STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 3, December 2025
18th article, Pages: 482 – 502
DOI: https://doi.org/10.32493/sm.v7i3.49999

KOMPARASI HASIL SEGMENTASI METODE *K-MEANS* DAN *AGGLOMERATIVE HIERARCHICAL* TERHADAP PROVINSI DI INDONESIA BERDASARKAN PROFIL PERJALANAN WISATA TAHUN 2024

Ni Luh Ayu Nariswari Dewi[1,*], Azizah Zalfa Assyadida[2], Steffany Marcellia Witanto[3], Muhammad Nasrudin[4], and Kartika Maulida Hindrayani[5]

[1,2,3,4,5] Program Studi Sains Data, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jawa Timur, Indonesia

[*]Correspondence author: *23083010068@student.upnjatim.ac.id*

ABSTRACT

Indonesia is a country with diverse natural and cultural resources, giving it enormous tourism potential. One important factor in the growth of the tourism sector is the movement of domestic tourists. Domestic tourists engage in various types of tourism activities, such as vacations, family visits, religious pilgrimages, and business trips. This diversity reflects the differing characteristics of tourist destinations across provinces, necessitating further analysis to group provinces based on travel profiles. This study aims to compare the results of segmentation using the K-Means method and Agglomerative Hierarchical Clustering (AHC) for provinces in Indonesia based on 2024 travel data sourced from the Central Statistics Agency (BPS). The evaluation of the cluster results using the K-Means method shows the formation of 3 clusters with a Silhouette Score of 0.662. Meanwhile, using the Agglomerative Hierarchical Clustering (AHC) method, 3 clusters were formed with a Silhouette Score of 0.9535 using the average linkage distance selection. This indicates that the objects or data are already in the appropriate clusters.
Keywords: *Travel Destination, K-Means, Agglomerative Hierarchical Clustering.*

ABSTRAK

Indonesia merupakan negara dengan kekayaan alam dan budaya yang beragam sehingga memiliki potensi pariwisata yang sangat besar. Salah satu faktor penting dalam pertumbuhan sektor pariwisata adalah pergerakan wisatawan nusantara. Kegiatan wisata yang dilakukan oleh wisatawan nusantara memiliki berbagai tujuan, seperti liburan, kunjungan keluarga, keagamaan, maupun urusan pekerjaan. Keanekaragaman tersebut mencerminkan adanya perbedaan karakteristik lokasi wisata di setiap provinsi sehingga diperlukan analisis lebih lanjut untuk mengelompokkan provinsi berdasarkan profil perjalanan wisata. Penelitian ini bertujuan untuk membandingkan hasil segmentasi menggunakan metode *K-Means* dan *Agglomerative Hierarchical Clustering (AHC)* terhadap provinsi di Indonesia berdasarkan data perjalanan wisata tahun 2024 yang bersumber dari Badan Pusat Statistik (BPS). Evaluasi hasil *cluster* dengan metode *K-Means* menunjukkan terbentuknya 3 *cluster* dengan *Silhouette Score* sebesar 0,662. Sedangkan, dengan metode *Agglomerative Hierarchical Clustering (AHC)* terbentuk 3 *cluster* yang memiliki nilai *Silhouette Score* sebesar 0,9535 menggunakan pemilihan jarak *average linkage.* Hal tersebut menunjukkan bahwa objek atau data sudah berada pada *cluster* yang sesuai.
Kata kunci: Tujuan Wisata, *K-Means, Agglomerative Hierarchical Clustering*.

ARTICLE INFO

Submission received: 14 June 2025	Accepted: 30 December 2025
Revised: 27 December 2025	Published: 31 December 2025

Available on: https://doi.org/10.32493/sm.v7i3.49999

StatMat
Jurnal Statistika dan Matematika
STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 3, December 2025
18th article, Pages: 482 – 502
DOI: https://doi.org/10.32493/sm.v7i3.49999

1. PENDAHULUAN

Sektor pariwisata merupakan salah satu sektor penting yang memberikan kontribusi signifikan terhadap pertumbuhan ekonomi berbagai negara di dunia (Wijaya et al., 2023). Bentuk kontribusi tersebut terlihat melalui perannya dalam meningkatkan Produk Domestik Bruto (PDB), menciptakan lapangan kerja, dan memberikan dampak positif pada sektor ekonomi lainnya (Prayitno et al., 2023). Indonesia menunjukkan potensi yang unggul dalam sektor pariwisata karena kekayaan alam dan budayanya yang melimpah. Hal tersebut menjadikan Indonesia sebagai destinasi unggulan, tidak hanya bagi wisatawan mancanegara tetapi juga bagi wisatawan nusantara. Dalam konteks nasional, wisatawan nusantara memegang peranan penting sebagai penggerak utama sektor pariwisata karena kontribusinya yang besar dalam mendukung pertumbuhan sektor pariwisata di berbagai daerah.

Menurut data Badan Pusat Statistik (2024), terdapat beragam tujuan perjalanan yang dilakukan oleh wisatawan nusantara, mulai dari liburan, kunjungan keluarga, kegiatan keagamaan, urusan pekerjaan, hingga keperluan pendidikan. Berbagai macam tujuan tersebut menunjukkan bahwa setiap provinsi memiliki karakteristik dan daya tarik wisata yang berbeda-beda. Hal tersebut menunjukkan bahwa perlu dilakukan analisis lebih lanjut untuk melakukan pengelompokan berdasarkan data yang ada untuk memahami bagaimana pola dan karakteristik di setiap provinsi berdasarkan profil perjalanan wisata. Pengelompokkan tersebut penting untuk dilakukan agar strategi pembangunan pariwisata dapat disusun secara efektif yang disesuaikan dengan karakteristik setiap provinsi.

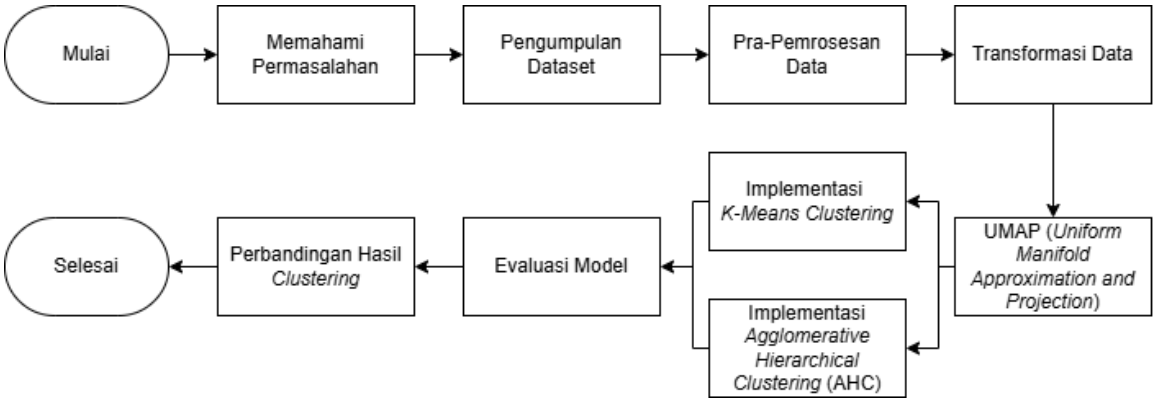
Salah satu bentuk pendekatan yang dapat digunakan untuk melakukan pengelompokan provinsi berdasarkan profil perjalanan wisata adalah dengan menggunakan teknik *clustering*. *Clustering* adalah sebuah teknik analisis yang bertujuan untuk mengelompokkan objek dalam data berdasarkan kemiripan karakteristik ke dalam satu kelompok atau *cluster,* sedangkan objek dengan karakteristik berbeda akan dikelompokkan ke dalam *cluster* lain (Aprilia & Sembiring, 2021). Dua metode yang populer dalam teknik *clustering* adalah *K-Means* dan *Agglomerative Hierarchical Clustering (AHC)*. *K-Means* bekerja dengan prinsip *partitioned clustering,* yaitu prinsip pengelompokan item secara acak berdasarkan nilai *centroid* (Nugroho & Hendrawan, 2022), sedangkan *Agglomerative Hierarchical Clustering (AHC)* memiliki kemampuan menggabungkan data dengan membuat hierarki dimana yang memiliki kemiripan akan ditempatkan di hierarki yang berdekatan dan yang tidak memiliki kemiripan ditempatkan pada hierarki yang berjauhan (Murtagh & Contreras, 2020).

Analisis ini bertujuan untuk membandingkan hasil pengelompokan provinsi di Indonesia berdasarkan profil perjalanan wisata tahun 2024 dengan menggunakan metode *K-Means* dan *Agglomerative Hierarchical*, dimana data yang digunakan bersumber dari publikasi Badan Pusat Statistik. Evaluasi terhadap hasil *clustering* dilakukan menggunakan nilai *Silhouette Score* untuk menilai kualitas pengelompokan antar *cluster*. Melalui pendekatan ini, diharapkan dapat dihasilkan pemetaan karakteristik provinsi dalam hal perjalanan wisata yang dapat dimanfaatkan sebagai dasar penyusunan kebijakan pengembangan pariwisata nasional yang lebih terarah dan berbasis data.

STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 3, December 2025
18th article, Pages: 482 – 502
DOI: https://doi.org/10.32493/sm.v7i3.49999

2. METODOLOGI

Penelitian ini mengikuti tahapan analisis yang sistematis untuk membandingkan performa dua metode segmentasi data, yaitu *K-Means clustering* dan *Agglomerative Hierarchical Clustering* (AHC) dalam mengelompokkan provinsi-provinsi di Indonesia berdasarkan karakteristik tujuan perjalanan wisata selama tahun 2024. Diagram alir berikut akan merangkum setiap tahapan penting dari proses analisis yang diharapkan dapat mempermudah pemahaman terkait alur penelitian.



Gambar 1. Diagram Alir Penelitian

Pengumpulan Data

Data yang diaplikasikan dalam analisis ini berasal dari sumber data sekunder berupa publikasi resmi dari situs Badan Pusat Statistik (BPS) Indonesia yang dapat diakses melalui laman www.bps.go.id. Data tersebut mencakup karakteristik tujuan perjalanan wisatawan domestik ke berbagai wilayah di Indonesia selama tahun 2024, seperti untuk berlibur, mudik, keperluan bisnis, keagamaan, kesehatan, kecantikan, pelatihan, atau tujuan-tujuan lainnya. Secara keseluruhan, terdapat 38 entitas wilayah yang dianalisis, masing-masing entitas merepresentasikan satu dari seluruh provinsi yang ada di Indonesia. Setiap provinsi memiliki atribut jumlah keseluruhan perjalanan yang dilakukan sepanjang tahun dan persentase distribusi tujuan perjalanan yang mencerminkan kecenderungan utama masyarakat dalam melakukan mobilitas wisata domestik.

Deskripsi Variabel Penelitian

Variabel penelitian yang akan digunakan dalam proses analisis disajikan dalam tabel 1 adalah sebagai berikut.

Tabel 1. Variabel Penelitian

Variabel	Keterangan	Skala Data
X_1	Jumlah Perjalanan Wisatawan	Rasio
X_2	Laki-Laki	Rasio
X_3	Perempuan	Rasio
X_4	Persentase Berlibur	Rasio

StatMat
Jurnal Statistika dan Matematika
STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 3, December 2025
18th article, Pages: 482 – 502
DOI: https://doi.org/10.32493/sm.v7i3.49999

X_5	Persentase Kesehatan dan Kecantikan	Rasio
X_6	Persentase Keagamaan	Rasio
X_7	Persentase Mengunjungi Teman	Rasio
X_8	Persentase Mudik	Rasio
X_9	Persentase Olahraga	Rasio
X_{10}	Persentase Belanja	Rasio
X_{11}	Persentase Bisnis	Rasio
X_{12}	Persentase MICE	Rasio
X_{13}	Persentase Pelatihan	Rasio
X_{14}	Persentase Tujuan Lainnya	Rasio

Data Pre-Processing

Tahap awal dalam analisis ini adalah proses *pre-processing* data atau tahapan pra-pemrosesan awal yang terdiri dari beberapa langkah penting, seperti pengecekan nilai yang hilang (*missing values*), pengecekan data duplikat, pengecekan *outlier,* transformasi data, normalisasi data, serta reduksi data menggunakan UMAP (*Uniform Manifold Approximation and Projection*). Tahapan ini dilakukan untuk memastikan kualitas, konsistensi dan kelengkapan data sebelum masuk ke tahap analisis *clustering*.

Outlier

Outlier atau pencilan dalam data merupakan nilai-nilai ekstrem yang ditemukan dalam objek pengamatan, baik karena nilainya jauh lebih rendah maupun jauh lebih tinggi, sehingga menimbulkan perbedaan yang signifikan dibandingkan dengan data lainnya (Fitrayana & Saputro, 2022). Terdapat beberapa cara untuk mendeteksi keberadaan *outlier* dalam data, salah satunya adalah dengan menggunakan metode grafik boxplot dan perhitungan nilai kuartil dan interkuartil data. Suatu data akan dikategorikan sebagai *outlier* ketika memiliki nilai observasi lebih kecil dari $Q1-1.5*IQR$ atau lebih besar dari $Q3+1.5*IQR$ (Sihombing, et al., 2022).

Transformasi Logaritma

Salah satu metode statistik yang dapat digunakan untuk mengurangi *skewness* atau ketidaksimetrisan distribusi dalam data adalah transformasi logaritma. Metode ini, dapat mengubah distribusi data yang tadinya tidak normal menjadi normal atau mendekati normal. Data akan diubah ke dalam bentuk logaritma (log) dengan formula berikut (Saputri, 2023).

$$X' = ln(1 + X) \quad (1)$$

StatMat
Jurnal Statistika dan Matematika
STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 3, December 2025
18th article, Pages: 482 – 502
DOI: https://doi.org/10.32493/sm.v7i3.49999

Normalisasi Data

Pada proses analisis *clustering*, perbedaan rentang nilai pada data dapat menyebabkan dominasi terhadap fitur tertentu, yang pada akhirnya akan sangat mempengaruhi hasil analisis dan pembentukan klaster. Hal ini dapat diatasi dengan melakukan normalisasi menggunakan metode *Min-Max Scaling* agar variabel numerik dalam data memiliki rentang nilai antara 0 hingga 1. Formula untuk melakukan normalisasi ini adalah (Allorerung, et al., 2024).

$$x' = \frac{x_i - min(x)}{max(x) - min(x)} \tag{2}$$

Keterangan ;

x' = Nilai hasil normalisasi
x_i = Nilai asli
$min(x)$ = Nilai minimum fitur
$max(x)$ = Nilai maksimum fitur

Reduksi Dimensi dengan UMAP

Uniform Manifold Approximation and Projection atau yang biasa dikenal sebagai UMAP, merupakan salah satu metode yang membantu mereduksi dimensi data menjadi lebih rendah tanpa menghilangkan struktur lokal dan global dari data. Menurut Suhandi, et al (2025) proses reduksi ini dimulai dengan pembangunan graf tetangga terdekat berdasarkan kedekatan jarak antar titik dalam ruang berdimensi tinggi. Kedekatan ini biasanya dihitung menggunakan jarak Euclidean, yaitu ukuran linear antara dua titik data yang dirumuskan sebagai:

$$euc = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{3}$$

Keterangan:

p_i = Koordinat ke-i dari titik pertama p
q_i = Koordinat ke-i dari titik kedua q

Setelah graf tetangga terdekat terbentuk, UMAP melanjutkan dengan menyusun struktur *simplicial,* yaitu representasi topologis yang menggambarkan keterkaitan lokal antar titik dalam ruang data. Proses ini kemudian diikuti oleh tahap optimasi topologi, di mana struktur dari ruang berdimensi tinggi diproyeksikan ke dalam dimensi yang lebih rendah dengan tujuan untuk mempertahankan sebanyak mungkin informasi lokal dalam data. Selain itu, UMAP juga mempertimbangkan distribusi jarak antar titik di ruang berdimensi rendah, yang diasumsikan mengikuti pola distribusi tertentu yang mencerminkan hubungan sebenarnya antardata.

Clustering

Analisis clustering merupakan teknik eksploratori dalam *data mining* yang bertujuan untuk mengelompokkan sejumlah data ke dalam beberapa kelompok (*cluster*) berdasarkan kemiripan karakteristik (Fauziah, & Basir, C., 2024). Metode ini memungkinkan objek-objek

StatMat
Jurnal Statistika dan Matematika
STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 3, December 2025
18th article, Pages: 482 – 502
DOI: https://doi.org/10.32493/sm.v7i3.49999

yang memiliki karakteristik serupa untuk berada dalam satu kelompok, sedangkan objek yang berbeda dikelompokkan secara terpisah. Dalam penelitian ini, dua algoritma yang digunakan untuk melakukan proses *clustering* adalah *K-Means Clustering* dan *Agglomerative Hierarchical Clustering* (AHC).

K-Means Clustering

Salah satu algoritma paling populer dan sederhana dalam analisis *clustering* adalah *K-Means Clustering* (Permata Sari, Y., & Basir, C., 2025). Algoritma ini termasuk dalam algoritma non hirarki yang bekerja dengan mengelompokkan data berdasarkan kemiripan karakteristik. Data-data dengan atribut atau pola serupa akan ditempatkan ke dalam satu kelompok yang sama, sedangkan data dengan karakteristik berbeda akan dimasukkan ke dalam kelompok lainnya. Menurut Sulistiyawati & Supriyanto (2021) metode ini pada dasarnya bertujuan untuk membentuk K kelompok dengan menentukan sejumlah pusat klaster (centroid) dari kumpulan data yang ada. Tujuan utama dari proses ini adalah untuk mengoptimalkan fungsi objektif, yaitu dengan cara mengurangi variasi di dalam setiap kelompok serta meningkatkan perbedaan antar kelompok, sehingga hasil pengelompokan menjadi lebih representatif dan terstruktur. Tahapan yang dilakukan dalam implementasi metode *K-Means* adalah sebagai berikut (Pribadi, et al., 2022).

