BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Sebelumnya

Dalam penelitian ini menggunakan studi literatur yang dilakukan sebagai media dalam mencari referensi sekaligus kelengkapan data untuk menjelaskan masalah yang akan dikaji. Berdasarkan tema dan metode dalam penelitian, maka terdapat lima jurnal yang dipilih. Berikut ini merupakan jurnal yang terkait:

Dalam kajian yang menganalisis *K-Means*, Siti Azizatus Sholihah mengelompokkan provinsi berdasarkan prevalensi covid-19 di setiap provinsi di Indonesia. Kesimpulan dari penelitian ini adalah rata-rata *Silhouette Coefficient* (SC) untuk k = 2 adalah 0,74 dengan 3 data tidak akurat karena 3 negatif. Namun *Silhouette Coefficien* = 0,74 termasuk dalam struktur kokoh. Provinsi Jawa Timur, Jawa Tengah, Jawa Barat, DKI Jakarta, Banten, dan Riau tergolong rawan penyebaran virus covid-19 berdasarkan teknik *K-Means*. Sementara 28 provinsi dinyatakan bebas dari risiko penularan covid-19 [7].

Tujuan penelitian yang berjudul "Penerapan Metode *Clustering* Untuk Pengelompokan Kecenderungan Siswa Drop Out Menggunakan Algoritma *K-Means++*" oleh Fakhri Mohammad Falahi adalah untuk mengkategorikan siswa yang berpotensi drop out. Berdasarkan temuan penelitian ini, model optimal dihasilkan dengan memanfaatkan tiga cluster dengan nilai SC dan nilai kemurnian 0,815% dan 1. Karena Silhouette Coefficient dan nilai kemurnian mendekati 1, kinerjanya sangat baik [8].

Kajian selanjutnya yang berjudul "Pengelompokan Penyebaran Virus Corona (covid-19) di DKI Jakarta Menggunakan Metode *K-Means*" dilakukan oleh Khansa Khairunnisa dan Achmad Solichin dengan tujuan untuk mengidentifikasi Wabah virus corona di Provinsi DKI Jakarta. ODP, PDP, kasus positif dan pasien sehat dan pasien meninggal semuanya diperhitungkan. Berdasarkan perhitungan SSE dan hasil uji klasterisasi kasus covid-19 di DKI Jakarta menggunakan teknik *K-means*, terindikasi sembilan klaster.. Berdasarkan hasil penelitian yang dilakukan pengelompokan yang menggunakan tiga klaster yaitu 19 kecamatan C1, 23 kecamatan C2, dan 2 kecamatan C3 telah dikelompokkan [1].

Penelitian Irman Budiman, Muliadi, dan Rendi selanjutnya yang berjudul "Model Pengelompokan Provinsi di Indonesia Berdasarkan Kualitas Air Sungai Menggunakan Algoritma *K-Means*++" bertujuan untuk mengkategorikan kualitas air sungai agar lebih mudah mendapatkan air bersih dan layak minum. Penelitian ini mengklasifikasikan provinsi-provinsi di Indonesia menjadi tiga kelompok tergantung pada kualitas sungai, dengan dua provinsi pada klaster pertama memiliki indeks rendah, satu provinsi di klaster kedua memiliki indeks ringan, tiga puluh provinsi memiliki indeks sedang, dan indeks tinggi [9].

Penelitian sebelumnya oleh Ariadi Retno Hayati dan Mamluatul Hani'ah, Ika Kusumaning berjudul "Perbandingan Hasil *Clustering* Studi Kasus Posyandu Dengan Metode Clustering *K-Means*++ *Scalable*" membandingkan hasil pengelompokan menggunakan metode pengelompokan *K-Means*++ *Scalable*, teknik pengelompokan *K-Means*++, dan pendekatan pengelompokan K-Means. Teknik clustering *K-Means*++ dan metode clustering K-Means++ Scalable memberikan tingkat kesalahan yang lebih rendah daripada metodologi K-Means Clustering. Dengan menggunakan pendekatan *Scalable K-Means*++, nilai error data uji 1 memiliki nilai error minimal 0,07, data uji 2 memiliki nilai error minimal 0,15, dan data uji 3 memiliki nilai error minimal 0,005. Nilai error minimal algoritma K-Means++ Clustering untuk data uji 4 selama clustering adalah 0,15 [10].

