

Pengelompokan Penanaman Modal Asing Di Surabaya Menggunakan Metode K-Means

Afzeni Fitri Hanik¹, Dian Yuliati, M.Si²

¹Matematika, Saintek, UIN Sunan Ampel Surabaya

²Matematika, Saintek, UIN Sunan Ampel Surabaya

[109040222049@student.uinsby.ac.id](mailto:09040222049@student.uinsby.ac.id), [2Dian.yuliati@uinsa.ac.id](mailto:Dian.yuliati@uinsa.ac.id)



All publications by Journal Of Information Technology is licensed under a [Lisensi Creative Commons Atribusi 4.0 Internasional](#). (CC BY 4.0)

Abstract— Surabaya City, as one of the centers of economic growth in Indonesia, continues to strive to attract investment from within and outside the country. This study uses investment realization data from DPMPTSP Surabaya for 2019-2024 to group investor countries using the K-Means method. Four variables used in the analysis are investment value, number of projects, number of Indonesian Migrant Workers (TKI), and number of Foreign Migrant Workers (TKA). Before clustering, the data was reduced using Principal Component Analysis (PCA) which produced two main components without losing important information. Determination of the optimal number of clusters was carried out using the Elbow and Silhouette methods which consistently showed that two clusters were the optimal number. The clustering results showed two groups, namely high and low investment clusters. Clustering validation used the Davies-Bouldin Index (DBI), with a value of 0.5084009 which indicated that the cluster results were quite good. This study aims to provide additional information that is useful in developing strategies to increase investment value in Surabaya, as well as helping to identify countries with high investment performance to be maintained.

Intisari— Kota Surabaya, sebagai salah satu pusat pertumbuhan ekonomi di Indonesia terus berupaya menarik investasi dari dalam maupun luar negeri. Penelitian ini menggunakan data realisasi investasi dari DPMPTSP Surabaya tahun 2019-2024 untuk mengelompokkan negara investor menggunakan metode K-Means. Empat variabel yang digunakan dalam analisis yaitu nilai investasi, jumlah proyek, jumlah Tenaga Kerja Indonesia (TKI), dan jumlah Tenaga Kerja Asing (TKA). Sebelum dilakukan klasterisasi, data direduksi menggunakan Principal Component Analysis (PCA) yang menghasilkan dua komponen utama tanpa kehilangan informasi penting. Penentuan jumlah klaster optimal dilakukan menggunakan metode Elbow, dan Silhouette yang secara konsisten menunjukkan bahwa dua klaster merupakan jumlah optimal. Hasil klastering menunjukkan adanya dua kelompok, yaitu klaster investasi tinggi dan rendah. Validasi klastering menggunakan Davies-Bouldin Index (DBI), dengan nilai sebesar 0,5084009 yang menunjukkan bahwa

hasil klaster cukup baik. Penelitian ini bertujuan untuk memberikan informasi tambahan yang bermanfaat dalam menyusun strategi peningkatan nilai investasi di Surabaya, serta membantu mengidentifikasi negara-negara dengan kinerja investasi yang tinggi untuk dipertahankan.

Kata Kunci— Investasi, PMA, Clustering, K-Means.

I. PENDAHULUAN

Investasi merupakan kegiatan penanaman modal yang bertujuan untuk mendorong pertumbuhan sektor ekonomi[1]. Salah satu bentuk investasi yang berpengaruh adalah investasi asing, yang tidak hanya berupa modal tetapi juga berupa teknologi, pengetahuan manajerial, memperluas koneksi pasar, dan mempercepat transfer pengetahuan[2]. Kehadiran investasi asing dapat mendorong terciptanya lapangan kerja, peningkatan produktivitas, serta meningkatkan pembangunan infrastruktur dan industri di Indonesia. Kota Surabaya sebagai salah satu pusat pertumbuhan ekonomi di Indonesia terus berupaya menarik investasi, baik dalam negeri maupun luar negeri. Untuk mendukung hal tersebut, pemerintah kota melalui Dinas Penanaman Modal dan Pelayanan Terpadu Satu Pintu (DPMPTSP) menyediakan kemudahan layanan perizinan usaha dan pembangunan, kini dapat diakses secara terpusat tanpa harus melalui dinas-dinas teknis secara terpisah[3]. Kemudahan ini bertujuan agar proses investasi berjalan lebih efisien, termasuk dalam pengurusan izin-izin seperti Nomor Induk Berusaha (NIB), Persetujuan Bangunan Gedung (PBG), hingga Sertifikat Laik Fungsi (SLF).

Berdasarkan data dari DPMPTSP Surabaya tahun 2019 hingga 2024, realisasi investasi yang masuk ke Surabaya mencakup berbagai sektor seperti kesehatan, perhotelan, restoran, dan sektor lainnya[4]. Setiap proyek investasi tidak hanya berkontribusi pada pertumbuhan ekonomi, tetapi juga

menciptakan peluang kerja bagi masyarakat lokal maupun tenaga kerja asing. Dalam konteks ini, Tenaga Kerja Indonesia (TKI) merujuk pada warga negara Indonesia yang dipekerjakan dalam proyek investasi, sedangkan Tenaga Kerja Asing (TKA) adalah warga negara asing yang bekerja di proyek tersebut. Data yang tersedia meliputi negara asal investor, nilai investasi, jumlah proyek, serta jumlah tenaga kerja Indonesia (TKI), dan tenaga kerja asing (TKA). Dalam penelitian ini, keempat variabel yang tersedia yaitu nilai investasi, jumlah proyek, TKI, dan TKA, akan dianalisis menggunakan *Principal Component Analysis* (PCA) proses reduksi bertujuan untuk meningkatkan kualitas hasil pengelompokan dalam metode K-means, dengan menghilangkan redundansi antar variabel tanpa kehilangan informasi penting dari data.

Meskipun DPMPTSP Kota Surabaya telah menerima penanaman modal asing dari berbagai negara, belum terdapat analisis yang mendalam untuk mengelompokkan negara-negara tersebut berdasarkan karakteristik atau nilai investasinya. Tanpa segmentasi ini, strategi promosi investasi cenderung bersifat umum dan belum terarah. Oleh karena itu, diperlukan pendekatan berbasis data seperti K-Means Clustering untuk mengidentifikasi kelompok negara dengan kinerja investasi yang tinggi, dan rendah. K-Means bekerja dengan cara mengelompokkan data berdasarkan kemiripan maksimal antar data[5]. Selain itu, algoritma ini memiliki keunggulan dalam menangani data berukuran besar secara efisien melalui proses iterasi sederhana[6]. Melalui hasil pengelompokan data tersebut, dapat dirancang strategi yang tepat untuk meningkatkan nilai investasi yang rendah serta mempertahankan investasi yang tinggi.

