Jurnal Informatika dan Sistem Informasi (JIFoSI)
Vol. 5 No.3. Desember 2024
e-ISSN : 2722-130
DOI : https://doi.org/10.33005/jifosi.v5i3.460

IMPLEMENTASI DATA MINING MENGGUNAKAN METODE K-MEANS CLUSTERING UNTUK ANALISIS TINGKAT PENGANGGURAN TERBUKA DI JAWA TIMUR

Andini Fitriyah Salsabilah[1)], Zain Muzadid Zamzani[2)], Wanda Gustrifa[3)]
E-mail : [1)]andiniafsh@gmail.com, [2)]zmz250203@gmail.com, [3)]wandagustrifa@gmail.com
[1,2,3]Informatika, Fakultas Ilmu Komputer, UPN Veteran Jawa Timur

(Naskah masuk: 25 November 2024, diterima untuk diterbitkan: 31 Desember 2024)

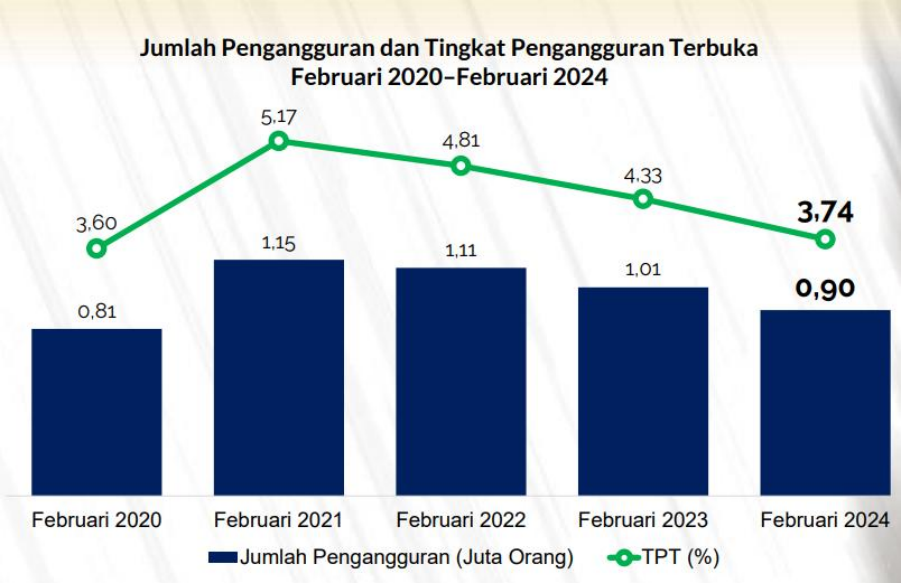
Abstrak

Penelitian ini menganalisis Tingkat Pengangguran Terbuka (TPT) di Provinsi Jawa Timur menggunakan metode K-Means *Clustering*. Data TPT diperoleh dari Badan Pusat Statistik (BPS) untuk periode 2019-2023, mencakup 38 kabupaten/kota di Jawa Timur. Penelitian ini bertujuan untuk mengelompokkan wilayah berdasarkan tingkat pengangguran guna mendukung pemerintah dan pemangku kepentingan dalam menentukan prioritas penanganan daerah dengan TPT tinggi. Proses analisis menggunakan metodologi SEMMA (*Sampling, Exploration, Modification, Modeling, dan Assessment)*. Penentuan jumlah cluster optimal dilakukan dengan metode Elbow menggunakan *Silhouette Score.* Hasil pengelompokan menunjukkan distribusi kabupaten/kota ke dalam beberapa klaster berdasarkan tingkat kemiripan TPT. Evaluasi menggunakan *Silhouette Score* menghasilkan nilai rata-rata sebesar 0,43 yang mengindikasikan bahwa klaster yang terbentuk memiliki tingkat pemisahan yang cukup baik tetapi masih dapat ditingkatkan. Hasil ini memberikan gambaran awal yang berguna dan dapat digunakan sebagai referensi untuk analisis lanjutan atau pengambilan kebijakan yang lebih tepat sasaran.
Kata kunci: *tingkat pengangguran terbuka, data mining, K-Means, SEMMA, silhouette score*

