SKRIPSI

IMPLEMENTASI ALGORITMA K-MEDOIDS DALAM PENGELOMPOKAN JENIS PELANGGARAN LALU LINTAS (STUDI KASUS KOTA MAKASSAR)

Disusun dan Diajukan Oleh:

ANDI IFFAT AINIYYAH HAMKA

D121181025

DEPARTEMEN TEKNIK INFORMATIKA

FAKULTAS TEKNIK

UNIVERSITAS HASANUDDIN

2022

LEMBAR PENGESAHAN SKRIPSI

IMPLEMENTASI ALGORITMA K-MEDOIDS DALAM PENGELOMPOKAN JENIS PELANGGARAN LALU LINTAS (STUDI KASUS KOTA MAKASSAR)

Disusun dan diajukan oleh

ANDI IFFAT AINIYYAH HAMKA

D121181025

Telah dipertahankan di hadapan Panitia Ujian yang dibentuk dalam rangka Penyelesaian Studi Program Sarjana Program Studi Teknik Informatika Fakultas Teknik Universitas Hasanuddin pada tanggal 25 November 2022 dan dinyatakan telah memenuhi syarat kelulusan.

Menyetujui,

Pembimbing Utama,	Pembimbing Pendamping,
Dr. Amil Ahmad Ilham, S.T., M.IT.	Anugrayani Bustamin, S.T., M.T.
NIP. 19731010 199802 1 001	NIP. 19901201 201807 4 001

Ketua Program Studi,

Dr. Indrabayu, S.T., M.T., M.Bus.Sys.
NIP. 19750716 200212 1 004

PERNYATAAN KEASLIAN

Yang bertanda tangan dibawah ini:

Nama	: Andi Iffat Ainiyyah Hamka
Nim	: D121181025
Program Studi	: Teknik Informatika
Jenjang	: S1

Menyatakan dengan ini karya tulisan saya berjudul:

IMPLEMENTASI ALGORITMA K-MEDOIDS DALAM PENGELOMPOKAN JENIS PELANGGARAN LALU LINTAS (STUDI KASUS KOTA MAKASSAR)

Adalah karya tulisan saya sendiri dan bukan merupakan pengambilan alihan tulisan orang lain bahwa skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri.

Apabila di kemudian hari terbukti atau dapat dibuktikan bahwa sebagian atau keseluruhan skripsi ini hasil karya orang lain, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Gowa, 2 Desember 2022

Yang Menyatakan,

Andi Iffat Ainiyyah Hamka

KATA PENGANTAR

Puji Syukur penulis panjatkan kepada Allah SWT atas segala rahmat, kemudahan, limpahan, dan karunia-Nya sehingga penulis dapat melaksanakan dan menyelesaikan penelitian yang berjudul **"Implementasi K-Medoids dalam Pengelompokan Jenis Pelanggaran Lalu Lintas (Studi Kasus Kota Makassar)"** sebagai bentuk memenuhi syarat kelulusan strata-1 di Prodi Teknik Informatika Fakultas Teknik Universitas Hasanuddin, Makassar.

Penelitian dan penulisan laporan ini tentunya melibatkan pihak yang berperan penting dalam membimbing, membantu, dan mendukung penulis selama penelitian ini berlangsung. Penulis mengucapkan banyak terima kasih kepada:

1. Allah SWT atas segala berkat, kemudahan, dan kelancaran yang dilimpahkan kepada penulis sehingga mampu menyelesaikan penelitian ini dengan baik.
2. Orang tua penulis Bapak H. Hamka Tere dan Ibu Hj. A. Besse Intan yang selalu memberikan doa, dukungan, kasih sayang, dan sabar dalam mendidik penulis.
3. Bapak Dr. Amil Ahmad Ilham, S.T., M.IT. selaku pembimbing I dan Ibu Anugrayani Bustamin, S.T., M.T. selaku pembimbing II, yang senantiasa memberikan waktu, tenaga, dan perhatian untuk membimbing penulis dalam menyelesaikan penelitian.
4. Seluruh staf dan dosen Departemen Teknik Informatika Fakultas Teknik

Universitas Hasanuddin yang selalu membimbing, mendidik, mengarahkan, dan membantu penulis selama masa perkuliahan.

5. Teman-teman SMA penulis Masitha, Indri, Alifia, Lea, Pute, Nisa, Rahma, Novi, Nadya, dan Amirah yang selalu memberikan semangat kepada penulis.
6. Teman-teman kampus penulis Tamara, Rofifah, Ainun, Dea, dan Ayu yang selalu memberikan bantuan, perhatian yang baik, dan mendukung penulis.
7. Seluruh teman angkatan Synchronous 2018 yang senantiasa menemani, memberikan dukungan, dan membantu penulis sejak maba.
8. Teman-teman belajar penulis Sandi, Emirat, Fadhil, dan Zahran yang selalu memberikan saran dan bantuan saat proses pengerjaan sistem.
9. Saudari penulis Ayi, Aicis, Mia, dan Restu yang selalu memberikan dukungan, inspirasi, dan mendengarkan keluh kesah penulis selama meneliti.
10. Pihak-pihak yang namanya tidak disebutkan namun membantu penulis baik dalam menyusun laporan maupun mengerjakan sistem.

Penulis berharap semoga Allah SWT membalas segala kebaikan dan jasa dari semua pihak yang banyak mendukung dan membantu penulis dalam menyelesaikan tugas akhir ini. Sebagai manusia biasa yang tidak luput dari kekurangan ataupun kesalahan, penulis menyadari tugas akhir ini masih jauh dari kata sempurna, maka penulis mengharapkan adanya saran dan kritik untuk penelitian ini. Semoga penelitian

ini bermanfaat untuk kedepannya. Terima kasih.