Penelitian Sylviani (2025) mengelompokkan karakteristik angkatan kerja di Indonesia berdasarkan variabel jenis kelamin, usia, tingkat pendidikan, jenis pekerjaan, dan pendapatan bulanan dengan menggunakan metode K-Means serta *Principal Component Analysis* (PCA) sebagai teknik reduksi dimensi[7]. Data yang digunakan berasal dari Survei Angkatan Kerja Nasional (SAKERNAS) tahun 2021 yang diselenggarakan oleh BPS. Hasil analisis menghasilkan enam klaster yang menunjukkan bahwa pekerja dengan tingkat pendidikan yang lebih tinggi dan usia yang lebih tua cenderung bekerja di sektor formal dengan pendapatan lebih tinggi, sedangkan pekerja dengan pendidikan rendah dan usia muda lebih banyak berada di sektor informal dengan pendapatan lebih rendah. Meskipun penelitian ini mampu menggambarkan variasi sosial ekonomi tenaga kerja secara umum, salah satu kekurangannya adalah belum dilakukan validasi eksternal terhadap hasil clustering dan belum mempertimbangkan faktor geografis atau wilayah yang dapat memberikan konteks lebih spesifik terhadap distribusi karakteristik tenaga kerja.

Oleh karena itu, diperlukan suatu pendekatan analisis yang mampu mengelompokkan negara-negara tersebut ke dalam kategori yang relevan, seperti klaster tinggi dan rendah. Metode K-Means dipilih dalam penelitian ini karena mampu mengelompokkan data secara objektif berdasarkan atribut

numerik. Melalui pendekatan ini, diharapkan dapat memberikan informasi tambahan yang bermanfaat dalam menyusun strategi peningkatan nilai investasi di Surabaya.

II. TINJAUAN PUSTAKA

A. Investasi

Investasi merupakan kegiatan penanaman modal yang bertujuan untuk meningkatkan pertumbuhan sektor ekonomi[8]. Investasi mengacu pada aset individu atau perusahaan untuk menyertakan modalnya sehingga memperoleh imbal hasil dalam jangka waktu tertentu. Investasi dikategorikan dua jenis yaitu investasi dalam negeri (PMDN) dan investasi modal asing (PMA)[9]. PMDN adalah investasi yang dilakukan oleh investor dalam negeri, yang seluruh modalnya berasal dari Indonesia. PMA diartikan sebagai penanaman modal dilakukan pihak swasta pemilik modal di negara asal, PMA salah satu cara para investor asing berinvestasi dengan cara membangun, membeli total atau mengakuisisi perusahaan[10].

Di kota Surabaya, Penanaman Modal Asing (PMA) memiliki peranan strategis seperti mentransfer teknologi dan keahlian, membuka akses ke pasar global, menciptakan lapangan pekerjaan serta penting untuk mendorong pertumbuhan ekonomi lokal[11]. Surabaya sebagai kota metropolitan dan salah satu pusat ekonomi di Jawa Timur memiliki potensi besar untuk menarik investor asing, terutama di sektor perdagangan, jasa, industri, serta bangunan dan lahan seperti perumahan, gedung perkantoran, pusat perbelanjaan[12]. Pemerintahan kota melalui DPMPTSP terus mendorong iklim investasi yang kondusif agar Surabaya tetap menjadi tujuan utama penanaman modal asing.

B. K-Means

K-Means clustering merupakan salah satu teknik dalam analisis data yang berfungsi untuk mengelompokkan sekumpulan data ke dalam beberapa klaster tertentu berdasarkan kemiripannya. Proses pengelompokan ini dilakukan dengan menghitung jarak antara data dan pusat klaster (centroid), lalu setiap data akan ditempatkan ke dalam klaster yang memiliki centroid terdekat. Algoritma ini termasuk kategori *unsupervised learning* karena tidak memerlukan data yang diberi label [13]. K-Means bertujuan untuk menimalkan variasi dalam klaster dan memaksimalkan perbedaan antar klaster[14].

Tahapan dari K-Means[15]:

1. Memilih jumlah klaster
2. Menginisialkan centroid secara acak
3. Menghitung centroid baru menggunakan rumus[16]:

$$\mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} x_q \quad (1)$$

Keterangan:

μ_k = titik centroid dari klaster ke-k

N_k = banyaknya data pada klaster ke-k

x_q = data ke-q pada klaster ke-k

4. Menghitung jarak data ke centroid menggunakan

rumus Euclidean distance

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

Keterangan:

X_{1i} = data ke-i

C_{1j} = centroid ke-j

5. Jika terdapat data yang berpindah cluster, maka proses perhitungan centroid diulang hingga data menempati cluster yang sama

C. Winsorizing Outlier

Winsorizing merupakan teknik yang digunakan untuk mengatasi outlier atau data yang nilainya ekstrem[17]. Teknik ini melakukan transformasi pada data yang terdeteksi sebagai outlier dengan menggantinya menggunakan nilai pada persentil tertentu, misalnya persentil ke-5 dan ke-95[18].

