P-ISSN: 2986-4178
E-ISSN: 2988-4004

Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025

Pengelompokan Provinsi di Indonesia Berdasarkan Tingkat Pengangguran Tahun 2023 Menggunakan *K-Medoids*

Salma Soesmono[1*], Riezki Pertiwi[1], Bening Saputri[1], Naomighina Putri[1], Edy Widodo[1]

[1] Program Studi Statistika, Universitas Islam Indonesia, Jl. Kaliurang KM 14,5, Kabupaten Sleman Daerah Istimewa Yogyakarta, 55584, Indonesia

*Corresponding author: 22611023@students.uii.ac.id

P-ISSN: 2986-4178
E-ISSN: 2988-4004

Riwayat Artikel

Dikirim: 19 September 2024
Direvisi: 19 Januari 2025
Diterima: 27 Januari 2025

ABSTRAK

Pengangguran di Indonesia mencapai 5.32% pada tahun 2023, lebih rendah dibandingkan tahun sebelumnya, namun masih belum mencapai target RPJM 2020-2024 yang berkisar antara 3.6-4.3%. Penelitian ini bertujuan untuk mengelompokkan provinsi-provinsi di Indonesia berdasarkan tingkat pengangguran menggunakan metode *K-Medoids*, serta menganalisis karakteristik pengangguran di setiap kelompok provinsi. Data yang digunakan mencakup tingkat pengangguran, jumlah angkatan kerja, dan indikator ekonomi lainnya yang diambil dari Badan Pusat Statistik (BPS). Analisis dilakukan dengan menggunakan *K-Medoids clustering* untuk mengidentifikasi kelompok provinsi yang memiliki pola pengangguran serupa. Hasil penelitian menunjukkan bahwa terdapat tiga *cluster* provinsi: *Cluster* 1 (24 provinsi) dengan pengangguran rendah, *Cluster* 2 (7 provinsi) dengan pengangguran sedang, dan *Cluster* 3 (3 provinsi) dengan pengangguran tinggi. Temuan ini diharapkan dapat menjadi dasar bagi kebijakan ketenagakerjaan berbasis data yang lebih tepat sasaran dalam mengurangi pengangguran dan mendukung pencapaian Visi Indonesia Emas 2045.

Kata Kunci: Analisis Komponen Utama, *K-Medoids*, Kebijakan Berbasis Data, Pengangguran, Pengelompokan Provinsi.

ABSTRACT

Unemployment in Indonesia reached 5.32% in 2023, lower than the previous year, but still has not reached the 2020-2024 RPJM target of 3.6-4.3%. This study aims to group provinces in Indonesia based on the unemployment rate using the *K-Medoids* method, and analyze the characteristics of unemployment in each provincial group. The data used includes unemployment rate, total labor force, and other economic indicators taken from the Central Bureau of Statistics (BPS). The analysis was conducted using *K-Medoids* clustering to identify provincial groups that have similar unemployment patterns. The results show that there are three provincial clusters: Cluster 1 (24 provinces) with low unemployment, Cluster 2 (7 provinces) with moderate unemployment, and Cluster 3 (3 provinces) with high unemployment. These findings are expected to be the basis for a more targeted data-based employment policy in reducing unemployment and supporting the achievement of the Golden Indonesia Vision 2045.

Keywords: Principal Component Analysis, *K-Medoids*, Data-Driven Policy, Unemployment, Provincial Clustering

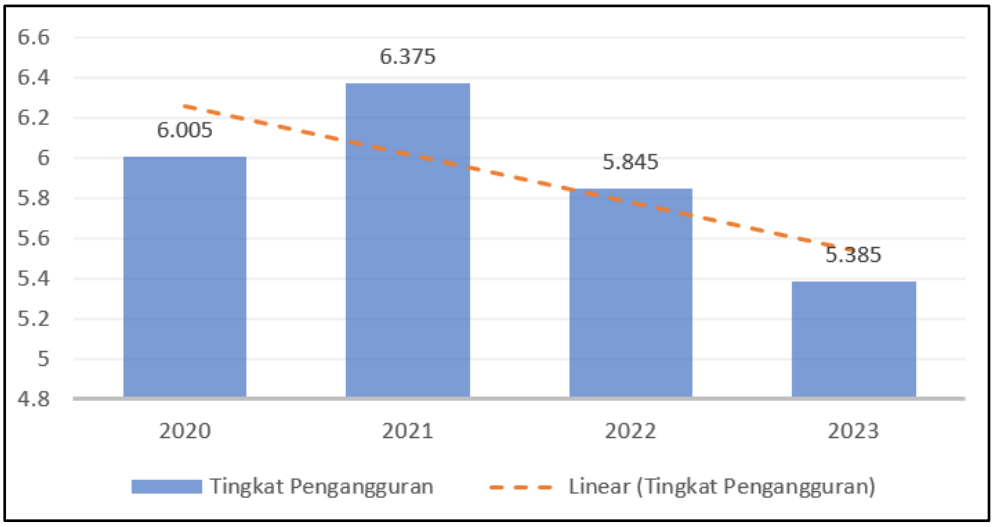
P-ISSN: 2986-4178
E-ISSN: 2988-4004

Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025

1. Pendahuluan

Menurut Badan Pusat Statistik, pengangguran merujuk pada individu yang tidak memiliki pekerjaan namun sedang aktif mencari pekerjaan, mempersiapkan usaha baru, atau telah diterima bekerja tetapi belum memulai tugasnya. Peristiwa ini umumnya terjadi akibat ketidakseimbangan antara jumlah lapangan kerja yang tersedia dan jumlah tenaga kerja yang membutuhkan pekerjaan. Pengangguran berdampak pada penurunan pendapatan dan produktivitas masyarakat, sehingga dapat memunculkan berbagai permasalahan ekonomi dan sosial, seperti kemiskinan dan kesenjangan sosial. Kondisi ini menjadi tantangan signifikan dalam mendukung tercapainya visi Indonesia Emas 2045 [1].

Salah satu tujuan utama dari visi Indonesia Emas 2045 adalah mengurangi kemiskinan dan ketimpangan sosial. Visi ke-2 Indonesia Emas 2045 menargetkan penurunan tingkat kemiskinan menjadi 0.5-0.8 persen serta mengurangi ketimpangan pendapatan antar wilayah. Untuk mencapai tujuan ini, kebijakan yang berbasis data sangat dibutuhkan, termasuk untuk mengatasi pengangguran dan meningkatkan kapasitas angkatan kerja di berbagai wilayah.



Gambar 1. Tingkat Pengangguran Indonesia Tahun 2020-2023

Gambar 1. adalah grafik tingkat pengangguran di Indonesia yang diambil dari Trading Economics. Berdasarkan data di atas, tingkat pengangguran di Indonesia menunjukkan tren penurunan selama tiga tahun terakhir dan mencapai persentase terendah pada tahun 2023 yaitu sebesar 5.385%. Namun, kondisi tersebut belum mencapai target RPJM 2020-2024 yang seharusnya berada di kisaran 3.6-4.3%. Selain itu, kesenjangan antar provinsi juga perlu diperhatikan, karena beberapa daerah tetap mengalami tingkat pengangguran yang lebih tinggi dibandingkan rata-rata nasional.

Belum tercapainya target yang ada, tingginya tingkat pengangguran di Indonesia menunjukkan diperlukannya kebijakan yang lebih spesifik dan tepat sasaran. Kegagalan menurunkan tingkat pengangguran hingga mencapai target RPJM tidak hanya menjadi tantangan dalam jangka pendek, tetapi juga dapat menghambat pencapaian visi Indonesia Emas 2045, mengingat adanya hubungan erat antara pengangguran, kemiskinan, dan ketimpangan sosial. Salah satu langkah strategis yang dapat dilakukan adalah dengan mengelompokkan provinsi-provinsi di Indonesia berdasarkan indikator tingkat pengangguran terbuka [2].

