STUDI KLASTERISASI METODE *K-MEANS* DAN *X-MEANS* PADA PELAYANAN SAMSAT BAPENDA PROVINSI LAMPUNG

Skripsi

Oleh

ILHAM
NPM. 2117031054

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2025

ABSTRACT

CLUSTERING STUDY OF K-MEANS AND X-MEANS METHODS ON SAMSAT BAPENDA SERVICES IN LAMPUNG PROVINCE

By

Ilham

This study discusses the analysis of public satisfaction with public services at the SAMSAT BAPENDA of Lampung Province using the K-Means and X Means clustering methods. The data used are secondary data derived from the 2024 Community Satisfaction Index (IKM) involving 1,610 respondents across 17 service units. Each respondent provided assessments of ten service indicators on a scale of 1–4.The research stages include data preprocessing, normalization, and the application of clustering algorithms. Evaluation was carried out using the Davies-Bouldin Index (DBI) to assess cluster quality. In addition, an analysis of service indicators (categories 1 and 2) was conducted to identify aspects of service that still need improvement.

The results show that the K-Means and X-Means methods are capable of clustering public satisfaction data, with X-Means being more adaptive in automatically determining the optimal number of clusters. The indicators of safety and comfort receive the highest number of low ratings and therefore become the main priorities for improvement, followed by service timeliness, the quality of facilities and infrastructure, and staff competence. In contrast, the indicator of compliance with requirements has the fewest low ratings and is considered to have been implemented well. Overall, this study provides a mapping of SAMSAT service quality based on IKM data along with recommendations for service improvement.

Keywords: Clustering, K-Means, X-Means, Community Satisfaction Index, public service.

ABSTRAK

STUDI KLASTERISASI METODE *K-MEANS* DAN *X-MEANS* PADA PELAYANAN SAMSAT BAPENDA PROVINSI LAMPUNG

Oleh

Ilham

Penelitian ini membahas analisis kepuasan masyarakat terhadap pelayanan publik pada SAMSAT BAPENDA Provinsi Lampung dengan menggunakan metode klasterisasi k-means dan x-means. Data yang digunakan merupakan data sekunder berupa Indeks Kepuasan Masyarakat (IKM) tahun 2024 dengan jumlah responden sebanyak 1.610 orang yang tersebar di 17 unit pelayanan. Setiap responden memberikan penilaian terhadap sepuluh indikator pelayanan dengan skala 1–4. Tahapan penelitian meliputi pra pemrosesan data, normalisasi, dan penerapan algoritma klasterisasi. Evaluasi dilakukan dengan Davies-Bouldin Index (DBI) untuk menilai kualitas klaster. Selain itu, dilakukan analisis indikator pelayanan (kategori 1 dan 2) pada setiap indikator untuk mengidentifikasi aspek pelayanan yang masih perlu ditingkatkan.

Hasil penelitian menunjukkan bahwa metode *k-means* dan *x-means* mampu mengelompokkan data kepuasan masyarakat, dengan *x-means* lebih adaptif dalam menentukan jumlah klaster optimal secara otomatis. Indikator keamanan dan kenyamanan memperoleh penilaian rendah terbanyak sehingga menjadi prioritas perbaikan, disusul ketepatan waktu pelayanan, kualitas sarana dan prasarana, serta kompetensi petugas. Sebaliknya, indikator kesesuaian persyaratan memiliki penilaian rendah paling sedikit dan dinilai telah berjalan dengan baik. Penelitian ini menghasilkan pemetaan kualitas pelayanan SAMSAT berbasis data IKM serta rekomendasi perbaikan layanan.

Kata kunci: Klasterisasi, *K-Means*, *X-Means*, Indeks Kepuasan Masyarakat, pelayanan publik.

STUDI KLASTERISASI METODE *K-MEANS* DAN *X-MEANS* PADA PELAYANAN SAMSAT BAPENDA PROVINSI LAMPUNG

ILHAM

Skripsi

Sebagai Salah Satu Syarat untuk Memperoleh Gelar
SARJANA MATEMATIKA

Pada

Jurusan Matematika

Fakultas Matematika dan Ilmu Pengetahuan Alam

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2025

Judul Skripsi	:	**STUDI KLASTERISASI METODE *K-MEANS* DAN *X-MEANS* PADA PELAYANAN SAMSAT BAPENDA PROVINSI LAMPUNG**
Nama Mahasiswa	:	Ilham
Nomor Pokok Mahasiswa	:	**2117031054**
Program Studi	:	**Matematika**
Fakultas	:	**Matematika dan Ilmu Pengetahuan Alam**

MENYETUJUI

1. Komisi Pembimbing

Dr. Subian Saidi, S.Si., M.Si.	**Misgiyati, S.Pd., M.Si.**
NIP 198008212008121001	NIP 198509282023212032

2. Ketua Jurusan Matematika

Dr.Aang Nuryaman, S.Si.,M.Si.
NIP. 197403162005011001

MENGESAHKAN

1. Tim Penguji

Ketua : **Dr. Subian Saidi, S.Si., M.Si.** .....

Sekretaris : **Misgiyati, S.Pd., M.Si.** .....

Penguji
Bukan Pembimbing : **Drs. Nusyirwan, M.Si.** .....

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Dr. Eng. Heri Satria, S.Si., M.Si.
NIP. 197110012005011002

Tanggal Lulus Ujian Skripsi: **16 oktober 2025**

PERNYATAAN SKRIPSI MAHASISWA

Yang bertanda tangan di bawah ini:

Nama	:	**Ilham**
Nomor Pokok Mahasiswa	:	**2117031054**
Jurusan	:	**Matematika**
Judul Skripsi	:	**Studi Klasterisasi Metode *K-Means* dan *X-Means* pada Pelayanan SAMSAT BAPENDA Provinsi Lampung**

Dengan ini menyatakan bahwa skripsi ini adalah hasil pekerjaan saya sendiri. Apabila kemudian hari terbukti bahwa skripsi ini merupakan hasil salinan atau dibuat oleh orang lain, maka saya bersedia menerima sanksi sesuai dengan ketentuan akademik yang berlaku.

Bandar Lampung, 16 Oktober 2025

Penulis,

Ilham

RIWAYAT HIDUP

Penulis memiliki nama Ilham yang lahir di Bagansiapiapi pada tanggal 03 Januari 2003, anak dari Bapak A.Rahman.S dan IbuNur'aini. Penulis tumbuh dan besar dalam keluarga sederhana yang selalu menanamkan nilai kejujuran, kedisiplinan, serta pentingnya pendidikan. Hal tersebut membentuk penulis menjadi pribadi yang mandiri, tekun, dan bersemangat dalam menempuh pendidikan.

Pendidikan dasar ditempuh penulis di SDN 026 Bagan Punak Pesisir hingga lulus pada tahun 2015. Pendidikan menengah pertama dilanjutkan di Madrasah Tsanawiyah Hubbul Wathan dan diselesaikan pada tahun 2018 . Selanjutnya, penulis melanjutkan ke SMAN 2 Bangko hingga lulus pada tahun 2021.

Pada tahun 2021, penulis diterima sebagai mahasiswa di Universitas Lampung, Fakultas Matematika dan Ilmu Pengetahuan Alam, Jurusan Matematika. Selama menempuh pendidikan tinggi, penulis aktif mengembangkan diri tidak hanya dalam bidang akademik, tetapi juga dalam bidang organisasi dan kegiatan kemahasiswaan.

Perjalanan organisasi penulis dimulai pada tahun 2021–2022 sebagai anggota Unit Kegiatan Mahasiswa Bidang Seni Universitas Lampung. Pada tahun 2022, penulis bergabung sebagai anggota kaderisasi Himpunan Mahasiswa Jurusan Matematika (HIMATIKA) serta dipercaya sebagai staff ahli Badan Eksekutif Mahasiswa (BEM) FMIPA Universitas Lampung. Pada tahun 2023, penulis mendapatkan amanah lebih besar sebagai Wakil Ketua Umum HIMATIKA. Tahun berikutnya, yaitu 2024, penulis dipercaya menjabat sebagai Kepala Dinas Pengembangan Sumber Daya Manusia BEM FMIPA Universitas Lampung. Puncaknya, pada tahun 2025, penulis terpilih sebagai Menteri Sosial Masyarakat BEM U KBM Universitas Lampung. Seluruh pengalaman organisasi tersebut memberikan pelajaran berharga mengenai kepemimpinan, kerja sama, komunikasi, serta pengabdian kepada mahasiswa dan masyarakat.

Dalam bidang akademik, penulis memiliki minat pada analisis data dan statistika. Hal ini diwujudkan melalui penyusunan skripsi dengan judul “Studi Klasterisasi Metode K-means dan X-means pada Pelayanan SAMSAT BAPENDA Provinsi Lampung”. Penelitian ini merupakan bentuk implementasi ilmu yang diperoleh selama perkuliahan untuk memberikan kontribusi terhadap analisis kualitas pelayanan publik.

KATA INSPIRASI

”Perjalanan ini adalah bukti bahwa mimpi membutuhkan keberanian untuk meninggalkan zona nyaman.”

PERSEMBAHAN

Dengan mengucap Alhamdulillah dan syukur kepada Allah SWT atas nikmat serta hidayah-Nya sehingga skripsi ini dapat terselesaikan dengan baik dan pada waktu yang tepat. Dengan rasa syukur dan Bahagia, saya persembahkan rasa terimakasih saya kepada:

Ayah dan Ibuku Tercinta

Ucapan terima kasih yang tak terhingga penulis sampaikan kepada kedua orang tua tercinta atas kasih sayang, motivasi, dukungan, serta doa yang tidak pernah putus, sehingga penulis dapat melalui berbagai tantangan selama masa perkuliahan. Setiap pelajaran yang diberikan menjadi penguat langkah bagi penulis hingga skripsi ini dapat diselesaikan dengan baik.

Dosen Pembimbing dan Pembahas

Terimakasih kepada dosen pembimbing dan pembahas yang sudah sangat membantu, memberikan motivasi, memberikan arahan serta ilmu yang berharga.

Sahabat-sahabatku

Terimakasih kepada semua orang-orang baik yang telah memberikan pengalaman, semangat, motivasinya, serta doa-doanya dan senantiasa memberikan dukungan dalam hal apapun.

