PENGELOMPOKAN DATA PENGANGGURAN DI INDONESIA MENGGUNAKAN ALGORITMA *K-MEANS CLUSTERING*

Novita Dwi Octaviana Ibrahim[1)], Tubagus M. Akhriza[2)], Rahayu Widayant[3)]
STMIK PPKIA Pradnya Paramita Malang
novitadwioctaviana@gmail.com[1)], akhriza@stimata.ac.id[2)], rahayu@stimata.ac.id[3)]

Abstract

Unemployment in Indonesia increased during 2019-2021 due to COVID-19. The government implemented PSBB to tackle the virus, but the economic downturn led to business efficiency measures and layoffs. This study analyzes unemployment data in Indonesian provinces using K-Means Clustering to see the changes in unemployment rates in Indonesia before the Covid-19 and after the Covid-19 pandemic in 2019 and 2021. Five clusters were identified in 2019, while four clusters existed in 2021. During 2019-2021, the provinces of West Kalimantan, NTT, NTB, and West Sulawesi moved from level 3 to level 1, indicating an increase in the unemployment rate in these areas. In addition, provinces such as South Sulawesi, Maluku, North Maluku, Aceh, North Kalimantan, East Kalimantan, Banten, Jambi, Riau, and North Sulawesi shifted from level 5 to level 3 indicating that the area experienced an increase in the unemployment rate, this also applied to East Kalimantan, DKI Jakarta, and Riau Islands that shifted from level 5 to level 4. The evaluation results show that 2019 produces a silhouette value of 0.36 and 0.35 for 2021. Both silhouette results are included in the weak cluster structure criteria because the silhouette value is in the range of 0.26-0.50.

Keywords: *Algorithm K-Means, Clustering, Elbow, Silhouette Coefficient*

1. PENDAHULUAN

Pengangguran merupakan masalah yang hingga saat ini belum bisa lepas dari negara berkembang maupun maju, termasuk Indonesia. Keberhasilan dalam pembangunan ekonomi suatu negara bisa ditunjukkan dari berbagai indikator ekonomi diantaranya yaitu tingkat pengangguran. Melalui tingkat pengangguran bisa dilihat kondisi perekonomian dalam suatu negara itu berkembang dengan baik, lambat ataupun sedang mengalami penurunan [1]. Menurut data Badan Pusat Statistik (BPS) tingkat pengangguran dari tahun 2019 hingga tahun 2021 mengalami peningkatan akibat adanya pandemi covid-19 yang terjadi di 34 provinsi Indonesia. Pemerintah mengeluarkan kebijakan Pembatasan Sosial Berskala Besar (PSBB) untuk mencegah persebaran Covid-19. PSBB ini membuat aktivitas masyarakat dan ekonomi menjadi terbatas. Penurunan pertumbuhan ekonomi membuat pelaku usaha melakukan efisiensi untuk mencegah kerugian. Akibatnya, pelaku usaha melakukan tindakan Pemutusan Hubungan Kerja (PHK) sehingga berdampak pada peningkatan jumlah pengangguran.

Permasalahan pengangguran memberikan pengaruh pada besarnya Tingkat Partisipasi Angkatan Kerja (TPAK), sebab tingkat partisipasi angkatan kerja dapat diartikan sebagai seberapa banyak tenaga kerja yang tersedia untuk proses produksi nantinya akan mengurangi tingkat pengangguran [2].

Permasalahan pengangguran lainnya memberikan pengaruh pada Indeks Pembangunan Manusia. Ketika Indeks Pembangunan Manusia mengalami mengalami peningkatan maka dapat diartikan pembangunan otonomi lebih baik. Peningkatan dapat disebabkan oleh faktor pendidikan, kesejahteraan masyarakat dan lain sebagainya, ketika beberapa hal tersebut mengalami peningkatan maka kualitas manusia akan mengalami peningkatan yang relatif baik sehingga kualitas dan kemampuan penduduk akan mengurangi jumlah pengangguran [3].

Permasalahan selanjutnya yaitu pendidikan dengan ditentukan berdasarkan Rata-rata Lama Sekolah (RLS), salah satu cara untuk meningkatkan kemampuan logika manusia yaitu pendidikan, dikarenakan adanya persaingan yang ketat dan kemajuan

teknologi. Pendidikan dapat memudahkan sumber daya manusia dalam mencari pekerjaan karena mempunyai nilai daya saing yang tinggi dan berakibat pada berkurangnya tingkat pengangguran [4].

Tingkat pengangguran juga dapat dipengaruhi oleh jumlah penduduk, hal ini dikarenakan semakin tinggi jumlah penduduk dalam suatu wilayah semakin besar kemungkinan terdapat jumlah orang yang menganggur, terutama jika tidak diimbangi dengan pertumbuhan ekonomi yang cukup kuat untuk menyerap tenaga kerja yang ada. Untuk mengurangi tingkat pengangguran dibutuhkan upaya yang komprehensif meliputi pembangunan ekonomi yang berkelanjutan, kebijakan pendidikan dan pelatihan kerja yang tepat sasaran, dan pengembangan lapangan kerja yang sesuai dengan potensi dan kebutuhan masyarakat [1].

Penelitian yang dilakukan oleh Fawaidul Badri dan Anang Habibi bertujuan untuk Implementasi Metode K-Mean *clustering* Analysis pada Pengelompokan Pengangguran di Indonesia sebagai Dampak dari Pandemi Covid-19 dengan data yang diperoleh dari BPS berupa atribut tingkat pengangguran menghasilkan 2 *cluster* yakni *cluster* tinggi terdiri 10 provinsi dan *Cluster* rendah terdiri dari 24 provinsi. Hasil dari *K-Means clustering* memiliki nilai Mean Square kurang dari 0,05, maka terbukti bahwa *cluster* yang terbentuk antara *cluster* 1 dan *cluster* 2 adalah menunjukkan hasil signifikan yang cukup baik [5]. Fadhillah Azmi Tanjung, dkk menggunakan metode *K-Means* untuk mengelompokkan pengangguran di Indonesia dengan mengambil atribut tingkat pengangguran tahun 2014-2019 [6]. Akramunnisa dan Fajriani melakukan *clustering* dengan menggunakan metode *K-Means* untuk menganalisis persebaran tingkat pengangguran Kabupaten/Kota di Sulawesi Selatan berdasarkan atribut Tingkat Pengangguran Terbuka (TPT), Upah Minimum Kabupaten/Kota (UMK), dan laju pertumbuhan Indeks Pembangunan Manusia (IPM). Hasil analisis dari *K-Means clustering* terdapat 2 *cluster* dengan kelompok tingkat pengangguran tinggi dan rendah [7].

