JSRCS Journal of Students' Research in Computer Science

e-ISSN: 2722 – 290X
Vol. 5 No. 2 (November 2024), Page: 109 – 122
https://doi.org/10.31599/wbpydb62
Available Online at http://ejurnal.ubharajaya.ac.id/index.php/JSRCS

Analisis Clustering *K-Means* untuk Pemetaan Tingkat Pengangguran Terbuka di Provinsi-Provinsi Indonesia Tahun 2013-2023

Alif Izzuddin Ramadhan[1], **Prima Dina Atika**[1,*], **Khairunnisa Fadhilla Ramdhania**[1]

* Korespondensi: e-mail: prima.dina@dsn.ubharajaya.ac.id

[1] Informatika; Fakultas Ilmu Komputer; Universitas Bhayangkara Jakarta Raya; Jl. Perjuangan No. 81, Marga Mulya, Bekasi Utara, Bekasi, Jawa Barat 17143, Telp/Fax: (021) 88955882; e-mail: alifizudin7758@gmail.com,prima.dina@dsn.ubharajaya.ac.id, khairunnisa.fadhilla@dsn.ubharajaya.ac.id

Submitted : **5 September 2024**
Revised : **8 Oktober 2024**
Accepted : **7 November 2024**
Published : **30 November 2024**

Abstract

This study analyzes unemployment rates in Indonesian provinces using data from the Central Statistics Agency (BPS) for the period 2013-2023 and the K-Means clustering algorithm. The aim is to group regions based on the Open Unemployment Rate (TPT). Two main clusters were produced: one with a high unemployment rate (cluster 0) and one with a low unemployment rate (cluster 1). Cluster 0 consists of 12 provinces, while cluster 1 consists of 22 provinces. The model evaluation shows a Davies-Bouldin Index score of 0.7041, indicating good clustering quality. The clustering results are visualized in the form of a map for easy interpretation. This research is expected to help policymakers design more effective policies in reducing unemployment in Indonesia, provide deep insights into regional differences in terms of unemployment, and support targeted decision-making.

Keywords: *Cluster Mapping, K-Means, Open Unemployment, Regional Analysis*

Abstrak

Penelitian menganalisis tingkat pengangguran di provinsi-provinsi Indonesia menggunakan data Badan Pusat Statistik (BPS) periode 2013-2023 dan algoritma *K-Means clustering*. Tujuannya mengelompokkan daerah berdasarkan Tingkat Pengangguran Terbuka (TPT). Dua *cluster* utama dihasilkan satu dengan tingkat pengangguran tinggi (*cluster* 0) dan satu dengan tingkat pengangguran rendah (*cluster* 1). *Cluster* 0 terdiri dari 12 provinsi, sementara *cluster* 1 terdiri dari 22 provinsi. Evaluasi model menunjukkan skor Davies-Bouldin Index sebesar 0.7041, menunjukkan kualitas *clustering* yang baik. Hasil *clustering* divisualisasikan dalam bentuk peta untuk memudahkan interpretasi. Penelitian diharapkan membantu pembuat kebijakan merancang kebijakan yang lebih efektif dalam mengurangi pengangguran di Indonesia, memberikan wawasan mendalam tentang perbedaan regional dalam hal pengangguran, dan mendukung pengambilan keputusan yang tepat sasaran.

Kata kunci: Analisis Regional, *K-Means*, Pemetaan *Cluster*, Pengangguran Terbuka

1. Pendahuluan

Pengangguran merupakan masalah yang kompleks dan krusial di Indonesia. Dengan tingkat pengangguran tinggi yang dapat menurunkan kesejahteraan masyarakat, memperburuk kondisi sosial-ekonomi, dan mengancam stabilitas politik serta kemajuan negara. Ada beberapa jenis pengganggguran, salah satunya pengangguran terbuka. Pengangguran terbuka adalah

tidak memiliki pekerjaan namun sedang mencari pekerjaan, serta sedang mempersiapkan usaha, atau sudah diterima bekerja tetapi belum mulai bekerja. Tujuan penelitian untuk mendukung pengambilan keputusan yang tepat, sehingga pemerintah dan pihak terkait dapat merumuskan kebijakan yang efektif untuk mengatasi masalah (Maliki, 2022). Keanekaragaman ekonomi, geografi, dan populasi di Indonesia menyebabkan kondisi dan masalah pengangguran yang berbeda di setiap daerah. Informasi ini dapat digunakan untuk merumuskan kebijakan yang lebih sesuai dengan kondisi spesifik setiap daerah. (Tanjung et al., 2021).

Pada tahun 2023 tingkat pengangguran terbuka (TPT) secara keseluruhan mengalami penurunan signifikan sebesar 5.32%, turun 0.54% dari tahun 2022, mencerminkan peningkatan kinerja pemerintah dalam menangani pengangguran di berbagai daerah (Badan Pusat Statistik, 2023). Namun, data BPS tidak memberikan label pada tiap daerah untuk mengindikasikan apakah TPT di daerah tersebut rendah, normal, atau tinggi. Label penting karena dapat memberikan wawasan baru bagi pihak terkait dan menjadi bahan pertimbangan atau kajian lebih lanjut dalam menangani masalah pengangguran. Hasil analisis dari penelitian ini dapat membantu pemerintah, pembuat kebijakan, dan peneliti memahami pola dan karakteristik pengangguran di seluruh wilayah Indonesia. Analisis ini memberikan wawasan yang bermanfaat bagi semua pihak, baik untuk pertimbangan pengambilan keputusan maupun sebagai sumber analisis lebih lanjut.