1. Menetapkan jumlah klaster (k) yang akan dibentuk.
2. Menginisialisasi pusat *cluster* (*centroid*) yang akan digunakan sebagai titik awal untuk mengukur kedekatan data.
3. Menghitung jarak antara setiap data ke masing-masing *centroid* terdekat menggunakan jarak *euclidean.*
4. Mengelompokkan data ke dalam klaster berdasarkan *centroid* yang memiliki jarak terdekat dengan data.
5. Menghitung ulang posisi centroid baru berdasarkan rata-rata dari data yang tergabung dalam masing-masing *cluster.* Secara matematis dapat dituliskan dengan (Supardi & Kanedi, 2020).

$$C_i = \frac{1}{M}\sum_{j=1}^{M} x_j \tag{4}$$

Keterangan:
C_i : *Centroid* fitur ke-i
M : Jumlah data dalam sebuah kelompok
i : Fitur ke-i dalam sebuah kelompok

6. Mengulangi proses dari langkah ketiga hingga kondisi *centroid* telah stabil dan seluruh data telah berada pada *cluster* yang tepat tanpa adanya perpindahan.

Elbow Method

Metode *Elbow* merupakan salah satu pendekatan umum yang digunakan untuk menentukan jumlah *cluster* optimal dalam algoritma *K-Means.* Teknik ini bekerja dengan menghitung nilai *Sum of Square Error* (SSE) dari tiap *cluster.* Lebih jelasnya, metode ini membandingkan nilai evaluasi tiap-tiap *cluster*, menambahkan jumlah *cluster* secara bertahap, lalu memvisualisasikannya dalam bentuk grafik. Titik di mana terjadi penurunan paling tajam dan membentuk sudut (*elbow*) menunjukkan jumlah *cluster* yang paling efisien, karena setelah titik tersebut, penambahan *cluster* tidak lagi memberikan peningkatan yang signifikan terhadap performa model (Harani, et al., 2020). Rumus untuk perhitungan SSE dapat dilihat di persamaan 5 (Riani, et al., 2023)

StatMat
Jurnal Statistika dan Matematika
STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 3, December 2025
18th article, Pages: 482 – 502
DOI: https://doi.org/10.32493/sm.v7i3.49999

$$SSE = \sum_{k=1}^{K} \sum_{x_i \in S_k} ||x_i - C_k||^2 \quad (5)$$

Keterangan:

x_i = Data ke-i dalam cluster ke-k
S_k = Sekumpulan data (anggota) pada cluster ke-k
C_k = Titik pusat cluster (centroid) ke-k

Agglomerative Hierarchical Clustering

Agglomerative Hierarchical Clustering (AHC) merupakan salah satu metode pengelompokkan hirarki yang bekerja dengan membentuk tingkatan tertentu seperti struktur pohon. Hasil implementasi dari teknik ini umumnya divisualisasikan dalam bentuk dendogram. Terdapat dua pendekatan utama dalam hierarchical clustering, yaitu *Agglomerative* dan *Divisive* tergantung pada proses dekomposisi hirarki yang dilakukan, baik secara *bottom-up* (penggabungan) atau *top-down* (pemisahan). Dari keduanya, metode *agglomerative* adalah metode yang paling umum digunakan, pendekatannya diawali dengan memperlakukan setiap objek sebagai *cluster* terpisah, lalu secara bertahap digabungkan hingga membentuk *cluster* yang lebih besar. Proses ini akan terus dilakukan sampai seluruh objek tergabung dalam satu *cluster* utama. Berikut adalah beberapa metode pemilihan jarak pada AHC (Nellie, et al., 2021).

a. *Single Linkage* (Jarak Terdekat)

Perhitungan jarak dengan *Single Linkage* berfokus pada pengukuran berdasarkan jarak terdekat. Artinya, apabila dua *cluster* berada pada jarak yang pendek, keduanya akan digabungkan ke dalam satu *cluster* yang sama.

$$d_{uv} = min\{d_{uv}\}, d_{uv} \epsilon\ D \quad (6)$$

Keterangan:
d = Kriteria
d_{uv} = Jarak data u ke v dari masing-masing *cluster*
ϵ = Eliminasi

b. *Average Linkage* (Jarak Rata-Rata)

Perhitungan jarak dengan *Average Linkage* berfokus pada pengukuran rata-rata jarak antara pasangan titik yang berasal dari dua *cluster* berbeda.

$$d_{uv} = \frac{1}{U_n \, X \, V_n} \Sigma d_{uv} \epsilon\ D \quad (7)$$

Keterangan:
d = Kriteria
d_{uv} = Jarak data u ke v dari masing-masing *cluster*
ϵ = Eliminasi
U_n , V_n = Jumlah data dalam masing-masing *cluster*

c. *Complete Linkage* (Jarak Terjauh)

Perhitungan jarak dengan *Complete Linkage* berfokus pada pengukuran berdasarkan jarak terjauh. Metode ini akan menggabungkan dua *cluster* terpisah yang memiliki jarak terjauh.

StatMat
Jurnal Statistika dan Matematika
STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 3, December 2025
18th article, Pages: 482 – 502
DOI: https://doi.org/10.32493/sm.v7i3.49999

$$d_{uv} = max\{d_{uv}\}, d_{uv} \epsilon\ D \quad (8)$$

Keterangan:
d = Kriteria
d_{uv} = Jarak data u ke v dari masing-masing *cluster*
ϵ = Eliminasi
U_n , V_n = Jumlah data dalam masing-masing *cluster*

d. *Ward*

Ward adalah metode pengelompokan yang mengukur jarak antar *cluster* berdasarkan peningkatan total varians dalam *cluster* setelah penggabungan. Metode ini bertujuan meminimalkan jumlah kuadrat galat (*error sum of squares*) di dalam setiap *cluster.*

Proses *clustering* menggunakan empat metode sebelumnya akan menghasilkan satu metode yang paling optimal. Dalam hal ini, ditentukan berdasarkan nilai indeks RMSSTD (*Root Mean Square Standard Deviation*) terkecil. Nilai RMSSTD terendah menunjukkan komposisi kelompok yang paling baik dan menjadi acuan dalam memilih jumlah *cluster* yang tepat (Matdoan & Delsen, 2020).

Silhouette Coefficient

Hasil pemodelan clustering menggunakan *K-Means* dan AHC akan dievaluasi menggunakan *Silhouette Coefficient*. Skor ini akan memvalidasi hasil pengelompokkan dengan mengukur seberapa cocok suatu objek dengan *cluster* yang ditempati dibandingkan dengan *cluster* lainnya. Adapun perhitungan *Silhouette* dapat dihitung dengan rumus (Drl et al., 2023):

$$S_i = \frac{b_i - a_i}{max\ (a_i - b_i)} \quad (9)$$

Keterangan:

S_i : Nilai *Silhouette Coefficient* untuk objek ke-i
a_i : Nilai kohesi objek ke-i
b_i : Nilai separasi objek ke-i

Silhouette Coefficient memiliki rentang nilai dari -1 hingga 1. Nilai silhouette yang semakin mendekati angka 1, menunjukkan bahwa hasil pengelompokan memberikan hasil yang semakin optimal. Berdasarkan Nurhaliza & Mukhti (2025), terdapat kriteria interpretasi nilai *silhouette* yang dapat dilihat pada tabel 2.

Table 2. Kategori Evaluasi Nilai *Silhouette Coefficient*

Nilai *Silhouette Coefficient*	Kriteria *Cluster*
$0{,}71 - 1{,}00$	Struktur Kuat
$0{,}51 - 0{,}70$	Struktur Baik
$0{,}26 - 0{,}50$	Struktur Lemah
$\leq 0{,}25$	Struktur Buruk

StatMat
Jurnal Statistika dan Matematika
STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 3, December 2025
18th article, Pages: 482 – 502
DOI: https://doi.org/10.32493/sm.v7i3.49999

3. HASIL DAN PEMBAHASAN

Penelitian ini bertujuan untuk melakukan komparasi antara dua algoritma *clustering*, yaitu *K-Means* dan *Agglomerative Hierarchical* guna mengevaluasi efektivitas masing-masing algoritma dalam proses pengelompokan data. Analisis yang dilakukan akan menghasilkan pemetaan Provinsi di Indonesia berdasarkan profil perjalanan wisata pada tahun 2024. Pada implementasinya, data harus melalui tahapan *preprocessing* dan *data understanding* terlebih dahulu sebelum dilakukan pembentukan *cluster* dengan kedua algoritma tersebut. Kemudian, hasil segmentasi dari masing-masing metode akan dianalisis dan dibandingkan untuk menentukan metode yang paling optimal dalam pengelompokan data yang diuji.

Data Pre-processing

Pemeriksaan *Missing Value*

Tahapan awal dalam proses *preprocessing* data adalah identifikasi terhadap keberadaan *missing value* atau data yang tidak terisi. Keberadaan nilai yang hilang dapat menurunkan kualitas dan validitas analisis data, sehingga perlu ditangani baik melalui imputasi maupun eliminasi entri yang bersangkutan. Berdasarkan hasil evaluasi terhadap dataset yang digunakan, tidak ditemukan *missing value* sehingga data dapat dinyatakan lengkap dan layak untuk diproses pada tahap berikutnya.