Berdasarkan penelitian terdahulu yang telah dilakukan, masih belum ada penelitian yang dilakukan dengan mengelompokan provinsi di Indonesia berdasarkan atribut jumlah penduduk, Tingkat Partisipasi Angkatan Kerja (TPAK), Indeks Pembangunan Manusia (IPM), dan rata-rata lama sekolah penduduk umur 15 tahun ke atas menggunakan Algoritma *K-Means* dan kaitannya sebagai dampak dari pandemi Covid-19 di Indonesia. Oleh karena itu, peneliti mengklaster tingkat pengangguran berdasarkan karakteristik setiap daerah untuk mengetahui perbandingan tingkat pengangguran di setiap provinsi sebelum pandemi tahun 2019 dan setelah pandemi tahun 2021.

2. KAJIAN LITERATUR

2.1 Data Mining

Data mining merupakan proses menemukan korelasi baru yang bermanfaat, pola dan trend dengan menambang sejumlah repository data dalam jumlah besar, dan menggunakan teknologi pengenalan pola seperti statistic dan teknik matematika [8]. Data mining disebut juga dengan *Knowledge discovery in database* (KKD) ataupun pengenalan pola. Data mining dapat dibagi menjadi empat kelompok, yaitu model prediksi *(prediction modelling)*, analisis kelompok *(Cluster analysis)*, analisis asosiasi *(association analysis)* dan deteksi anomaly *(anomaly detection)* [6].

Tahapan pada proses data mining diawali dari penyeleksian data, proses cleaning data, proses transformasi, proses data mining atau proses mencari pola atau informasi dari sebuah data terpilih dan tahap terakhir adalah tahap interpretasi dan evaluasi yang menghasilkan informasi - informasi baru yang bermanfaat. Secara detail dijelaskan sebagai berikut [9]:

1. Data *Selection*
   Pemilihan data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD *(Knowledge Discovery in Database).* Data hasil seleksi akan digunakan dalam proses data mining.
2. *Pre-processing / Cleaning*
   Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data.
3. Transformasi
   Proses transformasi ini adalah proses dimana data yang telah dipilih akan

diubah kedalam bentuk dimana data bisa diproses dalam data mining

4. Data Mining
   Data Mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu.
5. *Interpretation / Evaluation*
   Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan.

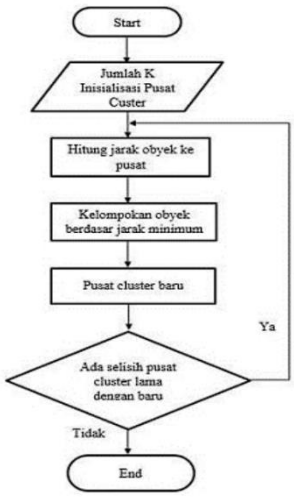
2.2 *Clustering*

Clustering merupakan suatu pengelompokkan data mining yang bertujuan untuk melakukan sebuah klasifikasi ataupun memprediksi nilai dari variabel target yang mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi berkelompok-kelompok yang memiliki kemiripan karakteristik sehingga objek yang di dalam *cluster* mirip satu sama dengan yang lainnya dan memiliki perbedaan dengan objek dari *cluster* yang lain [6]. Untuk melakukan *clustering* ada beberapa metode yang dapat digunakan, diantaranya metode *K-Means clustering*. Metode ini mampu mengelompokkan data dalam jumlah yang besar dan waktu yang cepat dan efisien. Metode *K-Means* adalah metode klastering berbasis jarak yang membagi data ke dalam sejumlah klaster dan algoritma ini bekerja pada atribut numerik [10]. Kekurangannya, hasil *clustering* bergantung pada penentuan awal pusat *cluster*, sehingga hasil perhitungan *clustering* dengan metode *K-Means* akan baik jika penentuan pusat *cluster* tepat [11].

2.3 Metode *K-Means*

Metode *K-Means* merupakan sebuah metode sederhana untuk membagi suatu kumpulan data dalam suatu angka spesifik dari *cluster*, yaitu *k*. Metode *K-Means* ditemukan oleh beberapa peneliti dengan disiplin ilmu berbeda-beda yaitu oleh Lloyd (1957, 1982), Forgery (1965), Friedman dan Rubin (1967), dan terakhir adalah McQueen (1967) [12]. Metode *K-Means* telah umum diterapkan dalam data mining dan pengenalan pola, dan telah didefinisikan sebagai salah satu pendekatan pengelompokan data mining paling sederhana yang mengimplementasikan fungsi jarak Euclidean [13]. Tujuan proses *clustering* adalah meminimalkan terjadinya *objective function* yang diset dalam proses *clustering*, yang pada umumnya digunakan untuk meminimalisasikan variasi dalam suatu *cluster* dan memaksimalkan variasi antar *cluster* [12].

Beberapa keuntungan dari *K-Means* adalah secara singkat diartikan efisien dan cepat. Namun, metode *K-Means* sebagian besar bergantung pada titik data awal dan varian dalam memilih sampel awal yang biasanya diarahkan ke berbagai hasil. Metode *K-Means* selalu menggunakan teknik gradien berdasarkan fungsi tujuan untuk mendapatkan nilai puncak. Fungsi pencarian tren dalam teknik gradien ini terutama diamati dalam proses komputasi di mana seluruh proses akan segera tenggelam ke titik terendah ketika titik fokus *cluster* awal mungkin tidak sesuai yang pada gilirannya menyebabkan pengurangan energi [13]. Berikut Gambar 1 merupakan *flowchart* dari algoritma *K-Means*.



