PESHUM : Jurnal Pendidikan, Sosial dan Humaniora
Vol.4, No.6, Oktober 2025

Analisis Klaster Data Ekonomi dengan Algoritma *K-Means* pada Beragam *Dataset* Nasional dan Internasional

Authira Dwi Khairunnisa[1], Muhamad Faizal Muzaki[2], Muhammad Fiqri Hardiansyah[3], Shifa Azzahra Ramadhanti[4], Siti Syahla[5], Zurnan Alfian[6]
[1,2,3,4,5,6]Universitas Pamulang, Indonesia
E-mail: authiradwik@gmail.com

Article History:
Received: 26 Juni 2025
Revised: 08 September 2025
Accepted: 23 September 2025

Keywords: *K-Means, Clustering, Data Mining, Silhouette Score, Economy.*

Abstract: *Penelitian ini mengkaji penerapan algoritma pengelompokan K-Means untuk menganalisis lima set data terkait isu ekonomi dan sosial dari berbagai domain dan periode waktu. Data yang digunakan meliputi skor GCG negara-negara Asia, realisasi investasi asing di Indonesia, indeks korupsi global, tingkat kesiapan e-Government, serta data pertumbuhan ekonomi dan tingkat pengangguran di Indonesia. Penelitian ini bertujuan untuk mengidentifikasi pola atau klaster yang dapat memberikan wawasan dalam pengambilan keputusan. Evaluasi kinerja model dilakukan dengan mengukur dua metrik, yaitu Silhouette Score dan Inertia. Hasil analisis menunjukkan bahwa algoritma K-Means dapat menghasilkan klaster yang informatif, yang berpotensi digunakan untuk mendukung kebijakan dan keputusan strategis. Skor evaluasi terbaik diperoleh pada set data terkait investasi asing, dengan Silhouette Score sebesar 0,671, sementara pengelompokan terlemah ditemukan pada data pertumbuhan ekonomi dan pengangguran dengan Silhouette Score yang lebih rendah, yaitu sebesar 0,291. Hasil ini memberikan pemahaman yang lebih mendalam tentang kekuatan dan keterbatasan K-Means dalam berbagai konteks ekonomi dan sosial.*

PENDAHULUAN

Dalam era transformasi digital yang semakin berkembang, data ekonomi telah menjadi salah satu aset yang berharga (Suryawijaya, 2023). Pemerintah, perusahaan, dan organisasi internasional kini memiliki akses ke volume data yang masif, yang mencakup berbagai aspek kehidupan ekonomi dan sosial. Pengolahan dan analisis data besar ini memerlukan teknik-teknik yang canggih untuk mengungkap pola-pola tersembunyi yang dapat membantu dalam proses pengambilan keputusan. Salah satu pendekatan yang paling banyak digunakan dalam analisis data besar adalah metode *clustering,* yang termasuk dalam kategori *unsupervised learning* (Langgeni dkk., 2010). Metode ini memungkinkan identifikasi grup atau klaster dalam data yang memiliki kesamaan karakteristik, tanpa memerlukan label atau informasi eksplisit tentang kategori data-data tersebut. Dalam konteks ini, algoritma *clustering* memberikan kesempatan untuk memahami hubungan yang ada antara variabel-variabel yang tidak langsung terlihat (Alam dkk., 2025).

Algoritma *K-Means* merupakan salah satu metode *clustering* yang paling populer karena

kesederhanaannya serta kemampuannya dalam menangani data berskala besar secara efisien (Adhitama dkk., 2020). Keunggulan utama dari *K-Means* terletak pada kemampuannya untuk membagi data ke dalam beberapa kelompok berdasarkan jarak minimal antara titik data dengan pusat klaster, yang dapat dengan cepat mengidentifikasi pola dalam *dataset* yang besar. Oleh karena itu, *K-Means* sering digunakan dalam berbagai aplikasi, seperti analisis pasar, segmentasi pelanggan, dan analisis perilaku sosial-ekonomi (Perdana dkk., 2022). Meskipun sederhana, algoritma ini memiliki potensi yang besar dalam mengungkap wawasan yang bermanfaat dari data yang kompleks. Namun, seperti halnya dengan teknik analisis data lainnya, *K-Means* juga memiliki bentuk keterbatasan, seperti kepekaannya terhadap pemilihan nilai awal klaster dan kebutuhan akan jumlah klaster yang sudah ditentukan sebelumnya.