1. PENDAHULUAN

Menurut Badan Pusat Statistik (BPS), pengangguran meliputi mereka yang aktif mencari pekerjaan, memulai usaha, atau yang belum memulai pekerjaan. Jawa Timur merupakan salah satu dari 38 Provinsi di Indonesia yang menempati posisi kedua di Indonesia dengan jumlah penduduk terbanyak [1]. Namun, Provinsi Jawa Timur tidak terlepas dari permasalahan pengangguran yang perlu mendapatkan penanganan yang serius. Menurut Badan Pusat Statistik (BPS), Pengangguran terbuka adalah angkatan kerja yang belum atau tidak bekerja, baik yang masih mempersiapkan usaha, mencari pekerjaan, dan sudah mendapat pekerjaan namun belum mulai bekerja.
Berdasarkan grafik pada Gambar 1, tingkat pengangguran terbuka di Provinsi Jawa Timur cenderung mengalami penurunan tiap tahunnya. Meskipun demikian, permasalahan pengangguran tetap harus dituntaskan mengingat bahwa Provinsi Jawa Timur memiliki tingkat kemiskinan sebesar 9,79% pada tahun 2024 [2]. Salah satu penyebab kemiskinan adalah minimnya tersedianya lapangan pekerjaan, tingginya tingkat pengangguran serta kualitas hidup manusia [3]. Penyebaran tingkat pengangguran di Provinsi Jawa Timur memerlukan perhatian dalam menentukan kabupaten mana yang perlu diprioritaskan dalam penangananya. Data TPT (Tingkat Pengangguran Terbuka) di Provinsi Jawa Timur

tersedia di website Badan Pusat Statistik (BPS) Jawa Timur, Agar mempermudah dalam mengetahui informasi yang berguna dapat menggunakan teknik
Data Mining, yakni *Clustering.* Sehingga kabupaten-kabupaten di Jawa Timur dapat dikelompokkan menjadi beberapa *cluster* berdasarkan jumlah TPT.



Gambar 1. Grafik Tingkat Pengangguran Terbuka di Provinsi Jawa Timur (Sumber : BPS)

Data mining adalah proses menemukan hubungan bermakna melalui pola dan tren dalam kumpulan data besar menggunakan berbagai metode atau algoritma[4]. SEMMA merupakan standar proses data mining yang digunakan sebagai strategi untuk memecahkan masalah dalam bisnis atau penelitian. Teknik data mining dalam metode SEMMA melibatkan lima tahap, yaitu: (1) Pengumpulan data (Sample), (2) Eksplorasi data (Explore), (3) Transformasi data (Modify), (4) Pemodelan data (Model), dan (5) Evaluasi data (Assess) [5].
Clustering adalah metode analisis yang mengelompokkan data berdasarkan kesamaan tertentu. Proses ini mengacu pada pengelompokan catatan, observasi, atau objek yang memiliki karakteristik serupa ke dalam satu kelas. *Cluster* merupakan kumpulan data yang memiliki kemiripan tinggi di dalamnya, tetapi berbeda dengan kelompok data lain. Tujuan clustering adalah membagi informasi ke dalam kelompok yang relatif homogen, dengan memastikan kesamaan di dalam kelompok tetap minimal [6].
K-means *clustering*, adalah model pembelajaran tanpa supervisi. Model pembelajaran tanpa supervisi digunakan untuk kumpulan data yang belum diberi label atau diklasifikasikan. Model ini mencatat titik-titik yang serupa dalam kumpulan data dan menanggapinya sesuai dengan pola kemiripan pada setiap titik data. Model ini menggunakan algoritma berbasis centroid atau berbasis jarak, di mana kita menghitung jarak untuk menentukan pengelompokan tiap titik. Langkah pertama adalah menetapkan nilai K, kemudian membagi data ke dalam K kelompok agar data dalam kelompok yang sama memiliki kesamaan yang lebih tinggi, yang memudahkan untuk membedakannya [7].
Beberapa penelitian terdahulu telah dilakukan dengan menggunakan metode pengelompokan atau *clustering* dengan algoritma K-Means *Clustering*. Penelitian-penelitian sebelumnya dapat dijadikan acuan atau referensi penulis dalam penelitian dengan topik yang serupa seperti penelitian K-Means *Clustering* Analisis pada Persebaran Tingkat Pengangguran Kabupaten/Kota di Sulawesi Selatan oleh [8] dengan mendapatkan hasil 24 Kabupaten/Kota di Sulawesi Selatan tersebar ke dalam 2 *cluster,*

