

Penerapan Algoritma K-MEANS Clustering untuk Pemetaan Data Tingkat Pengangguran Terbuka di Kota/Kabupaten di Indonesia

La Ode Muhammad Yudhy Prayitno^{#1}, Rama Qubra Putra^{#2}, Fiqrah Idhul Dwi Anugerah^{b3}

*#Teknik Informatika Universitas Halu Oleo
Kambu, Kendari, Sulawesi Tenggara 93231*

¹yudhyprayitno567@gmail.com

²ramaqubra.15@gmail.com

³emailnyafiqrah10@gmail.com

Abstrak

Tingkat pengangguran terbuka merupakan indikator penting dalam menilai kesehatan ekonomi suatu daerah. Penelitian ini bertujuan untuk menerapkan algoritma K-MEANS clustering dalam memetakan data tingkat pengangguran terbuka di kota/kabupaten di Indonesia. Data utama yang digunakan berasal dari Badan Pusat Statistik (BPS) dengan fokus pada tahun 2021, 2022, dan 2023. Metode Elbow digunakan untuk menentukan jumlah cluster optimal, dan hasilnya menunjukkan adanya tiga cluster: tingkat pengangguran rendah, sedang, dan tinggi. Evaluasi kualitas clustering dilakukan menggunakan Davies-Bouldin Index, yang menunjukkan bahwa hasil clustering memiliki kualitas yang baik. Visualisasi geospasial hasil clustering menunjukkan bahwa tingkat pengangguran tinggi lebih dominan di Pulau Jawa, sementara tingkat pengangguran rendah lebih banyak ditemukan di Pulau Sulawesi. Hasil penelitian ini diharapkan dapat memberikan wawasan berharga bagi pembuat kebijakan dalam merancang strategi pengentasan pengangguran yang lebih efektif.

Kata kunci: Analisis Data, K-MEANS Clustering, Kota/Kabupaten, Pemetaan Data, Pengangguran Terbuka

The Application of K-MEANS Clustering Algorithm for Mapping Open Unemployment Data in Cities/Regencies in Indonesia

Abstract

The open unemployment rate is an important indicator in assessing the economic health of a region. This study aims to apply the K-MEANS clustering algorithm to map open unemployment data in cities and regencies in Indonesia. The main data used comes from the Central Bureau of Statistics (BPS) focusing on the years 2021, 2022, and 2023. The Elbow method was used to determine the optimal number of clusters, resulting in three clusters: low, medium, and high unemployment rates. The quality of the clustering was evaluated using the Davies-Bouldin Index, which indicated that the clustering results were of good quality. The geospatial visualization of the clustering results shows that high unemployment rates are more dominant in Java Island, while low unemployment rates are more prevalent in Sulawesi Island. The results of this study are expected to provide valuable insights for policymakers in designing more effective unemployment alleviation strategies.

Keywords: Cities/Districts, Data Analysis, Data Mapping, Indonesia, K-MEANS Clustering, Open Unemployment

I. PENDAHULUAN

Tingkat pengangguran terbuka adalah salah satu indikator utama dalam menilai kesehatan ekonomi suatu daerah. Pengangguran terbuka didefinisikan ketika seseorang yang termasuk dalam usia kerja sedang mencari pekerjaan, sedang mempersiapkan usaha, atau tidak mencari pekerjaan karena merasa tidak mungkin mendapatkan pekerjaan [1]. Tingkat pengangguran diukur sebagai persentase jumlah orang yang berumur di atas 15

tahun yang termasuk dalam angkatan kerja dan tidak memiliki pekerjaan. Prevalensi pengangguran dihitung dengan rumus:

$$\text{Prevalensi Pengangguran}(\%) = \left(\frac{\text{Jumlah Penganggur}}{\text{Total Angkatan Kerja}} \right) \times 100$$

Sebagai contoh, di Simeulue (Kabupaten di Provinsi Aceh) pada tahun 2021, prevalensi pengangguran sebesar 5.71% berarti bahwa 5.71% dari angkatan kerja di Simeulue yang berumur di atas 15 tahun adalah pengangguran. Sebagai negara berkembang Indonesia

tidak akan lepas dari masalah pengangguran [2]. Pengangguran terbuka biasanya terjadi di kalangan generasi muda yang baru saja menyelesaikan pendidikan menengah dan tinggi. Ada kecenderungan bagi mereka yang baru saja menyelesaikan pendidikannya untuk mencari pekerjaan sesuai dengan keinginannya [3]. Badan Pusat statistik (BPS) merilis data Tingkat Pengangguran Terbuka (TPT) pada bulan Februari dan bulan Agustus tiap tahun. Pengangguran tidak hanya berdampak pada individu dan keluarga yang mengalaminya tetapi juga pada stabilitas ekonomi dan sosial suatu negara. Pemetaan tingkat pengangguran di berbagai wilayah sangat penting untuk memahami distribusi pengangguran dan mengidentifikasi daerah-daerah yang membutuhkan perhatian lebih.

