
PERBANDINGAN METODE AGGLOMERATIVE HIERARCHICAL *CLUSTERING* DAN METODE KMEDOIDS DALAM PENGELOMPOKAN DATA TITIK PANAS KEBAKARAN HUTAN DI INDONESIA

St Tuhpatussania^{1*}, Surni Erniwati², Zaenul Mutaqin³

¹Teknik Informatika, Fakultas Teknologi Informasi dan Komunikasi, Universitas Teknologi Mataram, Indonesia

²Manajemen Informatika, Fakultas Vokasi, Universitas Teknologi Mataram, Indonesia

³Teknik Informatika, Fakultas Teknologi Informasi dan Komunikasi, Universitas Teknologi Mataram, Indonesia

Email: ¹asna.tuhfah@gmail.ac.id, ²mentari1990@gmail.com, ³zaenulmutaqin2012@gmail.com

SEJARAH ARTIKEL

Diterima: 27.07.2024

Direvisi: 31.07.2024

Diterbitkan: 31.07.2024



Hak Cipta © 2024

Penulis: Ini adalah artikel akses terbuka yang didistribusikan berdasarkan ketentuan Creative Commons Attribution 4.0 International License.

ABSTRAK

Rusaknya keseimbangan alam, dikarenakan hutan termasuk pemasok oksigen serta penyedia kebutuhan bagi makhluk hidup lainnya. Asap yang dihasilkan dari kebakaran hutan pun juga dapat mengganggu aktivitas kehidupan manusia. Indonesia merupakan salah satu negara yang memiliki resiko terkait kebakaran hutan dan lahan. Diperlukan suatu upaya untuk mencegah terjadinya kebakaran hutan dan lahan yaitu melalui pemantauan titik hotspot atau titik panas bumi. Pengelompokan atau *Clustering* titik panas (hotspot) dilakukan untuk memudahkan dalam memantau area-area kawasan Indonesia yang memiliki potensi rawan kebakaran tertinggi. Parameter atau feature data yang digunakan dalam proses *Clustering* yaitu *Brightness*, *Brightness_t3*, *FRP (Fire Radiative Power)* dan *Confidence*. Metode atau algoritma yang digunakan dalam proses *Clustering* titik hotspot menggunakan algoritma *Kmedoids* dan *Agglomerative Hierarchical*. Output dari kedua algoritma menghasilkan suatu *cluster* atau pengelompok daerah yang memiliki potensi kebakaran hutan rendah, sedang dan tinggi. Evaluasi pengujian menggunakan metode *Silhouette Coefficient* dan *Davies Boulding Index*. Hasil pengujian nilai *Silhouette Coefficient* menggunakan algoritma *Kmedoids* sebesar 37% sedangkan *Agglomerative Hierarchical* sebesar 37.3%. Hasil pengujian menggunakan *Davies Boulding Index* menghasilkan nilai akurasi metode *Kmedoids* sebesar 90.3% sedangkan *Agglomerative Hierarchical* sebesar 90.9%. Hal ini menunjukkan bahwa proses *Clustering* menggunakan metode *Kmedoids* dan *Agglomerative Hierarchical* memiliki tingkat accuracy yang tidak jauh berbeda.

Kata Kunci: *Clustering, Hotspot, Kmedoids, Agglomerative Hierarchical, Silhouette Coefficient.*

ABSTRACT

Balance of nature is disrupted, as forests are crucial suppliers of oxygen and providers of resources for other living beings. The smoke produced from forest fires can also disrupt human activities. Indonesia is one of the countries at risk of forest and land fires. Efforts are needed to prevent these fires, which can be done through monitoring hotspot points or thermal anomalies. Clustering hotspot points is conducted to facilitate monitoring of regions in Indonesia with the highest fire risk potential. The data parameters or features used in the Clustering process are Brightness, Brightness_t3, FRP (Fire Radiative Power), and Confidence. The Clustering process uses the Kmedoids and Agglomerative Hierarchical algorithms. The output from both algorithms results in clusters indicating areas with low, medium, and high fire risk potential. Evaluation tests use the Silhouette Coefficient and Davies-Bouldin Index methods. The test results show that the Silhouette Coefficient value using the Kmedoids algorithm is 37%, while the Agglomerative Hierarchical algorithm is 37.3%. The Davies-Bouldin Index test results show the accuracy of the Kmedoids method at 90.3% and the Agglomerative Hierarchical method at 90.9%. This indicates that the Clustering process using the Kmedoids and Agglomerative Hierarchical methods has similar levels of accuracy.