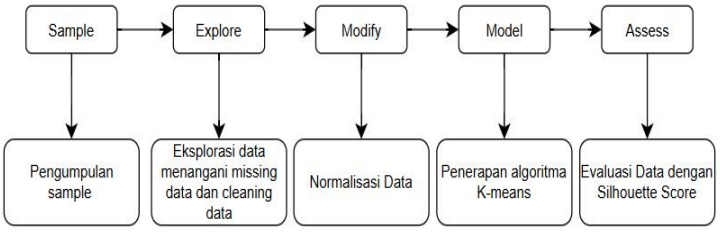
yaitu *cluster* dengan rata-rata tingkat pengangguran rendah dan tinggi. Dari penelitian [10] diperoleh hasil 13 provinsi masuk ke dalam *cluster* tinggi dan 21 provinsi masuk ke dalam kategori *cluster* rendah. Penelitian tentang penerapan metode K-Means *Clustering* pada data tingkat pengangguran terbuka menunjukkan bahwa Provinsi Riau naik dari *cluster* dengan tingkat pengangguran tinggi (2016-2018) ke *cluster* dengan tingkat pengangguran rendah (2019-2021). Sebaliknya, Provinsi Sumatera Barat turun dari *cluster* tingkat pengangguran rendah ke *cluster* tingkat pengangguran tinggi pada periode 2019-2021 [9].

Pada penelitian ini, metode *elbow* digunakan untuk menentukan jumlah *cluster* optimal. Namun, sebagai alternatif, digunakan pula *Silhouette Score* untuk mengevaluasi kualitas klaster yang terbentuk. Silhouette Score memperhitungkan jarak rata-rata antara setiap data dengan data lain dalam klaster yang sama (*intra-cluster distance*) dan jarak rata-rata terhadap klaster terdekat lainnya (*nearest-cluster distance*). Nilai ini berkisar antara -1 hingga 1, di mana nilai mendekati 1 menunjukkan klaster yang terdefinisi dengan baik, sedangkan nilai mendekati -1 menunjukkan data yang mungkin salah pengelompokan [10].

Dari penelitian di atas, para peneliti telah mendapatkan pengetahuan baru setelah melakukan *cluster* analisis pada penambangan data (data mining). Pengetahuan yang diperoleh yaitu berupa pemetaan tingkat pengangguran berdasarkan *cluster* data TPT di tingkat nasional, regional maupun di tingkat provinsi[9]. Penelitian ini dilakukan bertujuan menggali lebih dalam data TPT dengan menganalisa lebih dalam hasil penelitian [11]. Pada penelitian ini dilakukan untuk mengetahui persebaran TPT pada Provinsi Jawa Timur yang diharapkan pemerintah dan stakeholder lainnya dapat menyusun strategi spesifik dan dapat mengalokasikan sumber daya yang lebih efisien untuk mengurangi pengangguran di wilayah yang lebih membutuhkan intervensi.

2. METODOLOGI

Penelitian ini menerapkan metode SEMMA untuk pengolahan data. SEMMA mencakup beberapa tahap, yaitu *sample, explore, modify, model,* dan *asses*. Penelitian ini dilakukan dalam beberapa tahapan, seperti pada Gambar 2.



Gambar 2. Metode SEMMA

2.1 *Sample*

Pada tahap ini, pengumpulan data Tingkat Pengangguran Terbuka (TPT) Menurut Kabupaten/Kota (Persen) dihimpun dari situs resmi Badan Pusat Statistik Jawa Timur. Data ini mencakup informasi dari semua kota dan kabupaten di Jawa Timur untuk periode tahun 2019 hingga 2023. Dataset tersebut berbentuk file CSV yang memuat tingkat pengangguran dalam bentuk persentase.

2.2 *Explore*

Pada tahap ini, dataset melalui tahap pembersihan data seperti mengecek nilai null, NaN, dan duplikat data, serta pengisian nilai terhadap data yang null atau NaN. Selanjutnya, dilakukan pengecekan deskripsi statistik dataset untuk memahami sebaran dan karakteristik data.

2.3 *Modify*

Pada tahap *ini*, data dinormalisasi dengan melakukan deskripsi statistik dasar dan standarisasi skala. Standarisasi dilakukan pada data dari kolom tahun 2019 hingga 2023 menggunakan metode *StandardScaler*, yang mengubah setiap kolom sehingga memiliki rata-rata 0 dan standar deviasi 1. Selain itu, pada tahap ini juga dilakukan transformasi data untuk memastikan kolom tahun 2023 dikonversi menjadi tipe data numerik.

