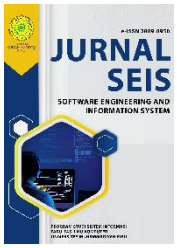
Jurnal Software Engineering and Information System (SEIS)

https://ejurnal.umri.ac.id/index.php/SEIS/index



OPTIMISASI ALGORITMA K-MEANS DENGAN METODE REDUKSI DIMENSI UNTUK PENGELOMPOKAN BIG DATA DALAM ARSITEKTUR CLOUD COMPUTING

Bayu Anugerah Putra[1], Harun Mukhtar[2], Elsi Titasari Br Bangun[3], Alris Gusnanda[4*], Adila Maisyarah[5], Muhammad Irgi Kurniawan[6], Raditya Pradipa[7], Zurrahman Muhammad Ali[8]
[1,2,3,4,5,6,7,8]Fakultas Ilmu Komputer, Universitas Muhammadiyah Riau
bayuanugerahputra@umri.ac.id[1], harunmukhtar@umri.ac.id[2], elsititasari@umri.ac.id[3], 220401001@student.umri.ac.id[4*], 220401118@student.umri.ac.id[5], 220401108@student.umri.ac.id[6], 220401091@student.umri.ac.id[7], 220401114@student.umri.ac.id[8]

Abstract

In the era of big data, data clustering becomes a major challenge due to the complexity and huge volume of data. The K-means algorithm is one of the clustering techniques that is often used due to its simplicity. However, K-means faces difficulties in handling high-dimensional and large-volume data. This study proposes an optimization of the K-means algorithm using the Principal Component Analysis (PCA) dimensionality reduction method to improve the efficiency and accuracy of big data clustering in cloud computing architecture. The KDD Cup 1999 dataset is used to test this method. The dataset undergoes pre-processing and dimensionality reduction using PCA, then K-means clustering is applied. The clustering results are evaluated using the Silhouette Score and Davies-Bouldin Index. The implementation is carried out in the Google Colab environment to utilize cloud computing resources. The results show that dimensionality reduction using PCA significantly reduces computational complexity and improves clustering quality. This method is effective in clustering big data, making it an efficient solution for data clustering in cloud computing architecture.

Keywords: *K-Means Optimization, Dimensionality Reduction, Principal Component Analysis (PCA), Big Data Clustering, Cloud Computing Architecture, KDD Cup 1999, Clustering Evaluation.*

Abstrak

Pada era big data, pengelompokan data menjadi tantangan utama karena kompleksitas dan volume data yang sangat besar. Algoritma K-means adalah salah satu teknik clustering yang sering digunakan karena kesederhanaannya. Namun, K-means menghadapi kesulitan dalam menangani data dengan dimensi tinggi dan volume besar. Penelitian ini mengusulkan optimisasi algoritma K-means menggunakan metode reduksi dimensi Principal Component Analysis (PCA) untuk meningkatkan efisiensi dan akurasi pengelompokan big data dalam arsitektur cloud computing. Dataset KDD Cup 1999 digunakan untuk menguji metode ini. Dataset tersebut mengalami pra-pemrosesan dan reduksi dimensi menggunakan PCA, kemudian diterapkan K-means clustering. Hasil pengelompokan dievaluasi menggunakan Silhouette Score dan Davies-Bouldin Index. Implementasi dilakukan di lingkungan Google Colab untuk memanfaatkan sumber daya komputasi cloud. Hasil penelitian menunjukkan bahwa reduksi dimensi menggunakan PCA secara signifikan mengurangi kompleksitas komputasi dan meningkatkan kualitas clustering. Metode ini efektif dalam mengelompokkan big data, menjadikannya solusi yang efisien untuk pengelompokan data dalam arsitektur cloud computing.

Kata Kunci: Optimisasi K-Means, Reduksi Dimensi, Principal Component Analysis (PCA), Pengelompokan Big Data, Arsitektur Cloud Computing, KDD Cup 1999, Evaluasi Clustering

PENDAHULUAN

Di era digital saat ini, big data menjadi salah satu tantangan utama dalam pengolahan informasi (Liu et al., 2023). Big data mencakup volume data yang sangat besar, kecepatan tinggi dalam pengumpulan, dan beragam jenis data yang berasal dari berbagai sumber seperti media sosial, sensor IoT, transaksi e-commerce, dan layanan cloud computing (Jin et al., 2024)(Gabrielli et al.,

2024). Untuk mengelola dan menganalisis data tersebut, diperlukan teknik yang efektif dan efisien. Salah satu metode yang populer untuk analisis data adalah pengelompokan (clustering), di mana algoritma K-means sering digunakan karena kesederhanaannya dan kemampuannya dalam mengelompokkan data ke dalam sejumlah cluster berdasarkan kemiripan (Askari, 2021).