D. Normalisasi Z-Score

Metode normalisasi Z-score yaitu teknik standarisasi data dengan mengubah skala nilai berdasarkan rata-rata (mean) dan simpangan baku (standard deviation) dari masing-masing atribut[19]. Penggunaan normalisasi ini bermanfaat karena mampu mengurangi dominasi atribut yang memiliki skala besar terhadap atribut skala kecil. Selain itu, metode ini juga membantu mengurangi efek keberadaan outlier sehingga hasil analisis tidak bias terhadap nilai-nilai yang sangat tinggi atau rendah[20]. Adapun persamaan yang digunakan[21]:

$$z = \frac{x - \mu}{\sigma} \quad (3)$$

Dimana:

x : nilai yang akan dinormalisasi

μ : rata-rata atribut

σ : standart deviasi atribut

E. PCA

Kompleksitas data dapat memengaruhi akurasi hasil klasifikasi, sehingga diperlukan proses reduksi dimensi untuk meminimalkan kesalahan klasifikasi. Salah satu metode yang dapat digunakan adalah **Principal Component Analysis (PCA)**. PCA bekerja dengan membentuk himpunan dimensi baru yang kemudian diurutkan berdasarkan besarnya variansi data. Adapun langkah algoritma PCA sebagai berikut [22]:

1. Menghitung mean data tiap dimensi

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (4)$$

n: jumlah data sampel

X_i : data sampel

2. Menghitung covariance matrix

$$C_x = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^T \quad (5)$$

n: jumlah data sampel

X_i : data sampel

\bar{X} : mean

3. Menghitung eigenvector dan eigenvalue dari covariance matrix
- $C_x v_m = \lambda_m v_m \quad (6)$
4. Urutkan eigenvalue secara descending. *Principal Component* (PC) adalah deretan eigenvector sesuai dengan urutan eigenvalue pada tahap ke-3
5. Menghasilkan dataset baru

F. Metode Elbow

Metode elbow merupakan metode yang digunakan untuk menentukan jumlah klaster data secara optimal[23]. Metode ini meghasilkan grafik dengan sumbu x yang merepresentasikan jumlah klaster dan sumbu y yang merepresentasikan nilai SSE (Sum of Squares Errors)[6]. Jumlah klaster yang optimal ditunjukkan oleh titik yang membentuk siku yaitu saat penurunan SSE mulai melambat secara signifikan[24].

$$SSE = \sum_{i=1}^K \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (7)$$

keterangan:

K = jumlah klaster

x = data dalam klaster

μ_i = centroid klaster ke - i

C_i = klaster ke - i

G. Metode Silhouette

Metode *Silhouette* merupakan metode yang digunakan untuk menentukan jumlah klaster optimal, dengan cara meggabungkan dua konsep utama yaitu *cohesion* dan *separation*, dimana *cohesion* mengukur seberapa dekat data dalam satu klaster, sedangkan *separation* mengukur seberapa jauh suatu klaster dengan klaster lainnya, sehingga nilai *Silhouette* yang tinggi menunjukkan klaster terpisah secara jelas [25]. Koefisien *Silhouette* memiliki rentang antara -1 hingga 1, dimana nilai yang semakin mendekati 1 menunjukkan kualitas pengelompokan yang semakin baik[26]. Adapun langkah-langkah dalam menghitung nilai *Silhouette* sebagai berikut[27]:

1. Menghitung jarak rata-rata antara titik i dan semua titik lain dalam klaster yang sama

$$a(i) = \frac{1}{|C_i|-1} \sum_{j \in C_i, j \neq i} d(i, j) \quad (8)$$

2. Menghitung jarak rata-rata antara titik i dan semua titik dalam klaster terdekat yang berbeda dari klasternya

$$b(i) = \min_{C_k \neq C_i} \left(\frac{1}{|C_k|} \sum_{j \in C_k} d(i, j) \right) \quad (9)$$

3. Nilai koefisien *Silhouette* dapat dihitung menggunakan persamaan ini:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (10)$$

Keterangan:

C_i = klaster tempat titik i berada
 $d(i, j)$ = Jarak antar titik i dengan j
 C_k = klaster lain selain C_i

H. Davies Bouldin Index

Davies-Bouldin Index (DBI) merupakan metode yang digunakan untuk mengevaluasi hasil K-Means tersebut, yaitu dengan mengukur seberapa baik klaster terpisah[28]. Semakin rendah nilai DBI (Davies-Bouldin Index) yang diperoleh (mendekati nol tidak negatif) ,maka semakin baik kualitas klaster yang dihasilkan[29][30].

Rumus metode DBI[31]:

$$\text{DBI} = \frac{1}{K} \sum_{i=1}^k \max_{i \neq j} \left(\frac{s_i + s_j}{M_{ij}} \right) \quad (11)$$

Keterangan:

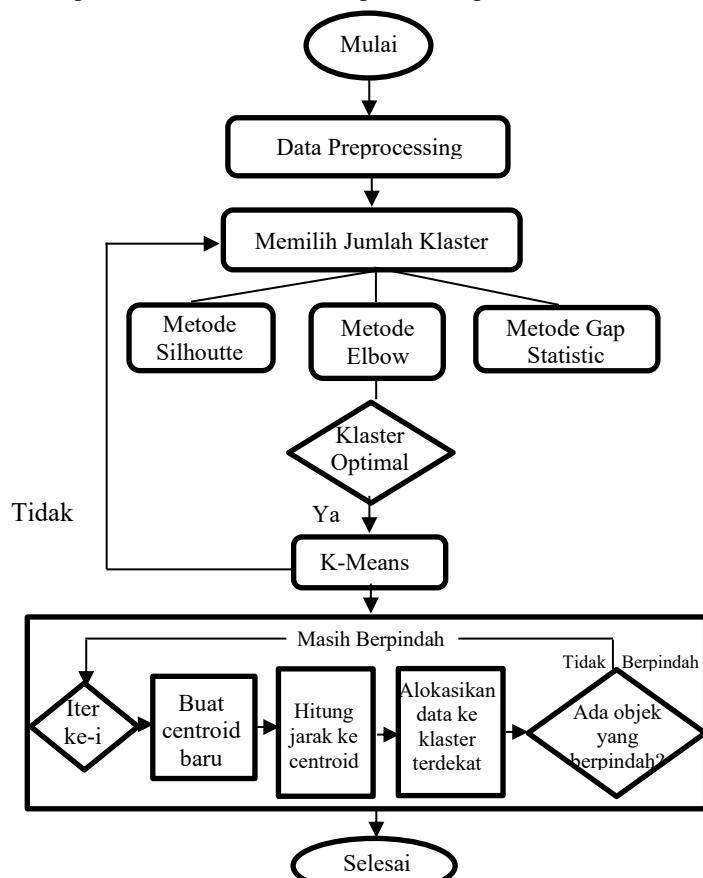
s_i = rata-rata jarak antara setiap titik dalam klaster i ke centroid klaster i

s_j = rata-rata jarak antara setiap titik dalam klaster i ke centroid klaster j

M_{ij} = jarak antara centroid klaster i dan centroid klaster j
 K = jumlah klaster

