

BAB I

PENDAHULUAN

1.1. Latar Belakang

Knowledge Discovery in Database (KDD), atau yang sering disebut sebagai *data mining*, merupakan proses untuk menemukan pola-pola baru, bermanfaat, dan mudah dipahami dari kumpulan data yang besar (Nanda & Rout, 2010). Proses ini mencakup beberapa tahapan, seperti prapemrosesan data, pemilihan fitur, penemuan pola, dan interpretasi (Kaur dkk., 2014). KDD menggabungkan berbagai teknik dari bidang pembelajaran mesin, statistika, dan sistem basis data untuk mengekstrak informasi berharga dari struktur data yang kompleks (Vasudha, 2020). Beberapa jenis pola yang sering dipelajari meliputi klasifikasi, aturan asosiasi, dan klasterisasi (Dong, 2002). *Data mining* memiliki peran penting dalam pengambilan keputusan bisnis dan ilmiah, menekankan nilai dari penemuan pengetahuan tingkat tinggi dari data (Kaur dkk., 2014).

Data mining melibatkan berbagai metode, seperti *supervised*, *unsupervised*, dan *semi-supervised learning*, yang digunakan untuk mengekstrak wawasan berharga dari data (Sivasankari & Sukumaran, 2024). *Supervised learning* bekerja dengan data yang telah diberi label untuk tugas klasifikasi atau regresi, sementara *unsupervised learning* digunakan pada data tanpa label, biasanya untuk klasterisasi (Jain dkk., 2022). *Semi-supervised learning* diterapkan ketika hanya tersedia sedikit data berlabel, tetapi terdapat banyak data tidak berlabel (Jain dkk., 2022). Pilihan antara *supervised* dan *unsupervised learning* bergantung pada faktor-faktor seperti tujuan aplikasi, ketersediaan data, dan karakteristik (Nurhalizah dkk., 2024). Pemilihan metode sangat bergantung pada tujuan penelitian, jenis data, dan sumbernya (Pireddu dkk., 2024).

Klasterisasi adalah salah satu teknik dasar dalam *unsupervised learning* yang digunakan untuk mengelompokkan data yang memiliki kemiripan ke dalam klaster

tertentu, tanpa perlu kelas yang telah ditentukan sebelumnya (Dalal & Harale, 2011). Teknik ini banyak digunakan dalam analisis data eksploratif, memungkinkan penemuan pola dan struktur yang bermakna dalam dataset (Caron dkk., 2018). Berbeda dengan klasifikasi yang termasuk dalam *supervised learning*, klasterisasi tidak memerlukan data berlabel, melainkan mengandalkan ukuran kemiripan atau jarak untuk membentuk kelompok. Algoritma klasterisasi dapat diaplikasikan pada berbagai jenis data, seperti dokumen atau dataset multidimensi, menjadikannya alat yang fleksibel untuk mengungkap hubungan tersembunyi dalam data yang kompleks (Mukhopadhyay, 2018).

Klasterisasi adalah metode untuk melakukan pengelompokan pada data ke dalam beberapa klaster atau kelompok sehingga data memiliki tingkat kemiripan yang paling tinggi dalam satu klaster dan memiliki tingkat kemiripan yang paling rendah antarklaster (P.-N. Tan dkk., 2006). Klasterisasi merupakan teknik yang tidak membuat asumsi awal mengenai jumlah kelompok atau struktur kelompok, informasi yang diperlukan untuk melakukan klasterisasi adalah ukuran kemiripan atau data yang kemiripannya dapat dihitung (Johnson & Wichern, 2007). Klasterisasi adalah proses untuk memberi kelompok pada data baik berdasarkan kemiripan maupun ketidaksamaan yang terdapat pada data tersebut (Han dkk., 2011).