Almamater Tercinta
Universitas Lampung

SANWACANA

Alhamdulillah, puji dan syukur penulis panjatkan kepada Allah SWT atas limpahan rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi ini yang berjudul ”Studi Klasterisasi Metode *K-Means* dan *X-Means* pada Pelayanan SAMSAT BAPENDA Provinsi Lampung” dengan baik dan pada waktu yang tepat. Shalawat serta salam semoga senantiasa tercurahkan kepada Nabi Muhammad SAW.

Dalam proses penyusunan skripsi ini, banyak pihak yang telah membantu memberikan bimbingan, dukungan, arahan, motivasi serta saran sehingga skripsi ini dapat terselesaikan. Oleh karena itu, dalam kesempatan ini penulis mengucapkan terimakasih kepada:

1. Ayah, Ibu dan keluarga tercinta atas segala doa, kasih sayang, dukungan moral maupun materiil, serta kesabaran yang tiada henti dalam setiap langkah perjalanan hingga terselesaikannya skripsi ini.
2. Bapak Dr. Subian Saidi, S.Si., M.Si. selaku Pembimbing I yang telah banyak meluangkan waktunya untuk memberikan arahan, bimbingan, motivasi, saran serta dukungan kepada penulis sehingga dapat menyelesaikan skripsi ini.
3. Ibu Misgiyati, S.Pd., M.Si. selaku pembimbing II yang telah memberikan arahan, saran, motivasi, serta dukungan kepada penulis sehingga dapat menyelesaikan skripsi ini
4. Bapak Drs. Nusyirwan, M.Si. selaku Penguji yang telah bersedia memberikan kritik dan saran yang membangun serta evaluasi kepada penulis sehingga dapat menjadi lebih baik.
5. Bapak Prof. Dr. La Zakaria, S.SI.,M.SC. selaku dosen pembimbing akademik.
6. Bapak Dr. Ahmad Faisol, S.Si., M.Sc. selaku Sekretaris Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.

7. Bapak Dr. Aang Nuryaman, S.Si., M.Si. selaku Ketua Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
8. Ibu Anita selaku admin Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung yang telah membantu penulis dalam mempersiapkan berkas seminar maupun wisuda.
9. Seluruh dosen Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung yang tidak dapat penulis uraikan satu persatu karena telah memberikan ilmu serta masukan kepada penulis selama masa perkuliahan.
10. Seluruh karyawan dan staff jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
11. Seluruh keluarga besar yang selalu memberikan doa, motivasi, dan semangat selama masa perkuliahan.
12. Rista Ariyani atas kebersamaan, dukungan, dan doa yang telah diberikan selama beberapa tahun perjalanan. Apa yang pernah kita perjuangkan bersama adalah pelajaran berharga yang akan selalu penulis kenang sebagai bagian dari perjalanan menuju penyelesaian studi ini.
13. Riski, Rizki, Eky, Dinda, Yomel, Nurul dan Fina yang selalu menjadi tempat berbagi cerita, canda, dan motivasi. Kehadiran kalian telah memberikan semangat tersendiri dalam menjalani proses perkuliahan hingga penyusunan skripsi ini.
14. Yudhi, Yusuf, Akbar dan Zul yang telah menjadi keluarga kedua selama menempuh studi. Kebersamaan, dukungan, serta suasana hangat yang tercipta di lingkungan kos telah memberikan banyak kenangan berharga dan semangat dalam menyelesaikan perjalanan akademik ini.
15. Aisyah yang hadir memberikan semangat, cerita, dan keceriaan selama proses penyusunan skripsi ini
16. Semua pihak yang telah memberikan dukungan dalam berbagai bentuk selama proses penyusunan skripsi ini, namun tidak dapat disebutkan satu per satu.
17. Teman-teman organisasi yang telah menjadi bagian penting dalam perjalanan perkuliahan. Kebersamaan, kerja sama, dan pengalaman berorganisasi bersama kalian tidak hanya memperkaya wawasan, tetapi juga mengajarkan arti

tanggung jawab, solidaritas, dan kepemimpinan. Semua pengalaman berharga tersebut menjadi bekal yang sangat berarti dalam menyelesaikan studi in

18. Almamater Universitas Lampung yang telah menjadi tempat untuk belajar, berkembang, mencari pengalaman dan menimba ilmu selama masa perkuliahan.

Semoga skripsi ini dapat bermanfaat bagi kita semua. Penulis menyadari bahwa skripsi ini masih jauh dari sempurna, sehingga penulis mengharapkan kritik dan saran yang membangun untuk menjadikan skripsi ini lebih baik lagi.

Bandar Lampung, 16 Oktober 2025

Ilham

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I

PENDAHULUAN

1.1 Latar Belakang & Masalah

Analisis klasterisasi merupakan salah satu metode yang bertujuan untuk mengelompokkan data atau objek berdasarkan kemiripan dan ketidakmiripan karakteristiknya sehingga objek yang berada pada satu klaster memiliki kemiripan yang besar dan sangat kecil bila dibandingkan dengan klaster lain. Dalam klasterisasi, tidak ada label atau kategori yang telah ditentukan sebelumnya, sehingga algoritma harus menemukan pola sendiri dari data mentah (Yusuf, *et al*., 2022).

Berkembangnya metode klaster hingga saat ini disebabkan oleh banyaknya bidang kehidupan yang membutuhkan analisis klaster untuk pengelompokan objek. Salah satu bidang yang sering membutuhkan hal tersebut adalah bidang pelayanan yang dikelompokkan berdasarkan indeks kepuasan masyarakat untuk melihat daerah mana saja yang membentuk kelompok dengan tingkat pelayanan yang sudah baik dan kurang baik. Informasi ini dapat menjadi dasar dalam mengambil kebijakan prioritas.

Survei yang dilakukan tahun 2022 oleh Edelmen Trust Barometer di 28 negara dengan 1.150 responden pada masing-masing negara, menunjukkan bahwa tingkat kepercayaan masyarakat terhadap institusi pemerintah lebih rendah dibandingkan dengan tingkat kepercayaan terhadap bisnis. Salah satunya Indonesia yang masyarakatnya 83% lebih percaya terhadap institusi bisnis (Javier, 2023).

Kualitas kepercayaan masyarakat terhadap institusi pemerintah tergantung pada pelayanan yang diberikan serta menjadi tolak ukur untuk melihat tingkat kepuasan

masyarakat terhadap kinerja suatu institusi. Pelayanan publik yang efektif dan efisien adalah elemen penting dalam meningkatkan kualitas hidup masyarakat serta mendorong kepuasan pengguna layanan.

Tahun 2021 Populi Center melakukan survei mengenai permasalahan pelayanan yang terjadi di Indonesia. Hasil survei tersebut menunjukkan 11.4% masyarakat mengeluhkan persyaratan yang berbelit, 8,6% mengeluhkan sarana dan prasarana yang tidak memadai, dan 2,7% menyatakan pelayanan yang kurang ramah (Annur, 2021).

Kantor Sistem Administrasi Manunggal Satu Atap (SAMSAT) merupakan salah satu layanan publik yang memiliki peran vital dalam administrasi perpajakan kendaraan bermotor, pengesahan STNK, dan penerbitan BPKB. Kinerja layanan SAMSAT sangat berpengaruh terhadap kepuasan dan kepercayaan masyarakat, terutama dalam hal kecepatan, ketepatan, dan kualitas pelayanan yang diberikan.

Untuk memahami karakteristik pengguna layanan dan mengidentifikasi faktor-faktor yang memengaruhi tingkat kepuasan masyarakat terhadap pelayanan SAMSAT, diperlukan analisis yang dapat mengelompokkan data berdasarkan kemiripan tertentu. Klasterisasi adalah salah satu metode yang dapat mengelompokkan data dengan tujuan untuk menemukan pola atau kelompok di antara data yang kompleks. Metode *k-means* dan *x-means*, sebagai algoritma klasterisasi, mampu memberikan pengelompokan data berdasarkan karakteristik yang relevan dengan kepuasan atau kinerja pelayanan.

Penelitian sebelumnya dengan metode klaster pernah dilakukan untuk membandingkan hasil pengelompokan menggunakan klaster berhirarki, klasterisasi *k-Means*, dan klasterisasi *ensamble* dengan studi kasus data indikator pelayanan kesehatan ibu hamil. Hasil yang diperoleh dari penelitian tersebut adalah dari beberapa metode yang diterapkan metode klasterisasi *ensamble* merupakan metode yang lebih tepat dalam mengelompokkan karakteristik pelayanan kesehatan ibu hamil (Suhaeni, *et al*., 2018).

Selain itu analisis klaster kepuasan pengguna terhadap layanan Shopee menggunakan algoritma *k-means* juga pernah dilakukan. Dari hasil penelitian tersebut diperoleh kesimpulan bahwa algoritma *k-means* dengan nilai k sama dengan 2 sampai 5, di mana data yang digunakan berasal dari kuisioner yang memiliki variabel data pribadi, reliabilitas (VI), *information quality*(V2), *web design and layout* (V3), *interaction quality*(V4), efisien(V5), dan kontak (V6)

menghasilkan nilai k optimal yang diperoleh adalah 2, dengan nilai *DBI* sebesar 1.5876178. Berdasarkan penelitian yang dilakukan perlu pengembangan lebih lanjut untuk mengoptimasi pada algoritma *k-means* (Patimah, *et al*., 2021).

Pada penelitian lainnya tentang analisis efektifitas pelayanan publik menggunakan klasterisasi *k-means* di Kecamatan Sukagumiwang didapatkan kesimpulan nilai k yang optimal untuk data survei kepuasan masyarakat menggunakan algorima *k-means* adalah 15 dengan nilai *DBI* 0,984 (Sofiyah, *et al*., 2023).

Yenik *et al*. (2025), melakukan penelitian tentang analisis kepuasan masyarakat terhadap pelayanan publik menggunakan klasterisasi *k-means* yang bertujuan untuk memahami pola tertentu yang terkandung dalam data kepuasan masyarakat atas layanan publik. Penelitian tersebut berhasil mengidentifikasi dua klaster yang memiliki karakteristik yang berbeda dalam menilai kualitas pelayanan. Hal ini dapat menjadi dasar bagi pihak pengelola layanan untuk melakukan evaluasi terhadap aspek-aspek yang masih kurang. Langkah ini dapat meningkatkan kualitas pelayanan secara keseluruhan dan memberi kepuasan yang lebih baik (Hariyanto, *et al*., 2025).

