NORMALISASI DATA UNTUK EFISIENSI K-MEANS PADA PENGELOMPOKAN WILAYAH BERPOTENSI KEBAKARAN HUTAN DAN LAHAN BERDASARKAN SEBARAN TITIK PANAS

(DATA NORMALIZATION FOR K-MEANS EFFICIENCY ON GROUPS OF AREAS WITH POTENTIAL FORES AND /LAND FIRE BASED ON HEAT SPOTS DISTRIBUTION)

Ahmad Harmain[1)], Paiman[2)], Henri Kurniawan[3)], Kusrini[4)], Dina Maulina[5)]
[1, 2, 3, 4, 5)] MTI Universitas Amikom Yogyakarta
e-mail: mail.ahmad.harmain@gmail.com [1)], fatihways@gmail.com[2)], henribatam@gmail.com[3)], kusrini@amikom.ac.id[4)], dina.m@amikom.ac.id[5)]

ABSTRAK

Kawasan indonesia merupakan bagian dari daerah tropis yang memiliki potensi kebakaran sangat tinggi terlebih pada musim kemarau, sehingga perlunya sebuah langkah kongkrit untuk dilakukan mitigasi supaya potensi-potensi kebakaran hutan itu menjadi terminimalisir. Untuk melakukan itu dibutuhkan suatu metode teknologi yang lebih mumpuni dan terbaru untuk memetakan wilayah-wilayah yang mempunyai potensi besar terjadinya kebakaran hutan. Sistem pencitraan dan Informasi dari sistem satelit (MODIS) adalah salah satu informasi tentang kondisi permukaan bumi, yaitu parameter Latitude, Longitude, Brightness, FRP (Fire Radiative Power), dan Confidence dapat dijadikan dasar pengelompokan suatu wilayah memiliki potensi kebakaran atau tidak. K-Means adalah salah satu metode dalam machine learning yang bisa digunakan sebagai salah satu metode dalam pengelompokan wilayah-wilayah tersebut. Akurasi dalam menguji hasil pengelompokan K-Means dapat diuji dengan metode Davies Bouldin Index (DBI) dan Silhouette Coefficient.

Kata Kunci: *K-Means, MODIS, Silhouette Coefficient, Davies Bouldin Index*

ABSTRACT

The Indonesian region is part of the tropics which has a very high fire potential, especially during the dry season, so it is necessary to take concrete steps to mitigate so that the potential for forest fires is minimized. To do this, a more advanced and up-to-date technological method is needed to map areas that have a high potential for forest fires. The imaging and information system from the satellite system (MODIS) is one of the information about the condition of the earth's surface, namely the parameters of Latitude, Longitude, Brightness, FRP (Fire Radiative Power), and Confidence, which can be used as the basis for grouping an area as having a fire potential or not. K-Means is a method in machine learning that can be used as a method for grouping these areas. Accuracy in testing the results of the K-Means grouping can be tested using the Davies Bouldin Index (DBI) and Silhouette Coefficient methods.

Keywords*: K-Means, MODIS, Silhouette Coefficient, Davies Bouldin Index*

I. PENDAHULUAN

Perubahan iklim sudah menjadi isu dunia dan masalah bersama berbagai bangsa. Seluruh komponen di dunia bersama-sama berkomitmen untuk mencegah pemasan global 1.5 derajat *celcius* pada sidang PBB COP26 [1]. Salah satu hal yang menjadi penting untuk mempertahankan bumi untuk tetap dingin adalah menjaga hutan-hutan untuk tetap sebagai paru-paru dunia. Terutama di negara-negera tropis khususnya Indonesia.

Adaptasi perubahan iklim sebagai sebuah isu dunia, melibatkan pemantauan dan antisipasi tindakan pengelolaan hutan [2]. Pengelola hutan umumnya akan mencari solusi yang membahas kedua tujuan utamanya yaitu mitigasi dan adaptasi. Dalam tindakan mitigasi dibutuhkan data-data pendukung tentang titik api satu wilayah yang mempunyai potensi kebakaran tinggi.

