

## Perbandingan *Algoritma K-Means* dan *Hierarchical* Untuk Klasterisasi Data Kehadiran Karyawan

Fathin Putri Azizah<sup>1\*</sup>, Shofa Shofiah Hilabi<sup>2</sup>, Tukino<sup>3</sup>, Agustia Hananto<sup>4</sup>

Sistem Informasi, Universitas Buana Perjuangan, Karawang, Indonesia

\*e-mail *Corresponding Author*: si21.fathinazizah@mhs.ubpkarawang.ac.id

### Abstract

*Employee attendance data analysis has an important role in human resource management to improve efficiency and productivity. This study compares the K-Means and Hierarchical methods for grouping employee attendance data to discover performance factors. The goal of this study is to compare the performance of the two algorithms utilizing the assessment metrics Silhouette Score, Calinski-Harabasz Index, and Dunn Index. The methods used include collecting employee attendance data, preprocessing data, applying clustering algorithms, and evaluating performance based on specified metrics. The evaluation results showed that K-Means produced a Silhouette Score of 0.46, a Calinski-Harabasz Index of 522.90, and a Dunn Index of 0.98, while Hierarchical obtained a score of 0.40, 452.85, and 0.86, respectively. These results indicate that K-Means is superior in forming clearer and separate clusters. Based on these findings, the K-Means technique is more recommended for employee attendance data analysis because it provides more optimal cluster separation.*

**Keywords:** *Clustering; K-Means; Hierarchical; Attendance Data; Evaluation*

### Abstrak

Analisis data kehadiran karyawan memiliki peran penting dalam manajemen sumber daya manusia untuk meningkatkan efisiensi dan produktivitas. Penelitian ini membahas perbandingan algoritma *K-Means* dan *Hierarchical* dalam klasterisasi data kehadiran karyawan guna mengidentifikasi karakteristik kinerja. Tujuan dari penelitian ini adalah mengevaluasi performa kedua algoritma menggunakan metrik evaluasi *Silhouette Score*, *Calinski-Harabasz Index*, dan *Dunn Index*. Metode yang digunakan meliputi pengumpulan data kehadiran karyawan, *preprocessing* data, penerapan algoritma klasterisasi, serta evaluasi performa berdasarkan metrik yang ditentukan. Hasil evaluasi menunjukkan bahwa *K-Means* menghasilkan *Silhouette Score* sebesar 0.46, *Calinski-Harabasz Index* sebesar 522.90, dan *Dunn Index* sebesar 0.98, sedangkan *Hierarchical* memperoleh nilai masing-masing sebesar 0.40, 452.85, dan 0.86. Hasil ini mengindikasikan bahwa *K-Means* lebih unggul dalam membentuk klaster yang lebih jelas dan terpisah. Berdasarkan temuan ini, algoritma *K-Means* lebih direkomendasikan untuk analisis data kehadiran karyawan karena memberikan pemisahan klaster yang lebih optimal.

**Kata kunci:** *Klasterisasi; K-Means; Hierarchica; Data Kehadiran; Evaluasi*

### 1. Pendahuluan

Dalam manajemen sumber daya manusia, analisis data kehadiran karyawan merupakan elemen penting yang dapat meningkatkan produktivitas dan efisiensi organisasi. Perusahaan yang mampu melakukan analisis kehadiran secara efektif akan dapat mengidentifikasi pola-pola kehadiran yang berkaitan dengan kinerja, sehingga mendukung pengambilan keputusan berbasis data [1]. Dengan kemajuan teknologi, pencatatan kehadiran kini telah beralih ke sistem digital, seperti fingerprint, RFID, dan aplikasi berbasis cloud, yang memungkinkan pengumpulan data dalam volume besar [2]. Data ini tidak hanya berguna untuk administrasi absensi, namun juga dapat dianalisis lebih lanjut untuk mendeteksi pola kehadiran dan tren kerja karyawan.

Salah satu metode yang banyak diterapkan dalam analisis data kehadiran adalah teknik klasterisasi, yang memungkinkan perusahaan untuk mengelompokkan karyawan berdasarkan karakteristik kehadiran mereka [3]. Melalui teknik ini, manajemen dapat mengenali kelompok karyawan dengan kebiasaan kehadiran yang serupa, seperti mereka yang selalu tepat waktu, sering terlambat, atau yang memiliki pola kerja fleksibel [4]. Hasil dari klasterisasi ini dapat

digunakan untuk merancang jadwal kerja yang lebih efisien, serta menetapkan kebijakan berbasis data yang bertujuan meningkatkan kinerja karyawan.

Namun, tantangan utama yang dihadapi dalam analisis data kehadiran adalah pemilihan algoritma klusterisasi yang tepat. Algoritma *K-Means* dikenal karena kemampuannya dalam mengelola data besar, tetapi memiliki kelemahan dalam menentukan total kluster yang terbaik dan sensitif atas pemilihan titik awal pusat kluster [5]. Di sisi lain, algoritma *Hierarchical* tidak memerlukan definisi jumlah kluster di awal dan lebih intuitif dalam visualisasi hubungan antar data, tetapi memiliki keterbatasan dalam kapasitas komputasi saat diterapkan pada dataset yang besar [6]. Oleh karena itu, studi yang membandingkan kedua algoritma ini dalam pengelolaan kehadiran karyawan sangat diperlukan untuk menentukan pendekatan yang lebih efektif.

Beberapa penelitian sebelumnya telah membandingkan metode klusterisasi dalam konteks yang berbeda. Misalnya, penelitian oleh Nahya Nur dan kolega (2023) tentang pengelompokan daerah berisiko stunting di Indonesia menggunakan *K-Means* dan *Hierarchical Clustering* membuktikan *K-Means* lebih optimal, dengan nilai *Silhouette Coefficient* sebesar 0.48 dan *Calinski-Harabasz Index* sebesar 10.49, dibandingkan dengan *Hierarchical Clustering* yang mencatat nilai 0.47 dan 9.54 [7]. Penelitian ini menggambarkan efektivitas *K-Means* dalam mengelompokkan wilayah berdasarkan parameter tertentu. Namun, penelitian tersebut tidak secara spesifik membahas pengelompokan data kehadiran karyawan, sehingga masih terdapat celah untuk penelitian lebih lanjut dalam penerapan kedua algoritma ini untuk analisis kehadiran. Selain itu, studi yang membandingkan algoritma klusterisasi dalam konteks data kehadiran masih terbatas, khususnya dalam penilaian performa berdasarkan *Silhouette Score*, *Calinski-Harabasz Index*, dan *Dunn Index*.