Clustering merupakan salah satu metode data mining terpercaya dan menjadi instrumen yang valid memecahkan masalah kompleks ilmu komputer dan statistik. Clustering bekerja dengan mengelompokkan titik-titik data dalam dua kelompok atau lebih, dimana titik-titik data didalam kelompok yang sama lebih mirip satu sama lain dibanding dengan kelompok data lainnya [4]. Penggunaan algoritma K-MEANS clustering [5] dalam penelitian ini diharapkan dapat memberikan gambaran yang lebih jelas tentang pola pengangguran di berbagai kota dan kabupaten di Indonesia. Dengan mengelompokkan wilayah-wilayah berdasarkan karakteristik tingkat penganggurannya, pemerintah dan pemangku kepentingan dapat merancang strategi yang lebih efektif dan tepat sasaran untuk mengatasi masalah pengangguran.

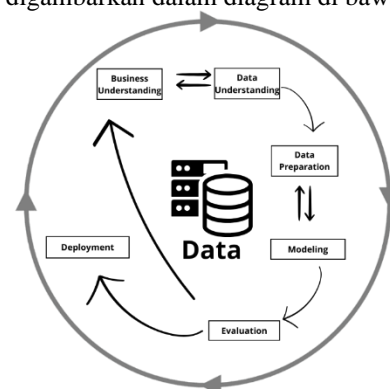
Penelitian terdahulu telah menunjukkan bahwa pengelompokan wilayah berdasarkan tingkat pengangguran dapat memberikan wawasan yang berharga. Misalnya, penelitian di Kalimantan Timur mengungkapkan bahwa meskipun provinsi ini memiliki Pendapatan Domestik Bruto Regional (PDRB) tinggi, tingkat penganggurannya masih signifikan, menunjukkan adanya dua klaster utama dengan karakteristik yang berbeda. Dalam penelitian ini, algoritma K-MEANS diterapkan untuk mengelompokkan data tingkat pengangguran di berbagai kabupaten/kota di Kalimantan Timur. Hasil klastering menunjukkan nilai koefisien Silhouette sebesar 0.45 dan nilai Davies-Bouldin Index (DBI) sebesar 1.25, yang menunjukkan bahwa klaster yang terbentuk cukup baik [6]. Penelitian serupa di Jawa Barat menggunakan data pengangguran dari tahun 2013-2021 untuk mengelompokkan kabupaten/kota berdasarkan tingkat pengangguran, yang terbagi menjadi klaster dengan tingkat pengangguran tinggi dan rendah. Penelitian ini menggunakan metode Elbow untuk menentukan jumlah klaster optimal, yang menunjukkan bahwa tiga klaster adalah yang paling optimal. Nilai koefisien Silhouette yang diperoleh adalah 0.51 dan DBI sebesar 1.10, yang menunjukkan bahwa klaster yang terbentuk cukup valid [7]. Selain itu, penelitian di Indonesia menunjukkan bahwa pengelompokan provinsi berdasarkan tingkat pengangguran dapat membantu pemerintah dalam memperluas lapangan pekerjaan untuk mengembangkan dan meningkatkan ekonomi di setiap provinsi. Penelitian

ini menemukan bahwa klaster dengan tingkat pengangguran tertinggi memiliki potensi untuk perbaikan ekonomi melalui peningkatan lapangan pekerjaan dan pelatihan kerja. Hasil klastering menunjukkan nilai koefisien Silhouette sebesar 0.58 dan DBI sebesar 0.95, yang menunjukkan klaster yang sangat baik [8]. Hal ini didukung oleh penelitian lain yang mengidentifikasi faktor-faktor yang mempengaruhi tingkat pengangguran terbuka di Indonesia, termasuk pertumbuhan ekonomi, tingkat pendidikan, dan pertumbuhan penduduk. Penelitian ini menggunakan algoritma K-MEANS untuk mengelompokkan data tingkat pengangguran berdasarkan provinsi, dengan hasil menunjukkan bahwa nilai koefisien Silhouette sebesar 0.53 dan DBI sebesar 1.05 [9]. Klasifikasi wilayah berdasarkan tingkat pengangguran juga membantu dalam merancang kebijakan yang lebih tepat sasaran untuk mengatasi pengangguran.

Bagaimana algoritma K-MEANS clustering dapat diterapkan untuk memetakan data tingkat pengangguran terbuka di kota/kabupaten di Indonesia? Tujuan penelitian ini adalah untuk: (1) menerapkan algoritma K-MEANS clustering pada data tingkat pengangguran terbuka di kota/kabupaten di Indonesia, (2) mengidentifikasi klaster-klaster utama yang menggambarkan tingkat pengangguran di berbagai wilayah, dan (3) memberikan rekomendasi kebijakan berdasarkan hasil klastering.

II. METODOLOGI PENELITIAN

Metode penelitian ini menggunakan salah satu teknik data mining, yaitu CRISP-DM (Cross Industry Standard Process for Data Mining). CRISP-DM merupakan metodologi standar yang terdiri dari enam tahapan dalam proyek data mining yaitu Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment. Persiapan data, pemilihan atribut pengklasifikasi, penggunaan K-Means Clustering untuk melakukan klasterisasi, dan menganalisis hasil klasterisasi. Proses ini digambarkan dalam diagram di bawah ini.