III. METODOLOGI PENELITIAN

Penelitian ini merupakan penelitian kuantitatif dengan pendekatan deskriptif, yang bertujuan untuk mengelompokkan data realisasi investasi asing menggunakan metode K-Means clustering. Metode ini dipilih karena mampu mengelompokkan data berdasarkan kemiripan karakteristiknya. Alur tahapan penelitian secara umum dapat dilihat pada flowchart berikut:



Gambar 1 Tahapan penelitian

A. Data preprocessing

Tahapan preprocessing dilakukan sebelum analisis klasterisasi, yang meliputi pengecekan missing value dan data duplikat. Hasil pemeriksaan menunjukkan bahwa tidak terdapat missing value maupun data duplikat pada data realisasi investasi asing. Selanjutnya dilakukan deteksi outlier menggunakan metode *boxplot*, outlier yang teridentifikasi kemudian ditangani menggunakan teknik *winsorizing*, setelah itu dinormalisasi menggunakan metode *z-score*. Selanjutnya, dilakukan reduksi dimensi menggunakan *Principal Component Analysis* (PCA) setelah melalui uji multikolinearitas, *Bartlett's Test of Sphericity*, dan uji *Kaiser-Mayer-Olkin* (KMO). Dari hasil PCA, diperoleh dua komponen utama yang digunakan dalam analisis klasterisasi.

B. Menentukan jumlah klaster

Penentuan jumlah klaster optimal dilakukan menggunakan tiga metode, yaitu *Elbow Method*, *Silhouette Coefficient*, dan *Gap Statistic*. Ketiga metode tersebut menunjukkan hasil yang konsisten, dimana jumlah klaster optimal adalah sebanyak dua klaster. Hasil ini menjadi dasar dalam penerapan algoritma K-Means pada tahap analisis klasterisasi data realisasi investasi asing di Surabaya.

C. Penerapan K-Means

Setelah melalui tahap preprocessing dan penentuan jumlah klaster optimal, algoritma K-Means diterapkan untuk mengelompokkan data realisasi investasi asing berdasarkan tiga variabel : nilai investasi, jumlah TKI, dan jumlah TKA. Proses ini dilakukan dengan iterasi hingga pusat klaster stabil dan tidak mengalami perubahan signifikan, sehingga diperoleh hasil pengelompokan yang optimal. Pada penelitian ini iterasi dilakukan sebanyak 5 kali.

D. Validasi

Validasi hasil klaster dilakukan menggunakan *Davies-Bouldin Index* (DBI) untuk mengukur kualitas pemisahan antar klaster. Nilai DBI yang lebih kecil menunjukkan bahwa klaster yang terbentuk semakin baik, dengan jarak antar klaster yang jelas dan tingkat kemiripan dalam klaster yang tinggi.

IV. HASIL DAN PEMBAHASAN**A. Deskripsi data**

Data yang digunakan dalam penelitian ini berasal dari tempat magang Saya yaitu Dinas Penanaman Modal dan Pelayanan Terpadu Satu Pintu Surabaya. Data yang tersedia merupakan data investasi asing (PMA) dari tahun 2019-2024 yang terdiri dari 5 kolom yaitu, negara, jumlah proyek, nilai investasi, TKI, dan TKA. Gambaran data dapat dilihat dalam tabel berikut:

Tabel 1 Data investasi 2019-2024

DATA REALISASI INVESTASI PENANAMAN MODAL ASING (PMA) BERDASARKAN NEGARA TAHUN 2019-2024					
No	Negara	Nilai Investasi	Jumlah Proyek	TKI	TKA
1	China	Rp 1,661,258,491,647	367	2073	100
2	India	Rp 856,043,828,627	73	377	10
3	Singapura	Rp 3,719,112,555,122	614	3248	19
4	Malaysia	Rp 1,997,608,210,457	61	483	9
5	Belanda	Rp 1,140,639,311,335	88	5273	9
:	:	:	:	:	:
40	Selandia Baru	Rp 543,140,000	6	1	1
41	Uni Emirat Arab	Rp 26,951,300,000	12	90	1
42	Vietnam	Rp 204,000,000	1	0	0
43	Finlandia	Rp 2,956,500,000	2	0	0
44	Nepal	Rp 18,903,000,000	17	0	0

B. Preprocessing data

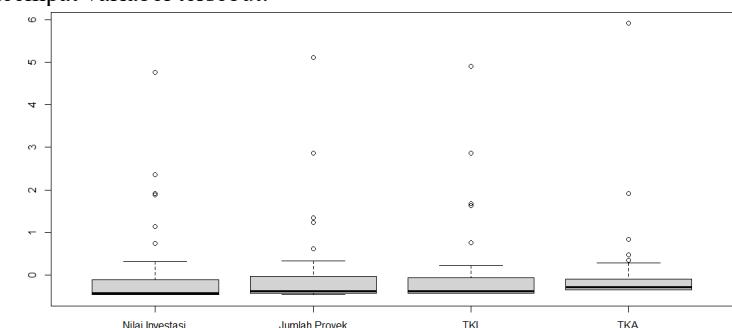
Sebelum dilakukan pengolahan dengan metode K-Means, data terlebih dahulu diperiksa untuk memastikan tidak terdapat missing value (nilai kosong). Berdasarkan hasil pengecekan (Gambar 2), tidak ditemukan nilai yang kosong. Selanjutnya, data juga diperiksa untuk mendeteksi kemungkinan adanya duplikasi. Jika ditemukan data ganda, maka data tersebut akan dihapus agar proses K-Means clustering dapat berjalan optimal. Hasil pengecekan (Gambar 2) menunjukkan bahwa tidak terdapat data duplikat dalam data realisasi investasi tersebut.

```
> #cek missing value
> any(is.na(PMA_Investasi2))
[1] FALSE
> #CEK DUPLIKAT
> any(duplicated(PMA_Investasi2))
[1] FALSE
```