Keywords: *Clustering, Hotspot, Kmedoids, Agglomerative Hierarchical, Silhouette Coefficient.*

1. PENDAHULUAN

Kebakaran hutan merupakan salah satu bencana yang mengambil perhatian penting di beberapa negara, termasuk Indonesia. Pada tahun 2015-2020, luas area kebakaran hutan dan lahan di Indonesia seluas 5.641.337 Ha. Nilai tersebut mengalami penurunan pada tahun 2020 hingga pada 31 Juli 2021, luas area kebakaran hutan dan lahan hanya mencapai 160.104 Ha. Meskipun mengalami penurunan, berbagai upaya tetap dilakukan untuk mencegah terjadinya kebakaran hutan dan lahan di Indonesia. Salah satu upaya pencegahan dini kebakaran hutan dan lahan yaitu dengan melakukan pengelompokan atau *Clustering* terhadap titik panas bumi atau disebut dengan hotspot. Hotspot adalah suatu area yang relatif suhunya lebih tinggi jika dibandingkan dengan area sekitarnya yang dideteksi satelit. Area tersebut diwakilkan dalam suatu titik yang memiliki koordinat tertentu [1]

Keberadaan titik panas dapat mengindikasikan adanya suhu tinggi yang bisa jadi berpotensi kebakaran ataupun bisa juga titik panas tersebut hanya menunjukkan suhu tinggi tanpa potensi kebakaran [2]. Untuk itu perlu dilakukan analisis dan validasi terhadap data hotspot untuk memastikan bahwa titik hotspot tersebut benar terjadi kebakaran atau hanya indikasi saja. Investigasi secara langsung akan membutuhkan banyak waktu, terlebih lagi apabila titik lokasi hotspot memiliki jumlah yang besar. Sehingga perlu adanya proses pengelompokan data titik panas yang dapat berguna dalam penanganan dini kebakaran hutan dan lahan.

Penelitian terdahulu yang meneliti terkait kebakaran di Kabupaten Bengkalis menggunakan teknik penginderaan jauh dengan citra Sentinel-2 dan indeks *Normalized Burn Ratio (NBR)* untuk menganalisis pengaruh curah hujan dan sebaran titik panas terhadap luas area kebakaran. Dengan model threshold pada indeks dNBR, akurasi mencapai antara 68,76% hingga 81,56% pada berbagai periode kebakaran. Meskipun akurasi model cukup tinggi, penelitian tersebut menunjukkan adanya variasi hasil korelasi di berbagai provinsi, yang menunjukkan kompleksitas hubungan antara variabel-variabel tersebut [3].

Proses pengelompokan data dapat dilakukan dengan beberapa metode, diantaranya yaitu *Kmedoids* dan *Agglomerative Hierarchical*. *Kmedoids* merupakan salah satu metode *Clustering* yang berfungsi untuk memecah dataset menjadi kelompok-kelompok. Kelebihan dari metode ini, mampu mengatasi kelemahan dari metode K-Means yang sensitif terhadap outlier [4]. Sedangkan *Agglomerative Hierarchical Clustering* mengelompokkan sejumlah data berdasarkan kemiripan yang membentuk pohon hierarki dari bawah ke atas [5].

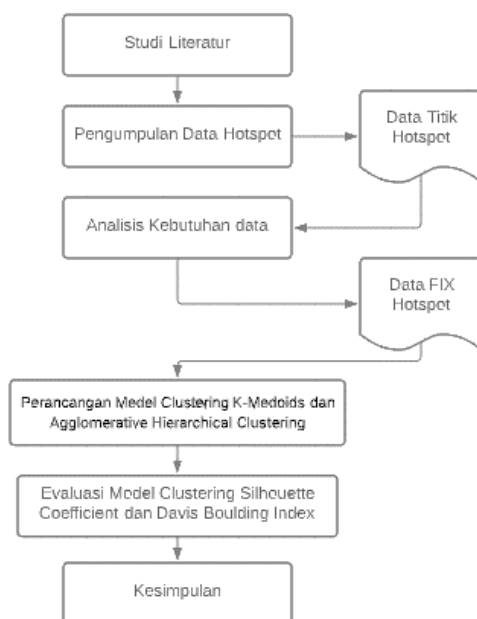
Penelitian sebelumnya yang melakukan perbandingan mengenai metode *Clustering* atau pengelompokan data memiliki tujuan untuk meningkatkan akurasi pengelompokan data dan membandingkan keefektifan metode yang berbeda. Penelitian pertama bertujuan untuk membandingkan metode *Hierarchical Clustering*, *K-Means*, *K-Medoids*, dan *Fuzzy C-Means* dalam pengelompokan provinsi di Indonesia berdasarkan indeks penanganan stunting. Metode K-Medoids ditemukan lebih unggul dalam beberapa kasus, meskipun semua metode memiliki kelebihan dan kekurangan tersendiri [6]. Penelitian lainnya membandingkan metode *Hierarchical Clustering* dan K-Means dalam pengelompokan data problem kerja praktek di Teknik Industri ITS, dan menemukan bahwa kombinasi *Single Linkage Clustering* dan K-Means menghasilkan *cluster* yang lebih baik berdasarkan parameter variance dan metode silhouette coefficient, namun metode ini memerlukan komputasi yang lebih kompleks [7].