Untuk memantau potensi titik api dapat menggunakan bantuan teknologi dengan penginderaan jauh. Salah satu fasilitas data yang bisa kita ambil adalah dari informasi satelit [3]. Salah satunya yaitu satelit *Terra/Aqua* dengan

bantuan sensor MODIS milik NASA. MODIS (*Moderate Resolution Imaging Spectroradiometer*) adalah salah satu instrumen utama yang dibawa *Earth Observing System* (EOS) *Terra satellite*, yang merupakan bagian dari program antariksa Amerika Serikat, *National Aeronautics and Space Administration* (NASA) [4].

Wilayah asia tenggara terutama Indonesia telah mengalami kenaikan titik api di beberapa wilayah, pada 27 Agustus 2019, dari 95 titik naik menjadi 266 titik pada 30 Agustus 2019 [5]. Indonesia mengalami kerugian mencapai Rp 209 Triliun selain memberikan efek perubahan iklim buat dunia.

Sebagaimana diatas pengelompokan titik api (*hotspot*), sebagai langkah awal untuk mengelompokkan wilayah-wilayah yang rawan terhadap kebakaran hutan. Adapun data diambil dari satelit MODIS Nasa yang tersedia setiap 24 jam sekali. Sehingga data yang terbaru bisa memberikan data terkini tentang kondisi titik api di beberapa wilayah asia tenggara, terutama indonesia.

II. STUDI PUSTAKA

Beberapa studi penelitian yang relevan dengan projek penelitian saat ini baik yang menngunakan algoritma K-Means dan menggunakan dataset yang sama dalam pengelompokan titik rawan kebakaran hutan dan lahan.

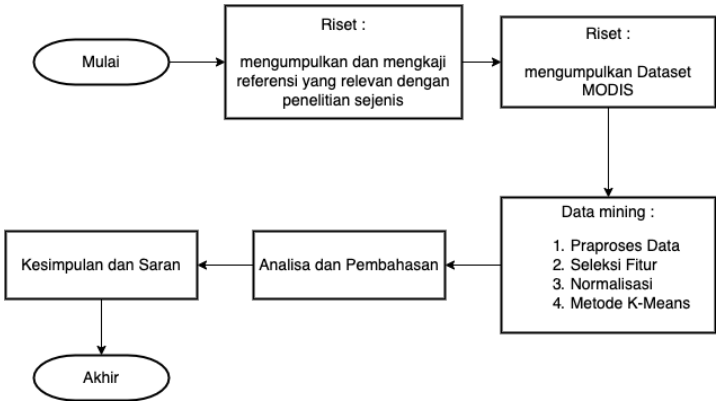
Pengelompokan daerah atau kawasan yang berstatus rawan titik api pada Provinsi Riau 2016 menggunakan *Chebysev Distance K-Means* [6], penelitian tersebut mengelompokkan data ke dalam tiga cluster diantaranya 133 titik kedalam *cluster* daerah sangat rawan, 101 titik kedalam *cluster* daerah rawan, dan 77 titik *cluster* daerah yang tidak rawan terhadap titik api.

Penelitian berikutnya dengan menggunakan dataset yang sama hasil citra Satelit MODIS parameter *latitude, longitude, brightness, fire radiation power* dan *confidence* dengan membandingkan kinerja dua algoritma antara K-Means dan K-Medoids [7]. Hasil penelitian tersebut bahwa K-Means menghasilkan skor *Silhouette Coefficient* lebih besar dari K-Medoids.

Pengelompokkan titik api menggunakan *Agglomerative Hierarchical Clustering* [8] pada penelitian tersebut menggunakan data hutan/lahan Provinsi Jambi, *Clustering* digunakan untuk mengelompokkan data menjadi 2 sampai 10 *cluster* kemudian dilakukan evaluasi dengan menghitung *Silhouette Coefficient* sehingga akan memilih koefisien terbaik dari setiap *cluster*.

III. METODE PENELITIAN

Adapun diagram alur penelitian yang menjadi dasar dalam isi dan pembahasan seperti pada gambar 1.