Gambar 2 Output preprocessing

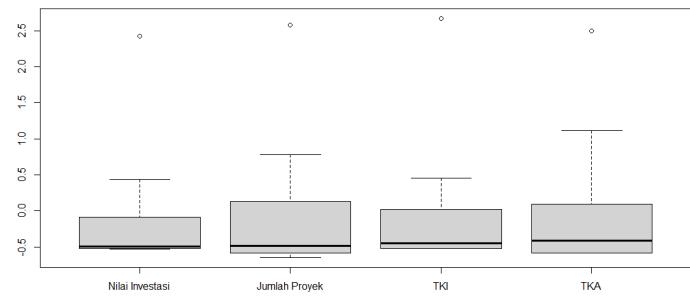
Selanjutnya dilakukan deteksi outlier menggunakan metode *boxplot* terhadap empat variabel, yaitu nilai investasi, jumlah proyek, jumlah Tenaga Kerja Indonesia (TKI), dan jumlah Tenaga Kerja Asing (TKA). Outlier yang teridentifikasi (Gambar 3) kemudian ditangani menggunakan teknik

winsorizing 95% yaitu dengan mengganti 2,5% nilai terendah dan 2,5% nilai tertinggi dengan nilai pada persentil ke-2,5 dan ke-97,5. Setelah proses penanganan outlier, data dinormalisasi menggunakan metode *z-score* untuk menyetarakan skala antar variabel (Gambar 4). Gambar 4 menunjukkan bahwa sudah tidak terdapat outlier lagi yang berarti, maka dilanjutkan analisis *Principal Component Analysis* (PCA) terhadap keempat variabel tersebut.



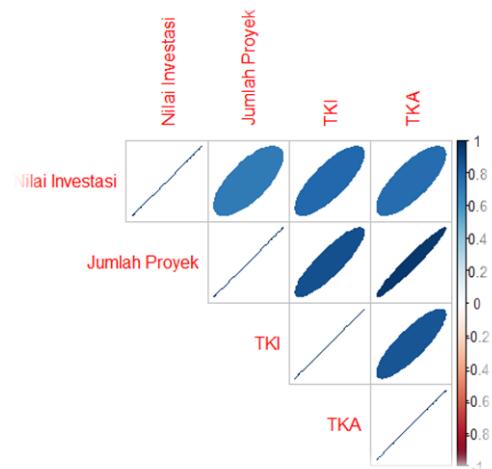
Gambar 3 Outlier

Boxplot Setelah Winsorizing



Gambar 4 Setelah Winsorizing

Sebelum menerapkan PCA, terlebih dahulu dilakukan pemeriksaan multikolinearitas melalui korelasi antar variabel serta pengujian *Bartlett's Test of Sphericity* untuk memastikan bahwa variabel-variabel memiliki korelasi yang memadai.



```
Bartlett's Test of Sphericity
call: bart_spher(x = data_normalisasi)
X2 = 225.191
df = 6
p-value < 2.22e-16
```

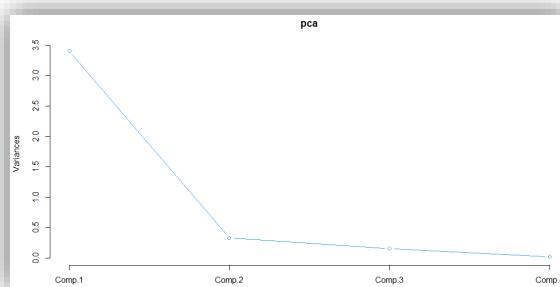
Gambar 5 Uji Multikolinearitas dan Uji Bartlett's

Hasilnya (Gambar 5) menunjukkan bahwa keempat variabel (nilai investasi, jumlah proyek, jumlah TKI, dan jumlah TKA) menunjukkan hubungan korelasi yang cukup kuat, yang ditandai dengan warna gradasi biru tua, mengindikasikan nilai korelasi positif tinggi. Selanjutnya, dilakukan uji *Bartlett's Test of Sphericity* untuk menguji apakah matriks korelasi secara keseluruhan berbeda secara signifikan dari matriks identitas. Hasil uji (Gambar 5) menunjukkan nilai p-value < 2.22e-16, yang berarti signifikan pada Tingkat kepercayaan 95%. Dengan demikian, data layak untuk dilakukan analisis PCA karena terdapat korelasi yang cukup antar variabel.

```
> KMO(data_normalisasi)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = data_normalisasi)
Overall MSA = 0.69
MSA for each item =
Nilai Investasi Jumlah Proyek TKI TKA
0.70 0.63 0.78 0.66
```

Gambar 6 Kaiser-Mayer-Olkin

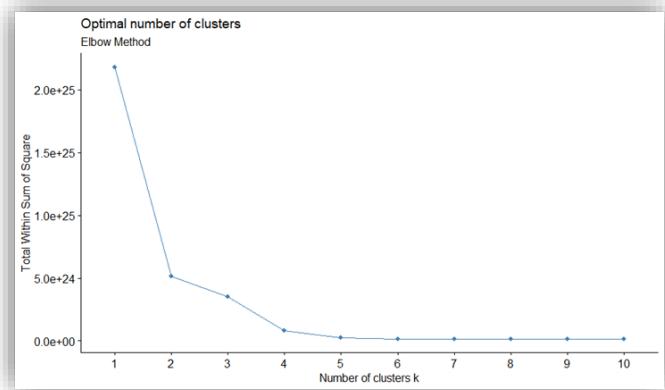
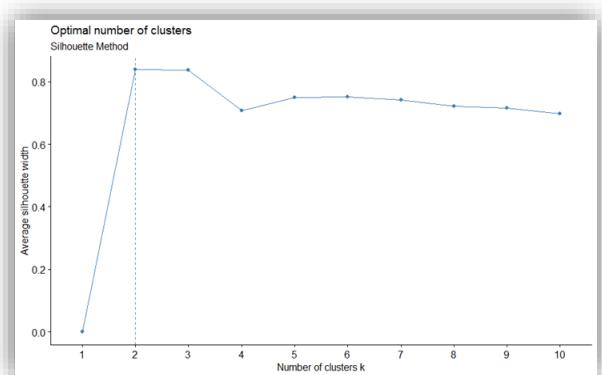
Selain itu, dilakukan uji *Kaiser-Mayer-Olkin* (KMO) untuk menilai kelayakan data dalam penerapan PCA, dengan hanya mempertahankan variabel yang memiliki nilai KMO diatas 0,5. Berdasarkan hasil pengujian (Gambar 6), seluruh variabel memiliki nilai KMO lebih dari 0,5, sehingga tidak ada variabel yang dieliminasi. Dengan demikian, seluruh variabel dapat dilanjutkan ke tahap PCA untuk mengatasi masalah multikolinearitas yang teridentifikasi sebelumnya.