Gambar 2. 1 CRISP-DM

A. Business Understanding

Pada tahap pertama, dilakukan analisis masalah atau pemahaman tentang masalah apa yang dapat diangkat dari penelitian ini. Setelah menetapkan tujuan dan masalah, tahap berikutnya adalah menemukan solusi untuk mengatasi masalah tersebut.

B. Data Understanding

Pada tahap kedua, dilakukan pengumpulan data, mendeskripsikan data, mencari data mana saja yang bermanfaat. Tahapan ini tujuannya untuk mendapatkan gambaran awal mengenai data. Data yang digunakan dalam penelitian ini adalah data dari BPS mengenai persentase tingkat pengangguran terbuka (TPT) berdasarkan kabupaten/kota di seluruh Indonesia.

C. Data Preparation

Pada tahap ini, berbagai langkah dilakukan, seperti memilih data dengan memilih kolom yang akan digunakan untuk proses modeling, membersihkan data, dan menggabungkan fitur ke dalam data baru. Pada proses ini, menyesuaikan dengan kebutuhan penelitian sangat penting.

D. Modeling

Pada tahap modeling, proses penentuan model dilakukan dan metode data mining digunakan sesuai dengan tujuan penelitian. K-Means Clustering adalah algoritma yang digunakan dalam penelitian ini. Proses ini mencakup penerapan metode elbow untuk menentukan jumlah cluster optimal. Metode Elbow merupakan salah satu metode untuk menentukan jumlah cluster yang tepat melalui persentase hasil perbandingan antara jumlah cluster yang akan membentuk siku pada suatu titik. Metode ini melibatkan perhitungan Within-Cluster Sum of Squares (WCSS) untuk berbagai jumlah klaster, mulai dari satu hingga sepuluh. WCSS dihitung menggunakan rumus berikut:

$$WCSS = \sum_{i=1}^n \sum_{j=1}^k (x_{ij} - c_j)^2$$

Keterangan:

$\sum_i^n = 1$ = adalah penjumlahan dari seluruh data point, di mana n adalah total jumlah data point.

$\sum_j^k = 1$ = adalah penjumlahan dari seluruh klaster, di mana k adalah jumlah klaster.

x_{ij} = adalah data point ke-i dalam klaster ke-j.

c_j = adalah centroid dari klaster ke-j.

$(x_{ij} - c_j)^2$ = adalah kuadrat jarak antara data point ke-i dengan centroid klaster ke-j.

Metode ini diterapkan pada data persentase tingkat pengangguran terbuka (TPT) dari semua kabupaten/kota di seluruh Indonesia yang telah dipersiapkan sebelumnya.

E. Evaluation

Pada tahap Evaluation, dilakukan analisis terhadap hasil yang telah diperoleh dari proses klasterisasi. Evaluasi ini merupakan bagian penting dari interpretasi model data mining yang digunakan. Dalam penelitian ini, metode evaluasi yang dilakukan menggunakan metode Davies Bouldin Index. Davies-Bouldin Index adalah metrik

evaluasi yang digunakan untuk menilai kualitas clustering dengan cara mengukur rata-rata rasio jarak intra-cluster dengan jarak inter-cluster. Evaluasi menggunakan Davies Bouldin Index ini memiliki skema evaluasi internal cluster, dimana baik atau tidaknya hasil cluster dilihat dari kuantitas dan kedekatan antar data hasil cluster. Semakin kecil nilai DBI diperoleh (non-negatif ≥ 0), maka semakin baik cluster yang diperoleh dari pengelompokan menggunakan algoritma clustering.

F. Deployment

Pada tahap terakhir, hasil dari pemodelan dan evaluasi yang telah dilakukan pada proses data mining disajikan. Hasil penelitian dibuat dalam bentuk visualisasi untuk memudahkan analisis dan interpretasi, sehingga dapat memberikan wawasan yang lebih komprehensif tentang pola pengangguran di berbagai wilayah kabupaten/kota di Indonesia. Pada tahap ini, penulis menggunakan visual dari data geospasial Indonesia, dimana data geospasial adalah informasi yang mencakup data lokasi atau posisi geografis dari suatu objek di bumi, seperti koordinat garis lintang dan bujur. Informasi ini dapat digunakan untuk berbagai keperluan seperti perencanaan, pengambilan keputusan, dan penanganan bencana. Data geospasial telah diproses sehingga dapat mendukung perumusan kebijakan dan operasi bumi antariksa serta vital untuk pengambilan keputusan global. Tahap analisis geospasial merupakan bagian penting dalam penelitian ini untuk memvisualisasikan hasil klasterisasi tingkat pengangguran di Indonesia secara geografis

III. HASIL DAN PEMBAHASAN

A. Business Understanding

Objek penelitian ini adalah tingkat pengangguran terbuka di seluruh Indonesia, dengan fokus utama pada data dari Kabupaten/Kota. Permasalahan utama yang ditemukan dalam penelitian ini adalah tingkat pengangguran yang tinggi di beberapa wilayah di Indonesia. Penelitian ini bertujuan untuk menggunakan algoritma K-Means untuk mengklasterisasi data pengangguran terbuka berdasarkan wilayah administratif Kabupaten atau Kota. Dengan mengelompokkan wilayah berdasarkan tingkat pengangguran, diharapkan bahwa pola-pola yang berguna akan ditemukan bagi pembuat kebijakan untuk membuat strategi penanganan pengangguran yang lebih tepat sasaran dan efisien dengan membagi wilayah berdasarkan tingkat pengangguran.

