Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

Comparison and Evaluation of Euclidean Distance and Divergence in Adaptive K-Means Algorithm for Clustering Human Development Index of Indonesia Province

Maria Claudia Purba[1], Zakarias Situmorang[2]
Faculty of computer sciene, University Katolik Santo Thomas, Medan, Indonesia

Article Info
Keywords:
K-Means Adaptive,
Divergence distance,
Euclidean distance,
Indeks Pembangunan Manusia,
Clustering

ABSTRACT
This research explores the application of the Adaptive K-Means clustering algorithm on Human Development Index (HDI) data across 34 provinces in Indonesia, comparing the performance of Euclidean and Divergence distance metrics. The HDI indicators used include life expectancy, years of schooling, and per capita expenditure. Data processing was conducted both manually on sample data and automatically using Python for the complete dataset. Results demonstrate that the choice of distance metric significantly impacts clustering effectiveness. Divergence outperformed Euclidean based on silhouette score evaluations, offering more representative cluster separation. Scatter plot visualizations tracked the iterative clustering process. The study contributes to optimizing clustering techniques for socio-economic indicators such as HDI.

Corresponding Author:
Maria Claudia Purba
University Katolik Santo Thomas, Medan, Indonesia
mariaccll033@gmail.com

INTRODUCTION

The Human Development Index (HDI) serves as a critical metric for assessing quality of life, encompassing health, education, and income components. Accurate clustering of HDI data helps policymakers identify disparities and formulate targeted development strategies. The K-Means algorithm is widely used due to its simplicity and efficiency in grouping data based on feature similarities. However, the performance of K-Means significantly depends on the choice of distance metric.

Several studies have focused on applying K-Means to socio-economic data, yet limited attention has been given to evaluating alternative distance metrics beyond the standard Euclidean measure. This research fills that gap by comparing Euclidean and Divergence distances within the Adaptive K-Means framework to determine which yields more meaningful clusters for HDI data.

The main objective of this study is to evaluate the effectiveness of both distance metrics in classifying Indonesian provinces by development level, thereby guiding future clustering applications in the field of socio-economic analysis.

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

METHODS

This study uses the Adaptive K-Means algorithm to cluster Human Development Index (HDI) data from 34 provinces in Indonesia. The data used is obtained from the official government portal, consisting of numerical indicators: Life Expectancy, Expected Years of Schooling, and Expenditure per Capita. Prior to clustering, the data was cleaned of duplicates and empty values. All attributes were standardized and ready to use without additional normalization.

The clustering process was performed by manually determining the initial centroid for the three clusters: Low, Medium and High HDI. Two distance measurement methods were used in the evaluation, namely:

1. Euclidean Distance
   This method calculates the distance between the data to the centroid using a formula:

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i|$$

   For manual calculations, its squared form is used:

$$D^2(x, c) = \sum_{i=1}^{n} (x_i - c_i)^2$$

2. Divergence Distance
   Divergence is used to measure the difference in distribution between two vectors, namely the data vector *vi* and the centroid *vij* with the formula:

$$D_{\text{div}}(v_i, v_j) = \sum_{k=1}^{R} \left(\frac{(v_{ik} - v_{jk})^2}{v_{ik} + v_{jk}} \right)$$

Each iteration involves the process of: distance calculation, cluster membership update, and centroid recalculation based on the average of the cluster members. Iteration continues until the clustering results are stable (convergent). To ensure accurate results, the process was performed manually on sample data and automatically on the whole data using Python in Google Colab, with scatter plot visualization for each iteration.

This observation can be described as follows:

1. Data Source
   Human Development Index (HDI) classification data is obtained from the official portal of the Indonesian government, https://data.go.id, covering 34 provinces with key indicators: life expectancy, schooling expectancy, average years of schooling, and per capita expenditure.expectancy, average years of schooling, and per capita expenditure.
2. Data Cleaning
   The dataset was checked to remove duplicates and incomplete data. Entries with blank values on key indicators were removed, resulting in 34 valid data.
3. Ready to Use Dataset

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

After cleaning, the HDI data is ready to be used in the clustering process using the Adaptive K-Means algorithm. Two distance measurement methods-Euclidean and Divergence-were used to compare their effectiveness in distinguishing development levels between provinces.