Klasterisasi banyak digunakan di berbagai bidang. Dalam konteks kendaraan listrik, metode klasterisasi telah digunakan untuk mengatasi tantangan yang muncul dari meningkatnya adopsi kendaraan listrik (Nazari dkk., 2023). Sementara itu, Fang & Liu (2021) telah menggunakan teknik klasterisasi untuk klasifikasi pelanggan ritel, meningkatkan efisiensi dan kemampuan mereka untuk mengidentifikasi jumlah klaster yang optimal. Demikian pula, Ahmed dkk. (2023) telah meninjau metode klasterisasi untuk analisis teks pendek, mengatasi masalah seperti kelangkaan data dan informasi yang terbatas dalam konten media sosial, dengan aplikasi dalam analisis sentimen dan pemfilteran spam.

Klasterisasi dapat dikategorikan secara luas ke dalam dua metode umum, yaitu metode partisi dan hirarki (Rani & Rohil, 2013). Metode partisi, secara khusus, membuat beberapa partisi data dan menilai kualitasnya menggunakan kriteria tertentu (Bano & Khan, 2018). Pendekatan ini melibatkan pengelompokan pola data ke dalam sejumlah klaster berdasarkan kesamaan mereka (Menéndez, 2021). Di antara berbagai teknik klasterisasi partisi, beberapa metode yang paling sering digunakan antara lain *K-Means*, *K-Medoids*, dan *Fuzzy K-Means* (Goel, 2014).

Istilah *k-means* digunakan untuk merujuk pada algoritma yang mengelompokkan setiap item ke dalam klaster dengan titik pusat (rata-rata) yang paling dekat (Macqueen, 1967). *K-means* merupakan salah satu metode klasterisasi data nonhirarki yang mempartisi data ke dalam bentuk satu atau lebih klaster, sehingga data yang memiliki karakteristik sama dikelompokkan dalam satu klaster, sedangkan data yang memiliki karakteristik berbeda dikelompokkan dalam klaster lain sehingga data yang berada dalam satu klaster memiliki tingkat variasi yang kecil (Agusta, 2007). Terdapat beberapa karakteristik data yang cocok untuk diklasterisasi menggunakan *k-means*. Metode *k-means* dikenal luas karena efisiensinya dalam mengelompokkan set data numerik (Hamzah dkk., 2017). Selain itu, *k-means* lebih efektif dalam ruang berdimensi rendah, karena data berdimensi tinggi dapat menyebabkan masalah seperti *curse of dimensionality*, yang mempengaruhi perhitungan jarak (Ikotun dkk., 2023).

Metode *k-means*, meskipun populer, memiliki keterbatasan yang signifikan, terutama sensitivitasnya terhadap pemilihan *centroid* awal. Inisialisasi acak dapat menyebabkan hasil pengelompokan yang tidak konsisten (Aldahdooh & Ashour, 2013). Terlepas dari keterbatasan ini, *K-means* dikenal karena kesederhanaan, efisiensi, dan kemudahan implementasinya (Bao, 2021). *K-means* juga dipuji karena efisiensi komputasi, fleksibilitas, dan kemampuan interpretasinya (Narang dkk., 2016). Selain itu, metode ini telah diadaptasi untuk memberikan perkiraan yang efisien dan dapat diskalakan, sehingga cocok untuk menangani dataset yang besar (Capó dkk., 2018).

Untuk mengatasi kelemahan *k-means* dalam menentukan pusat kluster awal, dapat digunakan metode penggabungan *k-means* dengan metode kluster lain yang termasuk dalam kluster hirarki. Klasterisasi hirarki adalah partisi rekursif dari sebuah dataset ke dalam kluster-kluster dengan granularitas yang semakin halus, dengan mengoptimalkan sebuah *cost function* (Cohen-Addad dkk., 2019). Klasterisasi hirarki terdiri dari dua algoritma yaitu divisif dan aglomeratif (Johnson & Wichern, 2007). Terdapat beberapa jenis metode klasterisasi secara hirarki dengan algoritma aglomeratif, yaitu *single linkage*, *complete linkage*, *average linkage*, *centroid linkage*, *median*, dan metode Ward.