B. Data Understanding

Pada tahap pemahaman data, penulis melakukan analisis data statistik deskriptif. Jenis data yang digunakan dalam penelitian ini adalah data sekunder. Data sekunder adalah data yang diperoleh bukan dari sumber langsung, melainkan sudah dikumpulkan oleh pihak lain dan sudah diolah serta memiliki keterkaitan dengan permasalahan yang diteliti.

Data utama yang digunakan dalam penelitian ini adalah persentase tingkat pengangguran terbuka di kabupaten/kota seluruh Indonesia, yang diperoleh dari situs resmi Badan

Pusat Statistik (BPS) di tiap provinsi. Data tersebut kemudian digabungkan dari 38 provinsi yang ada, mencakup setiap kabupaten/kota dan disimpan dalam satu file CSV. Fokus penelitian ini adalah data terbaru, yaitu tahun 2021, 2022, dan 2023.

BPS mencatat terdapat 514 kabupaten/kota di Indonesia pada tahun 2023, terdiri atas 416 kabupaten dan 98 kota yang tersebar di seluruh Nusantara. Data ini digunakan untuk menerapkan algoritma K-MEANS clustering guna memetakan tingkat pengangguran terbuka berdasarkan kelompok-kelompok tertentu.

Table 1. Data Persentase Tingkat Pengangguran Terbuka

No	Provinsi	Kab/Kota	2021	2022	2023
1	Aceh	Simeulue	5.71	6	5.85
2		Aceh Singkil	8.36	6.88	6.84
3		Aceh Selatan	6.46	4.82	4.73
....
....
513	Papua Barat	Tambrau	1.49	1.46	1.3
514	Daya	Kota Sorong	9.95	10.09	9.86

C. Data Preparation

Pada tahap ini, melakukan sebuah proses persiapan data yang terdiri dari data cleaning, dan data selection. Kedua proses ini dilakukan menggunakan Python untuk memastikan kualitas dan relevansi data yang akan dianalisis.

Proses data cleaning dimulai pada kolom tiap tahun 2023, 2022, dan 2021. Dari data yang ditemukan bahwa terdapat beberapa kabupaten yang memiliki nilai null atau kosong di tahun 2023, hal ini dikarenakan belum update-nya di website BPS kabupaten tersebut sampai dengan penelitian ini dibuat. Oleh karena itu, untuk memanipulasinya, penulis melakukan identifikasi dan pengisian data terbaru untuk setiap Kabupaten/Kota dengan prioritas menggunakan data tahun 2023 jika tersedia, beralih ke data tahun 2022 jika 2023 kosong, dan menggunakan data tahun 2021 jika keduanya kosong.

Kemudian, pada tahap data selection melibatkan penyederhanaan dataset menjadi dua kolom utama yaitu 'Kabupaten/Kota' dan 'Tahun Terbaru', dimana dari dataset kolom 'Provinsi' dihilangkan dengan tetap mempertahankan kolom 'Kabupaten/Kota' dan kemudian dibuatnya kolom 'Tahun Terbaru' yang berisi nilai pengangguran terbuka terkini sesuai prioritas yang telah ditentukan sebelumnya.

Proses ini diakhiri dengan pengecekan menyeluruh untuk memastikan tidak ada nilai null dalam dataset, menghasilkan dataset yang bersih dan siap untuk analisis lebih lanjut. Ini memastikan bahwa integritas dan relevansi data dalam penelitian tetap terjaga, dan memberikan dasar yang kuat untuk analisis selanjutnya.

Table 2. Hasil Data Preparation Tingkat Pengangguran Terbuka per Kabupaten/Kota

No	Kabupaten/Kota	Tahun Terbaru
1	Simeulue	5.85
2	Aceh Singkil	6.84
3	Aceh Selatan	4.73
....
....
513	Tambrau	1.3
514	Kota Sorong	9.86

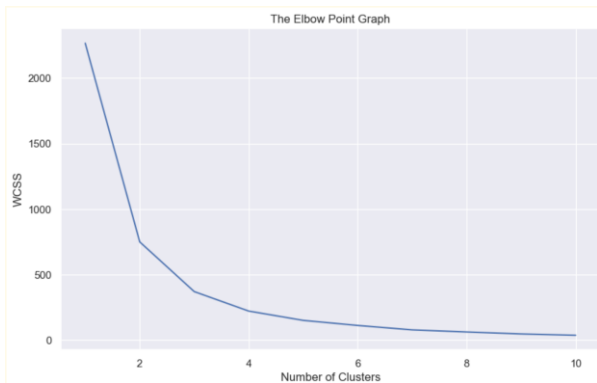
Data pada tabel 2 berisi atribut yaitu nama kabupaten kota, dan jumlah pengangguran yang akan digunakan pada proses clustering.

