster* lainnya sehingga dapat mencapai partisi terbaik. Ukuran evaluasi yang digunakan yaitu *connectivity coefficient (CC)*, *dunn index (DI)*, *silhouette coefficient (SC), davies bouldin index (DBI), silhouette index (SI)*, *calinski-harabasz index (CHI*), dan *xie and beni index (XBI)* yang dapat dilihat pada Tabel 2.1. Berdasarkan evaluasi yang telah dilakukan, dapat disimpulkan bahwa metode terbaik dalam mengelompokkan tingkat penanganan stunting pada provinsi-provinsi di Indonesia pada tahun 2018 yaitu menggunakan *Average Linkage* dengan empat *luster*.

Tabel 2.1 Evaluasi metode terbaik

Sumber: (Suraya & Wijayanto, 2022)

Metode	Jumlah *Cluster* (k)	CC	DI	SC	DBI	CHI	XBI
Average Linkage	4	15,4056	0,3850	0,2387	0,8840	6,8872	0,9261
K-Means	2	10,1933	0,3567	0,3722	1,1358	13,0889	1,1062
K-Medoids	2	8,0976	0,3328	0,3716	1,0256	11,5720	1,1426
Fuzzy C-Means (FCM)	3	33,5369	0,1485	0,0763	1,3880	12,1100	1,5376

Hal ini dapat dilihat bahwa metode *Average Linkage* memiliki tiga indeks dengan kriteria terbaik dibandingkan metode lainnya. Indeks tersebut terdiri dari *dunn index* dengan nilai tertinggi, sedangkan nilai *davies bouldin index* dan *xie and beni index* merupakan nilai terkecil (terbaik) di antara metode lainnya. Kemudian, metode terbaik berdasarkan *silhouette coefficient*, *calinski-harabasz index connectivity* yaitu metode *k-means* dengan dua *cluster* dan *untuk connectivity coefficient* metode terbaiknya yaitu *k-medoids* dengan dua *cluster*.

(Ermila, Fadhilah, Erdiani, & Saputri, 2025) dilakukan evaluasi dengan memanfaatkan *Davies-Bouldin Index* (DBI) untuk menentukan jumlah *cluster* yang optimal dari data yang telah di c*luster*. DBI digunakan untuk mengukur kualitas partisi *cluster* berdasarkan jarak antar *cluster* dan dalam c*luster*. Nilai DBI yang lebih rendah menandakan partisi yang lebih optimal, dimana *cluster*-c*luster* lebih padat dan terpisah dengan jelas di antara satu sama lain. Nilai DBI yang lebih rendah menunjukkan *clustering* yang lebih baik. Dalam analisis ini, diperoleh bahwa jumlah *cluster* yang optimal adalah 4, dengan nilai DBI terendah sebesar 1.171.

Tabel 2.2 Evaluasi metode terbaik

Sumber: (Suraya & Wijayanto, 2022)

Jumlah *Cluster*	2	3	4	5	6	7	8	9	10
Hasil DBI	1.198	1.231	1.171	1.187	1.178	1.253	1.403	1.296	1.344

1.15 Perceraian

Kata “cerai” menurut kamus besar Bahasa Indonesia berarti: pisah, putus hubungan sebagai suami istri, talak. Kemudian, kata “perceraian” mengandung arti: perpisahan, perihal bercerai (antara suami istri), perpecahan. Adapun kata “bercerai” berarti: tidak bercampur (berhubungan, bersatu) lagi, berhenti berlaki-bini (suami istri). Jadi secara yuridis istilah perceraian berarti putusnya perkawinan, yang mengakibatkan putusnya hubungan sebagai suami istri atau berhenti berlaki-bini (suam istri) sebagaimana diartikan dalam kamus besar Bahasa Indonesia (Sari, Saputra, & Gemasih, 2022).

Perceraian merupakan salah satu bentuk putusnya ikatan hukum antara suami dan istri dalam suatu perkawinan. Menurut Pasal 38 Undung-Undang Nomor 1 Tahun 1974 tentang Perkawinan, perkawinan dapat putus karena tiga hal, yaitu kematian, perceraian, dan putusan pengadilan, Sementara itu, Pasal 39 ayat (1) menegaskan bahwa: “Perceraian hanya dapat dilakukan di depan sidang pengadilan setelah pengadilan yang bersangkutan berusaha dan tidak berhasil mendamaikan kedua belah pihak” (peraturan.bpk.go.id, 2017). Dengan demikian, perceraian dalam konteks hukum Indonesia tidak dapat dilakukan secara sepihak, melainkan harus melalui proses hukum yang sah dan ditetapkan oleh pengadilan.