2.4 Model

Model yang digunakan dalam analisis ini adalah algoritma K-Means. Sebelum proses clustering dilakukan, jumlah cluster yang optimal ditentukan terlebih dahulu menggunakan metode seperti Elbow *Method.* Algoritma K-Means bekerja dengan menghitung jarak antara setiap data dengan centroid (pusat cluster) menggunakan jarak Euclidean, yang dinyatakan dengan persamaan (1):

$$D(i,j) = \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2 + i\ldots + (x_{ki} - x_{kj})^2} \quad (1)$$

Keterangan :
D (i,j) = Jarak data i ke pusat cluster j
Xki = Data ke i pada atribut data ke k
X kj = Jarak antara titik pusat ke j pada atribut k

Persamaaan (1) ini digunakan untuk mengukur seberapa dekat data dengan pusat cluster yang telah ditentukan. Dengan cara ini, data-data yang memiliki kemiripan akan dikelompokkan dalam cluster yang sama.

2.5 *Asses*

Untuk menilai kinerja hasil pengujian clustering pada metode yang digunakan dalam penelitian ini, dilakukan perhitungan *Silhouette Score*. *Silhouette Score* memberikan evaluasi kualitas cluster dengan mengukur seberapa baik data berada di dalam *cluster* mereka sendiri dibandingkan dengan jaraknya ke *cluster* terdekat lainnya. Jika nilai *Silhouette Score* mendekati 1 maka menunjukkan nilai *cluster* yang baik. Semakin tinggi nilai *Silhouette Score*, semakin baik kualitas *cluster* yang terbentuk. Dengan demikian, *Silhouette Score* membantu dalam menentukan jumlah *cluster* yang optimal berdasarkan kesesuaian data dengan *cluster* masing-masing [12]. Tabel 1 menjelaskan rentang nilai *Silhouette Score* untuk menentukan nilai *cluster* terbaik.

$$S = (b - a) : max(a, b) \quad (2)$$

Keterangan :
a = Rata-rata jarak antara setiap titik data dengan data lain dalam cluster yang sama (konsistensi intra-cluster).
b = Rata-rata jarak antara setiap titik data dengan data dalam cluster terdekat lainnya (separasi antar-cluster)
S = Nilai Silhouette Score untuk setiap data.

Table 1. Rentang Nilai *Silhouette Score* dan Interpretasi

Rentang Nilai	Interpretasi
0.71 - 1.0	*Strong Cluster*

0.51 - 0.70	*Good Cluster*
0.26 - 0.50	*Weak Cluster*
<0.25	*Bad Cluster*

3. HASIL DAN PEMBAHASAN

3.1 *Sample*

Dalam penelitian ini, penulis menggunakan dataset yang diperoleh dari Badan Pusat Statistik (BPS) Provinsi Jawa Timur terkait Tingkat Pengangguran Terbuka (TPT) Menurut Kabupaten/Kota. Pada dataset diperoleh data yang terdiri dari 6 kolom yaitu kabupaten/kota, 2019, 2020, 2021, 2022, dan 2023. Selain itu, dataset memiliki 38 baris yang merupakan nilai atau value dari kabupaten/kota dan tahun 2019 hingga 2023. (https://jatim.bps.go.id/id/statistics-table/2/NTQjMg==/tingkat-pengangguran-terbuka--tpt--menurut-kabupaten-kota.html). Data yang diperoleh akan dikelola menggunakan google collab, dataset dapat dilihat pada Tabel 2.

Table 2. Dataset Tingkat Pengangguran Terbuka (TPT) Menurut Kabupaten/Kota di Jawa Timur (Sumber: BPS)

	Kabupaten / Kota	**2019(%)**	**2020 (%)**	**2021 (%)**	**2022(%)**	**2023(%)**
1	Kabupaten Pacitan	0.91	2.28	2.04	3.65	1.83
2	Kabupaten Ponorogo	3.5	4.45	4.38	5.51	4.66
3	Kabupaten Trenggalek	3.36	4.11	3.53	5.37	4.52
4	Kabupaten Tulungagung	3.29	4.61	4.91	6.65	5.65
5	Kabupaten Blitar	3.05	3.82	3.66	5.45	4.91
6	Kabupaten Kediri	3.58	5.24	5.15	6.83	5.79
7	Kabupaten Malang	3.7	5.49	5.4	6.57	5.7
…	…	…	…	…	…	…
38	Kota Batu	2.42	5.93	6.57	8.43	4.52