Penelitian sebelumnya yang dilakukan oleh R.A. Indraputra dan R. Fitriana berjudul "*K-Means Clustering Data* covid-19*"* Untuk memerangi pandemi covid-19, penelitian ini bertujuan untuk mengumpulkan dan mengolah data Kaggle tentang covid-19 menggunakan metode *Data Mining K-Means Clustering*, yang mengelompokkan data berdasarkan mean terdekat. Hal ini memungkinkan pengelompokan wilayah sesuai dengan dampak *COVID*-nya, yang pada gilirannya memungkinkan penentuan prioritas dan penargetan bantuan *COVID*. Dua klaster data muncul dari analisis data, dengan klaster 2 memiliki jumlah orang yang terinfeksi dan mati lebih banyak daripada klaster 1, menunjukkan bahwa lokasi klaster ini harus diberikan prioritas untuk pengobatan [4].

Penelitian sebelumnya yang dilakukan oleh Ni Putu Eka Merliana, Ernawati, Alb. Joko Santoso berjudul "*ANALISA PENENTUAN JUMLAH CLUSTER TERBAIK PADA METODE K-MEANS CLUSTERING"* Pada contoh ini penelitian akan dilakukan di STAHN Tampung Penyang Palangka Raya untuk menentukan jumlah *cluster* pada proses *clustering* menggunakan pendekatan *K-Means*. Akibatnya, penelitian akan menggunakan metode *K-means* untuk artikel ini, identifikasi jumlah cluster yang sesuai. Ini dapat ditentukan dengan beberapa cara, salah satunya adalah pendekatan Elbow. Grafik Sum Square Error kumpulan data yang berbeda menunjukkan mengapa pendekatan ini dipilih. Dengan bertambahnya volume data, jumlah Square Error berkurang paling besar pada K = 3. dari SMA Hindu Negeri Tampung Penyang di Palangka Raya, di antara empat percobaan yang dijelaskan di atas. Ada tiga cluster yang harus digunakan dalam contoh ini untuk menetapkan karakteristik data: K = 3 [11].

Penelitian sebelumnya yang dilakukan oleh Dewa Ayu Indah Cahya Dewi, dan Dewa Ayu Kadek Pramita dengan judul *"Analisis Perbandingan Metode Elbow dan Sillhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali"* Dalam menentukan berapa banyak cluster yang dibutuhkan, penelitian ini membandingkan teknik elbow dan koefisien siluet. Sebuah indeks yang disebut Davies Bouldin Index digunakan untuk menguji hasil cluster. Dengan menggunakan pendekatan elbow, nilai DBI pada hasil uji clustering adalah 1,10. Sedangkan pada percobaan clustering diperoleh nilai DBI sebesar 1,06 dengan menggunakan *koefisien siluet*. Klustering *K-medoid* dengan koefisien siluet memiliki kualitas cluster yang lebih tinggi daripada k-medoid dengan pendekatan elbow karena nilai DBI yang lebih rendah [12].

Penelitian sebelumnya oleh Wiyli Yustanti, Naim Rahmawati, dan Yuni Yamasari yang berjudul *"Klastering Wilayah Kota/Kabupaten Berdasarkan Data Persebaran* covid-19 *di Propinsi Jawa Timur dengan Metode K-Means" K-Means*, pendekatan pengelompokan *non-hierarki*, digunakan untuk tujuan ini. Temuan pengelompokan diharapkan dapat mengungkapkan berapa banyak pengelompokan geografis yang berbeda secara substansial berbeda dari data Penyebaran virus covid-19 semakin meningkat.

Ada lima jumlah cluster ideal untuk *K-Means non-hierarki*, menurut pendekatan ini. Uji vektor rata-rata dijalankan menggunakan statistik *Wilks Lambda*, dan hasilnya menunjukkan bahwa lima klaster yang muncul berbeda secara statistik., dengan tingkat kepercayaan 95% [13].

Penelitian sebelumnya yang dilakukan oleh Arief Rachman, dan M. Reza Hidayat berjudul *"Klasterisasi Sumber Penyebaran Virus* covid-19 *dengan Menggunakan Metode K-MEANS Di Daerah Kota Cimahi dan Kab. Bandung Barat".* Ditemukan di lima klaster, yakni klaster 0, 1, 2, 3, dan 4, terdapat virus covid-19, khususnya di tempat kerja, restoran, rumah, dan pusat perbelanjaan (transportasi umum). Untuk Kota Cimahi dan Kabupaten Bandung Barat, Mei hingga Juli 2020. Klaster 1 sangat dirugikan oleh covid-19, agen Kecamatan. *cluster* 2 dan 3 rata-rata masing-masing 0,02 persen, 0,04 persen, dan 0,79 persen di Bandung Barat [14].

