ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v23i1.14464

APPLICATION OF K-MEANS AND Z-SCORE METHODS FOR UNEMPLOYMENT CLUSTERING IN REGENCIES AND CITIES OF WEST JAWA PROVINCE

Penerapan Metode *K-Means* Dan *Z-Score* Pada Pengelompokan Pengangguran Kabupaten Dan Kota Di Provinsi Jawa Barat.

Vanka Angelica Putri[1], Budi Santosa, S.Si, M.T.[2]

[1,2] Informatika, Universitas Pembangunan Nasional Veteran Yogyakarta, Indonesia

[1*]123200061@student.upnyk.ac.id, *[2]dissan@upnyk.ac.id*
**: Penulis korenspondensi (corresponding author)*

Informasi Artikel
Received:
Revised:
Accepted:
Published:

Abstract
Unemployment is one of the pressing issues faced by by Indonesia as a developing country. According to data from Badan Pusat Statistik (BPS), unemployment is a significant problem in West Java Province, which consistently reports an unemployment rate above the nasional average compared to other provinces in the country. Therefore, addressing this issue in West Jawa Province is critical. To tackle this challenge, an analysis of districts and cities within West Java Province is critical. To tackle this chalengge, an analysis of districts and cities within West Java is necessary to identify areas with high unemployment rates by applying clustering techniques. Clustring encompasses a variety of methods, particularly within the realms of supervised and unsupervised approaches. One prominent unsupervised clustering technique is K-Means. This method is widely used in addressing social issues due to its simplicity, quick convergence, computational efficiency, and ease of implementation. This study aims to employ Z-Score standardization prior to applying the K-Means algorithm for clustering and to evaluate the impact of Z-Score standardization on the clustering process itself. In this research, experiments were conducted by testing the hyperparameter k ranging from 2 to 10 using Silhouette Score for evaluation. The optimal result was achieved with k = 3, yielding a score of 4.305. The application of Z-Score standardization played a significant role in preventing data dominance and ensuring unbiased clustering outcomes, as it normalized the varying scales of the variables.

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v23i1.14464

Keywords: YOLOv8, EasyOCR, Vehicle License Plate Detection, Text Recognition, Indonesia
Kata kunci: YOLOv8, EasyOCR, Deteksi Plat Kendaraan, Pengenalan Karakter

Abstrak

Pengangguran menjadi salah satu masalah yang ada di Indonesia sebagai negara berkembang. Sesuai data Badan Pusat Statistik (BPS), pengangguran merupakan salah satu permasalahan uang ada di Provinsi Jawa Barat dengan salah satu persentase pengangguran yang tinggi berada di atas rata-rata nasional di antara provinsi lain yang ada di Indonesia. Sehingga menjadi salah satu permasalahan yang harus diselesaikan di Provinsi Jawa Barat. Oleh karena itu, perlu dilakukan analisis pada kabupaten dan kota di Jawa Barat untuk dapat mengetahui daerah yang memiliki yang memiliki tingkat pengangguran yang tinggi dengan menerapkan teknik *clustering*. *Clustering* memiliki banyak jenis, terutama dalam pendekatan *supervised* atau *unsupervised*. Salah satu teknik *clustering unsupervised* yaitu *K-Means*. *K-Means* populer digunakan dalam masalah sosial dan memiliki kelebihan seperti sederhana, konvergensinya cepat, komputasinya efisien, dan mudah diimplementasikan. Penelitian ini bertujuan untuk menerapkan standarisasi *Z-Score* sebelum menerapkan algoritma *K-Means* dalam pengelompokan dan bagaimana pengaruh *Z-Score* dalam *clustering* itu sendiri. Pada penelitian ini pengujian dilakukan dengan menguji *hyperparameter* k 2-10 menggunakan *Silhouette Score*. Berdasarkan pengujian hasil terbaik berada pada k-3 dengan nilai 4.305. Dengan penerapan standarisasi *Z-Score* membantu data untuk menghindari dominasi data dan memberikan hasil *cluster* yang tidak bias karena perbedaan skala nilai tiap variabel.