Tabel 2. menunjukan data yang didapatkan dari website Badan Pusat Statistik (BPS) Jawa Timur terkait Tingkat Pengangguran Terbuka (TPT) Menurut Kabupaten/Kota dengan terdapat data pengangguran dari 38 kabupaten/kota pada tahun 2019 - 2023. Dataset ini berisi angka Tingkat Pengangguran Terbuka (TPT) pada setiap wilayah yang disimpan dalam format csv dengan dataset direpresentasikan dalam bentuk persen agar dapat diproses ke tahap selanjutnya yaitu *Data Processing*. Setelah data melewati tahap pra proses data hingga dataset dapat diolah menggunakan data mining dengan menggunakan metode pengelompokan atau clustering. Algoritma yang digunakan dalam pengelompokan / *clustering* data pada penelitian ini adalah K Means *Clustering.*

3.2 Explore

Pada penelitian ini, dilakukan pembersihan data untuk memastikan tidak adanya data yang kosong dan duplikasi data. Untuk mengatasi data yang kosong, dilakukan imputasi dengan mean, di mana nilai kosong diisi dengan rata-rata dari data pada fitur tersebut.

Tabel 3. Kabupaten dengan Data yang Tidak Terbaca oleh *Machine Learning*

No	Kabupaten/Kota	Tahun 2019	Tahun 2020	Tahun 2021	Tahun 2022	Tahun 2023
24	Gresik	5.40	8.21	8.00	7.84	**6,82,**

Pada Tabel 3 terdapat kesalahan angka di kolom Tahun 2023 (6,82,) yang seharusnya tertulis (6.82). Kesalahan ini diubah pada dataset ke bentuk *NaN* dan diisi kembali dengan nilai mean dari fitur tersebut.

```
[ ]  df['2023'] = pd.to_numeric(df['2023'], errors='coerce')
     missing_values = df[df.isnull().any(axis=1)]
     missing_values
```

	Kabupaten/Kota Se Jawa Timur Tingkat Pengangguran Terbuka (TPT) Menurut Kabupaten/Kota (Persen)	2019	2020	2021	2022	2023
24	Kabupaten Gresik	5.4	8.21	8.0	7.84	NaN

```
df['2023'].fillna(df['2023'].mean(), inplace=True)
df['2023'] = df['2023'].round(2)
```

Gambar 3. Tahap *Explore*

Hasil dari proses pengisian *NaN* dengan mean dapat dilihat pada Tabel 3, di mana semua nilai telah diperbaiki dan dibulatkan hingga dua angka di belakang koma untuk konsistensi.

Tabel 4. Data Kabupaten Gresik setelah di *Explore*

Kabupaten/Kota	Tahun 2019	Tahun 2020	Tahun 2021	Tahun 2022	Tahun 2023
Gresik	5.40	8.21	8.00	7.84	4.64

3.3 Modify

Data Normalization

Dalam proses normalisasi, setiap fitur diubah ke skala yang seragam agar model machine learning dapat bekerja lebih optimal. Normalisasi dilakukan dengan menggunakan StandardScaler, yang mengubah data menjadi distribusi normal standar dengan nilai rata-rata 0 dan standar deviasi 1.

	2019	2020	2021	2022	2023
0	0.000000	0.000000	0.000000	0.307796	0.017673
1	0.521127	0.249712	0.265006	0.557796	0.434462
2	0.492958	0.210587	0.168743	0.538978	0.413844
3	0.478873	0.268124	0.325028	0.711022	0.580265
4	0.430584	0.177215	0.183465	0.549731	0.471281

Gambar 4. *Data Normalization*

Transformasi

Pada penelitian ini, transformasi data dilakukan untuk mengubah data mentah menjadi format yang sesuai bagi model machine learning. Salah satu langkah penting dalam proses ini adalah mengkonversi data per tahun yang terdiri beberapa kolom menjadi satu kolom agar dapat melihat tren angka pengangguran setiap tahun.

```
[417] # Mengubah bentuk DataFrame agar data per tahun menjadi satu kolom
      df_melted = df.melt(id_vars='Kabupaten/Kota',
                          value_vars=['2019', '2020', '2021',
                                      '2022', '2023'],
                          var_name='Tahun', value_name='Jumlah Pengangguran')

      # Memastikan nama kolom 'Kota/Kabupaten' benar di DataFrame baru
      df_melted.rename(columns={'Kota/Kabupaten': 'Kota_Kabupaten'}, inplace=True)

      # Mengganti nama kolom tahun menjadi format angka saja
      df_melted['Tahun'] = df_melted['Tahun'].str.extract('(\d+)').astype(int)
```