Sebelumnya, penelitian klasterisasi telah diterapkan dalam berbagai bidang, termasuk sektor kesehatan dan *e-commerce*. Namun, penelitian terkait analisis kepuasan masyarakat terhadap pelayanan SAMSAT menggunakan metode klasterisasi masih terbatas, khususnya di Provinsi Lampung. Dalam penelitian ini, metode *k-means* dan *x-means* akan digunakan untuk mengelompokkan tingkat kepuasan masyarakat berdasarkan data IKM SAMSAT BAPENDA Lampung. *K-means* merupakan metode klasterisasi yang membutuhkan jumlah klaster yang ditentukan terlebih dahulu, sedangkan *x-means* dapat secara otomatis menentukan jumlah klaster optimal berdasarkan *Bayesian Information Criterion* (BIC).

Penelitian ini diharapkan dapat memberikan pemetaan yang lebih jelas mengenai kelompok masyarakat dengan tingkat kepuasan berbeda. Dengan demikian, hasil klasterisasi ini dapat menjadi dasar dalam pengambilan keputusan untuk meningkatkan kualitas pelayanan SAMSAT di masa mendatang.

1.2 Tujuan Penelitian

1. Menganalisis kinerja pelayanan SAMSAT Provinsi Lampung berdasarkan data Indeks Kepuasan Masyarakat (IKM).
2. Menganalisis performa klasterisasi metode *k-means* dan *x-means*.

1.3 Manfaat Penelitian

1. Membantu SAMSAT dalam memahami area pelayanan yang perlu ditingkatkan.
2. Memilih metode klasterisasi yang lebih efektif dan akurat untuk analisis data yang besar dan kompleks.
3. Merumuskan kebijakan yang lebih tepat sasaran dalam memperbaiki layanan publik sesuai dengan kebutuhan masyarakat.
4. Berkontribusi pada pengembangan penerapan metode klasterisasi dalam bidang pelayanan publik.

BAB II

TINJAUAN PUSTAKA

2.1 *Knowledge Discovery in Database* dan Data *Mining*

Knowledge Discovery in Database (KDD) adalah salah satu metodologi yang digunakan untuk menganalisis dan memahami data dalam jumlah besar. KDD merupakan bidang interdisipliner yang menggabungkan unsur kecerdasan buatan, basis data, statistika, dan pembelajaran mesin (*machine learning*). Gagasan mengenai KDD pertama kali muncul pada akhir tahun 1980-an. Menurut Fayyad dkk dalam makalah tahun 1996, KDD didefinisikan sebagai proses *non-trivial* untuk mengidentifikasi pola-pola yang valid, baru, berpotensi berguna, dan dapat dipahami dari suatu data. Definisi ini juga sering diterapkan dalam konteks data *mining*. Bahkan, dalam literatur-literatur terbaru, istilah *data mining* dan KDD kerap digunakan secara bergantian atau tanpa pembedaan yang jelas. Namun demikian, dalam metodologi klasik KDD, *data mining* sebenarnya merupakan salah satu tahapan dalam keseluruhan proses KDD. Tahapan ini mencakup proses ekstraksi pengetahuan dari data, yang diawali dengan pemilihan (*selection*) dan pra-pemrosesan (*preprocessing*) data dari berbagai sumber, serta diakhiri dengan interpretasi hasil data *mining* secara tepat dan bermakna.

Dunham (2002) meringkas proses KDD dari berbagai tahapan, yaitu: seleksi data, pra-proses data, transformasi data, data *mining*, interpretasi dan evaluasi. Berikut adalah ilustrasi serta penjelasan mengenai proses KDD.

1. Data *cleansing*, proses pemilihan

2. Data *integration*, proses penggabungan data.

3. *Selection*, proses seleksi atau pemilihan data yang relevan terhadap analisis.
4. Data *transformation*, proses transformasi data terpilih ke dalam bentuk *mining procedure*.
5. Data *mining*, proses di mana dilakukan beragam teknik untuk mengekstrak pola-pola potensial sehingga menghasilkan data yang berguna.
6. *Pattern evolution*, proses di mana pola-pola yang telah diidentifikasi berdasarkan *measure* yang diberikan.
7. *Knowledge evolution*, proses visualisasi data agar lebih mudah dipahami sebelum mengambil tindakan berdasarkan analisis.

Sedangkan data *mining* meliputi deskripsi konsep, asosiasi aturan, klasifikasi dan prediksi, analisis klaster, analisis deret waktu, *text mining*, dan sebagainya. Secara garis besar, data *mining* dapat dikelompokkan menjadi 2 kategori utama yaitu:

1. *Deskriptive mining*, yaitu proses untuk menemukan karakteristik penting dari data dalam satu baris data. Teknik data *mining* yang termasuk *descriptive mining* adalah klasterisasi, *assosiation*, dan *sequential mining*.
2. *Predictive*, yaitu proses untuk menemukan pola dari data dengan menggunakan beberapa variabel lain di masa depan. Salah satu teknik yang terdapat dalam *predictive mining* adalah klasifikasi.

Istilah KDD dan data *mining* seringkali digunakan secara bergantian untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu basis data yang besar. Sebenarnya kedua istilah tersebut memiliki konsep berbeda, tetapi berkaitan satu sama lain, dan salah satu tahapan dalam keseluruhan proses KDD adalah data *mining*.

Data *mining* bukanlah suatu bidang yang sama sekali baru. Salah satu kesulitan untuk mendefinisikan data *mining* adalah kenyataan bahwa data *mining* mewarisi banyak aspek dan teknik dari bidang-bidang ilmu yang dulu sudah mapan. Data *mining* memiliki akar yang panjang dari bidang ilmu yang berbeda seperti kecerdasan buatan (*artificial intelligent*), *machine learning*, *statistic*, *database*, dan *information retrieval*.

Cara kerja data *mining* adalah dengan mencari pola atau informasi penting dalam data yang sebelumnya tidak diketahui, atau dengan memprediksi kejadian yang

akan datang. Teknik yang digunakan untuk melaksanakan proses ini disebut pemodelan. Dalam praktiknya, data *mining* mengekstraksi informasi yang bernilai melalui analisis pola-pola atau hubungan tertentu dalam kumpulan data berukuran besar.

Data *mining* memiliki keterkaitan erat dengan berbagai bidang ilmu, seperti sistem basis data, data *warehousing*, statistika, *machine learning*, *information retrieval*, dan komputasi tingkat tinggi. Selain itu, data *mining* juga didukung oleh disiplin ilmu lainnya seperti *neural network*, pengenalan pola, analisis data spasial, basis data citra, dan *signal processing* (Siregar & Puspabhuana, 2017).

2.2 *Principal Component Analysis* (*PCA*)

Principal Component Analysis (PCA) PCA merupakan salah satu teknik analisis statistik yang paling umum digunakan untuk mereduksi dimensi data. Metode ini pertama kali diperkenalkan pada awal abad ke-20 dan terus berkembang seiring dengan meningkatnya kompleksitas data yang dihadapi oleh para peneliti dan praktisi di berbagai bidang. Pada tahun 1901, Karl Pearson pertama kali memperkenalkan konsep PCA sebagai metode untuk menemukan ”garis terbaik” yang meminimalkan kesalahan kuadrat dalam data multidimensi (Pearson, 1901). Pada tahun 1933, Harold Hotelling mengembangkan ide Pearson dengan mengusulkan PCA sebagai metode transformasi variabel berkorelasi menjadi sekumpulan variabel yang tidak berkorelasi (komponen utama), yang dikenal sebagai transformasi ortogonal. Kontribusi Hotelling adalah memformalkan dasar matematika PCA dan memperkenalkan konsep transformasi ortogonal.

Tahun 1967, Gene H. Golub dan William Kahan memperkenalkan metode komputasi efisien untuk PCA menggunakan *Singular Value Decomposition* (SVD). SVD memungkinkan perhitungan komponen utama yang lebih efisien, sehingga PCA menjadi lebih praktis untuk analisis data besar. Pada tahun 1986, Ian Jolliffe mempopulerkan penggunaan PCA dalam berbagai disiplin ilmu melalui bukunya ”*Principal Component Analysis*,” yang menjadi referensi penting bagi para praktisi. PCA adalah metode statistik yang umum digunakan untuk mereduksi dimensi data dan menangani kompleksitas data. Dalam analisis data multidimensi, PCA berperan penting dalam mengidentifikasi pola tersembunyi dan mengurangi *noise*. Metode ini juga membantu mencegah *overfitting* dalam model pembelajaran

mesin yang beroperasi pada data berdimensi tinggi (Shalih *et al*, 2025).

2.3 Korelasi *Spearman*

Spearman adalah uji statistik untuk mengetahui hubungan antara dua atau lebih variabel berskala ordinal. Koefisien korelasi *spearman* adalah statistik non-parametrik, karena data yang didapat tidak berdistribusi normal. Keuntungan dari korelasi *spearman* adalah analisis berbasis peringkat lebih mudah dihitung dari pada analisis numerik dan juga dapat digunakan untuk menentukan korelasi non linear. Metode korelasi ini menghitung peringkat korelasi mulai dari -1 yang berarti korelasi sempurna dalam derajat kemiringan negatif dan +1 dimana yang berarti korelasi sempurna dalam derajat kemiringan sempurna. Selain nilai -1 dan +1, nilai diantara kedua angka tersebut, jika angka di atas 0,5 atau di bawah 0,5 maka itu dinamakan dengan hubungan yang moderat atau cukup kuat. Untuk menganalisis korelasi *spearman* dalam statistik uji menggunakan rumus :

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)}$$

Dengan:
r_s = Nilai korelasi *spearman*,
d_i = Selisih antara peringkat X dan Y,
n = Jumlah pasangan data.