Gowa, 1 Desember 2022

Penulis,

Andi Iffat Ainiyyah Hamka

ABSTRAK

Lalu lintas merupakan ruang gerak bagi para masyarakat untuk berkendara yang meliputi para pengemudi dan pejalan kaki. Berdasarkan data dari Badan Pusat Statistik tahun 2020 di Kota Makassar telah tercatat jumlah kendaraan bermotor menurut jenisnya, yaitu 248.682 unit mobil penumpang, 17.501 unit bus, 85.968 unit truk, dan 1.338.306 unit sepeda motor dan tahun berikutnya cenderung mengalami peningkatan. Padatnya pengguna kendaraan ini tentunya dapat mempengaruhi meningkatnya angka pelanggaran lalu lintas di jalan. Penelitian ini bertujuan mengelompokkan jenis pelanggaran lalu lintas di Kota Makassar dengan memanfaatkan algoritma K-Medoids dan melakukan visualisasi hasil klaster dengan menggunakan *Word Cloud* yang diharapkan memberikan informasi terkait pola-pola klaster jenis pelanggaran lalu lintas. Penelitian ini akan menggunakan studi kasus di Satlantas Polrestabes Kota Makassar pada tahun 2021 dengan total kasus pelanggaran lalu lintas sebanyak 5893 kasus. Data yang digunakan berupa data tilang yang terdiri dari fitur pasal dan jenis kendaraan. Hasil klasterisasi menunjukkan jenis kendaraan sepeda motor dan mini bus merupakan kendaraan yang paling sering dikenakan pelanggaran lalu lintas. Kendaraan r2 yaitu sepeda motor tidak hanya didominasi pelanggaran yang berkaitan dengan penggunaan helm standar SNI, namun juga cukup dominan pada pelanggaran tidak memenuhi kelengkapan dan kepemilikan SIM/STNK dan tidak memenuhi syarat teknik laik jalan seperti kaca spion, lampu utama, klakson, dll. Kendaraan r4 khusunya tipe mini bus dan mobil penumpang yang cukup dominan melanggar aturan lalu lintas. Aturan yang dilanggar tidak hanya berkaitan dengan penggunaan sabuk keselamatan bagi kendaraan r4, tetapi juga cukup dominan pada pelanggaran tidak memenuhi kelengkapan STNK, tidak dapat menunjukkan SIM, tidak memasang TKB (Tanda Kendaraan Bermotor), dan lain sebagainya. Hasil penelitian ini didapatkan hasil klaster dengan nilai k=9 dengan uji validasi nilai *Silhouette Score* sebesar 0.867.

Kata kunci = *pelanggaran, clustering, kmedoids, PCA, elbow method, silhouette score, word cloud*

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB I

PENDAHULUAN

1.1 Latar Belakang

Lalu lintas merupakan ruang gerak bagi para masyarakat untuk berkendara yang meliputi para pengemudi dan pejalan kaki. Penggunaan kendaraan dalam ruang berlalu lintas pun selalu meningkat, baik dalam penggunaan transportasi umum maupun transportasi pribadi. Berdasarkan data dari Badan Pusat Statistik tahun 2020 di Kota Makassar telah tercatat jumlah kendaraan bermotor menurut jenisnya, yaitu 248.682 unit mobil penumpang, 17.501 unit bus, 85.968 unit truk, dan 1.338.306 unit sepeda motor. Di tahun selanjutnya, telah terjadi peningkatan jumlah kendaraan yang cukup signifikan, yaitu 257.015 unit mobil penumpang, 17.582 unit bus, 88.359 unit truk, dan 1.377.837 unit sepeda motor. Kepadatan pengguna kendaraan ini dapat mempengaruhi meningkatnya pelanggaran lalu lintas di jalan.

Berbagai macam jenis pelanggaran yang dilakukan oleh pengendara seperti melanggar lampu lalu lintas, mengebut atau balapan, tidak menggunakan sabuk pengaman saat berkendara, serta melanggar rambu dan marka jalan. Adapun yang menjadi faktor penyebab terjadinya pelanggaran lalu lintas yaitu adanya pengaruh cuaca panas dan hujan, kondisi lalu lintas yang padat, dan dominannya penggunaan kendaraan pribadi (Ambo *et al*., 2020). Selain itu adanya kelemahan dari aparat

kepolisian yang dipicu oleh pengetahuan masyarakat yang minim dalam taat berlalu lintas, kurangnya kemampuan atau kompetensi yang dimiliki, serta sarana dan prasarana yang masih belum memadai (Anindhito & Maerani, 2018). Akibatnya, hal ini dapat memberikan dampak negatif dan ancaman keselamatan bagi para pengendara lainnya.

Dari hasil observasi, kasus pelanggaran lalu lintas di Makassar tahun 2020 berkisar hingga 15.000 kasus dan ada kemungkinan akan terus mengalami pertambahan. Pihak Satlantas Polrestabes Kota Makassar melakukan penindakan terhadap kasus pelanggaran dengan pencatatan tilang secara manual lalu menginput data melalui sistem tilang elektronik. Namun dalam sistem ini hanya mampu menyimpan data-data informasi pelanggaran lalu lintas yang bersifat deskriptif dari pelanggar. Maka akan dimanfaatkan analisis data mining untuk menghasilkan suatu informasi dan pengetahuan baru pada pola-pola yang akan terbentuk dari berbagai jenis pelanggaran lalu lintas.