Dengan demikian, studi ini bertujuan untuk membandingkan algoritma *K-Means* dan *Hierarchical* dalam klusterisasi data kehadiran karyawan, sehingga dapat menentukan metode mana yang paling efektif. Diharapkan temuan analisis ini dapat memberikan keterlibatan dalam proses analitik sumber daya manusia dan memberikan wawasan praktis bagi perusahaan dalam memilih algoritma klusterisasi yang sesuai untuk pengelolaan kehadiran karyawan.

## 2. Tinjauan Pustaka

Surya Darma dan rekan-rekannya melakukan sebuah studi yang berfokus pada analisis data kehadiran karyawan di Dinas Pekerjaan Umum dan Penataan Ruang Kabupaten Langkat, dengan menggunakan metode *K-Means clustering*. Studi ini bertujuan untuk mengkategorikan karyawan berdasarkan pola kehadiran mereka guna membantu dalam penilaian kinerja. Temuan penelitian mengungkapkan bahwa *K-Means* berhasil membuat tiga kluster utama: karyawan dengan kehadiran yang tidak tepat waktu, karyawan dengan kehadiran yang moderat, dan karyawan dengan kehadiran yang tepat waktu, sehingga membantu lembaga dalam pembuatan kebijakan untuk meningkatkan disiplin kerja [8].

Tulus Hastuti Meifera dan rekan-rekannya melakukan studi mengenai kategorisasi pembagian kerja di PT Anugrah Analisis Sempurna dengan meneliti elemen psikososial yang mempengaruhi stres karyawan, menggunakan teknik *K-Means Clustering*. Studi ini menggunakan Analisis Rasio Varians untuk mengidentifikasi jumlah kluster yang ideal, menghasilkan dua kelompok dengan rasio varians sebesar 1,42%. Temuan menunjukkan bahwa kluster pertama mengalami tingkat stres yang lebih rendah, sementara kluster kedua menghadapi stres sedang, yang menunjukkan bahwa pendekatan ini dapat membantu perusahaan dalam mengelola stres karyawan secara efektif [9].

Nathalia Clarissa Anggraini Suhartanto dan timnya melakukan studi mengenai pengelompokan kinerja pegawai di CV Mediatama Perkasa dengan menggunakan cara *K-Means Clustering*. Penelitian ini membandingkan pengelompokan berdasarkan Indikator Kinerja Utama (KPI) dan menilai kualitas pengelompokan dengan menggunakan *Silhouette Score* dan *Davies-Bouldin Index* (DBI). Temuan dari studi ini menunjukkan bahwa ada dua kelompok utama yang terbentuk, di mana kelompok pertama memiliki jumlah absensi yang lebih tinggi dan tingkat motivasi kerja yang lebih baik, sedangkan kelompok kedua mengalami lebih banyak kesalahan dalam pekerjaan. Dengan begitu, metode ini dapat membantu perusahaan dalam merumuskan strategi untuk meningkatkan kualitas sumber daya manusia [10].

*State of the art* dalam studi ini melibatkan penggunaan metode *K-Means* dan *Hierarchical Clustering* untuk menganalisis data kehadiran karyawan dalam mengidentifikasi karakteristik kinerja mereka. Beberapa penelitian terdahulu telah menerapkan metode klusterisasi dalam analisis kehadiran karyawan, namun studi ini secara khusus berfokus pada perbandingan

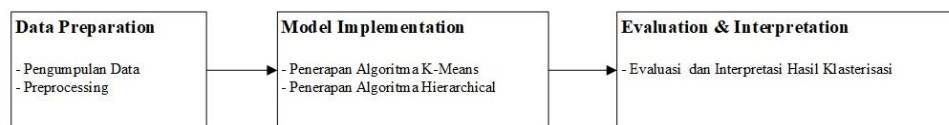
langsung antara *K-Means* dan *Hierarchical Clustering*. Evaluasi terhadap kedua metode ini bertujuan untuk menentukan algoritma yang paling optimal dalam mengelompokkan data kehadiran karyawan, sehingga dapat memberikan wawasan yang lebih mendalam bagi manajemen kehadiran karyawan berbasis data.

**3. Metodologi**

Metodologi penelitian ini menjelaskan proses yang digunakan untuk menerapkan teknik klusterisasi menggunakan algoritma *K-Means* dan *Hierarchical* dalam mengelompokkan data kehadiran karyawan berdasarkan karakteristik kehadiran mereka. Penelitian ini menggunakan pendekatan kuantitatif, dengan analisis berbasis data numerik yang dapat diolah secara matematis untuk mendapatkan pola klusterisasi yang optimal. Data yang diterapkan dalam studi ini merupakan data kehadiran karyawan yang didapat langsung dari perusahaan XYZ. Data ini dikategorikan sebagai data kuantitatif karena terdiri dari atribut numerik yang dapat dianalisis dengan metode komputasional.

**3.1 Tahapan Penelitian**

Studi ini meliputi berbagai proses utama yang disusun secara sistematis, seperti pada gambar 1. Setiap tahap dijelaskan dengan rinci untuk memberikan gambaran proses secara keseluruhan.



Gambar 1. Tahapan Penelitian

Berkaitan dengan gambar 1, tahapan penelitian yang dilaksanakan dapat dilihat penjelasan yang lebih jelas dari bagan tersebut berikut penjelasannya:

1) Pengumpulan Data

Data yang diterapkan dalam studi ini berasal dari perusahaan XYZ, yang mencatat data kehadiran karyawan selama periode Januari hingga Desember 2024. Dataset ini terdiri dari 441 data karyawan dan mencakup berbagai atribut yang menggambarkan karakteristik kehadiran mereka.