1. Pendahuluan

Pengangguran menjadi salah satu masalah yang ada di Indonesia sebagai negara berkembang. Keadaan saat seseorang ingin bekerja, atau sedang mencari pekerjaan tetapi tidak bisa mendapatkan pekerjaan yang sesuai disebut sebagai pengangguran (Tanjung et al., 2021). Sesuai Badan Pusat Statistik, pengangguran merupakan salah satu permasalahan yang ada di Provinsi Jawa Barat dengan salah satu persentase pengangguran yang tinggi berada di atas rata-rata nasional di antara provinsi lain yang ada di Indonesia. Sehingga, masalah pengangguran menjadi salah satu masalah berat yang harus diselesaikan, khususnya di Provinsi Jawa Barat (Azzahra et al., 2023). Oleh karena itu, perlu dilakukan analisis pada kabupaten dan kota di Jawa Barat untuk dapat mengetahui daerah yang memiliki tingkat pengangguran yang tinggi dengan menerapkan teknik clustering.

Clustering merupakan salah satu metode data mining yang memecahkan permasalahan menggunakan ilmu komputer dan statistik (Muharni et al., 2022). Pengelompokan atau clustering digunakan untuk mengetahui pengelompokan dari data atau variabel yang diteliti. Seperti pengelompokan pengangguran menggunakan K-Means di Jawa Barat menggunakan data kabupaten dan jumlah pengangguran tahun 2022 yang menghasilkan 3 cluster (Muthmainnah et al., 2023), penerapan metode K-Means pada data

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v23i1.14464

tingkat pengangguran terbuka 2016-2018 dan 2019-2021 yang mendapatkan hasil Provinsi Riau dari tingkat pengangguran tinggi naik cluster 1 sebagai tingkat pengangguran rendah, sedangkan Provinsi Sumatera Barat turun dari cluster 1 yang tingkat penganggurannya rendah menjadi tingkat pengangguran tinggi (Muharni et al., 2022). Membandingkan metode Hierarchical Clustering, K-Means, K-Medoids, dan Fuzzy C-Means dalam Pengelompokan Provinsi di Indonesia Menurut Indeks Khusus Penanganan Stunting berdasarkan indikator IKPS tahun 2018 dengan perbandingan metode average linkage sebagai yang terbaik menangani penelitian ini (Suraya & Wijayanto, 2022). Pengelompokan provinsi di Indonesia berdasarkan tingkat kemiskinan dengan Hierarchical Agglomerative Clustering menggunakan data tahun 2020 dengan beberapa variabel, yang menghasilkan tingkat kemiskinan tertinggi di Provinsi Papua (Widodo et al., 2021).