Prosedur uji hipotesis:

$$H_0 : X \text{ dan } Y \text{ saling bebas (tidak berkorelasi)}$$
$$H_1 : X \text{ dan } Y \text{ tidak saling bebas (berkorelasi)}$$

Kriteria penolakan H_0:

$$\text{Tolak } H_0 \quad \text{jika} \quad |r_s| > \alpha(2)$$

2.4 Klasterisasi

Sejarah algoritma klasterisasi pada awal abad ke-20, algorima klasterisasi berasal dari bidang antropologi dan psikologi yang diperkenalkan tahun 1932 untuk menyederhanakan ambiguitas tipologi budaya dan individu yang berbasis empiris. Pada bidang psikologi, algoritma ini digunakan oleh Cattell tahun 1943 untuk klasifikasi teori sifat dalam psikologi kepribadian, yang mengorganir dan menganalisis informasi bermakna dari data psikologi. Tahun 1960-an, klasterisasi semakin populer di berbagai bidang seperti ilmu sosial, kedokteran, biologi, dan geografi. Pada intinya klasterisasi adalah metode pembelajaran yang membantu dalam pengelompokan objek berdasarkan tingkat kesamaan di antara mereka. Seiring waktu, perkembangan algoritma ditandai dengan pengenalan metode baru untuk mengidentifikasi kesamaan yang diikuti oleh berbagai perbaikan dan adaptasi untuk menangani dataset yang lebih besar serta jenis data yang lebih kompleks.

Klasterisasi mirip dengan klasifikasi dalam hal data pengelompokan. Dalam klasterisasi pengelompokan dilakukan dengan menemukan kesamaan di antara data berdasarkan karakteristiknya. Kelompok-kelompok ini disebut klaster.

Beberapa penulis menganggap klasterisasi sebagai jenis khusus dari klasifikasi. Namun, menurut pandangan yang lebih konvensional keduanya berbeda. Sudah banyak definisi untuk klaster yang telah diajukan, diantaranya klaster adalah sekumpulan elemen yang serupa atau memiliki karakteristik yang sama. Jarak antara titik-titik dalam satu klaster lebih kecil dibandingkan jarak antara titik pada klaster tersebut dengan titik yang lain.

Istilah yang mirip dengan klasterisasi adalah segmentasi *database*, yaitu *tuple* (rekaman) serupa dalam *database* yang dikelompokkan bersama. Hal ini dilakukan untuk mempartisi atau membagi *database* menjadi komponen-komponen yang kemudian memberikan pandangan yang lebih umum kepada pengguna terhadap data tersebut (Dunham, 2006).

Saat klasterisasi diterapkan pada *database* dunia nyata, beberapa masalah menarik muncul, yaitu:

1. Penanganan *outlier*: elemen-elemen yang tidak secara alami termasuk dalam

klaster dapat dianggap sebagai klaster tersendiri. Namun, algoritma klasterisasi yang mencoba menemukan klaster yang lebih besar mungkin memaksa *outlier* tersebut masuk ke dalam salah satu klaster, yang dapat menghasilkan klaster yang buruk.

2. Data dinamis: keanggotaan klaster dapat berubah seiring waktu.
3. Interpretasi makna: dengan klasterisasi, label atau interpretasi setiap klaster tidak selalu jelas dan mungkin memerlukan ahli untuk memberikan makna atau label.
4. Tidak ada jawaban tunggal: tidak ada satu jawaban benar untuk masalah klasterisasi. Jumlah klaster yang diperlukan sering sulit ditentukan dan mungkin memerlukan keahlian domain.

Definisi klasterisasi: Diberikan sebuah *database* D yang terdiri atas sejumlah *tuple*, serta sebuah bilangan bulat positif k, permasalahan klasterisasi bertujuan untuk menentukan sebuah fungsi pemetaan:

$$f : D \to \{1, \ldots, k\} \tag{2.4.1}$$

yang mengalokasikan setiap *tuple* $t_i \in D$ ke salah satu dari k klaster yang tersedia.

Dalam konteks ini, *tuple* merujuk pada satuan data individu dalam *database* D. Sebagai contoh, jika D merupakan kumpulan data mengenai kinerja pelayanan SAMSAT, maka setiap *tuple* t_i dapat merepresentasikan satu unit pelayanan atau satu entri pengukuran seperti nilai Indeks Kepuasan Masyarakat (IKM), waktu pelayanan, jumlah pegawai, dan parameter lainnya. Dengan kata lain, *tuple* adalah baris data yang terdiri dari beberapa atribut atau variabel yang diamati.

Proses klasterisasi akan mengelompokkan seluruh *tuple* dalam D ke dalam k kelompok (klaster) berdasarkan tingkat kemiripan atau kedekatan antar *tuple*. Hasil dari pemetaan f adalah bahwa setiap *tuple* akan memiliki label klaster. Metode ini banyak digunakan dalam analisis data eksploratif, terutama ketika tidak terdapat label sebelumnya (*unsupervised learning*), dan bertujuan untuk menemukan pola atau struktur alami dalam data.

Secara umum metode klasterisasi dibagi menjadi dua yaitu klasterisasi hirarki dan klasterisasi partisi. Pada klasterisasi hirarki, data dikelompokkan melalui suatu bagan yang berupa hirarki, di mana terdapat penggabungan dua grup yang terdekat

disetiap iterasinya ataupun pembagian dari seluruh set data ke dalam klaster. Contoh metode *hierarchical* klasterisasi adalah *single linkage*, *complete linkage*, *average linkage*, *average group linkage*.

Sedangkan klasterisasi partisi, data dikelompokkan ke dalam sejumlah klaster tanpa adanya struktur hirarki antara satu dengan yang lainnya. Pada metode ini setiap klaster memiliki titik pusat klaster (*centroid*) dan secara umum metode ini memiliki fungsi tujuan yaitu meminimumkan jarak (*dissimilarity*) dari seluruh data ke pusat klaster masing-masing. Contoh metode klasterisasi partisi adalah *K-means*, *Fuzzy k-means* dan *Mixture modelling* (Irwansyah & Faisal 2012)

2.5 Klasterisasi Partisi

Klasterisasi partisi merupakan pendekatan yang membagi himpunan data ke dalam sejumlah klaster secara langsung dalam satu tahap proses. Berbeda dengan metode hierarki yang menyusun pohon klasterisasi secara bertahap (*bottom-up* atau *top-down*), metode partisi langsung menghasilkan satu himpunan klaster akhir berdasarkan jumlah klaster yang telah ditentukan sebelumnya.

Dalam pendekatan ini, hanya satu set klaster yang dihasilkan sebagai keluaran akhir, meskipun pada proses internal, algoritma tertentu mungkin melakukan inisialisasi atau pengulangan beberapa kali untuk mendapatkan hasil terbaik. Karena jumlah klaster tidak ditentukan secara otomatis, maka pengguna wajib memberikan parameter k, yaitu jumlah klaster yang diinginkan.

Untuk mengevaluasi kualitas hasil klasterisasi, digunakan suatu fungsi objektif atau metrik evaluasi yang disebut fungsi kriteria. Fungsi ini digunakan untuk mengukur seberapa baik pemisahan dan kekompakan klaster yang terbentuk. Salah satu metrik yang paling umum digunakan adalah *jumlah total kesalahan kuadrat* (*Total Within-Cluster Squared Error*), yang dihitung berdasarkan jarak antara setiap titik data dengan pusat klaster (*centroid*). Fungsi ini dirumuskan sebagai berikut:

$$\sum_{m=1}^{k} \sum_{t_{mi} \in K_m} \text{dis}(C_m, t_{mi})^2 \tag{2.5.2}$$

dengan:
k : jumlah klaster, ditentukan oleh pengguna
K_m : himpunan data (klaster) ke-m,
C_m : *centroid* (titik pusat) dari klaster K_m
t_{mi} : titik data ke-i dalam klaster K_m
dis(C_m, t_{mi}): jarak (biasanya Euclidean) antara titik data t_{mi} dan *centroid* C_m.

Fungsi objektif ini mengukur total penyimpangan data terhadap pusat klasternya. Semakin kecil nilai fungsi ini, semakin baik kualitas klasterisasi yang dihasilkan, karena data dalam klaster lebih kompak dan homogen.

Jika fungsi jarak menggunakan jarak *Euclidean*, maka $dis(C_m, t_{mi})$ dapat dituliskan sebagai:

$$dis(C_m, t_{mi}) = \|C_m - t_{mi}\|^2 = \sum_{j=1}^{n} (c_{mj} - t_{mij})^2 \tag{2.5.3}$$

dengan:
n : Dimensi data
c_{mj} : Koordinat *centroid* C_m pada dimensi j
t_{mij}: Koordinat titik data t_{mi} pada dimensi j.

Rumus ini digunakan dalam algoritma klasterisasi *k-Means* untuk meminimalkan jumlah total jarak kuadrat antara setiap titik data dengan *centroid*.

Masalah utama dalam algoritma klasterisasi partisi terletak pada meningkatnya jumlah kemungkinan pembentukan klaster seiring bertambahnya jumlah data yang diklasterkan. Dalam skenario umum, pencarian solusi optimal dengan mengevaluasi seluruh kemungkinan partisi data ke dalam klaster menjadi tidak praktis, terutama untuk jumlah data yang besar. Hal ini dikarenakan kompleksitas kombinatorial dari banyaknya cara pembagian elemen ke dalam klaster yang berbeda.

Salah satu pendekatan *brute-force* yang paling mendasar adalah dengan memeriksa semua kemungkinan partisi dari n elemen ke dalam k klaster. Jumlah total partisi tersebut diberikan oleh bilangan *stirling* kedua (*Stirling Number of the Second Kind*) yang dinotasikan dengan $S(n, k)$, dan dirumuskan sebagai berikut:

$$S(n, k) = \sum_{i=1}^{k} (-1)^{k-i} \binom{k}{i} i^n \tag{2.5.4}$$

Bilangan ini menunjukkan jumlah cara untuk membagi sebuah himpunan yang terdiri dari n elemen menjadi k himpunan tak kosong yang tidak berurutan.

Sebagai ilustrasi, untuk $n = 19$ dan $k = 4$, terdapat sebanyak 1.259.666.000 cara berbeda untuk membagi 19 elemen ke dalam 4 klaster. Angka ini menunjukkan bahwa eksplorasi menyeluruh terhadap seluruh ruang solusi tidak layak dilakukan secara komputasional.