Penerapan algoritma K-Means yang diimplementasikan untuk mengklasterkan wilayah-wilayah rawan pelanggaran lalu lintas menghasilkan 3 jenis klaster (Dewi *et al*., 2019). Penelitian lainnya melakukan klasterisasi dengan algoritma K-Means studi kasus data register perkara lalu lintas dan didapatkan hasil bahwa karakteristik pelanggar terbanyak yaitu usia remaja dan dewasa terdiri dari profesi pelajar, mahasiswa, dan pegawai swasta (Ramadhani et al., 2017). Atmaja *et al* melakukan penelitian dengan memanfaatkan algortima K-Medoids dalam menganalisa pola kejahatan hasilnya didapatkan 3 tingkat pola kriminalitas. Berdasarkan dari penelitian

terdahulu, beberapa penelitian dengan studi kasus yang sama melakukan proses pengelompokan dengan mengimplementasikan algoritma K-Means, sehingga dalam penelitian ini akan digunakan jenis algoritma yang berbeda yaitu K-Medoids Clustering. Selain itu, dalam penelitian oleh Marlina *et al* melakukan perbandingan antara algoritma K-Medoids dan K-Means pada kasus sebaran anak cacat, hasilnya berdasarkan pengujian nilai *Silhouette* performa algoritma K-Medoids lebih baik dibandingkan algoritma K-Means dengan jumlah 3 klaster sebaran wilayah. Selanjutnya di tahun 2019, Kurmiati *et al* melakukan perbandingan dua algoritma *clustering* yaitu metode K-Medoids dan K-Means yang bertujuan untuk menganalisa sebaran gempa dan pola spasial kejadian gempa di Indonesia. Hasilnya didapatkan nilai k=6 adalah nilai k terbaik dan nilai *Silhouette* tertinggi sebesar 0.4674067. Dari hasil perbandingan kedua algoritma, diketahui berdasarkan hasil perhitungan nilai *Silhouette* diperoleh hasil K-Medoids lebih baik bila dibandingkan dengan K-Means.

Berdasarkan uraian diatas, maka data tilang yang akan dikumpulkan dari Satlantas Polrestabes Kota Makassar akan diklasterkan dengan memanfaatkan algoritma K-Medoids. Algoritma ini akan mengelompokkan berbagai jenis pelanggaran yang terjadi di Kota Makassar.

1.2 Rumusan Masalah

1. Bagaimana mengelompokkan jenis pelanggaran di Kota Makassar dengan menggunakan algoritma K-Medoids?

2. Bagaimana memvisualisasi hasil pengelompokan jenis pelanggaran lalu lintas dalam bentuk sistem informasi?

1.3 Tujuan Penelitian

1. Untuk mengelompokkan jenis pelanggaran lalu lintas di Kota Makassar dengan menggunakan algoritma K-Medoids.
2. Memvisualisasi hasil pengelompokan jenis pelanggaran lalu lintas dalam bentuk sistem informasi.

1.4 Manfaat Penelitian

Penelitian ini bermanfaat untuk memberikan informasi terkait pola-pola klaster jenis pelanggaran lalu lintas yang diharapkan dapat membantu pihak Satlantas dalam menindaki dan mengurangi tingkat pelanggaran di Kota Makassar.

1.5 Batasan Masalah Penelitian

Penelitian ini akan menggunakan studi kasus di Satlantas Polrestabes Kota Makassar pada tahun 2021 dengan total kasus pelanggaran lalu lintas sebanyak 5893 kasus. Data yang digunakan berupa data tilang yang meliputi tanggal melanggar, alamat pelanggar, pasal yang dilanggar, barang bukti, jenis kendaraan, dan nomor polisi. Algoritma K-Medoids akan melakukan pengelompokan data pelanggaran kemudian hasil klaster akan divisualisasi berbentuk *Word Cloud*.

1.6 Sistematika Penulisan

Berikut uraian bab-bab tentang isi tulisan dalam penelitian ini:

BAB I PENDAHULUAN

Bab ini berisi tentang latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, dan batasan masalah penelitian.

BAB II TINJAUAN PUSTAKA

Bab ini membahas terkait landasan teori yang berhubungan dengan penyelasaian masalah yang diteliti dan berkaitan dengan data penelitian yang akan dianalisis yaitu fenomena pelanggaran lalu lintas, *data mining*, *unsupervised learning*, *clustering*, *principal component analysis*, algoritma K-Medoids, SSE, metode elbow, *Silhouette Score*, dan *Word Cloud*.

BAB III METODOLOGI PENELITIAN

Pada bab ini menguraikan tentang tahap-tahap penelitian, tempat dan waktu penelitian, teknik pengambilan data, instrumen penelitian, proses pembuatan sistem, implementasi K-Medoids, pengujian klaster, serta hasil visualisasi klaster.

BAB IV HASIL DAN PEMBAHASAN

Bab ini berisi mengenai hasil dari sistem yang telah dibuat dan pembahasan hasil penelitian.

BAB V PENUTUP

Bab ini berisikan kesimpulan yang didapatkan dari hasil penelitian serta saran-saran untuk penelitian selanjutnya.

BAB II

TINJAUAN PUSTAKA

2.1 Pelanggaran Lalu Lintas

Pelanggaran lalu lintas merupakan suatu kejadian yang seringkali terjadi dalam ruang gerak berlalu lintas masyarakat yang bertentangan dengan peraturan perundang undangan terkait lalu lintas. Kejadian ini melibatkan para pengendara dan para pengguna jalan yang mengakibatkan dampak yang cukup merugikan, baik itu korban jiwa ataupun materiil. Hal ini dapat berdampak pada beberapa aspek yaitu meningkatnya laju angka kecelakaan lalu lintas (lakalantas) dalam lingkup jalan raya dan persimpangan jalan lalu lintas, keselamatan para pengguna jalan menjadi terancam, dan adanya kebiasaan melanggar lalu lintas mengakibatkan terbentuknya budaya melanggar peraturan (Nurfauziah & Krisnani, 2021). Maraknya pelanggaran lalu lintas mampu mengusik kenyamanan dan keamanan dalam aktivitas berkendara.

Pihak kepolisian berwenang dalam upaya melakukan pencegahan dan penegakan hukum pelanggaran lalu lintas. Sebagaimana diatur dalam Undang-Undang Nomor 22 Tahun 2009 yang membahas tentang lalu lintas termasuk dalam sistem transportasi nasional yang harus ditingkatkan peranan dan potensinya guna menciptakan ketertiban, keamanan, kelancaran dan keselamatan dalam berlalu lintas. Kepolisian Republik Indonesia membagi pelanggaran lalu lintas dibagi menjadi 3 jenis

yakni pelanggaran ringan, pelanggaran sedang, dan pelanggaran berat. Pihak Korlantas Polri telah mengidentifikasi kasus tilang hingga Oktober 2021 sebanyak 1,77 juta bukti pelanggaran lalu lintas.