Ringkasan penelitian terdahulu pada Tabel 2.1 di bawah ini:

Tabel 2. 1 Penelitian Terdahulu

No	Peneliti	Judul	Tahun	Metode	Hasil	Perbedaan
1	Siti Azizatus Sholihah	Analisis *Cluster* Untuk Pemetaan Data Kasus Covid-19 Di Indonesia Menggunakan K - Means	2021	*K-Means*	Hasil penelitian ini berhasil mengelompokkan setiap provinsi di Indonesia berdasarkan zona sebagai berikut: Provinsi Jawa Timur, Jawa Tengah, Jawa Barat, DKI Jakarta, Banten, dan Riau ditetapkan sebagai lokasi rawan penularan virus *Covid-19* dengan menggunakan pendekatan *K-Means*. Sementara itu, 28 provinsi telah ditetapkan sebagai zona bebas covid-19.	Penelitian ini hanya menggunakan satu metode saja sehingga hasilnya tidak maksimal karena *K-Means* memiliki kekurangan dalam menentukan centroid(titik pusat), maka dari itu penulis melakukan penelitian dengan data yang sama tetapi menggunakan 2 metode berbeda sesuai dengan saran yang diberikan yaitu *K- Means* dan *K-Means++*
2	Fakhri Mohammad Falahi	Penerapan Metode *Clustering* Untuk Pengelompokan Mahasiswa Potensial Drop Out Menggunakan Algoritma K-Means++	2019	*K-Means++ Purity*	Temuan tersebut dibagi menjadi dua kelompok, dengan 87,1 persen potensi putus sekolah rendah dan 12,9 persen potensi putus sekolah tinggi. Hasil tersebut dibagi menjadi tiga klaster, masing-masing dengan proporsi potensi putus sekolah rendah 25,9%, potensi putus sekolah sedang 62,2 persen, dan potensi putus sekolah tinggi 12,9 persen. Hasilnya dibagi menjadi empat klaster potensi putus sekolah sangat rendah 25,5 persen, potensi putus sekolah 62,1 persen, potensi putus sekolah tinggi 6%, dan potensi putus sekolah sangat tinggi 6,4 persen. Hasil perhitungan	Perbedaannya penelitian ini menggunakan *Purity* yang digunakan untuk menentukan model dengan performa terbaik dan hanya menggunakan satu metode saja

No	Peneliti	Judul	Tahun	Metode	Hasil	Perbedaan
					menggunakan 5 klaster dengan potensi putus sekolah sangat rendah 25,4 persen, potensi putus sekolah rendah 17,9 persen, potensi putus sekolah 44,3 persen sedang, potensi putus sekolah 6% tinggi, dan potensi putus sekolah sangat tinggi 6,4 persen. Model terbaik terdiri dari tiga cluster dengan nilai SC 0,815 dan 1. Karena Silhoette Coefficient dan nilai kemurnian mendekati satu, kinerjanya sangat baik.	
3	Achmad Solichin, Khansa Khairunnia	Klasterisasi Persebaran Virus Corona (Covid-19) Di DKI Jakarta Menggunakan Metode K-Means	2020	*K-Means*	Jumlah cluster yang disarankan berdasarkan perhitungan SSE adalah 9 cluster, sesuai dengan hasil pengujian dan pengujian penerapan pendekatan K-Means untuk *clustering* kasus covid-19 di DKI Jakarta. Hanya 3 (tiga) klaster yang digunakan dalam penelitian ini, sehingga terdapat 19 kecamatan C1, 23 kecamatan C2, dan 2 kecamatan C3.	Perbedaan dengan penelitian ini yaitu penelitian ini dibuatkan menjadi system sedangkan penelitian yang penulis buat hanya sampai perhitungan saja