Oleh karena itu, algoritma klasterisasi partisi seperti *k-Means* umumnya tidak mencari solusi global dengan mengevaluasi semua kemungkinan, melainkan menggunakan pendekatan heuristik atau optimisasi lokal untuk menemukan solusi yang baik secara efisien. Algoritma ini hanya mengevaluasi sebagian kecil dari semua kemungkinan klasterisasi, dengan harapan dapat memperoleh partisi yang representatif dan memenuhi kriteria kualitas tertentu seperti minimisasi fungsi objektif (Dunham, 2006).

2.6 Klasterisasi *K-Means*

Diberikan sekumpulan objek numerik X dan sebuah bilangan bulat k ($k \leq n$), algoritma *k-means* mencari partisi dari X menjadi k klaster yang meminimalkan jumlah kuadrat galat dalam kelompok. Proses ini sering diformulasikan persamaan matematis berikut:

$$\textbf{P(W,Q)} = \sum_{i=1}^{n} \sum_{l=1}^{k} w_{i,l}\, d(x_i, Q_l) \tag{2.6.5}$$

dengan syarat:

1. $\sum_{l=1}^{k} w_{i,l} = 1$, untuk $1 \leq i \leq n$
2. $w_{i,l} \in \{0, 1\}$, untuk $1 \leq i \leq n$ dan $1 \leq l \leq k$

di mana **W** adalah matriks partisi berukuran $n \times k$, $Q = \{Q_1, Q_2, \ldots, Q_k\}$ adalah sekumpulan objek dalam domain objek yang sama, dan $d(\cdot, \cdot)$ adalah jarak *Euclidean* kuadrat antara dua objek.

Sifat penting dari algoritma *k-means* adalah:

1. Efisien untuk mengolah data berukuran besar.

2. Sering berakhir pada optimum lokal.

3. Hanya bekerja pada nilai numerik.

4. Klaster yang dihasilkan berbentuk konveks.

Ada beberapa varian algoritma *k-means* yang berbeda dalam pemilihan nilai awal k, perhitungan *dissimilarity*, dan strategi untuk menghitung rata-rata klaster. Varian terkenal termasuk algoritma ISODATA dan algoritma *fuzzy k-means*.

K-Means merupakan salah satu metode klasterisasi partisi yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih klaster/kelompok. Metode ini mempartisi data sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu klaster yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokan ke klaster yang lain. Tujuannya adalah untuk meminimalisasikan *objective function* dalam proses klasterisasi dan berusaha meminimalisasikan variasi di dalam suatu klaster serta memaksimalisasikan variasi antar klaster (Huang, 1998)

Algoritma *k-means* adalah salah satu metode klasterisasi yang paling umum digunakan untuk mengelompokkan data yang memiliki kemiripan tertentu. *K-Means* secara iteratif mempartisi data ke dalam k kelompok yang saling *eksklusif*, di mana k adalah jumlah klaster yang ditentukan sebelumnya. Algoritma ini bekerja dengan mengoptimalkan posisi *centroid* atau pusat dari setiap klaster, sehingga meminimalkan jarak antara titik-titik data dan *centroid* klaster masing-masing.

Algoritma *k-Means* bertujuan untuk mengelompokkan m titik data $\{x_1, x_2, \ldots, x_m\}$ ke dalam K klaster sedemikian rupa sehingga jumlah total jarak kuadrat antara titik-titik data dan *centroid* klaster masing-masing menjadi minimal. Fungsi objektif dari algoritma ini dapat dituliskan sebagai berikut:

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} \omega_{ik} \|x_i - \mu_k\|^2 \tag{2.6.6}$$

dengan:
x_i adalah titik data ke-i dalam himpunan data D
μ_k adalah *centroid* dari klaster ke-k
ω_{ik} adalah variabel indikator yang menunjukkan apakah titik x_i termasuk ke dalam

klaster ke-k:

$$\omega_{ik} = \begin{cases} 1, & \text{jika } x_i \in \text{klaster } k \\ 0, & \text{jika tidak} \end{cases} \qquad i = 1, 2, \ldots, M, \quad k = 1, 2, \ldots, K$$

Masalah optimasi ini dapat dipandang sebagai proses minimisasi dua tahap (dua variabel bebas), yaitu meminimalkan J terhadap ω_{ik} (penugasan klaster) dan μ_k (penentuan *centroid*).

Pendekatan yang digunakan oleh *k-means* untuk menyelesaikan masalah disebut dengan *Expectation-Maximization* (*EM*).

Langkah pertama adalah *e-step* (*expectation step*), menentukan klaster untuk setiap titik data dengan mengasumsikan bahwa posisi *centroid* μ_k sudah tetap. Penugasan ini dilakukan dengan memilih klaster yang memiliki jarak *Euclidean* terkecil terhadap titik data x_i. Secara matematis, ini dapat dituliskan sebagai:

$$\omega_{ik} = \begin{cases} 1, & \text{jika } k = \arg\min_j \|x_i - \mu_j\|^2 \\ 0, & \text{lainnya} \end{cases} \tag{2.6.7}$$

Dengan kata lain, titik data x_i akan ditugaskan ke klaster yang *centroid*-nya paling dekat secara kuadrat.

Langkah berikutnya adalah *m-step* (*maximization step*), memperbarui posisi *centroid* setiap klaster dengan mengasumsikan penugasan klaster (ω_{ik}) telah tetap. Untuk meminimalkan fungsi objektif terhadap μ_k, kita turunkan J terhadap μ_k dan disamakan dengan nol:

$$\frac{\partial J}{\partial \mu_k} = 2 \sum_{i=1}^{m} \omega_{ik}(\mu_k - x_i) = 0 \tag{2.6.8}$$

Menyelesaikan persamaan tersebut menghasilkan rumus pembaruan *centroid*:

$$\mu_k = \frac{\sum_{i=1}^{m} \omega_{ik} x_i}{\sum_{i=1}^{m} \omega_{ik}} \tag{2.6.9}$$

Artinya, *centroid* μ_k merupakan rata-rata dari semua titik data yang termasuk dalam klaster k.

Kedua langkah di atas *e-step* dan *m-step* dilakukan secara bergantian dan iteratif hingga terjadi konvergensi, yaitu ketika penugasan klaster tidak berubah lagi atau perubahan nilai *centroid* sangat kecil. Proses ini menjamin bahwa nilai fungsi objektif J akan menurun di setiap iterasi, meskipun tidak selalu mencapai minimum global.

2.7 Metode *Elbow*

Metode *Elbow* merupakan salah satu metode yang dapat digunakan untuk menentukan jumlah klaster terbaik, yaitu dengan cara melihat persentase setiap klaster yang akan membentuk siku pada suatu titik tertentu. Metode *Elbow* biasa disajikan dalam bentuk grafik untuk mengetahui lebih jelas siku yang terbentuk. Tujuan dari metode *Elbow* adalah untuk memilih nilai k yang kecil dan masih memiliki nilai kuadrat yang rendah. Nilai k pada kombinasi siku dengan *k-means* adalah grafik hubungan klaster dengan penurunan *error*. Jumlah klaster k yang dihasilkan dari pengujian dengan *k-means* dievaluasi dengan teknik *Sum of Square Error* (*SSE*). SSE merupakan rumus yang digunakan untuk mengukur perbedaan antara data yang telah dilakukan sebelumnya. Persamaan yang digunakan dalam metode *Elbow* yaitu nilai total *Within Cluster Sum of Squares* atau biasa disebut *Sum Square Error* (Ekasetya & Jananto, 2020).

Rumus SSE adalah sebagai berikut:

$$\text{SSE} = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \phi_k\|^2 \tag{2.7.10}$$

dengan:
C_k: Klaster ke-k yang terbentuk
K: Banyaknya klaster
x_i: Data x pada fitur ke-i
ϕ_k: Rata-rata klaster ke-k pada nilai k $(k = 1, 2, 3, \ldots, K)$

2.8 Klasterisasi *X-Means*

Klasterisasi *k-means* salah satu teknik klasterisasi yang paling sederhana dan efektif. Kesederhanaannya terletak pada pendekatannya yang intuitif. Dimulai dengan memilih pusat klaster secara acak, menetapkan titik ke klaster terdekat, memperbarui pusat klaster, dan mengulangnya hingga konvergen. Hal yang penting untuk dicatat di sini adalah bahwa jumlah klaster yang ingin dicari harus diberikan sebelum menjalankan algoritma. Kesederhanaan ini membuat *k-means* mudah dipahami, efisien secara komputasi, dan serbaguna untuk berbagai jenis data.

Meskipun sederhana, *k-means* memiliki keterbatasan. Keterbatasan pertama adalah bahwa algoritma ini lambat dan kurang dapat menangani peningkatan jumlah data dengan baik, terutama dalam hal waktu yang dibutuhkan untuk menyelesaikan setiap iterasi. Dan karena jumlah klaster harus diberikan oleh pengguna, pengguna harus memiliki gambaran tentang perkiraan jumlah klaster dalam dataset atau mencoba dengan berbagai nilai k (jumlah klaster) sebelum mencapai hasil yang baik. Keterbatasan lainnya adalah ketika dijalankan dengan nilai k yang tetap, biasanya *k-means* menemukan optimasi lokal terburuk.

Klasterisasi *x-means* sepenuhnya menyelesaikan dua keterbatasan pertama dan memberikan solusi sebagian untuk keterbatasan ketiga. *X-means* adalah pengembangan dari algoritma *k-means* yang secara otomatis menentukan jumlah klaster optimal dalam suatu dataset dengan menggunakan kriteria seperti *Akaike Information Criterion* (*AIC*) atau *Bayesian Information Criterion* (*BIC*) (Pelleg & Moore, 2000).

Algoritma *x-means* merupakan perluasan dari algoritma *k-means* yang dirancang untuk secara otomatis memperkirakan jumlah klaster optimal berdasarkan data yang diberikan. Berbeda dengan *k-means* yang membutuhkan jumlah klaster k sebagai input tetap, *x-means* mengevaluasi beberapa kemungkinan nilai k dalam suatu rentang tertentu dan memilih yang terbaik berdasarkan skor *Bayesian Information Criterion* (BIC).