Gambar 1. Diagram Alur Penelitian

A. *Pengumpulan Dataset*

Dalam pengumpulan dataset yang kami pilih diperoleh dari data sekunder (pihak lain), yaitu bersumber pada data yang diambil dari Server NASA berdasarkan data Satelit MODIS [4]. Data yang diambil adalah pada data tanggal 30/12/2021. Walaupun di aplikasi API (*Application Programming Interface*) yang kita gunakan bisa memilih 24jam terakhir sampai dengan satu minggu sebelumnya.

Tabel 1. Contoh dataset Satelit MODIS.

Brightness	Confidence	Bright_t31
314.01	72	294.50
315.70	68	298.96
310.95	30	297.35
309.06	62	291.62
305.79	51	291.49
….	….	….
….	….	….
330.75	87	299.92
308.58	22	294.44
309.37	48	293.55
332.71	88	296.72
323.92	83	293.51

B. *K-Means Clustering*

Algoritma K-Means merupakan algoritma *clustering*, dan implementasi algoritma tersebut dapat

menggunakan bahasa dan *library* dari bahasa pemrograman *python*. Dalam banyak contoh aplikasi seringkali data yang dibutuhkan untuk melatih model *machine leraning* tidak berlabel meskipun tersedia dalam jumlah cukup banyak. Proses pembelajaran *machine* menggunakan data tidak berlabel disebut sebagai pembelajaran secara *unsupervised learning* [9]. Pembelajaran secara *unsupervised* berbeda dengan secara *supervised*, karena pembelajaran ini berguna untuk menemukan pola atau struktur yang tersembunyi di dalam data input. Pembelajaran *unsupervised* banyak digunakan untuk penyelesaian masalah *clustering*. *Clustering* merupakan metode pengelompokan data kedalam beberapa *cluster* atau kelompok. Data dalam satu cluster memiliki tingkat kemiripan yang maksimum dan sebaliknya data antar *cluster* yang satu dengan yang lain memiliki kemiripan yang minimum.

C. Davies Bouldin Index (DBI)

David L. Davies dan Donald W. Bouldin memperkenalkan sebuah metode yang diberi nama dengan nama mereka berdua, yaitu *Davies-Bouldin Index* (DBI) yang digunakan untuk mengevaluasi *cluster* [10]. Evaluasi menggunakan *Davies Bouldin Index* ini memiliki skema evaluasi *internal cluster*, dimana baik atau tidaknya hasil *cluster* dilihat dari kuantitas dan kedekatan antar data hasil *cluster* [11].

Semakin kecil nilai DBI diperoleh (non-negatif >= 0), maka semakin baik *cluster* yang diperoleh dari pengelompokkan menggunakan algoritma *clustering* [7].

D. Normalisasi

Tahapan praproses (*prepocessing*) dalam *data mining*, sangat diperlukan untuk mendapatkan hasil yang optimal. Dengan adanya *prepocessing* menjadikan penerapan algoritma menjadi lebih efisien. Salah satu proses *preprosesing* data adalah normalisasi. Tujuan dari normalisasi data dalam dataset adalah untuk membentuk data dalam posisi nilai dengan rentang yang sama. Karena algortima K-Means sedikit sensitif dengan adanya data *outlier* (pencilan), sehingga dengan dilakukan normalisasi deviasi dari *outlier* makan distribusi data normal. Dalam percobaan ini kita akan menguji untuk 2 metode normalisasi dalam *data mining* yaitu Normalisasi L1 dan L2.

Normalisasi L1 (*library scikit-learn*), didasarkan pada penyimpangan absolute terkecil yang bekerja dengan memastikan bahwa jumlah nilai *absolute* adalah 1 dalam setiap baris. Normalisasi L2, yang mengacu pada kuadrat terkecil bekerja dengan memastikan bahwa jumlah kuadrat adalah 1. Secara umum, teknik Normalisasi L1 dianggap lebih kuat daripada teknik Normalisasi L2. Teknik normalisasi L1 kuat karena tahan terhadap *outlier* dalam data. Berikut sebaran dataset ketika belum dilakukan normalisasi.