No	Peneliti	Judul	Tahun	Metode	Hasil	Perbedaan
4	Irman Budiman, Muliadi, Rendi	Model Pengelompokan Provinsi di Indonesia Berdasarkan Kualitas Air Sungainya Menggunakan Algoritma *K-Means++*	2021	*K-Means++*	Penelitian ini menghasilkan pengelompokan provinsi-provinsi Indonesia menjadi tiga kelompok berdasarkan kualitas air sungai, dengan dua provinsi pada klaster pertama tergolong indeks rendah, satu provinsi pada klaster kedua tergolong indeks sedang, dan tiga puluh provinsi tergolong indeks tinggi	Perbedaan dengan penelitian ini yaitu penelitian ini tidak menggunakan *Silhouette Coefficient* untuk evaluasi, tetapi hanya mengelompokan data air sungai bersih.
5	Ariadi Retno Hayati, Mamluatul Hani'ah, Ika Kusumaning	Comparison of result *clustering* study case posyandu with the *scalable K Means ++ Clustering* Method	2020	*Scalable K-Means++ Clustering Method*	Hasil penelitian menghasilkan nilai error yang lebih kecil jika dibandingkan dengan metode *K-Means Clustering*. Nilai error data pengujian 1 mendapatkan error *clustering K- Means++ Scalable* minimal 0,07, data pengujian 2 mendapatkan error *K-Means++ Clustering* minimal 0,15, data pengujian 3 mendapatkan error *Scalable K-Means++ Clustering* minimal 0,005, dan data pengujian 4 menerima nilai kesalahan minimum 0,15 di *K-Means++ Clustering.*	Perbedaan dengan penelitian ini yaitu penelitian ini tidak dilakukan evaluasi, tetapi hanya membandingkan dari ketiga algoritma tersebut yang terbaik untuk pengelompokkan suatu dataset, dengan *Scalable K-Means++* yang merupakan pengembangan dari *K- Means++*

No	Peneliti	Judul	Tahun	Metode	Hasil	Perbedaan
6	R.A. Indraputra dan R. Fitriana	*K-Means Clustering Data* Covid-19	2020	*K-Means Clustering*	Hal ini memungkinkan pengelompokan wilayah sesuai dengan dampak covid-nya, yang pada gilirannya memungkinkan penentuan prioritas dan penargetan bantuan covid. Dua klaster data muncul dari analisis data, dengan klaster 2 memiliki jumlah orang yang terinfeksi dan mati lebih banyak daripada klaster 1, menunjukkan bahwa lokasi klaster ini harus diberikan prioritas untuk pengobatan	Penelitian ini dilakukan hanya untuk mengolah data covid-19 yang berasal dari Kaggle yang dibentuk *cluster* berdasarkan jumlah terjangkit dan meninggal sehingga tidak dilakukan pengujian untuk memastikan *cluster* tersebut sudah optimal atau belum.
7	Ni Putu Eka Merliana, Ernawati, Alb. Joko Santoso	Analisa Penentuan Jumlah Cluster Terbaik Pada Metode *K-Means Clustering*	2021	*K-Means Clustering*	Dengan jumlah data yang bervariasi dari SMA Hindu Negeri Tampung Penyang Palangka Raya, frekuensi kesalahan kuadrat menurun paling drastis pada K = 3, di antara empat percobaan yang dijelaskan di atas. Ada tiga cluster yang harus digunakan dalam contoh ini untuk menetapkan karakteristik data: K = 3	Penelitian ini dilakukan untuk mencari *cluster* terbaik tetapi menggunakan motode *Elbow* kurang maksimal sehingga disarankan menggunakan metode lainnya.

No	Peneliti	Judul	Tahun	Metode	Hasil	Perbedaan
8	Dewa Ayu Indah Cahya Dewi, dan Dewa Ayu Kadek Pramita	Analisis Perbandingan Metode *Elbow* dan *Sillhouette* pada Algoritma *Clustering K-Medoids* dalam Pengelompokan Produksi Kerajinan Bali	2019	*K-Medoids Clustering*	Dengan menggunakan pendekatan *elbow*, nilai DBI pada hasil uji clustering adalah 1,10. Sedangkan pada percobaan *clustering* diperoleh nilai DBI sebesar 1,06 dengan menggunakan *koefisien siluet*. Klustering *K-medoid* dengan koefisien siluet memiliki kualitas cluster yang lebih tinggi daripada *k-medoid* dengan pendekatan *elbow* karena nilai DBI yang lebih rendah	Penelitian ini menguji antara metode *Elbow* dan *Silhouette Coefficient* menggunakan *Davies Bouldin Index* (DBI) digunakan untuk menguji hasil *cluster Silhouette coefficient* memiliki kualitas *cluster* yang lebih tinggi daripada *Elbow* karena nilai DBI lebih rendah
9	Wiyli Yustanti, Naim Rahmawati, dan Yuni Yamasari	Klastering Wilayah Kota/Kabupaten Berdasarkan Data Persebaran Covid-19 di Propinsi Jawa Timur dengan Metode *K-Means*	2020	*K-Means Clustering*	Dengan menggunakan teknik *non-Hierarchical K-Means clustering*, ditentukan jumlah *cluster* yang optimal adalah lima cluster. Statistik *Wilks Lambda* digunakan untuk melakukan uji *vektor* temuan menunjukkan bahwa rata-rata lima kelompok sangat berbeda satu sama lain, dengan tingkat kepercayaan 95%	Penelitian ini dilakukan pengujian menggunakan statistic *Wilks Lambda* untuk mencari tingkat kepercayaan dari pengujian menggunakan *Wilks Lambda*
10	Arief Rachman, dan M. Reza Hidayat	Klasterisasi Sumber Penyebaran Virus Covid-19 dengan Menggunakan Metode *K-MEANS* Di Daerah Kota Cimahi dan Kab. Bandung Barat	2020	*K-Means Clustering*	Klaster 0 (perkantoran), klaster 1 (restoran), klaster 2 (rumah), klaster 3 (penyebaran virus melalui belanja), dan klaster 4 (penyebaran melalui pertokoan) merupakan lima klaster sumber virus covid-19 yang dianalisis (publik kendaraan). Mulai Mei hingga Juli 2020, Kota Cimahi dan Kabupaten Bandung Barat akan ditutup. covid-19 dari Kab. Bandung Barat paling banyak merugikan	Penelitian ini hanya mengelompokkan data covid-19 berdasarkan sumber penyebarannya, tidak diuji setiap clusternya