Table 1. data set

Indeks Pembangunan Manu-sia	Harapan Lama Sekolah (Ta-hun)	Pengeluaran per Kapita (Ta-hun)
74.11	14.37	9963
74.51	13.31	10848
75.16	14.1	11130
74.45	13.29	11158
73.11	13.05	10871
72.48	12.55	11109
73.68	13.68	10840
71.79	12.74	10336
73.5	12.18	13358
78.48	12.99	14469
82.77	13.08	18927
73.63	12.62	11277
72.8	12.81	11377
80.65	15.65	14482
74.05	13.37	11992
75.25	13.05	12216
77.4	13.48	13942
71.65	13.96	10681
67.63	13.21	7877
69.71	12.66	9355
73.17	12.75	11458
74	12.82	12469
77.36	13.84	12641
72.21	13.06	9350
74.52	12.95	11179
71.01	13.32	9696
73.96	13.53	11430
72.38	13.69	9708
70.62	13.12	9850
69.19	12.87	9358
72.04	14	8876
70.26	13.73	8398
66.72	13.21	8101
65.89	13.21	8101
62.16	13.10	7146

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

RESULTS AND DISCUSSION

Based on the clustering results, Euclidean distance generated clusters with overlapping characteristics among provinces with varying HDI scores. In contrast, Divergence distance provided clearer cluster boundaries, particularly distinguishing provinces with extremely high or low expenditure values.

Scatter plots of each iteration showed that Divergence resulted in faster convergence with more compact clusters. This is further supported by silhouette scores: 14.55% for Euclidean and 43.57% for Divergence. These results highlight that Divergence better captures variability among multi-dimensional indicators like HDI, which may have non-linear relationships.

Thus, Divergence is more suitable when clustering data with diverse attribute scales and ranges. However, Euclidean may still be effective in cases with relatively uniform data distributions.

1. Euclidean distante calculation for k-means

   The first step is to manually determine the first centroid point. Based on this study, the initial centroid (C1) was determined using data on one of the previously analyzed entries. The first centroid value is:

   C1 : 75; 13; 10848

   Perform calculations for the next centroid using the Adaptive K-Means method with the formula:

   D(Xi)2 = ||Xi-Cj||2

   As an example, two calculations of the squared distance to the initial centroid are shown. Other calculations were performed in Excel with similar formulas and the results are presented in the appendix.

Table 2. Example Distance calculation for sample data

$\|\|Xi-Cj\|\|^2$	$D(Xi)^2$
$((\|\| 75.16 - 74.51 \|\|^2) + (\|\| 14.1 - 13.31 \|\|^2)) + (\|\| 11130 - 10848 \|\|^2))$	282.002
$((\|\| 74.11 - 74.51 \|\|^2) + (\|\| 14.37 - 13.31 \|\|^2)) + (\|\| 9963 - 10848 \|\|^2))$	885.001

The following is the final result of 24 sample data in calculating the distance from the data to the initial centroid (C1).

Table 3. Determining te Centroid

Indeks Pembangunan Manusia 2022	Harapan Lama Sekolah 2022	Pengeluaran per Kapita 2022		d(xi)2
74.11	14.37	9963		885.001
74.51	13.31	10848	C1	0.000
75.16	14.1	11130		282.002
74.45	13.29	11158		310.000
73.11	13.05	10871		23.044
72.48	12.55	11109		261.009

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

Indeks Pembangunan Manusia 2022	Harapan Lama Sekolah 2022	Pengeluaran per Kapita 2022	d(xi)2
73.68	13.68	10840	8.051
71.79	12.74	10336	512.008
73.5	12.18	13358	2510.000
78.48	12.99	14469	3621.002
82.77	13.08	18927	8079.004
73.63	12.62	11277	429.001
72.8	12.81	11377	529.003
80.65	15.65	14482	3634.006
74.05	13.37	11992	1144.000
75.25	13.05	12216	1368.000
77.4	13.48	13942	3094.001
71.65	13.96	10681	167.026
67.63	13.21	7877	2971.008
69.71	12.66	9355	1493.008
73.17	12.75	11458	610.002
74	12.82	12469	1621.000
77.36	13.84	12641	1793.002
72.21	13.06	9350	1498.002
74.52	12.95	11179	331.000

Calculating Probability

$$P(xi) = \frac{D(Xi)^2}{\sum_{i=1}^{n} D(Xi)^2}$$

Table 4. Calculating Probability

Indeks Pembangunan Manusia 2022	Harapan Lama Sekolah 2022	Pengeluaran per Kapita 2022		d(xi)2	p(xi)
74.11	14.37	9963		885.001	0.024
74.51	13.31	10848	C1	0.000	0.000
75.16	14.1	11130		282.002	0.008
74.45	13.29	11158		310.000	0.008
73.11	13.05	10871		23.044	0.001
72.48	12.55	11109		261.009	0.007
73.68	13.68	10840		8.051	0.000
71.79	12.74	10336		512.008	0.014
73.5	12.18	13358		2510.000	0.068
78.48	12.99	14469		3621.002	0.097
82.77	13.08	18927	C2	8079.004	0.217