Gambar 1 *Flowchart* Algoritma *K-Means*

Algoritma dasar *clustering* data menggunakan metode *K-Means* dapat dilakukan dengan cara [14]:

1. Menentukan jumlah kelompok
2. Menentukan nilai centroid, untuk nilai centroid awal ditentukan secara acak. Selanjutnya untuk menentukan nilai centroid baru pada iterasi digunakan rata-rata dari kelompok ke i untuk variabel ke j dengan rumus sebagai berikut:

$$\overline{V_{IJ}} = \frac{1}{N_i}\sum_{k=0}^{N_i} X_{kj} \qquad (1)$$

Dimana:
$\overline{v_{IJ}}$ = centroid atau rata-rata kelompok ke-I untuk variabel ke j
N_i = jumlah data yang menjadi anggota kelompok ke-i
i, k = indeks dari kelompok
j = indeks dari variabel
X_{kj} = nilai ke-k yang ada didalam kelompok tersebut untuk variabel ke-j

3. Menentukan jarak antara titik centroid dengan setiap titik obyek. Jarak yang digunakan untuk mengukur titik centroid dengan titik obyek adalah jarak Euclid, dengan rumus sebagai berikut:

$$d_{ed} = \sum_{i=1}^{n}(x_i - y_i)^2 \qquad (2)$$

Dimana:
x_i = obyek pengamatan ke i
y_i = centroid ke i
n = banyaknya obyek yang menjadi kelompok
Alokasikan setiap data atau obyek ke *cluster* terdekat. Kedekatan dua obyek ditentukan berdasarkan jarak antar kedua obyek tersebut. Jarang paling dekat antara satu data dengan satu *cluster* tertentu akan menentukan suatu data masuk ke dalam *cluster* yang mana.

4. Pengelompokan obyek dalam kelompok yang sudah ditentukan berdasarkan jarak yang paling dekat atau minimum.
5. Melakukan iterasi, ulangi langkah 2 dengan menentukan centroid baru dan lanjutkan seterusnya hingga didapatkan hasil tidak terdapat perubahan obyek obyek yang terdapat dalam kelompok.

2.4 Metode *Elbow*

Metode *Elbow* merupakan suatu metode yang digunakan untuk menghasilkan informasi dalam menentukan jumlah *cluster* terbaik dengan cara melihat persentase hasil perbandingan antara jumlah *cluster* yang akan membentuk siku pada suatu titik. Untuk mendapatkan perbandingannya adalah dengan menghitung nilai SSE *(Sum of Square Error)* dari masing-masing *cluster*. Hasil persentase yang berbeda dari setiap nilai *cluster* dapat ditunjukkan dengan menggunakan grafik sebagai sumber informasinya. Jika nilai *cluster* pertama dengan nilai *cluster* kedua membentuk suatu siku dalam grafik atau nilainya mengalami penurunan paling besar, maka nilai klaster tersebut adalah yang terbaik [10]. Berikut rumus SSE pada *K-Means* adalah [14]:

$$SEE = \sum_{k=1}^{K}\sum_{x_i \in S_k}(x_i - C_k)^2 \qquad (3)$$

Dimana:
K adalah indeks untuk kelompok dari 1 sampai K, $x_i \in S_k$ adalah obyek ke-1 yang merupakan elemen kelompok S ke-k, C_k adalah centroid pada kelompok ke-k. Adapun langkah-langkah algoritma metode *elbow* dalam menentukan nilai K pada *K-Means* yaitu [12]:

1. Mulai
2. Inisialisasi awal nilai *k*
3. Naikkan nilai *k*
4. Hitung hasil *sum of square error* dari tiap nilai K
5. Melihat hasil *sum of square error* dari nilai K yang turun secara drastis
6. Tetapkan nilai *k* yang berbentuk siku
7. Selesai

2.4 Metode *Silhouette Coefficient*

Metode s*ilhouette coefficient* digunakan untuk mengevaluasi kualitas sebuah *clustering* dengan menguji seberapa baik *cluster* dipisahkan dan seberapa padat *cluster* tersebut. Biasanya metode ini digunakan ketika tidak terdapat klasterisasi yang ideal sebagai acuan [15]. Tahap perhitungan *silhouette coefficient* adalah [16]:

1. Menghitung rata-rata jarak tiap dokumen ke-i dengan semua dokumen yang berada dalam satu *cluster*. Nilai ini disebut a(i).
2. Kemudian menghitung rata-rata jarak tiap dokumen ke-i dengan semua dokumen di *Cluster* lain. Mengambil nilai terkecil dari semua jarak rata-rata tersebut. Nilai ini disebut b(i).
3. Kemudian menghitung nilai *silhouette coefficient* dengan menggunakan persamaan:

$$S(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \quad (1)$$

Dimana:
a(i) = rata-rata dokumen ke-I dengan semua dokumen pada satu *cluster* yang sama
b(i) = rata-rata dokumen ke-I dengan semua dokumen pada c*luster* yang berbeda
S(i) = nilai *Silhouette Coefficient*

Nilai *silhouette coefficient* berkisar antara -1 dan 1. Hasil *cluster* dikatakan baik jika nilai *Silhouette Coefficient* adalah 1, berarti dokumen ke-i sudah berada dalam c*luster* yang tepat. Jika nilai s*ilhouette coefficient* adalah 0, maka dokumen ke-i berada di antara dua *cluster*. Jika nilai *silhouette coefficient* adalah -1, artinya struktur *cluster* yang dihasilkan tidak baik, sehingga dokumen ke-i lebih tepat dimasukkan ke dalam *cluster* yang lain. Interpretasi subjektif dari *silhouette coefficient* yang didefinisikan sebagai lebar siluet rata-rata maksimal untuk seluruh kumpulan data ditunjukkan pada Tabel 1.