Penelitian ini yang berfokus pada evaluasi efektivitas algoritma *K-Means* dalam proses mengelompokkan data ekonomi yang berasal dari lima sumber yang berbeda. Data yang dianalisis mencakup berbagai indikator ekonomi dan sosial, seperti skor GCG negara-negara Asia, realisasi investasi asing di Indonesia, indeks korupsi global, kesiapan *e-Government*, serta data-data pertumbuhan ekonomi dan tingkat pengangguran Indonesia. Penelitian ini yang bertujuan untuk menunjukkan bagaimana K-Means dapat diterapkan untuk data yang berasal dari domain yang berbeda dan memiliki karakteristik yang sangat beragam. Dengan mengukur performa algoritma menggunakan metrik evaluasi seperti *Silhouette Score* dan *Inertia,* studi ini juga akan mampu mengidentifikasi variasi dalam efektivitas K-Means dalam mengelompokkan data sesuai dengan jenis dan kompleksitasnya. Selain itu, penelitian ini akan mengungkap potensi wawasan yang dapat diperoleh dari hasil klasterisasi yang dihasilkan, yang dapat membantu dalam pengambilan kebijakan atau keputusan ekonomi yang lebih baik.

LANDASAN TEORI

Data ekonomi dan sosial yang tersedia saat ini berjumlah sangat besar, mencakup berbagai aspek kehidupan yang dapat memberikan wawasan penting bagi pengambilan kebijakan dan keputusan bisnis. Untuk mengelola dan menganalisis data sebesar ini, diperlukan metode analisis yang efektif dan efisien. Salah satu pendekatan yang banyak digunakan dalam analisis data besar adalah *clustering*, yang merupakan bagian dari teknik *unsupervised learning* (Awalina & Rahayu, 2023). Dalam *clustering,* data dikelompokkan ke dalam grup atau klaster berdasarkan kesamaan karakteristik yang dimilikinya, tanpa adanya label atau informasi yang jelas mengenai kategori data. Proses ini yang memungkinkan peneliti atau analis untuk menemukan pola tersembunyi yang dapat memberikan pemahaman lebih mendalam mengenai hubungan antar variabel dalam data tersebut. Sebagai contoh, dalam data ekonomi, *clustering* dapat mengidentifikasi kelompok negara dengan kesamaan dalam pertumbuhan ekonomi, korupsi, atau tingkat pengangguran.

Algoritma *K-Means* merupakan salah satu metode *clustering* yang paling banyak digunakan karena kesederhanaannya dan kemampuannya dalam menangani data-data besar dengan efisien (Apriyani dkk., 2023). *K-Means* bekerja dengan membagi data menjadi k klaster, yang ditentukan sebelumnya, dengan cara meminimalkan jarak antara titik data dan pusat klaster. Proses ini dilakukan dalam beberapa iterasi sampai pusat klaster stabil. Salah satu keuntungan utama dari *K-Means* adalah kemampuannya untuk bekerja dengan cepat pada *dataset* yang besar, menjadikannya pilihan utama eksplorasi data awal. Namun, algoritma ini juga memiliki beberapa keterbatasan, seperti kepekaannya terhadap pemilihan nilai k yang tepat, serta kerentanannya terhadap *outliers* dan titik data yang tidak terdistribusi secara merata. Oleh karena itu, pemilihan nilai k dan evaluasi hasil klasterisasi menjadi aspek yang sangat penting dalam penerapan K-Means.