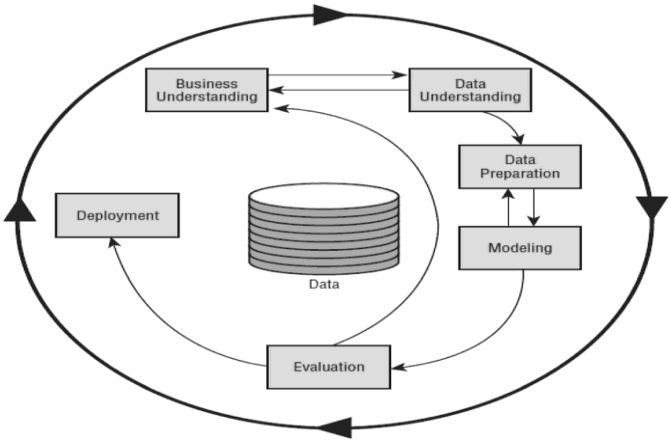
Algoritma *K-Means* merupakan alat analisis data yang efektif dalam mengidentifikasi pola dan kelompok dalam data pengangguran. Dengan mengelompokkan individu penganggur berdasarkan karakteristik serupa seperti usia, tingkat pendidikan, dan pengalaman kerja, algoritma ini membantu kita memahami dengan lebih baik akar permasalahan pengangguran. Hasil pengelompokan ini kemudian dapat digunakan sebagai dasar untuk merancang program dan kebijakan yang lebih tepat sasaran. Misalnya, jika ditemukan kelompok besar lulusan baru yang kesulitan mencari pekerjaan, maka program magang atau pelatihan keterampilan spesifik dapat menjadi solusi yang relevan. Namun, penerapan algoritma *K-Means* juga memiliki tantangan, seperti kualitas data yang digunakan dan kompleksitas dalam menginterpretasi hasil. Oleh karena itu, perlu adanya kolaborasi antara ahli data, pembuat kebijakan, dan para pemangku kepentingan lainnya untuk memastikan bahwa hasil analisis *K-Means* dapat diimplementasikan secara tepat dalam mengatasi masalah pengangguran.

Clustering merupakan instrumen untuk memecahkan masalah kompleks dalam ilmu komputer dan statistik. *Clustering* bekerja dengan mengelompokkan titik-titik data ke dalam dua kelompok atau lebih dan satu kelompok lebih mirip satu sama lain dibandingkan dengan kelompok data lainnya. Algoritma yang digunakan dalam penelitian ini adalah *K-Means*, karena algoritma ini dapat menghasilkan *cluster* yang optimal dengan konvergensi yang cepat (Akramunnisa & Fajriani, 2020). *K-Means* adalah metode pengelompokan dalam ilmu komputer dan statistik yang digunakan untuk menyelesaikan berbagai masalah pengelompokan data yang kompleks. Metode ini membentuk kelompok dari titik-titik data numerik yang memiliki

kemiripan tinggi satu sama lain dibandingkan dengan titik-titik data di kelompok lain. *K-Means* sangat cocok untuk data tanpa label atau *unsupervised learning*, menjadikannya pilihan yang efektif untuk analisis data yang membutuhkan identifikasi pola dan pengelompokan berdasarkan karakteristik intrinsik data (Badri & Habibi, 2022). Sehingga dengan beberapa hal di atas, maka peneliti melakukan penelitian yang berjudul "Pemetaan Pengangguran Terbuka di Provinsi Indonesia: Analisis *Cluster*ing *K-Means* Tahun 2013-2023".

2. Metode Penelitian

Kerangka kerja yang digunakan adalah *Cross Industry Standard Process for Data Mining* (*CRISP-DM*) (Larose, 2005). *CRISP-DM* merupakan metodologi terstruktur untuk ekstraksi pengetahuan yang populer di dunia industri. Dikembangkan oleh akademisi dan disempurnakan melalui pengalaman industri, dirancang untuk memenuhi kebutuhan industri namun tetap fleksibel untuk digunakan oleh pihak lain. Model *CRISP-DM* menawarkan beberapa keunggulan, seperti strukturnya yang jelas, cakupan semua tahapan penting dalam proses ekstraksi pengetahuan, fleksibilitasnya untuk berbagai proyek, dan efektivitasnya dalam menghasilkan model yang bermanfaat (Chapman et al., 2000).



Sumber: Chapman et al (2000)

Gambar 1. Tahap CRIPS-DM

2.1. Business Understanding

Business understanding pada penelitian adalah tujuan dari penelitian, yaitu membuat pengelompokan antar provinsi di Indonesia berdasarkan Tingkat Pengangguran Terbuka (TPT) menggunakan algoritma *K-Means Clustering*, sehingga hasil dari penelitian ini dapat dijadikan gambaran serta sumber penelitian lebih lanjut untuk mengetahui pengelompokan daerah berdasarkan tinggi dan rendahnya angka pengangguran.

2.2. Data Understanding

Data yang digunakan yaitu Provinsi, Tingkat Pengangguran Terbuka (TPT) – Februari, Tingkat Pengangguran Terbuka (TPT) Agustus, Tingkat Partisipasi Angkatan Kerja (TPAK) – Februari, Tingkat Partisipasi Angkatan Kerja (TPAK) – Agustus, Serta dengan jumlah baris atau *record* sebanyak total 385 *record* dari 11 *dataset* yang digunakan.

