

Evaluasi Kinerja Metode K-Means dan Hierarchical Clustering dalam Pengelompokan Data Pendidikan

Ninik Tri Hartanti^{*1}, Dina Maulina², Erni Seniwati³

¹Sistem Informasi, Universitas Amikom Yogyakarta, Indonesia

²Manajemen Informatika, Universitas Amikom Yogyakarta, Indonesia

³Teknik Informatika, Universitas Langlang Buana Bandung, Indonesia

Email: ¹ninik.t@amikom.ac.id, ²dina.m@amikom.ac.id, ³erni.seniwati@unla.ac.id

Abstrak

Seiring meningkatnya kebutuhan akan eksplorasi informasi tersembunyi dan pengenalan pola, penerapan *Educational Data Mining* dalam pengolahan data pendidikan menjadi semakin penting untuk mendukung kualitas pembelajaran. Urgensi penelitian ini didasarkan pada pentingnya pemanfaatan data pendidikan secara optimal untuk mendukung pengambilan keputusan yang lebih akurat serta pengembangan sistem pembelajaran adaptif berbasis data. Penentuan metode *clustering* yang tepat menjadi salah satu kendala utama, sebab kualitas hasil pengelompokan data berperan penting dalam menentukan akurasi interpretasi data dan keputusan dalam pendidikan. Penelitian ini bertujuan untuk mengevaluasi kinerja teknik pengelompokan data pendidikan dengan memanfaatkan metode K-Means dan Hierarchical Clustering. Dataset yang digunakan berasal dari data pendidikan publik yang mencakup atribut performa akademik dan aktivitas belajar siswa. Uraian langkah dalam penelitian meliputi preprocessing data, normalisasi data, penerapan algoritma clustering, kemudian evaluasi hasil cluster menggunakan *Silhouette Score* dan *Davies-Bouldin Index*. Temuan ini mengimplikasikan bahwa metode K-Means lebih direkomendasikan untuk analisis data pendidikan, khususnya dalam mendukung pengelompokan siswa dan perumusan strategi pembelajaran yang lebih efektif tanpa perlu mengacu pada detail nilai statistik secara berulang.

Kata kunci: *Davies-Bouldin Index, Educational Data Mining, Hierarchical Clustering, K-Means Clustering, Silhouette Score.*

Performance Analysis of K-Means and Hierarchical Clustering in Educational Data Clustering

Abstract

As the need for hidden information exploration and pattern recognition increases, the application of Educational Data Mining in educational data processing is becoming increasingly important to support the quality of learning. The urgency of this research is based on the importance of optimally utilizing educational data to support more accurate decision-making and the development of data-driven adaptive learning systems. Determining the appropriate clustering method is a major obstacle, because the quality of the data clustering results plays a crucial role in determining the accuracy of data interpretation and decisions in education. This study aims to evaluate the performance of educational data clustering techniques using the K-Means and Hierarchical Clustering methods. The dataset used comes from public education data that includes academic performance attributes and student learning activities. The steps in the study include data preprocessing, data normalization, application of the clustering algorithm, and evaluation of the cluster results using the Silhouette Score and Davies-Bouldin Index. These findings imply that the K-Means method is more recommended for educational data analysis, particularly in supporting student clustering and the formulation of more effective learning strategies without the need for repeated reference to detailed statistical values.

Keywords: *Davies-Bouldin Index, Educational Data Mining, Hierarchical Clustering, K-Means Clustering, Silhouette Score.*

1. PENDAHULUAN

Dunia pendidikan semakin berkembang pesat seiring dengan peningkatan volume data pendidikan, seperti data performa akademik, aktivitas belajar siswa, informasi fasilitas pendidikan, yang belum sepenuhnya

dimanfaatkan secara optimal dalam mendukung pengambilan keputusan di bidang pendidikan. Secara logika pemanfaatan data yang tepat dapat memberikan informasi penting terkait pola belajar siswa dan kualitas proses pembelajaran, sehingga dapat memberikan peluang besar dalam pengembangan sistem pembelajaran berbasis data. Oleh karena itu diperlukan sebuah pendekatan yang mampu untuk mengolah dan menganalisis data pendidikan secara efektif agar kualitas pendidikan menjadi lebih baik. Salah satu pendekatan yang relevan adalah Educational Data Mining (EDM), yang berfokus pada penerapan teknik data mining untuk menemukan pola tersembunyi dalam data pendidikan [1].

Penerapan EDM memungkinkan pengembangan sistem pembelajaran adaptif yang mampu menyesuaikan materi, metode, dan kecepatan belajar sesuai karakteristik siswa. Hal ini menjadi penting karena model pembelajaran tradisional cenderung bersifat seragam dan belum mampu mengakomodasi perbedaan individu siswa. Oleh karena itu, analisis data pendidikan menjadi kunci dalam mendukung transformasi menuju pembelajaran yang lebih adaptif dan berbasis data [2].

Data pendidikan publik merupakan salah satu penanda dari pesatnya perkembangan data pendidikan, sehingga dengan melalui data pendidikan tersebut akan dapat diperoleh pola pembelajaran siswa terkait pola perilaku belajar siswa, memprediksi performa siswa, mengelompokkan pola belajar siswa, dan mengidentifikasi siswa berisiko. Oleh karena itu, data yang digunakan dalam penelitian ini berdasarkan data pendidikan publik yang berasal dari <http://www.kaggle.com> dengan kata kunci *student performance*. Beragam instansi atau lembaga pendidikan saling bersaing dalam upaya untuk memajukan sistem pendidikan yang ditonjolkan, dengan harapan sistem pendidikan instansi tersebut akan dapat diterima masyarakat luas walaupun sistem pembelajaran masih belum bisa dikategorikan sebagai pembelajaran adaptif. Dalam proses pembelajaran diperlukan model pendekatan dalam sistem belajar, salah satunya adalah model pembelajaran adaptif. Terlepas dari model pembelajaran adaptif, terdapat model pembelajaran tradisional yang cenderung merata namun belum mampu mengakomodasi perbedaan karakteristik individu siswa, sehingga diperlukan pendekatan yang lebih adaptif dimana sistem pembelajaran dapat menyesuaikan materi, kecepatan, dan metode berdasarkan kebutuhan masing-masing siswa [3].