Dalam konteks aplikasi analisis data ekonomi, *K-Means* diterapkan untuk mengidentifikasi pola dalam data yang bersifat multidimensional dan beragam (Zhafar dkk., 2023). Data ekonomi yang kompleks, seperti skor GCG *(Good Corporate Governance),* realisasi investasi asing, indeks korupsi, kesiapan *e-Government*, dan data pertumbuhan ekonomi, seringkali memiliki hubungan yang tidak langsung antara variabel-variabelnya. Dengan menggunakan *K-Means*, berbagai negara atau wilayah dapat dikelompokkan berdasarkan pada kesamaan karakteristik ekonomi dan sosial mereka, yang membantu pengambil kebijakan dalam merancang strategi yang lebih tepat. Metrik evaluasi, seperti *Silhouette Score* dan *Inertia,* sering digunakan untuk mampu menilai kualitas klaster yang dihasilkan (Wijaya dkk., 2023; Hendrastuty, 2024). *Silhouette Score* yang mengukur seberapa baik tiap objek data dikelompokkan, sedangkan *Inertia* mengukur sejauh mana objek data berada dalam klaster yang sesuai. Kedua metrik ini membantu memastikan bahwa klaster yang dihasilkan oleh K-Means adalah informatif dan bermakna dalam konteks analisis data ekonomi.

METODE PENELITIAN

Metode penelitian ini yang melibatkan penggunaan lima *dataset* yang mencakup berbagai karakteristik dan *domain,* yang bertujuan untuk mengeksplorasi penerapan algoritma *K-Means* dalam analisis data ekonomi dan sosial. *Dataset* yang digunakan antara lain adalah GCG *Rating* dari 11 negara Asia pada tahun 2001, perkembangan investasi asing di Indonesia dari tahun 1985 hingga 2000, indeks korupsi global pada tahun 1995, global *e-Government readiness* dari tahun 2004 hingga 2008, serta data pertumbuhan ekonomi dan tingkat pengangguran di Indonesia dari tahun 1996 hingga 2005. Setiap *dataset* ini mewakili topik yang berbeda, tetapi semuanya memiliki potensi untuk dianalisis lebih mendalam menggunakan metode *clustering,* dengan tujuan untuk mengidentifikasi pola-pola yang mungkin tersembunyi dalam data.

Setiap *dataset* tersebut melalui serangkaian tahapan analisis yang dimulai dengan tahap *pre-processing*, yang bertujuan untuk mempersiapkan data sebelum dilakukan klasterisasi. Pada tahap ini, data dinormalisasi untuk memastikan keseragaman antar variabel yang memiliki skala berbeda dan juga dilakukan penanganan terhadap nilai-nilai yang hilang *(missing values)* agar kualitas data tetap terjaga. Setelah itu, algoritma K-Means diterapkan pada setiap dataset dengan jumlah klaster (K) yang telah ditetapkan sebesar tiga, dengan harapan bahwa hasil dari klasterisasi ini akan memberikan wawasan bermakna terkait karakteristik yang ada pada masing-masing kelompok data. Tahap berikutnya adalah evaluasi, di mana performa klasterisasi yang diukur menggunakan dua metrik utama, yaitu *Silhouette Score* dan *Inertia. Silhouette Score* digunakan mengevaluasi seberapa baik suatu objek data dikelompokkan dalam klaster yang sesuai, sementara *Inertia* mengukur seberapa erat objek data berkelompok dalam klaster yang sama, dengan nilai yang lebih rendah menunjukkan klaster yang lebih terorganisir dan padat.

Selain itu, perangkat lunak yang digunakan dalam penelitian ini melibatkan beberapa pustaka dan modul penting untuk memfasilitasi proses komputasi dan analisis data. Kode yang digunakan mengimpor pustaka seperti *Pandas* untuk pengelolaan dan manipulasi data, *KMeans* dari *Scikit-learn* untuk klasterisasi, *Matplotlib* untuk visualisasi hasil analisis, serta *StandardScaler* dari *Scikit-learn* untuk normalisasi data. Kode ini yang mencerminkan fondasi pemrograman yang diperlukan dalam proses analisis data, pemodelan menggunakan machine learning, serta visualisasi hasil klasterisasi yang telah dilakukan. Dengan menggunakan alat dan teknik ini, maka penelitian dapat mengidentifikasi dan mengevaluasi klaster-klaster yang dihasilkan, memberikan wawasan yang lebih mendalam mengenai hubungan antar variabel dalam data ekonomi dan sosial yang dianalisis, seperti proses klasterisasi seperti Gambar 1. berikut.

```
1   import pandas as pd
2   from sklearn.cluster import KMeans
3   import matplotlib.pyplot as plt
4   from sklearn.preprocessing import StandardScaler
```