Tabel 1. Atribut Dataset

No	Atribut	Keterangan
1	Total Sakit	Jumlah hari karyawan tidak masuk kerja karena sakit selama tahun 2024.
2	Total Cuti Mendadak	Jumlah hari cuti yang diambil secara mendadak tanpa perencanaan selama tahun 2024.
3	Total Cuti Terencana	Jumlah hari cuti yang telah direncanakan dan disetujui sebelumnya selama tahun 2024.
4	Total Tidak Masuk	Jumlah keseluruhan ketidakhadiran karyawan yaitu total dari sakit, cuti mendadak, dan cuti terencana selama tahun 2024.
5	Total Kehadiran	Jumlah hari kerja yang dihadiri oleh karyawan, dihitung dari total hari kerja (240 hari) dikurangi total tidak masuk.
6	Rasio Kehadiran	Proporsi kehadiran karyawan dalam setahun, dihitung sebagai total kehadiran dibagi 240 hari kerja.
7	Cuti Keseluruhan	Jumlah total cuti yang diambil oleh karyawan, termasuk cuti mendadak dan cuti terencana selama tahun 2024.
8	Persentase Cuti	Persentase cuti yang diambil oleh karyawan dalam setahun, dihitung sebagai (Cuti Keseluruhan ÷ 240) × 100%.

- 2) *Preprocessing*  
Proses *preprocessing* dilakukan dengan memeriksa apakah terdapat missing value dalam dataset. Selanjutnya, dipilih fitur yang paling relevan untuk analisis klasterisasi, yaitu Rasio Kehadiran dan Persentase Cuti. Pemilihan atribut ini bertujuan untuk menyederhanakan analisis serta lebih fokus dalam mengidentifikasi karakteristik kehadiran karyawan.
- 3) Implementasi Algoritma  
Pada tahap ini, dilakukan penerapan dua metode klasterisasi untuk membandingkan hasil yang diperoleh yaitu algoritma *K-Means* dan *Hierarchical*.
- 4) Evaluasi dan Interpretasi Hasil Klasterisasi  
Pada tahap ini, hasil klaster yang diperoleh dari algoritma *K-Means* dan *Hierarchical* dievaluasi menggunakan metrik *Silhouette Score*, *Calinski-Harabasz Index*, dan *Dunn Index*. Setelah evaluasi dilakukan, interpretasi hasil klasterisasi dilakukan untuk memahami karakteristik masing-masing klaster serta bagaimana informasi tersebut dapat digunakan untuk mendukung pengambilan keputusan terkait kehadiran karyawan.

### 3.2 Algoritma K-Means

*K-Means Clustering* adalah salah satu teknik pengelompokan yang paling banyak diterapkan dalam analisis data [11]. Algoritma ini berfungsi dengan membagi data menjadi K kelompok yang telah ditentukan sebelumnya, di mana K merujuk pada jumlah kelompok yang diinginkan [12]. Proses ini dimulai dengan pemilihan acak K titik pusat kelompok (*centroid*) di dalam ruang data. Selanjutnya, setiap titik data dikelompokkan ke dalam kelompok yang memiliki *centroid* terdekat [13]. Setelah itu, titik pusat setiap kelompok dihitung ulang berdasarkan nilai tengah dari titik data yang ada dalam kelompok tersebut [14]. Tahap ini akan diulang hingga tidak terdapat lagi perubahan dalam posisi titik data ke dalam kelompok, atau hingga batas iterasi yang ditentukan tercapai [15]. Proses *K-Means* terdiri dari langkah-langkah berikut:

- a. Menentukan jumlah klaster (k) yang optimal, menggunakan metode seperti *Elbow Method* atau *Silhouette Score* untuk menentukan jumlah klaster terbaik.
- b. Inisialisasi *Centroid*, memilih titik pusat (*centroid*) secara sembarang dari data yang ada.
- c. Menghitung jarak setiap data ke *centroid* terdekat, menggunakan *Euclidean Distance* yang dihitung dengan rumus:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \dots \dots \dots (1)$$

Keterangan:

$x$  dan  $y$  : dua titik data.

$x_i$  dan  $y_i$  : nilai dari masing-masing dimensi ke-  $i$ .

$n$  : jumlah dimensi dari dataset.

- d. Mengelompokkan setiap data ke pusat klaster terdekat.
- e. Menghitung ulang pusat klaster berdasarkan rata-rata data dalam klaster.
- f. Mengulangi proses hingga pusat klaster tidak berubah atau jumlah iterasi maksimal tercapai.

### 3.3 Algoritma Hierarchical Clustering

*Hierarchical Clustering* adalah algoritma yang membentuk struktur hierarki dalam proses pengelompokan data. Salah satu metode yang sering diterapkan adalah *Agglomerative Hierarchical Clustering*, di mana setiap data pertama-tama dianggap sebagai klaster pisah dan kemudian digabungkan secara bertahap hingga membentuk satu klaster besar [16]. Teknik yang digunakan dalam proses penggabungan klaster adalah *Ward's Linkage*, yang bertujuan untuk meminimalkan varians total dalam klaster. Rumus yang digunakan adalah:

$$d(C_i, C_j) = \frac{|C_i| + |C_j|}{T} \|\mu_i - \mu_j\|^2 \dots \dots \dots (2)$$

Keterangan:

$d(C_i, C_j)$  : jarak anatara dua klaster,

$|C_i|, |C_j|$  : jumlah data dalam klaster  $C_i$  dan  $C_j$ ,

$T$  : total jumlah data dalam dataset,

$\mu_i, \mu_j$  : *centroid* dari masing-masing klaster.

Adapun langkah kerja dari algoritma *Hierarchical Clustering* adalah sebagai berikut:

- 1) Menentukan tipe *Hierarchical Clustering*, Menggunakan metode *Agglomerative Clustering*, di mana setiap data awalnya dianggap sebagai kelompok terpisah. Kumpulan yang paling berdekatan akan digabungkan sampai hanya ada satu kumpulan utama yang tersisa.
- 2) Menetapkan cara mengukur kesamaan memilih cara untuk menghitung jarak antara data, seperti *Euclidean Distance*, *Manhattan Distance*, atau *Cosine Similarity*.
- 3) Membuat *Matriks Similaritas*, Menghitung matriks jarak antar setiap pasangan data untuk mengetahui tingkat kemiripan antar data. Matriks ini berfungsi untuk membuat dendrogram, yang merupakan struktur hierarki dari pengelompokan data.
- 4) Menyatukan Klaster Beralaskan Similaritas Terdekat, Setiap iterasi, dua klaster dengan jarak terdekat akan digabungkan hingga seluruh data berada dalam satu kelompok.
- 5) Menghitung Jarak Baru Antara Klaster yang Dibentuk, Setelah dua klaster digabungkan, jarak baru antara klaster yang tersisa dihitung menggunakan metode seperti *Ward's Method*.
- 6) Membentuk Dendrogram, Setiap kali klaster baru terbentuk, dendrogram diperbarui untuk merepresentasikan struktur hierarki dataset. Setelah dendrogram terbentuk, dilakukan pemotongan dendrogram pada level tertentu untuk menetapkan total klaster yang optimal.