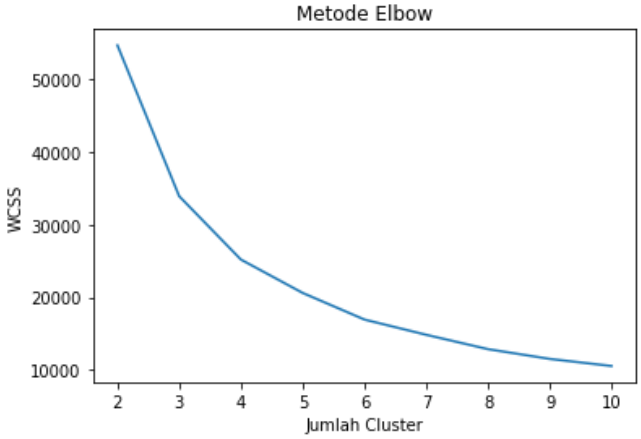
IV. HASIL DAN PEMBAHASAN

Berdasarkan hasil penelitian serta pengujian teknikal maka diperoleh beberapa hasil yang menjadi pembahasan di mana beberapa penjelasan banyak dilengkapi dengan visual berupa gambar dan tabel untuk memperkuat proses pengujian yang telah dilakukan.

A. Pengujian Jumlah Klaster

```
from sklearn.cluster import KMeans
wcss=[]
for i in range(1,11):
  kmeans=KMeans(
        n_clusters=i,
        init='k-means++',
        random_state=42)
   kmeans.fit(X)
   wcss.append(kmeans.inertia_)
```

Dalam metode *Elbow*, kita sebenarnya memvariasikan jumlah cluster (K) dari 2 – 11. Untuk setiap nilai K kita menghitung WCSS (*Within-Cluster Sum of Square*). WCSS adalah jumlah kuadrat jarak antara setiap titik dan pusat massa dalam sebuah *cluster*. Ketika kita diplot WCSS dengan nilai K, plotnya terlihat seperti *Elbow*. Dengan bertambahnya jumlah *cluster*, nilai WCSS akan mulai berkurang. Nilai WCSS terbesar ketika K=1. Ketika kita menganalisis grafik, dapat terlihat bahwa grafik akan berubah dengan cepat pada suatu titik dan dengan demikian menciptakan bentuk siku. Dari titik ini, grafik mulai bergerak hampir sejajar dengan sumbu X. Nilai K yang sesuai dengan titik ini adalah nilai K optimal atau jumlah *cluster* yang optimal. Berikut hasil pengujian data untuk metode *Elbow*.



Gambar 2. Grafik Penentuan K-*Cluster* (*Elbow*)

Dari hasil grafik *elbow* terlihat setelah pada K=4 tidak ada tekuk-an grafik sehingga artinya sudah mulai optimal. Maka pada data di atas pemilihan jumlah *cluster* (K) adalah 4. Jadi setelah lebih dari K=4 hasil akan cenderung stabil. Selanjutnya untuk melihat efisiensinya akan terlihat nanti pada hasil uji performa.