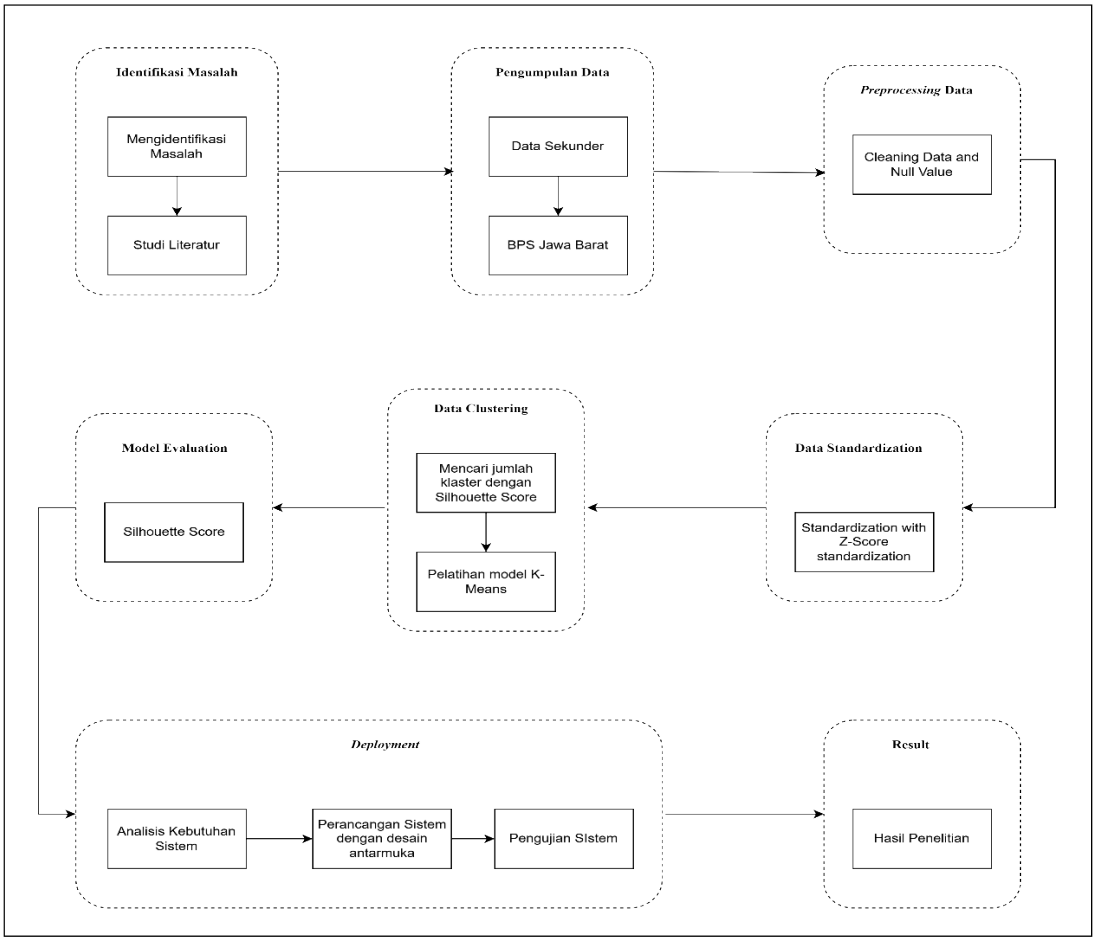
Clustering juga memiliki banyak jenis, contohnya seperti clustering hierarcy dan non-hierarcy. Pada clustering non-hierarcy contohnya yaitu K-Means (Syafiyah et al., 2022). Pada beberapa penelitian mengenai masalah sosial banyak menggunakan metode K-Means. K-Means memiliki kelebihan seperti sederhana, konvergensinya cepat, komputasinya efisien, dan mudah diimplementasikan (Mutrofin et al., 2023). K-Means Clustering menjadi algoritma yang populer karena efisiensi dan kesederhanaannya dalam mengelompokkan data (Patel & Mehta, 2011). Sebelum penerapan algoritma K-Means, terdapat beberapa tahap yang perlu dilakukan, seperti pengumpulan data, pra-pemrosesan data, kemudian dilakukan proses pengelompokan dengan K-Means. Pra-pemrosesan data merupakan langkah penting yang dapat mempengaruhi hasil dari algoritma clustering, bagian ini menghitung tuple dengan nilai yang hilang dengan menggunakan berbagai opsi, seperti nilai maksimum, minimum, konstan, rata-rata, dan standar deviasi (Patel & Mehta, 2011)