**3.4 Silhouette Score**

*Silhouette Score* menguji seberapa baik suatu objek ditempatkan dalam klaster dengan membandingkan jarak intra-klaster  $a(i)$  dan jarak inter-klaster terdekat  $b(i)$  [17]. Nilai *Silhouette Score* berkisar antara -1 hingga 1, dengan rumus:

$$S(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \dots\dots\dots(3)$$

Keterangan:

- $S(i)$  : skor silhouette untuk data  $i$ ,
  - $a(i)$  : rata-rata jarak antara  $i$  dan semua titik dalam klasternya sendiri,
  - $b(i)$  : rata-rata jarak antara  $i$  dan semua titik dalam klaster terdekat.
- Semakin tinggi nilai *silhouette score*, semakin baik pemisahan antar klaster.

**3.5 Calinski-Harabasz Index**

*Calinski-Harabasz Index* mengukur rasio antara variabilitas antar-klaster dan intra-klaster [18]. Nilai *Calinski-Harabasz* yang lebih tinggi menunjukkan bahwa klaster yang terbentuk lebih baik. Rumusnya adalah:

$$CH = \frac{T_r(B_k)/(k-1)}{T_r(W_k)/(n-k)} \dots\dots\dots(4)$$

Keterangan:

- $T_r(B_k)$  : jumlah variansi antar klaster.
- $T_r(W_k)$  : jumlah variansi intra klaster.
- $k$  : jumlah klaster.
- $n$  : jumlah total sampel.

**3.6 Dunn Index**

Dunn Index mengukur persentase antara rentang minimum antar-klaster dengan jarak maksimum intra-klaster [19]. Semakin tinggi nilai *Dunn Index*, semakin baik pemisahan antar klaster. Rumus Dunn Index adalah:

$$D = \frac{\min_{i \neq j} d(C_i, C_j)}{\max_k d(C_k)} \dots\dots\dots(5)$$

Keterangan:

- $d(C_i, C_j)$  adalah jarak antara pusat klaster  $C_i$  dan  $C_j$ .
- $d(C_k)$  adalah jarak maksimum antara titik dalam klaster  $C_k$ .

## 4. Hasil dan Pembahasan

### 4.1 Gambaran Umum Data

Studi ini menggunakan dataset kehadiran karyawan yang terdiri dari atribut seperti total cuti terencana, cuti mendadak, total tidak masuk, dan rasio kehadiran. Dataset ini terdiri dari 441 jumlah data, sebelum melewati tahap preprocessing untuk memastikan kualitas data yang lebih baik. Tabel 1 menunjukkan ringkasan statistik sebelum dilakukan klusterisasi.

Tabel 2. Ringkasan Statistik Data

Atribut	count	mean	std	min	25%	50%	75%	max
Total Sakit	441	2.97	5.39	0	0	1	4	52
Total Cuti Mendadak	441	2.69	2.90	0	0	2	4	15
Total Cuti Terencana	441	2.60	2.69	0	1	2	4	15
Total Tidak Masuk	441	8.26	7.28	0	3	6	11	64
Total Kehadiran	441	231	7.28	176	229	234	237	240
Rasio Kehadiran	441	96	3	73	95	97	98	100
Cuti Keseluruhan	441	5.29	3.69	0	2	5	7	17
Persentase Cuti	441	2.21	1.54	0	0	2	2	7

Berdasarkan Tabel 2, dapat dilihat bahwa terdapat variasi dalam jumlah cuti dan kehadiran karyawan. Rata-rata karyawan memiliki rasio kehadiran sekitar 96,56%, dengan beberapa karyawan memiliki tingkat kehadiran penuh (100%) dan yang terendah 73,3%. Selain itu, jumlah cuti yang diambil juga bervariasi, dengan nilai rata-rata cuti keseluruhan 5,29 hari, sementara beberapa karyawan tidak mengambil cuti sama sekali. Distribusi data ini menunjukkan adanya kelompok karyawan dengan pola kehadiran yang berbeda, yang menjadi dasar dalam proses klusterisasi. Dengan melihat penyebaran dan statistik deskriptif ini, analisis lebih lanjut dapat dilakukan untuk mengidentifikasi karakteristik kehadiran karyawan berdasarkan metode klusterisasi yang digunakan.

### 4.2 Preprocessing

Sebelum dilakukan klusterisasi, dataset kehadiran karyawan perlu melalui tahap preprocessing untuk memastikan kualitas data yang lebih baik dan meningkatkan akurasi hasil analisis. Tahapan *preprocessing* yang dilaksanakan dalam studi ini meliputi beberapa proses utama sebagai berikut:

- 1) Pembersihan Data, pada tahap ini dataset diperiksa untuk mengidentifikasi data yang hilang (*missing values*) atau data yang tidak valid. Hasil pemeriksaan menunjukkan bahwa tidak terdapat data yang hilang, sehingga tidak diperlukan imputasi nilai.
- 2) Pemisah Data untuk Klusterisasi, setelah proses pembersihan dan normalisasi, dataset siap digunakan untuk tahap klusterisasi dengan metode *K-Means* dan *Hierarchical Clustering*. Fitur yang ingin dianalisis dalam proses klusterisasi yaitu rasio kehadiran dan persentase cuti karena atribut ini yang paling relevan untuk mengidentifikasi karakteristik kehadiran karyawan agar analisis lebih jelas dan tidak terlalu kompleks.