Dengan mengelompokkan wilayah berdasarkan indikator tingkat pengangguran, sebagaimana disebutkan di atas dari penelitian ini, diharapkan dapat digunakan untuk mengidentifikasi pola-pola spasial dan karakteristik sosio-ekonomi yang terkait dengan

P-ISSN: 2986-4178
E-ISSN: 2988-4004

Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025

tingkat pengangguran. Salah satu metode pengelompokan yang digunakan dalam penelitian ini adalah *K-Medoids*.

K-Medoids adalah metode yang efisien dalam mengelompokkan data berdasarkan kesamaan karakteristik, menjadikannya ideal untuk mengidentifikasi kelompok wilayah dengan tingkat pengangguran yang serupa secara cepat dan akurat. Metode ini menggunakan medoid sebagai *centroid cluster*, yang membuatnya lebih tahan terhadap *outlier* dibandingkan algoritma K-Means [3].

Untuk meningkatkan akurasi dalam proses pengelompokan, penelitian ini memanfaatkan *Principal Component Analysis* (PCA). PCA merupakan teknik statistik yang bertujuan untuk mereduksi jumlah variabel dalam data tanpa menghilangkan informasi penting. Dengan metode ini, data yang memiliki banyak variabel dapat disederhanakan menjadi beberapa komponen utama yang mencerminkan sebagian besar variasi dalam data. Penggunaan PCA membantu membuat analisis lebih sederhana, mengurangi tumpang tindih antar variabel, serta mengatasi masalah multikolinearitas yang mungkin terjadi [4].

Sudah banyak penelitian terkait pengangguran, seperti yang dilakukan oleh Purba dan Karim et al. (2024) [5]; Putri (2024) [6]; Detrinidad dan López-Ruiz (2024) [7]; Almeira et al. (2021) [8]; serta Simanjuntak (2023) [9]. Sebagian besar penelitian tersebut hanya menggunakan metode *K-Medoids clustering* saja atau *Principal Component Analysis* (PCA) saja tanpa menggabungkan kedua metode.

Namun, penelitian-penelitian tersebut belum mengkaji lebih jauh tentang bagaimana integrasi antara metode *K-Medoids* dan PCA dapat memberikan hasil yang lebih optimal, terutama dalam mengelompokkan wilayah dengan tingkat pengangguran yang kompleks dan memiliki *outlier*. Padahal, kombinasi kedua metode ini berpotensi meningkatkan akurasi analisis, di mana PCA dapat mereduksi dimensi data yang kompleks, sedangkan *K-Medoids* lebih tahan terhadap *outlier*. Oleh karena itu, penelitian ini mencoba mengusulkan pendekatan yang mengintegrasikan metode *K-Medoids clustering* dengan PCA untuk menghasilkan pengelompokan wilayah yang lebih efektif dan informatif.

Berdasarkan latar belakang di atas, dalam makalah ini akan dibahas tentang pengelompokan provinsi di Indonesia berdasarkan indikator pengangguran tahun 2023 menggunakan *K-Medoids*.

2. Metodologi Penelitian

2.1. Jenis dan Sumber Data

Jenis data pada penelitian ini adalah data sekunder. Data penelitian yang digunakan bersumber dari Badan Pusat Statistik Indonesia, yaitu data Survei Ketenagakerjaan Indonesia tahun 2023. Unit observasi penelitian ini yaitu provinsi di Indonesia. Dengan variabel yang dilibatkan adalah Jumlah Bekerja (X1), Jumlah Pengangguran (X2), Jumlah Angkatan Kerja (X3), Jumlah Penduduk Usia 15 Tahun ke Atas (X4), Jumlah Tenaga Kerja Industri (X5), Lowongan Kerja Terdaftar Laki-Laki (X6), Lowongan Kerja Terdaftar Perempuan (X7), Penempatan/Pemenuhan Tenaga Kerja Laki-Laki (X8), dan Penempatan/Pemenuhan Tenaga Kerja Perempuan (X9). Pemilihan variabel ini, didasarkan pada beberapa penelitian sebelumnya, seperti yang dilakukan oleh Almeira et al. (2021) [8] dan Simanjuntak (2023) [9]. Analisis statistika yang digunakan pada data tersebut adalah analisis *K-Medoids* untuk mengelompokkan kabupaten/kota berdasarkan tingkat pengangguran. Berikut disajikan penjelasan mengenai setiap variabel pada **Tabel 1**.

P-ISSN: 2986-4178
E-ISSN: 2988-4004

Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025

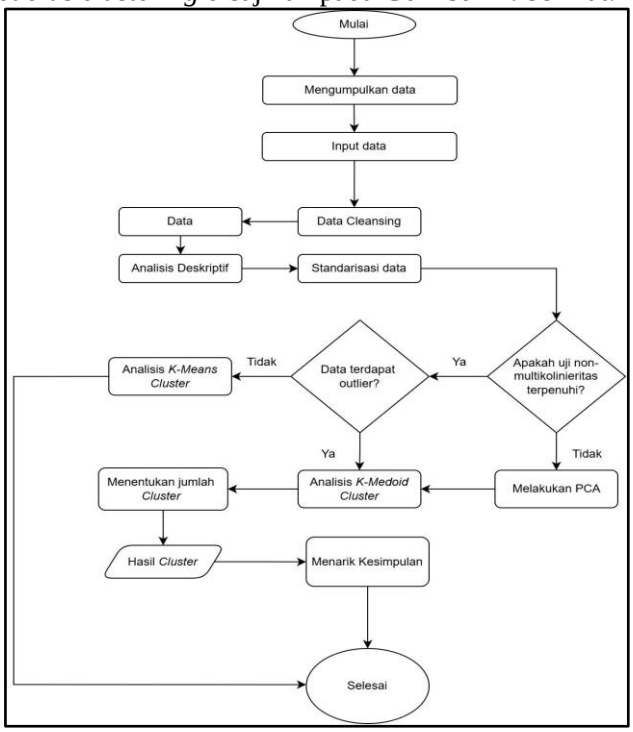
Tabel 1 Variabel

Nama Variabel	Kode	Keterangan
Jumlah Bekerja	X1	Jumlah penduduk yang bekerja di atas jam kerja normal (35 jam seminggu) atau di bawah jam kerja normal (<35 jam seminggu)
Jumlah Pengangguran	X2	Jumlah penduduk usia kerja (15 tahun ke atas) yang tidak bekerja
Jumlah Angkatan Kerja	X3	Jumlah penduduk usia 15 tahun ke atas yang aktif secara ekonomi
Jumlah Penduduk Usia 15 Tahun ke Atas	X4	Jumlah penduduk usia 15 tahun ke atas
Jumlah Tenaga Kerja Industri	X5	Jumlah rata-rata pekerja atau karyawan per hari kerja dalam industri
Lowongan Kerja Terdaftar Laki-Laki	X6	Jenis lapangan pekerjaan yang tersedia bagi para pencari kerja khususnya pengangguran laki-laki
Lowongan Kerja Terdaftar Perempuan	X7	Jenis lapangan pekerjaan yang tersedia bagi para pencari kerja khususnya pengangguran perempuan
Penempatan/Pemenuhan Tenaga Kerja Laki-Laki	X8	Jumlah penempatan atau pemenuhan tenaga kerja yang memiliki jenis kelamin laki-laki
Penempatan/Pemenuhan Tenaga Kerja Perempuan	X9	Jumlah penempatan atau pemenuhan tenaga kerja yang memiliki jenis kelamin perempuan

Sumber:https://www.bps.go.id/id/statistics-table?subject=520

2.2. Tahapan Analisis

Tahapan *K-Medoids clustering* disajikan pada **Gambar 2.** berikut.



Gambar 2. Tahapan *K-Medoids*

P-ISSN: 2986-4178
E-ISSN: 2988-4004

Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025

Gambar 2. merupakan diagram alur proses pengelompokkan data menggunakan metode *K-Medoids clustering*. Penelitian ini dimulai dengan mengumpulkan data dari sumber terpercaya, seperti Badan Pusat Statistik (BPS), yang mencakup 34 provinsi di Indonesia. Data terdiri dari sembilan variabel, yaitu: jumlah penduduk yang bekerja di atas atau di bawah jam kerja normal (X1), jumlah pengangguran usia kerja (X2), jumlah angkatan kerja (X3), jumlah penduduk usia 15 tahun ke atas (X4), jumlah tenaga kerja industri (X5), lowongan kerja terdaftar untuk laki-laki (X6) dan perempuan (X7), serta penempatan tenaga kerja laki-laki (X8) dan perempuan (X9). Data ini disusun dalam format tabel dengan 34 observasi dan 9 variabel.