Pada teknik pra-pemrosesan data diterapkan pada data mentah untuk membuat data bersih, bebas gangguan, dan konsisten. Dengan normalisasi atau standarisasi data, akan membakukan data mentah dengan mengubah menjadi rentang tertentu menggunakan transformasi linier yang dapat menghasilkan klaster berkualitas baik dan meningkatkan akurasi algoritma pengelompokan (Mohamad & Usman, 2013). Standardization atau normalisasi dapat dilakukan dengan beberapa metode contohnya seperti Z-Score, Min-Max, dan Decimal Scaling. Pada penelitian ini, menggunakan metode Z-Score yang bertujuan untuk menyamakan skala pada masing-masing variabel digunakan untuk atribut kategori numerik menjadi standar deviasi, dimana pada dataset pengangguran berisi numerik dan mudah diimplementasikan (Zulyanti & Noeryanti, 2022), kekurangan Z-Score kurang efektif untuk data yang cukup kompleks (Whendasmoro & Joseph, 2022). Sehingga pada penelitian ini menggabungkan K-Means dan Z-Score dalam pengelompokan pengangguran Kabupaten/Kota di Provinsi Jawa Barat. Menggabungkan K-Means dan Z-Score diharapkan agar dapat memberikan hasil pengelompokan yang lebih baik. Hasil dari clustering akan divisualisasikan dengan menggunakan library Geopandas dan Folium pada program sehingga akan menampilkan peta Provinsi Jawa Barat dengan hasil dari cluster kabupaten dan kota.

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v23i1.14464

2. Metode/Perancangan

Tahapan penelitian yang akan dilakukan dapat dilihat pada Gambar 1 berikut.



Gambar 1. Tahapan Penelitian

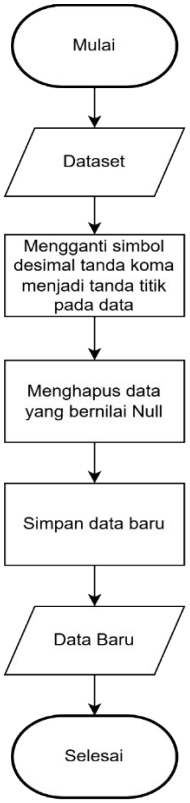
2.1. Pengumpulan Data

Data yang akan digunakan dalam penelitian ini menggunakan dataset sekunder yang diambil dari Badan Pusat Statistik Jawa Barat dan Jaringan Dokumentasi dan Informasi Humum Jawa Barat. Data berjumlah 27 data yang terdiri dari 18 kabupaten dan 9 kota sesuai yang ada di Provinsi Jawa Barat.

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v23i1.14464

2.2. Data Cleansing

Pada tahap ini dilakukan persiapan data dengan melakukan data cleansing terlebih dahulu, proses ini dilakukan dengan menyamakan pada format data, sehingga penyamaan format penulisan decimal diubah menjadi titik. Kemudian data untuk nilai rupiah dan miliar diberikan tanda titik untuk memudahkan mengetahui jumlah nominal. Selanjutnya untuk data bernilai null akan dihapus. Setelah selesai proses cleaning, maka data akan disimpan.

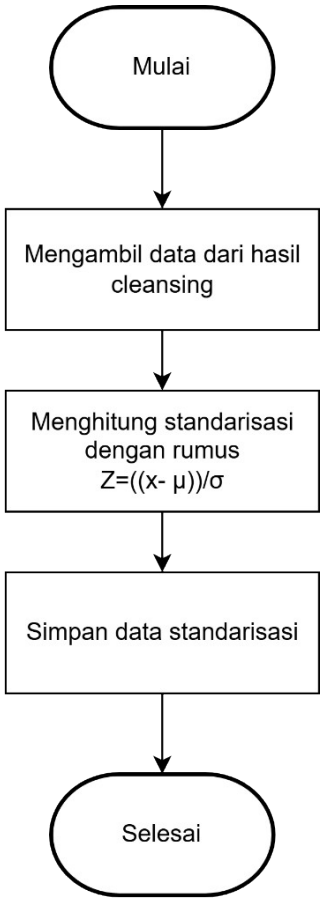


Gambar 2. Proses Data Cleansing

2.3. Standarisasi Data

Data yang telah melalui proses cleansing akan diganti dengan data dari hasil standarisasi data. Standarisasi data akan dilakukan dengan menggunakan Z-Score yang menghasilkan nilai ke dalam skala yang sama sehingga dapat diolah dalam proses klasterisasi. Proses ini dilakukan untuk menentukan standarisasi data yaitu dengan melakukan perhitungan pada data hasil cleansing. Dengan menggunakan rumus Z-Score, standarisasi dilakukan pada penelitian ini untuk mencegah bias pada cluster atau data yang mendominasi karena adanya perbedaan nilai yang jauh. Terutama pada variabel pengangguran ini memiliki satuan sampai jutaan maupun ratusan juta.