D. Modeling

Setelah data siap untuk digunakan, langkah berikutnya adalah pemodelan. Pada penelitian ini, algoritma yang digunakan yaitu algoritma K-Means yang kemudian akan dimodelkan berdasarkan data hasil transformasi. Jumlah cluster yang digunakan dalam penelitian ini adalah 3 cluster, dimana jumlah nilai cluster tersebut didapatkan dari hasil metode elbow.

```
wcss = []
for i in range(1,11):
    kmeans = KMeans(n_clusters=i,
init='k-means++', random_state=42)
    kmeans.fit(X)
    wcss.append(kmeans.inertia_)
```

Dalam metode Elbow, kita sebenarnya memvariasikan jumlah cluster (K) dari 1 – 11. Untuk setiap nilai K kita menghitung WCSS (Within-Cluster Sum of Square). WCSS adalah jumlah kuadrat jarak antara setiap titik dan pusat massa dalam sebuah cluster. Ketika kita diplot WCSS dengan nilai K, plotnya terlihat seperti Elbow. Dengan bertambahnya jumlah cluster, nilai WCSS akan mulai berkurang. Nilai WCSS terbesar ketika K=1. Ketika kita menganalisis grafik, dapat terlihat bahwa grafik akan berubah dengan cepat pada suatu titik dan dengan demikian menciptakan bentuk siku. Dari titik ini, grafik mulai bergerak hampir sejajar dengan sumbu X. Nilai K yang sesuai dengan titik ini adalah nilai K optimal atau jumlah cluster yang optimal. Berikut hasil pengujian data untuk metode Elbow.



Gambar 3. 1 Grafik Penentuan K-Cluster (Elbow)

Dari hasil grafik elbow terlihat setelah pada K=3 tidak ada tekuk-an grafik sehingga artinya sudah mulai optimal. Maka pada data di atas pemilihan jumlah cluster (K) adalah 3. Jadi setelah lebih dari K=3 hasil akan cenderung stabil. Selanjutnya untuk melihat efisiensinya akan terlihat nanti pada hasil uji performa. Sebelum melakukan proses clustering data dengan menggunakan k-means, yang harus dilakukan yaitu menentukan titik pusat centroid. Gambar 3.2 menunjukkan hasil centroid yang digunakan dengan menggunakan pemrograman python.

```
centroids = kmeans.cluster_centers_
print("Centroids:")
for i, centroid in enumerate(centroids):
    print(f"Centroid for cluster {i}: {centroid}")
```

```
Centroids:
Centroid for cluster 0: [2.5374026]
Centroid for cluster 1: [7.9025]
Centroid for cluster 2: [4.96095477]
```

Gambar 3. 2 Titik Pusat Centroid

Dari hasil centroid diatas menunjukkan bahwa cluster yang akan dibuat yaitu 3 cluster yaitu cluster 0, cluster 1, dan cluster 2. Langkah selanjutnya yaitu proses clustering dengan menggunakan k-means pada python. Dalam pembuatan model digunakan kelas KMeans dari library scikit learn.

```
kmeans = KMeans(
    n_clusters=3,
    init='k-means++',
    random_state=42
)
Y = kmeans.fit_predict(X)
```

```
[2 1 2 2 1 2 2 1 2 2 1 2 0 1 2 0 0 2 1 2 1 1 2 0 1 0 1 0 0 2 2 2 0 0 1 2 0
0 0 0 2 2 2 2 0 2 0 0 1 2 1 2 1 2 1 0 2 0 2 0 0 0 0 0 2 0 0 2 2 1 2 0 2 0
2 2 2 2 2 2 2 0 2 2 1 0 2 2 2 2 2 0 0 0 2 0 0 0 0 2 0 0 0 0 2 0 0 1 2
2 2 2 2 2 2 0 0 1 2 0 2 2 2 2 0 0 0 2 1 2 0 0 2 0 0 2 0 0 2 2 0 2 0 1 0 2
2 2 2 2 0 2 2 0 0 1 2 1 0 0 0 0 0 1 1 1 2 2 2 2 0 0 2 2 1 0 2 0 0 2 0 0
2 0 1 0 1 2 2 0 2 0 0 2 2 2 0 0 2 0 0 2 0 2 0 0 2 1 1 1 1 1 1 1 2 1 2 1 2
2 1 1 1 1 1 1 2 0 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2
2 2 0 0 2 2 2 0 0 2 0 0 2 2 0 2 2 0 1 1 1 2 2 2 2 2 0 2 2 0 2 2 2 2 2
2 2 0 2 2 2 0 0 2 1 2 2 2 2 0 2 2 1 2 0 0 0 2 2 1 2 2 2 2 1 2 0 0 0 0
0 0 0 0 0 0 0 0 0 0 2 0 0 2 0 0 2 0 0 2 0 0 0 0 2 0 2 0 0 0 0 0 0 2 0
0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 2 1 0 0 2 1 2 0 0 0 2 1
1 1 2 0 0 0 0 0 0 0 0 0 0 0 0 0 2 2 2 0 0 0 0 0 0 0 2 2 0 2 0 0 0 0 0
0 0 2 0 1 2 1 2 2 0 0 1 2 2 0 1 0 2 2 2 2 0 2 0 2 0 2 1 1 0 0 0 2 0 0 2 2
0 2 0 0 2 1 1 2 2 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 2 2 0 2 0 1]
```