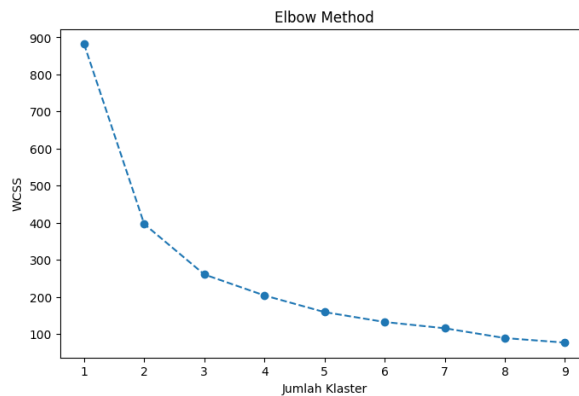
### 4.3 Implementasi Algoritma

Setelah melalui tahap *preprocessing*, dataset siap untuk dianalisis menerapkan metode klusterisasi. Studi ini membandingkan algoritma *K-Means* dan *Hierarchical Clustering* untuk mengelompokkan karyawan berdasarkan karakteristik kehadiran mereka. Implementasi klusterisasi dijalankan menggunakan Google Colab dengan bahasa pemrograman Python.

#### 4.3.1 Penentuan Jumlah Kluster

Menentukan jumlah kluster yang optimal merupakan tahap penting dalam proses klusterisasi, karena jumlah kluster yang tidak tepat dapat menyebabkan hasil klusterisasi yang kurang akurat. Pada penelitian ini, metode yang digunakan untuk menentukan jumlah kluster adalah *Elbow Method* untuk algoritma *K-Means* dan Dendrogram untuk algoritma *Hierarchical Clustering*.

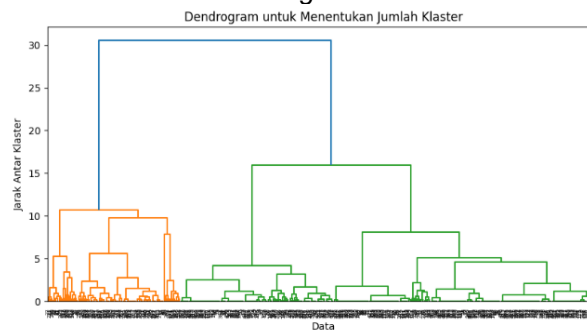
1) *Elbow Method* Untuk *K-Means*



Gambar 2. *Elbow Method*

Pemilihan total kluster dilakukan dengan menggunakan *Elbow Method*, yang didasarkan pada nilai *Within-Cluster Sum of Squares* (WCSS). Grafik menunjukkan bahwa seiring bertambahnya jumlah kluster, nilai WCSS terus menurun, namun dengan tingkat penurunan yang semakin kecil. Titik siku (*elbow point*) dalam grafik ini terlihat pada  $k=3$ , di mana setelah titik tersebut, penurunan WCSS mulai melambat secara signifikan. Oleh karena itu, jumlah kluster optimal yang dipilih adalah  $k=3$ , karena pada titik ini terjadi keseimbangan antara jumlah kluster yang cukup untuk menangkap variasi data tanpa menyebabkan pembentukan kluster yang terlalu kecil atau kurang bermakna.

2) Dendrogram Untuk *Hierarchical Clustering*



Gambar 3. Dendrogram

Pemilihan jumlah kluster dilakukan dengan memotong dendrogram pada tingkat tertentu berdasarkan jarak antar kluster. Pada dendrogram di atas, jumlah kluster optimal ditentukan dengan melihat titik di mana terdapat lonjakan signifikan dalam jarak penggabungan kluster sebelum titik tersebut. Dalam hal ini, pemotongan dilakukan pada tingkat yang menghasilkan tiga kelompok utama tanpa memecah kluster terlalu kecil. Dengan pemilihan ini, struktur hierarki data tetap terjaga, dan kluster yang terbentuk memiliki perbedaan yang jelas satu sama lain.

4.3.2 Distribusi Jumlah Anggota Tiap Kluster

Hasil klusterisasi menunjukkan bahwa data kehadiran karyawan terbagi ke dalam beberapa kelompok dengan distribusi jumlah anggota yang berbeda. Tabel berikut menunjukkan jumlah anggota dalam setiap kluster untuk masing-masing metode.

Tabel 3. Distribusi Jumlah Anggota Tiap Kluster

Metode	Klaster 0	Klaster 1	Klaster 2
<i>K-Means</i>	209	166	66
<i>Hierarchical</i>	212	123	106

Tabel 3 menunjukkan hasil klasterisasi data kehadiran karyawan menerapkan metode *K-Means* dan *Hierarchical Clustering*, yang masing-masing menghasilkan tiga klaster dengan jumlah anggota yang berbeda. Setiap klaster merepresentasikan kelompok karyawan dengan karakteristik kehadiran tertentu berdasarkan pola kehadiran mereka.

- 1) Klaster 0 merupakan kelompok karyawan dengan karakteristik tertentu dalam kehadiran. Jika melihat jumlah anggotanya yang paling besar dibandingkan klaster lain, kemungkinan besar klaster ini merepresentasikan karyawan dengan kehadiran yang lebih stabil atau sering hadir.
- 2) Klaster 1 memiliki jumlah anggota yang lebih sedikit dibandingkan klaster 0. Bisa jadi, klaster ini mewakili karyawan dengan pola kehadiran yang lebih bervariasi, misalnya ada beberapa keterlambatan atau ketidakhadiran dalam periode tertentu.
- 3) Klaster 2 memiliki jumlah anggota paling sedikit untuk metode *K-Means* tetapi lebih banyak dalam metode *Hierarchical*. Klaster ini mungkin berisi karyawan dengan karakteristik kehadiran yang lebih unik, misalnya memiliki tingkat kehadiran yang lebih rendah atau pola kehadiran yang tidak menentu.

Perbedaan jumlah anggota di setiap metode menunjukkan bahwa metode klasterisasi dapat menghasilkan pembagian yang berbeda tergantung pada cara metode tersebut membentuk klaster berdasarkan kedekatan data. Meskipun kedua metode membagi data menjadi tiga klaster, distribusi jumlah anggota dalam tiap klaster tidak selalu sama.

#### 4.3.3 Karakteristik Setiap Klaster

Untuk memahami lebih dalam tentang hasil klasterisasi, dilakukan analisis terhadap karakteristik dari masing-masing klaster berdasarkan atribut dalam dataset. Tabel 3 menunjukkan ringkasan statistik rata-rata dari setiap klaster yang terbentuk.

