BAB II

TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Dalam penelitian ini, penulis menggunakan 6 literatur/referensi dalam mendukung pernyataan tersebut. Oleh karena itu, peneliti memaparkan hasil beberapa penelelitian terdahulu yang berkaitan dengan penelitian yang dilakukan oleh peneliti. Hal ini dapat dilihat dari **tabel 2.1** sebagai berikut:

Tabel 2.1 Penelitian Terdahulu

Penulis dan Tahun	Judul	Perbedaan
(Dewi et al., 2019)	Analisa Metode K-*Means* pada Pengelompokan Kriminalitas Menurut Wilayah	Perbedaan antara penelitian tersebut dengan penulis adalah penggunaan data dan objek penelitian. Literatur ini menggunakan objek kriminalitas, sedangkan penulis menggunakan objek perguruan tinggi.
(Amri et al., 2019)	Analisis Metode K-*Means* Pada Pengelompokan Perguruan Tinggi Menurut Provinsi Berdasarkan Fasilitas yang Dimiliki Desa	Perbedaan penilitian tersebut dengan penelitian penulis yaitu data yang digunakan melalui situs Badan Pusat Statistik Indonesia dalam skala provinsi. Jika dalam penelitian penulis data yang digunakan melalui situs web Webometrics dan dalam skala seluruh wilayah Indonesia.
(Batubara et al., 2019)	Analisis Metode K-MEANS Pada Pengelompokan	Perbedaan penelitian tersebut dengan penelitian penulis meggunakan data Perangkingan

Tabel 2.1 Penelitian Terdahulu (Lanjutan)

Penulis dan Tahun	Judul	Perbedaan
	Keberadaan Area Resapan Air Provinsi	Webometrics, sedangkan pada penelitian tersebut menggunakan data persentase rumah tangga.
(Nabila et al., 2021)	Analisis Data Mining Untuk *Clustering* Kasus Covid-19 di Provinsi Lampung Dengan Algoritma K-*Means*	Perbedaan penelitian tersebut dengan penelitian penulis yaitu penggunaan atribut dalam penelitian tersebut menggunakan 6 atribut, sedangkan penulis menggunakan 3 atribut. Serta penelitian tersebut akan membagi data mejadi 4 *cluster* sedangkan peneliti hanya membagi data menjadi 3 *cluster*.
(Virgo et al., 2020)	Klasterisasi Tingkat Kehadiran Dosen Menggunakan Algoritma K-*Means Clustering*	Perbedaan penelitian tersebut dengan penilitian penulis yaitu perhitungan algoritma K-*Means Clustering* menggunakan *Excel* dan aplikasi *Rapid Miner,* sedangkan penilitian penulis menggunakan Bahasa pemograman *Python.*
(Muliono & Sembiring, 2019)	Data Mining *Clustering* Menggunakan Algoritma K-*Means* Untuk Klasterisasi Tingkat Tridarma Pengajaran Dosen	Perbedaan penelitian tersebut dengan penelitian penulis adalah data yang ada cukup banyak, sedangkan data pada penilitian tersebut relative sedikit dan Metode Algoritma K-*Means* akan efektif pada data yang cukup banyak.

Tabel 2.2 Penelitian Terdahulu (Lanjutan)

(Torres-Samuel et al., 2019)	*Clustering of Top 50 Latin American Universities in* SIR, QS, ARWU*, and* Webometrics *Rankings*	Perbedaan penelitian tersebut dengan penelitian penulis adalah hanya mengklasterisasi top 50 Universitas Amerika Latin sedangkan penelitian penulis mengklasteriasai perguruan tinggi se-Indonesia.