Satlantas Polrestabes Kota Makassar melakukan proses penindakan pelanggaran lalu lintas dengan memberikan sanksi berupa peringatan atau teguran secara langsung dan surat denda tilang berdasar dari pasal yang dikenakan kepada para pelanggar lalu lintas. Berdasarkan PP Nomor 80 Tahun 2012, pelaksanaan penindakan pelanggaran melalui proses pengisian dan tanda tangan pada Belangko Tilang dengan melengkapi data berikut:

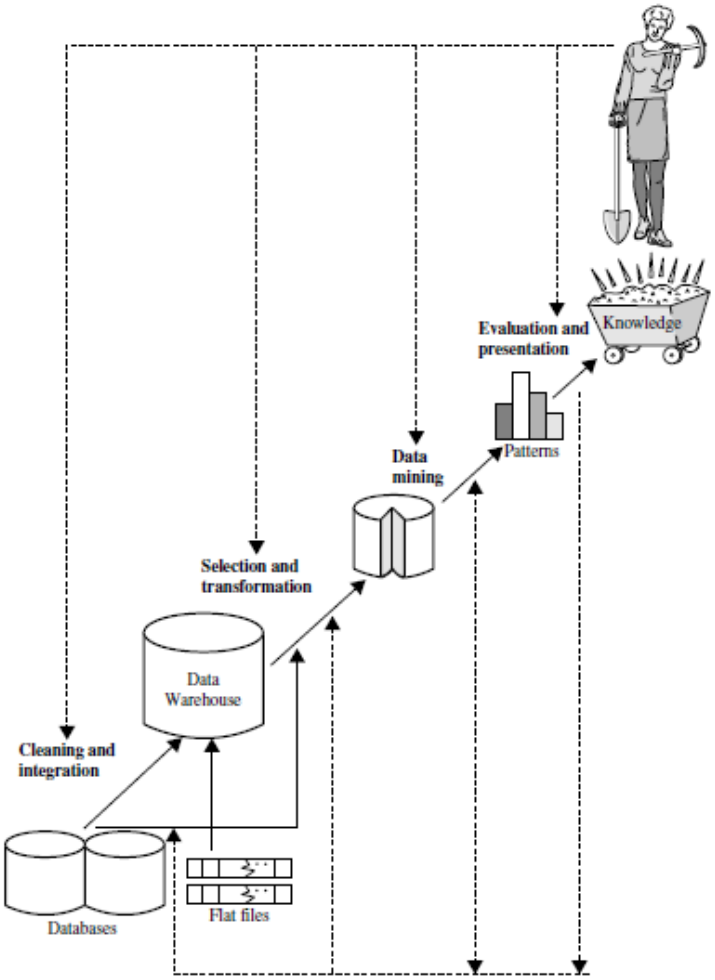
a) Identitas pelanggaran dan Kendaraan Bermotor yang digunakan;
b) Ketentuan dan pasal yang dilanggar;
c) Hari, tanggal, jam, dan tempat terjadinya pelanggaran;
d) Barang bukti yang disita;
e) Jumlah uang titipan denda ke Bank;
f) Tempat atau alamat dan/ atau nomor telpon pelanggar;
g) Pemberian kuasa;
h) Penandatanganan oleh pelanggar dan Petugas Pemeriksa;
i) Berita acara singkat penyerahan Surat Tilang kepada pengadilan;
j) Hari, tanggal, jam, dan tempat untuk menghadiri sidang pengadilan; dan
k) Catatan petugas penindak

2.2 Data Mining

Data mengalami perkembangan yang pesat, tersedia secara luas, dan memiliki jumlah yang sangat besar sehingga dapat disebut sebagai era data. Sebuah alat ataupun metode yang berguna dibutuhkan yang secara otomatis dapat mengungkap atau mengekstrak informasi berharga dari sejumlah besar data dan untuk mengubah data tersebut menjadi sebuah pengetahuan. Adanya kebutuhan inilah yang mendorong lahirnya *data mining*. *Data mining* merupakan tahapan mencari pola-pola dan pengetahuan menarik dari jumlah data yang besar (Han *et al*., 2012).

Mehmed Kantardzic berpendapat bahwa *data mining* dibagi menjadi dua kategori, yakni bersifat prediktif yang berarti teknik untuk menghasilkan suatu model yang dapat diterapkan dalam proses klasifikasi, prediksi, dan estimasi, kemudian bersifat deskriptif yang berarti dengan memanfaatkan kumpulan data yang besar untuk memperoleh suatu pengetahuan yang berasal dari pola dan hubungan yang dihasilkan (Kantardzic, 2020).

Menurut Jiawei *et al* terdapat pada Gambar 2. 1 mengilustrasikan tahapan-tahapan *data mining* dengan proses sebagai berikut:



Gambar 2. 1 Tahapan data mining (Han et al., 2012)

