

EVALUASI KINERJA *K-MEANS*, *K-HARMONIC MEANS*, DAN *K-MEDOIDS* PADA DATA YANG MENGANDUNG PENCILAN

NAUFALIA ALFIRYAL



PROGRAM STUDI MAGISTER STATISTIKA DAN SAINS DATA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2026

@Hak cipta milik IPB University

IPB University



IPB University
— Bogor Indonesia —

- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Perpustakaan IPB University



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

PERNYATAAN MENGENAI TESIS DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa tesis dengan judul “Evaluasi Kinerja *K-Means*, *K-Harmonic Means*, dan *K-Medoids* pada Data yang Mengandung Pencilan” adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir tesis ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Agustus 2026

Naufalia Alfiryal
M0501241074



RINGKASAN

NAUFALIA ALFIRYAL. Evaluasi Kinerja *K-Means*, *K-Harmonic Means*, dan *K-Medoids* pada Data yang Mengandung Pencilan. Dibimbing oleh ASEP SAEFUDDIN dan I MADE SUMERTAJAYA.

Data mining merupakan proses untuk menemukan pola dari sekumpulan data, salah satunya melalui teknik *unsupervised learning* seperti analisis gerombol. Analisis gerombol bertujuan untuk mengelompokkan objek ke dalam beberapa gerombol berdasarkan kemiripan karakteristik. Namun, dalam praktiknya, analisis gerombol sering menghadapi permasalahan berupa keberadaan pencilan (*outlier*), yaitu data yang memiliki karakteristik sangat berbeda dibandingkan data lainnya. Kehadiran pencilan dapat memengaruhi hasil pengelompokan sehingga gerombol yang terbentuk menjadi kurang representatif.

Metode *K-Means* merupakan salah satu metode penggerombolan berbasis partisi yang banyak digunakan karena sederhana dan efisien. Akan tetapi, metode ini sensitif terhadap pencilan dan sangat bergantung pada inisialisasi awal titik pusat gerombol. Untuk mengatasi kelemahan tersebut, dikembangkan metode *K-Harmonic Means* dan *K-Medoids* yang dinilai lebih stabil dan lebih kokoh terhadap pencilan. Meskipun demikian, kajian yang secara khusus membandingkan ketiga metode tersebut dalam kondisi data yang mengandung pencilan masih terbatas.

Penelitian ini bertujuan untuk mengevaluasi dan membandingkan kinerja metode *K-Means*, *K-Harmonic Means*, dan *K-Medoids* pada data yang mengandung pencilan. Evaluasi akan dilakukan menggunakan data simulasi yang dibangkitkan dalam beberapa skenario dengan variasi tingkat pencilan. Kinerja metode akan dibandingkan berdasarkan beberapa metrik evaluasi penggerombolan yang mencerminkan ketepatan dan kestabilan hasil gerombol.

Kajian simulasi dilakukan menggunakan data bangkitan yang disusun berdasarkan karakteristik data empiris dengan variasi proporsi pencilan (0%, 5%, 10%, dan 15%), tingkat kedekatan antargerombol (dekat, sedang, dan jauh), serta komposisi jumlah anggota gerombol (seimbang dan tidak seimbang). Setiap skenario dianalisis menggunakan ketiga metode dengan 10 kali perulangan dan dievaluasi menggunakan *Silhouette Coefficient*, *Gamma Index*, dan *Davies-Bouldin Index*, serta dianalisis menggunakan ANOVA. Hasil simulasi menunjukkan bahwa *K-Means* dan *K-Medoids* secara umum memberikan performa yang paling konsisten dan relatif setara pada berbagai kombinasi skenario. Sebaliknya, KHM menunjukkan perubahan performa yang lebih besar pada beberapa kombinasi proporsi pencilan dan tingkat kedekatan antargerombol. Hasil tersebut menunjukkan bahwa kualitas penggerombolan tidak hanya dipengaruhi oleh proporsi pencilan, tetapi juga oleh tingkat kedekatan antargerombol dan interaksinya dengan metode yang digunakan. Dengan demikian, *K-Means* dan *K-Medoids* tidak menunjukkan perbedaan performa yang besar sebagaimana dugaan awal bahwa *K-Medoids* akan lebih unggul pada data yang mengandung pencilan.