Setelah data terkumpul, dilakukan proses *data cleansing* untuk memastikan tidak ada nilai kosong (*missing values*), duplikasi, atau data yang tidak wajar. Selanjutnya, data dianalisis secara deskriptif untuk mendapatkan distribusi dan karakteristik tiap variabel, seperti rata-rata, nilai minimum, dan maksimum. Untuk memastikan variabel berada dalam skala yang sama sehingga dapat dibandingkan, dilakukan standarisasi data menggunakan metode seperti *z-score*.

Tahapan berikutnya adalah pemeriksaan terhadap *outlier* untuk mengidentifikasi nilai ekstrim yang dapat mempengaruhi hasil analisis. Jika ditemukan data ekstrem, digunakan metode *K-Medoid Cluster* karena lebih tahan terhadap *outlier*. Selain itu, dilakukan pemeriksaan multikolinieritas untuk memastikan variabel independen tidak memiliki korelasi yang terlalu tinggi. Jika ditemukan masalah multikolinieritas, maka digunakan *Principal Component Analysis* (PCA) untuk mereduksi dimensi data sekaligus mempertahankan informasi utama.

Setelah data siap, analisis *clustering* dilakukan menggunakan metode *K-Means Cluster* untuk mengelompokkan provinsi berdasarkan karakteristik tenaga kerja. Jika terdapat *outlier*, analisis *K-Medoid Cluster* digunakan sebagai alternatif. Penentuan jumlah cluster optimal dilakukan dengan metode *Gap Statistic*. Tahapan terakhir adalah interpretasi hasil analisis untuk mengidentifikasi pola atau kelompok provinsi yang memiliki karakteristik serupa dalam variabel tenaga kerja.

2.3. Cleaning Data

Cleaning data atau pembersihan data adalah salah satu dari teknik *preprocessing* dalam melihat dan memperbaiki kesalahan dan inkonsistensi dari data agar data tersebut akurat [10]. Salah satu langkah dalam *cleaning data* adalah menangani masalah *missing value*. Terdapat dua cara dalam menangani *missing value*, yaitu dengan menghapus sampel data yang berisi *missing value* dengan catatan bahwa proporsi dari *missing value* tidak terlalu banyak, dan dengan metode imputasi sebagai solusi lainnya.

2.4. Statistika Deskriptif

Statistik deskriptif adalah statistik yang digunakan untuk menganalisis data dengan cara memberikan gambaran atau deskriptif suatu data yang dilihat dari nilai rata-rata, maksimum, minimum, dan standar deviasi [11]. Berikut adalah rumus rata-rata:

$$\bar{X} = \frac{\Sigma x_i}{n} \tag{1}$$

P-ISSN: 2986-4178
E-ISSN: 2988-4004
Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025

dengan $\bar{X}$ adalah rata-rata, x_i adalah data ke-i, dan n adalah banyaknya data. Berikutnya disajikan rumus standar deviasi:

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} \tag{2}$$

dengan s adalah standar deviasi, x_i adalah data ke-i, $\bar{x}$ adalah rata-rata, dan n adalah banyaknya data.

2.5. Standarisasi Data

Standarisasi adalah teknik lain dalam perubahan skala data, di mana data yang dimiliki akan diubah sehingga memiliki rata-rata (mean) = 0 (terpusat) dan standar deviasi= 1. Teknik ini berguna untuk menghilangkan pengaruh skala yang berbeda antara variabel dalam analisis.

$$X' = \frac{X - \mu}{\sigma} \tag{3}$$

dengan X adalah nilai data asli, μ adalah nilai rata-rata dari data yang ada, dan σ adalah nilai standar deviasi dari data.

2.6. Uji Multikolinearitas

Multikolinearitas adalah fenomena statistik di mana dua atau lebih prediktor dalam model regresi berganda berkorelasi tinggi [12]. Dalam uji multikolinearitas, korelasi antara variabel bebas perlu diperhatikan untuk memastikan tidak adanya hubungan yang terlalu kuat antar variabel. Jika terdapat korelasi tinggi, maka pengelompokan data dapat menjadi kurang jelas dan sulit diinterpretasikan. Multikolinearitas yang signifikan dapat dideteksi dengan melihat nilai korelasi antar variabel, di mana korelasi yang tinggi menunjukkan potensi masalah dalam analisis klaster. Statistika uji yang digunakan dalam uji Henze Zirkler adalah sebagai berikut:

$$\text{HZ} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} e^{-\frac{\beta^2}{2}D_{ij}} - 2(1+\beta^2)^{-\frac{p}{3}}\frac{1}{n}\sum_{i=1}^{n} e^{-\frac{\beta^2}{2(1+\beta^2)}D_i} + (1+2\beta^2)^{-\frac{p}{2}} \tag{4}$$

dimana, $\beta = \frac{1}{\sqrt{2}}\left(\frac{n(2p+1)}{4}\right)^{\frac{1}{p+4}}$; $D_{ij} = (x_i - x_j)^T S^{-1}(x_i - x_j)$; $D_i = (x_i - \bar{x})^T S^{-1}(x_i - \bar{x})$;

p = jumlah variabel; dan S^{-1} = matriks varians kovarians [13].

2.7. *Principal Component Analysis* (PCA)

PCA merupakan teknik statistik yang dapat digunakan untuk menjelaskan struktur variansi-kovariansi dari sekumpulan variabel baru dimana variabel baru ini saling bebas dan merupakan kombinasi linier dari variabel asal atau dapat digunakan untuk data yang memiliki multikolinearitas [6]. *Principal Component Analysis* bertujuan untuk menyederhanakan dan menghilangkan faktor atau indikator *screening* yang kurang dominan dan kurang relevan tanpa mengurangi maksud dan tujuan dari data asli dari

P-ISSN: 2986-4178
E-ISSN: 2988-4004
Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025

variabel acak x (matrik berukuran n x n, dimana baris-baris yang berisi observasi sebanyak n dari variabel acak x) adalah sebagai berikut:

1. Menghitung matrik varians kovarian dari data observasi.

$$Var(x) = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(Z_{ij} - \mu_j)^2 \tag{5}$$

$$Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \mu_{xj})(y_{ij} - \mu_{yj}) \tag{6}$$

dengan μ_x dan μ_y merupakan rata – rata (mean) sampel dari variabel x dan y, dimana variabel x_i dan x_j merupakan nilai observasi ke-i dari variabel x dan y. Dari data nilai yang digunakan, maka diperoleh matrik kovarian berukuran n x n.

2. Mencari *eigen values* dan *eigen vector* dari matrik kovarian yang telah diperoleh yaitu [14]: Nilai *eigen* dan vektor *eigen* untuk matriks kovarians dihitung. Nilai *eigen* yang dikomputasi kemudian ditransformasikan (rotasi orthogonal varimax) menggunakan persamaan berikut:

$$Det(A - \lambda I) = 0 \tag{7}$$

dimana, A = matrik n × n; λ = nilai *eigenvalue*; I = matriks identitas (matriks persegi dengan elemen diagonal utama bernilai 1 sedangkan elemen lain bernilai 0)

3. Menentukan nilai proporsi *Principal Component* (proporsi *Principal Component* (%) dengan persamaan:

$$PC(\%) = \frac{Nilai\ Eigen}{Variance\ Covarian} x100\% \tag{8}$$

4. Menghitung bobot factor (*factor loading*) berdasarkan *eigenvector* dengan persamaan:

$$Ax = \lambda x \tag{9}$$

Sehingga diperoleh kombinasi linear yaitu:

$\lambda_1, \lambda_2, \lambda_3$ adalah *eigen value* matrik A dan x_1, x_2, x_3 adalah *eigen vector* sesuai *eigen value*-nya (λ_n)

Persamaan *eigen value* & *eigen vector* merupakan *Eigen Value Decomposition* (EVD), dengan persamaan sebagai berikut:

$$AX = XD \tag{10}$$

dengan, A = matrik n x n yang memiliki n *eigen value* (λ_n) dan D = *eigen value* dari *eigen vector*-nya.

$$A = XDX^{-1} \tag{11}$$

dengan, X = *eigen vector* dari matrik A dan X^{-1} = invers dari *eigen vector* X.