**Gambar 7 Hasil PCA**

Hasil dari PCA ditampilkan dalam screeplot (Gambar 7), pada screeplot terlihat kurva melandai pada komponen utama yang kedua. Maka akan digunakan dua komponen utama dalam analisis klaster dengan menggunakan metode K-Means. Kedua komponen utama tersebut sudah mewakili semua atribut yang digunakan dalam penelitian.

C. Menentukan jumlah klaster

Sebelum penerapan K-Means clustering, dilakukan terlebih dahulu analisis menggunakan metode elbow untuk menentukan jumlah klaster optimal. Berdasarkan hasil yang ditampilkan pada grafik (Gambar 8), terbentuk titik menyerupai siku pada jumlah klaster ke-2, dimana penurunan SSE mulai melambat.

**Gambar 8 Grafik elbow****Gambar 9 Grafik Silhouette**

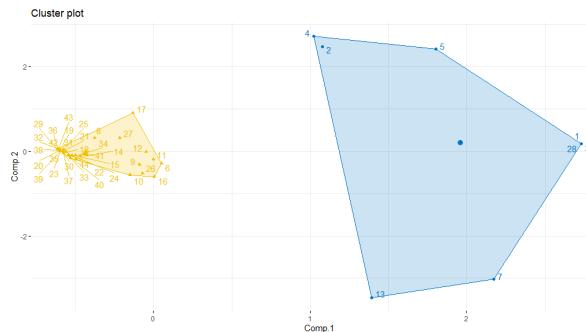
Selain metode elbow, analisis juga dilakukan menggunakan metode Silhouette untuk menentukan jumlah klaster yang optimal. Berdasarkan grafik yang ditampilkan (Gambar 9), nilai Silhouette score tertinggi diperoleh pada jumlah klaster ke-2. Karena dua metode *Elbow* dan *Silhouette* konstan menunjukkan 2 klaster, maka penelitian ini menggunakan 2 klaster.

D. Penerapan K-Means

Tabel 2 Hasil PCA KMeans

N o	Negara	Comp.1	Comp.2	Cluste r
1	China	5.091606247	0.10168403	1
2	India	2.011218029	1.43998682	6

3	Singapura	5.091606247	0.10168403	1		24	Denmark	-1.010153404	-	2	
4	Malaysia	1.904897915	1.58423230	1	7			0.04474429			
5	Belanda	3.36558216	1.40528052	1	9			2			
6	Inggris	0.095614487	-	2	0.16748544		25	Turki	-1.070084355	0.02595574	
					2			9		2	
7	Jepang	4.053602179	-	1	1.75723791		26	Taiwan	-0.130756193	-	
					3			0.30534604		2	
8	Belgia	-0.701254147	0.18520153	2	2			1			
9	British Virgin Islands	-0.165390737	-	2	0.18086428		27	Swiss	-0.400378746	0.18662668	
					2			5		2	
10	Thailand	-0.278475336	-	2	0.32310141		28	Hongkong	5.091606247	0.10168403	
					9			1			
11	Australia	-0.001526768	-	2	0.11242788		29	Timor Leste	-1.144561742	0.03176849	
					6			2			
12	Jerman	-0.087340642	-	2	0.01085377		30	Rusia	-1.073090211	-	
								0.02377979		2	
13	Korea Selatan	2.597465546	-	1	2.01481985		31	Spaniol	-1.083627682	0.00966616	
					5			5		2	
14	American Samoa	-0.798068071	-	2	0.05617919		32	Swaziland	-1.144671522	0.03157188	
					2			8		2	
15	Korea Utara	-0.969384769	-	2	0.05436726		33	Pakistan	-0.987173244	-	
								0.09313897		2	
16	Amerika Serikat	0.010660966	-	2	0.34982138		34	Rumania	-0.92626715	-	
					9			0.05651090		2	
17	Cayman Islands	-0.243309019	0.52715278	2	7		35	Austria	-1.110032557	0.00321248	
								6		2	
18	Prancis	-1.042436432	-	2	0.01078677		36	Eswatini	-1.13546463	0.02546190	
					3			9		2	
19	Yaman	-1.114832017	0.01721556	2	7		37	Maroko	-1.073439089	-	
								0.02302563		2	
20	Luxembour g	-1.116903231	0.01350630	2	8		38	Bulgaria	-1.12506991	0.01734229	
								2		2	
21	Swedia	-1.078289641	-	2	0.00030537		39	Samoa Barat	-1.110475024	0.00242008	
					4			8		2	
22	Italia	-0.922705232	-	2	0.10732651		40	Selandia Baru	-1.01330984	-	
					9			0.05718832		2	
23	Panama	-1.109911369	0.00205055	2	7		41	Uni Emirat Arab	-0.87210391	-	
								0.06262474		2	
								8			
							42	Vietnam	-1.144508856	0.0318632	2
							43	Finlandia	-1.133873021	0.02831226	2
							44	Nepal	-0.994991524	-	2
								0.06194393		2	



Gambar 10 Visualisasi hasil clustering

Berdasarkan hasil klastering menggunakan metode K-Means, diperoleh dua klaster. Pertama terdiri dari negara-negara dengan nilai investasi yang relatif rendah. Hal ini juga sejalan dengan jumlah proyek, jumlah tenaga kerja Indonesia maupun tenaga kerja asing yang cenderung sedikit. Negara-negara yang termasuk dalam klaster ini antara lain India, Inggris, Jepang, Belgia, dan beberapa negara lainnya yang dapat dilihat dalam Tabel 2.

Sementara itu, klaster kedua terdiri dari negara-negara dengan nilai investasi yang tinggi. Hal ini juga sejalan dengan jumlah proyek, jumlah tenaga kerja Indonesia maupun tenaga kerja asing yang cenderung banyak. Negara-negara yang tergolong dalam klaster investasi tinggi ini antara lain China, Singapura, Malaysia, Belanda, dan Hongkong.