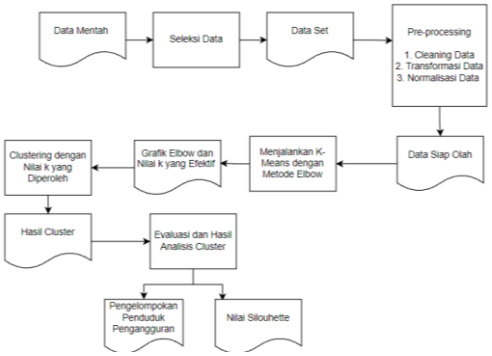
Tabel 1 Pengukuran *Silhouette Coefficient* (Sumber: *Finding Groups in Data: An Introduction to Cluster Analysis*, Leonard Kaufman dan Peter J. Rousseeuw, 2005 [17])

Silhouette Coefficient	**Interpretasi yang diusulkan**
0,71-1	Struktur kuat
0,51-0,70	Struktur baik
0,26-0,50	Struktur lemah
≤ 0,250	Tidak ada struktur substansial yang ditemukan

3. METODE PENELITIAN

3.1 Kerangka Kerja Penelitian

Dalam metodologi penelitian terdapat urutan *framework* penelitian. *Framework* penelitian ini merupakan langkah-langkah yang dilakukan dalam penelitian. Adapun *framework* penelitian yang digunakan dalam penelitian ini adalah seperti terlihat pada Gambar 2 dibawah ini.



Gambar 2 *Framework* Penelitian

3.2 Seleksi Data

Dalam penelitian ini menggunakan data sekunder yang bersumber dari website BPS (Badan Pusat Statistik) tahun 2019 dan tahun 2021. Data yang didapat berbentuk Microsoft Eel. Terdapat 4 atribut yang digunakan dalam penelitian ini yaitu jumlah penduduk, Tingkat Partisipasi Angkatan Kerja (TPAK), Indeks Pembangunan Manusia (IPM), dan rata-rata lama sekolah penduduk umur 15 tahun ke atas. Data tersebut terdiri dari 2 file Eel tahun 2019 dan 2021 dan masing-masing file terdiri 34 *record.*

3.3 Pre-processing

Pre-processing data mencakup antara lain membuang duplikasi data, melakukan integrasi data, merubah skala data dan seleksi fitur pada data, tujuannya agar data yang digunakan mendapatkan hasil informasi yang sesuai dan juga format yang baik tergantung terhadap metode yang digunakan [18]. Setelah mendapatkan data dari BPS (Badan Pusat Statistik) diperlukan adanya tahap *pre-processing* data. Adapun tahapan *pre-processing* data yang dilakukan sebagai berikut:

1. *Cleaning* data
   Setelah 4 atribut ditentukan, maka dilakukan cleaning data yang tidak sesuai sebelum proses pengolahan data lebih lanjut. Seperti data yang salah input karena kekeliruan atau data yang tidak mempunyai nilai.
2. Transformasi data
   Dari hasil proses *cleaning* data selanjutnya dilakukan proses mengubah data dari format file eel xlsx menjadi format csv, tujuannya untuk tahap proses data yang akan digunakan dalam tahap selanjutnya.

3. Normalisasi Data
   Tahap selanjutnya adalah normalisasai data. Normalisasi data adalah proses membuat skala nilai atribut ke dalam rentang yang lebih kecil dengan bobot yang sama. Penelitian ini menggunakan metode *Min-max Normalization* yaitu suatu metode yang melakukan transformasi linear dengan menggunakan nilai minimum dan maksimum yang menghasilkan keseimbangan antara data satu dengan yang lain pada rentang yang sama [19]. Selanjutnya menentukan nilai *k* dengan metode *elbow*.

3.4 Menjalankan *K-Means* dengan Metode Elbow

Penentuan optimal jumlah *cluster* menggunakan metode *elbow* dan *silhouette coefficient* dengan file data eel berbentuk format csv yang sebelumnya sudah di pre-processing. Dalam *clustering* menggunakan algoritma *K-Means* dibutuhkan library seperti Scikit-learn, Numpy, Pandas, Matplotlib, SNS, dan *K-Means* untuk membantu melakukan processing data. Selanjutnya menghitung jarak antara tiap titik data dalam sebuah *cluster* terhadap centroid dengan menggunakan WCSS *(Within-Cluster Sum of Squares)* dan menampilkan grafik nilai *elbow* dengan mencari titik siku. Kemudian mengukur sejauh mana titik dalam suatu *clustering* cocok dengan *cluster* berada dengan menghitung *silhouette coefficient* sehingga dapat menetapkan nilai *k* optimal yang dapat digunakan dalam *clustering*.

3.5 *Clustering* Nilai *k* yang diperoleh

Setelah memperoleh nilai *k* yang optimal dari metode *elbow* menggunakan algoritma *K-Means*. Setiap titik dalam data kemudian ditetapkan ke centroid terdekat dan setiap kumpulan titik yang ditetapkan ke centroid membentuk *cluster* dengan menginisialisasi *K-Means*++ yang berguna untuk memilih centroid *cluster* awal menggunakan pengambilan sampel secara random. Kemudian perhitungan jarak setiap *cluster* titik centroid yang berfungsi untuk memetakan setiap data dengan *cluster* terdekat.

3.6 Evaluasi dan Analisis Hasil *Cluster*

Tahapan evaluasi dan analisis hasil *cluster* berdasarkan data yang diperoleh dari implementasi menggunakan tool *jupyter notebook* berupa pengelompokan data pengangguran dengan karakteristik yang berbeda pada setiap *cluster*nya serta pemberian label disetiap *cluster*. Sebelum menentukan karakteristik setiap *cluster* dilakukan koefisien korelasi antara atribut-atribut dalam dataset. Koefisien korelasi dapat memberikan informasi tentang seberapa kuat hubungan linier antara atribut, dengan nilai berkisar antara -1 hingga 1. Kemudian menganalisis karakteristik setiap c*luster* yang diperoleh dan melihat nilai *Silhouette Coefficient* sehingga dapat mengetahui seberapa baik nilai *k* yang optimal dalam *clustering*.