Pengelompokan daerah titik hotspot menggunakan beberapa parameter data hotspot yang diambil dari situs yang menyediakan open dataset terkait titik kebakaran hutan di dunia yaitu NASA (<https://firms.modaps.eosdis.nasa.gov/>). Terdapat 11 fitur didalam data hotspot, yaitu yaitu *latitude*, *longitude*, *brightness*, *scan*, *track*, *acq_date*, *acq_time*, *satellite*, *confidence*, *version*, *bright_t31*, *frp*, *daynight*.

Berdasarkan permasalahan yang dijelaskan sebelumnya beserta paparan beberapa metode *Clustering*, maka dalam penelitian ini, peneliti mencoba melakukan Perbandingan Metode *Agglomerative Hierarchical Clustering* dan Metode *Kmedoids* dalam Pengelompokan Data Titik Panas Kebakaran Hutan di Indonesia. Diharapkan hasil penelitian ini dapat memudahkan untuk mengetahui daerah-daerah yang berpotensi terjadinya kebakaran hutan dan lahan. Sehingga penanggulangan bencana kebakaran hutan dan lahan dapat dilakukan lebih awal.

2. METODE PENELITIAN

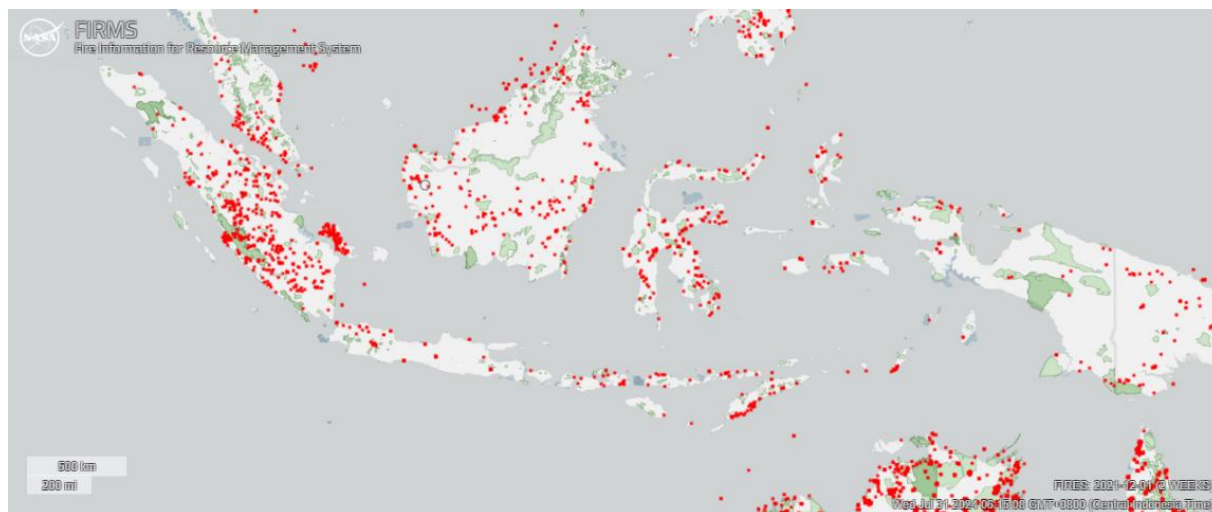
Tahapan penelitian yang digunakan pada penelitian dapat dilihat pada gambar 1, sebagai berikut :