Namun, ketika diterapkan pada data berdimensi tinggi dan berskala besar, algoritma K-means menghadapi tantangan yang signifikan. Masalah yang disebut "kutukan dimensi" (curse of dimensionality) menyebabkan penurunan performa algoritma karena meningkatnya kompleksitas komputasi dan menurunnya akurasi pengelompokan (Ran et al., 2021). Untuk mengatasi masalah ini, diperlukan teknik reduksi dimensi yang dapat mengurangi jumlah dimensi data tanpa kehilangan informasi penting. Teknik reduksi dimensi seperti Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), dan Linear Discriminant Analysis (LDA) dapat digunakan untuk tujuan ini (Tripathi & Singal, 2019), (Sarmina et al., 2023), (Li et al., 2020).

Selain itu, pengelompokan big data dalam arsitektur cloud computing menawarkan keuntungan dalam hal skalabilitas dan fleksibilitas sumber daya komputasi (Aceto et al., 2020). Platform cloud computing seperti Amazon Web Services (AWS), Google Cloud Platform (GCP), dan Microsoft Azure menyediakan lingkungan yang memungkinkan pemrosesan data besar secara paralel dan terdistribusi. Hal ini dapat mempercepat proses analisis dan pengelompokan, serta mengurangi biaya infrastruktur.(Neela et al., 2021), (Chen et al., 2023), (Poppe et al., 2024).

Penelitian ini bertujuan untuk mengoptimalkan algoritma K-means dengan metode reduksi dimensi untuk pengelompokan big data dalam arsitektur cloud computing. Kami akan mengevaluasi efektivitas dan efisiensi pendekatan yang diusulkan melalui eksperimen menggunakan dataset besar dengan berbagai karakteristik. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam meningkatkan performa pengelompokan data besar dan menawarkan solusi praktis yang dapat diterapkan dalam berbagai aplikasi seperti analisis pasar, deteksi penipuan, dan manajemen informasi.

RELATED WORK

Berbagai penelitian telah dilakukan untuk mengatasi tantangan yang dihadapi oleh algoritma K-means dalam menangani data berdimensi tinggi dan skala besar. Salah satu pendekatan yang sering digunakan adalah teknik reduksi dimensi (Ma & Yuan, 2019). Principal Component Analysis (PCA) adalah salah satu metode reduksi dimensi yang paling banyak digunakan(Minh et al., 2023). PCA bekerja dengan mentransformasikan data asli ke dalam ruang dimensi yang lebih rendah dengan memaksimalkan variansi (Anowar et al., 2021). Sebagai contoh, (Stahl et al., 2019) dalam penelitiannya menunjukkan bahwa penggunaan PCA sebelum pengelompokan K-means dapat meningkatkan akurasi dan mengurangi waktu komputasi.

Selain PCA, metode t-Distributed Stochastic Neighbor Embedding (t-SNE) juga populer dalam mengatasi masalah dimensi tinggi. t-SNE memproyeksikan data ke dalam ruang dua atau tiga dimensi dengan mempertahankan jarak probabilitas antar titik data. (Pezzotti et al., 2019) mengembangkan t-SNE dan menunjukkan efektivitasnya dalam visualisasi data berdimensi tinggi. Namun, t-SNE lebih cocok untuk visualisasi daripada pengelompokan karena komputasinya yang intensif.

Linear Discriminant Analysis (LDA) adalah metode lain yang digunakan untuk reduksi dimensi, terutama dalam konteks klasifikasi(Alahmadi et al., 2022). LDA mencari kombinasi linear dari fitur yang memaksimalkan separabilitas antar kelas. Penelitian oleh (Campbell et al., 2023) menunjukkan bahwa LDA dapat meningkatkan performa pengelompokan ketika digunakan bersama dengan K-means(Koehler et al., 2024)(Fabiyi et al., 2021).

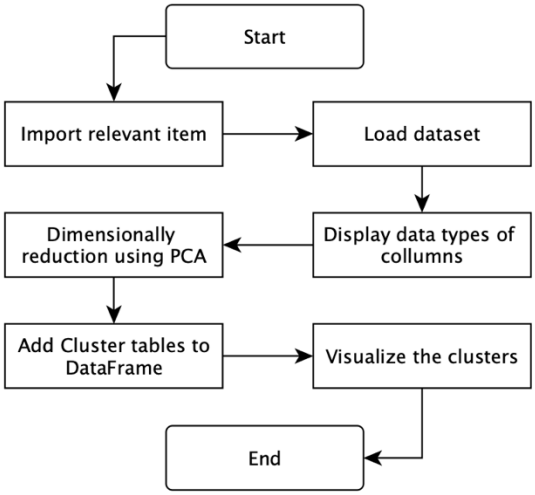
Di sisi lain, cloud computing telah menjadi platform yang penting untuk pemrosesan big data. Salah satu keunggulan utama cloud computing adalah kemampuannya untuk menskalakan sumber daya komputasi sesuai permintaan. (van der Vlist, 2022) dalam penelitiannya tentang MapReduce menunjukkan bagaimana pendekatan pemrosesan terdistribusi dapat digunakan untuk mengolah data dalam skala besar secara efisien. Hadoop dan Spark adalah contoh platform big data yang memanfaatkan konsep pemrosesan terdistribusi di cloud.

