

**PERBANDINGAN KINERJA *K-MEANS* DAN *K-MEDOIDS* PADA
TIME SERIES-BASED CLUSTERING DENGAN *DYNAMIC TIME
WARPING* PADA DATA JUMLAH PENDUDUK MISKIN DI
SULAWESI SELATAN TAHUN 2010-2023**

**NURUL MAGFIRAH
H051191081**



**PROGRAM STUDI STATISTIKA DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS HASANUDDIN
2024**

**PERBANDINGAN KINERJA *K-MEANS* DAN *K-MEDOIDS* PADA
TIME SERIES-BASED CLUSTERING DENGAN *DYNAMIC TIME
WARPING* PADA DATA JUMLAH PENDUDUK MISKIN DI
SULAWESI SELATAN TAHUN 2010-2023**

NURUL MAGFIRAH
H051191081



Diajukan sebagai salah satu syarat memperoleh gelar Sarjana Statistika

**PROGRAM STUDI STATISTIKA DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS HASANUDDIN
MAKASSAR
AGUSTUS 2024**

SKRIPSI

**PERBANDINGAN KINERJA *K-MEANS* DAN *K-MEDOIDS* PADA
TIME SERIES-BASED CLUSTERING DENGAN *DYNAMIC TIME
WARPING* PADA DATA JUMLAH PENDUDUK MISKIN DI
SULAWESI SELATAN TAHUN 2010-2023**

yang disusun dan diajukan oleh

NURUL MAGFIRAH
H051191081

Skripsi,

Telah dipertahankan di depan Panitia Ujian Sarjana Statistika pada 13 Agustus 2024
dan dinyatakan telah memenuhi syarat kelulusan

pada

Program Studi Statistika
Departemen Statistika
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Hasanuddin
Makassar

Mengesahkan:
Pembimbing tugas akhir,

Mengetahui:
Ketua Program Studi,


Sitti Sahriman, S.Si., M.Si
NIP. 198810182015042002



Dr. Anna Islamiyati, S.Si., M.Si.
NIP. 197708082005012002

PERNYATAAN KEASLIAN SKRIPSI DAN PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa, skripsi berjudul "Perbandingan Kinerja *K-Means* dan *K-Medoids* pada *Time Series-Based Clustering* dengan *Dynamic Time Warping* pada Data Jumlah Penduduk Miskin di Sulawesi Selatan Tahun 2010-2023" adalah benar karya saya dengan arahan dari pembimbing Sitti Sahriman, S.Si., M.Si. Karya ilmiah ini belum diajukan dan tidak sedang diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka skripsi ini. Apabila di kemudian hari terbukti atau dapat dibuktikan bahwa sebagian atau keseluruhan skripsi ini adalah karya orang lain, maka saya bersedia menerima sanksi atas perbuatan tersebut berdasarkan aturan yang berlaku.

Dengan ini saya melimpahkan hak cipta (hak ekonomis) dari karya tulis saya berupa skripsi ini kepada Universitas Hasanuddin.

Makassar, 13 Agustus 2024



Nurul Maqfirah
NIM H051191081

UCAPAN TERIMA KASIH

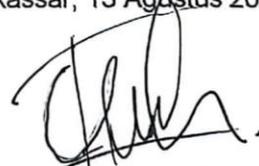
Puji dan syukur penulis panjatkan kepada Allah *Subhanahu wa Ta'ala* atas limpahan berkah, rahmat, dan karunia-Nya sehingga penulis mampu menyelesaikan skripsi yang berjudul "**Perbandingan Kinerja K-Means dan K-Medoids pada Time Series Based-Clustering dengan Dynamic Time Warping pada Data Jumlah Penduduk Miskin di Sulawesi Selatan Tahun 2010-2023**". Shalawat dan salam senantiasa tercurahkan kepada baginda Rasulullah *Shallallahu 'Alaihi wa Sallam* beserta keluarga dan para sahabatnya. Penulis menyadari sesungguhnya skripsi ini dapat terselesaikan tidak terlepas dari bantuan dan dukungan baik moril maupun materil dari berbagai pihak. Oleh karena itu, penulis mengucapkan terima kasih yang setulus-tulusnya serta penghargaan setinggi-tingginya kepada orang tua tercinta yakni Ibunda **Sukira** dan Ayahanda **Sallarong** atas segala doa, nasihat, kasih sayang, dukungan, dan motivasi yang diberikan kepada penulis demi kesuksesan dan kebahagiaan penulis.

Ucapan terima kasih juga penulis persembahkan kepada Ibu **Sitti Sahriman, S.Si., M.Si.** selaku dosen pembimbing tugas akhir penulis yang telah meluangkan banyak waktu, tenaga, dan pikiran untuk memberikan bimbingan, arahan, saran, serta motivasi sehingga skripsi ini dapat terselesaikan. Penulis juga ingin mengucapkan terima kasih kepada Bapak **Andi Kresna Jaya, S.Si., M.Si.** selaku dosen penasehat akademik penulis sekaligus dosen penguji dan Ibu **Anisa, S.Si., M.Si.** selaku dosen penguji penulis yang telah meluangkan waktu dalam memberikan saran dan kritikan yang membangun dalam penyempurnaan tugas akhir ini.

Ucapan terima kasih penulis juga haturkan kepada Adik **Nur Shihab** serta keluarga besar penulis khususnya Kakak **Nural Muhyida, Nurlaela, dan Thania Dwi Destriana A. R.** atas dukungan, nasihat, hiburan, dan bantuan yang telah diberikan sehingga penulis mampu melalui tahapan tugas akhir ini. Tak lupa penulis juga mengucapkan terima kasih kepada teman-teman **Statistika 2019** yang telah kebersamai serta memberikan dukungan dan bantuan kepada penulis untuk menyelesaikan tugas akhir ini khususnya **Dian Ayu Permata Sari Rusdy, Muharti Apriana, dan Amalia Mentari Djalumang** serta rekan-rekan seperbimbingan **Andi Apridhani Mattalatta, Melda Ftriyani Azis, dan Nurul Dwinilda Zhazhabilah.**

Akhirnya penulis juga mengucapkan terima kasih kepada sahabat sejak SMA **Aisyah Sri Rejeki** dan **Andi Fitriani** yang telah memberikan motivasi serta dukungan kepada penulis untuk menyelesaikan skripsi ini.

Makassar, 13 Agustus 2024



Nurul Magfirah

ABSTRAK

Nurul Magfirah. **Perbandingan Kinerja *K-Means* dan *K-Medoids* pada *Time Series Based-Clustering* dengan *Dynamic Time Warping* pada Data Jumlah Penduduk Miskin di Sulawesi Selatan Tahun 2010-2023** (dibimbing oleh Sitti Sahrinan).

Latar Belakang. *Time series-based clustering* merupakan analisis kluster yang mengelompokkan objek berdasarkan pola data *time series* yang dinamis sehingga membutuhkan metode jarak yang sesuai. *Dynamic time warping* (DTW) dipilih sebagai metode jarak karena kemampuannya menangani variasi temporal. Dalam penelitian ini, algoritma *K-means* digunakan dengan *centroid* sebagai pusat kluster sehingga lebih sensitif terhadap pencilan. Sementara itu, *K-medoids* menggunakan *medoid* sebagai pusat klusternya sehingga lebih *robust* terhadap pencilan. Untuk itu, *K-means* disempurnakan dengan DTW *Barycenter Averaging* (DBA) untuk menghasilkan *centroid* yang *robust* terhadap pencilan. **Tujuan.** Penelitian ini mengelompokkan kota/kabupaten di Sulawesi Selatan berdasarkan jumlah penduduk miskin tahun 2010-2023 dengan menggunakan algoritma *K-means* dan *K-medoids* dengan DTW sebagai metode jarak, serta membandingkan kinerja kedua algoritma tersebut dalam pengelompokan. **Metode.** Penelitian ini terdiri dari dua tahap utama, yaitu pengelompokan menggunakan metode *K-means* dan *K-medoids* berdasarkan jarak DTW, serta evaluasi kinerja masing-masing algoritma melalui koefisien *silhouette*. **Hasil.** Algoritma *K-means* dengan DBA menghasilkan 2 kluster. Masing-masing kluster terdiri dari 17 anggota dengan rata-rata 23.636 ribu jiwa dan 7 anggota dengan rata-rata 58.555 ribu jiwa. Untuk *K-medoids* juga diperoleh 2 kluster. Masing-masing kluster terdiri dari 16 anggota dengan rata-rata 22.688 ribu jiwa dan 8 anggota dengan rata-rata 56.087 ribu jiwa. Distribusi Kabupaten Maros menjadi perbedaan dari hasil kluster kedua algoritma. Pada *K-means* Maros masuk dalam kluster 1, tetapi pada *K-medoids* berada di kluster 2. Nilai koefisien *silhouette* *K-means* dengan DBA lebih tinggi dibandingkan *K-medoids* yakni 0,64588 dan 0,62385. **Kesimpulan.** Algoritma *K-means* dengan DBA memiliki kinerja yang lebih baik dibandingkan *K-medoids* dalam mengelompokkan kota/kabupaten berdasarkan data jumlah penduduk miskin tahun 2010 hingga 2023 berdasarkan nilai koefisien *silhouette*.