2.8. *K-Medoids Clustering*

Clustering adalah proses mengelompokkan sekumpulan objek data menjadi beberapa kelompok atau *cluster*. Objek dalam satu kelompok memiliki banyak persamaan

P-ISSN: 2986-4178
E-ISSN: 2988-4004

Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025

atau kemiripan, tetapi mereka sangat berbeda dari objek dalam kelompok lain. Didasarkan pada karakteristik yang menggambarkan objek, kemiripan dinilai dengan menggunakan pengukuran jarak. Dalam situasi ini, metode *clustering* yang berbeda dapat menghasilkan *cluster* yang berbeda pada kumpulan data yang sama; ini dilakukan melalui algoritma *clustering* dan tanpa mengetahui target kelas terlebih dahulu. Proses ini termasuk dalam kategori pembelajaran yang tidak diawasi yang tidak memiliki data latih. Oleh karena itu, data yang ada dikelompokkan menjadi dua, tiga, atau lebih bagian. Algoritma *clustering* dibuat dengan berbagai cara. Pendekatan hierarki dan pendekatan partisi adalah dua pendekatan utama untuk *clustering* data. Pendekatan partisi, yang juga dikenal sebagai *partition-based clustering*, memilah data yang akan dianalisis ke dalam cluster yang sudah ada. Salah satu algoritma yang termasuk dalam pendekatan ini yaitu algoritma *K-Medoids*.

Algoritma PAM (*Partioning Around Medoids*) atau biasa juga disebut dengan algoritma *K-Medoids*, merupakan algoritma yang diwakili oleh *cluster* yaitu medoid[14]. Metode *K-Medoids* muncul sebagai solusi untuk mengatasi kekurangan dari metode *K-Means* yang sensitif terhadap *outlier* dan bekerja dengan cara memisahkan *dataset* menjadi beberapa kelompok. Metode *K-Medoids* memiliki tujuan yang sama dengan *K-Means*, yaitu untuk meminimalkan jarak antara titik yang ditentukan sebagai pusat *cluster* dan titik data *cluster* [15]. Langkah-langkah pengelompokkan dalam metode *K-Medoids Clustering* sebagai berikut:

1. Menentukan banyaknya k *cluster*
2. Menentukan pusat awal *cluster* (medoid) secara acak
3. Menghitung jarak *euclidean* objek data pada *medoid* awal

$$D(x_i, x_j) = \sqrt{\sum_{m=1}^{n} (x_{im} - x_{jm})^2} \tag{12}$$

dimana, D = jarak *euclidean*; x_i = data ke-i; x_j = data ke-j; x_{im} = data ke-i atribut-m; x_{jm} = data ke-j atribut-m

4. Memilih objek data non-*medoid* pada masing-masing *cluster* sebagai kandidat *medoid* baru
5. Menghitung jarak *euclidean* setiap objek yang berada pada masing-masing *cluster* dengan kandidat *medoid* baru
6. Menghitung Total Simpangan (S), yaitu selisih dari total jarak terdekat objek pada *medoid* awal dan kandidat *medoid* baru

$$(S) = b - a \tag{13}$$

dimana, b = jumlah jarak terdekat antara objek ke *medoid* akhir; a = jumlah jarak terdekat antara objek ke *medoid* awal

jika S < 0, maka kandidat *medoid* baru diterima sebagai pembaruan *medoid* dan dilakukan pengulangan dari langkah 4

jika S ≥ 0, maka menolak kandidat *medoid* baru untuk pembaruan *medoid* dan hasil klasterisasi telah ditemukan.

Konsep mendapatkan nilai total simpangan paling kecil pada metode *K-Medoid clustering* melibatkan pemilihan *medoid* yang dapat menghasilkan klasterisasi dengan simpangan yang minimal. Total simpangan yang dimaksud adalah jumlah dari jarak antara setiap titik data dengan *medoid* yang terkait dalam *cluster* yang sesuai. Dalam setiap iterasi, *medoid* diganti dengan titik data lain dan total simpangan dihitung ulang. Jika total simpangan baru lebih kecil, maka medoid diubah. Jika tidak, *medoid* dikembalikan ke

P-ISSN: 2986-4178
E-ISSN: 2988-4004

Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025

medoid sebelumnya. Proses ini diulang untuk setiap *medoid* dalam *cluster* hingga konvergensi tercapai atau tidak ada perubahan lebih lanjut dalam total simpangan.

2.9. *Gap Statistic*

Gap Statistic digunakan untuk menentukan jumlah cluster optimal dengan membandingkan variasi dalam klaster antara data asli dan acak. Gap yang lebih besar menunjukkan pemisahan klaster yang lebih baik. Rumus Gap Statistic adalah:

$$Gap(k) = \frac{1}{B}\sum_{b=1}^{B} \log(W_k^b) - log(W_k) \quad (14)$$

Langkah-langkahnya:

1. Mengelompokkan data untuk berbagai k dan menghitung variasi intracluster W_k
2. Membuat data acak dan menghitung variasi intracluster W_{kb} .
3. Menghitung *Gap Statistic* dan standar deviasinya.
4. Memilih kk terkecil yang gap-nya berada dalam satu standar deviasi dari $k + 1$ [16].

3. Hasil dan Pembahasan

Berikut adalah data tenaga kerja menurut provinsi di Indonesia pada tahun 2023 yang disajikan pada **Tabel 2**. Tabel ini menunjukkan berbagai indikator terkait tenaga kerja, seperti jumlah penduduk yang bekerja, pengangguran, angkatan kerja, serta data terkait lapangan pekerjaan yang tersedia dan penempatan tenaga kerja di masing-masing provinsi.

Tabel 2 Data Tenaga Kerja

Provinsi	X1	X2	X3	X4	X5	X6	X7	X8	X9
Aceh	4.897.301	306.546	5.203.847	8.012.600	199.504	577	570	838	608
Sumatera Utara	15.010.693	884.611	15.895.304	22.441.918	265.159	3.693	2.203	4.862	2.611
Sumatera Barat	5.667.468	356.473	6.023.941	8.632.683	168.696	19.148	1.368	3.670	1.864
…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…
Papua Barat	1.149.612	66.380	1.215.992	1.780.160	11.645	206	168	277	197
Papua	4.870.071	154.694	5.024.765	6.526.278	28.166	344	320	436	341

Sumber: https://www.bps.go.id/id/statistics-table?subject=520

Tabel 2 menunjukkan distribusi berbagai variabel tersebut untuk setiap provinsi di Indonesia, yang memberikan gambaran tentang dinamika tenaga kerja pada tahun 2023. Dengan informasi ini, analisis lebih lanjut dapat dilakukan untuk mengevaluasi kesenjangan tenaga kerja, peluang kerja, serta penempatan tenaga kerja laki-laki dan perempuan di seluruh provinsi.

3.1. Statistika Deskriptif

Sebelum melakukan *K-Medoids clustering* perlu dilihat karakteristik data tenaga kerja di setiap provinsi di Indonesia pada tahun 2023. Berikut merupakan karakteristik data dari jumlah tenaga kerja, jumlah bekerja, jumlah pengangguran, serta informasi terkait lowongan kerja dan penempatan tenaga kerja berdasarkan *gender* yang disajikan pada **Tabel 3** berikut.