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v23i1.14464



Gambar 3. Proses Standarisasi Data

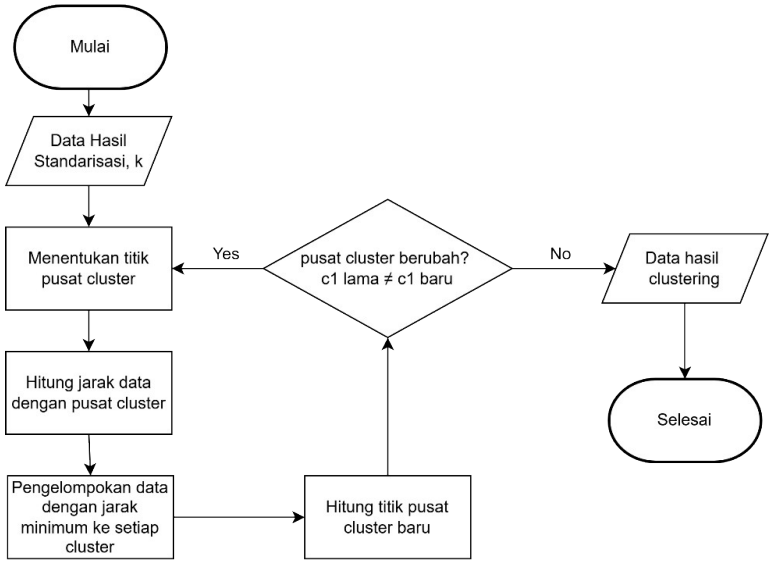
2.4. Pembuatan *Model Clustering K-Means*

Pada tahapan pembuatan model clustering akan dilakukan pembuatan model machine learning dengan melatih semua data yang telah melalui standarisasi dara. Pada tahapan ini akan dicari model dengan performa terbaik dengan cara menguji parameter penting dalam pelatihan model

2.4.1. Pelatihan Model K-Means

Pada tahapan ini digunakan algoritma k-means untuk melakukan pelatihan dengan dataset numerik yang terdiri dari 27 data dan 4 parameter. Data yang digunakan adalah data yang telah diproses sebelumnya dilakukan standarisasi data menggunakan z-score.

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v23i1.14464



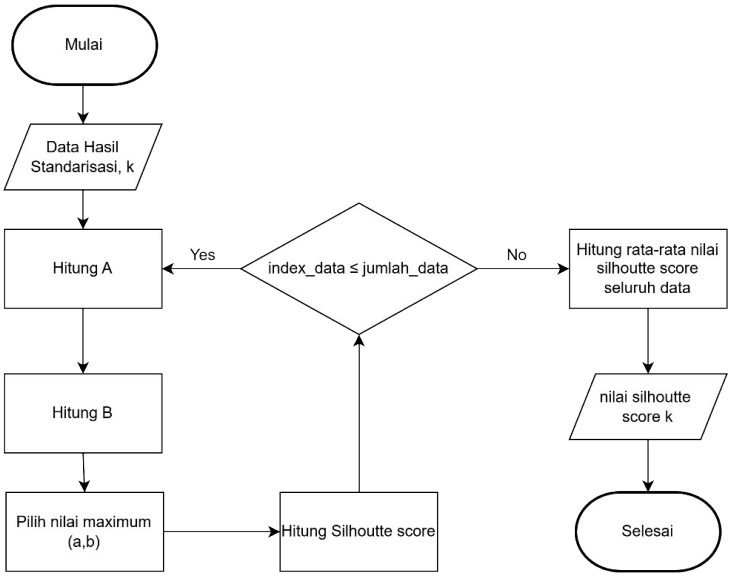
Gambar 4. Proses Data Clustering

2.4.2. Pemilihan K Parameter Optimum

Pemilihan nilai parameter k dilakukan menggunakan metode *silhouette*. Metode ini berfungsi sebagai *hyperparameter tuning* yang membantu dalam menemukan model terbaik dengan memilih parameter k yang paling optimal. Metode *silhouette* mengukur jarak rata-rata antara satu titik data dan lainnya dalam *cluster* yang sama (ukuran kohesi) serta jarak rata-rata antara *cluster* yang berbeda (ukuran pemisahan).