Kata Kunci: *DTW Barycenter Averaging, Dynamic Time Warping, K-Means, K-Medoids, Time Series-Based Clustering*

ABSTRACT

Nurul Magfirah. **Comparison of K-Means and K-Medoids Performance in Time Series-Based Clustering with Dynamic Time Warping on Poverty Data in South Sulawesi from 2010 to 2023** (supervised by Sitti Sahriman).

Introduction. Time series-based clustering is a clustering analysis that groups objects based on dynamic time series data patterns, requiring an appropriate distance metric. Dynamic Time Warping (DTW) is chosen as the distance method due to its ability to handle temporal variations. In this study, the K-means algorithm is used with centroids as the cluster centers, making it more sensitive to outliers. Meanwhile, K-medoids uses medoids as its cluster centers, making it more robust to outliers. Therefore, K-means is enhanced with DTW Barycenter Averaging (DBA) to produce centroids that are robust to outliers. **Aim.** This study aims to cluster cities/regencies in South Sulawesi based on the number of poor residents from 2010 to 2023 using the K-means and K-medoids algorithms with DTW as the distance method, and to compare the performance of these algorithms in the clustering process. **Methods.** This study consisted of two main stages: clustering using the K-means and K-medoids methods based on DTW distance, and evaluating the performance of each algorithm using the silhouette coefficient. **Results.** The K-means algorithm with DBA resulted in 2 clusters, where the first cluster consisted of 17 members with an average population of 23.636 thousand and the second cluster consisted of 7 members with an average population of 58.555 thousand. For K-medoids, 2 clusters were also formed, with the first cluster consisting of 16 members with an average population of 22.688 thousand and the second cluster consisting of 8 members with an average population of 56.087 thousand. The distribution of Maros Regency differed between the clustering results of the two algorithms. In K-means, Maros was included in cluster 1, but in K-medoids, it was placed in cluster 2. The silhouette coefficient for K-means with DBA was higher than for K-medoids, with values of 0.64588 and 0.62385, respectively. **Conclusion.** The K-means algorithm with DBA performed better than K-medoids in clustering the cities/regencies based on the number of poor people from 2010 to 2023 based on silhouette coefficient.

Keywords: *DTW Barycenter Averaging, Dynamic Time Warping, K-Means, K-Medoids, Time Series-Based Clustering.*

DAFTAR ISTILAH

Istilah	Arti dan Penjelasan
<i>Centroid</i>	Representasi dari pusat kluster yang merupakan titik rata-rata dari semua data dalam kluster
Inisialisasi	Proses mengeset nilai awal suatu variabel
Iterasi	Pengulangan langkah-langkah dalam proses algoritma untuk mencapai hasil konvergen
Kluster	Beberapa objek yang berkelompok menjadi satu
<i>Mean</i>	Rata-rata dari nilai dalam dataset
<i>Medoid</i>	Representasi dari pusat kluster yang merupakan titik yang benar-benar ada dalam data yang memiliki jumlah total jarak terdekat dari semua titik lain dalam kluster tersebut

DAFTAR LAMBANG/SINGKATAN

Lambang/Singkatan	Arti dan Penjelasan
DTW	<i>Dynamic Time Warping</i>
DBA	DTW <i>Barycenter Averaging</i>
x_{ij}	Nilai dari objek ke i pada periode j
n	Jumlah objek
μ_j	Rata-rata dari nilai semua objek pada periode j
σ_j	Standar deviasi dari nilai semua objek pada periode j
z_{ij}	Standarisasi data dari x_{ij}
$w_{i,j}$	Perbedaan abosolut tiap titik data dari dua <i>time series</i>
D	Matriks jarak kumulatif DTW
$d_{i,j}$	Elemen (i,j) pada matriks D
k	Jumlah kluster
C_k	Centroid pada kluste rk
O_i	Medoid pada kluster k
a_i^j	jarak rata-rata data ke- i dengan semua data di kluster yang sama
b_i^j	nilai minimum dari jarak rata-rata data ke- i dengan semua data di kluster berbeda
SI_i^j	<i>Silhouette Index</i> data ke- i pada kluster j
SI_j	rata-rata <i>silhouette index</i> kluster j
SI	<i>Silhouette Index</i>
SC	Koefisien <i>Silhouette</i>

DAFTAR ISI

	Halaman
HALAMAN JUDUL	i
HALAMAN PENGANTAR	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERNYATAAN KEASLIAN SKRIPSI	iv
UCAPAN TERIMA KASIH	v
ABSTRAK	vi
ABSTRACT	vii
DAFTAR ISTILAH	viii
DAFTAR LAMBANG/SINGKATAN	ix
DAFTAR ISI	x
DAFTAR TABEL	xii
DAFTAR GAMBAR	xiii
DAFTAR LAMPIRAN	xiv
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Batasan Penelitian	3
1.3 Tujuan Penelitian	3
1.4 Manfaat Penelitian	3
1.5 Teori	3
1.5.1 Analisis Kluster	3
1.5.2 <i>Time Series Based Clustering</i>	4
1.5.3 Standarisasi Data	5
1.5.4 <i>Dynamic Time Warping</i>	5
1.5.5 Algoritma <i>K-Means</i>	8
1.5.6 Algoritma <i>K-Medoids</i>	10
1.5.7 Koefisien Silhouette	11
1.5.8 Jumlah Penduduk Miskin	13
BAB II METODE PENELITIAN	14
2.1 Sumber Data	14
2.2 Variabel Penelitian	14
2.3 Struktur Data Penelitian	14
2.4 Metode Analisis	14
BAB III HASIL DAN PEMBAHASAN	17
3.1 Karakteristik Jumlah Penduduk Miskin di Sulawesi Selatan	17
3.1.1 Rata-Rata Jumlah Penduduk Miskin di Sulawesi Selatan	17
3.1.2 Tren Jumlah Penduduk Miskin	18
3.2 Pengujian Pencilan	18
3.2.1 Uji Pencilan dengan <i>Boxplot</i>	19
3.2.2 Uji Pencilan dengan <i>Robust Mahalanobis – Minimum Covarian Determinant</i>	19
3.3 Standarisasi Data	20
3.4 Pengukuran Jarak dengan <i>Dynamic Time Warping</i>	22
3.4.1 Jarak Lokal Antar Objek	22
3.4.2 Jarak Kumulatif Antar Objek	24
3.4.3 Hasil Jarak DTW	25
3.5 Analisis Kluster dengan Algoritma <i>K-Means</i>	25