Data empiris yang digunakan adalah data subindeks Indikator Pembangunan Teknologi, Informasi, dan Komunikasi (IP-TIK) per provinsi yang diperoleh dari publikasi BPS Indonesia. Hasil analisis dengan *K-Means* dan *K-Medoids* menghasilkan struktur penggerombolan yang hampir identik, yaitu membentuk lima gerombol dengan DKI Jakarta dan Papua sebagai gerombol tunggal. Hasil tersebut

menunjukkan adanya heterogenitas pembangunan TIK antarprovinsi di Indonesia. Papua masih menunjukkan capaian yang relatif rendah pada indikator akses, penggunaan TIK, dan pendidikan, sedangkan DKI Jakarta memiliki capaian tertinggi pada sebagian besar indikator yang digunakan dalam penelitian.

Kata kunci: IP-TIK, K-Harmonic Means, K-Means, K-Medoids, Pencilan.

@Hak cipta milik IPB University

IPB University



Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



SUMMARY

NAUFALIA ALFIRYAL. Performance Evaluation of K-Means, K-Harmonic Means, and K-Medoids on Data Containing Outliers. Supervised by ASEP SAEFUDDIN, and I MADE SUMERTAJAYA.

Data mining is a process of discovering patterns from a collection of data, one of which can be conducted through unsupervised learning techniques such as cluster analysis. Cluster analysis aims to group objects into several clusters based on the similarity of their characteristics. However, in practice, cluster analysis often encounters the problem of outliers, namely observations with characteristics that differ substantially from those of other observations. The presence of outliers can affect clustering results, causing the resulting clusters to become less representative.

K-Means is one of the widely used partition-based clustering methods due to its simplicity and efficiency. However, the method is sensitive to outliers and highly dependent on the initial selection of cluster centers. To address these limitations, K-Harmonic Means and K-Medoids have been developed with different characteristics in dealing with data containing outliers. Nevertheless, studies specifically comparing these three methods under conditions involving outliers remain limited. This study aims to evaluate and compare the performance of K-Means, K-Harmonic Means, and K-Medoids on data containing outliers. The methods are evaluated using simulated data generated under several scenarios with varying levels of outlier contamination. Their performance is compared using several cluster evaluation metrics that reflect the quality and stability of the resulting clusters.

The simulation study uses generated data constructed based on the characteristics of empirical data, with variations in outlier proportions (0%, 5%, 10%, and 15%), levels of inter-cluster proximity (near, moderate, and far), and cluster size compositions (balanced and unbalanced). Each scenario is analyzed using the three methods with 10 replications and evaluated using the Silhouette Coefficient, Gamma Index, and Davies-Bouldin Index, followed by analysis of variance (ANOVA). The simulation results show that K-Means and K-Medoids generally provide the most consistent and relatively comparable performance across various combinations of scenarios. In contrast, KHM exhibits greater performance variation under several combinations of outlier proportions and inter-cluster proximity levels. These results indicate that clustering quality is influenced not only by the proportion of outliers but also by the level of inter-cluster proximity and its interaction with the clustering method. Thus, K-Means and K-Medoids do not show substantial differences in performance, contrary to the initial expectation that K-Medoids would perform better on data containing outliers.

The empirical data consist of provincial sub-index data from the Information and Communication Technology Development Index (ICT Development Index), obtained from a publication by Statistics Indonesia (BPS). The analyses using K-Means and K-Medoids produce nearly identical clustering structures, forming five clusters, with DKI Jakarta and Papua each forming a single-member cluster. These results indicate heterogeneity in ICT development across provinces in Indonesia. Papua exhibits relatively low performance in indicators related to ICT access, ICT use, and education, whereas DKI Jakarta achieves the highest performance on most of the indicators included in the study.

Keywords: *IP-TIK, K-Harmonic Means, K-Means, K-Medoids, Outliers.*



@Hak cipta milik IPB University

IPB University



Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

© Hak Cipta milik IPB, tahun 2026
Hak Cipta dilindungi Undang-Undang

Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan atau menyebutkan sumbernya. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik, atau tinjauan suatu masalah, dan pengutipan tersebut tidak merugikan kepentingan IPB.

Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apa pun tanpa izin IPB.

EVALUASI KINERJA *K-MEANS*, *K-HARMONIC MEANS*, DAN *K-MEDOIDS* PADA DATA YANG MENGANDUNG PENCILAN

NAUFALIA ALFIRYAL

Tesis

sebagai salah satu syarat untuk memperoleh gelar

Magister Sains pada

Program Studi Statistika dan Sains Data

**PROGRAM STUDI MAGISTER STATISTIKA DAN SAINS DATA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2026**

@Hak cipta milik IPB University

IPB University





@Hak cipta milik IPB University

IPB University



IPB University
— Bogor Indonesia —

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Tim Penguji pada Ujian Tesis:

Dr. Ir. Indahwati. M.Si.