Berbagai penelitian sebelumnya menunjukkan bahwa metode clustering efektif dalam mengelompokkan data di berbagai bidang, termasuk pendidikan. Namun, jika ditinjau lebih lanjut berdasarkan penelitian yang telah dilakukan, masih terdapat beberapa keterbatasan yang belum teratasi.

Penelitian terkait pengelompokan mahasiswa [4] hanya berfokus pada penerapan clustering tanpa melakukan evaluasi kinerja algoritma secara komprehensif. Sementara itu, penelitian yang membandingkan K-Means dan Hierarchical Clustering pada data pengangguran [5] menunjukkan hasil kualitas cluster yang relatif sama, namun evaluasi yang digunakan masih terbatas pada satu metrik sehingga belum mampu menggambarkan kualitas cluster secara menyeluruh. Penelitian lain pada data log e-learning mampu mengelompokkan pola perilaku belajar mahasiswa menjadi 3 cluster optimal [6] hanya menggunakan metode K-Means tanpa melakukan perbandingan dengan metode clustering lainnya, sehingga belum dapat menunjukkan metode yang paling optimal.

Selain itu, penelitian yang membandingkan beberapa algoritma clustering seperti K-Means, GMM, dan BIRCH [7] menunjukkan bahwa performa algoritma sangat bergantung pada metrik evaluasi yang digunakan, dimana setiap metode memiliki keunggulan pada metrik tertentu. Hal ini dapat menegaskan bahwa performa algoritma clustering sangat bergantung pada karakteristik dataset yang digunakan, dan mengindikasikan bahwa penggunaan satu metrik saja tidak cukup untuk menilai kualitas clustering secara komprehensif.

Penelitian lain dalam konteks pendidikan untuk menganalisis dan mengelompokkan performa siswa berdasarkan data pendidikan menunjukkan hasil bahwa faktor jenis kelamin dan usia dapat diidentifikasi sebagai faktor yang mempengaruhi performa siswa [8] juga masih terbatas pada penggunaan satu metode clustering tanpa analisis komparatif. Sementara itu, penelitian pada data sosial salah satunya yakni kemiskinan, menunjukkan bahwa K-Means lebih unggul dalam efisiensi dan kecepatan, sedangkan *Hierarchical Clustering* lebih baik dalam mengungkap struktur data yang kompleks [9] dan penggunaan K-Means dalam bidang penelitian yang berfokus pada evaluasi kualitas clustering dalam konteks data pendidikan menggunakan metrik validitas cluster [10] telah membandingkan beberapa metode clustering, namun konteksnya belum secara spesifik pada data pendidikan.

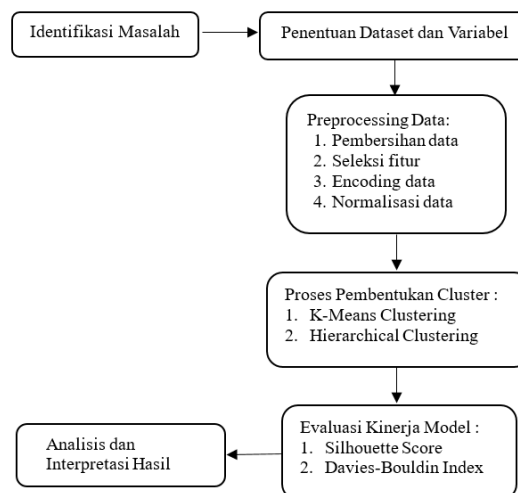
Metode K-Means dan *Hierarchical Clustering* telah banyak digunakan dalam berbagai penelitian, termasuk pada data pendidikan, sosial, dan kesehatan. Beberapa penelitian menunjukkan bahwa kedua metode tersebut mampu menghasilkan pengelompokan data yang baik, namun sebagian besar penelitian masih memiliki keterbatasan [11].

Berdasarkan uraian tersebut, terdapat beberapa *research gap* yang dapat diidentifikasi. Pertama, masih terbatasnya penelitian yang secara khusus membandingkan metode K-Means dan Hierarchical Clustering pada data pendidikan. Kedua, sebagian besar penelitian sebelumnya belum melakukan evaluasi kinerja clustering secara komprehensif dengan menggunakan lebih dari satu metrik validitas cluster. Ketiga, belum adanya kajian yang secara eksplisit mengaitkan hasil evaluasi metrik dengan pemilihan metode clustering yang paling sesuai untuk mendukung pengembangan sistem pembelajaran berbasis data.

Oleh karena itu, penelitian ini dilakukan untuk mengisi gap tersebut dengan melakukan evaluasi komprehensif terhadap kinerja metode K-Means dan Hierarchical Clustering pada data pendidikan publik menggunakan beberapa metrik validitas cluster, sehingga dapat memberikan dasar yang lebih kuat dalam pemilihan metode clustering yang optimal.

2. METODE PENELITIAN

Metode penelitian yang diterapkan merupakan metode eksperimen komparatif yang difokuskan untuk mengevaluasi kinerja algoritma Clustering dalam pengelompokan data pendidikan. Tahapan dalam penelitian terlampir pada Gambar 1.



Gambar 1. Langkah Penelitian

2.1. Identifikasi Masalah

Fokus permasalahan dalam penelitian ini terletak pada bagaimana melakukan pengelompokan data pendidikan dalam beberapa *cluster* optimal, dengan menggunakan 2 metode clustering yaitu K-Means dan *Hierarchical Clustering* sehingga melalui perbandingan hasil cluster diantara kedua metode tersebut, akan diperoleh hasil clustering optimal.

2.2. Penentuan Dataset dan Variabel

Informasi yang digunakan adalah data yang merepresentasikan performa akademik siswa, yang meliputi informasi terkait karakteristik siswa, kebiasaan belajar, serta hasil akademik siswa. Penelitian ini memanfaatkan data yang berasal dari publik melalui <http://www.kaggle.com> tentang *student performance*.

2.3. Preprocessing Data

Pada proses ini terdiri dari 4 langkah, yaitu pembersihan data (*data cleaning*), seleksi data (*feature selection*), transformasi data, normalisasi data, dan deteksi outlier [12].

2.3.1. Pembersihan Data

Pada tahap ini adalah tahap penting dalam preprocessing data sebelum proses analisis clustering. Tujuan dari pembersihan data adalah memperbaiki kualitas data sehingga data konsisten dan siap untuk dilakukan analisis. Langkah yang diterapkan adalah menghapus duplikasi data, memperbaiki skala *attendance*, dan validasi data.