3.5.1	Penentuan Nilai k Terbaik	25
3.5.2	Proses Kluster	26
3.5.3	Interpretasi Hasil Kluster.....	30
3.6	Analisis Kluster dengan Algoritma <i>K-Medoids</i>	31
3.6.1	Penentuan Nilai k Terbaik	31
3.6.2	Proses Kluster	31
3.6.3	Interpretasi Hasil Kluster.....	40
3.7	Evaluasi dan Perbandingan Hasil Kluster	41
3.7.1	Evaluasi Hasil Kluster	41
3.7.1.1	Evaluasi Hasil Kluster dari <i>K-Means</i> dengan DBA.....	41
3.7.1.2	Evaluasi Hasil Kluster dari <i>K-Medoids</i>	44
3.7.2	Perbandingan Hasil Kluster	47
3.7.2.1	Perbandingan Rata-Rata	47
3.7.2.2	Perbandingan Tren	48
BAB IV KESIMPULAN		50
4.1	Kesimpulan	50
4.2	Saran.....	50
DAFTAR PUSTAKA.....		51
LAMPIRAN		55

DAFTAR TABEL

Tabel	Halaman
1. Kriteria Koefisien Silhouette	13
2. Struktur Data Penelitian.....	14
3. Hasil Perhitungan MMCD	19
4. Data Nilai Z(1) dan Z(2)	22
5. Jarak Lokal Antar Objek	23
6. Jarak Kumulatif antar Objek	25
7. Jarak DTW antar 24 Kabupaten	25
8. Koefisien Silhouette K-Means	26
9. <i>Centroid</i> Awal	26
10. Perbandingan jarak antara <i>centroids</i> dan anggota kluster	27
11. Perbandingan jarak antara <i>centroids</i> dan anggota kluster iterasi 1	28
12. Perbandingan jarak antara <i>centroids</i> dan anggota kluster iterasi 2	29
13. Koefisien <i>Silhouette K-Medoids</i>	31
14. <i>Medoids</i> Awal	31
15. Perbandingan jarak antara <i>medoids</i> dan <i>non-medoids</i> lainnya	32
16. O_{random} iterasi 1	33
17. Perbandingan jarak antara O_{random} dan <i>non-medoids</i> lainnya iterasi 1	33
18. O_{random} iterasi 2	34
19. Perbandingan jarak antara O_{random} dan <i>non-medoids</i> lainnya iterasi 2	35
20. O_{random} iterasi 3	36
21. Perbandingan jarak antara O_{random} dan <i>non-medoids</i> lainnya iterasi 3	36
22. O_{random} iterasi 4	37
23. Perbandingan jarak antara O_{random} dan <i>non-medoids</i> lainnya iterasi 4	38
24. O_{random} iterasi 5	39
25. Perbandingan jarak antara O_{random} dan <i>non-medoids</i> lainnya iterasi 5	39
26. Hasil Pengelompokan Algoritma <i>K-Means</i> dengan DBA	47
27. Hasil Pengelompokan Algoritma <i>K-Medoids</i>	47

DAFTAR GAMBAR

Gambar	Halaman
1. Ilustrasi Pengaplikasian antara Jarak Euclidean dan DTW	6
2. Rata-rata Jumlah Penduduk Miskin di Sulawesi Selatan.....	17
3. Plot Tren Jumlah Penduduk Miskin Tahun 2010-2023	18
4. Boxplot Jumlah Penduduk Miskin.....	19
5. Hasil analisis kluster dengan metode K-means	30
6. Hasil analisis kluster dengan metode <i>k-medoids</i>	41
7. Plot <i>Silhouette Index</i> pada Algoritma <i>K-Means</i> dengan DBA	43
8. Plot <i>Silhouette Index</i> pada Algoritma <i>K-Medoids</i>	45
9. (a) & (b) Plot Tren Anggota Kluster 1 & 2 <i>K-means</i> dengan DBA. (c) & (d) Plot Tren Anggota Kluster 1 & 2 <i>K-medoids</i>	48

DAFTAR LAMPIRAN

Lampiran	Halaman
1. Data Penelitian Jumlah Penduduk Miskin	55
2. Hasil Standarisasi Data dengan <i>Z-Score</i>	57
3. Jarak Lokal antar Objek untuk Kepulauan Selayar dan Bulukumba.....	59
4. Jarak Kumulatif antar Objek untuk Kepulauan Selayar dan Bulukumba.....	60
5. Hasil Jarak antar Kabupaten/Kota.....	61
6. <i>Silhouette Index</i> pada Algoritma <i>K-Means</i> dengan DBA	63
7. <i>Silhouette Index</i> pada Algoritma <i>K-Medoids</i>	64

BAB I

PENDAHULUAN

1.1 Latar Belakang

Indonesia sebagai salah satu negara berkembang sedang menghadapi permasalahan kemiskinan. Berdasarkan data Badan Pusat Statistik (BPS) yang dirilis pada 17 Juli 2023, persentase jumlah penduduk miskin di Indonesia pada Maret 2023 sebesar 9,36 persen atau sebesar 25,90 juta jiwa. Angka ini cukup besar dan menjadi tantangan Indonesia ke depannya. Upaya penanganan kemiskinan ini dapat dimulai dari skala kecil, yaitu tingkat kabupaten/kota di berbagai provinsi di Indonesia salah satunya Sulawesi Selatan. Provinsi Sulawesi Selatan menyumbang persentase penduduk miskin sebesar 8,70 persen dari jumlah penduduknya atau sebesar 788,85 ribu jiwa penduduk. Untuk menanggulangi hal ini, pemerintah perlu mengetahui wilayah-wilayah yang termasuk pada kategori tingkat kemiskinan tinggi dan tingkat kemiskinan rendah dengan tujuan untuk menetapkan skala prioritas dalam penanggulangan ini (Bahauddin dkk., 2021). Untuk memahami secara mendalam pola kemiskinan di tingkat lokal, pendekatan statistika dapat digunakan sebagai langkah awal dalam proses penanggulangan kemiskinan dengan menentukan wilayah-wilayah yang memiliki tingkat kemiskinan tinggi dan rendah sehingga pemerintah dapat merancang strategi yang lebih terarah dan efektif untuk meningkatkan kesejahteraan masyarakat secara berkelanjutan. Salah satu pendekatan statistika yang dapat digunakan adalah analisis kluster.

Analisis kluster merupakan salah satu teknik dalam statistika yang digunakan untuk mengelompokkan variabel atau objek ke dalam beberapa kelompok (Soemartini & Supartini, 2017). Analisis kluster menemukan kelompok dalam kumpulan data dengan tujuan data yang berada pada satu kelompok mempunyai kemiripan yang dekat, dan memiliki perbedaan yang jelas dengan kelompok lain (Adnyani & Sihombing, 2021). Seiring dengan perkembangan teknologi dan kemajuan dalam pengumpulan data, selain pada data *cross section*, analisis kluster semakin berkembang dalam penerapannya pada bidang yang lain, salah satunya data *time series*. Data *time series* merupakan sekelompok data yang diperoleh dengan urutan pengamatan yang diambil secara berurutan dalam waktu yang diasumsikan memiliki struktur korelasi antara nilai data pengamatan setiap deret waktu (Riani & Sofro, 2023). Pengelompokan objek data yang mempertimbangkan data *time series* dikenal sebagai *time series-based clustering* (Setiawan & Zahra, 2023). Salah satu elemen kunci dalam *time series-based clustering* adalah pemilihan perhitungan jarak (Aghabozorgi dkk., 2015).

Data *time series* merupakan data yang berfluktuasi dan bervariasi dari waktu ke waktu sehingga dibutuhkan metode perhitungan jarak antar deret waktunya yang bersifat dinamis (Rizki dkk., 2021). Metode jarak yang dapat digunakan pada data yang dinamis adalah *dynamic time warping* (DTW). Metode ini merupakan salah satu metode perhitungan jarak yang memungkinkan untuk melihat perbandingan antar waktu dengan berbagai panjang periode yang berbeda. DTW menggunakan teknik pemrograman dinamis untuk menemukan semua jalur yang mungkin dan memilih salah satu jalur yang menghasilkan jarak minimum antara dua data *time series*, di mana jarak kumulatif setiap elemen matriks adalah nilai minimal dari tiga tetangga di

sekitarnya (Ayundari & Sutikno, 2019). Fleksibilitas DTW dalam menghitung jarak pada data yang dinamis baik dari segi temporal dan perbedaan dimensi data, membuatnya dapat diterapkan pada analisis kluster data *time series*.