Gambar 5. Data Transformation

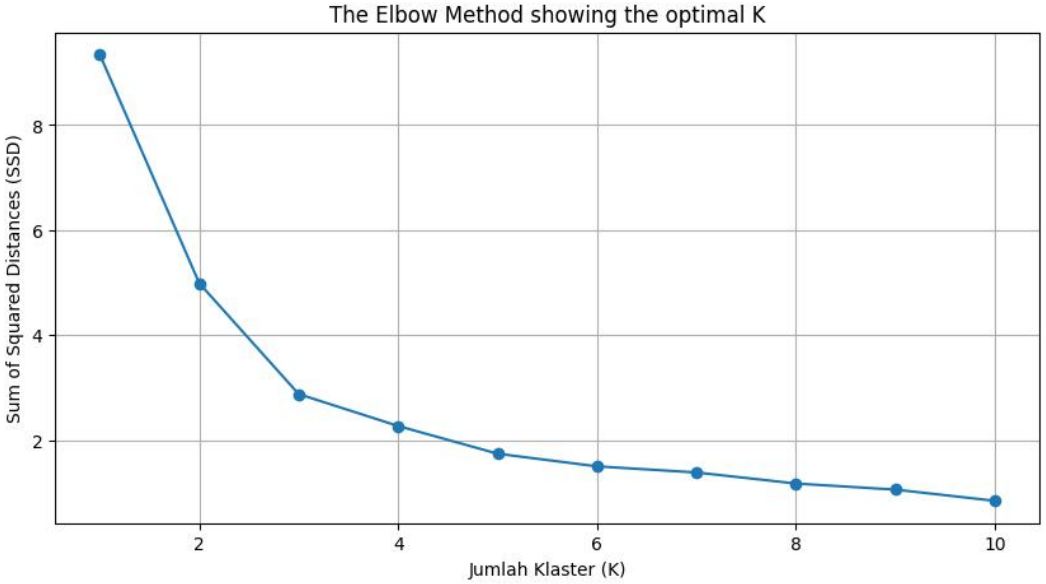
Selain itu, dilakukan *encoding* pada data kabupaten/kota menjadi sebuah format numerik agar semua tipe data seragam.

Tabel 5. Data Kabupaten dilakukan *Encoding*

Kabupaten/Kota	Kabupaten/Kota_Encoding
Kabupaten Pacitan	18
Kabupaten Ponorogo	21
Kabupaten Trenggalek	27
Kabupaten Tulungagung	29
Kabupaten Blitar	3
Kabupaten Kediri	9
Kabupaten Malang	14
…	…
Kota Batu	30

3.4 Model

Jumlah k cluster yang optimal akan diterapkan sebagai dasar dalam proses pengelompokan K-Means setelah menentukan jumlah k yang paling sesuai. Proses ini memanfaatkan *Google Collaboratory* dengan bahasa pemrograman Python untuk menemukan nilai k optimal, menggunakan pendekatan *Sum of Squared Distances* (SSD) dan Metode Elbow guna memastikan jumlah *cluster* yang tepat.



Gambar 6. Metode Elbow

Pada visualisasi Gambar 6, terlihat adanya elbow dalam data tingkat pengangguran terbuka di Provinsi Jawa Timur yang diuji menggunakan algoritma K Means dengan jumlah *n_clusters(K)=2* dan *n_clusters(K)=3*. Tujuannya adalah untuk memberikan gambaran yang lebih jelas dan meningkatkan pemahaman mengenai data tersebut. Dataset TPT (Tingkat Pengangguran Terbuka) di Provinsi Jawa Timur mencakup data dari 38 kabupaten. Proses analisis dilakukan menggunakan algoritma K Means. Pada Tabel 6 merupakan hasil *clustering* dari perhitungan menggunakan *n_clusters(K)=3* yang menunjukkan nama kabupaten/kota dan kelas *cluster*. Sedangkan pada Tabel 7 merupakan hasil *clustering* dari perhitungan menggunakan *n_clusters(K)=2* yang menunjukkan nama kabupaten dan kelas *cluster* .

Tabel 6. *Cluster* Data Tingkat Pengangguran Terbuka (TPT) Menurut Kabupaten/Kota di Jawa Timur Menggunakan K=3

No.	Kabupaten / Kota	2019	2020	2021	2022	2023	Cluster
1.	Kabupaten Pacitan	0.91	2.28	2.04	3.65	1.83	2
2.	Kabupaten Ponorogo	3.5	4.45	4.38	5.51	4.66	1
3.	Kabupaten Trenggalek	3.36	4.11	3.53	5.37	4.52	1
4.	Kabupaten Tulungagung	3.29	4.61	4.91	6.65	5.65	1
5.	Kabupaten Blitar	3.05	3.82	3.66	5.45	4.91	1