B. Model K-Means

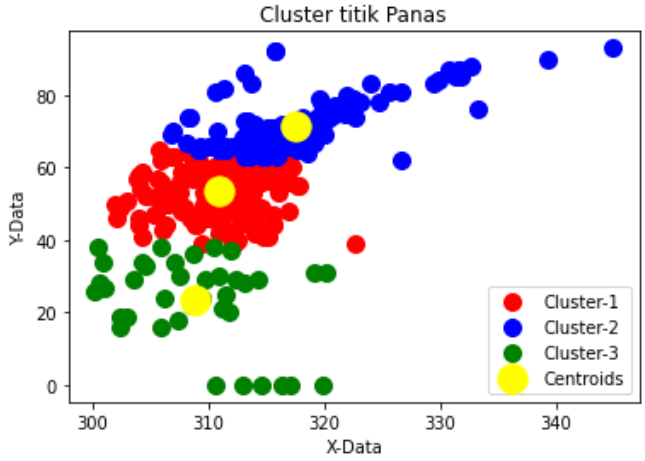
Dalam pembuatan model digunakan kelas KMeans dari *library scikit learn*.

```
kmeans = KMeans(
           n_clusters=3,
           init='k-means++',
           random_state=42
         )
y_kmeans = kmeans.fit_predict(X)
```

```
array([1, 1, 2, 0, 0, 0, 2, 2, 0, 1, 1, 2, 1, 2, 1, 0, 1, 0, 0, 0, 2, 2,
       0, 0, 0, 1, 0, 0, 1, 2, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 1, 1, 1, 0,
       1, 0, 1, 0, 2, 0, 2, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0,
       1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 1, 1, 0, 2,
       0, 1, 1, 2, 2, 1, 1, 0, 2, 1, 1, 0, 0, 0, 2, 1, 1, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1, 2, 1, 0, 0, 2, 0, 2, 0, 0, 1,
       0, 0, 0, 0, 0, 2, 1, 0, 1, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0,
       0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0,
       0, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0,
       2, 1, 0, 0, 0, 2, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 2, 1, 0, 1, 0, 1,
       1, 1, 0, 2, 1, 0, 0, 1, 0, 1, 1, 1, 2, 2, 1, 2, 0, 0, 1, 1, 0, 2,
       0, 0, 0, 0, 0, 1, 0, 1, 2, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1,
       0, 0, 2, 0, 1, 1, 1, 0, 2, 1, 0, 2, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0,
       0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1, 2, 1, 0, 1, 0, 0, 0, 2, 2, 0, 1,
       0, 1, 0, 1, 0, 1, 1, 0, 2, 1, 1, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 1,
       0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1, 0, 0, 2, 0, 1, 1, 2, 0, 1,
       1], dtype=int32)
```

Gambar 3. Hasil *cluster* data

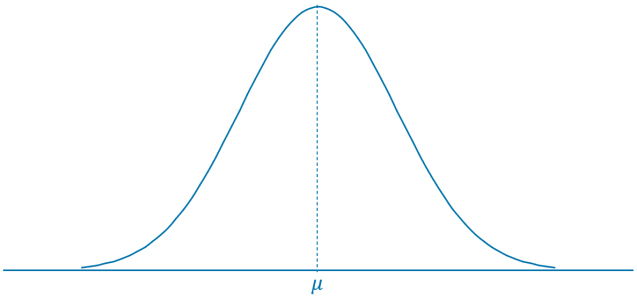
Mula – mula inisialisasi model algoritma K-Means dengan memberikan parameter jumlah *cluster* adalah 3.



Gambar 4. Bentuk sebaran antar *cluter*

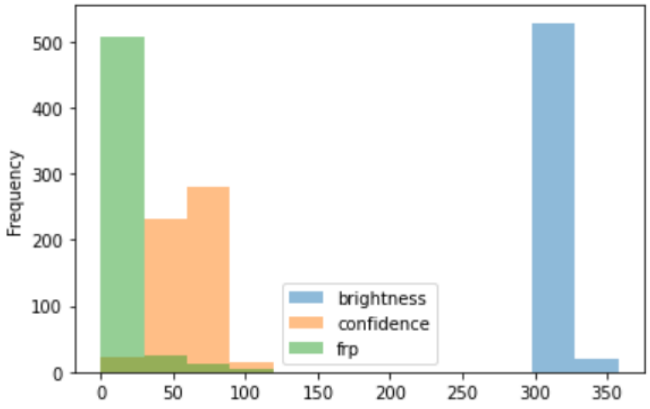
C. Pengujian Normalisasi

Normalisasi merupakan cara transformasi data agar distribusinya menjadi normal. Jadi data yang perlu dilakukan normalisasi harus dipastikan bahwa memang datanya belum berdistribusi normal. Distribusi normal itu secara sederhana digambarkan dalam kurva distribusi bentuknya cenderung seperti lonceng simetris.