Tabel 4. Karakteristik Setiap Klaster

Metode	Klaster	Rasio Kehadiran	Persentase Cuti
<i>K-Means</i>	0	98	0.96
	1	96	2.86
	2	91	4.53
<i>Hierarchical</i>	0	92	4.20
	1	98	0.98
	2	96	2.60

Dari tabel diatas dapat diidentifikasi karakteristik masing-masing klaster sebagai berikut:

- 1) Hasil dengan *K-Means* klaster 0 kemungkinan terdiri dari karyawan dengan tingkat disiplin tinggi, jarang mengambil cuti, dan memiliki komitmen kerja yang tinggi. Klaster 1 karyawan dalam klaster ini cenderung memiliki tingkat kehadiran yang cukup tinggi tetapi masih mengambil cuti dalam jumlah yang wajar. Klaster 2 Karyawan dalam klaster ini lebih sering mengambil cuti, yang dapat menunjukkan kebutuhan fleksibilitas kerja yang lebih tinggi atau adanya faktor eksternal yang memengaruhi kehadiran mereka.
- 2) Hasil dengan *hierarchical* klaster 0 mirip dengan Klaster 2 dari *K-Means*, karyawan dalam klaster ini mungkin memiliki pola cuti yang lebih sering dibanding klaster lainnya. Klaster 1 serupa dengan Klaster 0 dari *K-Means*, ini menunjukkan bahwa karyawan dalam klaster ini sangat disiplin dalam kehadiran dan jarang mengambil cuti. Klaster 2 mirip dengan klaster 1 dari *K-Means*.

Dengan adanya perbedaan karakteristik ini, perusahaan dapat menggunakan hasil klasterisasi untuk mengambil keputusan terkait kebijakan kehadiran karyawan.

#### 4.3.4 Evaluasi Dan Interpretasi Hasil Klasterisasi

Setelah melakukan proses klasterisasi menggunakan algoritma *K-Means* dan *Hierarchical*, dilakukan evaluasi terhadap kualitas klaster yang dihasilkan. Evaluasi ini menggunakan tiga metrik utama, yaitu *Silhouette Score*, *Calinski-Harabasz Index*, dan *Dunn Index*. Ketiga metrik ini digunakan untuk menilai seberapa baik algoritma dalam membentuk klaster yang terpisah dengan baik dan memiliki kesamaan internal dalam setiap klaster.



Tabel 5. Hasil Evaluasi dan Interpretasi

Metrik	<i>K-Means</i>	<i>Hierarchical</i>	Interpretasi
<i>Silhouette Score</i>	0.46	0.40	Nilai <i>Silhouette Score</i> yang lebih tinggi pada <i>K-Means</i> membuktikan kluster yang dihasilkan lebih kompak dan lebih terpisah dibandingkan dengan <i>Hierarchical</i> .
<i>Calinski-Harabasz Index</i>	522.90	452.85	<i>Calinski-Harabasz Index</i> yang lebih tinggi pada <i>K-Means</i> menunjukkan bahwa kluster yang terbentuk memiliki keseimbangan antara kepadatan dalam kluster dan keterpisahan antar kluster lebih baik dibandingkan <i>Hierarchical</i> .
<i>Dunn Index</i>	0.98	0.86	<i>Dunn Index</i> yang lebih tinggi pada <i>K-Means</i> mengindikasikan bahwa jarak antar kluster lebih besar, yang berarti kluster lebih terpisah dengan baik dibandingkan <i>Hierarchical</i> .

Berdasarkan hasil evaluasi di atas, dapat disimpulkan bahwa algoritma *K-Means* memiliki hasil yang lebih baik dibandingkan *Hierarchical* dalam menciptakan kluster yang lebih jelas dan terpisah. Hal ini dapat dijelaskan melalui beberapa metrik evaluasi berikut:

- 1) *Silhouette Score* yang lebih tinggi pada *K-Means* membuktikan bahwa kluster yang dihasilkan lebih kompak dalam satu kelompok dan lebih terpisah dari kluster lainnya dibandingkan dengan hasil *Hierarchical*. Namun, selisih skor yang tidak terlalu jauh menunjukkan bahwa perbedaan dalam pemisahan kluster mungkin tidak terlalu signifikan.
- 2) *Calinski-Harabasz Index* yang lebih besar pada *K-Means* mengindikasikan bahwa kluster yang terbentuk memiliki keseimbangan yang lebih baik antara kepadatan dalam kluster dan keterpisahan antar kluster. Indeks ini memperhitungkan rasio antara variabilitas antar-kluster dan intra-kluster, sehingga semakin tinggi nilainya, semakin baik kualitas klusterisasi yang dihasilkan.
- 3) *Dunn Index* yang lebih tinggi pada *K-Means* menyatakan bahwa jarak antar kluster lebih besar dibandingkan *Hierarchical*, yang berarti kluster lebih terpisah dengan baik. Nilai yang lebih tinggi menandakan bahwa titik data dalam suatu kluster lebih dekat satu sama lain, sementara kluster yang berbeda memiliki jarak yang lebih jauh satu sama lain.

Dari hasil ini, *K-Means* dianggap sebagai algoritma yang lebih unggul dalam melakukan klusterisasi data kehadiran karyawan dibandingkan dengan *Hierarchical*. Faktor yang membuat ketidaksamaan ini adalah fleksibilitas *K-Means* dalam menentukan jumlah kluster secara optimal, sedangkan *Hierarchical* menggunakan pendekatan berbasis dendrogram yang lebih cocok untuk dataset dengan jumlah sampel yang lebih kecil. Selain itu, *K-Means* lebih efisien secara komputasi dibandingkan *Hierarchical* karena tidak memerlukan penyimpanan struktur hierarki dari semua data.

Namun, meskipun *K-Means* memiliki keunggulan dalam penelitian ini, *Hierarchical* tetap memiliki kelebihan dalam visualisasi dendrogram yang dapat membantu memahami hubungan hierarki antar data. Oleh karena itu, pemilihan algoritma klusterisasi sebaiknya disesuaikan dengan karakteristik dataset dan tujuan analisis yang ingin dicapai.