6.	Kabupaten Kediri	3.58	5.24	5.15	6.83	5.79	1
7.	Kabupaten Malang	3.7	5.49	5.4	6.57	5.7	1
…	…	…	…	…	…	…	…
38.	Kota Batu	2.42	5.93	6.57	8.43	4.52	1

Tabel 7. *Cluster* Data Tingkat Pengangguran Terbuka (TPT) Menurut Kabupaten/Kota di Jawa Timur Menggunakan K=3

No.	**Kabupaten / Kota**	**2019**	**2020**	**2021**	**2022**	**2023**	**Cluster**
1.	Kabupaten Pacitan	0.91	2.28	2.04	3.65	1.83	1
2.	Kabupaten Ponorogo	3.5	4.45	4.38	5.51	4.66	1
3.	Kabupaten Trenggalek	3.36	4.11	3.53	5.37	4.52	1
4.	Kabupaten Tulungagung	3.29	4.61	4.91	6.65	5.65	1
5.	Kabupaten Blitar	3.05	3.82	3.66	5.45	4.91	1
6.	Kabupaten Kediri	3.58	5.24	5.15	6.83	5.79	1
7.	Kabupaten Malang	3.7	5.49	5.4	6.57	5.7	1
…	…	…	…	…	…	…	…
38.	Kota Batu	2.42	5.93	6.57	8.43	4.52	1

3.5 Asses

n_clusterS(K) = 2

```
[25] from sklearn.metrics import silhouette_score

     # Evaluasi menggunakan Silhouette Score
     sil_score = silhouette_score(normalized_df, clusters)
     print(f'Silhouette Score: {sil_score}')

     Silhouette Score: 0.5021940268351521
```

Gambar 7. Hasil Evaluasi K = 2 Menggunakan *Silhouette Score*

Penelitian ini menggunakan *Silhouette Score* dalam mengevaluasi model *Clustering*. *Silhouette Score* digunakan untuk mengukur dua aspek utama, kohesi dan separasi. Kohesi menunjukkan seberapa dekat data dalam satu *cluster* terhadap pusat *cluster*-nya, sedangkan separasi menunjukkan seberapa jauh *cluster* tersebut dari *cluster* lainnya. Evaluasi yang dilakukan pada *n_cluster(K) = 2* memperoleh hasil sebesar **0.50** Hasil tersebut menunjukkan bahwa pembagian *cluster* yang dilakukan dengan menggunakan metode K-Means memiliki kualitas yang cukup baik namun masih lemah. Nilai ini menunjukkan bahwa sebagian besar data dikelompokkan ke dalam *cluster* cukup sesuai tetapi terdapat beberapa data yang berada pada batas antara dua *cluster* atau berada di *cluster* yang tidak sesuai.

n_clusters(K) = 3

```
from sklearn.metrics import silhouette_score

# Evaluasi menggunakan Silhouette Score
sil_score = silhouette_score(normalized_df, df['Cluster'])
print(f'Silhouette Score: {sil_score}')
```

```
Silhouette Score: 0.43300169017310763
```

Gambar 8. Hasil Evaluasi Menggunakan Silhouette Score

Evaluasi yang dilakukan pada *n_cluster(K) = 3* memperoleh hasil sebesar **0.43.** Hasil tersebut menunjukkan bahwa pembagian cluster yang dilakukan dengan menggunakan metode K-Means memiliki kualitas yang kurang baik. Nilai ini menunjukkan bahwa sebagian besar data dikelompokkan ke dalam *cluster* belum sesuai karena terdapat beberapa data yang berada ada pada batas antara dua *cluster* atau berada di *cluster* yang tidak sesuai.. Dengan kata lain *cluster* yang terbentuk cukup berjauhan satu sama lain, namun masih ada kemungkinan untuk menambah jarak antar *cluster* untuk optimalisasi yang lebih baik.