2.5. Evaluasi Clustering

Pada tahap evaluasi dalam penelitian ini melibatkan pengujian model yang telah di bangun sebelumnya. Pengujian dilakukan dengan menentukan nilai *silhouette* dan menganalisis penggunaan nilai K terbaik pada model tersebut.

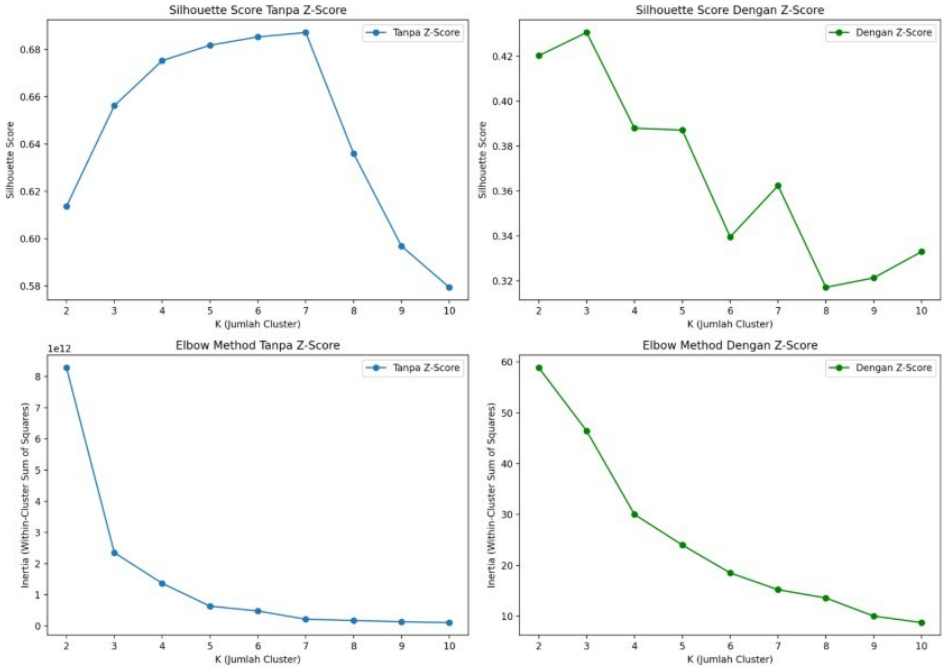


Gambar 5. Proses Silhouette Score

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v23i1.14464

3. Hasil dan Pembahasan

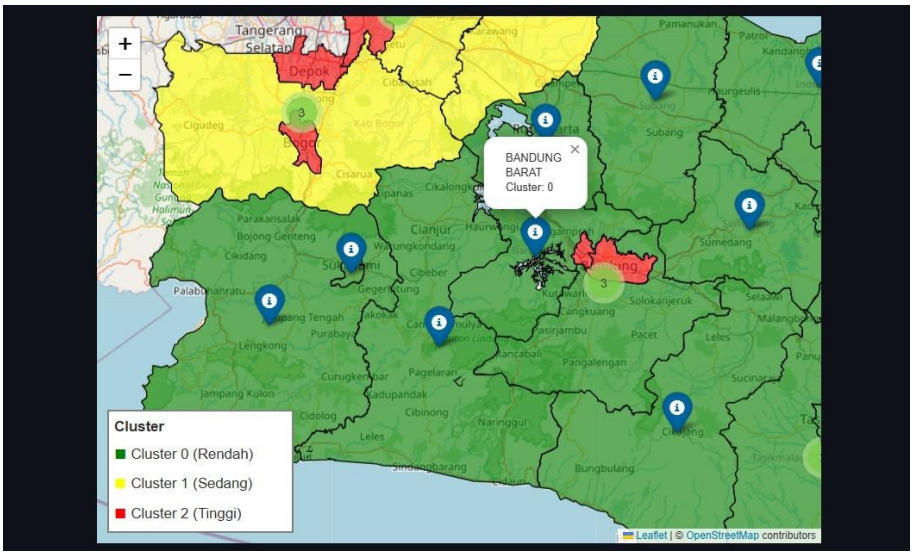
Bagian pembahasan berisi penjelasan hasil yang diperoleh dari masalah penelitian yang diangkat. Masalah pada penelitian ini yaitu apakah data yang memiliki jangkauan skala nilai yang jauh akan menunjukkan dominasi pada suatu variabel dan menghasilkan cluster yang bias dengan menerapkan proses standarisasi terlebih dahulu menggunakan metode Z-Score pada pra-pemrosesan yang digunakan untuk menyamakan skala nilai dari data. Pada penelitian ini juga membandingkan K-Means menggunakan Z-Score dan tanpa Z-Score.