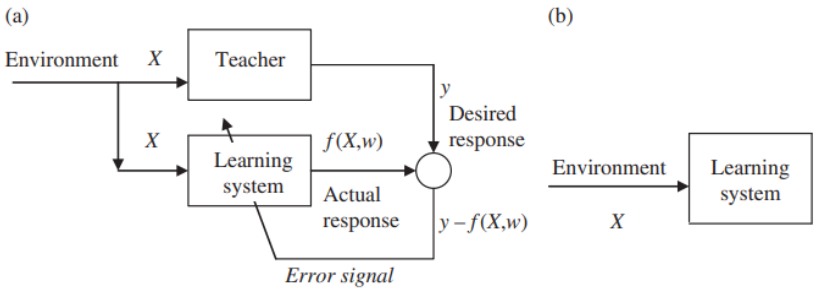
a. *Data Cleaning*, Tahap untuk membersihkan data yang memiliki nilai yang kosong (*missing value*), menghilangkan *noise*, dan tidak konsisten
b. *Data Integration*, Tahap menggabungkan beberapa sumber data yang memungkinkan

c. *Data Selection*, Tahap dimana dalam *database* akan diambil data yang relevan dengan sistem analisis
d. *Data Transformation*, Tahap melakukan operasi *summary* atau *aggregation* untuk mengubah data ke dalam bentuk yang sesuai
e. *Data Mining*. Tahap proses penting untuk ekstraksi pola data dengan menerapkan metode cerdas
f. *Pattern Evaluation*, Tahap dimana berdasarkan *interestingness measures* akan diidentifikasi pola-pola yang menarik yang mewakili pengetahuan
g. *Knowledge Presentation*, Tahap menyajikan pengetahuan kepada *user* dengan menggunakan teknik visualisasi dan representasi pengetahuan

2.3 Unsupervised Learning

Unsupervised learning adalah teknik penambangan tidak diawasi yang melakukan penyelesaian masalah dengan data yang tidak berlabel, sehingga tidak memerlukan adanya data latih dan data uji, contohnya klastering dan asosiasi. *Unsupervised learning* merupakan bentuk upaya dalam mengidentifikasi pola tersembunyi dari data tanpa memperkenalkan proses data latih yang meliputi masukan dan label kelas (Dua & Du, 2011).

Tanpa perlu adanya seorang guru atau secara harfiah set data yang perlu dilatih seperti metode *supervised*, model yang dibangun dengan metode *unsupervised* mengharuskan data mampu membentuk dan mengevaluasi dirinya sendiri. Dalam skema pembelajaran tanpa pengawasan, hanya sampel dengan nilai *input* yang diberikan ke *learning system* dan tidak ada gagasan *output* pasti yang dihasilkan selama proses pembelajaran yang bertujuan untuk menemukan struktur alami dari data *input* (Kantardzic, 2020). Pada Gambar 2. 2 (a) dan (b) mengilustrasikan blok diagram perbedaan alur pembelajaran *supervised* dan *unsupervised*.



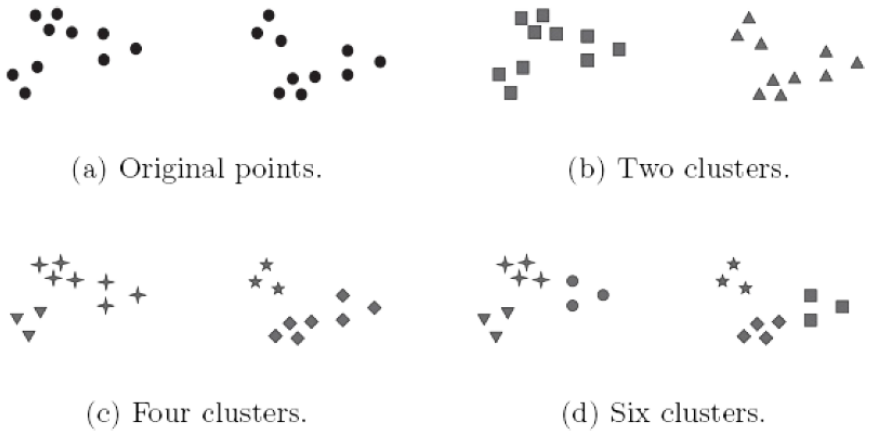
Gambar 2. 2 Ilustrasi blok diagram metode supervised (a) dan unsupervised (b) (Kantardzic, 2020)

2.4 Clustering

Clustering merupakan salah satu jenis metode *unsupervised learning* yang mampu mengelompokkan suatu objek pada set data yang dimana hasil dari pengelompokan yang diperoleh kemiripan dalam data satu sama lain untuk data yang berada dalam klaster yang sama. *Clustering* bertujuan untuk menemukan kelompok objek yang bermanfaat (*cluster*), dimana pemanfaatannya ditentukan oleh tujuan dari

analisis data (Tan *et al*., 2019). *Clustering* bertugas untuk mempartisi titik-titik data ke dalam kelompok yang disebut sebagai *cluster*, sedemikian rupa sehingga titik-titik pada suatu kelompok memiliki tingkat kemiripan yang sangat dekat, sedangkan titik-titik yang berada pada kelompok berbeda sebisa mungkin tidak sama (Zaki & Meira, 2020). Melalui *clustering*, dapat diketahui kesamaan apa yang dimiliki oleh tiap klaster.

Terdapat tiga cara yang berbeda dalam mengelompokkan kumpulan titik/data yang sama, pada Gambar 2. 3 menunjukkan bentuk penanda yang membentuk keanggotaan klaster menurut Pang Ning Tan *et al*.



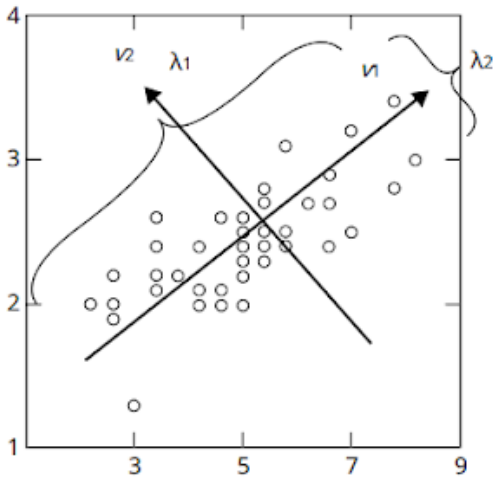
Gambar 2. 3 Bentuk keanggotaan cluster (Tan., et al 2019)

2.5 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) adalah suatu cara untuk mereduksi jumlah fitur atau variabel data yang besar menjadi beberapa komponen utama dengan tetap mempertahankan integritas dan esensi dari kumpulan data. PCA merupakan salah

satu teknik dalam mencari basis dimensi yang paling baik menemukan varians dalam data (Zaki & Meira, 2020). Diantara banyaknya variabel, PCA mempertahankan variasi dengan mengekspresikan kembali banyak variabel asli menjadi beberapa variabel baru sehingga sebagian besar variasi yang ada pada variabel asli tetap diperhitungkan ataupun dipertahankan oleh beberapa variabel baru yang tidak saling berkorelasi (Ratner, 2017).