No	Peneliti	Judul	Tahun	Metode	Hasil	Perbedaan
					Klaster 1, dengan rata-rata 0,55%; Klaster 2 memiliki rata-rata 0,02%; dan Klaster 3 memiliki rata-rata 0,40%; dan cluster 4 memiliki rata-rata 0,79 persen	

2.2 Dasar Teori

2.2.1 Covid – 19

Pandemi covid-19 adalah epidemi di seluruh dunia yang telah mempengaruhi orang-orang di seluruh dunia. Virus SARS-CoV2 (sindrom pernapasan akut yang parah) harus disalahkan atas pandemi ini,yang memiliki gejala termasuk suhu 37°C, batuk kering, ketidaknyamanan sendi, mudah lelah, tenggorokan sore, dan banyak lagigejala yang tidak diketahui. Penderita sesak napas harus menghindari virus ini karena menyerang saluran pernapasan, menyebabkan sesak napas, kesulitan berbicara, bahkan kematian. Sebelum menunjukkan gejala ini Orang yang baru saja mengunjungi tempat-tempat yang terkena dampak *covid–19* harus tinggal di rumah dan melakukan isolasi diri selama 14 hari, setelah itu dapat dilakukan pemeriksaan kesehatan untuk memastikan individu tersebut tidak tertular virus [15].

Virus ini menyebar melalui udara yang dihembuskan ketika seseorang batuk atau bersin. Untuk menghindari penularan, orang yang sedang flu atau batuk harus memakai masker yang memenuhi syarat kesehatan. Selain itu, virus ini memiliki istilah terkait antara lain:

- ***Suspect***

  Seseorang bisa disebut suspect jika terdapat beberapa kriteria di bawah ini :

  1. Mengalami demam atau masalah pernapasan
  2. Riwayat kontak langsung dengan orang yang terkena atau diduga terinfeksi covid-19.
  3. Riwayat perjalanan ke tempat-tempat umum atau tempat tinggal selama kurang dari 14 hari sebelum timbulnya gejala.

- ***Probable* (diduga)**

  Orang tersebut dirawat sebagai tersangka yang meninggal karena penyakit saluran pernapasan akut. Namun, tidak ada data tesyang menunjukkan dia telah terpapar covid-19.

- **Terkonfirmasi atau Positif**

  Pasien bisa dinyatakan positif terpapar virus *Covid – 19* yangtelah dibuktikan dengan pemeriksaan laboratorium.

- **Selesai Isolasi/sembuh**

  Orang sudah bisa dinyatakan selesai isolasi/sembuh dari virus covid-19 setelah dibuktikan dengan pemeriksaan laboratorium yang menyatakan sudah negative.

- **Kematian**

  Kondisi ketika orang yang meninggal terdapat dalam kasus diduga atau dinyatakan positif covid-19