2.3.2. Seleksi fitur

Proses seleksi fitur dimaksudkan untuk menentukan atribut yang paling relevan dari dataset yang digunakan untuk menghindari redundansi dan meningkatkan kualitas clustering. Berdasarkan dataset pendidikan publik <http://www.kaggle.com> *student performance* fitur terbaik yang digunakan adalah *overall_score* sebagai representasi performa akademik yang berisi nilai rata-rata dari beberapa mata pelajaran. Kolom *study_hours* dan

attendance_percentage digunakan sebagai indikator keterlibatan belajar, sedangkan *study_method* digunakan sebagai faktor pendukung yang mempengaruhi pola belajar siswa.

2.3.3. Encoding Data

Berdasarkan dataset yang telah diseleksi fiturnya, maka langkah selanjutnya adalah proses encoding data, yaitu mengubah dataset yang bertipe kategorikal ke numerik agar dapat diproses oleh algoritma clustering. Atribut yang dilakukan encoding adalah *study_method*.

2.3.4. Normalisasi Data

Setelah dilakukan proses pembersihan data, seleksi fitur dan encoding data, maka langkah selanjutnya adalah normalisasi data. Berdasarkan dataset yang digunakan, atribut *attendance_percentage* perlu dilakukan normalisasi data, sehingga nilai berkisar 1-100, sedangkan atribut *study_method* menggunakan angka 1 sampai angka 6. Proses normalisasi data dilakukan dengan menerapkan metode Min-Max Normalization sesuai dengan rumusan pada persamaan (1).

$$X' = (X - X_{min}) / (X_{maks} - X_{min}) \quad (1)$$

Keterangan

- X' : nilai setelah dinormalisasi
- X : nilai asli yang akan dihitung
- X_{min} : nilai minimal yang ada dalam dataset
- X_{maks} : nilai maksimal yang ada dalam dataset

2.4. Clustering

2.4.1. Algoritma K-Means

Penerapan clustering K-Means tidak luput dari metode untuk menentukan jumlah clusternya. Salah satu metode yang digunakan adalah metode Elbow yang diimplementasikan dalam menentukan banyaknya cluster terbaik untuk penerapan K-Means. Pada tahap ini terdapat perbandingan nilai untuk nilai k yang telah diuji menggunakan SSE (*Sum of Square Errors*), sehingga akan dihasilkan grafik berbentuk siku di titik tertentu. Semakin kecil nilai SSE maka semakin dekat cluster yang terbentuk. SSE (*Sum of Squared Errors*) merupakan jumlah total kuadrat jarak data terhadap pusat cluster pada masing-masing cluster [13].

Dalam metode Elbow, jumlah cluster yang optimal ditentukan dengan mengamati titik “siku” pada grafik SSE, adanya titik perubahan pada grafik SSE, setelah titik tersebut penurunan nilai SSE menjadi tidak terlalu signifikan meskipun jumlah cluster ditambah. Selain itu, dalam implementasi algoritma K-Means digunakan parameter *random_state* untuk mengontrol inisialisasi centroid secara acak. Parameter ini bertujuan untuk memastikan konsistensi hasil clustering sehingga eksperimen dapat direproduksi dengan hasil yang sama.

Metode Elbow menggunakan nilai SSE sebagai dasar dalam menentukan jumlah cluster optimal, dengan melihat pola penurunan SSE saat jumlah cluster bertambah. Perhitungan nilai SSE [14] pada metode Elbow menggunakan persamaan (2).

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

Keterangan :

- k : banyak cluster
- C_i : himpunan data pada cluster ke- i
- x : data dalam cluster
- μ_i : centroid (titik pusat) cluster ke- i
- $\|x - \mu_i\|^2$: jarak Euclidean kuadrat antara data dan centroid

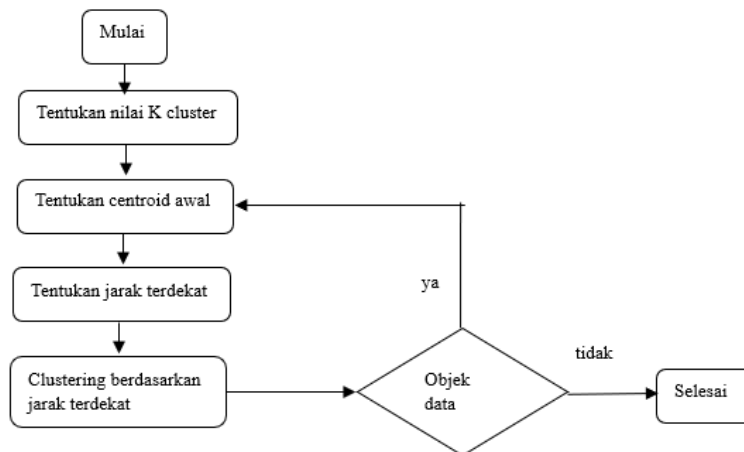
Hasil dari proses perhitungan dengan metode Elbow akan diperoleh nilai yang akan digunakan sebagai jumlah cluster optimal dan selanjutnya akan ditetapkan sebagai nilai k dalam perhitungan K-Means clustering. Tahapan dalam proses clustering K-Means dilakukan melalui beberapa langkah sebagai berikut [15]:

1. Penentuan nilai awal cluster

Pada langkah ini dilakukan penentuan nilai K sebagai nilai centroid awal secara acak dari dataset

2. Penentuan jarak terdekat.
Langkah ini dilakukan untuk setiap data, yang dialokasikan ke cluster sesuai dengan nilai centroid yang terdekat menurut perhitungan jarak *Euclidean Distance* di persamaan (3).
3. Proses Clustering
Menghitung ulang centroid sebagai rerata data dalam cluster.

Sesuai dengan alur di atas, Gambar 2 merupakan alur proses perhitungan clustering dengan metode K-Means.