Varians data mewakili variasi dari jumlah data yang dimuat data asli setelah menerapkan PCA. Komponen utama adalah ortogonal dalam ruang fitur, seperti yang ditunjukkan Gambar 2. 4, V1 merupakan komponen utama pertama yang mewakili varians asli dalam data dan V2 merupakan komponen utama kedua yang mewakili varians yang tersisa (Dua & Du, 2011).



Gambar 2. 4 Komponen PCA dua dimensi (Dua & Du)

Berikut langkah-langkah dalam penyelesaian PCA secara matematis (Tyagi, 2022) yaitu:

- Menghitung nilai rata-rata setiap fitur atau variabel dengan Formula (1):

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_M}{M} = \frac{\sum_{i=1}^{M} x_i}{M} \quad (1)$$

- Menghitung nilai varians yaitu jarak antara titik variabel itu dari nilai rata-rata variabel yang ditunjukkan dalam Formula (2):

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1} \quad (2)$$

Keterangan:

S^2 = varians sampel

x_i = nilai dari satu observasi

$\bar{x}$ = nilai dari semua observasi

n = jumlah observasi

- Menghitung matriks kovarians sebanyak n dimensi yang dinyatakan dalam Formula (3). Kovarians mengukur bagaimana dua variabel acak bervariasi satu sama lain dan digunakan dalam n-dimensi $n \geq 2$.

$$cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1} \quad (3)$$

Keterangan:

$cov(x,y)$ = kovarians antar variabel x dan y

x_i = nilai data x

y_i = nilai data y

$\bar{x}$ = rata-rata nilai data x

$\bar{y}$ = rata-rata nilai data y

n = jumlah nilai data

Untuk matriks kovarians 3 dimensi yang diilustrasikan dalam Formula (4):

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix} \quad (4)$$

- Menghitung nilai eigen dan vektor eigen dari matriks kovarians. Misalkan matriks kovarians dilambangkan dengan C, maka persamaan matriks kovarians ditunjukkan pada Formula (5), dimana λ adalah nilai eigen,

$$|C - \lambda I| \quad (5)$$

Kemudian untuk vektor eigen dinyatakan dalam Formula (6), dimana X merupakan *non-zero* vektor (mewakili komponen)

$$(C - \lambda I)X = 0 \quad (6)$$

- Urutkan nilai eigen dan sesuaikan vektor eigen yang diperoleh
- Membangun matriks fitur dengan memilih k vektor eigen dengan nilai eigen terbesar
- Transformasikan ke subruang baru untuk mengurangi jumlah dimensi

PCA memiliki sebuah *hyperparameter* "n_components" yang digunakan dala memutuskan banyak komponen yang ingin dipakai saat mereduksi jumlah fitur set data. Namun dalam penentuan jumlah komponen yang dipilih tentunya membutuhkan sebuah metode untuk mengetahui seberapa banyak komponen yang sebaiknya dipertahankan. Menurut Rukshan Pramoditha, dikarenakan "n_components" merupakan *hyperparameter* sehingga ia tidak belajar dari set data, maka cara penentuan nilainya dilakukan secara manual sebelum menerapkan fungsi PCA.

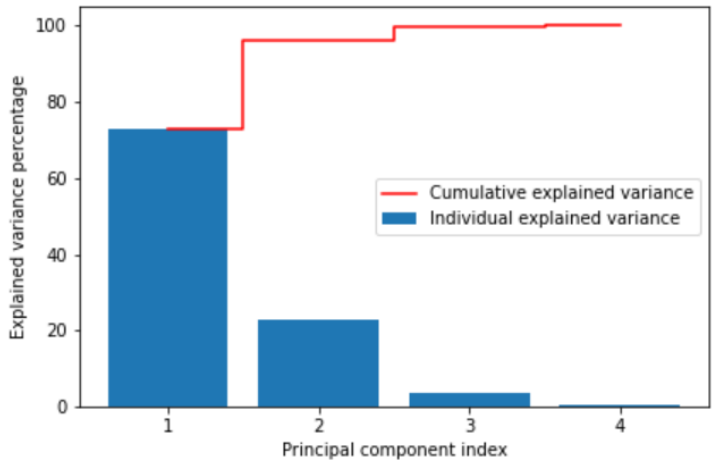
Berikut metode-metode untuk memilih jumlah komponen utama terbaik menurut Rukshan Pramoditha:

- Penggunaan komponen PCA untuk kebutuhan dalam visualisasi data. Dalam hal ini jumlah komponen utama bergantung pada plot yang dipilih. Untuk plot 2 dimensi memerlukan 2 komponen utama dan untuk plot 3 dimensi memerlukan 3 komponen utama,
- Penggunaan komponen PCA berdasarkan jumlah varians data. Metode ini adalah cara termudah untuk menentukan jumlah komponen terbaik pada set data. Misalnya, varians data asli yang ingin dipertahankan sebesar 85% setelah menerapkan PCA, maka dapat ditentukan untuk "n_components" sebesar 0.85. Setelah itu, secara otomatis akan dihasilkan jumlah komponen utama yang mempertahankan 85% varians data pada set data asli.