P-ISSN: 2986-4178
E-ISSN: 2988-4004

Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025

Tabel 3 Statistik Deskriptif

Variabel	Minimum	Q1	Median	Rata-rata	Q3	Maksimum
X1	730047	2242803	4581841	8190732	8028364	46798422
X2	30845	105243	176558	466010	383325	3893453
X3	760892	2339008	4740134	8656742	8389447	50791875
X4	1107825	3419190	6375160	12475774	12134016	76239197
X5	11125	58205	120835	289525	223393	2498925
X6	116	469.5	1227	3896.1	2907.8	29453
X7	29	200.8	710.5	2485.4	1549.8	21710
X8	257	1334	2088	5381	4574	39225
X9	165	418	961	3675	2837	34700

Berdasarkan **Tabel 3**, Jumlah Bekerja (X1) menunjukkan disparitas besar antar provinsi, dengan nilai minimum yang rendah (730.047) dan rata-rata yang tinggi (8.190.732), mencerminkan ketidakmerataan ketenagakerjaan di Indonesia. Jumlah Pengangguran (X2) memiliki rentang yang luas, dengan rata-rata 466.010, yang menunjukkan adanya variasi tingkat pengangguran di berbagai provinsi. Sementara itu, Jumlah Angkatan Kerja (X3) juga menunjukkan ketidakmerataan, dengan nilai minimum 760.892 dan rata-rata 8.656.742, mencerminkan perbedaan signifikan dalam jumlah angkatan kerja antar provinsi. Jumlah Tenaga Kerja Industri (X4) relatif rendah di beberapa provinsi, dengan nilai minimum 11.125 dan rata-rata 289.525, yang menunjukkan terbatasnya kontribusi sektor industri di wilayah tertentu. Lowongan Kerja Terdaftar Laki-Laki (X6) dan Perempuan (X7) menggambarkan rendahnya peluang kerja yang tersedia, dengan nilai minimum masing-masing 116 untuk laki-laki dan 29 untuk perempuan, serta rata-rata 3.896,1 untuk laki-laki dan 2.485,4 untuk perempuan. Terakhir, Penempatan Tenaga Kerja Laki-Laki (X8) dan Perempuan (X9) menunjukkan perbedaan dalam kapasitas penyerapan tenaga kerja antar provinsi, dengan nilai minimum masing-masing 257 untuk laki-laki dan 165 untuk perempuan.

Secara keseluruhan, analisis deskriptif ini menunjukkan ketidakmerataan distribusi tenaga kerja di Indonesia, baik dalam hal jumlah pekerjaan yang tersedia, sektor industri, maupun kesenjangan antara laki-laki dan perempuan. Hal ini menunjukkan perlunya kebijakan yang dapat memastikan kesempatan kerja lebih merata di seluruh provinsi dan mengurangi ketimpangan dalam penyerapan tenaga kerja.

3.2. Standarisasi data

Data yang digunakan dalam penelitian ini telah diperiksa dan tidak ditemukan adanya *missing values*. Oleh karena itu, data ini dapat langsung digunakan untuk tahap analisis dan pengujian selanjutnya tanpa memerlukan proses imputasi atau pengolahan tambahan.

Pada penelitian ini, dilakukan standarisasi data terhadap variabel X1, X2, X3, X4, X5, X6, X7, X8, dan X9. Tujuan dilakukannya standarisasi adalah untuk mengurangi ketidakseimbangan skala antar variabel, sehingga masing-masing variabel dapat dibandingkan dalam skala yang seragam. Standarisasi ini diperlukan karena variabel-variabel yang digunakan memiliki skala yang berbeda-beda, seperti jumlah penduduk yang bekerja, jumlah pengangguran, dan lowongan kerja terdaftar, yang masing-masing memiliki rentang nilai yang sangat bervariasi. Jika data tidak di standarisasi, variabel dengan skala besar dapat mendominasi hasil analisis, sementara variabel dengan skala kecil bisa terabaikan. Dengan melakukan standarisasi, seluruh variabel dapat dibandingkan pada skala yang seragam, sehingga memberikan hasil analisis yang lebih adil dan akurat. Proses standarisasi ini bertujuan untuk mengubah data ke dalam skala yang memiliki rata-rata 0

P-ISSN: 2986-4178
E-ISSN: 2988-4004

Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025

dan standar deviasi 1. Skala yang digunakan dalam tahap standarisasi pada penelitian ini adalah 0 ke 1, dengan rumus yang digunakan adalah Persamaan (1).

Tabel 4 Data Hasil Standarisasi

X1	X2	X3	X4	X5	X6	X7	X8	X9
-0.28	-0.20	-0.27	-0.25	-0.17	-0.47	-0.37	-0.50	-0.40
0.58	0.53	0.58	0.55	-0.04	-0.03	-0.06	-0.06	-0.14
-0.21	-0.14	-0.21	-0.21	-0.22	2.18	-0.22	-0.19	-0.24
...	…	…	…	…	…	…	…	...
...	…	…	…	…	…	…	…	...
-0.60	-0.50	-0.59	-0.59	-0.51	-0.53	-0.45	-0.56	-0.46
-0.28	-0.39	-0.29	-0.33	-0.48	-0.51	-0.42	-0.54	-0.44

3.3. Uji Multikolinearitas

Uji multikolinearitas bertujuan untuk mengetahui adanya korelasi antar variabel. Dalam penelitian ini, digunakan uji multikolinearitas dengan metode Henze-Zirkler untuk mengidentifikasi hubungan antar variabel independen. Jika *p-value* < α (0.05) maka tolak H_0 artinya data tidak berdistribusi normal multivariat, sedangkan jika p-value > α (0.05) maka gagal tolak H_0 artinya data berdistribusi normal multivariat [17].

```
            Test       HZ   p value MVN
1 Henze-Zirkler 0.969712 0.2259814 YES
```

Gambar 3. Hasil Analisis Korelasi

Berdasarkan hasil perhitungan pada **Gambar 3.**, diketahui bahwa *p-value* (0.2259814) > α (0.05) yang artinya gagal tolak H_0. Hal ini menunjukkan bahwa data berdistribusi normal multivariat. *Principal Component Analysis* (PCA) adalah salah satu pendekatan efektif untuk mengatasi masalah multikolinearitas.

3.4. *Principal Component Analysis* (PCA)

Untuk mengatasi masalah multikolinearitas yang terdeteksi melalui analisis korelasi, salah satu pendekatan yang umum digunakan adalah *Principal Component Analysis* (PCA). PCA dilakukan pada data yang telah disusun untuk mereduksi dimensi atau fitur dari data. Proses PCA menghasilkan sejumlah *principal component* (PC) yang jumlahnya dapat mencapai jumlah kolom maksimal dari data yang digunakan.

Tabel 5 Nilai Eigen, Proporsi Varians, dan Proporsi Kumulatif

Komponen Utama	Nilai Eigen	Proporsi Variansi	Proporsi Variansi Kumulatif
1	8.4795	0.942	0.942
2	0.2760	0.031	0.973
3	0.1559	0.017	0.990
4	0.0443	0.005	0.995
5	0.0213	0.002	0.997
6	0.0184	0.002	0.999
7	0.0045	0.001	1.000
8	0.0001	0.000	1.000
9	0.0000	0.000	1.000

P-ISSN: 2986-4178
E-ISSN: 2988-4004
Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025

Tabel 5 menunjukkan nilai *eigen* setiap komponen utama yang mencerminkan kontribusi terhadap varians data asli. Tiga komponen utama dipilih karena secara kumulatif menjelaskan 99,02% varians, dengan Komponen 1 menjelaskan 94,22%, Komponen 2 sebesar 3,07%, dan Komponen 3 sebesar 1,73%. Meskipun Komponen 1 merupakan komponen yang optimal dalam menjelaskan data dengan nilai *eige*n lebih dari 1, dua komponen tambahan tetap dipertimbangkan untuk menangkap variasi kecil namun relevan.

Data hasil proses PCA dihitung dengan menggunakan nilai *eigen* dan vektor *eigen*, kemudian disusun dalam tiga komponen utama, yang masing-masing memiliki 34 baris data. Data yang telah direduksi ini selanjutnya digunakan dalam proses *clustering*. Contoh baris data hasil proses PCA dapat dilihat pada **Tabel 6**.