Judul Tesis : Evaluasi Kinerja *K-Means*, *K-Harmonic Means*, dan *K-Medoids*
pada Data yang Mengandung Pencilan.

Nama : Naufalia Alfiryal

NIM : M0501241074

@Hak cipta milik IPB University

Disetujui oleh

Pembimbing 1:

Prof. Dr. Ir. Asep Saefuddin, M.Sc.

Pembimbing 2:

Prof. Dr. Ir. I Made Sumertajaya, M.Si.

Diketahui oleh

Ketua Program Studi:

Dr. Agus Mohamad Soleh, S.Si., M.T.

NIP. 19750315 199903 1 004

Dekan Sekolah Sains Data, Matematika, dan Informatika

Prof. Dr. Ir. Agus Buono, M.Si., M.Kom.

NIP. 19660702 199302 1 001

Tanggal Ujian:
11 Agustus 2026

Tanggal Lulus:

PRAKATA

Puji dan syukur penulis panjatkan kepada Allah subhanaahu wa ta'ala atas segala karunia-Nya sehingga proposal penelitian tesis dengan judul “Evaluasi Kinerja *K-Means*, *K-Harmonic Means*, dan *K-Medoids* Pada Data yang Mengandung Pencilan” dapat diselesaikan dengan baik.

Dalam penyusunan proposal tesis ini, penulis menyadari bahwa banyak pihak yang telah memberikan bantuan, bimbingan, serta dukungan baik secara langsung maupun tidak langsung. Oleh karena itu, penulis mengucapkan terima kasih sebesar-besarnya kepada:

1. Bapak Prof. Dr. Ir. Asep Saefuddin, M.Sc., dan Bapak Prof. Dr. Ir. I Made Sumertajaya, M. Si., selaku komisi pembimbing yang telah memberikan arahan, bimbingan, serta masukan yang berharga dalam proses perencanaan dan penyusunan karya ilmiah ini.
2. Seluruh dosen dan tenaga kependidikan di Departemen Statistika dan Sains Data IPB University yang telah memberikan ilmu dan dukungan selama proses perkuliahan.
3. Badan Pusat Statistik yang telah menyediakan data yang digunakan dalam penelitian ini.
4. Kedua orang tua dan Keluarga penulis yang senantiasa memberikan do'a, dukungan, dan semangat selama proses akademik berlangsung hingga penyusunan karya ilmiah ini.
5. Teman-teman Program Studi Statistika dan Sains Data Angkatan 61 serta semua pihak yang tidak dapat disebutkan satu per satu atas motivasi dan dukungannya.

Penulis tentu menyadari bahwa karya ilmiah ini masih sangat jauh dari sempurna. Oleh karena itu, penulis sangat mengharapkan kritik dan saran yang membangun demi perbaikan di masa yang akan datang. Akhir kata, penulis berharap karya ilmiah ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan serta pihak-pihak yang membutuhkan.

Bogor, Agustus 2026

Naufalia Alfiryal



@Hak cipta milik IPB University

IPB University



IPB University
— Bogor Indonesia —

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Perpustakaan IPB University

DAFTAR ISI

DAFTAR ISI	i
DAFTAR TABEL	ii
DAFTAR GAMBAR	iii
DAFTAR LAMPIRAN	iv
I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Tujuan Penelitian	3
1.4 Manfaat Penelitian	3
II TINJAUAN PUSTAKA	4
2.1 Pencilan	4
2.1.1 Deteksi Pencilan	5
2.1.2 Penyisipan Pencilan	6
2.2 Data Mining	6
2.3 Analisis Gerombol	7
2.3.1 Ukuran Jarak <i>Euclidean</i>	7
2.3.2 Dekomposisi Cholesky	8
2.4 <i>Principal Component Analysis</i> (PCA)	9
2.5 <i>K-Means</i>	9
2.6 <i>K-Harmonic Means</i>	10
2.7 <i>K-Medoids</i>	11
2.8 Metrik Evaluasi Hasil Analisis Gerombol	11
2.9 Analysis of Variance (ANOVA)	14
2.10 Indeks Pembangunan Teknologi Informasi dan Komunikasi (IP-TIK)	15
III METODOLOGI	17
3.1 Hipotesis Penelitian	17
3.2 Data	17
3.2.1 Data Simulasi	17
3.2.2 Data Empiris	18
3.3 Prosedur Analisis Data	18
IV HASIL DAN PEMBAHASAN	23
4.1 Eksplorasi Data Empiris	23
4.2 Kajian Data Simulasi	25
4.2.1 Karakteristik Data Simulasi	25
4.2.2 Hasil Evaluasi Penggerombolan	30
4.2.3 Pemilihan Algoritma Analisis Gerombol Terbaik	32
4.3 Kajian Data Empiris	36
4.3.1 Analisis Gerombol dengan Menggunakan Algoritma <i>K-means</i> dan <i>K-medoids</i>	36
4.3.2 Karakteristik Hasil Penggerombolan berdasarkan Algoritma <i>K-means</i> dan <i>K-medoids</i>	38
V SIMPULAN DAN SARAN	41
5.1 Simpulan	41
5.2 Saran	41
DAFTAR PUSTAKA	42
LAMPIRAN	45
RIWAYAT HIDUP	71