2.1.1 Literatur 1

Oleh (Dewi et al., 2019)dari jurusan Sistem Informasi, STIKOM Tunas Bangsa Pematangsiantar dengan judul Analisa Metode K-*Means* pada Pengelompokan Kriminalitas Menurut Wilayah. Dimana penelitian yang dilakukan penulis bagaimana menerapkan metode K-*Means* pada pengelompokam kriminalitas. Sumber data penelitian ini adalah kumpulan dari dokumen Badan Pusat Statistik Nasional yang menjelaskan tentang tindak pidana. Variable yang digunakan yaitu *cluster* tingkat tinggi dan *cluster* tingkat rendah. Hasil dari penelitian ini adalah analisa metode K-*Means* Pengelompokkan Kriminalitas Menurut Provinsi didapatkan hasil bahwa *cluster* tinggi (C1) = 8 provinsi, diantaranya Sumatera Utara, Sumater Selatan, Metro Jaya, Jawa Barat, Jawa Timur, Sulawesi Selatan, Sulawesi Tengah, dan Papua. Dan *cluster* rendah (C2) = 23 Provinsi diantaranya Aceh, Sumatera Barat, Riau, Jambi, Bengkulu, dan 18 provinsi lainnya.

2.1.2 Literatur 2

Dalam penilitian (Amri et al., 2019) dari jurusan Sistem Informasi, STIKOM Tunas Bangsa, Pematangsiantar, Indonesia. Membahas tentang Analisis Metode K-Means Pada Pengelompokan Perguruan Tinggi Menurut Provinsi Berdasarkan Fasilitas Yang Dimiliki Desa. Sumber data yang diperoleh dari data yang

dikumpulkan berdasarkan dokumen-dokumen tahun 2003 sampai tahun 2018 melaui situs Badan Pusat Statistik Indonesia. Data diolah menjadi 2 *cluster* yaitu *cluster* tingkat fasilitas tertinggi (C1) dan *cluster* tingkat fasilitas terendah (C2). Sehingga diperoleh dari 34 provinsi 3 provinsi dikolompokan dalam *cluster* tingkat fasilitas tinggi (C1) dan 31 provinsi dikelompokkan dalam *cluster* tingkat fasilitas rendah (C2). Hal ini dapat menjadi masukan pada pemerintah untuk provinsi yang memiliki perguruan tinggi yang masi memiliki fasilitas yg kurang memadai di setiap desa dan menjadi perhatian yang lebih dari pemerintah berdasarkan *cluster* yang dilakukan.

2.1.3 Literatur 3

Penelitian yang dilakukan (Batubara et al., 2019) Program Studi Sistem Informasi, STIKOM Tunas Bangsa dengan judul Analisis Metode K-*Means* Pada Pengelompokan Keberadaan Area Resapan Air Menurut Provinsi. Penelitian tersebut bertujuan untuk mengelompokan daerah resapan air menggunakan algortima data mining dengan metode K-*Means*. Data yang digunakan penelitian ini adalah data persentase rumah tangga berdasarkan provinsi dan keberadaan area resapan air pada tahun 2017 yang terdiri dari 34 provinsi di Indonesia. Berdasarkan data tersebut diperoleh pengelompokan area resapan air berdasarkan provinsi menjadi 2 *cluster* yaitu *cluster* tinggi (C1) dan *cluster* rendah menggunakan K-*Means Clustering*. Dari hasil penelitian diperoleh pengelompokan menjadi 16 provinsi dengan *cluster* tinggi (C1) dan 18 provinsi dengan *cluster* rendah (C2). Hal ini dapat menjadi masukan pada pemerintah untuk provinsi yang memiliki daerah resapan air rendah menjadi perhatian lebih berdasarkan *cluster* yang telah dilakukan.