Parameter utama dalam *x-means* adalah sebagai berikut:

1. K_{start}: Jumlah klaster awal (minimum) yang akan digunakan sebagai titik

awal klasterisasi.

2. K_{end}: Jumlah klaster maksimum yang diperbolehkan.

Algoritma *x-means* memulai proses klasterisasi dengan K_{start} klaster menggunakan algoritma *k-Means*. Kemudian, untuk setiap klaster hasil tersebut, algoritma mengevaluasi kemungkinan membaginya menjadi dua sub-klaster. Evaluasi ini didasarkan pada perbandingan nilai BIC antara model satu klaster dan dua klaster.

Algoritma akan memilih jumlah klaster $K_{\text{opt}} \in [K_{\text{start}}, K_{\text{end}}]$ yang menghasilkan nilai BIC tertinggi. Hal ini mencerminkan model dengan keseimbangan terbaik antara kompleksitas (jumlah klaster/parameter) dan kecocokan terhadap data.

Dengan pendekatan ini, *x-means* dapat menghindari pemilihan jumlah klaster yang terlalu kecil atau terlalu besar secara otomatis, memberikan hasil klasterisasi yang lebih adaptif terhadap struktur data.

Konsep klasterisasi *x-means* dirangkum ke dalam dua fase. Fase pertama adalah membangun klaster acak dengan *centroid* acak, kemudian menggunakan *k-means* untuk mengoptimalkan. Proses ini diulangi secara rekursif hingga salinan akhir *centroid* ditentukan, dan salinan akhir *centroid* ini disebut *centroid* induk. Pada fase kedua, *centroid* induk menghasilkan beberapa *centroid* turunan (C_j), di mana lokasi awal *centroid* turunan ditentukan seperti pada Persamaan 2.8.12 dengan $\theta = \frac{2\pi}{j}$ dan (d_m) adalah jarak *Euclidean* rata-rata antara induk dan node dalam klaster seperti pada Persamaan 2.8.13.

$$P_n = \alpha_{n-1}(x, y) + \frac{1}{S_n} \sum_{i=1}^{S_n} \beta_i(x, y) \quad \begin{cases} n \in 1, 2, 3, \ldots, k \\ i \in 1, 2, 3, \ldots, S_n \end{cases} \tag{2.8.11}$$

$$C_j = \begin{cases} X_j = d_m \cos\theta, & j \in 1, 2, 3, \ldots, P_n \\ y_j = d_m \sin\theta, \end{cases} \tag{2.8.12}$$

$$d_m = \frac{1}{S_n} \sum_{i=1}^{S_n} \sqrt{(x_i - x_P)^2 + (y_i - y_P)^2}, \quad P \in P_n \tag{2.8.13}$$

dengan:
P_n : *Centroid* induk ke-n
$a_{n-1}(x, y)$: Posisi x dan y dari *centroid* induk sebelumnya $(n - 1)$

S_n : Kumpulan node yang dialokasikan ke P_n
$b_i(x, y)$: Posisi x dan y dari node ke-i
d_m : Jarak rata-rata antara *centroid* induk dan node dalam klaster
C_j : Lokasi *centroid* turunan ke-j dari induk (P_n).

Setelah itu, *centroid* turunan baru C_j ditetapkan sebagai titik pusat klaster, kemudian node dikelompokkan ulang dan ditugaskan menggunakan teknik minimisasi, di mana node dikelompokkan berdasarkan rata-rata jarak minimum mereka dari *sink* dan *centroid* turunan (C_j) seperti pada persamaan berikut:

$$AD_{ij} = \min\left(\frac{d_{ij} + d_{\text{isink}}}{2}\right), \quad i \in 1, 2, 3, \ldots, S_n \tag{2.8.14}$$

$$AD_i = \min\{AD_{i1}, AD_{i2}, AD_{i3}, \ldots, AD_{ij}\}, \quad i \in S_n,\ j \in P_n \tag{2.8.15}$$

$$CC_j = \varphi(x, y) + \frac{1}{S_n}\sum_{i=1}^{S_n} \beta_i(x, y), \quad \begin{cases} n = 1, 2, 3, \ldots, k \\ i = 1, 2, 3, \ldots, S_n \end{cases} \tag{2.8.16}$$

dengan:
AD_{ij} : jarak rata-rata ke *centroid* turunan j dan $sink$
AD_i : jarak rata-rata minimum dari node i ke *centroid* turunan j
D_{ij} : jarak *Euclidean* node i ke *centroid* turunan j
D_{isink} : jarak *Euclidean* node i ke $sink$
CC_{nj} : posisi akhir dari *centroid* turunan j dari induk n
$\varphi(x, y)$: posisi dari sink

Algoritma *x-means* secara rekursif menjalankan Persamaan 2.6.16. hingga posisi pusat *centroid* turunan tetap. Proses rekursif ini memiliki tiga hasil: pertama, *centroid* turunan membentuk klaster mereka sendiri dan induk tereduksi; kedua, beberapa *centroid* turunan dan induk menyusut; dan ketiga, *centroid* turunan tereduksi, dan induk diposisikan kembali ke lokasi terbaik (Radwan *et al*, 2020).

2.9 *Bayesian Information Criterion* (*BIC*)

Salah satu kriteria informasi yang banyak digunakan untuk evaluasi model adalah *Bayesian Information Criterion* (BIC). Berbeda dengan *Akaike Information Criterion* (AIC) yang didasarkan pada prinsip maksimum *likelihood* dengan

penalti terhadap jumlah parameter, BIC diturunkan dalam kerangka kerja Bayesian dan dapat diinterpretasikan sebagai pendekatan terhadap logaritma dari *Bayes Factor* antara dua model yang bersaing.

Secara matematis, BIC didefinisikan sebagai:

$$\text{BIC} = -2 \cdot \log L + p \cdot \log n \tag{2.9.17}$$

dengan
L: Nilai maksimum dari fungsi *likelihood* dari model terhadap data
p: Jumlah parameter dalam model
n: Jumlah total observasi atau ukuran sampel.

Kriteria BIC memberikan penalti yang lebih besar terhadap kompleksitas model dibandingkan AIC, karena mengalikan jumlah parameter dengan logaritma ukuran sampel. Oleh karena itu, BIC cenderung lebih konservatif dalam memilih model yang lebih kompleks. Dalam konteks klasterisasi, BIC digunakan untuk menentukan jumlah klaster optimal dengan memilih model yang meminimalkan nilai BIC.

Secara sekilas, *BIC* berbeda dari *AIC* hanya pada bagian kedua dari rumusnya, yang kini bergantung pada ukuran sampel n. Model yang meminimalkan *Bayesian Information Criterion* dipilih dari perspektif *bayesian*, *BIC* dirancang untuk menemukan model yang paling mungkin berdasarkan data yang ada. Kinerja kriteria pemilihan model dalam memilih model yang baik untuk data yang diamati dievaluasi menggunakan studi simulasi. Perbandingan semacam ini tidaklah sederhana, dan bahkan relevansinya dapat dipertanyakan, mengingat kedua kriteria tersebut didasarkan pada motivasi teoritis dan tujuan yang berbeda. Namun, untuk tujuan aplikasi, *Akaike Information Criterion* (*AIC*) dan *Bayesian Information Criterion* (*BIC*) memiliki tujuan yang sama, yaitu mengidentifikasi model yang baik, meskipun definisi mereka tentang ”model yang baik” berbeda (Acquah, 2010).

BIC adalah ukuran parametrik tentang seberapa baik suatu model memprediksi data. Ini menggambarkan *trade-off* antara kemungkinan data di bawah model dan kompleksitas model. Model dengan lebih banyak parameter dapat memprediksi data dengan lebih baik, tetapi dapat menyebabkan *overfitting*. *BIC* telah digunakan dalam algoritma *x-means* untuk memilih jumlah klaster optimal dalam rentang

nilai yang diberikan sesuai dengan sifat intrinsik dari dataset yang diberikan. Teknik yang setara yang disebut *Minimum Description Length* (*MDL*).

2.10 Jarak *Euclidean*

Analisis klaster memberikan penugasan sekumpulan n kasus ke dalam kelompok atau klaster berdasarkan pengukuran ketidaksamaan (atau jarak) antara berbagai kasus, yang diukur pada sekumpulan p variabel. Pengukuran jarak membentuk dasar untuk mendefinisikan seberapa mirip atau tidak mirip berbagai kasus tersebut. Misalkan d_{ij} menyatakan jarak antara titik x_i dan x_j dalam ruang berdimensi p. Tidak ada kesepakatan mengenai pengukuran jarak yang paling tepat untuk digunakan, tetapi semuanya memiliki tiga sifat berikut.

1. Simetri. Jarak dari x_i ke x_j adalah sama dengan jarak dari x_j ke x_i, yaitu $d_{ij} = d_{ji}$.
2. Nonnegativitas. Jarak diukur sebagai kuantitas non-negatif, yaitu $d_{ij} \geq 0$.
3. Identifikasi. Jarak antara x_i dan x_i adalah nol, yaitu $d_{ii} = 0$.

Secara umum, pengukuran jarak dianggap ideal apabila memenuhi sifat-sifat metrik berikut:

1. *Definiteness*. Jika jarak antara x_i dan x_j adalah nol, maka x_i dan x_j adalah sama yaitu, $d_{ij} = 0$ hanya jika $x_i = x_j$.
2. Ketidaksamaan segitiga. Panjang satu sisi segitiga yang dibentuk oleh tiga titik tidak bisa lebih besar dari panjang total dua sisi lainnya yaitu, $d_{ij} \leq d_{ik} + d_{jk}$.

Jelas bahwa sifat-sifat ini menggambarkan karakteristik yang mendasar untuk suatu pengukuran jarak. Jarak yang bukan metrik memiliki masalah, diantaranya dapat memiliki jarak nol meskipun titik-titik tersebut tidak bertepatan, dan juga bahwa proyeksi dari n titik ke ruang berdimensi lebih rendah dapat menjadi masalah.

Secara geometris, jarak *Euclidean* antara dua titik adalah jarak terpendek yang mungkin antara kedua titik tersebut. Selain lima sifat di atas, pengukuran jarak

Euclidean tidak berubah di bawah transformasi ortogonal variabel (memutar titik-titik tidak mengubah jarak). Karena analisis komponen utama hanya merupakan proses pemusatan data, diikuti dengan rotasi sumbu, maka jarak *Euclidean* antara skor komponen utama adalah sama dengan yang ada di ruang asli.