#### 4.4 Pembahasan

Hasil studi ini menunjukkan bahwa algoritma *K-Means* lebih unggul dibandingkan dengan *Hierarchical* dalam mengidentifikasi karakteristik kinerja karyawan berdasarkan data kehadiran. Studi dilakukan menggunakan tiga ukuran utama, yaitu *Silhouette Score*, *Calinski-Harabasz Index*, dan *Dunn Index*, yang menunjukkan bahwa *K-Means* dapat mengelompokkan data dengan cara yang lebih jelas dan terpisah dibandingkan dengan *Hierarchical*.

Dalam konteks studi sebelumnya [7], mengkaji kemandirian algoritma *K-Means*. Temuan mereka menunjukkan bahwa *K-Means* unggul dalam membentuk kelompok yang lebih padat dan terpisah dibandingkan dengan metode lain yang diterapkan dalam penelitian tersebut. Hasil ini sejalan dengan penelitian ini, di mana *Silhouette Score* untuk *K-Means* mencapai 0.46, lebih tinggi dibandingkan dengan *Hierarchical* yang hanya 0.40. Perbedaan ini menegaskan bahwa *K-*

*Means* menciptakan kelompok yang lebih baik dalam hal kepadatan dan pemisahan antar kelompok.

Selain itu, ditemukan dalam studi sebelumnya [7] yang membandingkan *K-Means* dengan metode pengelompokan lainnya juga menunjukkan bahwa *Calinski-Harabasz Index* yang lebih tinggi pada *K-Means* menunjukkan keseimbangan antara kepadatan dalam kelompok dan pemisahan antar kelompok lebih optimal. Dalam studi ini, *Calinski-Harabasz Index* untuk *K-Means* adalah 522.90, lebih tinggi dari *Hierarchical* yang hanya 452.85, yang menunjukkan bahwa kelompok lebih jelas dan memiliki distribusi yang lebih baik.

Sementara itu, *Dunn Index* dalam studi ini juga menunjukkan bahwa *K-Means* memiliki pemisahan antar kelompok yang lebih baik dengan nilai 0.98 dibandingkan dengan 0.86 untuk *Hierarchical*. Ini menunjukkan bahwa kelompok yang terbentuk terpisah dengan baik dan memiliki jarak yang lebih ideal antar kelompok data.

Secara keseluruhan, hasil dari studi ini memperkuat temuan dari studi sebelumnya [7] yang menunjukkan bahwa *K-Means* lebih unggul dalam membangun kelompok yang padat dan terpisah dibandingkan metode *Hierarchical*. Namun, masih ada beberapa tantangan dalam penerapan algoritma ini, terutama dalam memilih jumlah kelompok yang tepat, yang dapat mempengaruhi keakuratan hasil pengelompokan. Karena itu, disarankan untuk melakukan studi lebih lanjut yang dapat mengeksplorasi algoritma lain seperti *DBSCAN* atau kombinasi metode pengelompokan dengan teknik optimasi untuk meningkatkan efektivitas pengelompokan dalam analisis data kehadiran karyawan.

Dengan demikian, studi ini memberikan kontribusi terhadap pengetahuan yang ada dengan menekankan bahwa *K-Means* merupakan metode yang lebih efektif dibandingkan *Hierarchical* dalam pengelompokan data kehadiran karyawan, serta memberikan pemahaman baru tentang pentingnya pemilihan ukuran evaluasi dalam menilai kualitas hasil pengelompokan.

## 5. Simpulan

Temuan penelitian membuktikan algoritma *K-Means* lebih unggul dibandingkan *Hierarchical* dalam mengidentifikasi karakteristik kinerja karyawan berdasarkan data kehadiran. Evaluasi menggunakan tiga metrik utama, yaitu *Silhouette Score*, *Calinski-Harabasz Index*, dan *Dunn Index*, menyatakan *K-Means* mampu membentuk klaster yang lebih jelas dan terpisah dengan nilai *Silhouette Score* sebesar 0.46, lebih tinggi dibandingkan *Hierarchical* yang hanya mencapai 0.40, yang menandakan klaster lebih kompak dengan pemisahan yang lebih baik. Selain itu, *K-Means* menghasilkan *Calinski-Harabasz Index* sebesar 522.90, lebih tinggi dibandingkan 452.85 pada *Hierarchical*, menunjukkan bahwa kepadatan dan pemisahan antar klaster lebih optimal. Dari segi pemisahan antar klaster, *K-Means* juga lebih unggul dengan *Dunn Index* sebesar 0.98, dibandingkan *Hierarchical* yang hanya mencapai 0.86. Berdasarkan hasil ini, *K-Means* lebih direkomendasikan dalam klasterisasi data kehadiran karyawan karena menghasilkan struktur klaster yang lebih optimal dan pemisahan yang lebih jelas. Untuk studi seterusnya, dianjurkan untuk mengeksplorasi algoritma lain, seperti *DBSCAN* atau kombinasi metode klasterisasi dengan teknik optimasi, guna meningkatkan akurasi dan efisiensi proses klasterisasi. Selain itu, penggunaan dataset yang lebih besar dan lebih variatif dapat memberikan pemahaman lebih luas perihal pola kehadiran karyawan serta pengaruhnya tentang keproduktifan kerja.

## Daftar Referensi

- [1] F. Wibowo and H. K. Tjahjono, "Optimalisasi kinerja organisasi melalui gaya kepemimpinan yang efektif untuk mencapai kepuasan kerja: sebuah studi literatur," *Entrepreneurship Bisnis Manajemen Akuntansi (E-BISMA)*, vol. 4, no. 1, pp. 129–142, Jun. 2023, doi: 10.37631/ebisma.v4i1.929.
- [2] E. H. Pratiwi and M. A. Ahmadi, "Pengaruh E-Recruitment dan Absensi Online Terhadap Kinerja Karyawan," *Neraca: Jurnal Ekonomi, Manajemen dan Akuntansi*, vol. 8, no. 1, pp. 629–638, 2025, [Online]. Available: <https://ojs.serambimekkah.ac.id/serambi-ekonomi-dan-bisnis/p629>
- [3] E. Arda, A. Aulia, O. Saputra, and J. Heikal, "Employee Performance Segmentation In The Public Housing Service And Payakumbuh City Residential Area With Using The K-Means Clustering Model," *Jurnal Ekonomi Dan Bisnis Digital*, vol. 01, no. 03, pp. 385–389, 2024.
- [4] A. Frianco Bunga, S. Yulianto, and J. Prasetyo, "Optimasi Penilaian Mutu Kerja Pegawai Dengan Metode Clustering Pada RRI Tual," *Jurnal Riset Sistem Informasi Dan Teknik*