Salah satu metode untuk melakukan analisis kluster adalah non-hirarki. Analisis kluster berbasis non-hirarki menghasilkan partisi dari data sehingga objek dalam kluster lebih mirip satu sama lain daripada objek yang ada dalam kluster lain (Triyanto, 2015). Di antara metode yang paling sering digunakan adalah *K-means* dan *K-medoids*. Kedua metode ini sering digunakan karena memiliki kemiripan konsep. Baik *K-means* maupun *K-medoids* melakukan partisi data menjadi beberapa k kluster yang telah ditentukan sebelumnya. Selain itu, keduanya melibatkan proses iteratif di mana titik data dikelompokkan ulang berdasarkan pusat kluster yang baru dihitung pada setiap iterasi. Tujuan dari kedua metode ini adalah untuk meminimalkan jumlah kesalahan pengelompokan dengan menempatkan titik data ke dalam kluster yang sesuai berdasarkan jarak ke pusat kluster. (Rodriguez dkk., 2019). Perbedaan utama antara kedua metode ini terletak pada representasi pusat kluster. *K-means* menggunakan *centroid* sebagai pusat kluster sehingga lebih sensitif terhadap data yang mengandung pencilan sementara *K-medoids* menggunakan *medoids* (titik data nyata) sehingga lebih *robust* terhadap pencilan (Larasati, 2017). Meskipun demikian, ketika menerapkan DTW sebagai metode jarak pada *K-means*, sebuah metode rata-rata yang lebih *robust* dapat diterapkan untuk memperbarui *centroid K-means*.

Perhitungan rata-rata untuk *time series* merupakan tantangan, terutama karena harus konsisten dengan kemampuan DTW untuk menyelaraskan *time series* tersebut. Oleh karena itu, metode pengelompokan *time series* lebih sering menggunakan algoritma *K-medoids* karena algoritma ini tidak terlalu bergantung pada rata-rata. Penggunaan DTW tanpa penyesuaian dapat menyebabkan *overfit*, yaitu keadaan dengan rata-rata yang dihasilkan terlalu terperinci dan tidak cukup sederhana untuk memberikan gambaran yang baik tentang data secara keseluruhan. Untuk menjawab tantangan ini, sebuah metode rata-rata global yang dapat diterapkan pada *K-means* diperkenalkan, DTW *barycenter averaging* (DBA). Metode ini merupakan perhitungan rata-rata yang terdiri dari penyempurnaan deret rata-rata awal secara iteratif untuk meminimalkan jarak DTW ke rata-rata *time series*. DBA lebih *robust* dalam menghasilkan rata-rata yang representatif karena mempertimbangkan urutan dan hubungan antar nilai pada *time series* sehingga lebih mampu menangkap dinamika dan pola yang ada dalam data (Petitjean dkk., 2011).

Beberapa penelitian mengenai DTW telah dilakukan. Di antaranya Ayundari dan Sutikno (2019) menggunakan metode *time series-based clustering* untuk menentukan zona musim di Mojokerto menurut karakteristik curah hujan. Hasil penelitian menunjukkan bahwa metode terbaik yang dipilih adalah pengelompokan dengan jarak DTW. Penelitian lain juga dilakukan oleh Adnyani dan Sihombing (2021) dan diperoleh hasil bahwa jarak DTW lebih baik dibandingkan jarak Euclidean pada pengelompokan data *time series*. Penelitian lain juga telah dilakukan oleh Aqsari (2023) dan diperoleh hasil bahwa metode *k-means* dengan jarak DTW lebih unggul dari jarak *Euclidean*. Penelitian lainnya juga pernah dilakukan oleh Radovanovic dkk (2020) dengan kesimpulan bahwa algoritma *K-medoids* dengan metode jarak DTW

lebih efektif dibandingkan metrik jarak *lock-step* seperti *Euclidean*, korelasi *Pearson* dan *Spearman*.

Berdasarkan uraian tersebut maka peneliti bermaksud untuk membandingkan kinerja algoritma *K-means* dengan DBA dan *K-medoids* pada *time series-based clustering* dengan metode jarak *dynamic time warping* (DTW) pada data jumlah penduduk miskin di Sulawesi Selatan tahun 2010-2023. Penelitian ini mengevaluasi hasil pengelompokan dan nilai koefisien *silhouette* dari kedua algoritma *K-means* dengan DBA dan *K-medoids*.

1.2 Batasan Penelitian

Batasan masalah pada penelitian ini adalah sebagai berikut:

1. Data yang digunakan adalah data jumlah penduduk miskin di Provinsi Sulawesi Selatan tahun 2010-2023
2. Metode pada perhitungan jarak lokal adalah jarak *Manhattan*
3. Penentuan jumlah kluster dilakukan dengan koefisien *silhouette*

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut:

1. Mengetahui hasil pengelompokan kota/kabupaten di Sulawesi Selatan berdasarkan jumlah penduduk miskin tahun 2010-2023 dengan algoritma *K-means* dan *K-medoids* dengan menggunakan DTW sebagai metode jarak.
2. Mengetahui hasil perbandingan kinerja algoritma *K-means* dan *K-medoids* dalam mengelompokkan kota/kabupaten di Sulawesi Selatan berdasarkan data *time series* jumlah penduduk miskin dengan menggunakan DTW sebagai metode jarak.

1.4 Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini adalah sebagai berikut:

1. Sebagai tambahan wawasan dan literatur dalam bidang analisis *time series*, khususnya dalam penerapan metode *Dynamic Time Warping* (DTW) pada algoritma *K-means* dan *K-medoids*.
2. Sebagai tambahan informasi kepada pemerintah dalam mengatasi penanggulangan kemiskinan yang lebih tepat sasaran.

1.5 Teori

1.5.1 Analisis Kluster

Analisis kluster adalah analisis yang bertujuan untuk mengelompokkan objek berdasarkan kemiripan karakteristik. Pada analisis kluster, tidak ada peubah respon sehingga data dari peubah peubah prediktor digunakan untuk mengelompokkan objek yang ada (Sirodj dkk., 2023). Objek yang termasuk ke dalam batas kemiripan dengan kelompoknya akan bergabung menjadi satu kelompok dan akan terpisah dengan kelompok yang berbeda jika keluar dari batas kemiripan kelompok tersebut. Kluster atau kelompok yang baik adalah yang memiliki homegenitas (kesamaan) yang tinggi antar anggota dalam satu kelompok dan heterogenitas (keberagaman) yang tinggi antar kelompok yang satu dengan kelompok yang lainnya. Kemiripan

antar objek yang satu dengan objek yang lainnya dalam pembentukan kelompok diperoleh dengan menggunakan ukuran jarak (*distance*) (Novidianto dkk., 2020).

Secara umum, ada empat jenis metode analisis kluster, yaitu metode non-hirarki, metode hirarki, metode *density-based* dan metode *grid-based*. Metode non-hirarki melakukan analisis kluster dengan pencarian kelompok yang terpisah secara langsung dan merelokasi titik data secara iteratif untuk memperbaiki kualitas kelompok dari langkah awal. Metode hirarki membentuk kelompok secara bertahap, yaitu melalui penggabungan kelompok yang kecil menjadi kelompok yang lebih besar, atau melalui pemecahan kelompok yang lebih besar menjadi kelompok yang lebih kecil. Untuk metode *density-based*, metode ini melakukan identifikasi kelompok dengan menghubungkan daerah-daerah yang memiliki titik data yang cukup banyak. Sedangkan pada metode *grid-based*, untuk meningkatkan efisiensi pengelompokannya, metode ini membagi ruang data menjadi sejumlah sel yang membentuk struktur *grid* yang rapat (terisi banyak) untuk membentuk kelompok (Bahauddin dkk., 2021).

1.5.2 Time Series Based Clustering

Time series based clustering adalah metode analisis kluster dengan mengelompokkan objek berdasarkan pola *time series* (Fransiska, 2021). Konsep dari metode ini adalah mengelompokkan berdasarkan kemiripan dari pola masing-masing *time series* sehingga objek yang memiliki kemiripan pola yang sama akan cenderung membentuk suatu himpunan (Zen dkk., 2023). Data *time series* merupakan data yang terdiri atas satu objek tetapi meliputi beberapa periode waktu misalnya data harian, mingguan, bulanan dan tahunan (Sumartini dkk., 2017). Data *time series* pada dasarnya digolongkan sebagai data dinamis karena nilai fiturnya berubah sebagai fungsi waktu, yang berarti bahwa nilai setiap titik deret waktu adalah satu atau lebih pengamatan yang dilakukan secara kronologis (Setiawan, 2018).