2.1.4 Literatur 4

Pada penelitian (Nabila et al., 2021) dari jurusan Sistem Informasi, Fakultas Teknik dan Ilmu Komputer, Universitas Teknokrat Indonesia dengan judul Analisis Data Mining Untuk *Clustering* Kasus Covid-19 di Provinsi Lampung dengan algoritma K-Means. Pada penelitian ini, algoritma K-Means digunakan untuk mengelompokkan dan mengagregasi data kasus Covid-19 di Provinsi Lampung, dengan menggunakan atribut Kabupaten/Kota, Suspek, Probable, Konfirmasi Positif, Selesai Isolasi, serta Kematian yang digunakan dalam proses perhitungan dan membagi data ke dalam 4 *cluster* yang diklasiikasikan menjadi Zona Merah, Zona Orange, Zona Kuning dan Zona Hijau. Juga divalidasi menggunakan *Davies-Bouldin Index* (DBI). Terdapat perbedaan antara hasil validasi DBI dengan perhitungan manual dan menggunakan bantuan *tools RapidMiner*. Pada masalah ini perhitungan manual memberikan hasil yang lebih baik dibandingkan dengan menggunakan *tools RapidMiner*, namun hasil dari kedua perhitungan tersebut sama-sama mendekati 0, menandakan bahwa *cluster* yang dievaluasi menghasilkan klaster yang baik.

2.1.5 Literatur 5

Penelitian (Virgo et al., 2020) dari Universitas Putra Indonesia YPTK Padang mengangkat judul Klasterisasi Tingkat Kehadiran Dosen Menggunakan Algoritma K-Means *Clustering* bertujuan untuk mengelompokkan data pertemuan dosen non pegawai negeri sipil menggunakan Knowledge Discovery in Database (KDD). Tahap selanjutnya adalah data mining menggunakan algoritma K-Means *Clustering*. Hasil penelitian ini pengelompokan dosen menjadi 3 kelompok yaitu 72 matakuliah yang diampu dosen non pns pada kelompok jarang melakukan pertemuan (4.7650%), 69 matakuliah yang diampu dosen non pns pada kelompok

sedang dalam melakukan pertemuan (4.5665%), dan 1370 matakuliah yang di-ampu dosen non pns pada kelompok rajin melakukan pertemuan (90.6684%). Sesuai dengan hasil penelitian didapatlah kesimpulan pada tahun akademik 2017/2018 semester gasal dan genap dosen non pns pengampu matakuliah tertentu rajin masuk pada setiap pertemuan dengan tingkat kehadiran 12-16 kali pertemuan per semester.

2.1.6 Literatur 6

Oleh (Muliono & Sembiring, 2019) dari Program Studi Teknik Informat-ika, Universitas Medan Area dengan judul Data Mining *Clustering* Menggunakan Algoritma K-Means Untuk Klasterisasi Tingkat Tridarma Pengajaran Dosen. Hasil klasterisasi mengacu pada pemberian besaran subsidi yang akan di berikan kepada dosen yang membentuk serta mengumpulkan dokumen pengajaran tese-but. Keakuratan prediksi yang dibuat oleh algortima K-means terhadap 15 data menunjukkan perbedaan ketepatan, dengan hanya 53.33% akurasi prediksi yang bernilai benar. Metode K-Means *Clustering* menggunakan pemrograman web perlu dilakukan optimasi lebih lanjut untuk permasalahan logika dan array berupa nilai decimal panjang. Dan uji K-*Means* dengan lebih banyak data untuk mendapatkan hasil prediksi yang lebih baik.

2.1.7 Literatur 7

Oleh (Torres-Samuel et al., 2019) dengan judul *Clustering of Top 50 Latin American Universities in* SIR, QS, ARWU, *and* Webometrics Rankings. Mengembangkan analisis deskriptif berdasarkan klaster universitas Amerika Latin yang menempati peringkat universitas dunia: ARWU, SIR Scimago, QS, dan Webometrics pada edisi 2019 dan menempati posisi Top 50. Delapan puluh lima universitas atau lembaga pendidikan tinggi diidentifikasi yang berada di