2.3. Data Preparation

Pengelolahan Data menggunakan *Visual code studio* dengan bahasa pemograman *Python* 3.11.6. Data dibersihkan dengan cara mengurangi atau memperbaiki untuk menyempurnakan data yang akan dipakai untuk tahap Modeling. Adapun langkah dalam melakukan *Data Cleaning* adalah (a) Melakukan *drop* kolom dan *row* yang tidak terpakai; (b) Pergantian nama kolom serta penyesuaian format beberapa *value* pada data untuk memudahkan proses pengolahan dan permodelan data; (c) *Handling missing value* atau mencari dan melakukan handling pada nilai yang hilang pada data; dan (d) *Data Integration* dengan melakukan *merging* pada keseluruhan *dataset* yang digunakan.

2.4. Modelling

Peneliti mulai melakukan permodelan berdasarkan model yang telah di tetapkan untuk penelitian pada tahapan ini, yaitu *K-Means Clustering*. Metode *K-Means Clustering* merupakan metode yang melakukan pengelompokan terhadap *data point* berdasarkan hasil pencarian pusat data atau *centroid* yang merepresentasikan sejumlah kelompok data (Chopra & Khurana, 2023) sedangkan menurut (Yuxi, 2019) [8] tujuan dari algoritma *K-Means* adalah model yang mempartisi data ke dalam *k cluster* berdasarkan fitur kesamaan. K adalah properti yang telah ditentukan dari model pengelompokan *K-Means*, masing-masing dari *k cluster* ditentukan oleh sebuah *centroid* (pusat *cluster*) dan setiap sampel data termasuk dalam *cluster* dengan *centroid* terdekat. Selama proses *training*, algoritma secara iteratif memperbarui k *centroid* berdasarkan data yang disediakan.

Metode *K-Means* sangat sederhana untuk dijalankan dan diimplementasikan, memiliki proses yang cepat, mudah dipergunakan, dan mudah berdaptasi serta umum digunakan dalam berbagai aplikasi kecil hingga menengah. Secara ilmiah, *K-Means* menjadi salah satu metode yang banyak digunakan dalam *data mining*. Metode *K-Means Clustering* merupakan salah satu dari teknik pengelompokan dengan metode nonhirarki yang tujuannya mengelompokkan objek yang diawali dengan mengindentifikasi data yang akan di *Cluster*. Langkah-langkah cara kerja metode *K-Means Clustering* (Aggarwal, 2015):

a. Tentukan Jumlah *cluster* (K), model *K-Means* perlu mengetahui berapa jumlah *cluster* yang perlu di proses sebagai hasil dari *modelling* yang dilakukan.
b. Menginisialisasi *centroids* atau titik pusat *cluster*, model akan memulai proses modelling dengan memilih data *point* secara random untuk digunakan sebagai *centroids* awal sesuai dengan K *cluster* yang ditentukan sebelumnya.
c. Hitung jarak antar setiap data ke dalam *cluster* dengan jarak yang paling pendek dengan menggunakan persamaan ukuran jarak *Euclidean Distance* dengan persamaan:

$$d(x_i, y_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - y_{jk})^2} \quad (1)$$

𝑑(x_i y_j) jarak antara objek data x_i dan y_j, n dimensi atau jumlah data, x_ik nilai dari titik data x_i pada dimensi ke-k, y_jk nilai dari titik data y_j pada dimensi ke- k, ∑_(k=1)^n

melakukan penjumlahan terhadap semua nilai dari k mulai dari 1 sampai dengan n yang merupakan jumlah dimensi.

d. Kelompokan data berdasarkan nilai jarak yang sudah dihitung pada tahap sebelumnya sehingga data menjadi berkelompok berdasarkan K klaster yang telah di tentukan sebelumnya.
e. Memperbarui *centroid*, dengan cara menghitung ulang *centroid* pada setiap *cluster* yang merupakan rata-rata dari semua sample yang ada di dalam *cluster*. *Centroid* di perbarui menjadi rata-rata *cluster* yang sesuai menggunakan persamaan:

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N_k} x_i \quad (2)$$

μ_k *centroid* cluster ke-k, N_k jumlah titik data dalam cluster ke- k, ∑_(i=1)^(N_k) penjumlahan dari seluruh titik data dalam cluster ke- k, x_i titik data ke- q dalam cluster ke-k.

f. Ulangi langkah 2 sampai dengan 5 hingga *centroid* tidak berubah lagi atau berubah dengan jarak yang sedikit, hasil cluster tidak berubah atau hingga iterasi telah dirasa cukup banyak.

2.5 Evaluation

Proses *modelling* yang telah dilakukan di ukur ke akuratan nya dalam memenuhi tujuan bisnis menggunakan metode *Davies Bouldin Index* untuk kemudian dilakukan evaluasi berdasarkan hasil yang didapatkan. *Davies Bouldin Index* merupakan metrik evaluasi model yang digunakan untuk mengevaluasi model *clustering. Davies Bouldin Index* bekerja dengan mengevaluasi seberapa baik proses *clustering* yang dilakukan berdasarkan fitur yang ada pada data yang digunakan (Tempola & Assagaf, 2018). Nilai yang dihasilkan dari evaluasi menggunakan *Davies Bouldin Index* didasarkan pada ukuran kesamaan *cluster* yang basisnya adalah ukuran dispresi atau dapat disebut perpindahan *cluster* (Tempola & Assagaf, 2018). Persamaan yang digunakan pada metrik evaluasi adalah k jumlah cluster yang digunakan pada modelling, R_(i,j) Rasio antara i dan j cluster.