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

Indeks Pembangunan Manusia 2022	Harapan Lama Sekolah 2022	Pengeluaran per Kapita 2022	d(xi)2	p(xi)
73.63	12.62	11277	429.001	0.012
72.8	12.81	11377	529.003	0.014
80.65	15.65	14482	3634.006	0.098
74.05	13.37	11992	1144.000	0.031
75.25	13.05	12216	1368.000	0.037
77.4	13.48	13942	3094.001	0.083
71.65	13.96	10681	167.026	0.004
67.63	13.21	7877	2971.008	0.080
69.71	12.66	9355	1493.008	0.040
73.17	12.75	11458	610.002	0.016
74	12.82	12469	1621.000	0.044
77.36	13.84	12641	1793.002	0.048
72.21	13.06	9350	1498.002	0.040
74.52	12.95	11179	331.000	0.009

The centroid is selected incrementally using a probabilistic approach based on squared distance. The data with the highest probability is set as the new centroid. In this study, three centroids were obtained:

C1 = (75, 13, 10848)
C2 = (82.77, 13.08, 18927)

Determining Clustering using Euclidean Distance

After determining the centroid, clustering is performed using Euclidean distance. The distance of each province's HDI data to C1, C2, and C3 was calculated, then the data was assigned to the cluster with the closest distance. Calculations were done manually for the sample and automatically with Python for all data, iteratively until convergent.

Table 5. Cluster

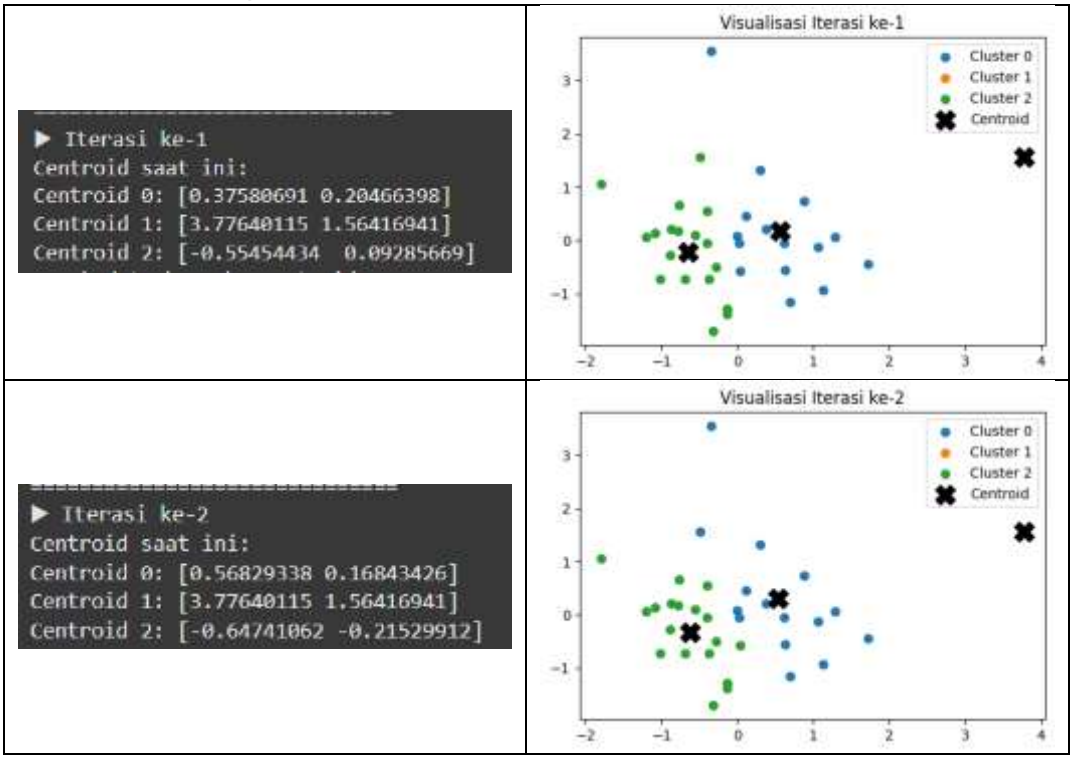
C1	C2	C3	Klaster
783226.13	80353304.76	20421367.66	C1
0	65270249.26	13205962.57	C1
79525.02	60793216.68	11235909.7	C1
96100.06	60357369.32	11048982.63	C1
530.46	64899145.66	13039328.98	C1
68123.6	61121134.3	11377137.74	C1
64.9669	65399578.11	12364171.24	C1
262147.04	73805292.1	17189325.33	C1
6300102.287	31013771.08	1263383.948	C3
13111645.07	19873768.3	172.432856	C3
652070249.3	0	19758028.33	C2
184042.3561	5822509.35	10272032.65	C1