Identifikasi Duplikasi Data

Pendeteksian data duplikat merupakan langkah untuk mengecek kemunculan entri yang tercatat lebih dari satu kali. Duplikasi data berpotensi menyebabkan bias dalam proses analisis dan menurunkan akurasi hasil *clustering*. Oleh karena itu, diperlukan verifikasi untuk memastikan keunikan setiap observasi dalam dataset. Hasil pemeriksaan menunjukkan bahwa tidak terdapat entri duplikat, sehingga tidak diperlukan proses eliminasi data ganda.

Deteksi *Outlier*

Analisis terhadap *outlier* dilakukan untuk mengidentifikasi nilai-nilai ekstrem yang secara signifikan menyimpang dari distribusi umum data. Dalam penelitian ini, metode *Interquartile Range* (IQR) diimplementasikan sebagai pendekatan statistik dalam menentukan batas bawah dan atas dari data yang dianggap wajar. Berdasarkan hasil deteksi, ditemukan adanya sedikit keberadaan *outlier* pada variabel Jumlah Perjalanan Wisatawan. Namun demikian *outlier* tidak dieliminasi dalam proses analisis, mengingat algoritma yang diadopsi pada penelitian ini dapat memanfaatkan keberadaan nilai ekstrem dalam mengidentifikasi pola-pola unik dan karakteristik khusus dalam data, sehingga berpotensi memperkaya hasil pengelompokan.

Transformasi Logaritma

Sebagian besar fitur dalam dataset menunjukkan signifikansi tingkat *skewness*, yang mencerminkan ketidaksesuaian terhadap distribusi normal. Untuk mengatasi kondisi ini dan mengurangi derajat *skewness* pada variabel numerik yang diuji, diterapkan transformasi logaritmik sebagai salah satu teknik penyesuaian distribusi. Transformasi ini berfungsi untuk menstabilkan varians antar variabel serta memperbaiki bentuk distribusi data agar lebih mendekati distribusi normal. Penyesuaian distribusi melalui transformasi logaritmik

StatMat
Jurnal Statistika dan Matematika
STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 3, December 2025
18th article, Pages: 482 – 502
DOI: https://doi.org/10.32493/sm.v7i3.49999

merupakan langkah krusial guna meningkatkan efektivitas algoritma *clustering* dalam mengidentifikasi dan memetakan struktur laten yang terdapat dalam data.

Normalisasi Skala Data

Perbedaan skala antar variabel data yang digunakan pada penelitian ini (ribu dan persentase) dapat menyebabkan ketidakseimbangan kontribusi fitur dalam proses pengelompokan. Oleh karena itu, dilakukan normalisasi data untuk menyetarakan rentang nilai setiap variabel ke dalam skala yang seragam, yakni antara 0 dan 1. Proses ini bertujuan untuk menghindari dominasi variabel tertentu yang memiliki rentang nilai lebih besar, sehingga semua variabel memiliki bobot yang setara dalam proses *clustering*.

Reduksi Dimensi dengan UMAP

Sebagai tahapan akhir dalam proses *preprocessing*, dilakukan reduksi dimensi dengan memanfaatkan teknik *Uniform Manifold Approximation and Projection* (UMAP). Metode ini diterapkan untuk menyederhanakan struktur data berdimensi tinggi menjadi representasi berdimensi lebih rendah tanpa menghilangkan informasi esensial data. Data yang digunakan dalam penelitian ini terdiri dari berbagai fitur yang merepresentasikan karakteristik perjalanan wisatawan dengan jumlah fitur yang cenderung tinggi dan bervariasi. Sehingga, penggunaan UMAP memungkinkan pemetaan ulang data ke dalam ruang berdimensi lebih rendah sambil tetap mempertahankan struktur global dan lokal dari data asli, serta memungkinkan visualisasi data dalam membantu interpretasi hasil clustering yang dihasilkan oleh algoritma yang digunakan dalam penelitian.

Eksplorasi Data Analisis

Pada penelitian ini, diimplementasikan analisis statistika deskriptif untuk meninjau distribusi nilai dari setiap variabel yang mencakup nilai minimum, maksimum, rata-rata, simpangan baku, serta kuartil. Melalui pendekatan ini, dapat dilakukan pemahaman data dengan identifikasi ketimpangan distribusi data seperti dominasi wilayah tertentu atau adanya nilai ekstrem (*outlier*) yang signifikan. Informasi ini penting sebagai dasar dalam menentukan strategi visualisasi yang tepat dan sebagai pertimbangan awal sebelum masuk ke tahap analisis *clustering*. Lebih lanjut, dilakukan visualisasi dalam bentuk peta tematik dari masing-masing hasil *cluster* yang memungkinkan penyajian perbedaan karakteristik antar wilayah secara lebih intuitif, misalnya dalam mengidentifikasi daerah dengan proporsi perjalanan wisata untuk tujuan berlibur yang lebih tinggi dibandingkan daerah dengan dominasi tujuan profesional. Setiap variabel diklasifikasikan ke dalam beberapa kategori berdasarkan *cluster* yang dihasilkan oleh setiap metode, kemudian hasilnya akan ditampilkan dengan gradasi warna berbeda pada peta. Hal ini bertujuan untuk mempermudah pembaca dalam menangkap pola spasial dari *cluster* berdasarkan profil perjalanan.

Statistika Deskriptif

Karakteristik data berdasarkan profil perjalanan wisata di Indonesia pada tahun 2024 menurut klasifikasi wilayah/provinsi disajikan dalam Tabel 3.1 sebagai berikut.

StatMat
Jurnal Statistika dan Matematika
STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 3, December 2025
18th article, Pages: 482 – 502
DOI: https://doi.org/10.32493/sm.v7i3.49999

Tabel 3. Karakteristik Dataset

Variabel	Jumlah	*Mean*	Varians	Minimum	Maksimum
X1	38	26870630	2.37E+15	178819	218711800
X2	38	58,65	110,78	0	67,57
X3	38	38,72	56,89	0	49,89
X4	38	27,25	140,35	0	54,72
X5	38	2,46	2,19	0	5,72
X6	38	1,51	2,02	0	6,27
X7	38	20,25	41,58	0	32,5
X8	38	4,77	6,58	0	10,4
X9	38	0,76	0,44	0	2,68
X10	38	2,44	3,06	0	7,59
X11	38	17,94	65,5	0	38,44
X12	38	5,7	4,37	0	9,27
X13	38	5,42	6,7	0	12,43
X14	38	4,81	3,78	0	9,21

Berdasarkan hasil perhitungan yang disajikan pada Tabel 3, data penelitian ini mencakup 38 observasi yang merepresentasikan kondisi profil perjalanan wisata berbagai wilayah di Indonesia. Variabel utama yang dianalisis meliputi jumlah perjalanan wisatawan, komposisi berdasarkan jenis kelamin, serta persentase tujuan perjalanan yang beragam. Rata-rata jumlah perjalanan wisatawan tercatat sebesar 26,87 juta dengan penyimpangan yang cukup besar, yaitu mencapai 48,68 juta. Hasil ini menjelaskan adanya ketimpangan yang relatif tinggi antar wilayah. Sebaran tersebut menunjukkan bahwa beberapa wilayah menjadi pusat aktivitas wisata yang sangat dominan dibandingkan dengan provinsi lainnya. Dari sisi jenis kelamin, terhitung bahwa rata-rata proporsi wisatawan berjenis kelamin laki-laki cenderung lebih tinggi di angka 58,65%, dibandingkan dengan perempuan yang hanya berkisar 38,72%.