Gambar. 1 Tahapan Metode Penelitian

2.1. Pengumpulan Data Hotspot

Pengumpulan data titik panas atau titik kebakaran hutan didapatkan dari situs dilakukan langsung dari situs resmi NASA LANCE – FIRM MODIS Active Fire (<https://firms.modaps.eosdis.nasa.gov/>) yang menyediakan.opendataset untuk titik kebakaran hutan yang dapat di filter berdasarkan negara tertentu. Data yang diperoleh dalam bentuk file *comma separated values (csv)*. Data dari NASA merupakan salah satu sumber data paling tepercaya dan diakui secara global serta pemantauan satelit menggunakan MODIS (Terra dan Aqua), yang dikenal memiliki kemampuan deteksi kebakaran yang tinggi dan menyediakan informasi yang akurat dan rinci [8]. Data berisikan informasi berupa *latitude, longitude, brightness, scan, track, acq_date, acq_time, satellite, confidence, version, bright_t31, frp dan daynight*. Data yang didapatkan dari website NASA merupakan data Titik panas untuk Wilayah Asia Tenggara dengan periode waktu dari tanggal 01 Desember 2021 hingga 14 Desember 2021.



Gambar 2. Dataset Titik Panas

2.2. Analisis Data Hotspot

Data yang didapatkan dari website NASA tersebut kemudian dianalisis terlebih dahulu. Data terlebih dahulu dilakukan proses penglabeling berdasarkan parameter confidence yang ada pada data titik hotspot. Berikut Tabel 1 *labeling confidence* :

Tabel 1. Labeling confidence

Tingkat Kepercayaan	Kelas	Tindakan
0% < c < 30%	Low	Important to note
30% < c < 80%	Medium	Warning
80% < c < 100%	Hight	Immediate Countermeasure

Setelah dilabeling berdasarkan parameter confidence, data kemudian di filter untuk meghasilkan data titik hotspot khusus daerah negara Indonesia. Untuk mendapatkan data titik hotspot negara Indonesia, maka digunakan library Geopy. Hasil dari proses analisis data hotspot akan menghasilkan data fix hotspot.

2.3. Perancangan Model Clustering

Clustering merupakan metode yang digunakan untuk mengklaster atau mengelompokkan kumpulan suatu data ke dalam sebuah himpunan atau klaster, mencari dan mengelompokkan data yang memiliki kesamaan atau kemiripan sifat atau ciri-ciri (similarity) antara data satu dengan data yang lainnya dalam sebuah dataset [9], [10].

Untuk model *Clustering* yang digunakan dalam penelitian ini yaitu, K-Medoids dan *Agglomerative Hierarchical*. Kedua metode tersebut dapat digunakan dalam python dengan menginstall library *sklearn.cluster* dan *sklearn_extra.cluster*.

Algoritma K-Medoids memiliki kelebihan dalam mengatasi kelemahan pada algoritma K-Means yang sensitive pada noise dan outlier. Dimana, objek dengan nilai besar yang memungkinkan terjadinya penyimpangan dari distribusi data. Sedangkan *Agglomerative Hierarchical* melakukan pengelompokkan data secara berulang-ulang berdasarkan pada tingkat kemiripan data satu sama lain sehingga membentuk sebuah hirarki [11].

Sebelum implementasi model, data hotspot yang telah dianalisis dan dicleaning ditransformasi terlebih dahulu menggunakan teknik normalisasi. Teknik Normalisasi merupakan proses penskalaan nilai atribut data sehingga dapat ditetapkan pada skala range tertentu [12], [13]. Proses normalisasi data dapat dilakukan dengan metode min-max, yaitu metode transformasi linear terhadap data asli. seperti pada rumus (1).

$$X' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

Dimana:

X' adalah data dinormalisasi

min (x) adalah nilai minimum dari range data normalisasi

max (x) adalah nilai maksimum dari range data normalisasi

2.4. Evaluasi Model Clustering

Hasil dari algoritma K-Medoids dan *Agglomerative Hierarchical* akan dievaluasi. Evaluasi dilakukan untuk mengetahui seberapa mirip atau dekat jarak antara objek dalam suatu *cluster* dan juga seberapa jauh jarak dari *cluster* yang berpisah dengan *cluster* lainnya [14]. Pada penelitian ini digunakan dua model evaluasi yaitu, Silhouette Coefficient dan David Boulding Index.