Penelitian lain oleh (Zaharia et al., 2019) memperkenalkan Apache Spark sebagai alternatif yang lebih cepat dari Hadoop untuk pemrosesan big data. Spark memungkinkan komputasi in-memory yang signifikan meningkatkan kecepatan pemrosesan data. Dalam konteks pengelompokan, Zaharia menunjukkan bahwa kombinasi Spark dengan algoritma K-means dapat mengatasi tantangan pengolahan data besar dengan lebih efektif.

Dalam penelitian ini, kami menggabungkan teknik reduksi dimensi dengan algoritma K-means dan mengimplementasikannya dalam arsitektur cloud computing untuk mengoptimalkan pengelompokan big data. Pendekatan ini bertujuan untuk mengatasi keterbatasan algoritma K-means dan memanfaatkan keunggulan cloud computing dalam pengolahan data besar.

METODOLOGI PENELITIAN

Penelitian ini menggunakan metode eksperimental yang melibatkan beberapa tahapan untuk mengoptimalkan algoritma K-means dengan menggunakan metode reduksi dimensi PCA dalam pengelompokan data besar di arsitektur cloud computing(Zhang et al., 2023).



Gambar 1. Skema Penelitian

Tahapan-tahapan dalam metodologi penelitian ini adalah sebagai berikut:

Pengumpulan Data: Dataset yang digunakan adalah KDD Cup 1999 yang diperoleh dari UCI Machine Learning Repository. Dataset ini terdiri dari sekitar 500.000 baris data dengan 41 fitur, yang mencakup fitur numerik dan kategorikal. Data ini dipilih karena ukurannya yang besar dan relevansinya dalam analisis keamanan jaringan(Roh et al., 2021).

Pra-pemrosesan Data: Langkah pertama adalah mengidentifikasi dan mengonversi fitur kategorikal menjadi format numerik menggunakan teknik one-hot encoding. Setelah itu, semua fitur dalam dataset distandarisasi menggunakan StandardScaler untuk memastikan skala yang konsisten antar fitur. Proses ini penting untuk menghindari bias dalam algoritma pengelompokan(Aldi et al., 2023).

Reduksi Dimensi: Principal Component Analysis (PCA) diterapkan untuk mengurangi jumlah fitur dari 41 menjadi 2 komponen utama. PCA dipilih karena kemampuannya untuk mempertahankan variabilitas data yang signifikan sambil mengurangi dimensi, sehingga memungkinkan pengolahan data yang lebih efisien dan peningkatan interpretabilitas hasil pengelompokan.

Penerapan Algoritma K-means: Algoritma K-means diterapkan pada data yang telah direduksi dimensinya. Algoritma ini dimulai dengan inisialisasi k centroid secara acak dan melakukan iterasi untuk memperbarui posisi centroid hingga konvergensi tercapai(Subedi et al., 2019). Dalam penelitian ini, jumlah cluster (k) dipilih berdasarkan eksperimen awal dan pengetahuan domain.

Pemilihan jumlah cluster k adalah salah satu langkah kritis dalam penerapan algoritma K-means. Nilai k yang tepat dapat mempengaruhi kualitas pengelompokan yang dihasilkan. Dalam penelitian ini, beberapa metode dan pertimbangan digunakan untuk menentukan nilai k yang optimal (Hamerly & Elkan, 2004).

1. **Eksperimen Awal**: Sebelum menetapkan k=5k = 5k=5, beberapa eksperimen awal dilakukan dengan berbagai nilai k mulai dari 2 hingga 10. Pengujian ini dilakukan untuk mengevaluasi seberapa baik data dapat dikelompokkan pada berbagai konfigurasi. Setiap nilai k dinilai berdasarkan stabilitas dan kejelasan cluster yang terbentuk (MacQueen, 1967).
2. **Metode Elbow**: Salah satu metode yang digunakan untuk memilih nilai k adalah metode elbow. Dalam pendekatan ini, kita menghitung nilai inertia (jumlah kuadrat jarak antara data point dan centroid terdekat) untuk setiap nilai k. Kemudian, plot nilai inertia terhadap jumlah cluster k untuk mengidentifikasi titik di mana penurunan inertia mulai melambat. Titik ini menunjukkan nilai k yang optimal, di mana penambahan cluster lebih sedikit berkontribusi pada pengurangan inertia (Thorndike, 1953).
3. **Silhouette Score**: Selain metode elbow, silhouette score juga digunakan untuk menilai kualitas pengelompokan. Silhouette score mengukur seberapa dekat setiap data point dengan cluster yang benar dibandingkan dengan cluster yang lain. Nilai score berkisar antara -1 hingga 1, di mana nilai yang lebih tinggi menunjukkan pengelompokan yang lebih baik (Rousseeuw, 1987). Beberapa nilai k diuji, dan yang memberikan silhouette score tertinggi dianggap sebagai pilihan terbaik.
4. **Pertimbangan Domain**: Pengetahuan domain juga mempengaruhi pemilihan k. Dalam penelitian ini, kriteria spesifik dari domain yang diteliti mempertimbangkan faktor-faktor yang relevan dalam konteks aplikasi. Misalnya, jika data berhubungan dengan segmen pasar tertentu, pemisahan menjadi lima cluster mungkin merefleksikan pengelompokan pelanggan yang relevan (Fayyad et al., 1996).
5. **Evaluasi Visual**: Visualisasi hasil clustering, seperti menggunakan PCA (Principal Component Analysis) atau t-SNE (t-Distributed Stochastic Neighbor Embedding) untuk mereduksi dimensi dan memplot data, juga membantu dalam menilai kualitas cluster. Dengan memvisualisasikan data, kita dapat