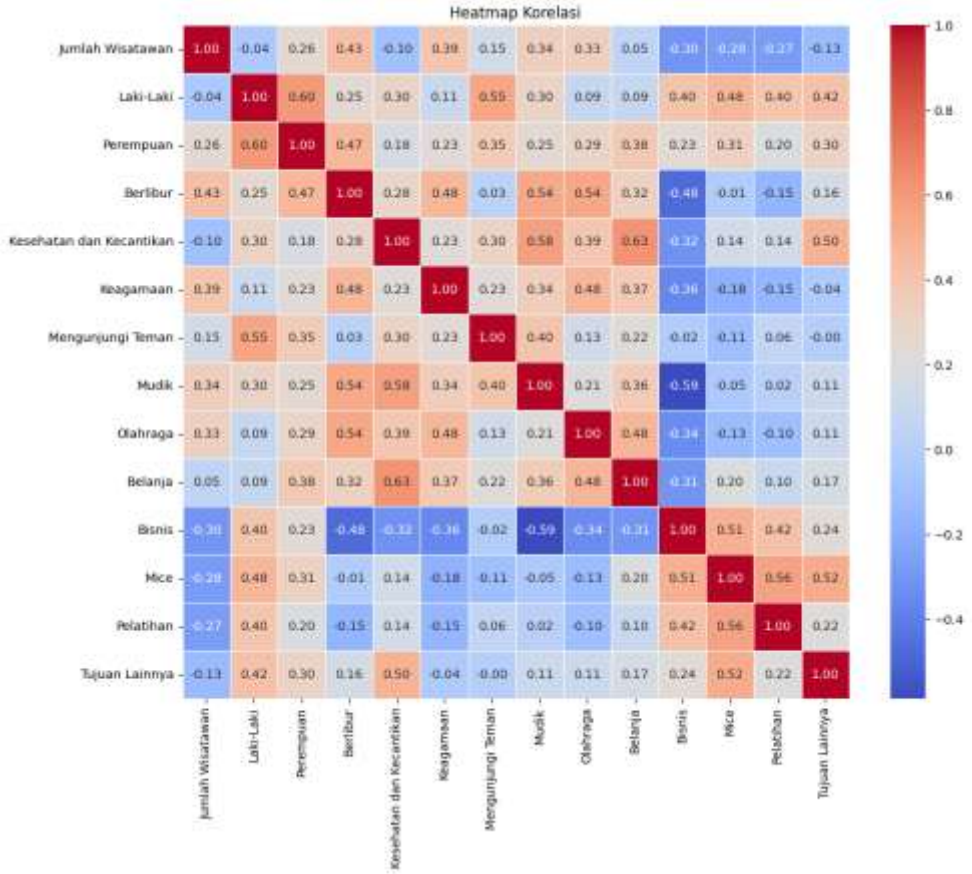
Lebih lanjut ditinjau berdasarkan tujuan perjalanan, aktivitas berlibur menjadi kegiatan yang paling umum dilakukan dengan rata-rata tertinggi, yaitu sebesar 27,25%. Hasil ini diikuti oleh kunjungan ke teman dan keluarga dengan angka 20,25%, serta tujuan bisnis di angka 17,94%. Profil perjalanan lain, seperti pelatihan, MICE (*Meeting, Incentive, Convention, and Exhibition*), dan belanja menunjukkan persentase yang lebih kecil namun tetap signifikan dalam data distribusi nasional. Beberapa tujuan seperti olahraga, kesehatan dan kecantikan, serta keagamaan memiliki rata-rata yang cenderung rendah, namun tetap menunjukkan keragaman preferensi perjalanan di tiap wilayah. Keragaman data ini memperlihatkan adanya perbedaan karakteristik wilayah dalam menarik jenis wisatawan tertentu dengan tujuan perjalanan yang berbeda-beda. Secara keseluruhan, distribusi data menunjukkan adanya disparitas yang cukup menonjol, terutama dalam jumlah perjalanan

StatMat
Jurnal Statistika dan Matematika
STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 3, December 2025
18th article, Pages: 482 – 502
DOI: https://doi.org/10.32493/sm.v7i3.49999

wisata dan tujuan utama. Hasil ini dapat memberikan gambaran awal terkait pola perjalanan wisatawan yang tidak merata di Indonesia.

Korelasi Pearson

Dalam mengidentifikasi hubungan linier antara variabel-variabel numerik dalam data, diterapkan analisis korelasi Pearson. Metode ini bertujuan untuk mengukur sejauh mana perubahan pada satu variabel berkaitan secara linier dengan perubahan pada variabel lainnya.



Gambar 2. *Heatmap* Korelasi Pearson

Gambar 2 berikut menyajikan hasil perhitungan korelasi Pearson antara variabel-variabel yang dianalisis. Berdasarkan *heatmap* yang disajikan, diketahui bahwa fitur Jumlah Perjalanan Wisatawan memiliki korelasi positif sedang dengan Persentase Berlibur (0,43), Persentase Keagamaan (0,39), serta Persentase Mudik (0,34). Hasil ini menjelaskan bahwa semakin tinggi jumlah perjalanan, maka proporsi perjalanan yang dilakukan untuk tujuan tersebut akan cenderung semakin tinggi. Di sisi lain, fitur Jumlah Perjalanan Wisatawan berkorelasi negatif dengan Persentase Bisnis (-0,30), MICE (-0,28), dan Pelatihan (-0,27). Hasil ini menunjukkan bahwa wilayah dengan volume perjalanan yang tinggi cenderung memiliki proporsi perjalanan bisnis yang lebih rendah.

Lebih lanjut, korelasi antara Laki-Laki dan Perempuan berada pada angka 0,60 yang menunjukkan hubungan positif sedang, maka hasil ini menjelaskan bahwa distribusi jenis kelamin wisatawan pada dataset ini relatif sejalan. Variabel Persentase Kesehatan dan Kecantikan memiliki korelasi cukup tinggi dengan Persentase Belanja (0,63). Hasil ini menjelaskan adanya keterkaitan tujuan wisata untuk konsumsi dan perawatan. Sementara

StatMat
Jurnal Statistika dan Matematika
STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 3, December 2025
18th article, Pages: 482 – 502
DOI: https://doi.org/10.32493/sm.v7i3.49999

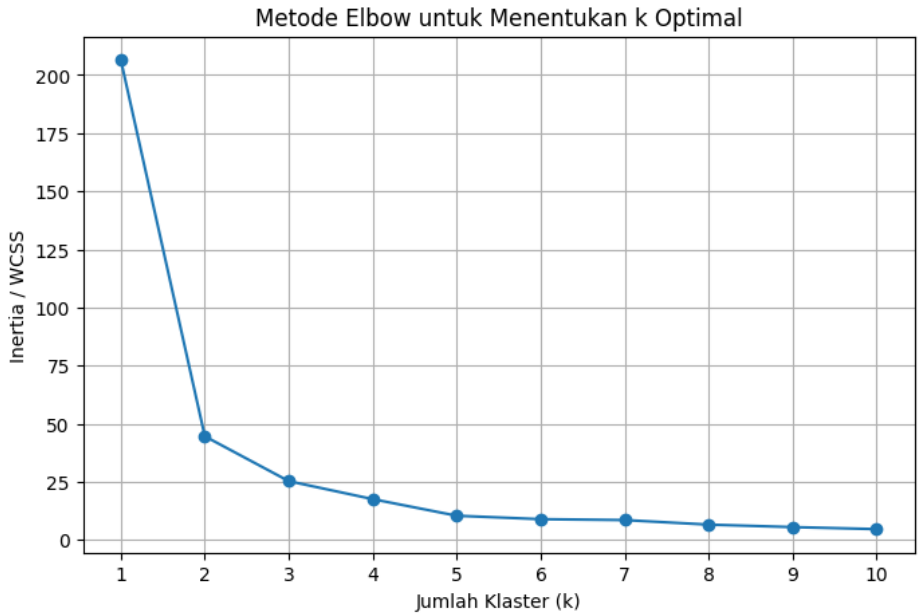
itu, Persentase Bisnis berkorelasi kuat dengan MICE (0,51) dan Pelatihan (0,42) yang menunjukkan bahwa perjalanan untuk tujuan profesional seringkali berkaitan satu sama lain. Sehingga, secara keseluruhan analisis korelasi Pearson ini memberikan gambaran awal mengenai hubungan linier antara tujuan perjalanan yang dapat digunakan untuk mendukung interpretasi pola-pola perjalanan wisata di berbagai wilayah di Indonesia.

Modelling

Dalam tahap ini dilakukan proses pemodelan untuk mengelompokkan data berdasarkan karakteristik tertentu menggunakan metode *clustering*. Untuk memperoleh hasil klasterisasi yang optimal dan komprehensif, penelitian ini mengimplementasikan dua metode yang berbeda, yaitu *K-Means Clustering* dan *Agglomerative Hierarchical Clustering*. Kedua metode ini dipilih karena memiliki pendekatan yang berbeda, dimana *K-Means* bersifat partisional dan iteratif, sementara *Agglomerative Hierarchical* membentuk hirarki dari penggabungan data secara bertahap. Pengujian ini dilakukan untuk mengetahui metode yang memberikan hasil klasterisasi paling representatif berdasarkan karakteristik data yang dianalisis.

K-Means Clustering

Tahap *clustering* dengan pendekatan *K-Means* diawali dengan penentuan nilai *k* yang paling optimal. Pada penelitian ini, metode yang digunakan untuk menentukan nilai *k* adalah metode *elbow*. Metode *elbow* mengukur seberapa besar penurunan nilai *Within-Cluster Sum of Squares* (WSS) seiring dengan bertambahnya jumlah klaster (*k*). Indikator WSS menunjukkan tingkat kedekatan data terhadap pusat klasternya, sehingga semakin rendah nilai WSS maka semakin baik pembentukan klaster tersebut. Nilai *k* yang optimal ditentukan pada titik di mana penurunan WSS mulai melambat atau membentuk siku pada grafik yang menandakan bahwa penambahan jumlah klaster selanjutnya tidak memberikan peningkatan signifikan dalam partisi data. Dengan demikian, metode *elbow* membantu memastikan pembentukan klaster yang efisien dan representatif terhadap struktur data. Berikut ini merupakan hasil visualisasi nilai *elbow* pada data yang digunakan.