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij} \quad (3)$$

Nilai pada persamaan diatas diperoleh dari komponen kohesi atau kedekatan antar data *point* dalam satu cluster dan separasi atau seberapa jauh jarak antar cluster yang berbeda. Hasil cluster yang baik adalah cluster dengan kohesi terendah dan separasi tertinggi.

3. Hasil dan Pembahasan

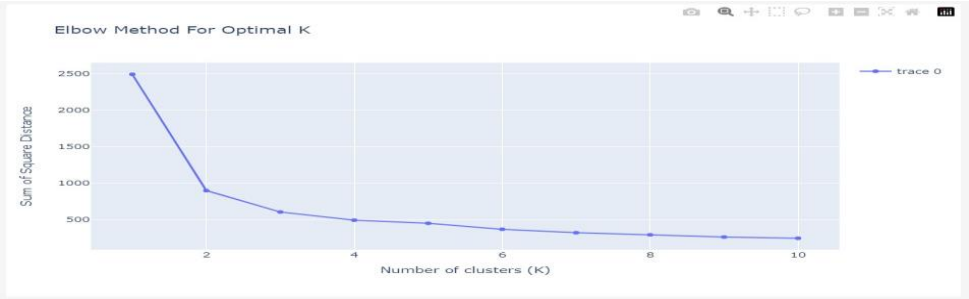
Penelitian menggunakan algoritma *K-Means cluster*ing untuk proses pemodelan. Algoritma ini melibatkan beberapa langkah penting untuk mencapai hasil akhir berupa *cluster*. Langkah-langkah tersebut termasuk menentukan jumlah optimal dari K atau *cluster* untuk model, melakukan pemodelan, dan melakukan evaluasi dengan menggunakan *Davies Bouldin Index* untuk menilai kualitas model dan *cluster* yang dihasilkan.

Tujuan dari pemodelan dengan algoritma *K-Means clustering* adalah untuk menemukan pola dan menentukan *cluster* berdasarkan data yang digunakan. Algoritma bekerja dengan cara menentukan K atau jumlah *cluster* yang akan digunakan, yang telah ditentukan dalam proses sebelumnya. Setelah jumlah K *cluster* ditentukan, langkah berikutnya adalah menentukan *centroid* atau titik pusat data awal secara acak. Kemudian, menghitung jarak setiap titik data terhadap *centroid* menggunakan euclidean distance dan menentukan *cluster* data berdasarkan nilai minimum hasil perhitungan titik data ke setiap *centroid*. Proses ini diulangi sampai posisi *centroid* tidak berubah lagi.

Penelitian berfokus pada penggunaan data sekunder yang diperoleh dari situs resmi Badan Pusat Statistik. Badan Pusat Statistik secara rutin merilis informasi dan data statistik terkait kondisi ekonomi, sosial, dan demografi Indonesia, serta beberapa data spesifik seperti investasi, populasi, dan lainnya, termasuk Tingkat Pengangguran Terbuka tahun 2013-2023 yang digunakan pada penelitian ini. Seluruh data yang disediakan oleh Badan Pusat Statistik di situs resminya dapat diakses secara bebas oleh publik untuk keperluan penelitian, informasi, dan berbagai kebutuhan lainnya, sehingga peneliti dapat memberikan pemahaman yang lebih mendalam hasil analisis pengelompokan dengan menggunakan data resmi dari Badan Pusat Statistik.

3.1. *Elbow Method*

Penelitian menggunakan *elbow method* sebagai metode untuk mencari jumlah *cluster* yang optimal. Cara kerja *elbow method* adalah dengan melakukan iterasi sebanyak K=1 sampai dengan K=n untuk mencari titik “siku” pada grafik yang menampilkan nilai *Within Cluster Sum of Square* (WCSS) atau secara singkat dapat diartikan sebagai penjumlahan jarak kuadrat antara seluruh titik data dengan *centroid* yang ditentukan berdasarkan nilai dari K yang digunakan.



Sumber: Hasil Penelitian (2024)

Gambar 2. Grafik Elbow Method

Dari visualisasi grafik dapat dilihat semakin banyak jumlah K yang digunakan, maka jarak antara data dengan *centroid* akan semakin kecil. Sebagai contoh pada K=2 total jarak setiap data dengan *centroid* adalah 893.8832, berbeda dengan K=3 yang memilki total jarak setiap data pada *centroid* sebesar 600.8467. Begitupun seterusnya, semakin banyak jumlah K yang digunakan maka akan semakin kecil pula jumlah jarak setiap data terhadap *centroid* yang digunakan. Elbow method dinamakan demikian dikarenakan cara menentukan K yang optimal secara visual adalah dengan melihat garis “siku” pada scatter plot yang telah dibuat. Seperti

pada Gambar 2 dapat dilihat garis siku terletak pada K=2 sehingga dapat disimpulkan jumlah *cluster* yang akan digunakan adalah 2 *cluster*.