E. Validasi K-Means Clustering

Untuk memastikan hasil klastering yang diperoleh sudah optimal dan dapat diinterpretasikan dengan baik, dilakukan proses validasi terhadap hasil penerapan metode K-Means. Validasi ini bertujuan untuk mengevaluasi sejauh mana klaster yang terbentuk memiliki pemisahan yang jelas antar klaster dan kemiripan dalam masing-masing klaster. Metode validasi yang digunakan dalam penelitian ini adalah *Davies-Bouldin Index* (DBI). Hasilnya menunjukkan bahwa nilai DBI sebesar 0,5084009. Nilai ini mengindikasikan bahwa kualitas klastering cukup baik. Mengingat semakin kecil nilai DBI, maka semakin baik pula pemisahan antar klaster dan konsistensi dalam klaster tersebut. Oleh karena itu, pembagian negara ke dalam dua klaster berdasarkan nilai investasi, jumlah proyek, jumlah tenaga kerja Indonesia, dan jumlah tenaga kerja asing dapat dinyatakan valid dan representatif terhadap pola dalam data.

V. KESIMPULAN DAN SARAN

Berdasarkan data realisasi investasi di Kota Surabaya tahun 2019-2024, dilakukan analisis klaster menggunakan metode K-Means. Hasil analisis menunjukkan bahwa data dapat dikelompokkan ke dalam dua klaster, yaitu klaster satu berisi negara-negara dengan nilai investasi rendah, yang juga ditandai dengan jumlah proyek, jumlah tenaga kerja Indonesia dan tenaga kerja asing yang relatif sedikit. Sementara itu, klaster dua

menunjukkan negara-negara dengan nilai investasi tinggi, dengan jumlah proyek, jumlah tenaga kerja Indonesia dan tenaga kerja asing yang lebih besar dibandingkan klaster pertama. Pengelompokan ini memberikan gambaran mengenai pola kontribusi investasi asing terhadap perekonomian lokal di Surabaya, khususnya dari sisi jumlah investasi dan distribusi tenaga kerja. Selain itu, hasil validasi menggunakan metode *Davies-Bouldin Index* (DBI) menunjukkan nilai sebesar 0,5084009, yang menunjukkan bahwa klastering cukup baik. Dengan demikian, metode K-Means terbukti efektif dalam mengelompokkan negara investor berdasarkan karakteristik nilai investasi, jumlah proyek dan tenaga kerja

UCAPAN TERIMA KASIH

Peneliti mengucapkan terima kasih kepada Dinas Penanaman Modal dan Pelayanan Terpadu Satu Pintu (DPMPTSP) Surabaya yang telah berkenan memberikan data sehingga memungkinkan penelitian ini dapat terlaksana. Ucapan terimakasih juga disampaikan kepada Ibu Dian Yuliati, M.Si selaku dosen pembimbing atas arahan, masukan, dan bimbingannya selama proses penyusunan penelitian ini. Penelitian ini merupakan bagian dari proses pembelajaran dan pengembangan kemampuan analisis data, khususnya dalam penerapan metode clustering. Semoga hasil dari penelitian ini dapat menjadi referensi awal untuk pengembangan riset di masa datang.

REFERENSI

- [1] H. Kampono, "Pengaruh Investasi Asing dan Investasi Dalam Negeri terhadap Pertumbuhan Ekonomi Indonesia Elyzabet Indrawati Marpaung," *J. Akunt.*, vol. 12, no. 1, pp. 137–145, 2020, [Online]. Available: <http://journal.maranatha.edu>
- [2] D. A. Sinulingga, "Bentuk Penanaman Modal Asing Bangkok Bank Di Indonesia Berdasarkan Hukum Penanaman Modal," *J. Huk. Pembang.*, vol. 52, no. 3, p. 10, 2022, doi: 10.21143/jhp.vol52.no3.3370.
- [3] A. Aziz and S. Zakir, "Evaluasi Kebijakan One Day Service (ODS) dalam Meningkatkan Pelayanan Perizinan di Kota Surabaya," vol. 2, no. 3, pp. 1030–1037, 2022.
- [4] A. Wulansari, M. Efendi, S. Anisa, and N. Faujiah, "Implementasi Akad Murabahah dan Akad Mudharabah Terhadap Profitabilitas (Studi Komparasi pada Bank Mega Syariah Surabaya dan Bank KB Bukopin Syariah Surabaya)," vol. 3, no. 1, pp. 429–438, 2025.
- [5] M. Siahaan, "Data Mining Strategi Pembangunan Infrastruktur Menggunakan Algoritma K-Means," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 11, no. 3, pp. 316–324, 2022, doi: 10.32736/sisfokom.v11i3.1453.
- [6] W. T. A. Shabila Oktavia, "Analisis Segmentasi Pelanggan Pembiayaan Berdasarkan Demografi untuk Memprediksi Tingkat Kredit Menggunakan Algoritma K-Means," vol. 7, pp. 394–402, 2025.
- [7] P. Data, S. Kerja, and A. Nasional, "Analisis k-means clustering angkatan kerja berdasarkan jenis kelamin, usia, pendidikan, dan pendapatan pada data satuan kerja angkatan nasional tahun 2021," vol. 2, no. 1, pp. 942–951, 2025.
- [8] & M. Y. Adriansyah Hidayat, "Strategi Dinas Penanaman Modal dan Pelayanan Terpadu Satu Pintu dalam Meningkatkan Investasi di Provinsi Sumatera Utara," *J. Islam. Econ. Financ.*, vol. 1, no. 4, 2023.
- [9] T. Zabilla Buciarda, W. Priana, and M. Wahed, "Analisis Pengaruh PMA, PMDN dan Konsumsi Rumah Tangga terhadap Pertumbuhan Ekonomi Di Kota Surabaya," *J. Syntax Admiration*, vol. 2, no. 6, pp. 1176–1190, 2021, doi: 10.46799/jsa.v2i6.244.