Beberapa penelitian telah mengusulkan penggabungan *k-means* dengan metode kluster hirarki untuk mengatasi kelemahan *k-means* dalam menentukan pusat kluster awal. Arai & Barakbah (2007) mengusulkan pendekatan baru untuk mengoptimalkan titik pusat awal untuk *k-means*. Pendekatan ini memanfaatkan semua hasil kluster *k-means* dalam waktu tertentu, meskipun beberapa di antaranya mencapai titik optimal lokal. Kemudian, kita transformasikan hasilnya dengan menggabungkan kepada algoritma hirarki untuk menentukan titik pusat awal kluster dalam metode *k-means*. Di samping itu, Cheng dkk. (2012) mengusulkan algoritma *k-means* dua tahap yang disempurnakan yang pertama-tama memilih sejumlah besar *centroid* awal, menerapkan algoritma *k-means* dasar untuk mendapatkan kluster antara, dan kemudian menggabungkan kluster antara tersebut ke dalam k kluster akhir menggunakan pengelompokan hirarki aglomeratif, dalam rangka mengatasi kelemahan bahwa algoritma *k-means* standar sensitif terhadap pemilihan *centroid* awal.

Klasterisasi dengan penggabungan dua metode dari dua pendekatan klasterisasi yang berbeda sudah dilakukan oleh beberapa peneliti sebelumnya. Sintiya dkk. (2021) melakukan klasterisasi terhadap wilayah desa di Kabupaten Pematang berdasarkan angka kemiskinan dengan menggabungkan *k-means* dengan metode klasterisasi hirarki yaitu metode *single linkage*. Klasterisasi dengan menggabungkan kedua metode

tersebut terbukti efektif karena dapat mengatasi kelemahan *k-means* dalam menentukan pusat awal klaster dan jumlah klaster serta menghasilkan empat klaster optimal untuk data yang diteliti. Selain itu, Turnip (2023) melakukan klasterisasi terhadap provinsi di Indonesia berdasarkan Indeks Pembangunan Kebudayaan (IPK) dengan menggabungkan *k-means* dengan metode klasterisasi hirarki yaitu metode ward. Klasterisasi dengan menggabungkan kedua metode tersebut terbukti efektif karena dapat mengatasi kelemahan *k-means* dalam menentukan pusat awal klaster dan jumlah klaster serta menghasilkan dua klaster optimal untuk data yang diteliti.

Selain menggabungkan *k-means* dengan metode hirarki baik itu metode *single linkage* maupun ward, *k-means* dapat digabungkan dengan metode klasterisasi hirarki yang lain. Pemilihan metode klasterisasi, baik *k-means* maupun metode hirarki, harus mempertimbangkan tipe data, distribusi, dan ukuran dataset. Mereka membahas bagaimana beberapa metode hirarki lebih cocok untuk data dengan karakteristik tertentu, seperti data spasial atau data berbasis fitur yang sangat beragam (Xu & Wunsch, 2005). Pada penelitian ini, akan dilakukan klasterisasi yang menggabungkan *k-means* dengan metode *centroid linkage*.

Kombinasi metode *k-means* dan metode *centroid linkage* cocok digunakan karena keduanya beroperasi dengan prinsip yang serupa dalam hal klasterisasi data berdasarkan jarak ke pusat (centroid). Metode *centroid linkage* menyajikan beberapa keuntungan yang membuatnya menjadi pilihan menarik dibandingkan dengan metode klasterisasi lainnya. Salah satu kekuatan utamanya adalah beban komputasi yang relatif lebih rendah, karena tidak memerlukan konstruksi matriks jarak yang besar, sehingga cocok untuk komputasi penelitian dasar (Kellom & Raymond, 2017). Dalam hal akurasi, *centroid linkage* mengungguli klasterisasi *k-means*, dengan penelitian yang menunjukkan akurasi 87% untuk *centroid linkage* dibandingkan 81% untuk *k-means* (Liantoni & Cahyani, 2017). Kemajuan terbaru telah meningkatkan efisiensinya lebih jauh, dengan algoritma baru yang mencapai kecepatan hingga 36x lipat dengan tetap