3.2. K-Means modelling

Tahap pertama pada proses *modelling* setelah menentukan jumlah *cluster* adalah menentukan *centroid* awal dengan memilih data secara acak pada *dataset* yang digunakan, pada penelitian ini 2 *centroid* dipilih secara acak dari *dataset* yang telah diproses sebelumnya, adapun data yang dipilih oleh peneliti adalah:

Tabel 1. *Centroid* awal

No	Provinsi	*Centroid*
1	Sulawesi Utara	0
2	Kalimantan Tengah	1

Sumber: Hasil Penelitian (2024)

Centroid yang telah di pilih secara acak dari *dataset* Tingkat Pengangguran Terbuka, akan digunakan sebagai *centroid* iterasi 1 yang selanjutnya akan dihitung jaraknya terhadap setiap titik data yang ada menggunakan rumus *euclidean distance* mulai dari data ke-1 terhadap *centroid cluster* 0 dan 1, hingga data ke-n terhadap *centroid cluster* 0 dan 1. Sebagai contoh menghitung jarak antara data ke-1, 2 dan 3 terhadap *centroid cluster* 0 dan 1.

Data ke-1 adalah provinsi Aceh terhadap Sulawesi Utara dengan *centroid* 0 (C_0).

$$d(x_i, y_j) = \sqrt{\sum_{k=1}^{2} (x_{ik} - y_{jk})^2}$$

$$d(x_1, y_0) = \sqrt{(x_{1_1} - y_{0_1})^2 + (x_{1_2} - y_{0_2})^2 + \cdots + (x_{1_{22}} + y_{0_{22}})^2}$$

$$d(x_1, y_0) = \sqrt{(8.34 - 7.50)^2 + (10.12 - 6.79)^2 + \cdots + (6.03 - 6.10)^2}$$

$$d(x_1, y_0) = 4.846730857$$

Jarak data ke-1 yaitu provinsi Aceh terhadap Kalimantan Tengah dengan *centroid* 1 (C_1).

$$d(x_i, y_j) = \sqrt{\sum_{k=1}^{2} (x_{ik} - y_{jk})^2}$$

$$d(x_1, y_1) = \sqrt{(x_{1_1} - y_{1_1})^2 + (x_{1_2} - y_{1_2})^2 + \cdots + (x_{1_{22}} + y_{1_{22}})^2}$$

$$d(x_1, y_1) = \sqrt{(8.34 - 1.81)^2 + (10.12 - 3.00)^2 + \cdots + (6.03 - 4.10)^2}$$

$$d(x_1, y_1) = 17.35603353$$

Selanjutnya jarak data ke-2 yaitu provinsi Sumatera Utara terhadap C_0

$$d(x_i, y_j) = \sqrt{\sum_{k=1}^{2} (x_{ik} - y_{jk})^2}$$

$$d(x_2, y_0) = \sqrt{(x_{2_1} - y_{0_1})^2 + (x_{2_2} - y_{0_2})^2 + \cdots + (x2_{22} + y0_{22})^2}$$

$$d(x_2,y_0) = \sqrt{(6.09-7.50)^2 + (6.45-6.79)^2 + \cdots + (5.89-6.10)^2}$$

$$d(x_2,y_0) = 5.235943086$$

Jarak data ke-2 yaitu provinsi Sumatera Utara terhadap C_1

$$d(x_i,y_j) = \sqrt{\sum_{k=1}^{2}(x_{ik}-y_{jk})^2}$$

$$d(x_2,y_1) = \sqrt{(x_{2_1}-y_{1_1})^2 + (x_{2_2}-y_{1_2})^2 + \cdots + (x_{2_{22}}+y_{1_{22}})^2}$$

$$d(x_2,y_1) = \sqrt{(6.09-1.81)^2 + (6.45-3.00)^2 + \cdots + (5.89-4.10)^2}$$

$$d(x_2,y_1) = 11.26048844$$

Selanjutnya jarak data ke-3 yaitu Sumatera Barat terhadap C_0

$$d(x_i,y_j) = \sqrt{\sum_{k=1}^{2}(x_{ik}-y_{jk})^2}$$

$$d(x_3,y_0) = \sqrt{(x_{3_1}-y_{0_1})^2 + (x_{3_2}-y_{0_2})^2 + \cdots + (x_{3_{22}}+y_{0_{22}})^2}$$

$$d(x_3,y_0) = \sqrt{(6.39-7.50)^2 + (7.02-6.79)^2 + \cdots + (5.94-6.10)^2}$$

$$d(x_3,y_0) = 5.064711245$$

Jarak data ke-3 yaitu Sumatera Barat terhadap C_1

$$d(x_i,y_j) = \sqrt{\sum_{k=1}^{2}(x_{ik}-y_{jk})^2}$$

$$d(x_3,y_1) = \sqrt{(x_{3_1}-y_{1_1})^2 + (x_{3_2}-y_{1_2})^2 + \cdots + (x_{3_{22}}+y_{1_{22}})^2}$$

$$d(x_3,y_1) = \sqrt{(6.39-1.81)^2 + (7.02-3.00)^2 + \cdots + (5.94-4.10)^2}$$

$$d(x_3,y_1) = 11.76636732$$

Hasil perhitungan jarak *data point* terhadap masing masing *centroid cluster* menggunakan rumus *euclidean distance* pada data ke-1 yaitu provinsi Aceh menghasilkan nilai jarak tehadap C0 = 4,85 dan jarak data ke-1 terhadap C1 = 17,36 kemudian Jarak data ke-2 terhadap C0 = 5,24 dan jarak data ke-2 terhadap C1 = 11,26. Adapun jarak data ke-3 terhadap C0 = 5,06 dan C1 = 11,77. Selanjutnya perhitungan jarak menggunakan *eudlidean distance* akan dilakukan pada data ke-4 hingga data ke-n atau pada *dataset* yang digunakan hingga data ke-34, kemudian dari hasil yang didapatkan langkah selanjutnya adalah mengambil nilai minimum atau terkecil di antara jarak data terhadap C0 dan C1 untuk kemudian ditetapkan *cluster* sementara bedasarkan nilai terkecil yang didapatkan.