Gambar 2. Alur Proses K-Means Clustering

Proses perhitungan jarak dengan *Euclidean Distance* menggunakan persamaan (3).

$$d(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \dots + (a_n - b_n)^2} \quad (3)$$

Keterangan

- $d(a,b)$: jarak antara titik a dengan titik b
- a : nilai titik a
- b : nilai titik b

2.4.2. Algoritma Hierarchical Clustering

Dalam data mining, *Hierarchical Clustering* digunakan sebagai metode untuk mengelompokkan data berdasarkan kesamaan karakteristik antar objek, tanpa perlu adanya penentuan jumlah cluster di awal. Struktur hierarki yang dihasilkan metode ini direpresentasikan dalam bentuk dendrogram, yakni diagram pohon yang memperlihatkan proses penggabungan atau pemisahan data[16] Tahapan dalam proses clustering dengan metode *Hierarchical Clustering* adalah :

1. Inisialisasi Cluster
Setiap objek data pendidikan dianggap sebagai satu cluster awal, sehingga jika terdapat n data, maka akan terbentuk n cluster awal.
2. Penentuan jarak
Metode *Euclidean Distance* berfungsi untuk menghitung jarak antar data sebagai indikator kemiripan. Metode tersebut menghitung jarak antar objek berdasarkan atribut numerik, dan data dianggap memiliki kemiripan yang lebih tinggi apabila jarak objek lebih kecil.
3. Penggabungan Cluster
Cluster yang memiliki jarak terdekat akan dikombinasikan menjadi satu cluster baru. Tahapan ini dilakukan secara iteratif hingga terbentuk satu cluster besar atau hingga jumlah cluster yang diinginkan tercapai.
4. Penentuan Metode Linkage
Metode linkage digunakan untuk menentukan jarak antar cluster dalam proses penggabungan cluster, adapun macam-macam metode linkage:

- a) Single Linkage : perhitungan berdasarkan jarak terdekat antar anggota cluster
 - b) Complete Linkage : perhitungan berdasarkan jarak terjauh
 - c) Average Linkage : perhitungan berdasarkan rata-rata jarak
 - d) Ward Method : untuk minimasi variansi
5. Pembentukan Dendrogram
Dendrogram merupakan visualisasi dari hasil clustering, ibarat pohon yang mempunyai hubungan data yang semakin dekat cabang maka semakin mirip datanya.
 6. Penentuan jumlah Cluster optimal
Penentuan dilakukan melalui analisis dendrogram.

Metode yang digunakan pada penelitian ini adalah metode Ward, yakni pendekatan penggabungan cluster dengan meminimalkan peningkatan variansi dalam cluster (*within-cluster variance*). Secara matematis, metode ini memilih pasangan cluster yang menghasilkan kenaikan jumlah kuadrat error terkecil setelah penggabungan. Adapun keunggulan metode Ward adalah kemampuannya menghasilkan cluster yang lebih kompak dan homogen dibandingkan metode linkage lainnya, seperti single atau complete linkage. Oleh karena itu, metode ini sangat sesuai digunakan pada data numerik yang telah melalui proses normalisasi.

Dendrogram merupakan representasi visual dari proses penggabungan cluster, di mana sumbu vertikal menunjukkan jarak atau tingkat ketidakmiripan antar cluster. Jumlah cluster ditentukan dengan memotong dendrogram pada titik dengan jarak vertikal terbesar (*longest vertical distance*) yang tidak memotong banyak cabang. Titik ini menunjukkan adanya perbedaan yang signifikan antar cluster sehingga dianggap sebagai batas pemisahan cluster yang optimal.

Pendekatan ini memungkinkan penentuan jumlah cluster dilakukan secara lebih objektif berdasarkan struktur data yang terbentuk, tanpa harus ditentukan di awal seperti pada metode K-Means. Dengan demikian, proses Hierarchical Clustering dalam penelitian ini tidak hanya mempertimbangkan kedekatan jarak antar data, tetapi juga struktur hierarki yang terbentuk, sehingga menghasilkan pengelompokan yang lebih representatif terhadap pola data.

2.5. Evaluasi Kerja

Hasil dari semua proses clustering akan dievaluasi di tahap ini. Metode clustering merupakan metode unsupervised learning yang tidak memiliki label acuan, oleh karena itu untuk proses evaluasi kinerja model clustering diperlukan pendekatan penilaian clustering dengan metrik Silhouette Score dan Davies-Bouldin Index (DBI). Persamaan untuk proses perhitungan dengan *Silhouette Score* adalah persamaan (4).

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (4)$$

Keterangan

- $a(i)$: rata-rata jarak antara satu data (ke-i) dengan data lain dalam cluster yang sama
- $b(i)$: nilai rata-rata jarak data ke-i ke cluster terdekat

Persamaan untuk proses perhitungan *Davies-Bouldin Index (DBI)* adalah di persamaan (5).

$$DBI = \frac{1}{k} \sum_{i=1}^k \max \left(\frac{S_i + S_j}{M_{ij}} \right) \quad (5)$$

Keterangan

- k : banyaknya cluster
- S_i : nilai rata-rata jarak data pada cluster ke-i terhadap centroidnya
- S_j : nilai rata-rata jarak pada cluster ke-j
- M_{ij} : jarak antara pusat cluster ke-i dengan pusat cluster ke-j

2.6. Analisis dan Interpretasi Hasil

Berdasarkan semua hasil perhitungan dengan langkah-langkah penelitian, tahap akhir adalah analisis dan interpretasi hasil pengelompokan menggunakan algoritma K-Means dan Hierarchical Clustering agar hasilnya dapat dimanfaatkan secara optimal sesuai kebutuhan.

3. HASIL DAN PEMBAHASAN

3.1. Dataset yang digunakan

Berdasarkan sumber publik melalui <http://www.kaggle.com> tentang *student performance*, atribut dalam dataset terdiri atas aspek demografis, aktivitas belajar, perolehan nilai siswa pada mata pelajaran Matematika, Sains, dan Bahasa Inggris. Terdapat indikator pendukung seperti tingkat kehadiran siswa dan durasi waktu belajar yang digunakan untuk menganalisis pola performa siswa. Setiap data merepresentasikan satu siswa dengan sejumlah atribut numerik yang digunakan untuk proses clustering untuk identifikasi pola kemiripan antar siswa.

Terdapat sumber data yakni sebanyak 500 data siswa dengan 16 kolom yaitu *student_id*, *age*, *gender*, *school_type*, *parent_education*, *study_hours*, *attendance_percentage*, *internet_access*, *travel_time*, *extra_activities*, *study_method*, *math_score*, *science_score*, *english_score*, *overall_score*, dan *final_grade*. Pada proses clustering akan menggunakan data yang bertipe numerik, oleh karena itu tidak semua kolom dalam dataset akan digunakan. Kolom yang akan digunakan dalam proses clustering berdasarkan dataset tersebut adalah *age*, *overall_score*, *study_hours*, *attendance_percentage*, *study_method*, *internet_access*.