4. HASIL DAN PEMBAHASAN

4.1 Pre-Processing

Pada tahapan *pre-processing* yang dilakukan pertama kali yaitu membersihkan data yang tidak sesuai, tidak lengkap atau data yang tidak mempunyai nilai (kosong) sebelum proses pengolahan data lebih lanjut. Untuk data yang digunakan dalam penelitian ini tidak ada data yang kosong atau tidak mempunyai nilai. Selanjutnya dilakukan Transformasi data adalah proses mengubah data yang sesuai dengan kebutuhan, seperti mengubah format data. Penelitian ini dilakukan proses mengubah dataset dari format file eel xlsx menjadi format csv, tujuannya untuk tahap proses data yang akan digunakan dalam *jupyter notebook*. Kemudian tahapan Normalisasi data pada tahapan ini melakukan normalisasi *min-max* guna untuk mengubah setiap nilai pada variabel data menjadi nilai yang berada pada rentang nilai 0-1, semua nilai atribut mempunyai nilai minimum 0 dan maksimum 1.

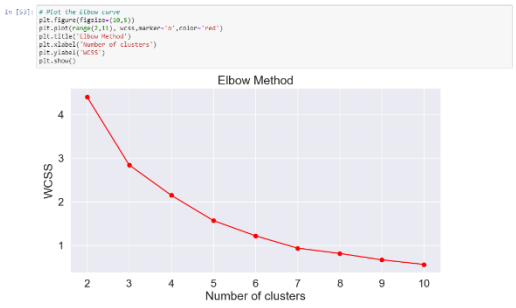
```
In [5]: from sklearn.preprocessing import MinMaxScaler
        # perform Min-max normalization on data
        scaler = MinMaxScaler()
        data_norm = scaler.fit_transform(x)
        data_norm

Out[5]: array([[0.09561203, 0.        , 0.55522088, 0.64319249],
               [0.28852664, 0.52463768, 0.54718876, 0.6713615 ],
               [0.09898898, 0.3442029 , 0.57981928, 0.55633803],
               [0.12703921, 0.13115942, 0.61044177, 0.58685446],
               [0.05939877, 0.19275362, 0.52309237, 0.47183099],
               [0.16143156, 0.32898551, 0.46084337, 0.41079812],
               [0.02640727, 0.50434783, 0.52058233, 0.52347418],
               [0.16061216, 0.42971014, 0.43825301, 0.35446009],
               [0.01563289, 0.28768116, 0.5251004 , 0.35211268],
               [0.03199   , 0.11304348, 0.73493976, 0.76995305],
               [0.20295856, 0.0557971 , 1.        , 1.        ],
               [1.        , 0.13478261, 0.56174699, 0.45539906],
               [0.70057069, 0.41449275, 0.54668675, 0.27699531],
               [0.06565606, 0.69492754, 0.96134538, 0.69953052],
               [0.80801033, 0.46956522, 0.53514056, 0.29577465],
               [0.24869226, 0.05072464, 0.58232932, 0.52112676],
               [0.07586555, 0.77101449, 0.72991968, 0.54929577],
               [0.09222059, 0.45942029, 0.36646586, 0.26525822],
               [0.09811371, 0.52246377, 0.22038153, 0.26525822],
               [0.09001275, 0.38985507, 0.34186747, 0.22300469],
               [0.04043652, 0.44637681, 0.50552209, 0.46478873],
               [0.07285071, 0.40869565, 0.49598394, 0.4084507 ],
               [0.0605058 , 0.20507246, 0.79166667, 0.71126761],
               [0.        , 0.17826087, 0.51757028, 0.56103286],
               [0.03721476, 0.05869565, 0.60993976, 0.65258216],
               [0.04855404, 0.3384058 , 0.43473896, 0.5       ],
               [0.16810063, 0.04855072, 0.54317269, 0.44131455],
               [0.04072414, 0.43333333, 0.52008032, 0.56338028],
               [0.00994877, 0.30797101, 0.38403614, 0.29577465],
               [0.01373128, 0.48985507, 0.24548193, 0.32159624],
               [0.02220056, 0.06086957, 0.43222892, 0.74647887],
               [0.01117581, 0.13913043, 0.39457831, 0.57981221],
               [0.00554549, 0.33188406, 0.1937751 , 0.72065728],
               [0.05486513, 1.        , 0.        , 0.        ]])
```

Gambar 3 Hasil Proses Normalisasi Data *Min-Max*

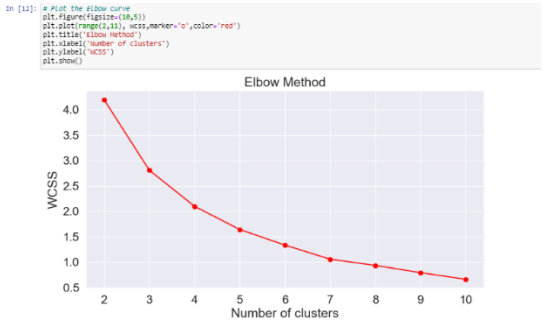
4.2 Menjalankan *K-Means* dengan Metode *Elbow*

Pada tahapan ini dalam pengolahan program *clustering* pada data pengangguran di Indonesia menggunakan alat bantu *jupyter notebook* dengan menggunakan bahasa pemrograman *Python*. Hal yang pertama dilakukan yakni menyiapkan dataset dengan format csv. Sebelum mengimport dataset perlu menginstall library yang dibutuhkan. Selanjutnya menentukan nilai *k* optimal dilakukan dengan menerapkan metode *elbow* dari perhitungan WCSS *(Within-Cluster Sum of Squares)* untuk mengukur seberapa baik titik dalam *cluster* tersebut saling berdekatan dan menghitung nilai *silhouette coefficient* sebagai acuan ketika tidak terdapat klasterisasi yang ideal.