Hasil perhitungan yang telah didapatkan dari perhitungan di iterasi ke-1 akan dijadikan sebagai acuan untuk melakukan perhitungan *centroid* baru di iterasi ke-2, dan begitu seterusnya perhitungan terus dilakukan sampai nilai *centroid cluster* atau anggota *cluster* tidak berubah. Pada iterasi ke-2 langkah pertama yang harus dilakukan adalah dengan menentukan

nilai *centroid* baru untuk C0 dan C1, yang didapatkan dari menghitung rata-rata dari seluruh data *point* yang menjadi anggota *cluster* sementara yang dihasilkan perhitungan iterasi-1. Adapun perhitungan yang dilakukan adalah:

$$C0 = \left(\frac{8.34 + 6.09 + \cdots + 4.36}{13}\right), \left(\frac{10.12 + 6.45 + \cdots + 4.40}{13}\right), \ldots\ldots\ldots$$

$$\left(\frac{6.03 + 5.89 + \cdots + 5.38}{13}\right) = \{(7.15), (7.40), \ldots (5.99)\}$$

$$C1 = \left(\frac{2.89 + 5.41 + \cdots + 2.91}{21}\right), \left(\frac{4.76 + 4.84 + \cdots + 3.15}{21}\right), \ldots\ldots\ldots$$

$$\left(\frac{2.50 + 3.84 + \cdots + 3.48}{21}\right) = \{(3.62), (4.07), \ldots (3.76)\}$$

Proses perhitungan terus dilakukan sampai perhitungan mencapai konvergen, atau dengan kata lain *centroid* tidak mengalami perubahan perhitungan iterasi yang dilakukan. Adapun pada penelitian, perhitungan mencapai konvergen pada iterasi ke-3 dengan hasil pada Tabel 2.

Tabel 2. Hasil Perhitungan Iterasi Terakhir

No	Provinsi	C0	C1	Min	Cluster
1	Aceh	4.59	15.51	4.59	0
2	Sumatera Utara	5.16	9.53	5.16	0
3	Sumatera Barat	4.91	9.91	4.91	0
4	Riau	7.94	8.71	7.94	0
5	Jambi	14.02	2.70	2.70	1
6	Sumatera Selatan	11.48	3.35	3.35	1
7	Bengkulu	17.47	3.76	3.76	1
8	Lampung	12.00	3.45	3.45	1
9	Kepulauan Bangka Belitung	13.98	3.70	3.70	1
10	Kepulauan Riau	7.03	17.76	7.03	0
11	Dki Jakarta	6.29	17.44	6.29	0
12	Jawa Barat	7.41	21.50	7.41	0
13	Jawa Tengah	9.43	5.82	5.82	1
14	Di Yogyakarta	16.89	3.00	3.00	1
15	Jawa Timur	12.61	2.82	2.82	1
16	Banten	8.20	22.31	8.20	0
17	Bali	22.13	9.06	9.06	1
18	Nusa Tenggara Barat	14.43	4.32	4.32	1
19	Nusa Tenggara Timur	18.15	4.01	4.01	1
20	Kalimantan Barat	12.41	3.55	3.55	1
21	Kalimantan Tengah	15.95	3.00	3.00	1
22	Kalimantan Selatan	13.33	2.60	2.60	1
23	Kalimantan Timur	4.19	15.81	4.19	0
24	Kalimantan Utara	10.18	5.63	5.63	1

No	Provinsi	C0	C1	Min	Cluster
25	Sulawesi Utara	2.87	13.91	2.87	0
26	Sulawesi Tengah	17.13	3.13	3.13	1
27	Sulawesi Selatan	8.29	6.71	6.71	1
28	Sulawesi Tenggara	16.26	2.71	2.71	1
29	Gorontalo	16.50	3.45	3.45	1
30	Sulawesi Barat	20.74	6.68	6.68	1
31	Maluku	5.01	16.94	5.01	0
32	Maluku Utara	10.88	4.96	4.96	1
33	Papua Barat	7.53	10.29	7.53	0
34	Papua	17.13	3.32	3.32	1

Sumber: Hasil Penelitian (2024)

3.3. Hasil *Cluster*ing

Hasil modelling yang telah dilakukan sebelumnya akan dianalisis lebih lanjut untuk menetukan *cluster* mana yang akan dikategorikan sebagai *cluster* dengan Tingkat Pengangguran Terbuka yang tinggi dan *cluster* mana yang akan dikategorikan sebagai *cluster* dengan Tingkat Pengangguran Terbuka yang rendah. Tabel 3 adalah *dataset* akhir yang dihasilkan dari proses *modelling* yang telah dilakukan.