Gambar 3. 3 Hasil cluster data

Mula – mula inialisasi model algoritma KMeans dengan memberikan parameter jumlah cluster adalah 3. Setelah inialisasi model algoritma KMeans dengan jumlah cluster yang ditentukan sebanyak 3, langkah selanjutnya adalah melakukan fitting dan prediksi cluster menggunakan metode `fit_predict()`. Metode ini diterapkan pada data input X, yang merupakan dataset yang telah dipersiapkan sebelumnya. Hasil dari proses ini adalah array Y yang berisi label cluster untuk setiap data point.

Penggunaan parameter `'init=k-means++'` pada inialisasi model bertujuan untuk mengoptimalkan pemilihan centroid awal, yang dapat meningkatkan kecepatan dan kualitas hasil clustering. Sementara itu, parameter `'random_state=42'` digunakan untuk memastikan reproducibility hasil, sehingga setiap kali kode dijalankan akan menghasilkan output yang konsisten.

Setelah proses clustering selesai, label cluster untuk setiap data point dapat dilihat melalui perintah `print(Y)` yang dapat dilihat pada gambar 3.3. Kemudian, dari hasil clustering dibuatkan kolom terbaru dari dataset untuk menampung masing-masing nilai dari cluster tiap kabupaten dan kota tersebut. Selain itu, ditambahkan juga sebuah kolom kategori untuk merepresentasikan tiap nilai clustering tersebut. Berdasarkan titik pusat centroid, didapatkan bahwa untuk cluster 0 itu sebagai kategori rendah, disusul oleh cluster 2 sebagai kategori menengah, dan cluster 1 itu sebagai kategori tinggi. Tabel 3 merupakan tabel hasil dari proses clustering dan pengkategorian yang dijalankan dengan python.

Table 3. Hasil proses clustering dengan kmeans menggunakan python

No	Kabupaten/Kota	Tahun Terbaru	Cluster	Kategori
1	Simeulue	5.85	2	Menengah
2	Aceh Singkil	6.84	1	Tinggi
3	Aceh Selatan	4.73	2	Menengah
....
....
513	Tambora	1.3	0	Rendah
514	Kota Sorong	9.86	1	Tinggi

E. Evaluation

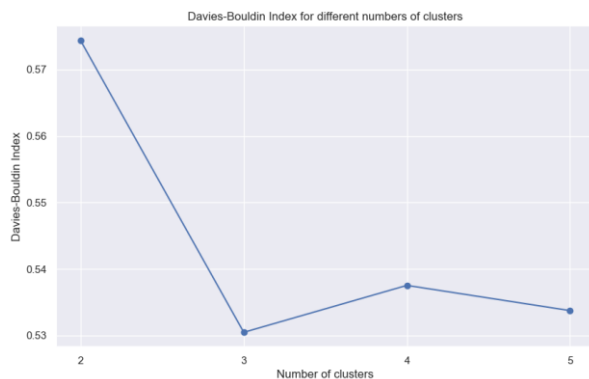
Pada tahap evaluasi dilakukan untuk mengukur kinerja hasil cluster jumlah pengangguran dengan jumlah cluster = 3 menggunakan Davies Bouldin Index. Hasil adalah angka,

yang jika hasilnya rendah maka kualitas cluster tersebut semakin baik. Tabel 4 merupakan tabel hasil pengujian dengan menggunakan Davies Bouldin Index yang dijalankan pada python:

Table 4. Hasil Pengujian

No	Jumlah Cluster	Davies Bouldin Indeks
1	2	0.5743434451856256
2	3	0.5304718608280816
3	4	0.5375292293570424
4	5	0.5337380541187509

Gambar 3.4 merupakan visualisasi terbaik dari hasil pengujian dengan menggunakan Davies Bouldin Index.



Gambar 3. 4 Visualisasi terbaik dari hasil pengujian

Dari hasil pengujian Davies-Bouldin Index yang ditunjukkan pada gambar, dapat disimpulkan bahwa jumlah cluster optimal adalah tiga. Nilai Davies-Bouldin Index terendah diperoleh ketika menggunakan tiga cluster, yaitu sebesar 0.5304718608280816. Nilai ini menunjukkan bahwa tiga cluster menghasilkan clustering yang paling baik dibandingkan dengan jumlah cluster lainnya, berdasarkan metrik Davies-Bouldin Index. Dengan menggunakan tiga cluster, dapat diinterpretasikan bahwa pemisahan antar cluster cukup baik dan masing-masing cluster memiliki keseragaman internal yang baik. Oleh karena itu, penggunaan tiga cluster dalam penelitian ini adalah yang paling efektif untuk mengelompokkan data tingkat pengangguran terbuka di Kabupaten/Kota di Indonesia.