posisi lima puluh teratas (50 Teratas) di masing-masing peringkat Amerika Latin. *Cluster* dibangun dengan mempertimbangkan frekuensi kehadiran universitas di 50 Besar dari empat peringkat, posisi mereka, dan negara untuk menerapkan analisis statistik deskriptif. Perlu dicatat bahwa beberapa universitas dapat menempati posisi yang sama dalam suatu pemeringkatan, sehingga dalam kasus SIR lebih dari 50 universitas diidentifikasi dalam sampel masing-masing. Dengan hasil ditemukan di Brasil, Chili, Argentina, Meksiko, dan Kolombia mewakili 89% (*cluster* 1). Di antaranya, 22 universitas (29%) yang muncul secara bersamaan di 50 Besar dari empat peringkat yang dipelajari dan yang terletak di negara-negara tersebut menonjol. Brasil memimpin 50 Besar dengan hampir 45% universitas. Pada *cluster* kedua terdapat 9 universitas Amerika Latin Top 50 yang berlokasi di Peru, Venezuela, Costa Rica, Cuba, Puerto Rico, Uruguay, dan Jamaica. Mengenai peringkat, ada positioning yang relevan (72%) dari universitas *cluster* 1 di SIR Scimago, selain itu, 60% berada di 50 Besar dari peringkat 3 atau 4. Untuk universitas klaster kedua, posisi tertingginya berada di peringkat QS (78%) dan 66% berada di 50 Besar peringkat tunggal,

2.2 Perangkingan perguruan tinggi

Perangkingan perguruan tinggi ialah pemberian peringkat atau urut-urutan sesuai kriteria tertentu pada perguruan tinggi. Pemeringkatan perguruan tinggi dapat dilakukan oleh lembaga pemerintah, swasta, dalam, ataupun luar negeri (Hermawan, 2018). Praktik pemeringkatkan universitas dimulai pada tahun 1925, Profesor Donald Hughes mengklasifikasikan program pascasarjana di AS berdasarkan reputasi rekan kerja. Praktik ini mendapatkan pupularitas dan berlanjut hingga tahun 1980-an seiring dengan pertum-buhan pasar pendidikan tinggi. Hal ini terlihat dari banyaknya siswa yang mencari perguruan tinggi. Pada tahun 1983, pemeringkatan perguruan tinggi dipublikasikan untuk

pertama kalinya di Amerika Serikat. Karena meningkatnya permintaan, banyak lembaga akademis, badan independen dan organisasi media melakukan pemeringkatan, yang selanjutnya mendorong pengembangan dan publikasi peringkat nasional dan internasional. Yang kemudian juga ada Webometrics berskala nasional dan terkemuka yang membuat pemeringkatan sendiri.

2.3 Webometrics

Webometrics merupakan situs pemeringkatan akademik terbesar dari Institusi Pendidikan Tinggi yang didirikan sejak 2004. Lembaga ini menerbitkan peringkat universitas setiap 2 kali setahun, yaitu pada Januari dan Juli. Tujuan pemeringkatan ini adalah meningkatkan daya saing, keberadaan (eksitensi) dan peran perguruan tinggi melalui publikasi dan penelitian ilmiah perguruan tinggi. Dengan itu, ilmu di perguruan tinggi bermanfaat bagi masyarakat dan transparan sebagai sumber ilmiah. Webometrics diterbitkan oleh *Cybermetrics Labs* di bawah naungan *Spanish National Research Council* (CSIC).

Terdapat 4 indikator berbobot pada penilaian untuk meningkatkan rank pada Webometrics yaitu sebagai berikut:

1. *Presence*

   Yaitu jumlah halaman webhost dalam webdomain utama (termasuk semua subdomain dan direktori) yang diindeks oleh mesin pencari Google. (bobot 10%)

2. *Visibility – Web contents Impact*

   Yaitu jumlah jaringan eksternal (subnet) yang terhubung ke web institusi. (bobot 50%)

3. *Openness*

   Yaitu jumlah kutipan dari 210 penulis teratas yang tertangkap. (bobot 10%)

4. *Excellence*

Yaitu jumlah makalah antara 10% teratas yang paling banyak dikutip di masing-masing dari 27 disiplin ilmu dari database lengkap. (bobot30%)