mempertahankan kualitas pengelompokan (Bateni dkk., 2024). Namun, batasan yang perlu diperhatikan adalah bahwa hasil *centroid linkage* dapat bergantung pada urutan input, meskipun masalah ini dapat diatasi dengan menggabungkan hasil dari beberapa iterasi (Kellom & Raymond, 2017). Meskipun algoritma berbasis tautan sering kali memiliki sifat teoritis formal yang lebih kuat, metode berbasis pusat seperti *k-means* lebih sering digunakan dalam praktiknya. Untuk menjembatani kesenjangan ini, kerangka kerja teoretis baru telah dikembangkan untuk memberikan wawasan tentang kondisi di mana paradigma pengelompokan yang berbeda harus diterapkan (Ackerman dkk., 2021). Kemajuan ini menyoroti keseimbangan yang terus berkembang antara ketelitian teoritis dan kegunaan praktis dalam metodologi pengelompokan.

Data yang sesuai untuk dikelompokkan menggunakan metode klasterisasi yang menggabungkan *k-means* dan metode *centroid linkage* adalah data dengan karakteristik numerik karena bekerja dengan menghitung *centroid* dari kelompok data (Han dkk., 2011). Selain itu, data yang digunakan memiliki jumlah observasi yang lebih kecil karena kompleksitas komputasionalnya meningkat secara eksponensial dengan bertambahnya jumlah data (Bishop, 2006). *Centroid linkage* adalah metode klasterisasi hirarki yang dapat digunakan untuk membentuk klaster awal dengan menghitung jarak rata-rata antara semua pasangan titik data dalam dua klaster. Metode ini membantu dalam mengidentifikasi klaster yang terpisah dengan baik pada awalnya, yang dapat bermanfaat untuk penyempurnaan selanjutnya menggunakan *k-means* (Arai & Ridho Barakbah, 2007).

Data yang memenuhi karakteristik tersebut adalah data jumlah tenaga kesehatan tahun 2023 yang diakses melalui (Badan Pusat Statistik Indonesia, 2025). Data ini mencakup berbagai jenis tenaga kesehatan, termasuk perawat, bidan, tenaga kefarmasian, tenaga kesehatan masyarakat, tenaga kesehatan lingkungan, serta tenaga gizi. Selain itu, terdapat kategori tenaga medis seperti dokter dan dokter spesialis, serta tenaga kesehatan lainnya seperti tenaga psikologi klinis, keterampilan fisik, keteknisan

medis, teknik biomedika, dan tenaga kesehatan tradisional. Data ini menampilkan jumlah tenaga kesehatan berdasarkan provinsi dalam berbagai jenis tenaga kesehatan.

Indonesia menghadapi tantangan yang signifikan dalam hal ketersediaan tenaga kesehatan. Meskipun memiliki lebih dari satu juta tenaga kesehatan pada tahun 2020, masih terdapat kekurangan tenaga kesehatan, terutama di daerah pedesaan dan terpencil (Sukmawarni dkk., 2022). Distribusi yang tidak merata ini dipengaruhi oleh beberapa faktor seperti konsentrasi penduduk, kondisi geografis, dan perbedaan gaji (Purwaningsih, 2023). Indonesia bagian timur, khususnya provinsi seperti Nusa Tenggara Timur, Maluku, dan Papua Barat, mengalami kekurangan dokter, bidan, dan tenaga kesehatan masyarakat yang cukup parah (Hikmah dkk., 2020). Untuk mengatasi masalah ini, berbagai solusi telah diusulkan, termasuk perencanaan tenaga kerja yang lebih baik berdasarkan kompetensi dan beban kerja, memprioritaskan tenaga kesehatan lokal, dan menawarkan insentif yang lebih tinggi untuk penempatan yang menantang (Hidayanti, 2019).