Definisi 1: Analisis kluster time series, diberikan dataset n data *time series* $D = \{F_1, F_2, F_3, \dots, F_n\}$ dilakukan partisi secara *unsupervised* D terhadap $C = \{C_1, C_2, C_3, \dots, C_k\}$ sedemikian rupa sehingga dalam kelompok *time series* homogen, pengelompokan dilakukan berdasarkan ukuran kesamaan tertentu. Selanjutnya, C_i disebut sebagai kluster, di mana $D = \bigcup_{i=1}^k C_i$ dan $C_i \cap C_j = \emptyset$ untuk setiap $i \neq j$.

Salah satu tantangan dalam *time series-based clustering* berkaitan dengan pengukuran kesamaan yang digunakan untuk membuat kluster atau kelompok. Untuk melakukan hal ini, diperlukan identifikasi *time series* yang serupa melalui proses perhitungan kesamaan di antara semua *time series* menggunakan metrik kesamaan tertentu. Proses ini disebut sebagai “*whole sequence matching* (pencocokan urutan lengkap)”, yaitu seluruh panjang *time series* dipertimbangkan saat menghitung jarak (Aghabozorgi et al., 2015).

Proses pengelompokan data *time series* memperhatikan sifat dinamis data *time series*. Penggunaan jarak dalam *time series based clustering* dibagi menjadi tiga kategori, yaitu (Rahkmawati & Annisa, 2023):

1. *Raw data* (data mentah), menghitung jarak berdasarkan data mentah

2. *Feature data*, menghitung jarak fitur data sebagai representasi karakteristik data
3. Parameter model, menghitung jarak koefisien model *time series*

1.5.3 Standarisasi Data

Standarisasi data merupakan proses yang sangat penting dalam analisis kluster. Nilai-nilai atribut data yang berbeda-beda rentangnya sering kali perlu dilakukan standarisasi agar proses pengelompokan tidak bias. Standarisasi akan membuat atribut memiliki bobot yang sama sehingga memperkecil perbedaan antar kelompok. Salah satu teknik standarisasi yang dapat digunakan adalah standarisasi *Z-Score* di mana metode tersebut sering disebut pula dengan *zero-mean* karena dalam prosesnya atribut diubah ke dalam bentuk data standar berdasarkan nilai rata-rata sama dengan nol serta deviasi standar sama dengan satu (Zen dkk., 2023).

Misalkan terdapat data *time series* $Z_t = Z_1, Z_2, Z_3, \dots, Z_n$ dan dilakukan proses standarisasi dengan menghitung nilai rata-rata dan standar deviasi per periode. Formula dari rata-rata untuk setiap periode j dituliskan pada Persamaan (1).

$$\mu_j = \left(\frac{1}{n}\right) \sum_{i=1}^n x_{ij} \quad (1)$$

Sedangkan formula untuk standar deviasi dituliskan pada Persamaan (2).

$$\sigma_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \mu_j)^2} \quad (2)$$

Standarisasi data runtun waktu Z_t dapat diperoleh dengan Persamaan (3)

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (3)$$

Keterangan:

x_{ij} adalah nilai dari objek ke i pada periode j

n adalah jumlah objek

μ_j adalah rata-rata dari nilai semua objek pada periode j

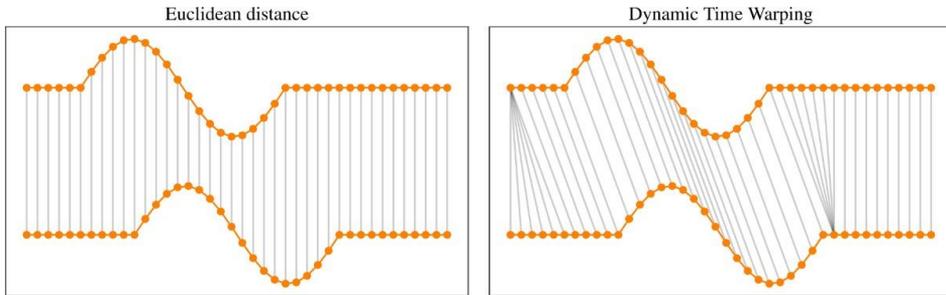
σ_j adalah standar deviasi dari semua objek pada periode j

z_{ij} adalah standarisasi data dari x_{ij}

1.5.4 Dynamic Time Warping

Dynamic time warping (DTW) adalah metode yang digunakan untuk mengukur jarak antara dua rangkaian waktu yang mungkin memiliki kecepatan atau skala yang berbeda (Puspita & Zulkarnain, 2020). DTW disebut juga sebagai *non-linear sequence alignment*, sehingga perhitungan jarak dengan DTW ini lebih realistis untuk

digunakan dalam kemiripan suatu pola, daripada hanya menggunakan jarak linear seperti *Euclidean*, *Manhattan*, *Canberra* dan lainnya (Aqsari, 2023).



Gambar 1. Ilustrasi Pengaplikasian antara Jarak Euclidean dan DTW (Sumber: Rakhmawati & Annisa, 2023)

Gambar 1 menunjukkan bahwa jarak *Euclidean* menghitung jarak berdasarkan pada waktu yang sama sehingga ilustrasi jaraknya terlihat garis-garis lurus. Berdasarkan keadaan tersebut maka jarak *euclidean* disebut sebagai jarak linear. Sedangkan jarak DTW menghitung jarak optimum dari dua data *time series* sehingga observasi ke- i pada data *time series* 1 belum tentu berpasangan dengan observasi ke- i dari data *time series* 2 (Aqsari, 2023). Konsep utama dari DTW merupakan proses yang digunakan antara data *time series* untuk menemukan *warping path* (kombinasi jalur) optimal dengan panjang yang berbeda. Proses DTW menerapkan teknik pemrograman yang disusun dalam matriks jarak untuk memperoleh *warping path* yang mungkin dan akan dipilih salah satu jalur yang mempunyai jarak terkecil antara dua rangkaian data *time series*. Elemen yang terisi dalam matriks jarak adalah jarak kumulatif dari minimal tiga tetangga terdekatnya (Zen dkk., 2023).

Misalkan terdapat dua deret waktu $Z(1)_t = Z_1, Z_2, Z_3, \dots, Z_n$ dan $Z(2)_t = Z_1, Z_2, Z_3, \dots, Z_m$ di mana $i = 1, 2, 3, \dots, n$ dan $j = 1, 2, 3, \dots, m$ akan membentuk sebuah matriks D berukuran $n \times m$ dengan elemen ke- i, j adalah selisih antara $Z(1)_i$ dengan $Z(2)_j$, yang dituliskan pada Persamaan (4) sebagai berikut (Novidianto & Dani, 2020):

$$w_{i,j} = |Z(1)_i - Z(2)_j| \quad (4)$$

Setelah mendapatkan jarak kumulatif berupa $w_{i,j}$, kemudian ditambahkan nilai minimum dari tiga elemen yang berdekatan dengan elemen (i, j) yaitu

$$\min \begin{cases} d_{(i-1)(j-1)} \\ d_{(i-1)j} \\ d_{i(j-1)} \end{cases}$$

di mana $0 < i \leq n$ dan $0 < j \leq m$ sehingga terbentuk matriks D sebagai berikut:

$$d_{i,j} = w_{ij} + \min \begin{cases} d_{(i-1)(j-1)} \\ d_{(i-1),j} \\ d_{i(j-1)} \end{cases} \quad (5)$$

Setelah mendapatkan matriks D , maka selanjutnya adalah menentukan jarak DTW antara dua data *time series* $Z(1)_t$ dan $Z(2)_t$ dengan Persamaan (9) sebagai berikut (Riani & Sofro, 2023):

$$d_{DTW}(Z(1)_t, Z(2)_t) = \min_{\forall w \in P} \left\{ \sum_{i,j=1}^k d_{i,j} \right\} \quad (6)$$

Keterangan:

p adalah kemungkinan dari sekumpulan semua *warping path*

$d_{i,j}$ adalah elemen (i, j) pada matriks D

k adalah panjang dari *warping path*

Ada beberapa kondisi yang harus dipenuhi untuk memastikan perhitungan jarak antar dua *time series* berjalan dengan benar. Berikut adalah beberapa batasan yang dipertimbangkan untuk mengurangi jalur yang akan ditentukan selama proses penyelarasan data *time series* (Spiegel, 2015):

1. *Boundary conditions*

Jalur harus dimulai pada $w_1 = (1,1)$ dan berakhir pada $w_k = (n, m)$, yaitu *warping path* harus dimulai dari elemen matriks $(1,1)$ hingga elemen matriks terakhir (n, m) . Hal ini untuk memastikan jalur *warping path* dimulai dari data awal hingga akhir.