Tabel 6 Data Hasil PCA

Provinsi	Comp.1	Comp.2	Comp.3
Aceh	-0.983298994	0.107716505	0.186616332
Sumatera Utara	0.658313717	0.458967827	0.027519362
Sumatera Barat	0.127151287	-1.457079855	-1.733818195
…	…	…	…
…	…	…	…
Papua Barat	-1.622294784	0.005733309	0.034816952
Papua	-1.240795619	0.136850565	0.105110420

Tabel 6 menunjukkan hasil *Principal Component Analysis* (PCA) yang mereduksi dimensi data dan menghasilkan tiga komponen utama (Comp.1, Comp.2, dan Comp.3). Komponen pertama (Comp.1) mengandung variasi terbesar dalam data, dengan provinsi seperti Aceh, Papua Barat, dan Papua yang memiliki nilai negatif, sementara provinsi seperti Sumatera Utara memiliki nilai positif, menunjukkan perbedaan karakteristik yang mencolok antara provinsi-provinsi tersebut. Komponen kedua (Comp.2) menjelaskan variasi tambahan, di mana Sumatera Barat memiliki nilai sangat negatif, sementara provinsi lain menunjukkan nilai positif, yang menunjukkan adanya perbedaan dalam variabel yang dijelaskan oleh komponen kedua. Komponen ketiga (Comp.3) juga menunjukkan variasi yang lebih kecil namun penting, dengan Sumatera Barat memiliki nilai yang sangat negatif, sementara provinsi lain memiliki nilai positif yang lebih kecil. Berdasarkan hasil PCA ini, kita dapat mengelompokkan provinsi-provinsi berdasarkan kesamaan nilai pada ketiga komponen utama, yang memungkinkan pemisahan provinsi berdasarkan karakteristik yang serupa. Penggunaan PCA membantu mereduksi dimensi data dan mengatasi masalah multikolinearitas, sehingga memudahkan analisis lanjutan seperti clustering.

3.5. Analisis *K-Medoids Clustering*

Proses *clustering* dilakukan pada data dari 34 provinsi di Indonesia menggunakan algoritma *K-Medoids*, karena uji asumsi non-multikolinearitas tidak terpenuhi dan data menunjukkan adanya *outlier*. Berdasarkan hasil pengujian Henze-Zirkler untuk *multivariate normality*, diperoleh nilai statistik HZ sebesar 0.969712 dengan *p-value* sebesar 0.2259814. Dengan *p-value* > 0.05, data dinyatakan memenuhi asumsi normal multivariat (*MVN: YES*). Namun, meskipun data multivariat normal, ditemukan *outlier* pada beberapa variabel yang dapat mempengaruhi hasil analisis *clustering*. Oleh karena itu, algoritma *K-Medoids* dipilih karena lebih tahan terhadap *outlier* dibandingkan algoritma seperti *K-Means*.

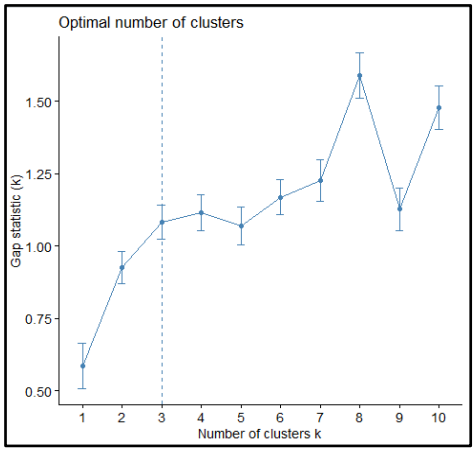
3.5.1 Penentuan Jumlah *Cluster*

Penentuan jumlah *cluster* yang optimal dilakukan terlebih dahulu untuk mengidentifikasi jumlah kelompok yang paling tepat berdasarkan data yang digunakan.

P-ISSN: 2986-4178
E-ISSN: 2988-4004

Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025

Dalam penelitian ini, metode *Gap Statistic* diterapkan untuk menentukan jumlah *cluster* yang optimal, dengan hasil pada **Gambar 4.**



Gambar 4. *Cluster* optimal

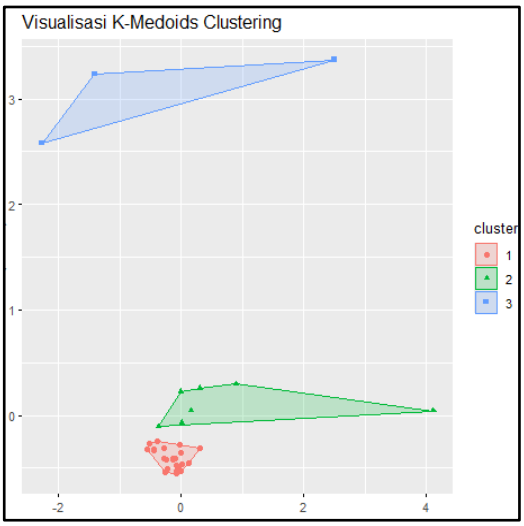
Berdasarkan **Gambar 4.**, grafik tersebut menampilkan *Gap Statistic* untuk menentukan jumlah *cluster* optimal dalam analisis *cluster*. Sumbu X menunjukkan jumlah *cluster* (k), sedangkan sumbu Y menunjukkan nilai *Gap Statistic* untuk setiap jumlah *cluster* (k). Garis dengan titik-titik vertikal merepresentasikan nilai *Gap Statistic* beserta *error bars* (kesalahan atau variasi), sementara garis putus-putus vertikal menunjukkan jumlah *cluster* optimal yang diusulkan berdasarkan metode *Gap Statistic*. *Gap Statistic* mengukur seberapa jauh hasil *cluster* yang dihasilkan lebih baik dibandingkan dengan data acak. Jumlah *cluster* optimal dipilih sebagai *k* di mana *Gap Statistic* pertama kali mencapai nilai maksimum sebelum menurun atau saat nilai tersebut tetap stabil dalam kisaran tertentu. Dalam grafik ini, jumlah *cluster* optimal adalah 3, karena garis putus-putus vertikal menunjukkan titik tersebut, dan *Gap Statistic* pada $k = 3$ relatif tinggi dibandingkan nilai *k* sebelumnya. Meskipun nilai *Gap Statistic* pada $k = 4$ sedikit lebih tinggi dibandingkan dengan $k = 3$, peningkatannya sangat kecil dibandingkan kenaikan antara $k = 2$ dan $k = 3$. Hal ini menunjukkan bahwa $k = 3$sudah memberikan pemisahan klaster yang baik tanpa menambah kompleksitas.

3.5.2 Hasil Pengelompokan

Pengelompokan provinsi berdasarkan faktor-faktor yang mempengaruhi pengangguran, yang didasarkan pada data tenaga kerja, kemudian dikelompokkan menjadi tiga kelompok, dengan visualisasi *plot cluster* pada **Gambar 5.**

P-ISSN: 2986-4178
E-ISSN: 2988-4004

Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025



Gambar 5. *Plot Cluster*

Berdasarkan **Gambar 5.** merupakan visualisasi hasil *K-Medoids Clustering* yang menggambarkan pembagian data ke dalam tiga klaster berbeda berdasarkan hasil algoritma. Warna (merah, hijau, dan biru) digunakan untuk membedakan setiap klaster, sementara simbol (bulatan untuk klaster 1, segitiga untuk klaster 2, dan kotak untuk klaster 3) membantu memperjelas identifikasi klaster. Setiap klaster ditampilkan dalam bentuk *polygon* yang mencakup data poin di dalamnya, menggambarkan batas atau ruang cakupan klaster yang dihasilkan oleh algoritma. Klaster merah (1) terlihat lebih terkonsentrasi di satu area, menunjukkan data yang lebih homogen, sedangkan klaster hijau (2) dan biru (3) tampak lebih tersebar, mencerminkan adanya variasi data yang lebih besar. Pemisahan antara klaster biru dan hijau cukup jelas, sementara klaster merah tampak lebih terisolasi, yang mungkin mengindikasikan karakteristik unik dari data dalam klaster tersebut. Visualisasi ini memberikan gambaran tentang pola atau struktur dalam dataset berdasarkan hasil pengelompokan dengan algoritma *K-Medoids*.