Studi terbaru telah mengeksplorasi teknik pengelompokan untuk menganalisis distribusi tenaga kesehatan dan kebutuhan layanan kesehatan di seluruh Indonesia. Metode Ward digunakan untuk mengelompokkan kabupaten di Kalimantan Barat berdasarkan rasio tenaga kesehatan, yang menunjukkan adanya kesenjangan antar wilayah dan korelasi antara distribusi tenaga kesehatan dan angka harapan hidup (Saraswi dkk., 2024). Pengelompokan *K-Means* mengidentifikasi daerah dengan tenaga kesehatan yang tidak mencukupi di Kabupaten Karawang (Sitinjak dkk., 2022) dan memetakan kebutuhan layanan kesehatan yang tidak terpenuhi di seluruh provinsi di Indonesia, menyoroti kebutuhan yang tidak terpenuhi yang lebih tinggi di luar Pulau Jawa (Kusmanto dkk., 2023). Studi lain menerapkan *K-Means* untuk mengelompokkan kabupaten di Jawa Tengah dan Yogyakarta berdasarkan indikator kematian, yang memberikan wawasan tentang profil kesehatan daerah (Atthina & Iswari, 2014). Pendekatan-pendekatan pengelompokan ini menawarkan alat yang berharga bagi para

pembuat kebijakan untuk mengidentifikasi area-area prioritas untuk intervensi layanan kesehatan, meningkatkan alokasi sumber daya, dan mengatasi kesenjangan akses dan kualitas layanan kesehatan di berbagai daerah di Indonesia.

Klasterisasi data jumlah tenaga kesehatan berdasarkan provinsi bertujuan untuk mengelompokkan wilayah dengan karakteristik tenaga kesehatan yang serupa, sehingga dapat digunakan sebagai dasar dalam perencanaan dan pengambilan kebijakan di sektor kesehatan. Dengan teknik ini, pemerintah dan pemangku kepentingan dapat mengidentifikasi daerah yang memiliki kekurangan tenaga kesehatan serta menentukan strategi distribusi yang lebih merata. Selain itu, klasterisasi dapat membantu dalam analisis tren tenaga kesehatan, seperti perbedaan jumlah tenaga medis dan nonmedis di berbagai wilayah, serta hubungan antara jumlah tenaga kesehatan dengan indikator kesehatan masyarakat. Menurut Badan Pusat Statistik (BPS), analisis berbasis data sangat penting dalam meningkatkan efektivitas kebijakan publik, termasuk dalam sektor kesehatan (BPS, 2023). Dengan demikian, hasil klasterisasi ini dapat berkontribusi dalam meningkatkan akses dan kualitas layanan kesehatan bagi masyarakat secara lebih adil dan merata.

1.2. Rumusan Masalah

Permasalahan yang akan dibahas pada penelitian ini adalah bagaimana mengklasterisasi provinsi di Indonesia berdasarkan jumlah tenaga kesehatan tahun 2023 menggunakan metode *k-means* dengan *centroid linkage*?

1.3. Tujuan Penelitian

Berdasarkan rumusan masalah di atas, tujuan dari penelitian ini yaitu mengembangkan dan menerapkan metode *k-means* dengan *centroid linkage* untuk mengelompokkan provinsi di Indonesia berdasarkan jumlah tenaga kesehatan tahun 2023.

1.4. Manfaat Penelitian

Penelitian ini dapat memberikan pengetahuan yang lebih terkait penggabungan *k-means* dan metode *centroid linkage*, serta dapat menumbuhkan motivasi pembaca serta peneliti lain untuk melakukan penelitian dengan metode yang lain.