3. HASIL DAN PEMBAHASAN

3.1 Analisis Data Hotspot

Data hotspot yang didapatkan dalam situs resmi NASA, akan dilakukan praprocessing dan labeling. Proses pengumpulan data yang dilakukan sejak tanggal 01 Desember 2021 dan 14 Desember 2021. Dari 2216 data yang dikumpulkan, hanya 210 data yang akan digunakan. Hal itu dikarenakan fokus daerah titik hotspot pada penelitian ini di negara Indonesia. Adapun data hotspot yang telah kami lakukan pembersihan dan labeling, dapat dilihat pada Tabel 3 berikut.

Tabel 3. Data hotspot setelah dicleaning

<i>latitude</i>	<i>longi tude</i>	<i>bright ness</i>	<i>con fidence</i>	<i>bright_t 31</i>	<i>frp</i>	<i>categor y</i>	<i>action</i>	<i>location</i>
- 8.1020 899999 99999	127.2 2121 0000 0000 1	313.2 6	48	298.99	7.35	Medium	Warnin g	Wonreli, Maluku Barat Daya, Moluccas, Indonesia

- 8.22327	124.0 332	311.7 8	51	294.76	13.39	Medium	Warnin g	Blang Merang, Alor, East Nusa Tenggara, Indonesia
0.48553	127.9 9257	318.0 7	54	299.13	11.51	Medium	Warnin g	Halmahera Tengah, North Maluku, Indonesia
0.48401000000000005	128.0 029	319.3 3	55	298.97	12.58	Medium	Warnin g	Halmahera Tengah, North Maluku, Indonesia
- 2.91125	129.2 9607	315.7 2	69	297.12	7.96	Medium	Warnin g	Maluku Tengah, Moluccas, Indonesia

Kemudian melakukan pemilihan parameter yang digunakan untuk proses *Clustering* data. Pada penelitian ini digunakan empat parameter, diantaranya dapat dilihat pada tabel berikut:

Tabel 4. Parameter proses *Clustering*

brightne ss	bright_t 31	frp	confide nce
313.26	298.99	7.35	48
311.78	294.76	13.39	51
318.07	299.13	11.51	54
319.33	298.97	12.58	55
315.72	297.12	7.96	69

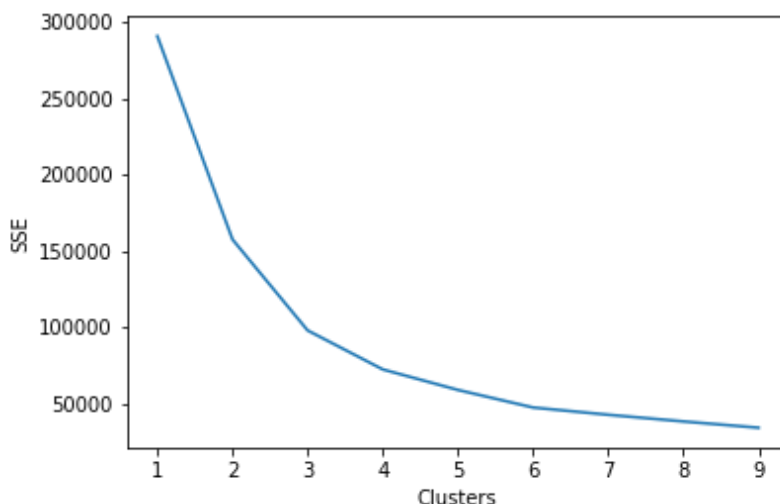
Selanjutnya data diatas diproses melalui transformasi data dengan teknik normalisasi. Dimana, data yang numerik diubah menggunakan metode min-max agar menghasilkan nilai yang memiliki skala antar 0 hingga 1. Berikut adalah hasil normalisasi data:

Tabel 5. Normalisasi data

brightness	bright_t31	frp	confidence
0.19488717	0.86469345	0.02598635	0.48
0.17153227	0.75290698	0.06182142	0.51
0.27079059	0.86839323	0.05066746	0.54
0.29067382	0.8641649	0.05701572	0.55
0.2337068	0.81527484	0.02960546	0.69
0.25059176	0.45586681	0.37881934	0.63

3.2 Penerapan Model *Clustering*

Pemilihan jumlah *cluster* menggunakan Elbow method untuk mencari jumlah *cluster* optimal [15]. Berdasarkan gambar berikut pola yang membentuk siku terdapat di *cluster* yang dapat digunakan 2, 3, 4 dan 6.