Gambar 6. Perbandingan Evaluasi Hasil Clustering

Pada penelitian ini digunakan empat variabel data yang mempengaruhi dan dipengaruhi oleh pengangguran. Data yang diambil berjumlah 27 data sesuai dengan jumlah kabupaten dan kota yang ada di Jawa Barat. Data penelitian ini dilakukan proses cleansing pada tahap awal untuk membersihkan data yang kosong. Selanjutnya dilakukan proses data preprocessing yang bertujuan untuk melakukan proses standarisasi data dengan menggunakan metode Z-Score pada data hasil cleansing. Setelah data hasil standarisasi Z-Score dilakukan maka nilai data akan berada di antara -3 sampai 3. Data hasil standarisasi akan digunakan untuk membuat model dengan nilai parameter k = 3. Pengujian model menerapkan nilai Silhouette Score 0.4305.

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v23i1.14464



Gambar 7. Visualisasi Hasil Clustering

Berdasarkan hasil dari clustering, kategori cluster dibagi menjadi tingkat pengangguran tinggi, tingkat pengangguran sedang, dan tingkat pengangguran rendah. Kemudian hasil dari pemodelan clustering divisualisasikan pada peta kabupaten dan kota Provinsi Jawa Barat dan keterangan dari cluster ditampilkan pada kiri bawah peta. Pada visualisasi peta, setiap kabupaten dan kota memiliki pin yang jika ditekan akan menunjukkan nama kabupaten/kota dan cluster-nya.

4. Kesimpulan dan Saran

Berdasarkan hasil dari penelitian, dapat ditarik kesimpulan sebagai berikut :

1. Proses *clustering* pengangguran di Provinsi Jawa Barat dengan menerapkan standarisasi *Z-Score* sebelum algoritma *K-Means* menghasilkan k=3 *cluster*. *Cluster* 0 (rendah) dengan rata-rata *centroid* -0,43225 menghasilkan 19 anggota *cluster*. *Cluster* 1 (sedang mendekati tinggi) dengan rata-rata centroid 0,96235 menghasilkan 5 anggota *cluster*. Pada *cluster* 1 menunjukkan nilai *centroid* tertinggi pada variabel Indeks Pembangunan Manusia (IPM). Sementara itu, *cluster* 2 (tinggi) dengan rata-rata *centroid* 1,133625 menghasilkan 3 anggota *cluster*. *Cluster* 2 memiliki nilai *centroid* tertinggi untuk variabel lainnya selain IPM. Nilai *silhouette score* yang diperoleh adalah 0.4305 dengan nilai metode *elbow* sebesar 46.4512
2. Hasil perbandingan *clustering* tanpa *Z-Score* dan dengan standarisasi Z-Score dilakukan dengan menggunakan k terbaik pada masing-masing proses *clustering*. Hasil pada visualisasi analisis bias menunjukkan bahwa standarisasi membantu menghasilkan pembagian *cluster* yang lebih stabil berdasarkan jarak antar data. Meski *silhouette score* pada data yang menerapkan *standarisasi* cenderung rendah, namun penerapan standarisasi *Z-Score* menghasilkan analisis visualisasi data *cluster* yang lebih kompak dan jarak antar kelompok yang lebih jelas, sehingga bias hasil *clustering* dapat diminimalkan.