  Dalam beberapa kasus batas variasi yang sering digunakan adalah 70% hingga 90% dan bisa saja lebih tinggi atau lebih rendah lagi bergantung dari kumpulan data yang ada. Dalam kasus seperti nilai variansi yang lebih besar dari 90% sudah mewakili sumber variasi sangat jelas dan dominan namun dalam kasus struktur yang dihasilkan kurang jelas, maka diperlukan batas yang lebih tinggi dari 90%. Sebaliknya, jika menetapkan nilai variasi sebesar 70% dan PC yang dihasilkan dapat digunakan dalam

analisis lebih lanjut, maka ambang batasnya dapat ditetapkan lebih rendah (Jolliffe, 2002).

- Penggunaan komponen PCA berdasarkan hasil plot persentase varians pada masing-masing komponen dan persentase total varians yang dimiliki oleh setiap komponen utama.



Gambar 2. 5 Plot komponen PCA berdasarkan explained variance (Rukshan., 2021)

Jumlah bar sama dengan jumlah variabel atau fitur dalam set data asli. Dalam plot Gambar 2. 5 setiap bar memperlihatkan persentase varians dari masing-masing komponen dan plot langkah memperlihatkan persentase dari varians secara kumulatif. Dengan melihat hasil plot dapat ditentukan berapa komponen yang ingin digunakan, kasus yang Gambar 2. 5 hanya dua komponen pertama yang menangkap hampir semua varians dalam kumpulan data.

2.6 K-Medoids Clustering

K-Medoids merupakan varian lain dari algoritma pengelompokan (*clustering*) yang dapat dimanfaatkan untuk mendapatkan hasil pengelompokan dalam kumpulan data. Metode K-Medoids ini sangat mirip dengan K-Means, namun ada sedikit perbedaan dari segi fungsi optimasinya (Malik & Tuckfield, 2019). Dalam algoritma K-Medoids, salah satu diantara anggota dalam klaster dijadikan sebagai perwakilan yang disebut medoid. Proses pengelompokan K-Medoids dengan melakukan iterasi dalam mengatur item-item data ke dalam klaster dan menominasikan medoid untuk setiap klaster, sampai klaster medoid menjadi konvergen (Jo, 2019).

Menurut Haesang dan Chihyuck (Park & Jun, 2009), Langkah-langkah algortima K-Medoids terdiri dari 3 tahap, yaitu:

Langkah 1: Menentukan inisialisasi medoid

- Memilih titik medoid sebanyak k secara acak
- Menghitung jarak antara tiap pasangan semua objek berdasarkan ketidakmiripan (*dissimilarity*) dengan memilih metode pengukuran jarak (dalam kasus ini pengukuran jarak Euclidean).

  Pengukuran jarak Euclidean antara objek i dan objek j yang dinyatakan dalam Formula (7):

$$d_{ij} = \sqrt{\sum_{a=1}^{p} (x_{ia} - x_{ja})^2}, \quad i = 1, \ldots . n; \; j = 1, \ldots . n \qquad (7)$$

- Hasil klaster awal akan didapatkan dengan menetapkan setiap objek ke medoid terdekatnya
- Hitung jumlah jarak dari semua objek ke medoidnya, dengan menghitung V_j untuk semua objek, yang dinyatakan dalam Formula (8):

$$V_j = \sum_{i=1}^{n} d_{ij} \qquad (8)$$

Langkah 2: Melakukan pembaruan medoid

- Temukan medoid baru dalam setiap klaster, yaitu objek yang meminimalkan jarak total ke objek lain dalam klasternya.
- Hitung jumlah jarak dari semua objek ke medoidnya, jika jumlahnya lebih besar dari sebelumnya, maka hentikan iterasi algoritma. Jika tidak, lakukan pembaruan pada medoid saat ini dengan medoid baru di setiap klaster dan kembali ke Langkah temukan medoid baru.

Langkah 3: Penetapan objek ke medoid

- Lakukan penetapan pada setiap objek ke medoid terdekat dan dapatkan hasil klaster.

Menurut Mark dan Tuckfield dalam beberapa aspek, kedua algoritma ini yaitu K-Medoids dan K-Means memiliki beberapa perbedaan, diantaranya:

- Kompleksitas komputasi, dari kedua metode komputasi K-Medoids lebih mahal. Saat set data yang dimiliki terlalu besar (>10.000 poin) dan ingin menghemat waktu komputasi, metode K-Means lebih disarankan.
- Kehadiran *outlier*, bila dibandingkan metode K-Means lebih sensitif terhadap *outlier* daripada metode K-Medoids.
- Pusat klaster, untuk kedua algoritma ini dalam proses cara menemukan titik pusat klasternya berbeda. Pusat klaster K-Means tidak memerlukan titik data dalam *dataset*, sedangkan K-Medoids menggunakan titik data dalam kumpulan data sebagai pusatnya.

2.7 Sum of Squared Error

Dalam metrik kualitas internal (*internal quality internal*), biasanya melakukan pengukuran kekompakan pada suatu klaster dengan menggunakan beberapa perhitungan persamaan (*similarity*). Beberapa diantaranya dengan mengukur homogenitas intra-cluster, keterpisahan antar-cluster, atau perpaduan keduanya. Metode ini hanya menggunakan data itu sendiri tanpa adanya informasi yang bersifat eksternal. Salah satunya metode *Sum of Squared Error* (SSE). SSE merupakan kriteria pengukuran dalam pengelompokan yang paling sederhana dan banyak digunakan (Maimon & Rokach, 2010).