2. *Continuity conditions*

Langkah-langkah dalam *warping path* hanya boleh menuju sel yang berdekatan, termasuk sel yang berdekatan secara diagonal. Artinya, jika berada di suatu titik (i, j) di matriks, titik sebelumnya bisa jadi adalah $(i - 1, j - 1)$, $(i - 1, j)$ atau $(i, j - 1)$. Hal ini untuk memastikan setiap langkah hanya bisa ke sel yang berdekatan.

3. *Monotonic conditions*

Titik-titik dalam *warping path* harus berurutan dalam waktu. Ini berarti jika berada di titik (i, j) , titik sebelumnya (i', j') harus memenuhi kondisi bahwa $i' \leq i$ dan $j' \leq j$, sehingga jalur tidak pernah bergerak mundur.

Untuk kejelasan lebih lanjut, berikut adalah langkah-langkah pada DTW (Aqsari, 2023):

1. Membentuk matriks dengan ukuran $n \times n$

Sebagai contoh jika memiliki data *time series* 1 yaitu 1,3,5,2,7,1 dan *time series* 2 yaitu 2,5,2,5,3,6. Maka matriks yang terbentuk adalah sebagai berikut:

6						
3						

5						
2						
5						
2						
	1	3	5	2	7	1

2. Mengisi matriks tersebut dengan tiga rumus yang ditunjukkan pada Persamaan (5) sehingga nilai dari matriks tersebut adalah sebagai berikut:

6	17	9	7	9	6	11
3	12	6	6	5	8	6
5	10	6	4	5	4	8
2	6	4	5	2	7	8
5	5	3	2	5	7	11
2	1	2	5	5	10	11
	1	3	5	2	7	1

3. Menentukan *warping path* yaitu dilakukan dengan cara menandai elemen matriks yang paling akhir yaitu titik a , kemudian berjalan sampai dengan titik b . Jalannya *warping path* yaitu memilih angka yang minimum dari sekitar titik. Sebagai contoh titik a , 11, angka yang berada di sekitarnya adalah 6,6 dan 8. Nilai terkecil dari ketiganya adalah 6 sehingga *warping path* berjalan dari titik 11 ke titik 6.

					a	
6	17	9	7	9	6	11
3	12	6	6	5	8	6
5	10	6	4	5	4	8
2	6	4	5	2	7	8
5	5	3	2	5	7	11
2	1	2	5	5	10	11
b	1	3	5	2	7	1

1.5.5 Algoritma *K-Means*

K-means merupakan metode pengelompokan yang paling banyak digunakan diberbagai bidang karena sederhana dan mudah diimplementasikan. *K-means* merupakan metode pengklasteran secara non-hirarki yang memisahkan data ke dalam kelompok yang berbeda. Tujuan dari pengklasteran data ini adalah untuk meminimalisasikan fungsi objektif dalam proses pengelompokkan, yang pada umumnya berusaha meminimalisasikan ragam di dalam suatu kelompok dan memaksimalkan ragam antar kelompok (Soemartini & Supartini, 2017). Ada dua hal yang harus ditentukan terlebih dahulu jika menggunakan analisis kluster *partitioning* yaitu jumlah kelompok dan pemilihan pusat kelompok atau *centroid* (Aqsari, 2023). Langkah-langkah algoritma *K-means* adalah sebagai berikut (Bangun, 2016):

1. Menentukan k sebagai jumlah kluster yang ingin dibentuk
2. Bangkitkan k *centroid* awal secara random. Penentuan *centroid* awal dilakukan secara random/acak dari objek-objek yang tersedia sebanyak k kluster, kemudian untuk menghitung *centroid* kluster ke- i berikutnya, digunakan persamaan sebagai berikut:

$$v = \frac{\sum_{i=1}^n x_i}{n} \quad (7)$$

Keterangan:

v adalah *centroid* pada kluster

x_i adalah objek ke- i , $i = 1, 2, 3, \dots, n$

n adalah banyaknya objek yang menjadi anggota kluster

3. Hitung jarak setiap objek ke masing-masing *centroid*
4. Kelompokkan setiap data berdasarkan jarak terdekat antara data dengan *centroid*-nya
5. Tentukan posisi *centroid* baru C_k dengan cara menghitung nilai rata-rata dari data-data yang ada pada *centroid* yang sama dengan persamaan sebagai berikut:

$$C_k = \left(\frac{1}{n_k}\right) \sum d_i \quad (8)$$

Keterangan:

n_k adalah jumlah anggota dalam kluster k

d_i adalah anggota dalam kluster k

Langkah 3 memerlukan cara untuk menghitung *mean* dari sekumpulan *time series*. Sebuah metode disebut sebagai DTW *barycenter averaging* (DBA). Metode ini merupakan metode yang menghitung rata-rata dari sekumpulan *time series* dengan menggunakan teknik DTW. Dengan menggunakan jarak DTW untuk membandingkan setiap data *time series* dengan rata-rata, dimungkinkan untuk menghubungkan titik-titik data *time series* ke setiap titik rata-rata dan menghitung *barycenter* semua titik dari barisan yang terhubung dengan setiap titik dari rata-rata sehingga menghasilkan lintasan rata-rata baru (Palaude & Viéville, 2023).

Tujuan DBA adalah untuk meminimalkan jumlah kuadrat jarak DTW dari barisan rata-rata ke himpunan barisan. Jumlah tersebut dibentuk oleh jarak tunggal antara setiap koordinat rata-rata dan koordinat barisan yang saling terhubung. Prinsip DBA adalah menghitung setiap koordinat barisan rata-rata sebagai *barycenter* dari koordinat-koordinat terkait dari himpunan barisan tersebut (Anh & Thanh, 2015).

Diberikan fungsi *barycenter* (Palaude & Viéville, 2023):

$$Barycenter(\{X_1, X_2, \dots, X_l\}) = \frac{X_1 + X_2 + \dots + X_l}{l} \quad (9)$$

Jika setiap titik dari himpunan merupakan vektor berdimensi n , tanda penjumlahan sesuai dengan penjumlahan vektor. Diberikan S , deret i yang panjangnya m yang dapat dituliskan sebagai:

$$S_i = s_{i1}s_{i2} \dots s_{im} \quad (10)$$

Di mana setiap titik adalah sebuah vektor di ruang n dimensi:

$$s_{ij} = (s_{ij1}, s_{ij2}, \dots, s_{ijn}) \quad (11)$$

Barisan rata-rata A yang panjangnya m' dituliskan:

$$A = a_1 a_2 \dots a_{m'} \quad (12)$$

Di mana

$$a_j = (a_{j1} a_{j2} \dots a_{jn}) \quad (13)$$

Setiap titik di A akan memiliki kumpulan titik dari deret S yang terhubung saat menerapkan DTW antara A dan deret dari S . Ruang yang terbentuk yaitu:

$$Assoc(a_j)$$

Algoritma DBA menghitung rata-rata setiap titik A secara individual dengan menggantinya dengan *barycenter* dari semua titik yang terhubung. Dengan kata lain, pada setiap iterasi algoritma DBA, setiap titik a pada deret A senantiasa diperbarui:

$$a_j = \text{Barycenter}(Assoc(a_j)) \quad (14)$$

1.5.6 Algoritma K-Medoids

Algoritma *K-medoids* atau yang juga dikenal sebagai *partitioning around medoids* (PAM) merupakan salah satu metode pengelompokan yang titik pusatnya menggunakan objek representatif (*medoids*) (Zen dkk., 2023). Dalam algoritma ini, data yang terdiri dari n objek akan dipartisi menjadi k kluster di mana jumlah $k \leq n$. *Medoids* adalah objek yang dianggap mewakili kluster sekaligus sebagai pusat kluster. Algoritma ini membentuk suatu kluster dengan menghitung suatu kluster dengan cara menghitung jarak kemiripan yang dimiliki antara *medoids* dengan objek *non-medoids* (Dewi & Pramita., 2019). Algoritma *K-medoids* adalah sebagai berikut (Sihombing dkk., 2019):