Tabel 7 Hasil klasterisasi

Cluster	Jumlah	Anggota *Cluster*
1	24	Aceh, Riau, Jambi, Bengkulu, Kepulauan Bangka Belitung, Kepulauan Riau, DI Yogyakarta, Bali, Nusa Tenggara Barat, Nusa Tenggara Timur, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku, Maluku Utara, Papua Barat, Papua.
2	7	Sumatera Utara, Sumatera Barat, Sumatera Selatan, Lampung, DKI Jakarta, Banten, Sulawesi Selatan.
3	3	Jawa Barat, Jawa Tengah, Jawa Timur.

Tabel 7 menunjukkan hasil pengelompokan menggunakan metode *K-Medoids*. Berdasarkan tabel tersebut, akan dilakukan profilisasi untuk setiap cluster yang disajikan pada Tabel 8 & 9 berikut.

P-ISSN: 2986-4178
E-ISSN: 2988-4004

Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025

Tabel 8 Profilisasi *cluster* dari data asli

Provinsi	X1	X2	X3	X4	X5	X6	X7	X8	X9	Cluster
Aceh	4.897.301	306.546	5.203.847	8.012.600	199.504	577	570	838	608	1
Sumatera Utara	15.010.693	884.611	15.895.304	22.441.918	265.159	3.693	2.203	4.862	2.611	2
Sumatera Barat	5.667.468	356.473	6.023.941	8.632.683	168.696	19.148	1.368	3.670	1.864	2
…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…
Papua Barat	1.149.612	66.380	1.215.992	1.780.160	11.645	206	168	277	197	1
Papua	4.870.071	154.694	5.024.765	6.526.278	28.166	344	320	436	341	1

Tabel 8 menunjukkan profilisasi *cluster* berdasarkan data asli dari variabel-variabel seperti jumlah penduduk (X1), pencari kerja terdaftar (X2), tenaga kerja terlibat (X3), pengangguran terbuka (X5), dan lainnya. Setiap provinsi dikelompokkan ke dalam tiga *cluster* hasil pengelompokan metode *K-Medoids*. *Cluster* 1, yang mencakup provinsi seperti Aceh dan Papua Barat, memiliki karakteristik tingkat pengangguran rendah dan penempatan tenaga kerja yang stabil. *Cluster* 2, seperti Sumatera Utara dan Sumatera Barat, menunjukkan jumlah penduduk lebih besar, peluang kerja tinggi, tetapi tantangan pemenuhan tenaga kerja masih ada. Profilisasi ini memberikan gambaran rinci tentang perbedaan karakteristik antar *cluster* sebelum analisis lanjutan menggunakan PCA.

Tabel 9 Profilisasi *cluster* dari PCA

	CLUSTER		
KOMPONEN UTAMA	**1**	**2**	**3**
Comp.1	-1,216353993	0,290897107	9,05207203
Comp.2	0,003899311	0,004761249	-0,042304071
Comp.3	0,066182961	-0,293906321	0,156317729

Tabel 9 menunjukkan nilai rata-rata dari tiga komponen utama (Comp.1, Comp.2, dan Comp.3) untuk setiap *cluster*. Angka-angka ini menggambarkan karakteristik masing-masing *cluster*. Misalnya, *Cluster* 1 memiliki nilai negatif pada Comp.1, sementara *Cluster* 3 memiliki nilai positif yang tinggi, menunjukkan perbedaan karakteristik antara *cluster*. Nilai pada Comp.2 dan Comp.3 juga berbeda untuk setiap *cluster*, yang menggambarkan variasi tambahan dalam data. Terdapat tiga kelompok provinsi dengan karakteristik dari masing-masing kelompoknya adalah sebagai berikut.

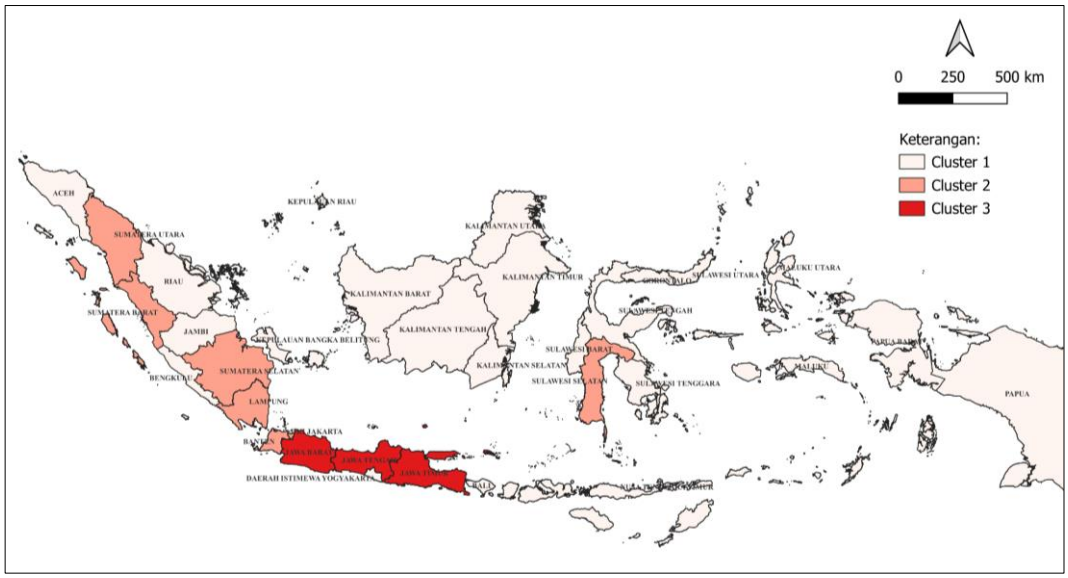
1. *Cluster* 1, memiliki 19 anggota yaitu Provinsi yang terdapat pada *cluster* pertama ini, seperti Aceh, Bali, Bengkulu, dan lain-lain, memiliki tingkat pengangguran yang jauh lebih rendah dibandingkan dengan provinsi dalam kelompok cluster lainnya. Selain itu, provinsi-provinsi pada *cluster* ini menunjukkan angka pengangguran yang relatif lebih rendah dan stabil dalam berbagai sektor tenaga kerja, seperti jumlah tenaga kerja industri, lowongan kerja terdaftar, serta penempatan tenaga kerja laki-laki dan perempuan.
2. *Cluster* 2, dengan 6 anggota, termasuk provinsi seperti Banten, DKI Jakarta, dan Lampung, menunjukkan tingkat pengangguran yang lebih tinggi dibandingkan dengan *Cluster* 1. Meskipun memiliki potensi tinggi dalam hal lowongan pekerjaan, *cluster* ini masih menghadapi tantangan dalam pemenuhan tenaga kerja, terutama di sektor-sektor tertentu.
3. *Cluster* 3, terdiri dari 6 anggota, termasuk provinsi seperti Jawa Barat, Jawa Tengah, dan Jawa Timur, menunjukkan pengangguran yang paling tinggi di antara ketiga *cluster*. Meskipun jumlah penduduk dan tenaga kerja yang terlibat sangat

P-ISSN: 2986-4178
E-ISSN: 2988-4004

Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025

besar, tingkat pengangguran tetap menjadi isu yang signifikan, dengan jumlah pengangguran dan lowongan kerja yang lebih tinggi dibandingkan *cluster* lainnya.

3.6. Pemetaan Hasil *Cluster*



Gambar 6. Pemetaan *Cluster*

Gambar 6. menunjukkan hasil pemetaan dari *cluster* menggunakan *software* QGIS. Peta Indonesia ditampilkan dengan tiga warna berbeda untuk membedakan tiap *cluster* yang terbentuk. Warna merah mewakili *cluster* ketiga dengan tingkat pengangguran tinggi, warna pink menunjukkan *cluster* kedua dengan tingkat pengangguran sedang, dan warna putih merepresentasikan *cluster* pertama dengan tingkat pengangguran rendah.