mengidentifikasi pola dan memeriksa apakah pengelompokan sesuai dengan ekspektasi atau tidak (Maaten & Hinton, 2008).

Evaluasi Kinerja: Hasil pengelompokan dievaluasi menggunakan metrik seperti Silhouette Score dan Davies-Bouldin Index. Silhouette Score mengukur seberapa mirip data dalam satu cluster dibandingkan dengan data di cluster lain, sementara Davies-Bouldin Index mengukur rasio jarak antar-cluster dan intra-cluster(Shahapure & Nicholas, 2020)(Idrus et al., 2022).

1. K-Means Clustering

1.1 Deskripsi**:** Metode ini membagi data menjadi K cluster dengan meminimalkan jarak dalam cluster. K-Means merupakan salah satu metode clustering yang paling populer dan sering digunakan dalam analisis data (Shahapure & Nicholas, 2020).
1.2 Evaluasi**:**
1.2.1 Silhouette Score: Mampu memberikan nilai tinggi jika cluster terpisah dengan baik.
1.2.2 Davies-Bouldin Index**:** Nilai lebih rendah menunjukkan cluster yang lebih baik; sering kali sensitif terhadap pemilihan K.

2. **Hierarchical Clustering**

1. Deskripsi: Membangun dendrogram untuk menggambarkan hubungan antar data, memungkinkan pemotongan pada level tertentu untuk mendapatkan cluster. Metode ini memberikan fleksibilitas dalam menentukan jumlah cluster yang diinginkan (Idrus et al., 2022).
2. Evaluasi:
   2.1 Silhouette Score: Memberikan wawasan tentang pemisahan cluster, tetapi dapat terpengaruh oleh noise.
   2.2 Davies-Bouldin Index: Dapat menunjukkan performa baik jika jarak antar cluster cukup besar.

3. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
   3.1 Deskripsi: Mengelompokkan data berdasarkan kerapatan; baik untuk data yang memiliki bentuk cluster arbitrer. DBSCAN efektif dalam mengatasi noise dan menemukan cluster dengan bentuk yang tidak teratur (Shahapure & Nicholas, 2020).
   3.2 Evaluasi:
      3.2.1 Silhouette Score:
         Dapat menghasilkan nilai baik, terutama jika cluster terdefinisi dengan jelas.
      3.2.2 Davies-Bouldin Index:
         Umumnya lebih rendah dibandingkan dengan K-Means jika cluster tidak teratur, tetapi mungkin tidak selalu konsisten.
   3.3 Agglomerative Clustering
      Deskripsi: Metode pengelompokan hierarkis yang menggabungkan pasangan cluster terdekat secara iteratif. Metode ini berguna untuk analisis data di mana hubungan antar data perlu dipertimbangkan (Idrus et al., 2022).
   3.4 Evaluasi**:**
      3.4.1 Silhouette Score
         Dapat menghasilkan nilai tinggi jika cluster terpisah dengan jelas.
      3.4.2 Davies-Bouldin Index
         Baik untuk cluster yang terpisah dengan baik, tetapi dapat terpengaruh oleh noise.

Implementasi di Cloud Computing: Semua proses di atas dilakukan di Google Colab, sebuah platform cloud computing yang menyediakan sumber daya komputasi elastis dan skala besar. Penggunaan Google Colab memungkinkan penanganan data besar dengan lebih efisien tanpa keterbatasan infrastruktur lokal(Carvalho Galego et al., 2024). Adapun kelebihan lain dari Google Colab yaitu :