Apabila nilai pada suatu klaster semakin mendekati 0, maka titik-titik data pada kumpulan data yang ditetapkan dalam suatu klaster mempunyai kedekatan yang signifikan. Proses perhitungan SSE yang dinyatakan dalam Formula (9) (Thinsungnoen *et al*., 2015) :

$$SSE = \sum_{i=1}^{K} \sum_{x_j \in C_i} \left\| x_j - m_i \right\|^2 \quad (9)$$

Keterangan:

K = Jumlah Klaster

X = {x1, x2, x3, ….., xn}

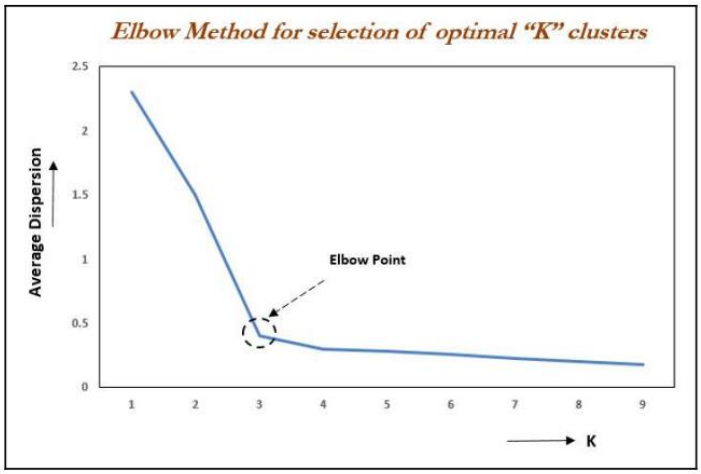
C = {C1, C2, C3, ….., Cn}

M = Titik pusat dari klaster

‖ ‖ = perhitungan jarak Euclidean

2.8 Metode Elbow

Metode Elbow merupakan suatu cara untuk menentukan jumlah klaster yang paling optimal dalam proses klastering. Metode ini melakukan plot nilai fungsi yang dihasilkan oleh nilai-nilai yang berbeda dari nilai k. Seperti yang diketahui apabila k meningkat, maka distorasi rata-rata akan mengalami penurunan. Elbow atau siku adalah nilai dari k dimana tingkat distorsinya paling menurun, yang dimana proses pembagian data menjadi kelompok harus berhenti (Dangeti, 2017). Pada Gambar 2. 6 mengilustrasikan contoh penerapan metode elbow.



Gambar 2. 6 Elbow Method (Dangeti, 2017)

2.9 Silhouette Score

Silhouette digunakan untuk mengukur seberapa baik kualitas dari klaster yang dihasilkan. Metrik *Silhouette* bekerja dengan menganalisis seberapa baik suatu titik dalam klasternya. Metrik ini berkisar dari -1 hingga 1, apabila skor siluet rata-rata pada

seluruh pengelompokan adalah satu maka dicapai hasil pengelompokan yang sempurna tetapi terdapat sedikit kesulitan untuk mengetahui tentang titik mana termasuk dimana (Johnston *et al*., 2019).

Secara matematis, perhitungan *Silhouette Score* dilakukan dengan mudah melalui *Simplified Silhouette Index* (SSI) yang dinyatakan dalam Formula (10):

$$SSI_i = \frac{b_i - a_i}{\max(a_i b_i)} \tag{10}$$

Dimana, a_i merupakan jarak dari titik i untuk klaster centroidnya dan b_i merupakan jarak dari titik i ke klaster terdekat centroidnya. a_i mewakili bagaimana titik kohesif klaster i sebagai klaster yang jelas dan b_i mewakili seberapa jauh letak jarak klaster.

Kaufman dan Rousseeuw mengemukakan bahwa berdasarkan dari pengalaman mereka, *Silhouette* menghasilkan suatu interpretasi yang subjektif seperti yang ditampilkan dalam Tabel 2. 1.

Tabel 2. 1 Interpretasi Silhouette Score (Kaufman & Rousseuw., 1990)

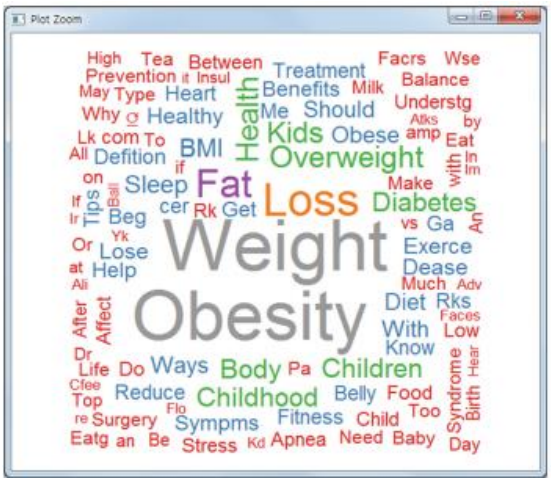
Silhouette	**Interpretasi**
0.71-1.00	Telah didapatkan sebuah struktur yang kuat
0.51-0.70	Telah didapatkan sebuah struktur yang masuk akal
0.26-0.50	Telah didapatkan struktur yang lemah dan mungkin buatan. Cobalah metode tambahan pada set data
≤ 0.25	Tidak didapatkan struktur yang substansial

2.10 Word Cloud

Text mining pada dasarnya menampilkan data yang terstruktur ataupun tidak terstruktur ke dalam model yang telah disederhanakan, salah satunya adalah *Word Cloud*. Visualisasi *Word Cloud* berbentuk gambar atau grafik yang memuat kata-kata yang digunakan dalam teks atau subjek tertentu, dimana ukuran setiap kata menunjukkan frekuensi atau kepentingannya (James., *et al* 2017). *Word Cloud* ini juga banyak dimanfaatkan dalam kebutuhan visualisasi data.

James *et al* mengemukakan bahwa *Word Cloud* merupakan teknik yang memvisualisasikan kata kunci (*keyword*) dalam teks guna memahami maknanya. Misalnya, ukuran kata akan ditingkatkan tergantung frekuensi penggunaannya apabila

kata sering disebutkan. Salah satu contoh kasus dalam penerapan *Word Cloud* dalam proses pencarian kata-kata obesitas selama musim dingin.



Gambar 2. 7 Word Cloud (James et al., 2017)

Gambar 2. 7 menunjukkan bahwa kata-kata yang paling banyak digunakan adalah *Diet, Health, Danger, Obesity, dan Effectiveness*. Hasil dari analisis dari James *et al*, dimungkinkan untuk mengetahui bahwa orang gemuk selama musim dingin, menunjukkan minat yang tinggi pada subjek-subjek tersebut.