Gambar.2 Grafik Elbow Methode

Setelah pemilihan *cluster*, berikutnya implementasi model. Implementasi model dilakukan untuk mencari perbandingan dari masing-masing model untuk nilai $k=2$, $k=3$ dan $k=4$. Hasil dari implementasi model sebagai berikut:

Tabel 6. Jumlah daerah dengan nilai cluser $k=2$

K = 2	Cluster 1	Cluster 2
Kmedoids	113	97
Agglomerative hierarchical	173	37

Tabel 7. Jumlah daerah dengan nilai cluser $k=3$

K = 3	Cluster 1	Cluster 2	Cluster 3
Kmedoids	91	71	48
Agglomerative hierarchical	37	75	98

Tabel 8. Jumlah daerah dengan nilai cluser $k=3$

K = 4	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Kmedoids	87	15	37	71
Agglomerative hierarchical	75	29	98	8

Berdasarkan tabel diatas, dengan menggunakan metode Kmedoids dan $K=2$, maka daerah *cluster 1* sebanyak 113, sedangkan *cluster 2* sebanyak 97. Sedangkan menggunakan metode *Agglomerative Hierarchical* dengan nilai K yang sama, maka daerah *cluster 1* sebanyak 97 dan *cluster 2* sebanyak 37. Metode Kmedoids untuk $k=3$, maka daerah *cluster 1* sebanyak 91, *cluster 2* sebanyak 71 dan *cluster 3* sebanyak 48. Untuk metode *Agglomerative Hierarchical* dengan nilai K sama, daerah *cluster 1* sebanyak 37, *cluster 2* sebanyak 75 dan *cluster 3* sebanyak 98. Metode Kmedoids untuk $k=4$, maka daerah *cluster 1* sebanyak 87, *cluster 2* sebanyak 15, *cluster 3* sebanyak 37 dan *cluster 4* sebanyak 71. Sedangkan *Agglomerative Hierarchical* dengan daerah K yang sama, daerah *cluster 1* sebanyak 75, *cluster 2* sebanyak 29, *cluster 3* sebanyak 98 dan *cluster 4* sebanyak 8.

3.3 Evaluasi Model Clustering

Setelah dilakukan permodelan menggunakan metode *Clustering*, selanjutnya dilakukan evaluasi model. Evaluasi dilakukan dengan membandingkan nilai rata-rata dari setiap model evaluasi dan model *Clustering* serta masukkan nilai jumlah *cluster* yang berbeda [16]. Pengujian dilakukan menggunakan jumlah inputan *cluster* $k=2$, $k=3$, dan $k=4$. Untuk hasilnya dapat bisa dilihat dalam tabel berikut:

Tabel 9. Evaluasi model *Clustering*

K	Kmedoids	Agglomerative hierarchical
----------	-----------------	-----------------------------------

	Silhouette Coefficient	DBI	Silhouette Coefficient	DBI
2	37.1%	96%	47.8%	94%
3	37%	90%	37.3%	91%
4	37.1%	96%	37.5%	94%

Berdasarkan tabel diatas, jumlah *cluster* terbaik yang digunakan untuk model *Clustering Kmedoids* yaitu k=2 atau k=4. Sedangkan untuk model *Agglomerative Hierarchical*, jumlah *cluster* terbaik berada pada nilai k=2.

4. KESIMPULAN

Berdasarkan hasil pengujian dan analisis diatas, maka dapat disimpulkan bahwa model *Clustering Kmedoids* dan *Agglomerative Hierarchical* sama baik nya digunakan untuk pengelompokkan data titik hotspot Indonesia. Evaluasi model terbaik dengan menggunakan jumlah *Clustering* sebanyak k=2, dimana nilai silhouette coefficient sebesar 37.1% dan DBI sebesar 96% untuk model *Clustering Kmedoids*. Sedangkan model *Agglomerative Hierarchical*, nilai silhouette coefficient sebesar 47.8% dan DBI sebesar 94%. Nilai evaluasi model akan bertambah apabila jumlah data untuk *Clustering* semakin banyak. Dan disarankan untuk penelitian kedepannya, perlu ditambahkan parameter dari data hotspot untuk proses *Clustering*