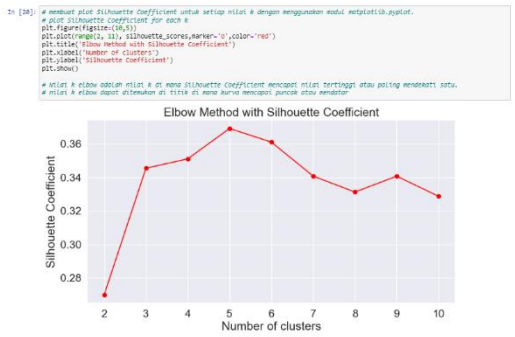
Berdasarkan hasil perhitungan WCSS dari *k=2* hingga *k=10* terlihat ada tiga bentuk sudut siku yang terbentuk pada grafik di *k=3*, di *k=4*, dan *k=5*. Kemudian hasil *silhouette coefficient* terlihat nilai yang tinggi di tahun 2019 menunjukkan *k=5* dan tahun 2021 menunjukkan *k=4*. Dapat disimpulkan bahwa jumlah k optimal dapat ditentukan pada tahun 2019 *k=5* dan 2021 *k=4*.



Gambar 4 Hasil Nilai *k* Data Tahun 2019



Gambar 5 Hasil Nilai *k* Data Tahun 2021



Gambar 6 Grafik *Silhouette Coefficient* Tahun 2019

Gambar 7 Grafik *Silhouette Coefficient* Tahun 2021

4.4 *Cluster*ing dengan nilai *k* yang diperoleh

Hasil dari perhitungan *K-Means* data tahun 2019 terdapat 5 *cluster*, yaitu *cluster 0* sebanyak 2 provinsi terdiri dari: Bali dan DI

Yogyakarta. *Cluster 1* sebanyak 16 provinsi terdiri dari: Kalimantan Selatan, Kalimantan Tengah, Kalimantan Barat, Nusa Tenggara Timur, Nusa Tenggara Barat, Papua Barat, Sulawesi Tenggara, Gorontalo, Sulawesi Tengah, Kep. Bangka Belitung, Lampung, Bengkulu, Sumatera Selatan, Sumatera Barat, Sumatera Utara, Sulawesi Barat. *Cluster 2* sebanyak 3 provinsi terdiri dari: Jawa Tengah, Jawa Timur, dan Jawa Barat. *Cluster 3* sebanyak 12 provinsi yaitu Sulawesi Selatan, Maluku, Maluku Utara, Aceh, Kalimantan Utara, Kalimantan Timur, Banten, DKI Jakarta, Kep. Riau, Jambi, Riau, Sulawesi Utara. *Cluster 4* sebanyak 1 provinsi yaitu Papua. Berikut Tabel 4 Hasil *Clustering* data tahun 2019.

Tabel 2 Hasil *Clustering* Data Tahun 2019

Cluster	Provinsi
Cluster_0	Bali dan DI Yogyakarta.
Cluster_1	Kalimantan Selatan, Kalimantan Tengah, Kalimantan Barat, Nusa Tenggara Timur, Nusa Tenggara Barat, Papua Barat, Sulawesi Tenggara, Gorontalo, Sulawesi Tengah, Kep. Bangka Belitung, Lampung, Bengkulu, Sumatera Selatan, Sumatera Barat, Sumatera Utara, Sulawesi Barat.
Cluster_2	Jawa Tengah, Jawa Timur, Jawa Barat
Cluster_3	Sulawesi Selatan, Maluku, Maluku Utara, Aceh, Kalimantan Utara, Kalimantan Timur, Banten, DKI Jakarta, Kep. Riau, Jambi, Riau, Sulawesi Utara.
Cluster_4	Papua.

Selanjutnya hasil perhitungan *K-Means* data tahun 2021 terdapat 4 *Cluster*, yaitu *Cluster 0* berjumlah 21 provinsi terdiri dari: Aceh, Sulawesi Tengah, Kalimantan Selatan, Kalimantan Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Papua Barat, Banten, Kalimantan Utara, Sulawesi Utara, Maluku Utara, Kep. Bangka Belitung, Lampung, Bengkulu, Sumatera Selatan, Jambi, Riau, Sumatera Barat, Sumatera Utara, dan Maluku. *Cluster 1* sebanyak 3 provinsi terdiri dari: Jawa Barat, Jawa Timur, dan Jawa Tengah. *Cluster 2* sebanyak 5 provinsi terdiri dari: Sulawesi Barat, Papua, Kalimantan Barat, Nusa Tenggara Timur, dan Nusa Tenggara Barat. *Cluster* 3 sebanyak 5 provinsi terdiri dari: DI Yogyakarta, DKI Jakarta, Kep. Riau, Kalimantan Timur, dan Bali. Berikut Tabel 5 Hasil *clustering* data tahun 2021.

Tabel 3 Hasil *Clustering* Tahun 2021

Cluster	Provinsi
Cluster_0	Aceh, Sulawesi Tengah, Kalimantan Selatan, Kalimantan Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Papua Barat, Banten, Kalimantan Utara, Sulawesi Utara, Maluku Utara, Kep. Bangka Belitung, Lampung, Bengkulu, Sumatera Selatan, Jambi, Riau, Sumatera Barat, Sumatera Utara, dan Maluku
Cluster_1	Jawa Barat, Jawa Timur, dan Jawa Tengah
Cluster_2	Sulawesi Barat, Papua, Kalimantan Barat, Nusa Tenggara Timur, dan Nusa Tenggara Barat.
Cluster_3	DI Yogyakarta, DKI Jakarta, Kep. Riau, Kalimantan timur, dan Bali

4.5 Evaluasi dan Analisis Hasil *Cluster*

Pada tahapan ini dilakukan proses koefisien korelasi antara atribut-atribut dalam dataset. Koefisien korelasi dapat memberikan informasi tentang seberapa kuat hubungan linier antara atribut, dengan nilai berkisar antara -1 hingga 1. Hal ini dapat membantu dalam identifikasi hubungan antar atribut sebagai patokan untuk pemberian label tiap *cluster*. Dari hasil yang diperoleh hubungan atribut yang kuat yaitu antara atribut Indeks Pembangunan Manusia (IPM) dan rata-rata lama sekolah penduduk umur 15 tahun ke atas dengan nilai 0.690.