Karena banyaknya sifat berguna dari metrik *Euclidean*, mayoritas analisis klaster dalam literatur klimatologi didasarkan pada jarak *Euclidean*, dan perkembangan terbaru dalam algoritma klaster sebagian besar melibatkan penggunaan jarak *Euclidean*. Jarak kuadrat *Euclidean* tidak memiliki sifat ketidaksamaan segitiga dan oleh karena itu dianggap tidak cocok untuk digunakan dalam sebagian besar aplikasi klimatologi.

Salah satu masalah dengan pengukuran jarak *Euclidean* adalah bahwa ia tidak mempertimbangkan korelasi antara variabel. Di tempat di mana ada variabel yang sangat berkorelasi, variabel-variabel ini pada dasarnya mengukur karakteristik yang sama. Dalam situasi ini, jarak *Euclidean* memberikan bobot yang sama untuk setiap variabel, dengan demikian memberikan bobot tambahan pada karakteristik tunggal yang diukur oleh variabel-variabel yang berkorelasi. Secara efektif, jarak *Euclidean* memberikan bobot berlebih pada variabel yang berkorelasi (Mimmack *et al*, 2001).

Algoritma klasterisasi memulai prosesnya dengan menentukan sejumlah k *centroid* awal secara acak, yang merepresentasikan pusat dari masing-masing klaster. Selanjutnya, algoritma menghitung jarak antara setiap titik data $\mathbf{x}_i \in \mathbb{R}^n$ terhadap masing-masing *centroid* $\boldsymbol{\mu}_j \in \mathbb{R}^n$ menggunakan metrik *Euclidean*, yang didefinisikan sebagai berikut:

$$d(\mathbf{x}_i, \boldsymbol{\mu}_j) = \sqrt{\sum_{l=1}^{n} (x_{il} - \mu_{jl})^2} \tag{2.10.18}$$

dengan:
$\mathbf{x}_i \quad = (x_{i1}, x_{i2}, \ldots, x_{in})$ adalah vektor data ke-i pada ruang berdimensi n
$\boldsymbol{\mu}_j \quad = (\mu_{j1}, \mu_{j2}, \ldots, \mu_{jn})$ adalah *centroid* dari klaster ke-j
$d(\mathbf{x}_i, \boldsymbol{\mu}_j)$ = jarak *Euclidean* antara titik data $\mathbf{x}_i$ dan *centroid* $\boldsymbol{\mu}_j$.

Fungsi jarak ini digunakan untuk menentukan klaster mana yang paling dekat dengan masing-masing titik data, sehingga setiap titik dapat dialokasikan ke

klaster dengan *centroid* terdekat.

2.11 *Davies-Bouldin Index* (*DBI*)

Davies Bouldin Index DBI adalah sebuah ukuran untuk mengevaluasi kinerja klasterisasi. Gagasan dasarnya adalah mengevaluasi pemisahan antara klaster ke-i dan klaster ke-j, di mana jarak antar klaster seharusnya sebesar mungkin, sementara jarak dalam klaster seharusnya sekecil mungkin. DBI memiliki korelasi positif untuk kasus ”dalam kelas” dan korelasi negatif untuk kasus ”antar kelas”. Dalam metode ini semakin kecil nilai yang mendekati 0, semakin optimal klasterisasi yang dihasilkan.

Rumus *DBI* adalah sebagai berikut:

$$\text{DBI} = \frac{1}{K} \sum_{i=1}^{K} \max_{i,j \neq i} \frac{S_i + S_j}{d_{i,j}}$$

dengan:

$$S_i = \frac{1}{|C_i|} \sum_{x_j \in C_i} \|x_j - v_i\|$$

adalah ukuran dispersi dalam klaster i,

dengan:
K = jumlah klaster
x_j = vektor fitur berdimensi n yang ditugaskan ke klaster i
v_i = pusat klaster ke-i
$\| \cdot \|$ = jarak *Euclidean*
$d_{i,j} = \|v_i - v_j\|$ = jarak antara pusat klaster i dan j.

Secara umum, validasi klasterisasi, yang biasanya terbagi menjadi dua kategori utama validasi eksternal dan validasi internal digunakan untuk menilai kinerja hasil klasterisasi. Rendon (2011) menyajikan studi komparatif antara empat indeks eksternal: *F-measure*, *NMI Measure Entropy*, *Purity*, dan lima indeks internal: *BIC*, *CH*, *DBI*, *SIL*, *DUNN*.

Mereka menguji algoritma *k-means* dan *bisecting k-means* pada 12 kumpulan data sintetis. Hasilnya:

- Untuk algoritma *bisecting k-means*, tingkat keberhasilan mencapai 86% dengan indeks internal, dan 51,9% dengan indeks eksternal.
- Untuk algoritma *k-means*, akurasi mencapai 76,9% dengan indeks internal, dan 61,5% dengan indeks eksternal.

Disisi lain, DBI menunjukkan kinerja terbaik pada kedua algoritma tersebut. Faktanya, tidak hanya hasil penelitian yang menunjukkan bahwa indeks eksternal kurang efektif, tetapi juga kebutuhan akan pengetahuan awal dalam indeks eksternal menunjukkan bahwa indeks tersebut tidak cocok untuk data nyata. Perbedaan utama antara indeks eksternal dan internal adalah apakah diperlukan pengetahuan awal. Namun, untuk sebagian besar aplikasi nyata, pengetahuan awal biasanya tidak tersedia. Oleh karena itu, indeks eksternal tidak cocok untuk menentukan jumlah klaster. Disisi lain, DBI memiliki kemampuan diskriminasi yang lebih baik dibandingkan metrik lainnya. *Dunn Index* juga merupakan metrik evaluasi internal. Dibandingkan dengan *Dunn Index*, DBI tidak sensitif terhadap titik batas dan dapat menentukan jumlah klaster dengan benar ketika lebih dari 2. Dapat disimpulkan bahwa DBI adalah indeks yang sesuai untuk menentukan jumlah klaster dan mudah diintegrasikan ke dalam algoritma *k-means* dan *x-means* (Eppstein, 2025).

2.12 *Silhouette Score*

Salah satu metode evaluasi yang banyak digunakan dalam analisis klaster adalah *silhouette score*. Konsep ini pertama kali diperkenalkan oleh Rousseeuw (1987) dalam artikelnya *Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis Silhouette score* digunakan untuk mengukur kualitas pengelompokan dengan melihat tingkat kesesuaian setiap objek terhadap tempat klaster tersebut bergabung dibandingkan dengan klaster lain yang paling berdekatan.

Secara matematis, nilai *silhouette* untuk suatu objek i dihitung menggunakan dua parameter utama, yaitu $a(i)$ dan $b(i)$. Nilai $a(i)$ merepresentasikan rata-rata jarak objek i terhadap semua anggota dalam klaster yang sama, sedangkan $b(i)$ adalah rata-rata jarak terkecil dari objek i terhadap anggota klaster lain yang paling dekat. Dengan demikian, nilai *silhouette* didefinisikan sebagai:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{2.12.19}$$

Interpretasi nilai $s(i)$ adalah sebagai berikut: jika mendekati $+1$, objek berada dalam klaster yang sesuai; jika mendekati 0, objek terletak di batas antara dua klaster; sedangkan jika mendekati -1, objek cenderung salah ditempatkan dalam klaster yang tidak tepat. Untuk menilai kualitas keseluruhan, digunakan nilai rata-rata dari seluruh objek yang dikenal dengan *Average Silhouette Width (ASW)*. Semakin tinggi nilai rata-rata ini (mendekati 1), semakin baik kualitas klaster yang terbentuk.

Selain sebagai ukuran numerik, *silhouette* juga diperkenalkan sebagai alat visualisasi untuk mengevaluasi kualitas klaster. Plot ini menampilkan distribusi nilai *silhouette* pada tiap klaster, sehingga dapat digunakan untuk mengidentifikasi klaster yang kompak maupun objek yang berpotensi sebagai outlier. Hingga saat ini, *silhouette score* masih menjadi salah satu metrik paling populer dalam validasi klaster karena kemudahan perhitungan, interpretasi yang intuitif, serta dapat diterapkan pada berbagai algoritma klasterisasi seperti *k-means*, *hierarchical*, maupun DBSCAN (Rousseeuw, 1987).

BAB III

METODE PENELITIAN

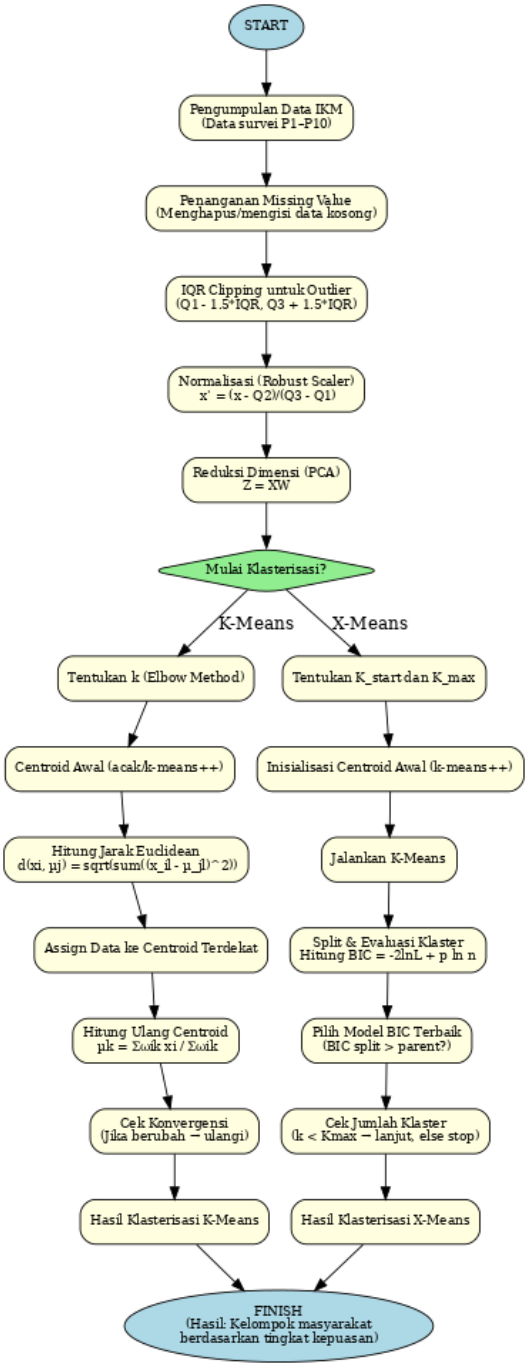
3.1 Waktu dan Tempat Penelitian

Penelitian ini dilakukan pada semester genap tahun ajaran 2024/2025 di Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung Jalan Prof. Dr. Ir. Soemantri Brojonegoro, Gedong Meneng, Kecamatan Rajabasa, Kota Bandar Lampung, Lampung.