DAFTAR PUSTAKA

- [1] I. S. Sitanggang *et al.*, "Indonesian Forest and Land Fire Prevention Patrol System," *Fire*, vol. 5, no. 5, 2022, doi: 10.3390/fire5050136.
- [2] P. A. G. Moreira, T. A. Mendes, and D. F. Dos Santos, "Assessment of potential sites for the installation of observation towers for forest fire risk prevention," *Cienc. Florest.*, vol. 30, no. 4, 2020, doi: 10.5902/1980509839686.
- [3] E. Widayanti, A. Sukmono, and F. Hadi, "Analisis Pengaruh Curah Hujan Dan Sebaran Titik Panas Terhadap Luas Area Kebakaran Hutan Dan Lahan Di Kabupaten Bengkalis Menggunakan Indeks Normalized Burn Ratio," *J. Geod. Undip*, vol. 12, no. 4, 2023.
- [4] Nurhayati, N. S. Sinatrya, L. K. Wardhani, and Busman, "Analysis of K-Means and K-Medoids's Performance Using Big Data Technology," in *2018 6th International Conference on Cyber and IT Service Management, CITSM 2018*, 2019. doi: 10.1109/CITSM.2018.8674251.
- [5] P. Bhattacharjee and P. Mitra, "A survey of density based *Clustering* algorithms," 2021. doi: 10.1007/s11704-019-9059-3.
- [6] G. R. Suraya and A. W. Wijayanto, "Comparison of Hierarchical *Clustering*, K-Means, K-Medoids, and Fuzzy C-Means Methods in Grouping Provinces in Indonesia according to the Special Index for Handling Stunting," *Indones. J. Stat. Its Appl.*, vol. 6, no. 2, 2022, doi: 10.29244/ijsa.v6i2p180-201.
- [7] T. Alfina and B. Santosa, "Analisa Perbandingan Metode Hierarchical *Clustering*, K-Means dan Gabungan Keduanya dalam Membentuk *Cluster* Data (Studi Kasus : Problem Kerja Praktek Jurusan Teknik Industri ITS)," *Anal. PerbandinganMetode Hierarchical Clust. K-means dan Gabungan Keduanya dalam Clust. Data*, vol. 1, no. 1, 2012.
- [8] L. Giglio, L. Boschetti, D. P. Roy, M. L. Humber, and C. O. Justice, "The Collection 6 MODIS burned area mapping algorithm and product," *Remote Sens. Environ.*, vol. 217, 2018, doi: 10.1016/j.rse.2018.08.005.
- [9] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means *Clustering* algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci. (Ny)*, vol. 622, 2023, doi: 10.1016/j.ins.2022.11.139.
- [10] G. J. Oyewole and G. A. Thopil, "Data *Clustering*: application and trends," *Artif. Intell. Rev.*, vol. 56, no. 7, 2023, doi: 10.1007/s10462-022-10325-y.
- [11] X. Wang, X. Wang, and D. M. Wilkes, "An Efficient K-Medoids *Clustering* Algorithm for Large Scale Data," in *Machine Learning-based Natural Scene Recognition for Mobile Robot Localization in An Unknown Environment*, 2020. doi: 10.1007/978-981-13-9217-7_5.
- [12] N. Shah and K. Shah, "Introduction to Data Mining," in *Practical Data Mining Techniques and Applications*, 2023. doi: 10.1201/9781003390220-1.
- [13] S. R. Durugkar, R. Raja, K. K. Nagwanshi, and S. Kumar, "Introduction to data mining," 2022. doi: 10.1002/9781119792529.ch1.
- [14] X. Ran, X. Zhou, M. Lei, W. Tepsan, and W. Deng, "A novel K-means *Clustering* algorithm with a noise algorithm for capturing urban hotspots," *Appl. Sci.*, vol. 11, no. 23, 2021, doi: 10.3390/app112311202.
- [15] I. M. S. Bimantara and I. M. Widiartha, "Optimization Of K-Means *Clustering* Using Particle Swarm Optimization Algorithm For Grouping Traveler Reviews Data On Tripadvisor Sites," *J. Ilm. Kursor*, vol. 12, no. 1, 2023, doi: 10.21107/kursor.v12i01.269.
- [16] B. M. Knisely and H. H. Pavliscsak, "Research proposal content extraction using natural language

processing and semi-supervised *Clustering*: A demonstration and comparative analysis,” *Scientometrics*, vol. 128, no. 5, 2023, doi: 10.1007/s11192-023-04689-3.