2.2.2 Vaksinasi

Vaksinasi adalah memberikan suntikan untuk secara aktif membangun atau meningkatkan kekebalan seseorang terhadap suatu penyakit, sehingga jika mereka terkena di kemudian hari, mereka hanya memiliki penyakit ringan atau tetap sehat, dan tidak menjadi sumber dari transmisi, untuk mencegah penyebaran covid-19 masyarakat diberikan 2 dosis vaksin yang memiliki fungsi vaksinasi awal berfungsi sebagai dosis utama atau dosis utama untuk mengantarkan virus covid-19 yang tidak aktif ke dalam tubuh penerima. Sedangkan dosis vaksinasi kedua berfungsi sebagai dosis booster atau penambah efektifitas vaksin. Beberapa jenis vaksin yang masuk di Indonesia: Vaksin Sinovac, Vaksin Astrazeneca, Vaksin Moderna, Vaksin Sinapharm, Vaksin Pfizer, Vaksin Novavax, Vaksin Sputnik-V, Vaksin Jassen, Vaksin Convidencia, Vaksin Zifivax. Dosis pertama dan kedua vaksin covid-19 memiliki keuntungan sebagai berikut: Mereka meningkatkan fungsi kekebalan tubuh, menurunkan risiko penularan, mengurangi keparahan virus, dan menciptakan kekebalan kelompok.

2.2.3 Clustering

Clustering adalah proses pengelompokan pola-pola yang tidak memiliki nama dan dilakukan secara mandiri menjadi suatu kelompok dengan kualitas tertentu. Pengelompokan sangat penting dalam berbagai situasi, seperti analisis pola, pengambilan keputusan, pembelajaran mesin, dan penambangan data. Teknik *clustering* termasuk teknik data mining yang terkenal dan umum digunakan [16].

Data mining adalah proses memperoleh pengetahuan dari sejumlah besar data dengan menggunakan metode dari berbagai disiplin ilmu, termasuk pemrosesan sinyal, basis data dan teknologi pergudangan data, statistik, pengenalan pola, jaringan saraf, dan visualisasi data. Data mining dengan demikian dianggap sebagai salah satu bidang pengetahuan yang paling penting dalam database dan sebagai salah satu bidang yang paling menjanjikan dalam pengembangan teknologi informasi *interdisipliner*. Metode *k-means* memiliki banyak manfaat, termasuk konvergensi yang cepat dan implementasi yang sederhana [11].

2.2.4 Analisis Cluster

Dimungkinkan untuk menggunakan pendekatan analisis klaster untuk mengevaluasi banyak faktor dan kemudian membuat k klaster dengan (kn) berdasarkan variabel p, menghasilkan distribusi objek yang lebih seragam. Untuk proses pengelompokan data, terdapat dua bentuk data klaster yaitu data klaster hierarkis (hierarkis) dan data klaster non-hierarkis (non-hierarkis) [17]. Struktur pohon Dendogram, misalnya, dapat digunakan untuk menunjukkan hasil pengelompokan menggunakan metodologi hierarkis. Metode hirarki adalah prosedur pengelompokan yang dilakukan secara bertahap. Metode aglomerasi (juga dikenal sebagai metode aglomerasi) dan metode devisive adalah metode yang dapat digunakan dalam metode hierarkis. Dendogram adalah bagaimana cluster tumbuh dan bagaimana nilai koefisien jarak berubah pada setiap fase analisis cluster ditunjukkan dalam penggambaran grafik ini (metode pembagian). Strategi cluster yang digunakan dalam metode non-hierarki melibatkan penentuan secara manual berapa banyak cluster yang harus ada. Dimulai dengan pemilihan nilai cluster awal tergantung pada jumlah yang diperlukan, proses non-hierarki kemudian mengelompokkan objek yang diamati ke dalam cluster tersebut. Pendekatan sekuensial termasuk dalam pendekatan non-hierarki ini antara lain *sequential threshold, parallel threshold*, dan *optimizing partitioning* [18].

Kemudian menghitung jarak Euclidean setiap titik data dari titik awal pusat cluster, memilih sampel terdekat, dan mengalokasikannya ke cluster yang tepat. Pusat cluster diperbarui sampai kesalahan kuadrat rata-rata turun di bawah ambang batas tertentu atau pusat cluster berhenti bergerak, yaitu mencapai tengah. Semua titik data memiliki jarak minimum dari pusat titik pada saat ini [19]. Kelebihan dari analisis cluster antara lain : Memiliki kemampuan untuk menggabungkan sejumlah besar kumpulan data observasional. Lebih mudah untuk menganalisis data yang telah dipecah ke dalam kategori yang berbeda. Diterapkan pada tipe data ordinal, interval, dan rasio [20].

Tujuan analisis cluster untuk membentuk kumpulan item terkait. Semakin dekat dua objek, semakin mirip nilainya semakin terpisah mereka, semakin bervariasi nilainya. Beberapa proses yang dilakukan untuk analisis cluster [21].