1. Pilih sembarang k objek dari n objek yang ada sebagai inialisasi *medoid*. Misalkan setiap *medoids* sebagai O_i .
2. Hitung jarak setiap objek terhadap setiap *medoid* dan tempatkan setiap objek ke kluster yang terdekat dengan *medoid*-nya.
3. Pilih sembarang objek *non-medoid* sebagai O_{random} .
4. Hitung total jumlah jarak terdekat dan S dari pertukaran *medoid* O_i dan O_{random} yang dinyatakan dengan Persamaan (15) berikut:

$$\text{Jumlah Jarak Terdekat} = \sum d_{ij} \quad (15)$$

di mana d_{ij} adalah jarak terdekat *non-medoid* dengan *medoid*-nya

5. Jika $S < 0$, maka tukar O_i dengan O_{random} untuk membentuk diri sebagai *medoid* baru. Nilai S dinyatakan dengan Persamaan (16) berikut:

$$S = \text{Jumlah jarak terdekat baru} - \text{jumlah jarak terdekat lama} \quad (16)$$

Jumlah jarak terdekat baru = jumlah jarak terdekat *non-medoid*
 Jumlah jarak terdekat lama = jumlah jarak terdekat *medoid*.

6. Ulangi langkah 2-5 sampai tidak ada perubahan

1.5.7 Koefisien Silhouette

Koefisien *silhouette* digunakan untuk melihat kualitas dan kekuatan kluster, seberapa baik atau buruknya suatu objek ditempatkan dalam suatu kluster (Dewi & Pramita, 2019). Langkah-langkah perhitungan koefisien *silhouette* adalah sebagai berikut (Riani & Sofro, 2023):

1. Menghitung a_i^j yaitu jarak rata-rata data ke- i dengan semua data di kluster yang sama. Untuk menghitung a_i^j dapat dilihat pada Persamaan (17)

$$a_i^j = \frac{1}{m_j - 1} \sum_{\substack{r=1 \\ r \neq i}}^{m_j} d(x_i^j, x_r^j) \quad (17)$$

Keterangan:

j adalah kluster

i adalah index data ($i = 1, 2, \dots, m_j$)

a_i^j adalah rata-rata jarak data ke- i dengan semua data di kluster yang sama

m_j adalah jumlah data dalam kluster ke- j

$d(x_i^j, x_r^j)$ adalah jarak data ke- i dengan data ke- r dalam satu kluster j

2. Menghitung nilai b_i^j yaitu nilai minimum dari jarak rata-rata data ke- i dengan semua data di kluster berbeda. Formula b_i^j dituliskan pada Persamaan (18)

$$b_i^j = \min_{\substack{n=1, \dots, k \\ n \neq j}} \left\{ \frac{1}{m_n} \sum_{\substack{r=1 \\ r \neq i}}^{m_n} d(x_i^j, x_r^n) \right\} \quad (18)$$

Keterangan

j adalah kluster

i adalah index data ($i = 1, 2, \dots, m_j$)

b_i^j adalah rata-rata jarak data ke- i dengan semua data di kluster yang berbeda

m_j adalah jumlah data dalam kluster ke- j

$d(x_i^j, x_r^n)$ adalah jarak data ke- i dengan data ke- r dalam satu kluster n

3. Menghitung SI_i^j yaitu *Silhouette Index* data ke- i yang dituliskan pada persamaan (19)

$$SI_i^j = \frac{b_i^j - a_i^j}{\max\{a_i^j, b_i^j\}} \quad (19)$$

Keterangan:

SI_i^j adalah *Silhouette Index* data ke- i dalam satu kluster

a_i^j adalah rata-rata jarak data ke- i dengan semua data di kluster yang sama

b_i^j adalah rata-rata jarak data ke- i dengan semua data di kluster yang berbeda

Nilai koefisien *silhouette* dari setiap objek dalam suatu kluster merupakan suatu ukuran yang menunjukkan seberapa dekat kemiripan data dikelompokkan dalam satu kluster tersebut. Nilai SI_i^j berada pada rentang:

$$-1 < SI_i^j \leq 1$$

Nilai SI_i^j mendekati -1 menunjukkan bahwa jarak antar objek dalam a_i^j lebih besar dibandingkan b_i^j , sehingga dikatakan bahwa terjadi salah pengelompokan atau keragu-raguan dalam pengelompokan yang dilakukan. Sedangkan jika nilai SI_i^j mendekati 1 menunjukkan bahwa jarak antar objek dalam a_i^j jauh lebih kecil dibandingkan b_i^j , sehingga dikatakan pengelompokan dilakukan dengan baik.

4. Menghitung SI_j yaitu rata-rata *silhouette index* kluster j dengan Persamaan (20) sebagai berikut:

$$SI_j = \frac{1}{m_j} \sum_{i=1}^{m_j} SI_i^j \quad (20)$$

5. Selanjutnya menghitung nilai SI global dengan Persamaan (21) berikut (Fransiska, 2021):

$$SI = \frac{1}{I} \sum_{i=1}^I SI_i^j \quad (21)$$

dengan SI adalah rata-rata *silhouette index* dari dataset.

Langkah terakhir yaitu menentukan koefisien *silhouette* (SC) yang diperoleh dengan mencari nilai maksimal dari *Silhouette Index Global* dari jumlah kluster 2 sampai jumlah kluster ke- $q - 1$. Formula koefisien *silhouette* dapat dituliskan pada persamaan (22) sebagai berikut:

$$SC = \max_k SI(k) \quad (22)$$

Keterangan:

SC adalah koefisien *silhouette*

SI adalah *Silhouette Index Global*

k adalah kluster ke- k ($k = 2, 3, \dots, q - 1$) dengan q adalah jumlah kluster

Kriteria koefisien *silhouette* yang ditetapkan oleh Kaufman & Rousseeuw (1990) disajikan pada Tabel 1 sebagai berikut:

Tabel 1. Kriteria Koefisien Silhouette

Nilai Koefisien <i>Silhouette</i>	Kriteria Kluster
0,71 – 1,00	<i>Strong</i>
0,51 – 0,70	<i>Good</i>
0,26 – 0,50	<i>Weak</i>
0,00 – 0,25	<i>Bad</i>

1.5.8 Jumlah Penduduk Miskin

Secara umum kemiskinan didefinisikan sebagai kondisi di mana seseorang atau sekelompok orang tidak mampu memenuhi hak-hak dasarnya untuk mempertahankan dan mengembangkan kehidupan yang bermartabat. Definisi yang sangat luas ini menunjukkan bahwa kemiskinan merupakan masalah multidimensional, sehingga tidak mudah untuk mengukur kemiskinan dan perlu kesepakatan pendekatan pengukuran yang dipakai (Badan Pusat Statistik (BPS), 2011). Indikator kemiskinan bisa dilihat dari beberapa sisi, yaitu persentase penduduk miskin, pendidikan (khususnya angka buta huruf), kesehatan (angka kematian bayi dan anak balita kurang gizi), ketenagakerjaan dan ekonomi (konsumsi per kapita). Diperlukan tolak ukur yang jelas untuk menentukan seseorang dapat dikatakan miskin atau tidak. Berbagai pendekatan atau konsep digunakan sebagai bahan perhitungan dan penentuan batas kemiskinan (Mahendra, 2016).