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

C1	C2	C3	Klaster
279842.96	57002510.04	9641033.348	C1
13205967.62	19758033.72	0	C3
1308736.464	48094233.8	6200106.983	C1
1871424.808	45037528.52	513478.16	C3
9572838.919	24850230.53	291607.9589	C3
27892.2825	67996527.89	1447612.86	C1
8826847.89	122102515.2	43626043.97	C1
2229054.223	91623197.24	26286148.88	C1
372101.6536	55785970.71	9144591.89	C1
2627641.75	41705772.84	4052183.659	C1
3214852.131	39513801.99	3389287.566	C1
2244006.363	91718939.56	26337439.15	C1
109561.1369	60031512.27	10909822.42	C1

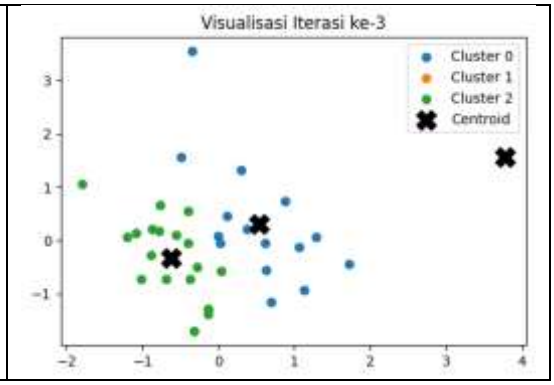
To simplify the calculation od Euclidean distance using big data, Google Colab can be used to generate a scatter plot output in the form of data distribution options.

Figure 1. Euclidean Distance Output



Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

```
==============================
▶ Iterasi ke-3
Centroid saat ini:
Centroid 0: [0.53300419 0.31017562]
Centroid 1: [3.77640115 1.56416941]
Centroid 2: [-0.61955074 -0.3272002 ]
```



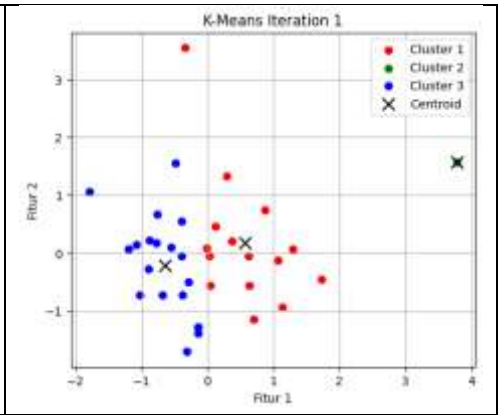
The output obtained from Google Colab is 3 iterations. The iteration stops when the cluster point no longer changes.

Testing with Divergence Distante

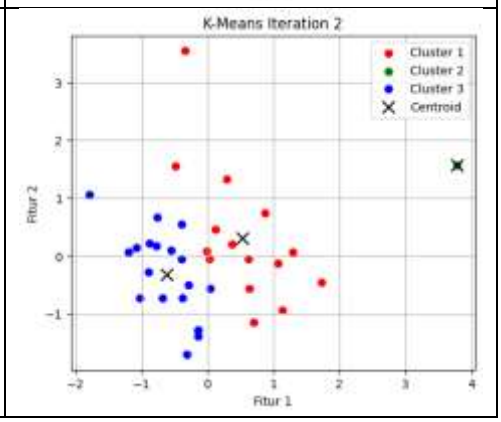
To shorten the testing time, Google Colab tools are used. The following are the iteration results and scatter plot distribution points:

Figure 2. Divergence Distance Output

```
Ringkasan Rata-rata Divergence per Iterasi:
    Iterasi  Rata-rata Divergence
0         1              0.242373
1         2              0.242988
2         3              0.242889
3         4              0.242927
4         5              0.243074
5         6              0.243310
6         7              0.242695
7         8              0.242837
8         9              0.242904
9        10              0.242949
10       11              0.242844
11       12              0.243016
12       13              0.242807
13       14              0.242802
```




```
Detail Divergence Setiap Iterasi:
     Iterasi Pasangan Data  Divergence
0          1          1-2    0.306091
1          1          2-3    0.229840
2          1          3-4    0.235958
3          1          4-5    0.094926
4          1          5-6    0.147299
..       ...          ...         ...
471       14        30-31    0.353218
472       14        31-32    0.115052
473       14        32-33    0.229403
474       14        33-34    0.036091
475       14        34-35    0.189808
```



Evalution of Euclidean Distance and Divergence Distance

After testing Adaptive K-Means with Euclidean distance and Divergence, an evaluation using silhouette score is performed to assess the clustering accuracy. A score value close to 1 indicates a good cluster, while a value close to -1 indicates possible misplacement. This

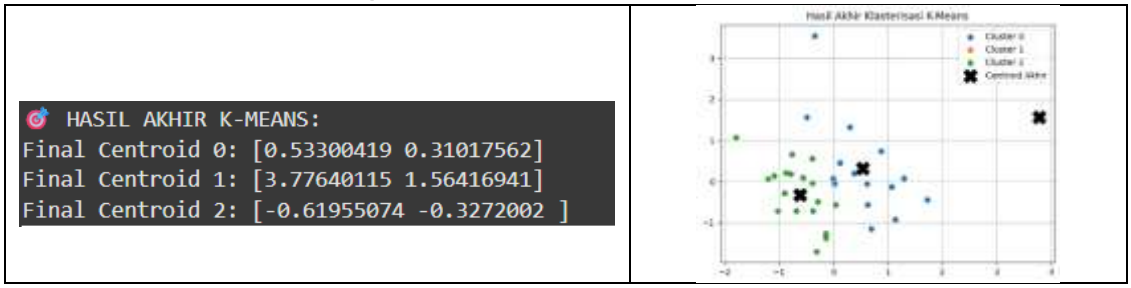
Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

evaluation compares the effectiveness of the two methods in differentiating development levels between provinces.

a. Evaluation of Euclidean distance

The evaluation was carried out using the silhouette score. The following is the output.

Figure 3. Evaluation of Euclidean Distance

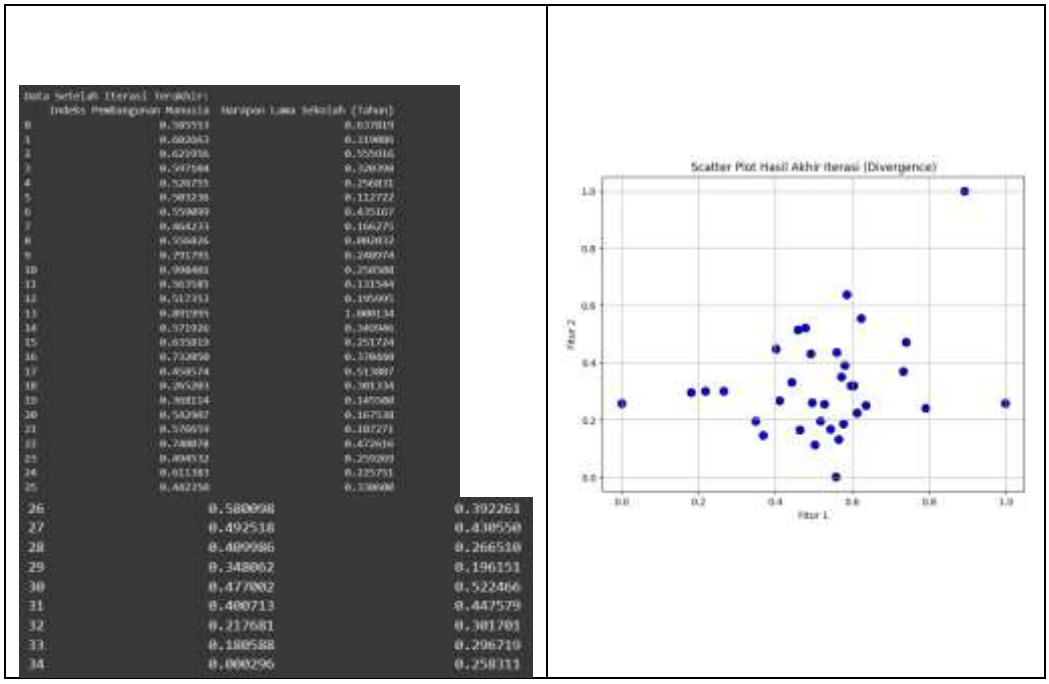


Thus, from the evaluation of the Euclidean distance, the average obtained was 14.55%.

b. Divergence distance evaluation

The following are the results of the silhouette score evaluation calculation:

Figure 4. divergence distance evaluation



Thus, from the evaluation results of the divergence distance, the average obtained was 43.57%.