- Informatika (JURASIK*, vol. 9, no. 1, pp. 391–397, 2024, [Online]. Available: <https://tunasbangsa.ac.id/ejurnal/index.php/jurasik>
- [5] F. Salsabila, N. Azise, and M. A. Ridla, “Perbandingan Kinerja Algoritma Clustering K-Means Dan Kmedoids Pada Popularitas Line Webtoon,” *Prosiding Seminar Nasional Sains dan Teknologi Seri 02*, vol. 1, no. 2, pp. 3047–6569, 2024.
- [6] H. Februariyanti, J. S. Wibowo, D. B. Santoso, and M. Sukur, “Analisis Kecenderungan Informasi Menggunakan Algoritma Hierarchical Agglomerative Clustering,” *Jurnal Informatika, Manajemen dan Komputer*, vol. 13, no. 1, pp. 9–17, 2021.
- [7] Indra, N. Nur, M. Iqram, and N. Inayah, “Perbandingan K-Means dan Hierarchical Clustering dalam Pengelompokan Daerah Beresiko Stunting,” *Jurnal Inovtek Polbeng - Seri Informatika*, vol. 8, no. 2, pp. 356–367, 2023.
- [8] S. Darma, Y. Yusman, and J. Hendrawan, “Analisis Data Tingkat Kehadiran Pegawai dengan Menggunakan Clustering K-Means Pada Dinas Pekerjaan Umum dan Penataan Ruang Kabupaten Langkat,” *Jurnal Minfo Polgan*, vol. 13, no. 1, pp. 1106–1116, Aug. 2024, doi: 10.33395/jmp.v13i1.13958.
- [9] T. H. Meifera, A. Andriyati, and A. Sumarsa, “Pengelompokan Divisi Kerja Pt. Anugrah Analisis Sempurna Berdasarkan Faktor Psikososial Yang Mempengaruhi Tingkat Stres Karyawan Menggunakan Metode K-Means Clustering,” *Interval: Jurnal Ilmiah Matematika*, vol. 4, no. 2, pp. 51–61, 2024.
- [10] N. C. A. Suhartanto, S. Hadi Wijoyo, and W. Purnomo, “Strategi Peningkatan SDM Berdasarkan Pengelompokan Kualitas Kinerja Pegawai CV Mediatama Perkasa Bogor Menggunakan K-Means Clustering,” *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer (J-PTIIK)*, vol. 9, no. 5, pp. 1–7, 2025, [Online]. Available: <http://j-ptiik.ub.ac.id>
- [11] D. D. Satrio, F. A. Akbar, and M. M. Al Haromainy, “Pengembangan Bot Discord Sebagai Pemutar dan Rekomendasi Musik Menggunakan Metode K-Means,” *Jutisi: Jurnal Ilmiah Teknik Informatika dan Sistem Informasi*, vol. 13, no. 1, pp. 95–106, Apr. 2024, doi: 10.35889/jutisi.v13i1.1681.
- [12] T. Widyanti, S. S. Hilabi, A. Hananto, Tukino, and E. Novalia, “Implementasi K-Means dan K-Nearest Neighbors pada Kategori Siswa Berprestasi,” *Jurnal Informasi & Teknologi (JIdT)*, vol. 5, no. 1, pp. 75–82, 2023.
- [13] A. Lia Hananto *et al.*, “Analysis Of Drug Data Mining With Clustering Technique Using K-Means Algorithm,” *J Phys Conf Ser*, vol. 1908, no. 1, pp. 1–8, Jun. 2021, doi: 10.1088/1742-6596/1908/1/012024.
- [14] N. Kastiawan, B. Huda, E. Novalia, and F. Nurapriani, “Klasterisasi Data Obat dengan Algoritma K-Means (Kasus pada UPTD Puskesmas Curug),” *Jurnal Sains Komputer & Informatika (J-SAKTI)*, vol. 8, no. 1, pp. 120–130, 2024.
- [15] N. Hendrastuty, “Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Dalam Evaluasi Hasil Pembelajaran Siswa,” *Jurnal Ilmiah Informatika dan Ilmu Komputer (JIMA-ILKOM)*, vol. 3, no. 1, pp. 46–56, Mar. 2024, doi: 10.58602/jima-ilkom.v3i1.26.
- [16] N. Satyahadewi, S. J. Sinaga, and H. Perdana, “Hierarchical Cluster Analysis Of Districts/Cities In North Sumatra Province Based On Human Development Index Indicators Using Pseudo-F,” *Barekeng: Jurnal Ilmu Matematika dan Terapan*, vol. 17, no. 3, pp. 1429–1438, Sep. 2023, doi: 10.30598/barekengvol17iss3pp1429-1438.
- [17] D. Astuti and Muqorobin, “Optimasi Metode K-Means Clustering untuk Pengelompokan Obat Di Puskesmas Mertoyudan I Magelang,” *Sistemasi: Jurnal Sistem Informasi*, vol. 13, no. 5, pp. 2144–2160, 2021, [Online]. Available: <http://sistemasi.ftik.unisi.ac.id>
- [18] I. F. Ashari, E. Dwi Nugroho, R. Baraku, I. Novri Yanda, and R. Liwardana, “Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood-Affected Areas in Jakarta,” *Journal of Applied Informatics and Computing*, vol. 7, no. 1, pp. 89–97, Jul. 2023, doi: 10.30871/jaic.v7i1.4947.
- [19] N. A. Kilo, M. R. Katili, and I. K. Hasan, “Perbandingan Metode K-Means dan K-Medoids Dengan Validitas Davies-Bouldin Indeks, Dunn Indeks dan Indeks Connectivity Pada Pengelompokan Masyarakat Penerima Bantuan Langsung Tunai,” *Research in the Mathematical and Natural Sciences*, vol. 4, no. 1, pp. 8–15, 2025, doi: 10.55657/rmns.v4i1.190.