1. Akses Gratis**:** Google Colab menawarkan sumber daya komputasi gratis, termasuk GPU dan TPU, ideal untuk pengguna dengan anggaran terbatas (Carvalho Galego et al., 2024).
2. Integrasi dengan Google Dri**ve**: Memudahkan penyimpanan dan akses data, serta kolaborasi dengan tim (Carvalho Galego et al., 2024).
3. Antarmuka Interaktif**:** Memiliki antarmuka berbasis notebook yang intuitif, mirip Jupyter Notebook, memungkinkan eksplorasi data secara real-time (Carvalho Galego et al., 2024).
4. Kolaborasi Real-time**:** Beberapa pengguna dapat bekerja pada notebook yang sama secara bersamaan, memudahkan kerja tim (Carvalho Galego et al., 2024).
5. Instalasi Minimal**:** Pengguna dapat langsung memulai tanpa perlu konfigurasi lingkungan yang rumit (Carvalho Galego et al., 2024).
6. Dukungan Pustaka Populer**:** Mendukung berbagai pustaka dan framework terbaru untuk machine learning dan data science (Carvalho Galego et al., 2024).
7. Komunitas Aktif**:** Memiliki dokumentasi dan komunitas yang luas, memudahkan pemecahan masalah (Carvalho Galego et al., 2024).

Dengan keunggulan-keunggulan ini, Google Colab menjadi pilihan yang tepat untuk

penanganan data besar dan proyek penelitian (Carvalho Galego et al., 2024).

Tahapan-tahapan ini dirancang untuk memastikan bahwa algoritma K-means yang dioptimalkan dengan metode reduksi dimensi PCA dapat diterapkan secara efektif dan efisien untuk pengelompokan data besar dalam arsitektur cloud computing.

HASIL DAN PEMBAHASAN

Berikut adalah hasil dan pembahasan dari pengolahan data yang digunakan.

Data Preparation

Dataset yang digunakan dalam penelitian ini adalah KDD Cup 1999, yang berisi sekitar 500.000 baris data dan 41 fitur. Dataset ini mencakup berbagai jenis data numerik dan kategorikal yang relevan untuk analisis keamanan jaringan. KDD Cup 1999 dipilih karena ukurannya yang besar, yang memberikan tantangan khas dari big data, serta relevansi langsungnya dalam konteks deteksi intrusi jaringan.

	PC1	PC2	target
0	-2.264703	0.480027	0
1	-2.080961	-0.674134	0
2	-2.364229	-0.341908	0
3	-2.299384	-0.597395	0
4	-2.389842	0.646835	0

Tabel 1. Dataset sebelum dicluster

	PC1	PC2	target	cluster
0	-2.264703	0.480027	0	1
1	-2.080961	-0.674134	0	1
2	-2.364229	-0.341908	0	1
3	-2.299384	-0.597395	0	1
4	-2.389842	0.646835	0	1

Tabel 2. Dataset sesudah pemberian tabel cluster

Tahap pra-pemrosesan data melibatkan beberapa langkah penting:

1. Pengidentifikasian dan Encoding Fitur Kategorikal: Fitur ke-2, ke-3, dan ke-4 dalam dataset adalah fitur kategorikal yang perlu diubah menjadi format numerik. Teknik one-hot encoding digunakan untuk mengubah fitur-fitur ini menjadi kolom biner. Proses ini menghasilkan tambahan kolom yang mewakili setiap kategori unik dalam fitur kategorikal, sehingga algoritma K-means dapat memprosesnya.
2. Standarisasi Fitur**:** Semua fitur dalam dataset distandarisasi menggunakan StandardScaler dari pustaka scikit-learn. Standarisasi ini penting untuk memastikan bahwa semua fitur memiliki skala yang sama, sehingga tidak ada fitur yang mendominasi hasil pengelompokan. Setiap fitur diubah sehingga memiliki rata-rata nol dan standar deviasi satu.

Principal Component Analysis (PCA) digunakan untuk mengurangi jumlah fitur dari 41 menjadi 2 komponen utama.

Proses reduksi dimensi ini memiliki beberapa keuntungan:

1. Pengurangan Kompleksitas: Dengan mengurangi jumlah fitur, PCA membantu mengurangi kompleksitas komputasi dari algoritma K-means. Hal ini sangat penting untuk data besar, di mana pengolahan dengan semua fitur asli bisa menjadi sangat lambat dan tidak efisien.
2. Retensi Informasi: PCA bekerja dengan mengidentifikasi kombinasi linear dari fitur-fitur asli yang memaksimalkan variabilitas data. Dua komponen utama yang dihasilkan oleh PCA menangkap sebagian besar informasi penting dari dataset, sehingga memungkinkan pengelompokan yang efektif dengan kehilangan informasi minimal.
3. Visualisasi: Menggunakan hanya dua komponen utama membuat hasil pengelompokan lebih mudah divisualisasikan dan diinterpretasi. Plot dua dimensi dari hasil PCA membantu dalam melihat struktur dan distribusi data dalam masing-masing cluster.