DAFTAR TABEL

1	Kriteria struktur gerombol	13
2	Subindeks dan Indikator Pembangun IP-TIK	15
3	Hipotesis penelitian	17
4	Skenario data simulasi	17
5	Peubah yang digunakan dalam penelitian	18
6	Statistika deskriptif	23
7	Profil karakteristik gerombol hasil analisis gerombol skenario jumlah anggota gerombol seimbang	27
8	Profil karakteristik gerombol hasil analisis gerombol skenario jumlah anggota gerombol tidak seimbang	27
9	Tabulasi silang kelompok acuan dan hasil pengelompokan pada kedekatan dekat dengan proporsi pencilan 0%	29
10	Tabulasi silang <i>ground truth</i> dan hasil pengelompokan pada kedekatan dekat dengan proporsi pencilan 5%	30
11	Hasil uji ANOVA pada jumlah anggota gerombol tidak seimbang	34
12	Sebaran Anggota Setiap Gerombol	40

DAFTAR GAMBAR

1.	Contoh pencilan dalam dataset 2-dimensi (Chandola, 2009)	4
2.	Visualisasi boxplot dengan data pencilan (Smiti, 2020)	5
3.	Titik pusat pada K-Medoids	11
4.	Diagram alir penelitian kajian simulasi	19
5.	Diagram alir penelitian kajian empiris	20
6.	Diagram kotak garis untuk setiap peubah	24
7.	Diagram T^2 Hotelling	24
8.	Matriks korelasi	25
9.	Plot Sebaran data dua komponen utama untuk skenario pencilan 5% untuk masing-masing variasi kedekatan dan komposisi jumlah anggota geromb	26
10.	Perbandingan evaluasi performa algoritma <i>K-Means</i> , <i>K-Medoids</i> , dan KHM pada skenario jumlah anggota gerombol seimbang (a) dan jumlah anggota gerombol tidak seimbang (b)	31
11.	Plot interaksi proporsi pencilan, tingkat kedekatan, dan algoritma pada skenario jumlah anggota gerombol seimbang	33
12.	Plot interaksi proporsi pencilan, tingkat kedekatan, dan algoritma pada skenario jumlah anggota gerombol tidak seimbang	35
13.	Perbandingan nilai <i>Silhouette Score</i> per jumlah gerombol untuk <i>K-means</i> dan <i>K-medoids</i>	37
14.	Visualisasi hasil analisis gerombol	38
15.	Karakteristik gerombol berdasarkan kelompok satuan indikator pada (a) indikator X1-X3, (b) indikator X4, dan (c) indikator X5-X6	39
16.	Peta sebaran anggota gerombol	40

DAFTAR LAMPIRAN

1.	Plot Sebaran Data Simulasi	45
2.	Profil Gerombol Untuk Setiap Skenario Data Simulasi (Jumlah Gerombol Seimbang)	47
3.	Profil Gerombol Untuk Setiap Skenario Data Simulasi (Jumlah Gerombol Tidak Seimbang)	53
4.	Tabulasi Silang Kelompok Acuan dan Hasil Penggerombolan pada Skenario dengan Jumlah Anggota Gerombol Seimbang	59
5.	Tabulasi Silang Kelompok Acuan dan Hasil Penggerombolan pada Skenario dengan Jumlah Anggota Gerombol Tidak Seimbang	65

@Hak cipta milik IPB University

IPB University



Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :

- Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
- Pengutipan tidak merugikan kepentingan yang wajar IPB University.

2. Dilarang mengumunkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.