Tabel 3. Hasil *Labelling*

No	Provinsi	Cluster	Label
1	Aceh	0	Tinggi
2	Sumatera Utara	0	Tinggi
3	Sumatera Barat	0	Tinggi
4	Riau	0	Tinggi
5	Jambi	1	Rendah
6	Sumatera Selatan	1	Rendah
7	Bengkulu	1	Rendah
8	Lampung	1	Rendah
9	Kepulauan Bangka Belitung	1	Rendah
10	Kepulauan Riau	0	Tinggi
11	Dki Jakarta	0	Tinggi
12	Jawa Barat	0	Tinggi
13	Jawa Tengah	1	Rendah
14	Di Yogyakarta	1	Rendah
15	Jawa Timur	1	Rendah
16	Banten	0	Tinggi
17	Bali	1	Rendah
18	Nusa Tenggara Barat	1	Rendah
19	Nusa Tenggara Timur	1	Rendah
20	Kalimantan Barat	1	Rendah
21	Kalimantan Tengah	1	Rendah

No	Provinsi	Cluster	Label
22	Kalimantan Selatan	1	Rendah
23	Kalimantan Timur	0	Tinggi
24	Kalimantan Utara	1	Rendah
25	Sulawesi Utara	0	Tinggi
26	Sulawesi Tengah	1	Rendah
27	Sulawesi Selatan	1	Rendah
28	Sulawesi Tenggara	1	Rendah
29	Gorontalo	1	Rendah
30	Sulawesi Barat	1	Rendah
31	Maluku	0	Tinggi
32	Maluku Utara	1	Rendah
33	Papua Barat	0	Tinggi
34	Papua	1	Rendah

Sumber: Hasil Penelitian (2024)

Langkah terakhir pada proses *modelling* di penelitian ini adalah membuat visualisasi graf dalam bentuk *geo map* menggunakan data hasil *modelling* yang telah dilakukan. Visualisasi dilakukan menggunakan aplikasi Tableau yang biasa digunakan untuk membuat *dashboard* hasil analisis data. Gambar 3 hasil visualisasi yang dilakukan menggunakan aplikasi Tableau. Visulisasi yang dihasilkan akan menampilkan bagaimana pesebaran *cluster* Tingkat Pengangguran Terbuka, dengan *map* berwarna hijau menandakan daerah dengan *cluster* 0 atau Tingkat Pengangguran Terbuka yang tinggi sementara warna merah menandakan daerah dengan *cluster* 1 atau Tingkat Pengangguran Terbuka yang tinggi berdasarkan hasil dari *script*.

Sumber: Hasil Penelitian (2024)

Gambar 3. Hasil Visualisasi *Geo Map*

3.4. *Evaluation*

Proses evaluasi merupakan proses terakhir yang dilakukan pada penelitian, evaluasi dilakukan dengan tujuan melihat kualitas dari kinerja model *clustering* dengan cara menghitung kedekatan jarak antar *data point,* kedekatan jarak antara *data point* terhadap *centroid cluster*-nya, atau kedekatan jarak antar *cluster.* Penelitian menggunakan *Davies Bouldin Index* (DBI) sebagai metriks evaluasi untuk mengukur kualitas model yang telah dibuat. DBI melakukan

evaluasi dengan mengukur seberapa baik *cluster* telah dipisahkan satu sama lain, semakin rendah atau mendekati nilai 0 *score* DBI maka semakin bagus model tersebut. Langkah-langkah dalam menghitung DBI.

a. Menghitung *Sum of Square Within Cluster* (*SSW*) atau mengukur seberapa mirip atau dekat *data point* dalam suatu *cluster* terhadap *centroid* nya, dapat disebut juga sebagai nilai kohesi. Hal ini dilakukan dengan persamaan 4.

$$SSW = \frac{1}{m} \sum_{i=1}^{k} d\left(x_i\, y_j\right) \quad (4)$$

Sebagai contoh menghitung SSW pada data ke-1 terhadap *centroid cluster* nya

$$d(x_i, y_j) = \sqrt{\sum_{k=1}^{2} (x_{ik} - y_{jk})^2}$$

$$d(x_i, y_j) = \sqrt{(x_{i1} - y_{j1})^2 + (x_{i2} - y_{j2})^2 + \cdots + (x_{i22} + y_{j22})^2}$$

$$d(x_i, y_j) = \sqrt{(8.34 - 7.26)^2 + (10.12 - 7.59)^2 + \cdots + (6.03 - 6.12)^2}$$

$$d(x_i, y_j) = 4.5878693$$

Selanjutnya perhitungan persamaan 4 dilakukan pada seluruh *data point* yang ada di *dataset.* Langkah selanjutnya adalah menentukan nilai SSW pada masing-masing *cluster* dengan cara menghitung rata-rata jarak dari tiap anggota *cluster* yang ada. SSW *cluster* 0

$$SSW_0 = \frac{4.5878693 + 5.1555386 + \cdots + 7.5259437}{12} = 5.926915686$$

SSW *cluster 1*

$$SSW_1 = \frac{2.6969678 + 3.3481608 + \cdots + 3.3183387}{22} = 4.170091675$$

b. Menghitung *Sum of Square Between Cluster* (SSB) untuk mengetahui seberapa jauh *centroid* antar *cluster* tersebar atau dapat disebut juga sebagai nilai separasi, dengan menggunakan persamaan.

$$SSB_{ij} = d(y_i, y_j) \quad (5)$$

$$SSB = \sqrt{(7.26 - 3.73)^2 + (7.59 - 4.12)^2 + \cdots + (6.12 - 3.79)^2}$$

$$SSB = 14.33960862$$

Perhitungan persamaan 5 dilakukan sebanyak *cluster* yang ada, pada penelitian dikarenakan hanya menggunakan 2 *cluster* maka perhitungan cukup dilakukan satu kali dengan hasil Tabel 4.