Untuk mengukur kemiskinan, BPS menggunakan konsep kemampuan memenuhi kebutuhan dasar (*basic need approach*). Dengan pendekatan ini, kemiskinan dipandang sebagai ketidakmampuan dari sisi ekonomi untuk memenuhi kebutuhan dasar makanan dan bukan makanan yang diukur menurut garis kemiskinan (Badan Pusat Statistik (BPS), 2023). Garis kemiskinan merupakan tingkat minimum pendapatan yang dianggap perlu dipenuhi untuk memperoleh standar hidup yang mencukupi di suatu negara. Garis kemiskinan digunakan sebagai perangkat ekonomi untuk mengukur rakyat miskin dan mempertimbangkan pembaharuan sosio-ekonomi, misalnya seperti program peningkatan kesejahteraan dan asuransi pengangguran untuk menanggulangi kemiskinan (Aprilia & Sembiring, 2021). Garis kemiskinan terdiri dari dua komponen, yaitu Garis Kemiskinan Makanan (GKM) dan Garis Kemiskinan Bukan Makanan (GBKM). GKM merupakan nilai pengeluaran kebutuhan minimum makanan yang disetarakan dengan 2100 kkalori per kapita per hari. GBKM merupakan kebutuhan minimum untuk perumahan, sandang, pendidikan dan kesehatan (Badan Pusat Statistik (BPS), 2023).

Menurut data BPS per Maret 2023, persentase penduduk miskin di Indonesia sebesar 9,36 persen, menurun 0,21 persen terhadap September 2022. Jumlah total penduduk miskin yang tercatat sebesar 25,90 juta jiwa. Garis Kemiskinan tercatat sebesar Rp. 550.048,-/kapita/bulan dengan komposisi Garis Kemiskinan Makanan sebesar Rp. 408.522,- (74,21 persen) dan Garis Kemiskinan Bukan Makanan sebesar Rp. 141.936,- (25,79 persen). Faktor-faktor yang berpengaruh terhadap tingkat kemiskinan adalah Tingkat Pengangguran Terbuka (TPT), Nilai Tukar Petani (NTP), laju inflasi dan konsumsi rumah tangga.

BAB II METODE PENELITIAN

2.1 Sumber Data

Data yang digunakan adalah data jumlah penduduk miskin dalam ribu jiwa di 24 kabupaten/kota di Sulawesi Selatan. Data ini merupakan data sekunder yang dikumpulkan dari *website* Badan Pusat Statistik (BPS) Sulawesi Selatan pada situs <https://sulsel.bps.go.id>. Data ini mencakup periode tahunan dari tahun 2010 hingga 2023. Jumlah data yang didapatkan untuk 14 periode dari 24 kabupaten/kota adalah 336 observasi sebagaimana terlampir dalam Lampiran 1.

2.2 Variabel Penelitian

Variabel yang digunakan dalam penelitian adalah jumlah penduduk miskin yang merujuk pada jumlah individu yang hidup di bawah garis kemiskinan di setiap kabupaten/kota di Sulawesi Selatan dalam rentang waktu dari tahun 2010 hingga 2023.

2.3 Struktur Data Penelitian

Terdapat 24 kabupaten yang selanjutnya digunakan sebagai banyaknya observasi. Variabel yang digunakan adalah tahun jumlah penduduk miskin diobservasi yaitu sejak tahun 2010 sampai dengan 2023. Struktur data penelitian ditunjukkan pada Tabel 2.

Tabel 2. Struktur Data Penelitian

Kabupaten/ Kota	Jumlah Penduduk Miskin (Ribu Jiwa)							
	2010	2011	2012	2013	2014	2015	...	2023
1	$Z_{1,1}$	$Z_{2,1}$	$Z_{3,1}$	$Z_{4,1}$	$Z_{5,1}$	$Z_{6,1}$...	$Z_{14,1}$
2	$Z_{1,2}$	$Z_{2,2}$	$Z_{3,2}$	$Z_{4,2}$	$Z_{5,2}$	$Z_{6,2}$...	$Z_{14,2}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
24	$Z_{1,24}$	$Z_{2,24}$	$Z_{3,24}$	$Z_{4,24}$	$Z_{5,24}$	$Z_{6,24}$...	$Z_{14,24}$

2.4 Metode Analisis

Dalam penelitian ini, pengolahan data dilakukan dengan bantuan *software R-Studio*. Berikut adalah langkah-langkah analisis data yang dilakukan dalam penelitian ini:

1. Mendeskripsikan data jumlah penduduk miskin dengan:
 - a. Menghitung statistik deskriptif *mean* untuk melihat karakteristik dari jumlah penduduk miskin di Sulawesi Selatan
 - b. Membuat plot *time series* untuk melihat tren jumlah penduduk miskin dari tahun ke tahun tiap kabupaten
 - c. Memvisualisasikan *mean* jumlah penduduk miskin per kabupaten
2. Mendeteksi pencilon pada data *time series* dengan metode MMCD
3. Menstandarisasi data dengan menggunakan *Z – Score* menggunakan Persamaan (3)

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (3)$$

4. Menghitung jarak data *time series* antar kabupaten/kota dengan pendekatan *Dynamic Time Warping* (DTW) dengan Persamaan (6) dan prosedur sebagai berikut:

$$d_{DTW}(Z(1)_t, Z(2)_t) = \min_{vw \in P} \left\{ \sum_{i,j=1}^k d_{i,j} \right\} \quad (6)$$

- Membentuk matriks **D** berukuran 14×14 (jumlah titik data *time series* tiap kabupaten)
- Mengisi matriks **D** dengan jarak lokal untuk setiap pasangan tiap titik dengan Persamaan (4):

$$w_{i,j} = |Z(1)_i - Z(2)_j| \quad (4)$$

- Menjumlahkan w_{ij} dengan nilai minimum dari tiga tetangga terdekat dengan menggunakan Persamaan (5):

$$d_{i,j} = w_{ij} + \min \begin{cases} d_{(i-1)(j-1)} \\ d_{(i-1),j} \\ d_{i(j-1)} \end{cases} \quad (5)$$

- Menganalisa *warping path* yang terbentuk pada matriks **D**
- Ulangi langkah 4.a sampai 4.e untuk menghitung jarak DTW antar dua kabupaten selanjutnya
- Mengisi matriks jarak DTW berukuran 24×24 (jumlah kabupaten yang diobservasi).

5. Melakukan pengelompokan dengan algoritma *K-means* dengan DBA dengan prosedur sebagai berikut.

- Menentukan jumlah k optimal dengan koefisien *silhouette*
- Memilih k *centroid* awal secara acak dari data
- Menghitung jarak tiap objek ke masing-masing *centroid* menggunakan jarak DTW yang diperoleh pada langkah 4.
- Mengelompokkan tiap data berdasarkan jarak terdekat antara objek dengan *centroid*-nya
- Tentukan *centroid* baru dengan Persamaan (14)

$$a_j = \text{Barycenter}(\text{Assoc}(a_j)) \quad (14)$$

- Hitung jarak DTW tiap objek dengan *centroid* baru
- Ulangi langkah d dan e hingga tidak ada perubahan signifikan pada pengelompokan

6. Melakukan pengelompokan dengan algoritma *K-medoids* dengan prosedur sebagai berikut.

- Menentukan jumlah k optimal dengan koefisien *silhouette*
- Memilih k *medoids* awal secara acak dari data
- Menghitung jarak tiap objek ke masing-masing *medoid* menggunakan jarak DTW yang diperoleh pada langkah 4.

- d. Mengelompokkan tiap data berdasarkan jarak terdekat antara objek dengan *medoids*-nya
 - e. Secara acak memilih k *medoids* baru dari data
 - f. Ulangi langkah c hingga e hingga tidak ada perubahan signifikan pada pengelompokan
7. Membandingkan hasil pengelompokan *K-means* dan *K-medoids* dengan koefisien *silhouette*.
 8. Membentuk plot data pada masing-masing kluster
 9. Menganalisis karakteristik masing-masing kluster menggunakan nilai *mean*.