3.2. Processing Data

Hasil preprocessing data yang telah dilakukan adalah terdapat 500 data dengan 4 atribut dalam dataset yaitu *overall_score*, *study_hours*, *attendance_percentage*, dan *study_method*. Atribut yang dilakukan encoding adalah *study_method* dengan penomoran angka sesuai tabel 1.

Tabel 1. Encoding data untuk atribut *study_method*

angka	data
1	notes
2	textbook
3	group study
4	coaching
5	mixed
6	online video

Setelah dilakukan encoding data, proses selanjutnya adalah normalisasi data dengan menerapkan persamaan (1) sehingga menghasilkan data pada tabel 2.

Tabel 2. Hasil Normalisasi Data

Overall_score	Study_hours	Attendance_percentage	Study_method
0.39	0.35	0.69	0.00
0.50	0.43	0.76	0.20
0.87	0.99	0.31	0.00
0.25	0.08	0.16	0.00
0.04	0.11	0.22	0.40
0.53	0.44	0.39	0.60
...	...	...	...
...	...	...	...
...	...	...	...
0.42	0.33	0.72	0.80
0.81	0.77	0.67	0.00
0.62	0.65	0.75	0.00
0.40	0.68	0.55	0.80

3.3. Proses Clustering dengan K-Means

Pengelompokan data menggunakan K-Means didasarkan pada hasil perhitungan yang mengikuti persamaan (2) menghitung nilai SSE (*Sum of Square Errors*). Hasil nilai SSE ada pada tabel 3.

Tabel 3. Nilai SSE per uji coba

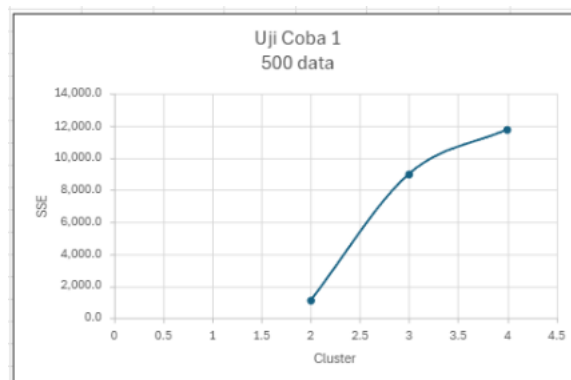
Uji coba	Jumlah data	Jumlah Cluster = 2	Jumlah Cluster = 3	Jumlah Cluster = 4
1	500	1,121.8	9,020.8	11,790.7
2	300	234.2	719.5	3,415.8
3	200	799.3	3,426.1	10,893.4

Tabel 3 menunjukkan bahwa dilakukan uji coba sebanyak 3 kali dengan jumlah data berbeda beda, uji coba1 menggunakan 500 data, uji coba2 menggunakan 300 data, dan uji coba3 menggunakan 200 data. Setiap uji coba menghasilkan nilai SSE yaitu untuk uji coba1 dengan jumlah cluster = 2 nilai SSE = 1,121.8 untuk jumlah cluster adalah 3 nilai SSE = 9,020.8, untuk jumlah cluster = 4 nilai SSE = 11,790.7 dan untuk jumlah cluster = 5 nilai SSE = 2,156.9. Jumlah cluster optimal akan terdeteksi setelah adanya perhitungan selisih nilai SSE dalam setiap uji coba untuk setiap clusternya, sesuai pada tabel 4.

Tabel 4. Nilai Selisih SSE Uji Coba1

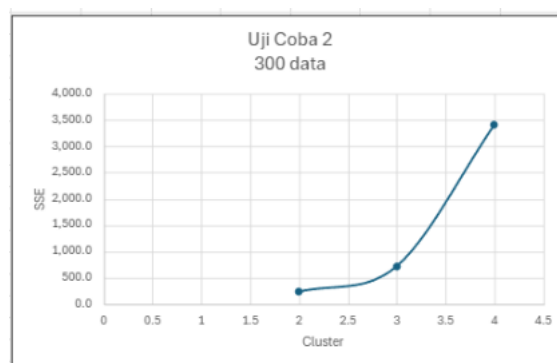
Cluster	SSE	Selisih
2	1,121.8	1,121.8
3	9,020.8	7,899.0
4	11,790.7	2,769.9

Langkah selanjutnya adalah membuat grafik berdasarkan data pada tabel 4. Grafik inilah yang akan menunjukkan nilai jumlah cluster optimal dengan memperhatikan posisi garis yang melengkung. Pada grafik uji coba1 diperoleh posisi garis melengkung berada di angka 3, sehingga sesuai nilai SSE yang diperoleh dan grafik tersebut menunjukkan bahwa jumlah cluster optimal = 3. Setiap hasil dari ujicoba akan memberikan data nilai selisih SSE dan akan membentuk grafik sesuai gambar 4.



Gambar 3. Grafik Uji Coba1

Hasil grafik pada gambar 4 menunjukkan bahwa posisi garis melengkung ada di angka 3, sehingga jumlah cluster optimal adalah 3. Perolehan jumlah cluster optimal tersebut akan dijadikan nilai k dalam proses perhitungan K-Means.



Gambar 4. Grafik Uji Coba2

Penentuan jumlah cluster (k) pada metode K-Means dilakukan menggunakan metode Elbow dengan memanfaatkan nilai Sum of Squared Errors (SSE). Metode ini bekerja dengan mengamati perubahan nilai SSE terhadap jumlah cluster, di mana titik siku (elbow point) menunjukkan jumlah cluster optimal karena setelah titik tersebut penurunan SSE tidak signifikan. Berdasarkan hasil grafik pada Gambar 3 dan Gambar 4, diperoleh bahwa nilai optimal berada pada $k = 3$.

Pada tahap proses clustering dengan K-Means dengan melibatkan bahasa python dengan kode:

```
import pandas as pd
from sklearn.cluster import KMeans
data = pd.read_excel("data_penelitian.xlsx")
data.head()

Dengan menentukan nilai k atau jumlah cluster = 3
kmeans = KMeans(n_clusters=, random_state=42)
kmeans.fit(X)

Disertai distribusi dan kategori untuk cluster
summary = data.groupby('cluster').agg({
    'overall_score': 'mean',
    'study_hours': 'mean',
    'attendance_percentage': 'mean'
}).round(2)
print(summary)
```

Hasil output ditampilkan pada tabel 5.