Gambar 3. Grafik yang Dibentuk dengan Metode *Elbow* untuk Menentukan Nilai k

StatMat
Jurnal Statistika dan Matematika
STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 3, December 2025
18th article, Pages: 482 – 502
DOI: https://doi.org/10.32493/sm.v7i3.49999

Berdasarkan Gambar 3, titik *elbow* teridentifikasi pada $k = 3$ yang ditandai dengan mulai melambatnya penurunan nilai *Within-Cluster Sum of Squares* (WSS). Hal ini menunjukkan bahwa menambahkan jumlah klaster setelah $k = 3$ tidak akan memberikan peningkatan signifikan dalam partisi data. Oleh karena itu, 3 klaster dipilih sebagai jumlah optimal untuk pembentukan kelompok menggunakan *K-Means Clustering*. Setelah proses *clustering* dilakukan, data akan terbagi menjadi tiga kelompok utama yang memiliki karakteristik unik. Setiap klaster menunjukkan pola khas berdasarkan rata-rata dari variabel seperti jumlah perjalanan wisatawan, proporsi jenis kelamin, serta persentase tujuan perjalanan. Hasil ini disajikan dalam Tabel 4 yang memberikan gambaran awal mengenai segmentasi wilayah berdasarkan perilaku perjalanan wisata.

Tabel 4. Karakteristik Tiap *Cluster* Berdasarkan Nilai Rata-Ratanya

Fitur	***Cluster* 1**	***Cluster* 2**	***Cluster* 3**
Jumlah Wisatawan	87039880	2472560	18134320
Laki-Laki	58,3775	56,73143	60,45875
Perempuan	41,6225	36,12571	39,54125
Berlibur	39,89875	16,94429	29,95438
Kesehatan dan Kecantikan	1,82375	1,617857	3,525625
Keagamaan	3,34875	0,47	1,49875
Mengunjungi Teman	20,26625	18,565	21,71875
Mudik	5,83625	3,280714	5,55
Olahraga	1,04625	0	1,27375
Belanja	2,31125	1,498571	3,326875
Bisnis	13,10125	23,35071	15,62625
Mice	4,35375	6,397143	5,77
Pelatihan	3,94375	5,971429	5,67125
Tujuan Lainnya	4,07	4,525	5,42625

Berdasarkan Tabel 4, dapat diketahui bahwa *Cluster* 1 memiliki karakteristik jumlah perjalanan wisatawan yang paling tinggi dengan proporsi perempuan yang cenderung lebih besar daripada laki-laki. Tujuan utama perjalanan dalam cluster ini adalah untuk berlibur dengan persentase tertinggi, diikuti oleh kunjungan keagamaan dan mudik. Aktivitas belanja dan olahraga tergolong cukup dominan, meskipun tidak sekuat aktivitas berlibur. Lebih lanjut, *Cluster* 2 ditandai oleh jumlah perjalanan wisatawan yang paling rendah dibandingkan klaster lainnya dengan proporsi laki-laki lebih tinggi dibandingkan perempuan. Tujuan perjalanan dalam *cluster* ini lebih beragam dengan dominasi pada perjalanan bisnis dan pelatihan, serta aktivitas MICE (*Meeting, Incentive, Convention, and*

StatMat
Jurnal Statistika dan Matematika
STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 3, December 2025
18th article, Pages: 482 – 502
DOI: https://doi.org/10.32493/sm.v7i3.49999

Exhibition) yang relatif lebih tinggi dibandingkan *cluster* lainnya. Sementara itu, *Cluster* 3 menunjukkan karakteristik jumlah perjalanan yang sedang, dengan proporsi gender yang relatif seimbang. Motif perjalanan dalam *cluster* ini didominasi oleh tujuan berlibur, namun lebih seimbang dengan tujuan lainnya seperti keagamaan, bisnis, dan belanja. *Cluster* ini juga memiliki persentase kesehatan dan kecantikan tertinggi dibanding *cluster* lain. Untuk memperoleh pemahaman yang lebih mendalam mengenai hasil *cluster*, visualisasi data dilakukan melalui peta tematik yang menggambarkan sebaran data dari masing-masing *cluster*. Penyajian tersebut dapat dilihat pada Gambar 3.3 yang menampilkan representasi visual hasil klasterisasi dalam ruang dua dimensi.

Gambar 4. Peta Tematik Hasil Segmentasi Wilayah di Indonesia Berdasarkan Profil Perjalanan Tahun 2024 dengan Metode *K-Means*

Berdasarkan visualisasi dan hasil pengelompokan yang dilakukan menggunakan metode *K-Means Clustering*, diperoleh tiga *cluster* utama yang merepresentasikan karakteristik perjalanan wisata di berbagai provinsi. *Cluster* 1, yang diberi label Rekreasi dan Sosial, terdiri atas 16 provinsi dengan dominasi aktivitas wisata yang berfokus pada tujuan liburan, kunjungan sosial, dan mudik. *Cluster* 2, dinamakan Perjalanan Profesional yang mencakup 14 provinsi menunjukkan kecenderungan pada perjalanan dengan motif bisnis, pelatihan, serta kegiatan formal lainnya. Sementara itu, *Cluster* 3 dengan jumlah anggota 8 provinsi diberi nama Wisata Multifungsi karena wilayah-wilayah ini menunjukkan keragaman yang relatif seimbang dalam berbagai jenis aktivitas wisata. Informasi lengkap mengenai provinsi yang termasuk dalam tiap cluster dapat dilihat pada Tabel 5.

StatMat
Jurnal Statistika dan Matematika
STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 3, December 2025
18th article, Pages: 482 – 502
DOI: https://doi.org/10.32493/sm.v7i3.49999

Tabel 5. Daftar Provinsi/Wilayah pada Tiap *Cluster*

Cluster	Nama Cluster	Provinsi/Wilayah	Jumlah Provinsi
1	Rekreasi dan Sosial	DI Yogyakarta, Gorontalo, Kalimantan Utara, Kalimantan Tengah, Kep.Riau, Jambi, Papua Selatan, Papua Barat, dan Papua Tengah.	16
2	Perjalanan Profesional	Jawa Tengah, Maluku, Papua Barat Daya, DKI Jakarta, Riau, Sumatera Selatan, Bengkulu, Lampung, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Kalimantan Barat, dan Nusa Tenggara Timur.	14
3	Wisata Multifungsi	Sumatera Utara, Jawa Timur, Banten, Jawa Barat, Kalimantan Selatan, Bali, Maluku Utara, Sulawesi Barat, Papua, Papua Pegunungan, Kep. Bangka Belitung, Kalimantan Timur, Nusa Tenggara Barat, Sumatera Barat, dan Aceh.	8

Evaluasi terhadap hasil *K-Means Clustering* dilakukan untuk mengetahui sejauh mana kualitas pengelompokan data yang telah terbentuk. Metrik evaluasi yang digunakan adalah *Silhouette Score* dan *Calinski-Harabasz Index* yang mengukur sejauh mana suatu data cocok berada dalam cluster-nya dan seberapa jauh jaraknya terhadap cluster lain. Pada analisis ini diperoleh nilai *Silhouette Score* sebesar 0,662, nilai ini menunjukkan bahwa struktur *cluster* yang terbentuk tergolong baik dan data cenderung berada dalam *cluster* yang tepat. Sementara itu, *Calinski-Harabasz Index* menunjukkan nilai sebesar 549,875 yang memperkuat kualitas hasil *clustering*, karena nilai ini merefleksikan pemisahan *cluster* yang jelas. Secara keseluruhan, metrik ini menjelaskan bahwa hasil *clustering* yang terbentuk cukup optimal dan representatif.