2.4 Data Mining

Data mining adalah proses ekstraksi suatu data (sebelumnya tidak diketahui, bersifat implisit, dan dianggap tidak berguna) menjadi informasi atau pengetahuan atau pola dari data yang jumlahnya besar. Data-data yang dianggap "sampah" karena tidak terpola/tidak terstruktur, diolah (*filter)* sehingga membentuk informasi atau pengetahuan atau pola baru yang berguna. Secara umum terdapat 5 (lima) peranan dalam data mining, yaitu estimasi (numerik), prediksi (numerik), klasifikasi (numerik/kategorial), *Clustering (numerik),* dan asosiasi. Proses pengolahan data dalam data mining membutuhkan algoritma-algoritma untuk melakukan ekstraksi menjadi informasi/pola/pengetahuan. Peranan *Clustering* digunakan salah satunya algoritma K-*Means* (Joko Suntoro, 2019). Istilah data mining memiliki beberapa pandanan, seperti *knowledge discovery* ataupun *pettern recognition*. Istilah *knowledge discovery* atau inovasi pengetahuan tepat dipergunakan karena tujuan utama data mining memang untuk mendapatkan pengetahuan yang masih tersembunyi di dalam bongkahan data. Sedangkan istilah untuk *pettern recognition* atau pengenalan pola tepat untuk digunakan karena guna menemukan pola yang tersembunyi di dalam bongkahan data (Nabila et al., 2021).

Data mining di kenal sebagai sinonim untuk istilah *Knowledge Discovery in Databases* (KDD) yang bertugas untuk mengekstrak pola atau model dari data dengan menggunakan suatu algoritma yang spesifik. Adapun tahapan dari *Knowledge Discovery in Databases* (KDD) sebagai berikut (Nabila et al., 2021) :

1. Pembersihan data (*Data Cleaning*)

   Pembersihan data untuk menghilangkan noise dan data yang tidak konsisten.

2. Integrasi data (*Data Integration*)

Integrasi data ialah penggabungan dari beberapa sumber data.

3. Seleksi data (*Data Selection*)

   Pengambilan data yang relevan untuk analisis yang diambil dari *database.*

4. Transformasi data (*Data Transformation*)

   Data yang diubah atau digabungkan ke dalam format yang sesuai dan diproses ke dalam data mining.

5. Data Mining

   Proses penting yang menerapkan metode untuk mengekstrak pola dari suatu data.

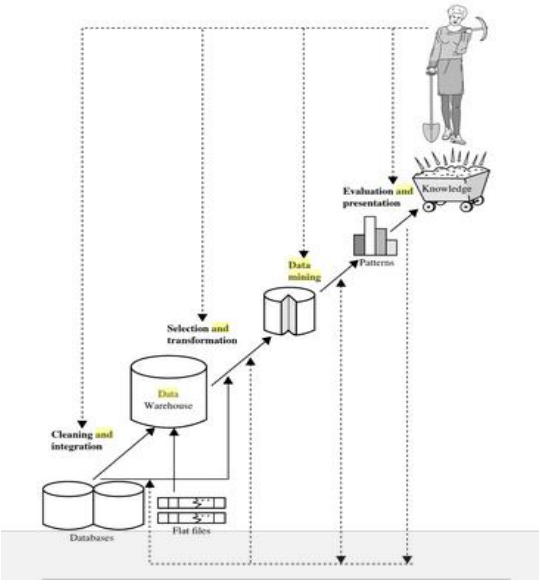
6. Evaluasi pola (*Pattern Evaluation*)

   Mengidentifikasi pola-pola menarik ke dalam *knowledge based* yang ditemukan.

7. Presentasi pengetahuan (*Knowledge Presentation*)

   Teknik visualisasi dan representasi pengetahuan digunakan untuk menyajikan pengetahuan yang diperoleh.