F. Deployment

Setelah menyelesaikan seluruh tahapan dari proses awal hingga pemodelan, hasil analisis divisualisasikan menggunakan data geospasial untuk memberikan representasi visual yang komprehensif dan intuitif. Visualisasi ini memungkinkan pemahaman yang lebih mendalam terhadap pola spasial tingkat pengangguran terbuka di seluruh Indonesia.

Gambar 3.5 menunjukkan hasil clustering tingkat pengangguran terbuka di kabupaten/kota di Indonesia

berdasarkan algoritma K-MEANS. Peta ini mengelompokkan wilayah ke dalam tiga kategori: tingkat pengangguran rendah (hijau), menengah (kuning), dan tinggi (merah).



Gambar 3. 5 Visualisasi Geospasial Hasil Clustering Tingkat Pengangguran Terbuka Kabupaten/Kota di Indonesia

Dari peta tersebut, terlihat bahwa wilayah dengan tingkat pengangguran tinggi (warna merah) lebih dominan di Pulau Jawa, dengan konsentrasi yang signifikan di daerah-daerah tertentu. Hal ini menunjukkan bahwa Pulau Jawa menghadapi tantangan pengangguran yang lebih serius dibandingkan dengan daerah lain di Indonesia. Faktor-faktor seperti urbanisasi yang tinggi, densitas penduduk yang besar, dan migrasi masuk yang tinggi dapat berkontribusi pada tingginya tingkat pengangguran di pulau ini.

Sebaliknya, wilayah dengan tingkat pengangguran rendah (warna hijau) terlihat lebih banyak tersebar di Pulau Sulawesi. Ini menunjukkan bahwa beberapa daerah di Sulawesi berhasil mengatasi masalah pengangguran dengan lebih baik. Kemungkinan faktor yang berkontribusi pada rendahnya tingkat pengangguran di wilayah ini termasuk adanya kegiatan ekonomi lokal yang kuat, sektor pertanian dan perikanan yang dominan, serta inisiatif lokal yang efektif dalam menciptakan lapangan pekerjaan.

Selain itu, terdapat beberapa wilayah dengan tingkat pengangguran menengah (warna kuning) yang tersebar di berbagai pulau, menunjukkan variasi dalam kondisi ekonomi dan kesempatan kerja di berbagai daerah. Pola ini memberikan wawasan penting bagi pembuat kebijakan untuk memahami distribusi tingkat pengangguran dan mengidentifikasi area-area yang memerlukan intervensi lebih lanjut.

Dengan visualisasi ini, pemetaan data tidak hanya memberikan informasi numerik, tetapi juga menyajikan gambaran visual yang memudahkan analisis lebih lanjut serta pengambilan keputusan yang lebih efektif. Visualisasi ini dapat digunakan sebagai alat komunikasi yang kuat untuk menyampaikan temuan penelitian kepada berbagai pemangku kepentingan, termasuk pemerintah, akademisi, dan masyarakat luas.

IV. KESIMPULAN

Dalam penelitian ini, algoritma K-MEANS clustering telah berhasil diterapkan untuk memetakan tingkat pengangguran terbuka di kota/kabupaten di Indonesia. Proses ini melibatkan beberapa tahap penting, termasuk pemilihan data, pemodelan, dan evaluasi hasil clustering. Pada tahap pemodelan, metode Elbow digunakan untuk menentukan jumlah cluster yang optimal, yang kemudian

ditemukan sebanyak tiga cluster. Cluster pertama mencakup wilayah dengan tingkat pengangguran rendah, cluster kedua mencakup wilayah dengan tingkat pengangguran sedang, dan cluster ketiga mencakup wilayah dengan tingkat pengangguran tinggi. Evaluasi kualitas clustering dilakukan menggunakan Davies-Bouldin Index, yang menunjukkan bahwa hasil clustering memiliki kualitas yang baik dengan nilai DBI yang rendah. Visualisasi hasil clustering memberikan gambaran yang jelas mengenai distribusi tingkat pengangguran di berbagai wilayah di Indonesia. Peta yang dihasilkan menunjukkan bahwa tingkat pengangguran tinggi (ditandai dengan warna merah) lebih dominan di Pulau Jawa, sementara tingkat pengangguran rendah (ditandai dengan warna hijau) lebih banyak ditemukan di Pulau Sulawesi. Kesimpulannya, algoritma K-MEANS clustering terbukti efektif dalam memetakan data tingkat pengangguran terbuka di Indonesia. Hasil dari penelitian ini dapat memberikan wawasan berharga bagi pembuat kebijakan dalam merancang strategi pengentasan pengangguran yang lebih tepat sasaran dan efisien .