Agglomerative Hierarchical Clustering

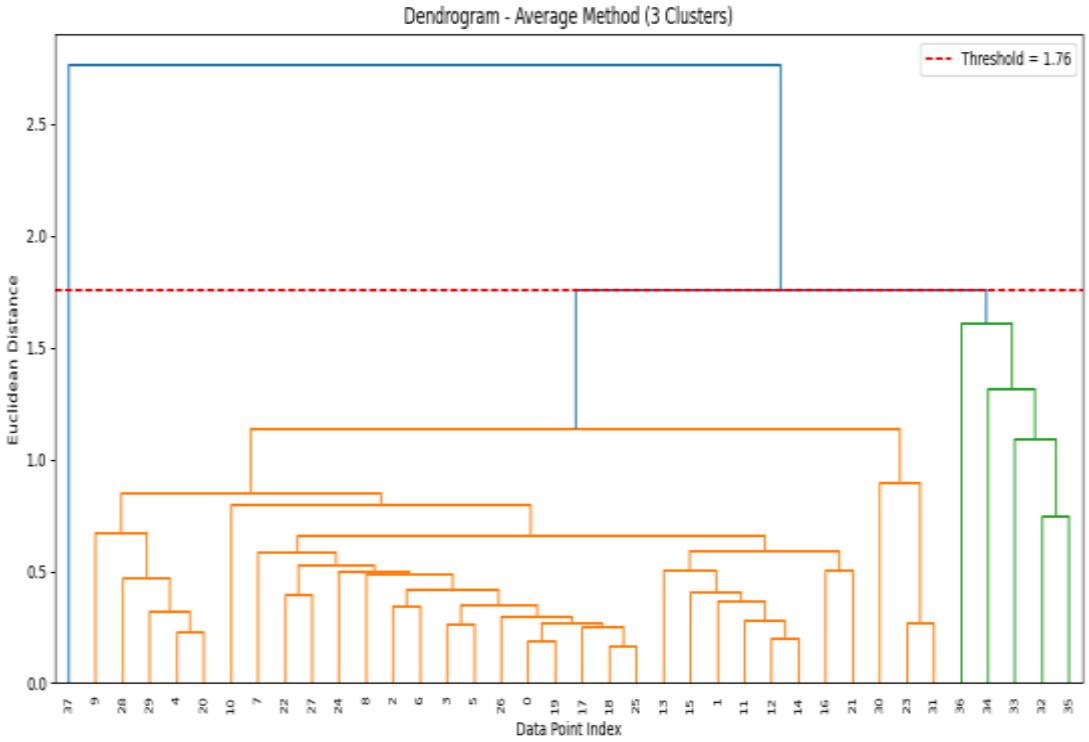
Tahap *clustering* dengan pendekatan *Agglomerative Hierarchical* dilakukan dengan mengukur jarak menggunakan metrik *Euclidean* dan tingkat kemiripan data dihitung menggunakan metode *single linkage, average linkage, complete linkage,* serta *ward method*. Kinerja masing-masing metode kemudian dievaluasi dengan menghitung korelasi antara *cophenetic distance* dan jarak asli yang hasilnya disajikan pada Tabel 6.

Tabel 6. Evaluasi Metode *Agglomerative Hierarchical* dengan *Cophenetic Distance*

Jenis Metode	Korelasi *Cophenetic Distance* dengan Jarak Sebenarnya
Single Linkage	0,9451
Average Linkage	0,9535
Complete Linkage	0,9206
Ward	0,8458

Penentuan jumlah *cluster* optimum metode ini dilakukan dengan menggunakan teknik pemotongan dendogram berdasarkan *threshold* untuk menghasilkan 3 *cluster*. Dari beberapa metode *linkage* yang diuji, hierarchical clustering dengan metode *average linkage* dipilih karena memiliki nilai korelasi *cophenetic* tertinggi seperti yang ditunjukkan pada Tabel 6. Oleh karena itu, *clustering* yang digunakan adalah *average linkage* dengan jumlah *cluster* sebanyak 3 kelompok. Berikut merupakan hasil visualisasi *cluster* dengan dendogram yang disajikan pada Gambar 5.

StatMat
Jurnal Statistika dan Matematika
STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 3, December 2025
18th article, Pages: 482 – 502
DOI: https://doi.org/10.32493/sm.v7i3.49999



Gambar 5. Visualisasi Hasil *Cluster* Metode *Average Linkage* dengan Dendogram

Berdasarkan dendogram yang disajikan pada Gambar 5, hasil pengelompokan hierarkis menghasilkan tiga *cluster* utama dengan jumlah anggota yang berbeda. *Cluster* 1 terdiri dari 32 provinsi, *Cluster* 2 terdiri dari 5 provinsi, dan *Cluster* 3 mencakup 1 provinsi. Masing-masing *cluster* telah diberi label sesuai dengan karakteristik dominan dari aktivitas perjalanan wisata yang tercermin pada data. Informasi lengkap mengenai provinsi yang termasuk ke dalam masing-masing *cluster* disajikan pada Tabel 7.

Tabel 7 Karakteristik Tiap *Cluster* yang Berdasarkan Nilai Rata-Ratanya

Fitur	***Cluster* 1**	***Cluster* 2**	***Cluster* 3**
Jumlah Wisatawan	31796190	645306	379292
Laki-Laki	60,12344	60,93	0
Perempuan	39,87656	39,07	0
Berlibur	30,63719	11,058	0
Kesehatan dan Kecantikan	2,926563	0	0
Keagamaan	1,792188	0	0
Mengunjungi Teman	21,02469	19,35	0
Mudik	5,527812	0,906	0

StatMat
Jurnal Statistika dan Matematika
STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 3, December 2025
18th article, Pages: 482 – 502
DOI: https://doi.org/10.32493/sm.v7i3.49999

Olahraga	0,898438	0	0
Belanja	2,896875	0	0
Bisnis	16,38531	31,482	0
Mice	5,8325	6,014	0
Pelatihan	5,46375	6,21	0
Tujuan Lainnya	5,141875	3,638	0

Berdasarkan Tabel 7, diketahui bahwa *Cluster* 1 memiliki karakteristik fitur Jumlah Perjalanan Wisatawan tertinggi dengan proporsi Laki-Laki sedikit lebih besar daripada Perempuan. Tujuan utama perjalanan dalam *cluster* ini adalah untuk berlibur dengan persentase yang dominan, diikuti oleh aktivitas mengunjungi teman, bisnis, serta motif lain seperti mudik, belanja, serta kesehatan dan kecantikan yang juga menunjukkan kontribusi signifikan. Sementara itu, *Cluster* 2 ditandai oleh Jumlah Perjalanan Wisatawan yang rendah dibandingkan *cluster* lainnya, namun aktivitas yang tercatat cukup beragam. Perjalanan dalam cluster ini didominasi oleh bisnis, pelatihan, dan MICE, dengan proporsi Laki-Laki lebih tinggi daripada Perempuan. Aktivitas berlibur, mudik, dan mengunjungi teman masih terlihat namun dalam persentase yang jauh lebih rendah dibandingkan *Cluster* 1.

Sementara itu, *Cluster* 3 menunjukkan tidak adanya aktivitas perjalanan wisata yang tercatat, dengan seluruh nilai variabel perjalanan bernilai nol. Hal ini menjelaskan adanya kemungkinan rendahnya mobilitas wisatawan di wilayah yang termasuk dalam *cluster* ini. Untuk memperoleh pemahaman yang lebih mendalam terhadap hasil klasterisasi, visualisasi data dilakukan melalui peta tematik yang menggambarkan sebaran masing-masing *cluster* dalam ruang dua dimensi. Penyajian tersebut dapat dilihat pada Gambar 3.5 yang menampilkan representasi visual hasil pengelompokan wilayah.

Gambar 6. Peta Tematik Hasil Segmentasi Wilayah di Indonesia Berdasarkan Profil Perjalanan Tahun 2024 dengan Metode *Hierarchical Average Linkage*

StatMat
Jurnal Statistika dan Matematika
STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 3, December 2025
18th article, Pages: 482 – 502
DOI: https://doi.org/10.32493/sm.v7i3.49999

Berdasarkan visualisasi dan hasil perhitungan yang telah dilakukan, segmentasi wilayah menggunakan metode *Average Linkage* menghasilkan tiga *cluster* utama. *Cluster* 1 diberi label Wisata Multifungsi yang mencakup 32 provinsi dan menggambarkan wilayah dengan keberagaman motif perjalanan wisata, mulai dari liburan hingga kegiatan sosial dan profesional. *Cluster* 2 diberikan nama Perjalanan Profesional yang terdiri dari 5 provinsi dengan dominasi aktivitas perjalanan dengan tujuan bisnis, pelatihan, dan MICE. Sementara itu, *Cluster* 3 yang disebut Non-aktif hanya mencakup 1 provinsi, yaitu Papua Pegunungan dan menunjukkan minimnya aktivitas wisata berdasarkan data yang tersedia. Daftar lengkap provinsi yang termasuk dalam tiap *cluster* disajikan pada tabel 8.

Tabel 8. Daftar Wilayah/Provinsi Pada Tiap *Cluster*.

Cluster	**Nama *Cluster***	**Kabupaten/Kota**	**Jumlah Provinsi**
1	Wisata Multifungsi	Aceh, Sumatera Utara, Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, Lampung, Kep. Bangka Belitung, Kep.Riau, DKI Jakarta, Jawa Barat, Jawa Tengah, DI Yogyakarta, Jawa Timur, Banten, Bali, Nusa Tenggara Barat, Nusa Tenggara Timur, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku, Papua.	32
2	Perjalanan Profesional	Maluku Utara, Papua Barat, Papua Barat Daya, Papua Selatan, Papua Tengah.	5
3	Non-aktif	Papua Pegunungan	1

Evaluasi terhadap hasil segmentasi menggunakan metode *Hierarchical Clustering,* yaitu *average linkage* dilakukan untuk menilai kualitas pengelompokan data yang telah terbentuk. Metrik evaluasi yang digunakan adalah *Silhouette Score* dan *Calinski-Harabasz Index*. Pada analisis ini diperoleh nilai *Silhouette Score* sebesar 0,507 yang menunjukkan bahwa struktur *cluster* yang terbentuk berada dalam kategori cukup baik dengan sebagian besar data berada pada *cluster* yang tepat, meskipun masih terdapat beberapa data yang tidak sepenuhnya terpisah secara optimal. Sementara itu, nilai *Calinski-Harabasz Index* sebesar 21,111 menjelaskan adanya pemisahan *cluster* yang cukup jelas, meskipun tidak sekuat hasil *clustering* sebelumnya dengan metode *K-Means*. Secara keseluruhan, hasil evaluasi ini menunjukkan bahwa pengelompokan data dengan pendekatan *average linkage* cukup representatif, meskipun kualitasnya masih dapat ditingkatkan jika dibandingkan dengan metode sebelumnya, terutama dalam hal pemisahan antar cluster.