3.2 Data Penelitian

Data yang digunakan dalam penelitian ini merupakan data sekunder, yaitu data Indeks Kepuasan Masyarakat (IKM) terhadap layanan SAMSAT yang diperoleh dari Kantor Sistem Administrasi Manunggal Satu Atap (SAMSAT), Badan Pendapatan Daerah (BAPENDA) Provinsi Lampung, tahun 2024.

3.3 Metode Penelitian



Gambar 3.1 Metode Penelitian

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian mengenai klasterisasi data Indeks Kepuasan Masyarakat (IKM) terhadap pelayanan publik menggunakan metode *K-Means* dan *X-Means*, maka dapat disimpulkan beberapa hal sebagai berikut:

1. Metode *k-means*, dengan jumlah klaster yang telah ditentukan sebelumnya k=5, menghasilkan pengelompokan yang stabil dan cukup baik secara metrik evaluasi seperti inertia 162.52, *silhouette score* 0.7463, dan DBI 0.5662. Klasterisasi ini membentuk segmentasi mutu pelayanan dari kategori A hingga C, namun kurang mampu menangkap variasi lokal yang lebih halus.
2. Metode *x-means* secara otomatis membentuk 12 klaster berdasarkan nilai BIC tertinggi. Meskipun jumlah klaster lebih banyak, distribusi dominan dari unit pelayanan menunjukkan bahwa hanya 5 klaster utama yang merepresentasikan sebagian besar unit pelayanan. Dengan performa metrik evaluasi yang lebih unggul: *silhouette score* sebesar 0.8496, dan DBI sebesar 0.5040.
3. Hasil klasterisasi unit pelayanan SAMSAT menggunakan metode *k-means* dan *x-means* menghasilkan lima klaster (A–E) dengan komposisi berbeda. Unit seperti Metro, Pesawaran, dan Bandar Lampung (Mall) tergolong dalam klaster A atau B yang mencerminkan kualitas layanan baik, sedangkan Pringsewu secara konsisten berada dalam klaster tersendiri (klaster E), menunjukkan karakteristik yang sangat berbeda dari unit lain.

Perbedaan pembagian antara kedua metode mencerminkan fleksibilitas model dalam menangkap pola data, namun konsistensi klaster tertentu mengindikasikan adanya unit-unit yang secara signifikan menonjol, baik positif maupun negatif. Hasil ini dapat digunakan sebagai dasar evaluasi kinerja pelayanan dan prioritas peningkatan kualitas layanan di unit-unit dengan performa yang masih rendah.

4. Indikator keamanan dan kenyamanan merupakan kelemahan utama yang memerlukan prioritas perbaikan. Sebaliknya, indikator kesesuaian persyaratan pelayanan telah berjalan dengan baik dan perlu dipertahankan kualitasnya.

5. Hasil klasterisasi menunjukkan adanya hubungan antara jenis pekerjaan, pendidikan terakhir, jenis kelamin, serta jenis layanan dengan distribusi klaster. Misalnya, profesi seperti wiraswasta dan petani cenderung berada di klaster dengan nilai IKM lebih rendah, sedangkan profesi seperti ASN/PNS cenderung berada di klaster dengan nilai IKM yang lebih tinggi.

5.2 Saran

Berdasarkan hasil penelitian, disarankan agar pemilihan metode klasterisasi disesuaikan dengan tujuan analisis. Jika tujuan utamanya adalah memperoleh segmentasi yang lebih rinci dan adaptif terhadap variasi data, maka *x-means* menjadi pilihan yang lebih tepat dibandingkan *k-means* yang cenderung konservatif. Hasil klasterisasi yang lebih mendalam dari *x-means* dapat dimanfaatkan oleh instansi terkait, seperti Badan Pendapatan Daerah atau Dinas Pelayanan Publik, sebagai dasar perumusan kebijakan peningkatan mutu pelayanan secara lebih terarah, terutama untuk unit-unit yang masuk dalam kategori mutu rendah.Pihak BAPENDA disarankan memberikan prioritas utama pada perbaikan indikator keamanan dan kenyamanan. Pada indikator kesesuaian persyaratan perlu terus dipertahankan kualitasnya agar kepuasan masyarakat tetap konsisten.

DAFTAR PUSTAKA

Acquah, H. D. G. 2010. Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship. *Journal of Development and Agricultural Economics*. **2**(1): 1-6

Agusta, Y. 2007. K-Means-Penerapan, Permasalahan dan Metode Terkait. *Jurnal Sitem Informatika*. **3**(1): 47 – 60.

Annur, M. C. 2021. Persyaratan Berbelit, Keluhan Utama Masyarakat Terhadap Pelayanan Publik. *Databoks*. https://datAboks.katadata.co.id/politik/statisti/65fa06b690e14d2/persyaratan-berbelit-keluhan utama-masyarakat-terhadap-pelayanan-publik.

Dunham, M. H. 2006. *Data Mining: Introductory and Advanced Topics*. Pearson Education. ISBN: 978-8177587852

Dunham, M. H. 2002. *Data Mining: Introductory and Topics*. Prentice-Hal. Engle wood Cliffs.

Ekasetya, A. V., & Jananto, A. 2020. Klusterisasi optimal dengan elbow method untuk pengelompokkan data kecelakaan lalu lintas di Kota Semarang. *Dinamika Informatika*. **12**(1): 20 – 28.

Eppstein, D. 2025. Cluster Analysis *In Wikipedia. https://en.wikipedia.org/wiki/Clusteranalysis*.

Hariany, Z., & Matondang, R.A. 2014. Analisis Indeks Kepuasan Masyarakat (IKM) Terhadap Pelayanan Publik di Puskemas XXX. *e-Jurnal Teknik Industri FT USU*. **5**(2): 17 – 21.

Hariyanto, Y., Primadewi, A., & Hanafi, M. 2025. Analisis Kepuasan Masyarakat terhadap Pelayanan Publik menggunakan K-Means Clustering. *Journal of information system research*. **6**(2): 1067 – 1076.

Hourdakis, N., Argyriou, M., Petrakis, G. M. E., & Milios, E. E. 2010. Hierarchical Clustering in Medical Document Collections: the BIC-Means Method. *Journal of Digital Information Management*. **8**(2): 71 – 77.

Huang, Z. 1998. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*. **2**(1): 283 – 304.

Irwansyah, E., & Faisal, M. 2012. *Advanced Clustering: Teori dan Aplikasi*. Deepublish. Yogyakarta.

Javier, F. 2023. Tingkat kepercayaan Masyarakat terhadap Institusi Bisnis dan Pemerintah,Siapa yang Lebih Tinggi?. *Tempo.co.https://www.tempo.co/data/data/tingkat-kepercayaan-masyarakat-terhadap-institusi-bisnis-dan-pemerintah-siapa-yang-lebih-tinggi–9936474.*

Mimmack, M. G., Mason, J. S., & Galpin, S. J. 2001. Choice of Distance Matrices in Cluster Analysis: Defining Regions. *Journal of Climate*. **14**(12): 2790–2797.

Nurrizka, H.R., & Saputra, W. 2011. Pengukuran Indeks Kepuasan Masyarakat Terhadap Pelayanan Kesehatan. *Jurnal Manajemen Pelayanan Kesehatan*. **14**(1): 11–19.

Patimah, E., Ernatita., & Chamidah, N. 2021. Analisis Cluster Kepuasan Pengguna Terhadap Layanan Shopee Menggunakan Algoritma K-Means. *Jurnal informatik*. **17**(3): 209–217.

Pelleg, D., & Moore, A. 2000. X-means: Extending k-means with efficent estimation of the number of cluster. *School of computer science*. Carnegia Mellon University, Pittsburgh.

Radwan, A., Kamarudin, N., Solihin, I. M., Leong, H., Rizon, M., Desa, H., & Azizi, M. 2020. X-means clustering for wireless sensor network. *Journal of Robotics, Networking and Artificial Life*. **7**(2): 111–115.

Rousseeuw, P.J. 1987. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*. **20**(1): 53–65.

Shalih, A.F., Ramadhan, A.F., & Shalaisha, N. 2025. Tinjauan Komprehensif tentang Aplikasi dan Perkembangan Principal Component Analysis (PCA). *Jurnal EurekaMatika*. **13**(1): 25–34.

Siregar, M. A., Puspabhuana, A. 2017. *Data Mining: Pengolahan data menjadi informasi dengan rapidminder*. CV.Kekata Group. Surakarta.

Sofiyah, O. S., Nining, R., & Dana, D. R. 2023. Analisis efektivitas pelayanan publik menggunakan k-means clustering Di Kecamatan Sukagumiwang. *Jurnal mahasiswa teknik informatika*. **7**(2): 1291-1295.

Suhaeni, C., Kurnia, A., & Ristiyanti. 2018. Perbandingan Hasil Pengelompokan menggunakan Analisis Cluster Berhirarki, K-Means Cluster, dan Cluster Ensemble (Studi Kasus Data Indikator Pelayanan Kesehatan Ibu Hamil). *Jurnal media infotama*. **14**(1): 31-38.

Yusuf, B., Mahara, R., Ahmadian, H., Wahyuni, S., & Khairan, A.R. 2022. Analisis Clustering Penduduk Miskin Di Provinsi Aceh Menggunakan Algoritma K-Means Dan X-Means. *Jurnal Nasional Komputasi dan Teknologi Informasi*. **51**(1): 26-35.