Dari langkah-langkah di atas, dapat diilustrasikan pada **Gambar 2.1** dibawah ini:



Gambar 2. 1 Tahapan *Knowledge Discovery in Databases (KDD)*

(Nabila et al., 2021)

Ada beberapa tugas yang dapat dilakukan data mining dalam proses pemecahan masalah dan pencarian pengetahuan yang baru, meliputi (Amri et al., 2019):

1. Estimasi (*Estimation*)

   Digunakan, memprediksi atau menilai sesuatu hal yang belum pernah ada sebelumnya yang disajikan dalam bentuk hasil kuantitatif (angka).

2. Prediksi (*Predictions*)

   Digunakan, memperkirakan atau meramalkan suatu kejadian yang belum pernah terjadi.

3. Klasifikasi (*Classification*)

   Digunakan untuk menemukan model atau fungsi yang menjelaskan atau membedakan antara konsep atau kelas data untuk tujuan menyimpulkaan kelas suatu objek dengan label yang tidak diketahui.

4. Klastering (*Clustering*)

   Digunakan untuk mengelompokkan atau mengidentifikasi data. Dimana data yang memiliki karakteristik data yang memiliki karakteristik tertentu.

5. Asosiasi (*Association*)

   Digunakan untuk mengatasi masalah bisnis yang khas, yakni dengan menganalisa tabel transaksi penjualan dan mengidentifikasi produk-produk yang seringkali dibeli bersamaan oleh pelanggan.

2.5 *Clustering*

Clustering merupakan metode pengelompokan data. *Clustering* adalah proses membuat kelas-kelas berupa klaster berdasarkan kemiripannya. *Clustering* berbeda dengan klasifikasi dari sisi data, dimana pada *cluster* data tidak memiliki kelas (sering

diistilahkan dengan label atau target), sehingga *Clustering* masuk kedalam kategori pembelajaran tak terpandu (*unsupervised learning)* (Rahmadya Trias Handayanto Herlawati, 2020).

Clustering dalam data mining membantu menemukan pola distribusi pada sebuah data set yang berguna untuk proses analisa data. Kesamaan objek biasanya muncul dari kedekatan nilai atribut yang menggambarkan objek data, sedangkan objek data umumnya direpresentasikan sebagai sebuah titik ruang multidimensi (Nabila et al., 2021).

2.6 Algoritma K-*Means*

Algoritma K-*Means* merupakan salah satu penerapan data mining *Clustering. Clustering* termasuk dalam kelompok teknik *unsupervised learning*. *Unsupervised learning* ialah pembelajaran untuk mencari pola dari semua variable yang menjadi label/class tidak ditentukan (tidak ada) (Joko Suntoro, 2019). Algoritma K-*Means* merupakan teknik *Clustering* berbasis jarak yang membagi data ke dalam beberapa *cluster* dan algoritma ini hanya bekerja dengan nilai numeric atau atribut numeric (Nabila et al., 2021).

Istilah-istilah dalam algoritma K-*Means Clustering* (Muliono & Sembiring, 2019):

1. *Cluster* : *Cluster* merupakan kelompok atau grup.
2. *Centroid* : *Centroid* adalah titik pusat untuk menetukan *Euclidian distance* (kedekatan nilai jarak dari dua variabel).
3. *Iterasi* : *Iterasi* ialah pengulangan dari suatu proses. Dan berhenti ketika hasil *Iterasi* telah konvergensi.

Langkah-langkah algoritma K-*Means* adalah sebagai berikut:

1. Tentukan nilai k, atau jumlah *Cluster* dalam data set.
2. Menentukan nilai pusat (centroid). Penentuan nilai centroid pada tahap awal dilakukan secara random, sedangkan pada tahap iterasi digunakan rumus seperti

dibawah ini:

$$V_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i} Xkj \text{.....(1)}$$

Keterangan:

V_{ij} = *Centroid* rata-rata *cluster* ke-I untuk variabel ke-i

N_i = Jumlah anggota *cluster* ke-i

i, k = Indeks dari *cluster*

j = Indeks dari variabel

Xkj = nilai data ke-k variabel ke-j untuk *cluster* tersebut

3. Menghitung jarak antara titik *centroid* dengan titik tiap objek menggunakan *Euclidean Distance*. *Euclidean Distance* adalah jarak garis lurus biasa antara dua titik dalam ruang *Euclidean* dengan rumus seperti dibawah ini:

$$De = \sqrt{(x_i - s_i)}^2 + \sqrt{(y_i - t_i)}^2 \text{.....(2)}$$

Keterangan:

De = *Euclidean Distance*

i = Banyaknya objek

(x, y) = Koordinat objek

(s, t) = Koordinat *centroid*

4. Kelompokkan objek berdasarkan jarak ke centroid terdekat.

5. Ulangi langkah ke-2 hingga ke-4, lakukan iterasi hingga centroid bernilai optimal.

2.7 *Silhouette Coefficient*

Silhouette Coefficient merupakan metode untuk menguji ketepatan sebuah *cluster* yang sudah terbentuk dari proses *Clustering*. *Silhouette Coefficient* adalah gabungan dari 2 metode yaitu metode *separation* dan *cohesion*. Pada *Shilhouette Coefficient* terdapat angka antara nilai -1 sampai 1 jika nilai *Silhouette Coefficient* mendekati angka 1, maka

akan semakin baik pengelompokkan objek-objek sebuah *cluster* dan sebaliknya dimana *Shilouette Coefficient* mendekati angka -1, akan semakin buruk pengelompokkan data pada *cluster* tersebut.

Tahapan perhitungan *Silhouette Coefficient* adalah sebagai berikut :

1. Hitung rata-rata jarak dari suatu data dengan semua data lain yang berada di satu *cluster* :

$$a(i) = \frac{1}{|A| - 1} \sum j \in A, j \neq 1 \; d(i,j)$$

dimana :

a(i = Perbedaan rata-rata objek (i) ke semua objek lain pada A

d(i,j) = jarak antara data i dengan j

A = *Cluster*

2. Menghitung rata-rata jarak data *i* dengan semua data di *cluster* lain dan mengambil nilai terkecilnya :

$$d(i,C) = \frac{1}{|C|} \sum j \epsilon C \; d(i,j)$$

dimana:

d(*i*,*C*) = Perbedaan rata-rata objek (i) ke semua aobjek lain pada C

C = *cluster* lain selain *cluster* A

3. Setelah melakukan perhitungan *d(i,C)*, akan diambil nilai terkecil :

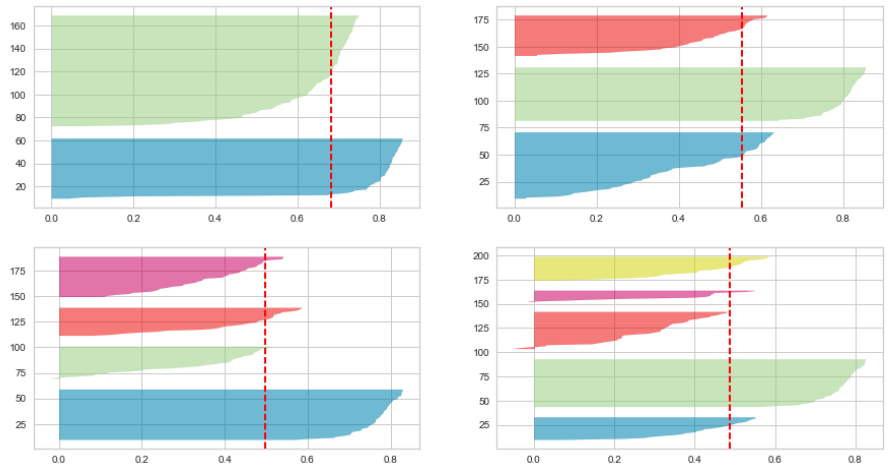
$$b(i) = \; min_{C \neq A} \, d(i,j)$$

4. Untuk menentukan nilai *Silhouette Coefficient* adalah sebagai berikut :

$$s(i) = \frac{(b(i) - a(i))}{\max a(i), b(i)}$$

Nilai *Silhouette Coefficient* rata-rata pada setiap data atau objek di sebuah *cluster* adalah sutau takaran yang menunjukan seberapa erat objek-objek dikelompokkan pada suatu *cluster*. *Silhouette Coefficient* menurut Kaufman dan Rousseeuw sebagai berikut :