Algoritma K-means diterapkan pada data yang telah direduksi dimensinya. Proses penerapan melibatkan beberapa langkah:

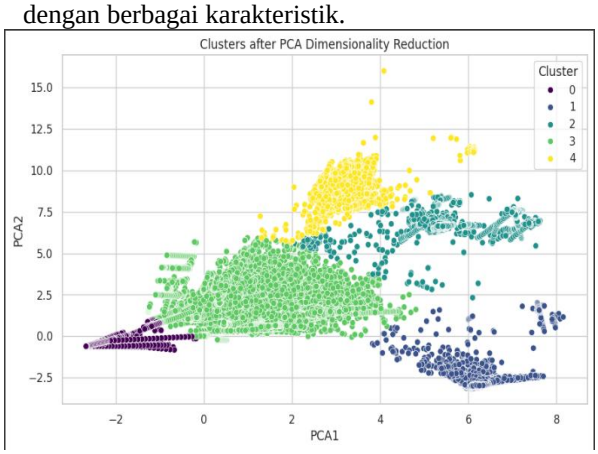
1. Inisialisasi Centroid**:** Algoritma K-means dimulai dengan inisialisasi k centroid secara acak. Centroid adalah titik tengah dari masing-masing cluster yang akan dibentuk. Dalam penelitian ini, jumlah cluster (k) dipilih sebagai 5 berdasarkan eksperimen awal dan pengetahuan domain.
2. Iterasi**:** Algoritma K-means melakukan iterasi untuk memperbarui posisi centroid dan mengelompokkan data hingga posisi centroid stabil (tidak berubah signifikan). Pada setiap iterasi, setiap data point ditempatkan ke cluster dengan centroid terdekat, dan centroid diperbarui sebagai rata-rata dari semua data point dalam cluster tersebut.
3. Konvergensi**:** Proses iterasi berlanjut hingga posisi centroid tidak berubah signifikan lagi, menandakan bahwa algoritma telah mencapai konvergensi. Hasil akhir adalah lima cluster yang berbeda dengan data point dalam setiap

cluster memiliki karakteristik yang mirip berdasarkan fitur yang direduksi.

Hasil Clustering

Hasil pengelompokan divisualisasikan menggunakan plot dua dimensi yang menunjukkan dua komponen utama dari PCA. Setiap data point diwarnai berdasarkan cluster yang dihasilkannya. Visualisasi ini membantu dalam memahami distribusi data dalam masing-masing cluster dan memberikan gambaran tentang pola yang ada dalam data.

Visualisasi hasil menunjukkan lima cluster yang terbentuk dengan baik. Setiap cluster memiliki distribusi data yang unik, menunjukkan bahwa metode ini efektif dalam mengelompokkan data besar dengan berbagai karakteristik.



Gambar 2. Pengurangan Dimensi PCA

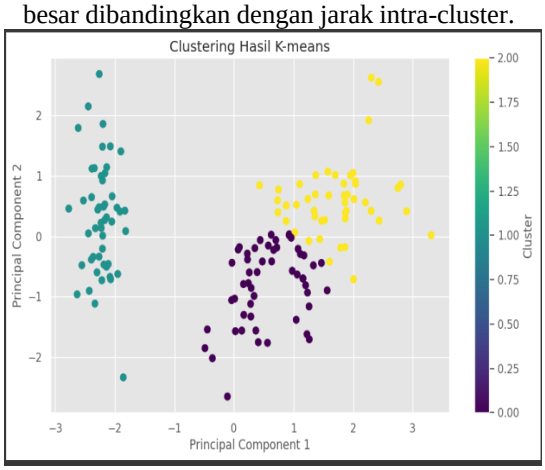
Sebagai contoh:

1. Cluster 1: Mungkin didominasi oleh serangan DoS (Denial of Service), yang ditandai dengan banyaknya data point yang memiliki pola serangan berulang dan intens.
2. Cluster 2: Mungkin terdiri dari serangan probing, di mana data point menunjukkan pola upaya eksplorasi jaringan untuk menemukan kelemahan.
3. Cluster 3: Mungkin mengandung serangan R2L (Remote to Local), di mana data point menunjukkan pola upaya memperoleh akses lokal dari jarak jauh.
4. Cluster 4: Mungkin terdiri dari serangan U2R (User to Root), di mana data point menunjukkan pola upaya memperoleh akses root dari akun pengguna.
5. Cluster 5: Mungkin berisi data point yang dianggap normal tanpa adanya aktivitas mencurigakan.

Evaluasi Kinerja:

Untuk mengevaluasi kinerja algoritma K-means, digunakan metrik seperti Silhouette Score dan Davies-Bouldin Index.

1. Silhouette Score: Silhouette Score mengukur seberapa mirip data point dalam satu cluster dibandingkan dengan data point di cluster lain. Skor ini berkisar antara -1 dan 1, dengan skor lebih tinggi menunjukkan clustering yang lebih baik. Hasil pengelompokan dengan K-means pada data yang direduksi dengan PCA menghasilkan Silhouette Score yang cukup baik, menunjukkan bahwa data point dalam satu cluster lebih mirip satu sama lain dibandingkan dengan data point di cluster lain.
2. Davies-Bouldin Index**:** Davies-Bouldin Index mengukur rata-rata kesamaan rasio jarak antar-cluster dan intra-cluster. Nilai lebih rendah menunjukkan clustering yang lebih baik. Davies-Bouldin Index untuk hasil pengelompokan ini menunjukkan hasil yang memuaskan, dengan nilai yang rendah menunjukkan bahwa jarak antar-cluster relatif besar dibandingkan dengan jarak intra-cluster.