Selanjutnya untuk membedakan hasil pengelompokan yang terbentuk maka dilakukan pelabelan dengan mencari nilai rata-rata setiap atribut. Penamaan *cluster* akan ditentukan dari atribut yang diperoleh dari koefisien korelasi yaitu Indeks Pembangunan Manusia dan rata-rata lama sekolah penduduk umur 15 tahun ke atas dengan kriteria pelabelan yaitu level 1 menempati level tertinggi sedangkan level terendah adalah level 5 yang terdampak pengangguran.

Dapat diketahui nilai rata-rata setiap atribut tahun 2019 dengan hasil data yang sudah di normalisasi sebelumnya pada *Cluster_0* terdapat jumlah penduduk rendah dengan Tingkat Partisipasi Angkatan Kerja tinggi tetapi Indeks Pembangunan Manusia sangat tinggi dan rata-rata lama sekolah penduduk umur 15 tahun ke atas tinggi. Berdasarkan hal tersebut, maka *Cluster_0* sebagai **level 4 yang terdampak pengangguran.**

Pada *Cluster_1* terdapat jumlah penduduk sedang dengan Tingkat Partisipasi Angkatan Kerja sedang tetapi Indeks Pembangunan Manusia rendah dan rata-rata lama sekolah penduduk umur 15 tahun ke atas sedang. Berdasarkan hal tersebut, maka *Cluster_1* sebagai **level 3 yang terdampak pengangguran.**

Pada *Cluster_2* terdapat jumlah penduduk sangat tinggi dengan Tingkat Partisipasi Angkatan Kerja rendah tetapi Indeks Pembangunan Manusia sedang dan rata-rata lama sekolah penduduk umur 15 tahun ke atas rendah. Berdasarkan hal tersebut, maka *Cluster_2* sebagai **level 2 yang terdampak pengangguran.**

Pada *Cluster_3* terdapat jumlah penduduk tinggi dengan Tingkat Partisipasi Angkatan Kerja sangat rendah tetapi Indeks Pembangunan Manusia tinggi dan rata-rata lama sekolah penduduk umur 15 tahun ke atas sangat tinggi. Berdasarkan hal tersebut, maka *Cluster_3* sebagai **level 5 yang terdampak pengangguran.**

Pada *Cluster_4* terdapat jumlah penduduk sangat rendah dengan Tingkat Partisipasi Angkatan Kerja sangat tinggi tetapi Indeks Pembangunan Manusia sangat rendah dan rata-rata lama sekolah penduduk umur 15 tahun ke atas sangat rendah. Berdasarkan hal tersebut, maka *Cluster_4* sebagai **level 1 yang terdampak pengangguran.**

Tabel 4 Nilai Rata-Rata *Cluster* Tahun 2019

Cluster	Jumlah Penduduk	Angkatan Kerja Terhadap Penduduk Usia Kerja (TPAK)	Indeks Pembangunan Manusia	Rata-Rata Lama Sekolah Penduduk Umur 15 Tahun keatas
Cluster_0	0,071	0,733	0,846	0,624
Cluster_1	0,079	0,409	0,424	0,431
Cluster_2	0,836	0,340	0,548	0,343
Cluster_3	0,089	0,103	0,608	0,640
Cluster_4	0,055	1,000	0,000	0,000

Dapat diketahui nilai rata-rata setiap atribut tahun 2021 dengan hasil data yang sudah di normalisasi sebelumnya pada *Cluster_0* terdapat jumlah penduduk sedang dengan Tingkat Partisipasi Angkatan Kerja rendah tetapi Indeks Pembangunan Manusia sedang dan rata-rata lama sekolah penduduk umur 15 tahun ke atas tinggi. Berdasarkan hal tersebut, maka *Cluster_0* sebagai **level 3 yang terdampak pengangguran.**

Pada *Cluster_1* terdapat jumlah penduduk sangat tinggi dengan Tingkat Partisipasi Angkatan Kerja sedang tetapi Indeks Pembangunan Manusia tinggi dan rata-rata lama sekolah penduduk umur 15 tahun ke atas sedang. Berdasarkan hal tersebut, maka *Cluster_1* sebagai **level 2 yang terdampak pengangguran.**

Pada *Cluster_2* terdapat jumlah penduduk rendah dengan Tingkat Partisipasi Angkatan Kerja sangat tinggi tetapi Indeks Pembangunan Manusia rendah dan rata-rata lama sekolah penduduk umur 15 tahun ke atas rendah. Berdasarkan hal tersebut, maka *Cluster_2* sebagai **level 1 yang terdampak pengangguran.**

Pada *Cluster_3* terdapat jumlah penduduk tinggi dengan Tingkat Partisipasi Angkatan Kerja tinggi tetapi Indeks Pembangunan Manusia dan rata-rata lama sekolah penduduk umur 15 tahun ke atas sangat tinggi. Berdasarkan hal tersebut, maka *Cluster_2* sebagai **level 4 yang terdampak pengangguran.**

Tabel 5 Nilai Rata-Rata *Cluster* Tahun 2021

Cluster	Jumlah Penduduk	Angkatan Kerja Terhadap Penduduk Usia Kerja (TPAK)	Indeks Pembangunan Manusia	Rata-Rata Lama Sekolah Penduduk Umur 15 Tahun keatas
Cluster_0	0,084	0,300	0,507	0,537
Cluster_1	0,862	0,373	0,568	0,362
Cluster_2	0,077	0,627	0,251	0,218
Cluster_3	0,088	0,405	0,845	0,767