Terdapat proses analisis cluster yaitu [7] :

1. Menentukan kemiripan antar objek.

   Memanfaatkan kedekatan antar item dan mengukur jarak antara dua hal, pendekatan ini digunakan untuk menilai sebagian besar kesamaan antara objek.

2. Proses standarisasi data.

   Disparitas antara nilai masing-masing variabel yang disebabkan oleh perbedaan skala merupakan salah satu faktor yang dipertimbangkan saat melakukan standarisasi data. Metode modifikasi yang dikenal sebagai "normalisasi data" menyesuaikan nilai dari -1 ke 1 atau 0 ke 1. Salah satunya, pendekatan Normalisasi *Min-Max* dengan rumus, digunakan untuk menormalkan data. Berikut perhitungan dengan Persamaan (2.1) :

$$X^1 = \frac{X - X_{min}}{X_{max} X_{min}} \tag{2.1}$$

   di mana,

   $X1$ = Data atribut yang akan dinormalisasikan.

   $Xmin$ = Nilai terkecil atribut tersebut

   $Xmax$ = Nilai tertinggi atribut tersebut.

   Metode normalisasi *Min-Max* digunakan untuk menormalkan data. Normalisasi akan memodifikasi data asli secara linier untuk membangun keseimbangan nilai perbandingan antara data sebelum dan sesudah operasi.

3. Melakukan pengelompokan.

   Pengelompokan dilakukan dengan cara menghitung ukuran jarak dan melihat sifat-sifat barang tersebut.

4. Melakukan validasi *cluster*

   Validitas hasil selanjutnya dievaluasi pada *cluster* yang telah dikembangkan. *Cluster* dianggap asli jika mereka dapat menentukan apakah ide jarak cluster antara *cluster* tetangga memiliki nilai minimum dan jarak antar *cluster* memiliki nilai maksimum.

2.2.5 Algoritma *K-Means*

Menggunakan jarak *Euclidean*, Algoritma *K-Means Clustering* adalah cara yang terkenal untuk mengklasifikasikan data. Dalam metode ini, data dengan kualitas yang sama dikelompokkan menjadi satu, sedangkan data dengan karakteristik yang bervariasi dipisahkan menjadi beberapa kelompok. Pengorganisasian data ini ke dalam kelompok memiliki dua tujuan: untuk mengurangi jumlah variasi dalam satu kelompok dan untuk mendapatkan informasi terbanyak tentang perbedaan antar kelompok. Adapun beberapa tahapan dalam penggunaan algoritma dari *K-Means Clustering* [22]:

1. Memasukkan data yang sudah dipilih variable nya untuk di klaster
2. Menentukan banyaknya klaster yang akan dibentuk
3. Mengambil data secara sembarang sebanyak jumlah klasteruntuk dijadikan titik pusat klaster (*Centroid)*.
4. Menghitung jarak Euclidean dengan Persamaan (2.2) :

$$d_{(x,y)} = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2} \tag{2.2}$$

Penjelasan:

$d(x, y)$ = Jarak data ke x ke pusat cluster y

x_i = data x pada observasi ke-i

y_i = titik pusat ke y observasi ke-i

n = banyaknya observasi

5. Dihitung kembali titik pusat klater menggunakan anggota klasteryang baru berdasarkan kedekannya dengan *centroid* (jarak terkecil)
6. Memperbaharui nilai *centroid*. Nilai *centroid* baru akan didapatkan dari rata – rata *cluster* yang berhubungan dengan Persamaan (2.3):

$$V_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i} X_{kj} \tag{2.3}$$

Penjelasan:

V_{ij} = centroid rata – rata pada cluster ke-I untuk variable ke – j

N_i = jumlah anggota *cluster* ke-*i*

i,k = indeks dari cluster

j = indeks variable

X_{kj} = nilai data ke – k variable ke -j untuk cluster tersebut

7. Apabila titik pusat tidak ada perubahan antara nilai titik pusat baru dengan nilai titik pusat lama terakhir, jika prosedur clustering belum dianggap selesai, ulangi langkah 3 sampai tidak ada perubahan pada cluster center.