Tabel 5. Hasil Kmeans

Cluster	Overall score	Study hours	Attendance percentage	Study method	Jumlah	Kategori
1	0.295	0.207	0.454	0.51	198	Rendah
2	0.703	0.722	0.527	0.83	166	Tinggi
3	0.686	0.680	0.528	0.18	136	Sedang

Berdasarkan hasil clustering menggunakan metode K-Means, diperoleh 3 kelompok utama dengan karakteristik yang berbeda. Cluster 2 menunjukkan kelompok mahasiswa dengan performa akademik tinggi yang didukung oleh jam belajar yang tinggi serta metode belajar yang efektif. Cluster 3 memiliki performa yang cukup baik dengan jam belajar yang relatif tinggi, namun metode belajar yang kurang optimal. Sementara itu, Cluster 1 mahasiswa dengan performa rendah yang ditandai dengan rendahnya jam belajar dan nilai keseluruhan. Perbedaan ini menunjukkan bahwa metode belajar dan intensitas belajar memiliki pengaruh terhadap hasil akademik mahasiswa.

Pola cluster yang dihasilkan menunjukkan adanya pemisahan yang cukup jelas antar kelompok siswa berdasarkan performa akademik dan kebiasaan belajar. Hal ini mengindikasikan bahwa atribut yang digunakan memiliki kontribusi signifikan dalam membentuk struktur cluster yang kompak, sehingga mendukung efektivitas algoritma K-Means dalam mengidentifikasi pola pada data.

3.4. Proses Clustering dengan Hierarchical Clustering

Berbeda dengan K-Means, metode Hierarchical Clustering merupakan pendekatan clustering yang mengelompokkan data dalam bentuk tingkatan hierarkis. Hierarchical Clustering digunakan untuk menemukan pola dan hubungan dalam data tanpa perlu menentukan jumlah cluster di awal. Pendekatan yang dilakukan dalam penelitian ini yakni pendekatan *bottom-up* yang menggabungkan data dari individu menjadi kelompok yang lebih besar secara bertahap. Proses penggabungan cluster ditentukan menggunakan metode *linkage*.

Parameter *linkage = 'ward'* dipilih karena metode Ward bertujuan untuk meminimalkan jumlah variansi dalam cluster (within-cluster variance). Secara matematis, metode ini menggabungkan dua cluster yang menghasilkan peningkatan variansi terkecil. Oleh karena itu, metode Ward cenderung menghasilkan cluster yang lebih kompak dan homogen, terutama pada data numerik yang telah dinormalisasi. Adapun kode dengan menggunakan google colab dan bahasa python berikut:

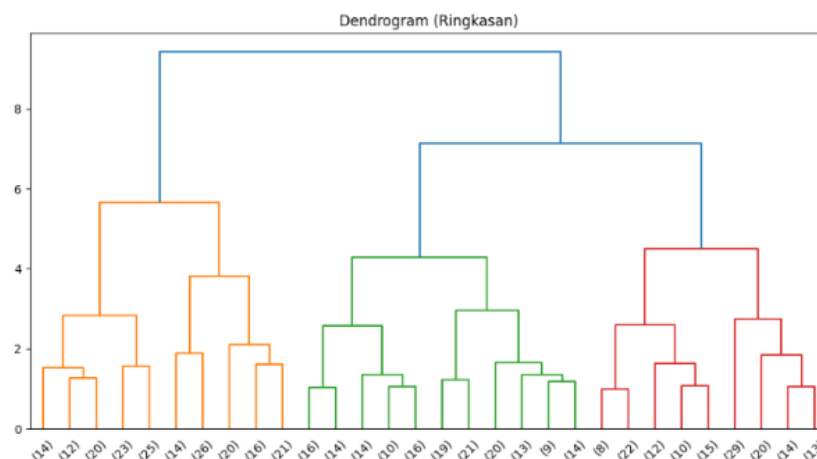
```
import pandas as pd
import matplotlib.pyplot as plt
from scipy.cluster.hierarchy import dendrogram, linkage
from sklearn.preprocessing import MinMaxScaler
df = pd.read_excel('data_penelitian.xlsx')
df.head()
```

Proses hierarchical clustering, dengan menentukan linkage ward
`Z = linkage(X_scaled, method='ward')`

Menampilkan dendrogram dengan kode
`plt.figure(figsize=(12,6))`
`dendrogram(Z, truncate_mode='lastp', p=30)`
`plt.title('Dendrogram (Ringkasan)')`
`plt.show()`

```
from scipy.cluster.hierarchy import fcluster
clusters = fcluster(Z, t=3, criterion='maxclust')
df['cluster'] = clusters
df.head()
df['cluster'].value_counts()
```

Sehingga hasil dendrogram akan seperti pada gambar 5.



Gambar 5. Tampilan Dendrogram

Menampilkan nilai atribut per cluster
`df.groupby('cluster').mean()`
 Menampilkan jumlah cluster dan anggotanya
`df.head()`
`df['cluster'].value_counts()`

Penentuan jumlah cluster pada metode Hierarchical Clustering dilakukan berdasarkan analisis dendrogram. Dendrogram merepresentasikan proses penggabungan cluster dalam bentuk struktur pohon, di mana sumbu vertikal menunjukkan jarak atau dissimilarity antar cluster. Jumlah cluster optimal ditentukan dengan memotong dendrogram pada titik dengan jarak vertikal terbesar (longest vertical distance) yang tidak memotong banyak cabang. Berdasarkan hasil visualisasi dendrogram pada Gambar 5, diperoleh bahwa pemotongan yang paling optimal menghasilkan 3 cluster, sehingga nilai tersebut digunakan dalam proses pembentukan cluster menggunakan fungsi `fcluster` dengan parameter `criterion = 'maxclust'`.

Hasil cluster akan menjadi seperti pada tabel 6.