- [10] J. K. Putri, "Peran Penanaman Modal Asing Dalam Membangun Perekonomian Di Indonesia," *J. Soc. Res.*, vol. 1, no. 3, pp. 201–212, 2022, doi: 10.55324/josr.v1i3.55.
- [11] E. Prikafais and G. V. Sinardi, "Studi Literatur Terhadap Kebijakan Investasi Berbasis Lingkungan Hidup pada Pembangunan Kota Surabaya," *Fair Value J. Ilm. Akunt. dan Keuang.*, vol. 5, no. 5, pp. 2195–2200, 2022, doi: 10.32670/fairvalue.v5i5.2453.
- [12] Anugerah Christian Putra, "Pengaruh Pertumbuhan Pdrb Usaha Mikro Kecil Menengah (Umkm) Terhadap Pertumbuhan Ekonomi Surabaya," *J. Lemhannas RI*, vol. 10, no. 2, pp. 65–78, 2022, doi: 10.55960/jlri.v10i2.278.
- [13] F. P. Ferdy Pangestu, N. Y. Nur Yasin, R. C. Ronald Chistover Hasugian, and Y. Yunita, "Penerapan Algoritma K-Means Untuk Mengklasifikasi Data Obat," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 12, no. 1, pp. 53–62, 2023, doi: 10.32736/sisfokom.v12i1.1461.
- [14] A. Aditya Nugroho and P. Paduloh, "Analisis Clustering Kasus Covid – 19 di Indonesia Menggunakan Algoritma K-Means," *J. Eng. Environmental Energy Sci.*, vol. 2, no. 2, pp. 111–118, 2024, doi: 10.31599/eppng886.
- [15] D. Tohendry and D. Jollyta, "Penerapan Algoritma K-Means Clustering Untuk Pengelompokan Saham Berdasarkan Price Earning Ratio Dan Price To Book Value," *J. Mhs. Apl. Teknol. Komput. dan Inf.*, vol. 5, no. 1, pp. 3–9, 2023.
- [16] Y. Hartati, S. Defit, and G. W. Nurcahyo, "Klasterisasi Bibit Terbaik Menggunakan Algoritma K-Means dalam Meningkatkan Penjualan," *J. Inform. Ekon. Bisnis*, vol. 3, pp. 4–10, 2021, doi: 10.37034/infeb.v3i1.56.
- [17] D. N. Indahsari and P. Fitriandi, "Pengaruh Kebijakan Insentif Pajak Di Masa Pandemi Covid-19 Terhadap Penerimaan Ppn," *J. Pajak dan Keuang. Negara*, vol. 3, no. 1, pp. 24–36, 2021, doi: 10.31092/jpkn.v3i1.1202.
- [18] P. R. Sihombing, S. Suryadiningrat, D. A. Sunarjo, and Y. P. A. C. Yuda, "Identifikasi Data Outlier (Penculan) dan Kenormalan Data Pada Data Univariat serta Alternatif Penyelesaiannya," *J. Ekon. Dan Stat. Indones.*, vol. 2, no. 3, pp. 307–316, 2023, doi: 10.11594/jesi.02.03.07.
- [19] E. N. R. Khakim, A. Hermawan, and D. Avianto, "Implementasi Correlation Matrix Pada Klasifikasi Dataset Wine," *JKKO (Jurnal Informat. dan Komputer)*, vol. 7, no. 1, p. 158, 2023, doi: 10.26798/jkko.v7i1.771.
- [20] I. Permana and F. N. S. Salisah, "Pengaruh Normalisasi Data Terhadap Performa Hasil Klasifikasi Algoritma Backpropagation," *Indones. J. Informat. Res. Softw. Eng.*, vol. 2, no. 1, pp. 67–72, 2022, doi: 10.57152/ijirse.v2i1.311.
- [21] I. M. Karo Karo and H. Hendriyana, "Klasifikasi Penderita Diabetes menggunakan Algoritma Machine Learning dan Z-Score," *J. Teknol. Terpadu*, vol. 8, no. 2, pp. 94–99, 2022, doi: 10.54914/jtt.v8i2.564.
- [22] P. Di, K. Bojonegoro, D. Hediyati, and I. M. Suartana, "Penerapan Principal Component Analysis (PCA) Untuk Reduksi Dimensi Pada Proses Clustering Data Produksi," vol. 05, pp. 49–54, 2021.
- [23] H. W. S. Rina Yuliana Sari, Hardian Oktavianto, "Algoritma K-Means dengan Metode Elbow untuk Mengelompokkan Kabupaten/Kota Di Jawa Tengah Berdasarkan Komponen Pembentuk Indeks Pembangunan Manusia," vol. 3, no. 2, pp. 104–108, 2022.
- [24] N. T. Hartanti *et al.*, "Metode Elbow K-Means dalam Implementasi Data Mining pada Pemetaan Penyebaran Guru SMK," vol. 18, no. x, pp. 501–512, 2024.
- [25] E. Sulistiawan, A. Hapsery, and L. J. A. Arifahanum, "Perbandingan Metode Optimasi Untuk Pengelompokan Provinsi Berdasarkan Sektor Perikanan Di Indonesia (Studi Kasus Dinas Kelautan dan Perikanan Indonesia)," *J. Gaussian*, vol. 10, no. 1, pp. 76–84, 2021, doi: 10.14710/j.gauss.v10i1.30936.
- [26] H. Lailatul Ramadhania, L. Zakaria, and D. Nusyirwan, "Aplikasi Metode Sillhouette Coefficient, Metode Elbow dan Metode Gap Statictic dalam Menentukan K Optimal pada Analisis K-Medoids," *J. Siger Mat.*, vol. 04, no. 01, pp. 1–10, 2023.
- [27] M. Guntara and N. Lutfi, "Optimasi Cacah Klaster pada Klasterisasi dengan Algoritma KMeans Menggunakan Silhouette Coeficient dan Elbow Method," *JuTI "Jurnal Teknol. Informasi."*, vol. 2, no. 1, p. 43, 2023, doi: 10.26798/juti.v2i1.944.
- [28] V. No and A. Hal, "Klasterisasi Supplier Berdasarkan Kinerja Menggunakan Algoritma K-," vol. 7, no. 2, pp. 334–341, 2025.
- [29] F. Salsabila, T. Ridwan, and H. H, "Analisa Volume Penyebaran Sampah Di Karawang Menggunakan Algoritma K-Means Clustering," *J. Informat. dan Tek. Elektro Terap.*, vol. 12, no. 2, 2024, doi: 10.23960/jitet.v12i2.4226.
- [30] S. Kurniawan, A. M. Siregar, and H. Y. Novita, "Penerapan Algoritma K-Means dan Fuzzy C-Means Dalam Mengelompokan Prestasi Siswa Berdasarkan Nilai Akademik," *Sci. Student J. Information, Technol. Sci.*, vol. IV, no. 1, pp. 73–81, 2023.
- [31] S. Di and S. Utara, "Optimasi Jumlah Cluster Metode K-Medoids," no. July, 2022, doi: 10.20527/klik.v6i3.