2.2.6 Algoritma *K-Means++*

Dengan *K-Means++*, pendekatan *K-means* telah ditingkatkan. Dibandingkan dengan pendekatan *K-means*, satu set nilai awal baru digunakan. *K-Means++* dikembangkan untuk mengurangi efek buruk algoritma *K-means*, yang didasarkan pada nilai awal dari [6]. Menghitung *K-Means++*menggunakan Persamaan (2.4) :

$$K = \frac{D(x)^2}{\sum_{x \epsilon X} D(x)^2} \tag{2.4}$$

$D(x)^2$ = Jarak *Euclidean distance*

$\sum_{x \epsilon X} D(x)^2$ = jumlah jarak *Euclidean distance*

Adapun beberapa Langkah – Langkah algoritma *K-Means++*sebagai berikut :

1. Memilih nilai k pusat *cluster* pertama secara acak dari titik data
2. Untuk setiap data akan diamati x, untuk dihitung jarak $D(x)$ kepusat *cluster* terdekat.
3. Memilih *cluster* baru yang berasal dari antara titik data, dengan probabilitas x dipilih proporsional dengan $D(x)^2$.
4. Ulangi Langkah 2 dan 3 sampai k pusat *cluster* telah dipilih
5. Apabila titik pusat tidak ada perubahan antara nilai titik pusat baru dengan nilai titik pusat lama maka proses klaster dinyatakan selesai

2.2.7 Silhouette Coefficient

Pengujian diperlukan dalam sebuah model untuk mengumpulkan informasi tentang kedekatan hubungan antara satu item dengan item lainnya dalam sebuah *cluster*, serta jarak antar *cluster*. Dalam hal ini, metode pengujiannya adalah teknik Kohesi yang mengukur seberapa dekat objek terkait dalam sebuah *cluster*, dan metode Pemisahan, yang menentukan seberapa jauh jarak satu item dari yang lain, digabungkan untuk membuat Koefisien Silhouette berbeda dari kelompok lain [10].

Adapun langkah untuk menghitung *Silhouette coefficient* di awali dengan mencari jarak rata-rata data ke-*i* dengan semua data dalam *cluster* yang sama, disini anggap saja data ke-*i* terletak pada *cluster* A. Rumus dari *a(i)* terdapat di Persamaan (2.5) [23].

$$a(i) = \frac{1}{|A|-1} \sum_{j \epsilon A, j \neq 1} d(i,j) \tag{2.5}$$

di mana,
A = total data di *cluster* A

kemudian tentukan nilai *b(i)* yang merupakan nilai minimal dari rata-rata jarak data ke-*i* antara semua data pada setiap *cluster*. Sekarang mari kita perhatikan *cluster* lain selain A dan *cluster* C. Untuk menghitung jarak rata-rata data ke-*i* dengan semua data di *cluster* C, menggunakan Persamaan (2.6).

$$d(i,C) = \frac{1}{|C|} \sum_{j \epsilon C} d(i,j) \tag{2.6}$$

di mana,
C = total data di *cluster* C

Setelah menghitung *d(i,C)* untuk semua *cluster* $C \neq A$, Langkah selanjutnya untuk memilih jarak paling minimum sebagai nilai *b(i)*, menggunakan Persamaan (2.7)

$$b(i) = \min_{C \neq A} d(i,j) \tag{2.7}$$

Cluster terbaik kedua untuk data ke-*i*, setelah *cluster* A, adalah *cluster* B, yang disebut sebagai tetangga dari data ke-*i* jika *cluster* B memiliki nilai jarak minimal, yaitu *d(i,B) = b*(*i*). *Silhouette coefficient* ditentukan pada langkah terakhir setelah *a(i)* dan *b(i)* diketahui.

Jika nilai indeks *Silhouette* mendekati 1 berarti proses *clustering* berhasil, dan jika mendekati 0, berarti prosedur *clustering* tidak berhasil. Berikut rumus *silhouette coefficient*, pada Persamaan (2.8)

$$S(i) = \frac{b(i)-a(i)}{\max(a(i),(b(i)} \quad (2,8)$$

$s(i)$ = nilai *silhouette coefficient* pada object ke-i

(i) = object yang akan diteliti

$a(i)$ = rata-rata jarak antar anggota dalam *cluster* yang sama.

$b(i)$ = nilai minimal jarak rata-rata antara item ke-i dengan objek pada cluster lain

Nilai *s(i)* berletak diantara -1 dan 1, di mana setiap nilai ditafsirkan sebagai berikut:

s(i) = 1 => data ke-*i* digolongkan dengan baik (dalam A)
s(i) = 0 => data ke-*i* terletak di tengah antara dua *cluster* (A dan B)
s(i) = -1 => data ke-*i* tergolong lemah (dekat ke *cluster* B daripada A)

Tabel 2. 2 Silhouette Coefficient

Range	Interpretasi
0.71 – 1.00	Struktur yang dihasilkan kuat
0.51 – 0.70	Struktur yang dihasilkan baik
0.26 – 0.50	Struktur yang dihasilkan lemah
≤ 0.25	Tidak terstruktur