CONCLUSION

This study applied the Adaptive K-Means algorithm to cluster HDI data from 34 Indonesian provinces using Euclidean and Divergence distances. The analysis found that Divergence achieved a higher silhouette score (43.57%) compared to Euclidean (14.55%), indicating more effective and representative cluster formations. Future research may explore additional

distance metrics such as Mahalanobis or Cosine similarity. Moreover, incorporating other socio-economic indicators like poverty rates or employment ratios could enhance the robustness of clustering results. Further studies may also test the method's performance on time-series HDI data.

REFERENCE

Bagnall, A., & Janacek, G. (2005). Clustering time series with clipped data. *Machine Learning*, *58*(2–3), 151–178. https://doi.org/10.1007/s10994-005-5825-6

Basbug, M. E., & Engelhardt, B. (2015). *AdaCluster : Adaptive Clustering for Heterogeneous Data*. *17*, 1–34. http://arxiv.org/abs/1510.05491

Biabiany, E., Page, V., Bernard, D., & … (2020). Using an expert deviation carrying the knowledge of climate data in usual clustering algorithms. *ArXiv Preprint ArXiv …*, 1–9. https://arxiv.org/abs/2006.05603%0Ahttps://arxiv.org/pdf/2006.05603

Bora, M. D. J., & Gupta, D. A. K. (2014). *Effect of Different Distance Measures on the Performance of K-Means Algorithm: An Experimental Study in Matlab*. *5*(2), 2501–2506. http://arxiv.org/abs/1405.7471

Dani, A. T. R., Putra, F. B., Fauziyah, M., Sifriyani, Suyitno, & Fathurahman, M. (2023). K-Means Algorithm for grouping provinces in Indonesia based on macroeconomic and criminality indicators. *Jurnal Statistika*, *11*(2), 12–21. https://doi.org/10.14710/JSUNIMUS.11.12.-21

Fahmiyah, I., & Ningrum, R. A. (2023). Human Development Clustering in Indonesia: Using K-Means Method and Based on Human Development Index Categories. *Journal of Advanced Technology and Multidiscipline*, *2*(1), 27–33. https://doi.org/10.20473/jatm.v2i1.45070

Ha, J., Kambe, M., & Pe, J. (2011). Data Mining: Concepts and Techniques. In *Data Mining: Concepts and Techniques*. https://doi.org/10.1016/C2009-0-61819-5

Hedar, A. R., Ibrahim, A. M. M., Abdel-Hakim, A. E., & Sewisy, A. A. (2018). K-means cloning: Adaptive spherical K-means clustering. *Algorithms*, *11*(10), 1–21. https://doi.org/10.3390/a11100151

Holmström, L. (2008). Nonlinear Dimensionality Reduction by John A. Lee, Michel Verleysen. *International Statistical Review*, *76*(2), 308–309. https://doi.org/10.1111/j.1751-5823.2008.00054_10.x

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Nurhasanah, N., Salwa, N., Ornila, L., Hasan, A., & Mardhani, M. (2021). Classifying regencies and cities on human development index dimensions: Application of K-Means cluster analysis. *Jurnal Sains Sosio Humaniora*, *5*(2), 913–918. https://doi.org/10.22437/jssh.v5i2.15801

Rahardja, U., Aini, Q., & Iqbal, M. (2020). Analisis Cluster dalam Pengelompokan Provinsi di Indonesia Berdasarkan Variabel Penyakit Menular Menggunakan Metode Complete Linkage, Average Linkage dan Ward. *InfoTekJar : Jurnal Nasional Informatika Dan*

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

Teknologi Jaringan, *5*(1), 40–43.

Ram, A., Jalal, S., Jalal, A. S., & Kumar, M. (2010). A Density Based Algorithm for Discovering Density Varied Clusters in Large Spatial Databases. *International Journal of Computer Applications*, *3*(6), 1–4. https://doi.org/10.5120/739-1038

Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, *11*, 2837–2854.