Tabel 4. Hasil Perhitungan SSB

SSB	C0	C1
C0	0	1.433.960.862
C1	1.433.960.862	0

Sumber: Hasil Penelitian (2024)

c. Setelah nilai SSW dan SSB telah ditemukan, maka langkah selanjutnya adalah pengukuran nilai *rasio* untuk mengetahui nilai perbandingan variabilitas antara nilai dalam *cluster* dengan nilai antar *cluster,* semakin kecil nilai rasio maka semakin baik. Pengukuran rasio dapat dilakukan dengan persamaan.

$$R_{ij} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \quad (6)$$

$R_{ij} = \frac{5.926915686 + 4.170091675}{14.33960862} = \mathbf{0.704134097}$

d. Nilai rasio yang telah didapatkan, selanjutnya digunakan untuk menghitung nilai DBI dengan persamaan.

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij} \quad (7)$$

$DBI = \frac{1}{2}(0.704134097 + 0.704134097) = \mathbf{0.704134097}$

4. Kesimpulan

Berdasarkan penelitian yang telah dilakukan, ditemukan bahwa metode Elbow yang digunakan untuk menentukan jumlah *cluster* optimal menunjukkan "siku" pada K=2. Hal ini mengindikasikan bahwa K=2 adalah *jumlah cluster* yang optimal, yang juga didukung oleh skor *Davies Bouldin Score* yang lebih tinggi dibandingkan dengan jumlah K lainnya. Hasil *cluster*ing menghasilkan dua *cluster*, yaitu *cluster* 0 dengan 12 anggota dan *cluster* 1 dengan 22 anggota. Proses *labelling* dilakukan berdasarkan hasil modelling dengan melakukan *Exploratory Data Analysis (EDA)* dan analisis statistik. Dari hasil analisis, bahwa *cluster* 0 memiliki nilai mean atau rata-rata yang lebih besar dibandingkan dengan *cluster* 1, sehingga *cluster* 0 diberi label "Tinggi" dan *cluster* 1 diberi label "Rendah" dalam konteks Tingkat Pengangguran Terbuka. Evaluasi model dilakukan menggunakan Davies Bouldin Index dengan nilai akhir sebesar 0.7041340965671974, yang menunjukkan kualitas model yang baik mengingat semakin mendekati nilai 0 pada skor DBI, maka semakin baik model *cluster*ing tersebut.

Daftar Pustaka

Aggarwal, C. C. (2015). *Data Mining*. Springer International Publishing. https://doi.org/10.1007/978-3-319-14142-8

Akramunnisa, & Fajriani. (2020). K-Means Clustering Analysis pada PersebaranTingkat Pengangguran Kabupaten/Kota di Sulawesi Selatan. *Jurnal Varian*, *3*(2), 103–112. https://doi.org/10.30812/varian.v3i2.652

Badan Pusat Statistik. (2023). *Infografis Tingkat PengangguranTerbuka 2023*. Badan Pusat Statistik.

Badri, F., & Habibi, A. (2022). Implementasi Metode K-Means Clustering Analysis pada Pengelompokan Pengangguran di Indonesia sebagai Dampak dari Pandemi Covid-19.

ILKOMNIKA: Journal of Computer Science and Applied Informatics, *4*(2), 171–179. https://doi.org/10.28926/ilkomnika.v4i2.471

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinatz, T., Searer, C., & Wirth, R. (2000). *CRIPS DM 1.0 Step by Step Data Mining Guide*.

Chopra, D., & Khurana, R. (2023). Introduction to Machine Learning with Python. In *Introduction to Machine Learning with Python* (1 st). BENTHAM SCIENCE PUBLISHERS. https://doi.org/10.2174/9789815124422123010l

Larose, D. T. (2005). *An Introduction to Data Mining. In Discovering Knowledge in Data: An Introduction to Data Mining*. *4*(3).

Maliki, R. (2022). Perbandingan Tingkat Pengangguran Terbuka Provinsi di Indonesia Berbasis Metode K-Means Clustering. *Computer Science (CO-SCIENCE)*, *2*.

Tanjung, F. A., Windarto, A. P., & Fauzan, M. (2021). Penerapan Metode K-Means Pada Pengelompokkan Pengangguran Di Indonesia. *Jurasik (Jurnal Riset Sistem Informasi Dan Teknik Informatika)*, *6*(1), 61. https://doi.org/10.30645/jurasik.v6i1.271

Tempola, F., & Assagaf, A. F. (2018). Clustering of Potency of Shrimp In Indonesia With K-Means Algorithm And Validation of Davies-Bouldin Index. *Proceedings of the International Conference on Science and Technology (ICST 2018)*. https://doi.org/10.2991/icst-18.2018.148

Yuxi, H. L. (2019). *Python machine learning by example : easy-to-follow examples that get you up and running with machine learning*. Packt Publishing.