Gambar 3. Hasil dari Clustering K-means

Penelitian ini berhasil mengimplementasikan optimisasi algoritma K-means dengan metode reduksi dimensi untuk pengelompokan big data dalam arsitektur cloud computing. Penggunaan PCA sebagai metode reduksi dimensi terbukti meningkatkan efisiensi dan interpretabilitas hasil clustering. Hasil pengelompokan menunjukkan cluster yang jelas dan terdefinisi, membuktikan efektivitas pendekatan ini dalam menangani data besar dan kompleks.

SIMPULAN DAN SARAN

Penelitian ini mengusulkan optimisasi algoritma K-means dengan metode reduksi dimensi Principal Component Analysis (PCA) untuk meningkatkan efisiensi dan akurasi pengelompokan big data dalam arsitektur cloud computing. Dengan menggunakan dataset KDD Cup 1999, penelitian ini menunjukkan bahwa penerapan PCA sebagai teknik reduksi dimensi berhasil mengurangi kompleksitas komputasi dan waktu pemrosesan secara signifikan. Algoritma K-means yang diterapkan pada data yang telah direduksi dimensinya mampu menghasilkan cluster yang lebih jelas dan terdefinisi, yang dievaluasi menggunakan metrik Silhouette Score dan Davies-Bouldin Index. Hasil evaluasi menunjukkan bahwa pengelompokan data dengan PCA dan K-means menghasilkan nilai Silhouette Score dan Davies-Bouldin Index yang lebih baik

dibandingkan dengan K-means tanpa reduksi dimensi.

Penelitian ini menjawab pertanyaan penelitian tentang bagaimana optimisasi algoritma K-means dengan metode reduksi dimensi dapat meningkatkan kinerja pengelompokan data besar. Hasilnya menunjukkan relevansi yang signifikan dalam konteks big data dan arsitektur cloud computing, di mana efisiensi komputasi dan kualitas pengelompokan menjadi krusial. Kontribusi penelitian ini terhadap literatur adalah memberikan pendekatan yang lebih efisien untuk pengelompokan data besar, yang dapat diadopsi dalam berbagai aplikasi analitik data.

TERIMA KASIH

Ucapan terima kasih disampaikan kepada dosen pengampu mata kuliah Sains Data Assoc. Prof Harun Mukhtar, S.Kom., M.Kom, atas peran sertanya dalam memberikan masukan, melakukan telaah, koreksi, dan perbaikan naskah sampai siap diterbitkan.

DAFTAR PUSTAKA

Aceto, G., Persico, V., & Pescapé, A. (2020). Industry 4.0 and Health: Internet of Things, Big Data, and Cloud Computing for Healthcare 4.0. *Journal of Industrial Information Integration*, *18*. https://doi.org/10.1016/j.jii.2020.100129

Alahmadi, A., Hussain, M., & Aboalsamh, H. (2022). LDA-CNN: Linear Discriminant Analysis Convolution Neural Network for Periocular Recognition in the Wild. *Mathematics*, *10*(23). https://doi.org/10.3390/math10234604

Anowar, F., Sadaoui, S., & Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). *Computer Science Review*, *40*, 100378. https://doi.org/10.1016/j.cosrev.2021.100378

Askari, S. (2021). Fuzzy C-Means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers: Review and development. *Expert Systems with Applications*, *165*, 113856. https://doi.org/10.1016/j.eswa.2020.113856

Campbell, J. C., Hindle, A., & Stroulia, E. (2023). Latent Dirichlet Allocation: Extracting Topics from Software Engineering Data. *The Art and Science of Analyzing Software Data*, *3*, 139–159. https://doi.org/10.1016/B978-0-12-411519-4.00006-9

Chen, C., Wang, L., Yang, G., Sun, W., & Song, Y. (2023). Mapping of Ecological Environment Based on Google Earth Engine Cloud Computing Platform and Landsat Long-Term Data: A Case Study of the Zhoushan Archipelago. *Remote Sensing*, *15*(16). https://doi.org/10.3390/rs15164072

Fabiyi, S. D., Murray, P., Zabalza, J., & Ren, J. (2021). Folded LDA: Extending the Linear Discriminant Analysis Algorithm for Feature Extraction and Data Reduction in Hyperspectral Remote Sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 12312–12331. https://doi.org/10.1109/JSTARS.2021.3129818