Tabel 6. Hasil Hierarchical Clustering

Cluster	Overall score	Study hours	Attendance percentage	Study method	Jumlah	Kategori
1	0.299	0.230	0.432	0.582	191	Rendah
2	0.633	0.616	0.497	0.187	166	Sedang
3	0.740	0.740	0.589	0.732	143	Tinggi

Berdasarkan tabel 6, diperoleh 3 kelompok mahasiswa dengan karakteristik yang berbeda. Cluster 1 dengan jumlah siswa 191 menunjukkan kelompok mahasiswa dengan tingkat performa rendah yang ditandai dengan nilai rata-rata (*overall score*), jam belajar (*study hours*), dan kehadiran (*attendance percentage*) yang relatif rendah. Cluster 2 dengan 166 siswa merepresentasikan kelompok dengan performa sedang, mahasiswa memiliki jam belajar yang cukup tinggi namun belum diimbangi dengan metode belajar yang efektif. Sementara itu, Cluster 3 sebanyak 143 siswa merupakan kelompok dengan performa tinggi yang ditunjukkan oleh tingginya nilai pada seluruh variabel, termasuk jam belajar, metode belajar, dan tingkat kehadiran.

3.5. Evaluasi dan Analisis Kinerja Model Clustering

Pada tahap evaluasi kinerja model clustering dilakukan untuk menilai kualitas hasil pengelompokan yang telah diperoleh dengan perhitungan kedua metode clustering, yaitu metode K-Means dan Hierarchical Clustering. Proses evaluasi kinerja model clustering menggunakan *Davies-Bouldin Index (DBI)* dan *Silhouette Score*. Penerapan kode python dalam proses perhitungan metrik evaluasi cluster metode K-Means seperti berikut:

1. Perhitungan DBI untuk KMeans

```
# perhitungan DBI
import pandas as pd
from sklearn.metrics import davies_bouldin_score
df = pd.read_excel('hasil_kmeans.xlsx')
X = df[['overall_score','study_hours','attendance_percentage','study_method']]
labels = df['cluster']
dbi = davies_bouldin_score(X, labels)
print("Nilai DBI:", dbi)
Nilai DBI: 1.163210435583218
```

2. Perhitungan Silhouette Score untuk KMeans

```
# perhitungan Silhouette
import pandas as pd
from sklearn.metrics import silhouette_score
from sklearn.preprocessing import StandardScaler
# Load data
df = pd.read_excel('hasil_kmeans.xlsx')
# Ambil fitur
X=df[['overall_score','study_hours','attendance_percentage','study_method']]
# 4. Ambil label cluster
labels = df['cluster']
# Normalisasi
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
# Hitung Silhouette Score
score = silhouette_score(X_scaled, labels)
print("Silhouette Score:", score)
Silhouette Score: 0.2763987193448641
```

Sedangkan penerapan kode python pada proses perhitungan metrik evaluasi cluster metode Hierarchical seperti berikut:

1. Perhitungan DBI untuk Hierarchical Clustering

```
#import library
import pandas as pd
from sklearn.preprocessing import MinMaxScaler
from sklearn.cluster import AgglomerativeClustering
```

```

from sklearn.metrics import davies_bouldin_score
# Load data (contoh dari Excel)
df = pd.read_excel('data_penelitian.xlsx')
# Ambil kolom yang digunakan (sesuaikan dengan data Anda)
X = df[['overall_score', 'study_hours',
        'attendance_percentage', 'study_method']]
# Normalisasi
scaler = MinMaxScaler()
X_scaled = scaler.fit_transform(X)
Nilai Davies-Bouldin Index: 1.3535044744163134

```

2. Perhitungan Silhouette Score untuk Hierarchical Clustering

```

import pandas as pd
from sklearn.cluster import AgglomerativeClustering
from sklearn.metrics import silhouette_score
data = pd.read_excel('data_penelitian.xlsx')
# Ambil kolom fitur (sesuaikan)
X = data[['overall_score', 'study_hours', 'attendance_percentage', 'study_method']]
hc = AgglomerativeClustering(n_clusters=3, linkage='ward')
labels = hc.fit_predict(X)
score = silhouette_score(X, labels)
print("Silhouette Score:", score)
Silhouette Score: 0.225539250804675

```

Hasil perhitungan dengan menggunakan kedua metrik evaluasi cluster untuk metode K-Means dan Hierarchical clustering ada pada tabel 7.

Tabel 7. Hasil Evaluasi Clustering

Metode	DBI	Silhouett Score
K-Means	1.163	0.276
Hierarchical clustering	1.353	0.225

Berdasarkan hasil evaluasi pada Tabel 7, metode K-Means memiliki performa lebih baik dibandingkan Hierarchical Clustering, ditunjukkan oleh nilai DBI yang lebih rendah (1.163) dan Silhouette Score yang lebih tinggi (0.276). Hal ini menunjukkan bahwa K-Means menghasilkan cluster yang lebih homogen dan memiliki pemisahan yang lebih jelas.

Keunggulan K-Means dipengaruhi oleh karakteristik dataset yang digunakan, yaitu data numerik yang telah dinormalisasi dan cenderung membentuk pola yang kompak. Kondisi ini sesuai dengan cara kerja K-Means yang mengelompokkan data berdasarkan kedekatan terhadap centroid menggunakan jarak Euclidean. Selain itu, jumlah data yang cukup besar (500 data) membuat K-Means bekerja lebih optimal karena algoritma ini mampu secara iteratif meminimalkan variasi dalam cluster, sehingga menghasilkan cluster yang stabil dan konsisten.

Sebaliknya, Hierarchical Clustering bekerja dengan menggabungkan data secara bertahap. Proses ini sangat bergantung pada keputusan awal dan metode linkage (Ward), sehingga pada data yang besar dan bervariasi dapat menghasilkan cluster yang kurang optimal. Metode ini juga lebih sensitif terhadap noise, sehingga pemisahan cluster menjadi kurang tegas. Dengan demikian, K-Means lebih unggul pada penelitian ini karena sesuai dengan karakteristik data, sedangkan Hierarchical Clustering lebih cocok untuk data berukuran kecil atau untuk analisis struktur hierarki.

4. KESIMPULAN

Penelitian yang telah dilakukan dengan menggunakan dataset pendidikan publik tentang *student performance*, dan diolah dengan 2 metode clustering, yaitu K-Means dan Hierarchical Clustering telah memberikan hasil bahwa kinerja clustering dengan metode K-Means lebih efektif daripada Hierarchical Clustering. Keunggulan kinerja ini dipengaruhi oleh kesesuaian antara karakteristik dataset yang berupa data numerik terstandarisasi dengan mekanisme kerja K-Means yang berbasis centroid dan optimasi variansi, sehingga mampu menghasilkan cluster yang lebih kompak dan terpisah dengan baik.