Gambar 1. Potongan Kode dalam Proses Klasterisasi

HASIL DAN PEMBAHASAN

1. Evaluasi Hasil Analisis

ID	Nama Mahasiswa	*Dataset*	*K-Means*	*Silhouette Score*	*Inertia*	Catatan
1.	Authira Dwi Khairunnisa	GCG *Rating* 11 Negara Asia (2001)	3	0.310	30.04	Mengelompokkan negara berdasar skor GCG: unggul, sedang, lemah
2.	Siti Syahla	Investasi Asing Indonesia (1985–2000)	3	0.671	1,051,346.94	Periode pertumbuhan investasi rendah, sedang, tinggi
3.	Shifa Azzahra Ramadhanti	Indeks Korupsi Dunia (1995)	3	0.608	10.94	Klaster skor korupsi rendah, sedang, tinggi
4.	Muhammad Fiqri Hardiansyah	*E-Government Readiness* (2004–2008)	3	0.539	7335.25	Negara diklasterkan berdasar kesiapan *E-Gov*
5.	Muhamad Faizal Muzaki	Ekonomi & Pengangguran Indonesia (1996–2005)	3	0.291	6.70	Mengelompokkan tahun berdasar tren ekonomi dan pengangguran

Berdasarkan data tabel ini yang menunjukkan hasil analisis *clustering* menggunakan metode *K-Means* yang dilakukan oleh lima mahasiswa dengan menggunakan berbagai *dataset* dan masing-masing mengelompokkan data ke dalam tiga klaster. Setiap mahasiswa yang mengevaluasi hasil klasterisasi dengan dua metrik, yaitu *Silhouette Score* dan *Inertia,* untuk menilai kualitas klaster yang terbentuk. Authira Dwi Khairunnisa menggunakan *dataset* GCG Rating dari 11 negara Asia pada tahun 2001. Hasil *clustering*-nya yang menunjukkan *Silhouette Score* sebesar 0.310, yang menunjukkan kualitas pemisahan klaster yang relatif rendah. *Inertia* yang dihasilkan adalah 30.04, yang menunjukkan bahwa klaster-klaster tersebut memiliki jarak antar data yang tidak terlalu kecil. Proyek ini mengelompokkan negara berdasarkan skor GCG ke dalam tiga kategori: unggul, sedang, dan lemah.

Siti Syahla menggunakan *dataset I*nvestasi Asing di Indonesia dari tahun 1985 hingga 2000. Hasil *clustering*-nya menunjukkan *Silhouette Score* tertinggi di antara semua proyek, yaitu 0.671, yang menandakan pemisahan klaster yang sangat baik. *Inertia* sebesar 1,051,346.94 menunjukkan bahwa skala data investasi yang besar mempengaruhi jarak antar data dalam klaster. Proyek ini mengelompokkan tahun-tahun berdasarkan periode pertumbuhan investasi menjadi rendah, sedang, dan tinggi. Untuk. Shifa Azzahra Ramadhanti menggunakan *dataset* Indeks

Korupsi Dunia tahun 1995. Clustering yang dilakukan menghasilkan *Silhouette Score* sebesar 0.608, menunjukkan kualitas klaster yang baik. *Inertia* yang dihasilkan adalah 10.94, yang cukup kecil, menunjukkan penyebaran data relatif terpusat. Dalam proyek ini, negara-negara dikelompokkan berdasarkan skor indeks korupsi menjadi tiga kategori: rendah, sedang, dan tinggi.

Muhammad Fiqri Hardiansyah menggunakan *dataset E-Government Readiness* dari tahun 2004 hingga 2008. *Silhouette Score* diperoleh adalah 0.539, yang menunjukkan kualitas pemisahan klaster yang cukup baik. *Inertia* yang dihasilkan sebesar 7335.25 menunjukkan bahwa distribusi data antar klaster relatif lebih besar. Negara-negara dalam *dataset* ini dikelompokkan berdasarkan tingkat kesiapan dalam implementasi *e-government.* Muhamad Faizal Muzaki yang menggunakan *dataset* yang berkaitan dengan Ekonomi dan Pengangguran Indonesia dari tahun 1996 hingga 2005. Hasil analisisnya menunjukkan *Silhouette Score* terendah, yaitu 0.291, yang menunjukkan bahwa pemisahan klaster kurang optimal. *Inertia* sebesar 6.70 menunjukkan bahwa data dalam klaster sangat terpusat, mungkin karena tren ekonomi dan pengangguran yang lebih konsisten. Proyek ini mengelompokkan tahun-tahun berdasarkan tren ekonomi dan pengangguran.