4. Kesimpulan

Berdasarkan hasil pengelompokan tingkat pengangguran di provinsi-provinsi Indonesia menggunakan metode *K-Medoids Clustering*, dapat disimpulkan bahwa penelitian ini Metode *Gap Statistic* digunakan untuk menentukan jumlah *cluster* yang optimal dengan membandingkan hasil *clustering* pada data asli dengan hasil *clustering* pada data acak. *Gap Statistic* mengukur perbedaan atau "kesenjangan" antara kedua hasil tersebut, dan jumlah *cluster* yang optimal adalah jumlah yang meminimalkan kesenjangan ini. Dalam penelitian ini, *Gap Statistic* menunjukkan bahwa jumlah *cluster* yang paling optimal adalah tiga, yang digunakan untuk mengelompokkan tingkat pengangguran di provinsi-provinsi Indonesia dan menghasilkan tiga kelompok provinsi. *Cluster* 1 (24 provinsi) memiliki tingkat pengangguran paling rendah dan stabil, *Cluster* 2 (7 provinsi) menunjukkan tingkat pengangguran sedang dengan tantangan pemenuhan tenaga kerja di sektor tertentu, dan *Cluster* 3 (3 provinsi) memiliki tingkat pengangguran tertinggi meskipun jumlah penduduk dan tenaga kerjanya besar. Visualisasi spasial melalui peta QGIS menunjukkan distribusi pengangguran yang jelas, dengan *Cluster* 1 (pengangguran rendah) ditandai dengan warna putih, *Cluster* 2 (pengangguran sedang) dengan warna pink, dan *Cluster* 3 (pengangguran tinggi) dengan warna merah.

Berdasarkan hasil tersebut, dapat disimpulkan bahwa tingkat pengangguran di provinsi-provinsi Indonesia telah tergambarkan dengan jelas, sehingga dapat dijadikan landasan dalam merumuskan strategi pengurangan pengangguran yang disesuaikan dengan karakteristik masing-masing *cluster*. Sebagai contoh, provinsi pada *Cluster* 1 dapat diarahkan pada pengembangan tenaga kerja dengan keahlian khusus guna meningkatkan daya saing. Sementara itu, provinsi pada *Cluster* 2 memerlukan perhatian pada sektor-

sektor tertentu yang mengalami kekurangan tenaga kerja. Adapun provinsi pada *Cluster* 3, prioritas utama adalah meningkatkan peluang kerja melalui investasi yang signifikan pada sektor padat karya serta penguatan pendidikan vokasi. Dengan menerapkan rekomendasi tersebut, diharapkan Indonesia dapat mencapai target RPJM 2020-2024 yang sejalan dengan visi Indonesia Emas 2045.

5. Daftar Pustaka

[1] Direktorat Statistik Kependudukan dan Ketenagakerjaan, Booklet Sakernas: Agustus 2021. Jakarta: Badan Pusat Statistik, Desember 2021, ISBN: 2714-853X, No. Publikasi: 04100.2114, Katalog: 2303014, ii + 22 halaman.

[2] I. E. Mufrida, "Banten Jadi Provinsi Paling Banyak Pengangguran, Mayoritas Lulusan SMK," GoodStats, 10 Juni 2024. [Online]. Available: https://goodstats.id/article/banten-jadi-provinsi-paling-banyak-pengangguran-mayoritas-lulusan-smk-E1ZKC. [Accessed: 07-Jan-2025]

[3] Rina, "Clustering menggunakan Algoritma K-Medoids," Medium, 2020. [Online].Tersedia:https://esairina.medium.com/clustering-menggunakan-algoritma-K-Medoids-67179a333723.

[4] P. Di, K. Bojonegoro, D. Hediyati, and I. M. Suartana, "Penerapan Principal Component Analysis (PCA) Untuk Reduksi Dimensi Pada Proses Clustering Data Produksi," vol. 05, pp. 49–54, 2021.

[5] A. Karim, S. Esabella, S. Suryadi, and E. Mardinata, "Clusterisasi Tingkat Pengangguran Terbuka Menurut Provinsi di Indonesia Menggunakan Algoritma K-Medoids," vol. 6, no. 3, pp. 1341–1351, 2024, doi: 10.47065/bits.v6i3.6198.

[6] M. P. Putri, et al., "Analisis Faktor yang Mempengaruhi Kualitas Hidup di Indonesia Tahun 2020 Menggunakan Analisis Komponen Utama dan Analisis Faktor," Jurnal Matematika dan Ilmu Pengetahuan Alam, vol. 4, no. 3, 2024.

[7] E. Detrinidad and V. R. López-Ruiz, "The Interplay of Happiness and Sustainability: A Multidimensional Scaling and K-Means Cluster Approach," Sustain., vol. 16, no. 22, 2024, doi: 10.3390/su162210068.

[8] D. Almeira and G. Graciella Juanda, "Analisis Multidimensional Scaling dan k-Means Clustering untuk Pengelompokan Provinsi Berdasarkan Tingkat Pengangguran," E-Prosiding Nas. | Dep. Stat. FMIPA Univ. Padjadjaran, vol. 10, p. 08, 2021, [Online]. Available: http://prosiding.statistics.unpad.ac.id/index.php/prosidingnasional/article/view/75

[9] D. S. M. Simanjuntak, I. Gunawan, S. Sumarno, P. Poningsih, and I. P. Sari, "Penerapan Algoritma K-Medoids Untuk Pengelompokkan Pengangguran Umur 25 tahun Keatas Di Sumatera Utara," J. Krisnadana, vol. 2, no. 2, 2023, doi: 10.58982/krisnadana.v2i2.264.

[10] D. Widyadhana, R. B. Hastuti, I. Kharisudin, and F. Fauzi, "Perbandingan Analisis Klaster K-Means dan Average Linkage untuk Pengklasteran Kemiskinan di Provinsi Jawa Tengah," Prism. Pros. Semin. Nas. Mat., vol. 4, no. 2, pp. 584–594, 2021.

[11] N. Rosdiani and A. Hidayat, "Pengaruh Derivatif Keuangan, Konservatisme Akuntansi dan Intensitas Aset Tetap terhadap Penghindaran Pajak," Journal of Technopreneurship on Economics and Business Review, vol. 1, no. 2, pp. 1-10, 2020. E-ISSN: 2716-0092, P-ISSN: 2716-0106. [Online]. Tersedia: https://jtebr.unisan.ac.id/index.php/jtebr.

[12] J. I. Daoud, "Multicollinearity and Regression Analysis," J. Phys. Conf. Ser., vol. 949, no. 1, 2018, doi: 10.1088/1742-6596/949/1/012009.
[13] M. Muhajir, Modul Praktikum Statistika Multivariat Terapan, 1st ed., Yogyakarta, Universitas Islam Indonesia, 2021.
[14] D. U. Iswavigra, S. Defit, and G. W. Nurcahyo, "Data Mining dalam Pengelompokan Penyakit Pasien dengan Metode K-Medoids," J. Inf. dan Teknol., vol. 3, pp. 181–189, 2021, doi: 10.37034/jidt.v3i4.150.
[15] F. Harahap, "Perbandingan Algoritma K-Means dan K-Medoids untuk Clustering Kelas Siswa Tunagrahita," TIN Terap. Inform. Nusant., vol. 2, no. 4, pp. 191–197, 2021, [Online]. Available: https://ejurnal.seminar-id.com/index.php/tin/article/download/873/599
[16] H. Lailatul Ramadhania, L. Zakaria, and Nusyirwan, "Aplikasi Metode Sillhouette Coefficient, Metode Elbow dan Metode Gap Staticstic dalam Menentukan K Optimal pada Analisis K-Medoids," J. Siger Mat., vol. 04, no. 01, pp. 1–10, 2023.
[17] M. Z. Nasution, A. A. Nababan, K. U. Syaliman, M. S. Novelan, and M. Jannah, "Penerapan Principal Component Analysis (PCA) Dalam Penentuan Faktor Dominan Yang Mempengaruhi Pengidap Kanker Serviks (Studi Kasus: Cervical Cancer Dataset)," J. Mantik Penusa, vol. 3, no. 1, pp. 204–210, 2019.