Penelitian ini memiliki beberapa keterbatasan, yakni dataset yang digunakan masih terbatas pada satu sumber data publik dengan jumlah variabel yang relatif sedikit, sehingga belum merepresentasikan kondisi data pendidikan

yang lebih kompleks. Keterbatasan berikutnya, yakni metode clustering yang digunakan hanya terbatas pada K-Means dan Hierarchical Clustering, tanpa membandingkan dengan metode lain yang mungkin memberikan hasil yang berbeda. Kemudian, evaluasi yang digunakan hanya berfokus pada metrik internal, sehingga belum mempertimbangkan validasi eksternal berbasis label atau konteks nyata.

Saran untuk penelitian selanjutnya, dapat menggunakan dataset yang lebih beragam dan kompleks, serta menambahkan variabel lain yang relevan agar hasil clustering lebih representatif. Selanjutnya, penggunaan metode clustering lain seperti DBSCAN atau metode berbasis model dapat dipertimbangkan untuk memperoleh perbandingan yang lebih komprehensif. Dari segi pengembangannya, dapat diarahkan pada penerapan hasil clustering dalam sistem pendukung keputusan di bidang pendidikan, sehingga memberikan manfaat yang lebih aplikatif.

DAFTAR PUSTAKA

- [1] V. Ashrafimoghari, "Big Data and Education: using big data analytics in language learning," 2022, doi: 10.13140/RG.2.2.21946.06080.
- [2] I. Papadogiannis, M. Wallace, and G. Karountzou, "Educational Data Mining: A Foundational Overview," *Encyclopedia*, vol. 4, no. 4, pp. 1644–1664, 2024, doi: 10.3390/encyclopedia4040108.
- [3] A. D. A. Systematic, "G-Tech : Jurnal Teknologi Terapan Educational Data Mining in Online Learning : Data Mining Techniques and," vol. 10, no. 1, pp. 190–205, 2026.
- [4] E. L. Cahapin, B. A. Malabag, C. S. Santiago, J. L. Reyes, G. S. Legaspi, and K. L. Adrales, "Clustering of students admission data using k-means, hierarchical, and DBSCAN algorithms," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 6, pp. 3647–3656, 2023, doi: 10.11591/eei.v12i6.4849.
- [5] A. Alif *et al.*, "Penerapan Algoritma K-means Clustering dan Hierarchical Clustering dalam Mengelompokkan Data Pengangguran di Karawang," *Jurnal Algoritma*, vol. 21, no. 2, pp. 209–220, Nov. 2024, doi: 10.33364/algoritma/v.21-2.2155.
- [6] S. N. Safitri, H. Setiadi, and E. Suryani, "Educational Data Mining Using Cluster Analysis Methods and Decision Trees based on Log Mining," *Jurnal RESTI*, vol. 6, no. 3, pp. 448–456, Jun. 2022, doi: 10.29207/resti.v6i3.3935.
- [7] R. A. N. Diaz, N. L. G. P. Suwirmayanti, E. Setyaningsih, and I. W. B. Sentana, "Comparing K-Means, GMM and BIRCH for Student Academic Performance Data: Evaluation on Two Public Datasets," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 25, no. 1, pp. 217–230, Nov. 2025, doi: 10.30812/matrik.v25i1.5843.
- [8] S. J. Sultan Alalawi, I. N. Mohd Shahrane, and J. Mohd Jamil, "CLUSTERING STUDENT PERFORMANCE DATA USING k-MEANS ALGORITHMS," *Journal of Computational Innovation and Analytics (JCIA)*, vol. 2, no. 1, pp. 41–55, Jan. 2023, doi: 10.32890/jcia2023.2.1.3.
- [9] I. P. Putra and A. Fadhillah, "Perbandingan Metode K-Means dan Hierarchical Clustering dalam Pengelompokan Data Penduduk Miskin di Kabupaten Cianjur," *LANCAH: Jurnal Inovasi dan Tren*, vol. 3, no. 1, Apr. 2025, doi: 10.35870/ljit.v3i1.4028.
- [10] N. T. Hartanti, E. Seniwati, R. Pramitasari, I. R. Wulandari, and U. A. Yogyakarta, "Jumlah Cluster Optimal dalam Pengelompokan Siswa SMK dengan Metode Elbow K-Means Clustering," Dec. 2024. [Online]. Available: <http://jurnal.bsi.ac.id/index.php/jais>
- [11] A. Agung, A. Daniswara, I. Kadek, and D. Nuryana, "Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru," *Journal of Informatics and Computer Science*, vol. 05, 2023.
- [12] N. A. Maori, "METODE ELBOW DALAM OPTIMASI JUMLAH CLUSTER PADA K-MEANS CLUSTERING," *Jurnal SIMETRIS*, vol. 14, 2023.
- [13] N. T. Hartanti, D. Maulina, E. Seniwati, and I. R. Wulandari, "Penerapan Teknik Clustering untuk Identifikasi Pola Perilaku Belajar Siswa Menggunakan Dataset Publik Learning Analytics Universitas Amikom Yogyakarta," *Journal of Innovative and Creativity*, vol. 5, no. 3, pp. 14651–14663, Dec. 2025.
- [14] N. A. Maori, "METODE ELBOW DALAM OPTIMASI JUMLAH CLUSTER PADA K-MEANS CLUSTERING," *Jurnal SIMETRIS*, vol. 14, 2023.
- [15] J. Homepage *et al.*, "MALCOM: Indonesian Journal of Machine Learning and Computer Science Clustering Analysis Using the K-Means Method to Identify Alumni Satisfaction Pattern Analisis Clustering Menggunakan Metode K-Means untuk Mengidentifikasi Pola Kepuasan Alumni," vol. 6, no. 1, pp. 31–41, 2026, doi: 10.57152/malcom.v6i1.2401.

- [16] N. I. Boyko and O. A. Tkachyk, "Hierarchical clustering algorithm for dendrogram construction and cluster counting," *Informatics and mathematical methods in simulation*, vol. 13, no. 1–2, pp. 5–15, Apr. 2023, doi: 10.15276/imms.v13.no1-2.5.