3. Hasil dari pengelompokan pengangguran didasarkan oleh empat indikator yang mempengaruhi pengangguran. Sehingga hasil penelitian ini tidak dapat dilihat dari salah satu indikator atau salah satu variabel saja, namun berdasarkan keseluruhan indikator (TPT, UMK, IPM, dan PDRB) yang digunakan.

Berdasarkan hasil dari penelitian, dapat ditarik saran sebagai berikut :

1. Penambahan database yang dapat menyimpan data dari hasil modeling dan pengelompokan sebelumnya.
2. Penambahan penanganan outlier untuk berkemungkinan dapat menghasilkan peningkatan kualitas pada clustering.

Daftar Pustaka

[1] Akramunnisa, A., & Fajriani, F. (2019). Hierarchical Clustering Analysis dalam Pengelompokan Tingkat Pengangguran di Sulawesi Selatan. Prosiding Seminar Nasional Teknologi Informasi Dan Komputer, 2(1), 270–275. https://journal.uncp.ac.id/index.php/semantik/article/view/1525/1336%0Ahttps://journal.uncp.ac.id/index.php/semantik/article/view/1525

[2] Aprizkiyandari, S., Satyahadewi, N., Pratama, A. N., Rivaldo, R., Nurdiansyah, S. I., & Helena, S. (2023). Implementasi K-Means Cluster untuk Menentukan Persebaran Tingkat Pengangguran. Empiricism Journal, 4(2), 400–406. https://doi.org/10.36312/ej.v4i2.1518

[3] Azzahra, Q. N., Pratama, M. N., Prionggo, E. A., Nurillatiffah, T., Pravitasari, A. A., & Indrayatna, F. (2023). Pengelompokan Tingkat Pengangguran, Kemiskinan, Dan Pendapatan Di Kabupatenkota Provinsi Jawa Barat. BIAStatistics: Jurnal Statistika Teori Dan Aplikasi: Biomedics, Industry & Business And Social Statistics, 6274(2), 67–80.

[4] Hanifah, S., & Primandari, A. H. (2023). Implementasi Metode K-Means Clustering dalam Pengelompokan Kabupaten/ Kota di Provinsi NTB Berdasarkan Indikator Pendidikan. Emerging Statistics and Data Science Journal, 1(3), 378–393. https://doi.org/10.20885/esds.vol1.iss.3.art44

[5] Harits, D., Puji, A. A., Andivas, M., Dermawan, D., & Thoriq, E. A. (2022). Analisis Klaster Tingkat Pengangguran di Provinsi Kalimantan Timur Menggunakan Algoritma K-Means. Jurnal Surya Teknika, 9(2), 456–460. https://doi.org/10.37859/jst.v9i2.4430

[6] Mohamad, I. Bin, & Usman, D. (2013). Standardization and its effects on K-means clustering algorithm. Research Journal of Applied Sciences, Engineering and Technology, 6(17), 3299–3303. https://doi.org/10.19026/rjaset.6.3638

[7] Muthmainnah, T. N., Indriyana, S., & Enri, U. (2023). Penerapan Algoritme K-Means Dalam Mengelompokkan Data Pengangguran Terbuka Di Provinsi Jawa Barat. Jurnal Informatika Dan Rekayasa Perangkat Lunak, 5(2), 122. https://doi.org/10.36499/jinrpl.v5i2.8736

[8] Tanjung, F. A., Windarto, A. P., & Fauzan, M. (2021). Penerapan Metode K-Means Pada Pengelompokkan Pengangguran Di Indonesia. Jurasik (Jurnal Riset Sistem

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v23i1.14464

Informasi Dan Teknik Informatika), 6(1), 61. https://doi.org/10.30645/jurasik.v6i1.271