5. KESIMPULAN

Berdasarkan hasil penelitian dan pembahasan yang telah dilakukan, *clustering* tahun 2019 menghasilkan *k=5* dan tahun 2021 menghasilkan *k=4*. Kriteria pelabelan yang digunakan pada penelitian ini yaitu level 1 menempati level tertinggi sedangkan level terendah adalah level 5 yang terdampak pengangguran. Terdapat perpindahan provinsi pada tahun 2019 dan 2021 yaitu provinsi Kalimantan Barat, NTT, NTB, dan Sulawesi Barat dari level 3 berpindah ke level 1 terdampak pengangguran. Provinsi Sulawesi Selatan, Maluku, Maluku Utara, Aceh, Kalimantan Utara, Kalimantan Timur, Banten, Jambi, Riau, dan Sulawesi Utara dari yang memiliki level 5 berpindah ke level 3 terdampak pengangguran. Provinsi Kalimantan Timur, DKI Jakarta, dan Kep. Riau dari yang memiliki level 5 berpindah ke level 4 terdampak pengangguran. Hasil evaluasi menunjukkan tahun 2019 menghasilkan nilai *silhouette* 0.36 dan 2021 menghasilkan nilai *silhouette* 0.35. Kedua hasil *silhouette* tersebut termasuk dalam kriteria struktur *cluster* lemah karena nilai *silhouette* berada pada rentang nilai 0,26-0,50.

6. REFERENSI

[1] G. D. B. Patra, I. Nuraini and M. K. Fuddin, "Faktor Yang Mempengaruhi Tingkat Pengangguran Di Beberapa Negara Asean," *Jurnal Ilmu Ekonomi (JIE)*, p. 409~420, 2022.

[2] A. A. R. Syahputra, S. Syahnaztia and Nurfahmiyati, "Faktor Penentu Tingkat Pengangguran Usia Produktif Awal Pada Masa Pandemi Covid-19 Di Kecamatan Sukajadi," *Bina Ekonomi,* pp. 50-62, 2022.

[3] R. Firdhania and F. Muslihatinningsih, "Faktor-Faktor yang Mempengaruhi Tingkat Pengangguran di Kabupaten Jember," *e-Journal Ekonomi Bisnis dan Akuntansi*,, pp. 117-121, 2017.

[4] F. N. A. Zahro, F. Abimanyu, A. N. Azhar and E. Widodo, "Analisis Faktor-Faktor yang Mempengaruhi Tingkat Pengangguran di Sulawesi Utara Pada Masa Pandemi COVID-19 Tahun 2020," *Jurnal Ekonomi Pembangunan,* pp. 180-193, 2021.

[5] F. Badri and A. Habibi, "Implementasi Metode K-Mean Clustering Analysis pada Pengelompokan Pengangguran di Indonesia sebagai Dampak dari Pandemi Covid-19," *ILKOMNIKA: Journal of Computer Science and Applied Informatics,* pp. 171-179, 2022.

[6] F. A. Tanjung, A. P. Windarto and M. Fauzan, "Penerapan Metode K-Means Pada Pengelompokkan Pengangguran Di Indonesia," *Jurnal Riset Sistem Informasi Dan Teknik Informatika,* pp. 61-74, 2021.

[7] Akramunnisa and Fajriani, "K- Means Clustering Analysis Pada Persebaram Tingkat Pengangguran Kabupaten/Kota Di Sulawesi Sela," *Jurnal Varian,* pp. 103-112, 2020.

[8] K. Fatmawati and A. P. Windarto, "Data Mining: Penerapan Rapidminer Dengan K-Means Cluster Pada Daerah Terjangkit Demam Berdarah Dengue (Dbd) Berdasarkan Provinsi," *CESS (Journal of Computer Engineering System and Science),* pp. 173-178, 2018.

[9] A. Winarta and W. J. Kurniawan, "Optimasi Clustering K-Means Menggunakan Metode Elbow Pada Pengguna Narkoba Dengan Pemrograman Python," *Jurnal Teknik Informatika Kaputama (JTIK)*, pp. 113-119, 2021.

[10] S. C. Sihombing and D. A. Sihombing, "Pengelompokan Tingkat Kesejahteraan Masyarakat di Sumatera Utara dengan Metode K-Means Clustering," *Jurnal Matematika Integratif*, pp. 127-135, 2021.

[11] N. T. Hartanti, "Metode Elbow dan K-Means Guna Mengukur Kesiapan Siswa SMK Dalam Ujian Nasional," *Jurnal Nasional Teknologi dan Sistem Informasi ,* pp. 82-89, 2020.

[12] E. Muningsih and S. Kiswati, "Sistem Aplikasi Berbasis Optimasi Metode Elbow Untuk Penentuan Clustering Pelanggan," *JOUTICA ,* 2018.

[13] A. Rahmadhani, "Pendekatan Clustering untuk Menganalisis Efisiensi dan Kinerja Mahasiswa Berdasarkan Data Menerapkan Metode K-Means," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, pp. 2461-2468, 2022.

[14] H. E. Prastyo and F. Ilfana, "Pengelompokan Kabupaten Dan Kota Di Jawa Timur Berdasarkan Indeks Pembangunan Manusia Dengan Menggunakan Metode K-Means Tahun 2020-2021," *Jurnal Ilmiah Komputasi dan Statistika*, pp. 22-32, 2022.

[15] M. Orisa and A. Faisol, "Analisis Algoritma Partitioning Around Medoid untuk Penentuan Klasterisasi," *Jurnal Teknologi Informasi dan Terapan (J-TIT)*, pp. 86-90, 2021.

[16] D. A. C. Rachman, R. Goejantoro and F. D. T. Amijaya, "Implementasi Text Mining Pengelompokkan Dokumen Skripsi Menggunakan Metode K-Means Clustering," *Jurnal EKSPONENSIAL*, pp. 167-174, 2020.

[17] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Canada: simultraneosly in Canada, 2005.

[18] I. R. Drl, Y. H. Chrisnanto and F. R. Umbara, "Analisis Cluster Pada Kelompok Masyarakat Yang Rentan Terhadap Paparan COVID-19

Menggunakan Metode K-Means Clustering Dan Visualisasi Dengan SIG," *Informatics And Digital Expert (INDEX)*, pp. 61-69, 2022.

[19] G. A. B. Suryanegara, Adiwijaya and M. D. Purbolaksono, "Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi," *Jurnal Resti*, pp. 114-122, 2021.