4. KESIMPULAN DAN SARAN

Penelitian ini menggunakan algoritma K-Means *Clustering* untuk mengelompokkan kabupaten/kota di Jawa Timur berdasarkan Tingkat Pengangguran Terbuka (TPT) periode 2019 hingga 2023 dengan metodologi SEMMA. Evaluasi jumlah *cluster* optimal dilakukan menggunakan metode Elbow, diikuti dengan perbandingan kualitas *cluster* menggunakan *Silhouette Score*. Pada *n_clusters* = 2, diperoleh nilai *Silhouette Score* sebesar 0.50, yang menunjukkan kualitas *cluster* yang cukup baik. Sementara itu, pada *n_clusters* = 3, diperoleh nilai sebesar 0.43, yang menunjukkan kualitas yang lebih rendah. Meskipun demikian, pembagian ke dalam 3 *cluster* dipilih karena memberikan interpretasi praktis yang lebih bermakna. *Cluster* 1 mencakup daerah dengan TPT rendah, yang menunjukkan kondisi ekonomi yang relatif stabil. *Cluster* 2 mencakup daerah dengan TPT sedang, menunjukkan wilayah yang memerlukan perhatian lebih untuk mencegah peningkatan pengangguran. *Cluster* 3 mencakup daerah dengan TPT tinggi, yang menjadi prioritas untuk intervensi kebijakan seperti penciptaan lapangan kerja baru atau pelatihan keterampilan. Hasil ini memberikan gambaran pemetaan wilayah berdasarkan tingkat pengangguran, yang dapat menjadi referensi bagi pemerintah dalam mengalokasikan sumber daya secara lebih efektif. Untuk optimalisasi, disarankan untuk meningkatkan *Silhouette Score* dengan mencoba parameter atau algoritma lain, menambahkan variabel seperti tingkat pendidikan atau akses ke layanan publik untuk memberikan informasi yang lebih komprehensif, serta menggunakan hasil *cluster* ini sebagai dasar untuk pembuatan kebijakan daerah yang lebih berbasis data.

5. DAFTAR RUJUKAN

[1] Badan Pusat Statistik. (2024). Jumlah penduduk menurut provinsi di Indonesia. Retrieved from https://sulut.bps.go.id/id/statistics-table/ (Diakses pada tanggal 28 September 2024)

[2] Badan Pusat Statistik. (n.d.). Data penduduk Indonesia. Retrieved from https://web-api.bps.go.id/download (Diakses pada tanggal 28 September 2024)

[3] Ristika, E. D., Primandhana, W. P., & Wahed, M. (2021). Analisis pengaruh jumlah penduduk, tingkat pengangguran terbuka, dan indeks pembangunan

manusia terhadap tingkat kemiskinan di Provinsi Jawa Timur. Eksis: Jurnal Ilmiah Ekonomi dan Bisnis, 12(2), 129–136.
[4] Putri, S. U., Irawan, E., & Rizky, F. (2021). Implementasi data mining untuk prediksi penyakit diabetes dengan algoritma C4.5. Kesatria: Jurnal Penerapan Sistem Informasi (Komputer dan Manajemen), 2(1), 39–46.
[5] Cui, M. (2020). On the Elbow Method. 5–8. https://doi.org/10.23977/accaf.2020.010102
[6] Sari, N., Handayani, H. H., & Siregar, A. M. (2023). Implementasi clustering data kasus Covid-19 di Indonesia menggunakan algoritma K-Means. Bianglala Informatika, 11(1), 7–12.
[7] Santoso, H., Magdalena, H., & Wardhana, H. (2022). Aplikasi Dynamic Cluster pada K-Means Berbasis Web untuk Klasifikasi Data Industri Rumahan. MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, 21(3), 541–554. https://doi.org/10.30812/matrik.v21i3.1720
[8] Akramunnisa, A., & Fajriani, F. (2020). K-Means Clustering Analysis pada Persebaran Tingkat Pengangguran Kabupaten/Kota di Sulawesi Selatan. Jurnal Varian, 3(2), 103–112. https://doi.org/10.30812/varian.v3i2.652
[9] Muharni, S., & Sigit, A. (2022). Penerapan Metode K-Means Clustering Pada Data Tingkat Pengangguran Terbuka Tahun 2016–2018 dan 2019–2021. Jurnal Informatika, 22(1).
[10] Punhani, A., Faujdar, N., Mishra, K. K., & Subramanian, M. (2022). Binning-based silhouette approach to find the optimal cluster using K-means. IEEE Access, 10, 115025–115032.
[11] Tanjung, F. A., Windarto, A. P., & Fauzan, M. (2021). Penerapan Metode K-Means pada Pengelompokkan Pengangguran di Indonesia. JURASIK (Jurnal Riset Sistem Informasi dan Teknologi Informasi), 6(1), 61. https://doi.org/10.30645/jurasik.v6i1.271
[12] Aji, B. G., Sondawa, D. C. A., Gifari, M. R., & Wijayanto, S. (2023). Penerapan Algoritma K-Means untuk Clustering Harga Rumah di Bandung. Jurnal Ilmiah Informatika Global, 14(2), 17–23. https://doi.org/10.36982/jiig.v14i2.3189