4. KESIMPULAN

Berdasarkan hasil analisis clustering terhadap data karakteristik perjalanan wisatawan domestik antar provinsi di Indonesia tahun 2024 menggunakan dua metode, yaitu *K-Means* dan *Agglomerative Hierarchical Clustering* (AHC), proses yang dimulai dari tahapan *preprocessing* hingga pembentukan klaster menghasilkan struktur pengelompokan yang berbeda, baik dari segi jumlah provinsi dalam setiap cluster maupun karakteristik dominan tiap kelompok.

StatMat
Jurnal Statistika dan Matematika
STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 3, December 2025
18th article, Pages: 482 – 502
DOI: https://doi.org/10.32493/sm.v7i3.49999

Metode *K-Means* menunjukkan kecenderungan membentuk cluster dengan distribusi anggota relatif seimbang dan lebih adaptif terhadap sebaran data, sehingga lebih mampu mengidentifikasi pola aktivitas wisata secara eksplisit. Di sisi lain, metode *Agglomerative Hierarchical Clustering* khususnya dengan pendekatan *average linkage* menghasilkan struktur klaster yang lebih hirarkis dan mampu merefleksikan kedekatan antar wilayah berdasarkan keseluruhan dimensi. Dengan demikian, pemilihan metode clustering sangat bergantung pada kebutuhan analisis. *K-Means* lebih unggul dalam efisiensi komputasi dan interpretabilitas jika jumlah cluster telah ditentukan sebelumnya, sedangkan metode AHC menawarkan fleksibilitas dalam eksplorasi jumlah cluster melalui dendrogram. Oleh karena itu, pemanfaatan keduanya secara komplementer dapat memberikan pemahaman yang lebih menyeluruh terhadap pola segmentasi yang terbentuk.

5. REFERENSI

1. Wijaya A, Fasa H, Berliandaldo M, Andriani D, Prasetio A, Strategis BK, Film G, Indonesia P. Implikasi Penerapan Kebijakan Golden Visa dalam Rangka Mendorong Pengembangan Investasi pada Sektor Pariwisata. Jurnal Kepariwisataan [Internet]. 2023;22(2):159–175. https://doi.org/10.52352/jpar.v22i2.1117
2. Prayitno ARD, Purwantoro A, Astuti NW, Haryanto T. Analisis Produktivitas Pariwisata: Studi Kasus pada Beberapa Negara berdasarkan Perbedaan Karakter Wilayah. Jurnal Pendidikan Ekonomi (JUPE) [Internet]. 2023;11(3):304–312. https://doi.org/10.26740/jupe.v11n3.p304-312
3. Aprilia K, Sembiring F. Analisis Garis Kemiskinan Makanan Menggunakan Metode Algoritma K-Means Clustering. Seminar Nasional Sistem Informasi dan Manajemen Informatika (SISMATIK). 2021;1:1–10.
4. Nugroho MR, Hendrawan IE, Purwantoro. Penerapan Algoritma K-Means Untuk Klasterisasi Data Bbat pada Rumah Sakit ASRI. Jurnal Nuansa Informatika [Internet]. 2022;16(1):125–133. https://doi.org/10.25134/nuansa.v16i1.5294
5. Murtagh F, Contreras P. Hierarchical clustering: A survey. Wiley Interdiscip Rev Data Min Knowl Discov. 2020;10(1):e1254. doi:10.1002/widm.1254
6. Fitrayana, P., & Saputro, D. Algoritma Clustering Large Application (CLARA) untuk Menangani Data Outlier. PRISMA, Prosiding Seminar Nasional Matematika [internet]. 2022;5;721-725.
7. Sihombing PR, Suryadiningrat S, Sunarjo DA, Yuda YP. Identifikasi Data Outlier (Pencilan) dan Kenormalan Data Pada Data Univariat serta Alternatif Penyelesaiannya. Berdikari: Jurnal Ekonomi Dan Statistik Indonesia [Internet] 2023;2(3);307-316. https://doi.org/10.11594/jesi.02.03.07
8. Saputri NA. Pengaruh Media Exposure, Profitabilitas Dan Ukuran Perusahaan Terhadap Carbon Emission Disclosure. Jurnal Ilmu dan Riset Akuntansi [Internet]. 2023;12(8).
9. Allorerung, P. P., Erna, A., Bagussahrir, M., & Alam, S. Analisis Performa Normalisasi Data untuk Klasifikasi K-Nearest Neighbor pada Dataset Penyakit. JISKA (Jurnal Informatika Sunan Kalijaga) [Internet. 2024;9(3);178–191. https://doi.org/10.14421/jiska.2024.9.3.178-191
10. Nazori S, Rendra G, & Destria, A. Klasifikasi Penyakit TBC Menggunakan Metode UMAP dan K-NN. Bit-Tech [Internet] 2025;7(3);843–852. https://doi.org/10.32877/bt.v7i3.2227
11. Sulistiyawati A & Supriyanto E. Implementasi Algoritma K-Means Clustering dalam

STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 3, December 2025
18th article, Pages: 482 – 502
DOI: https://doi.org/10.32493/sm.v7i3.49999

Penentuan Siswa Kelas Unggulan. Jurnal Tekno Kompak [Internet]. 2021;15(2);25-36.

12. Pribadi WW, Yunus A, Wiguna AS. Perbandingan Metode K-Means Euclidean Distance Dan Manhattan Distance Pada Penentuan Zonasi Covid-19 Di Kabupaten Malang. JATI (Jurnal Mahasiswa Teknik Informatika) [Internet] 2022;6(2);493-500. https://doi.org/10.36040/jati.v6i2.4808
13. Supardi R & Kanedi I. Implementasi Metode Algoritma K-Means Clustering Pada Toko Eidelweis. Jurnal Teknologi Informasi [Internet]. 2020;4(2);270-277
14. Harani NH, Prianto C, Nugraha FA. Segmentasi Pelanggan Produk Digital Service Indihome Menggunakan Algoritma K-Means Berbasis Python. Jurnal Manajemen Informatika (JAMIKA) [Internet]. 2022;10;2;133-146. https://doi.org/10.34010/jamika.v10i2.2683
15. Riani AP, Voutama A, Ridwan T. Penerapan K-Means Clustering Dalam Pengelompokan Hasil Belajar Peserta Didik Dengan Metode Elbow. Jurnal Teknologi Sistem Informasi dan Sistem Komputer TGD [Internet]. 2023;6(1);164-172. https://doi.org/10.53513/jsk.v6i1.7351
16. Nellie V, Mawardi VC, Perdana NJ. IMPLEMENTASI METODE AGGLOMERATIVE HIERARCHICAL CLUSTERING UNTUK SISTEM REKOMENDASI FILM. Jurnal Ilmu Komputer dan Sistem Informasi [Internet].
17. Matdoan, M. Y., & Noya Van Delsen, M. S. Penerapan Analisis Cluster Dengan Metode Hierarki Untuk Klasifikasi Kabupaten/Kota Di Provinsi Maluku Berdasarkan Indikator Indeks Pembangunan Manusia. Statmat : Jurnal Statistika Dan Matematika [Internet] 2020;2(2);123–130. https://doi.org/10.32493/sm.v2i2.4740
18. Drl, I. R., Chrisnanto, Y. H., & Umbara, F. R. Analisis Cluster Pada Kelompok Masyarakat Yang Rentan Terhadap Paparan Covid-19 Menggunakan Metode K-Means Clustering Dan Visualisasi Dengan Sig. Informatics and Digital Expert (INDEX) [Internet] 2024;4(2);61–69. https://doi.org/10.36423/index.v4i2.885
19. Nurhaliza S & Mukhti TO. Clustering Regions in West Sumatera Based on the Special Protection Index for Children Using K-Means Clustering with Silhouette Coefficient . UNP Journal of Statistics and Data Science [Internet] 2025;3(1);123–129. https://doi.org/10.24036/ujsds/vol3-iss1/356
20. Fauziah, & Basir, C. (2024). Analisis Klaster dengan Metode K-Means Berdasarkan Usia Warga yang Divaksin Covid-19 di Kelurahan Grogol Selatan. *Jurnal Bayesian: Jurnal Ilmiah Statistika dan Ekonometrika, 4(1), 53-62*. https://bayesian.lppmbinabangsa.id/index.php/home/article/view/76
21. Permata Sari, Y., & Basir, C. (2025). Analisis Klaster dengan Metode K-Means pada Persebaran Kasus Covid-19 Berdasarkan Desa di Kecamatan Kemang-Bogor. *Pelita : Jurnal Penelitian Dan Karya Ilmiah, 24(2), 33–45*. https://doi.org/10.33592/pelita.v24i2.5504