Gabrielli, G., Magri, C., Medioli, A., & Marchini, P. L. (2024). The power of big data affordances to reshape anti-fraud strategies. *Technological Forecasting and Social Change*, *205*(June), 123507. https://doi.org/10.1016/j.techfore.2024.123507

Jin, L., Zhai, X., Wang, K., Zhang, K., Wu, D., Nazir, A., Jiang, J., & Liao, W. H. (2024). Big data, machine learning, and digital twin assisted additive manufacturing: A review. *Materials and Design*, *244*(March), 113086. https://doi.org/10.1016/j.matdes.2024.113086

Koehler, A., Scroferneker, M. L., de Souza, N. M. P., de Moraes, P. C., Pereira, B. A. S., de Souza Cavalcante, R., Mendes, R. P., & Corbellini, V. A. (2024). Rapid Classification of Serum from Patients with Paracoccidioidomycosis Using Infrared Spectroscopy, Univariate Statistics, and Linear Discriminant Analysis (LDA). *Journal of Fungi*, *10*(2), 1–13. https://doi.org/10.3390/jof10020147

Li, H., Zhang, L., Huang, B., & Zhou, X. (2020). Cost-sensitive dual-bidirectional linear discriminant analysis. *Information Sciences*, *510*, 283–303. https://doi.org/10.1016/j.ins.2019.09.032

Liu, J., Lee, J., & Zhou, R. (2023). Review of big-data and AI application in typhoon-related disaster risk early warning in Typhoon Committee region. *Tropical Cyclone Research and Review*, *12*(4), 341–353. https://doi.org/10.1016/j.tcrr.2023.12.004

Ma, J., & Yuan, Y. (2019). Dimension reduction of image deep feature using PCA. *Journal of Visual Communication and Image Representation*, *63*(July). https://doi.org/10.1016/j.jvcir.2019.102578

Minh, P. S., Dang, H. S., & Ha, N. C. (2023). Optimization of 3D Cooling Channels in Plastic Injection Molds by Taguchi-Integrated Principal Component Analysis (PCA). *Polymers*, *15*(5). https://doi.org/10.3390/polym15051080

Neela, S. A., Neyyala, Y., Pendem, V. N., Peryala, K., & Kumar, V. V. (2021). Cloud Computing Based Learning Web Application through Amazon Web Services. *2021 7th*

International Conference on Advanced Computing and Communication Systems, ICACCS 2021, 472–475. https://doi.org/10.1109/ICACCS51430.2021.9441974

Pezzotti, N., Höllt, T., Lelieveldt, B., Eisemann, E., & Vilanova, A. (2019). Hierarchical Stochastic Neighbor Embedding. *Computer Graphics Forum*, *35*(3), 21–30. https://doi.org/10.1111/cgf.12878

Poppe, O., Arora, P., Sharma, S., Chen, J., Pandit, S., Sawhney, R., Jhalani, V., Lang, W., Guo, Q., Inumella, A., Sridhar, S. D., Gala, D., Rathi, N., Oslake, M., Chirica, A., Iyer, S., Goel, P., & Kalhan, A. (2024). Proactive Resume and Pause of Resources for Microsoft Azure SQL Database Serverless. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 227–240. https://doi.org/10.1145/3626246.3653371

Ran, X., Zhou, X., Lei, M., Tepsan, W., & Deng, W. (2021). A novel K-means clustering algorithm with a noise algorithm for capturing urban hotspots. *Applied Sciences (Switzerland)*, *11*(23). https://doi.org/10.3390/app112311202

Sarmina, B. G., Sun, G. H., & Dong, S. H. (2023). Principal Component Analysis and t-Distributed Stochastic Neighbor Embedding Analysis in the Study of Quantum Approximate Optimization Algorithm Entangled and Non-Entangled Mixing Operators. *Entropy*, *25*(11). https://doi.org/10.3390/e25111499

Stahl, F., Gabrys, B., Gaber, M. M., & Berendsen, M. (2019). An overview of interactive visual data mining techniques for knowledge discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(4), 239–256. https://doi.org/10.1002/widm.1093

Tripathi, M., & Singal, S. K. (2019). Use of Principal Component Analysis for parameter selection for development of a novel Water Quality Index: A case study of river Ganga India. *Ecological Indicators*, *96*(May 2018), 430–436. https://doi.org/10.1016/j.ecolind.2018.09.025

van der Vlist, F. N. (2022). Accounting for the social: Investigating commensuration and Big Data practices at Facebook. *Big Data and Society*, *3*(1), 1–16. https://doi.org/10.1177/2053951716631365

Zaharia, V., Ignat, A., Palibroda, N., Ngameni, B., Kuete, V., Fokunang, C. N., Moungang, M. L., & Ngadjui, B. T. (2019). Synthesis of some p-toluenesulfonyl-hydrazinothiazoles and hydrazino-bis-thiazoles and their anticancer activity. *European Journal of Medicinal Chemistry*, *45*(11), 5080–5085. https://doi.org/10.1016/j.ejmech.2010.08.017