Silhouette Coeffisien	**Struktur**
0.7 < SC < = 1	Struktur Kuat
0.5 < SC < = 0.7	Struktur Sedang
0.25 < SC < = 0.5	Struktur Lemah
SC < = 0.25	Tidak terstruktur



Gambar 2. 2 Grafik perhitungan *Silhouette Coefficient*

2.8 *Elbow Method*

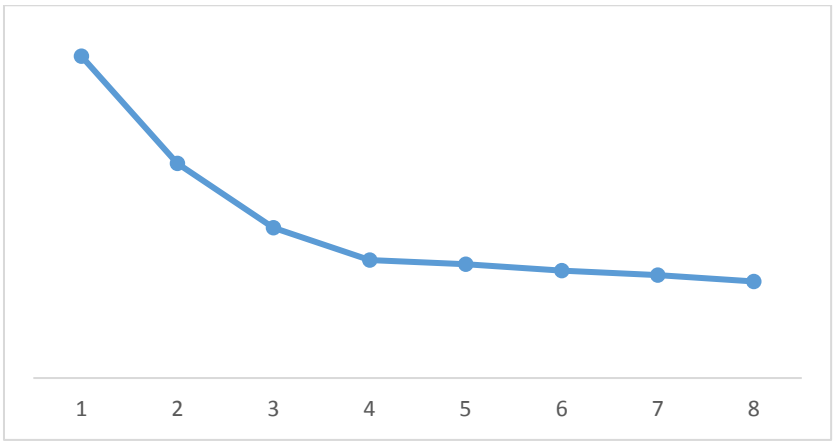
Elbow Method merupakan suatu metode yang digunakan untuk menghasilkan informasi dalam menentukan jumlah *cluster* terbaik dengan cara melihat presentase hasil perbandingan antara jumlah *cluster* yang membentuk siku pada suatu titik. Metode ini memberikan ide dengan memilih nilai *cluster* dan kemudian menambah nilai *cluster* tersebut untuk dijadikan model data dalam penentuan *cluster* terbaik. Hasil presentase yang berbeda dari setiap nilai *cluster* dapat ditunjuk dengan menggunakan grafik sebagai sumber informasinya (Putu et al., n.d.).

Elbow Method berperan penting dalam proses pengujian jumlah k pada model *Clustering*. Algoritma K-*Means Clustering* memiliki kelemahan saat menentukan jumlah k terbaik dari n percobaan (Zartesya & Komalasari, 2021).

Untuk mendapatkan perbandingannya yaitu dengan menghitung SSE (*Sum of Square Error*) dari tiap-tiap nilai *cluster*. Karena semakin besar jumlah *cluster* K akan semakin kecil nilai SSE. Berikut merupakan rumus SSE pada K-*Means* :

$$\sum_{k=1}^{k} \sum_{xi \in Sk} \|Xi - Ck\|_2^2$$

Setelah dilihat ada beberapa nilai K akan mengalami penurunan yang sangat signifikan yang kemudian turun secara perlahan sampai hasil dari nilai K akan stabil. Contohnya nilai *cluster* K=2 ke K=3, lalu dari K=3 ke k=4, sangat terlihat penurunan yang membentuk siku pada titik K=3 maka nilai *cluster* k idealnya adalah K=3.



Algortima Metode *Elbow* dalam menentukan nilai K pada K-Means :

1. Mulai
2. Inisiasi awal nilai K
3. Naikkan nilai K
4. Hitung hasil *sum of square error* dari tiap nilai K
5. Melihat hasil *sum of square error* dari nilai K yang turun secara drastic

6. Tetapkan niai K yan berbentuk siku
7. Selesai