Kualitas klasterisasi yang dihasilkan mahasiswa ini sangat bervariasi, dengan Siti Syahla memperoleh hasil terbaik, sementara Muhamad Faizal Muzaki mendapatkan hasil yang paling lemah. Variasi *Inertia* dan *Silhouette Score* menggambarkan perbedaan dalam skala data, jumlah data, dan tingkat keberhasilan dalam pemisahan klaster.

2. Interpretasi Hasil

Hasil klasterisasi menunjukkan bahwa *dataset* terkait investasi asing memberikan performa terbaik, dengan *Silhouette Score* sebesar 0.671. Nilai ini menunjukkan bahwa klaster-klaster yang terbentuk sangat terpisah dan padat, mencerminkan adanya perbedaan yang jelas antara periode ekonomi yang tercatat dalam data investasi. Klasterisasi yang baik ini mencerminkan bahwa K-Means berhasil mengidentifikasi kelompok data yang memiliki karakteristik yang sangat berbeda, yang memungkinkan pengambilan keputusan yang lebih terarah dan berbasis data. Hal ini yang mengindikasikan bahwa data investasi asing memiliki pola yang lebih konsisten dan mudah untuk dikelompokkan, memberikan gambaran yang jelas mengenai fluktuasi investasi asing di Indonesia dalam rentang waktu tersebut.

Di sisi lain, *dataset* yang berkaitan dengan indeks korupsi global dan kesiapan *e-Government* memiliki hasil evaluasi yang moderat, dengan *Silhouette Score* masing-masing sebesar 0.608 dan 0.539. Meskipun klaster-klaster yang terbentuk cukup padat dan terpisah, perbedaan antar klaster tidak sejelas pada data investasi asing, yang mengindikasikan bahwa data ini memiliki sedikit kerumitan dalam pemisahan yang lebih tajam. Keterbatasan ini mungkin disebabkan oleh adanya faktor eksternal atau variabel lain yang mempengaruhi data, yang membuat perbedaan antar kelompok kurang signifikan. Sebaliknya, *dataset* yang berkaitan dengan GCG dan data ekonomi-pengangguran menunjukkan performa klasterisasi yang lebih rendah, dengan *Silhouette Score* 0.310 dan 0.291. Hal ini menunjukkan bahwa *K-Means* kesulitan dalam menemukan pemisahan klaster yang jelas, kemungkinan besar karena distribusi data yang tidak homogen atau pola yang kurang terdefinisi dengan baik. Keterbatasan ini mencerminkan tantangan dalam menganalisis data dengan distribusi yang lebih kompleks atau kurang konsisten, yang membuat klasterisasi menjadi kurang efektif.

KESIMPULAN

Penelitian ini menunjukkan bahwa algoritma K-Means merupakan alat eksplorasi data yang fleksibel dan efektif dalam mengidentifikasi struktur tersembunyi dalam data ekonomi yang kompleks. Hasil klasterisasi yang diperoleh untuk mampu memberikan wawasan kontekstual