DAFTAR PUSTAKA

- [1] Wardah Bonitta, N., & Primandari, A. H. (2024). Analisis Clustering Tingkat Pengangguran Terbuka di Provinsi DIY Tahun 2010-2022 dengan Dynamic Time Warping. *Emerging Statistics and Data Science Journal*, 2(1).
- [2] Aziz, A., Norhasanah, S., Darwan Ali, U., Batu Berlian No, J., Baru Hulu, M., Mentawa Baru Ketapang, K., Kotawaringin Timur, K., & Tengah, K. (2023). *Klasterisasi Data Tenaga Kerja Terbuka Menurut Provinsi Dengan Penggunaan Algoritma K-Means* (Vol. 10, Issue 3). <http://jurnal.mdp.ac.id>
- [3] Pastia, N. I., & Dikananda, F. N. (2023). Pengelompokan Data Pengangguran Terbuka Menggunakan Algoritma K-Means Berdasarkan Provinsi Jawa Barat. *Jurnal Dinamika Informatika*, 12(1).
- [4] Muharni, S., & Andriyanto, S. (2022). PENERAPAN METODE K-MEANS CLUSTERING PADA DATA TINGKAT PENGANGGURAN TERBUKA TAHUN 2016-2018 DAN 2019-2021. *Jurnal Informatika*, 22(1), 89–99.
- [5] Kholila, N., Mujiono, M., & Wahyudi, D. (2023). Pemetaan Kondisi Lingkungan Tanam menggunakan K-Means Clustering. *JSITIK: Jurnal Sistem Informasi Dan Teknologi Informasi Komputer*, 1(2), 137–147. <https://doi.org/10.53624/jsitik.v1i2.182>
- [6] Wahyudi, M., & Pujiastuti, L. (2020). Penerapan Data Mining Dalam Mengelompokkan Data Pengangguran Terbuka Menurut Provinsi Menggunakan Algoritma K-Means. *Prosiding Seminar Nasional Riset Dan Information Science (SENARIS)*, 2, 432–440.
- [7] Harits, D., Andriyas Puji, A., Andivas, M., Dermawan, D., Thoriq, E. A., Raya, J. P., Bahagia, G., & Timur, K. (2022). Analisis Klaster Tingkat Pengangguran di Provinsi Kalimantan Timur Menggunakan Algoritma K-Means. *Jurnal Surya Teknika*, 9(2), 456–460.
- [8] Kusuma Arum, S., Astuti, R., & Muhammad Basysyar, F. (2024). PENERAPAN ALGORITMA K-MEANS PADA DATASET PENGANGGURAN TERBUKA BERDASARKAN PENDIDIKAN DI PROVINSI JAWA BARAT. In *Jurnal Mahasiswa Teknik Informatika* (Vol. 8, Issue 2).
- [9] Tanjung, F. A., Windarto, A. P., Fauzan, M., Studi, M. P., Informasi, S., & Tunas Bangsa, S. (2021). Penerapan Metode K-Means Pada Pengelompokan Pengangguran Di Indonesia. *Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)*, 6(1), 61–74. <https://tunasbangsa.ac.id/ejurnal/index.php/jurasik>
- [10] Ervi Lianita, Angga Pratama, & Ananda Faridhatul Ulva. (2024). Penerapan Metode K-Means Clustering Untuk Pemetaan Produktivitas Sayur-Sayuran Berbasis Sistem Informasi Geografis Di Provinsi Sumatera Utara. *JUSTIN (Jurnal Sistem Dan Teknologi Informasi)*, 12(2), 232–238.
- [11] Budiarti, T. N., Kesehatan, P., Cendekia, K., Indriani, D., & Melaniani, S. (2022). *Penelitian Asli K-MEANS CLUSTERING UNTUK*. <https://www.researchgate.net/publication/37380685>
- [12] Hardiani, T. (2022). Analisis Clustering Kasus Covid 19 di Indonesia Menggunakan Algoritma K-Means. *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, 11(2), 156–165. <https://doi.org/10.23887/janapati.v11i2.45376>
- [13] Suprihatin, S., Utami, Y. R. W., & Nugroho, D. (2019). K-MEANS CLUSTERING UNTUK PEMETAAN DAERAH RAWAN DEMAM BERDARAH. *Jurnal Teknologi Informasi Dan Komunikasi (TIKOMSiN)*, 7(1). <https://doi.org/10.30646/tikomsin.v7i1.408>
- [14] Lestari, W., Bina, S., & Kendari, B. (2019). Clustering Data Mahasiswa Menggunakan Algoritma K-Means Untuk Menunjang Strategi Promosi (Studi Kasus : STMIK Bina Bangsa Kendari). In *SIMKOM* (Vol. 4, Issue 2). <http://ejurnal.stmikbinsa.ac.id/index.php/simkom35>
- [15] Haq, M. A., Purnomo, W., & Setiawan, N. Y. (2023). *Analisis Clustering Topik Survey menggunakan Algoritme K-Means (Studi Kasus: Kudata)* (Vol. 7, Issue 7).