yang berharga, seperti bentuk pengelompokan berdasarkan kinerja investasi asing, tingkat korupsi, atau kesiapan teknologi di berbagai negara dan periode waktu. Meskipun demikian, efektivitas K-Means sangat bergantung pada karakteristik intrinsik dari *dataset* yang digunakan. *Dataset* dengan pola yang jelas, seperti data investasi asing, menghasilkan klaster yang lebih terpisah dan padat, sementara *dataset* dengan distribusi data yang lebih heterogen, seperti data GCG dan ekonomi-pengangguran, menunjukkan klasterisasi yang kurang jelas. Hal ini mengindikasikan bahwa K-Means lebih efektif digunakan pada *dataset* yang memiliki pola yang lebih homogen dan terdefinisi dengan baik. Di sisi lain, untuk *dataset* dengan pola yang kurang tajam atau distribusi data yang tidak homogen, algoritma K-Means mungkin tidak optimal dalam menghasilkan klaster yang tajam. Oleh karena itu, penelitian mendatang disarankan untuk mempertimbangkan penggunaan teknik *clustering* alternatif seperti DBSCAN atau *Gaussian Mixture Model*, yang dapat lebih efektif dalam menangani data dengan karakteristik yang lebih kompleks dan memberikan hasil klasterisasi yang lebih informatif dan tajam.

DAFTAR REFERENSI

Adhitama, R., Burhanuddin, A., & Ananda, R. (2020). Penentuan jumlah cluster ideal SMK di Jawa Tengah dengan Metode X-means clustering dan K-means clustering. *JIKO (Jurnal Informatika dan Komputer)*, *3*(1), 1-5. https://doi.org/10.33387/jiko.v3i1.1635

Alam, R. M., Hazriani, H., Arda, A. L., & Mardin, M. I. (2025). Identifikasi Pola Prilaku Belajar Mahasiswa Pada Platform Learning Management System Dengan Algoritma K-Means. *Jurnal JEETech*, *6*(1), 11-24. https://doi.org/10.32492/jeetech.v6i1.6102

Apriyani, P., Dikananda, A. R., & Ali, I. (2023). Penerapan algoritma K-Means dalam klasterisasi kasus stunting balita desa Tegalwangi. *Hello World Jurnal Ilmu Komputer*, *2*(1), 20-33. https://doi.org/10.56211/helloworld.v2i1.230

Awalina, E. F. L., & Rahayu, W. I. (2023). Optimalisasi strategi pemasaran dengan segmentasi pelanggan menggunakan penerapan K-means clustering pada transaksi online retail. *Jurnal Teknologi Dan Informasi*, *13*(2), 122-137. https://doi.org/10.34010/jati.v13i2.10090

Hendrastuty, N. (2024). Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Dalam Evaluasi Hasil Pembelajaran Siswa. *Jurnal Ilmiah Informatika dan Ilmu Komputer (JIMA-ILKOM)*, *3*(1), 46-56. https://doi.org/10.58602/jima-ilkom.v3i1.26

Langgeni, D. P., Baizal, Z. A., & AW, Y. F. (2010). Clustering Artikel Berita Berbahasa Indonesia Menggunakan Unsupervised Feature Selection. Dalam *Seminar Nasional Informatika (SEMNASIF)* (Vol. 1, No. 4). https://jurnal.upnyk.ac.id/index.php/semnasif/article/view/1175

Perdana, S. A., Florentin, S. F., & Santoso, A. (2022). Analisis Segmentasi Pelanggan Menggunakan K-Means Clustering Studi Kasus Aplikasi Alfagift. *Sebatik*, *26*(2), 446-457. https://doi.org/10.46984/sebatik.v26i2.1991

Suryawijaya, T. W. E. (2023). Memperkuat Keamanan Data melalui Teknologi Blockchain: Mengeksplorasi Implementasi Sukses dalam Transformasi Digital di Indonesia. *Jurnal Studi Kebijakan Publik*, *2*(1), 55-68. https://doi.org/10.21787/jskp.2.2023.55-68

Wijaya, M. A., Prayoga, D. S., Rahman, A. K., & Sari, A. P. (2023). Perbandingan Algoritma K-Means dan DBSCAN dalam Metode Clustering dengan PCA untuk Analisis Data Statistik Negara Dunia. In *Prosiding Seminar Nasional Informatika Bela Negara* (Vol. 3, pp. 63-70). https://santika.upnjatim.ac.id/submissions/index.php/santika/article/view/195

Zhafar, M. N., Usman, K., & Akhyar, F. (2023). Penerapan Metode Clustering Dengan Algoritma

K-Means Untuk Analisa Persebaran Varian Covid-19 (Studi Kasus Kelurahan Antapani Kidul). *eProceedings of Engineering*, *10*(5). https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/21215