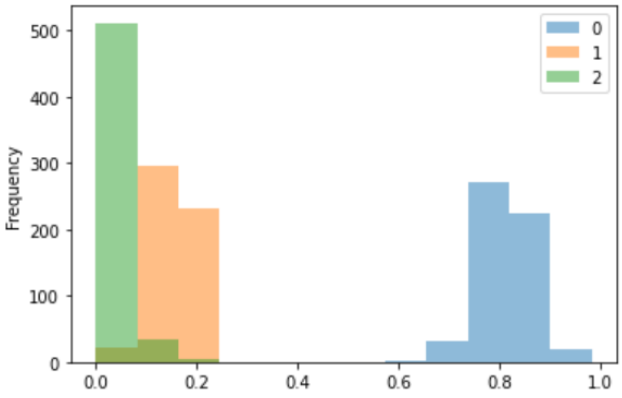
Gambar 5. Grafik distribusi normal

Untuk membandingkan apakah dataset yang kita lakukan normalisasi bisa memberikan distribusi atau bentuk yang yang lebih cenderung normal dibandingkan dengan yang tanpa normalisasi bisa menggunakan histogram atau bisa juga menggunakan plot data. Berikut perbandingan sebaran data sebelum Normalisasi, Normalisasi L1 dan Normalisasi L2.



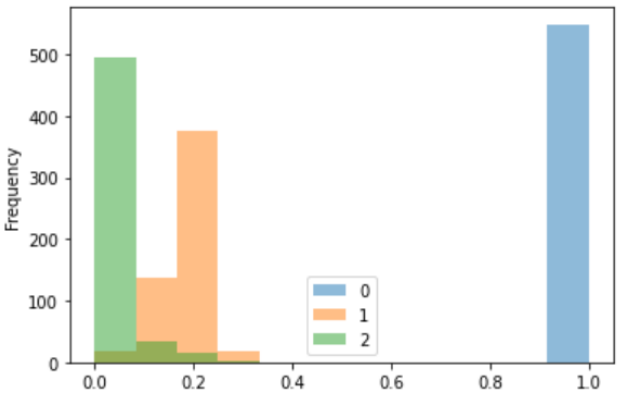
Gambar 6. Grafik distribusi data sebelum Normalisasi

Dari Gambar 6. terlihat distribusi datanya cenderung menceng ke kiri atau menceng negatif terutama pada data *brightness*.



Gambar 7. Grafik Distribusi data setelah Normalisasi L1

Setelah dilakukan Normalisasi L1 terlihat untuk distribusi datanya cendrung lebih seimbang dan menuju ke normal terutama terlihat pada data *brightness* pada kurva warna biru. Data yang lain juga secara grafik lebih bagus distribusinya walaupun tidak seperti signifikan yang data *brightness*.



Gambar 8. Grafik distribusi data setelah Normalisasi L2

Pada Normalisasi data L2 juga terlihat distribusinya jauh lebih bagus dibandingkan dengan tanpa normalisasi.

D. Pengujian Performa

```
from sklearn.metrics import davies_bouldin_score
results={}
for i in range(2,11):
  kmeans = KMeans(
         n_clusters=i,
         random_state=30
         )
  labels=kmeans.fit_predict(X)
  db_index=davies_bouldin_score(
           X,
           labels)
  results.update({i:db_index})
```

Untuk menguji performa algoritma *clustering* kita gunakan *Davies Bouldin Index* (DBI) untuk beberapa *cluster* antara 2 sampai 4. Dimana akan kita lihat nilai DBI nya apakah setelah di akukan normalisasi ada perbaikan, dimana nilai DBI yang bagus adalah mendekati nilai 0. Berikut perbandingan data-data DBI sebelum normalisasi dan setelah dilakukan normalisasi L1 dan L2.

Tabel 2. Perbandingan DBI sebelum dan setelah normalisai L1 dan L2

Cluster	DBI Tanpa Normalisasi	DBI (Normal L1)	DBI (Normal L2)	Efisiensi (L1)	Efisiensi (L1)
2	0.7777659124086238	0.8834960803826888	0.9622571843844187	-0.13%	-0.24%
3	0.7990618543943127	0.6961881111379767	0.6881686575199727	0.13%	0.14%
4	0.8546964498023958	0.7032974907656754	0.7542798669760952	0.18%	0.12%
5	0.8782098168768172	0.691155376536404	0.7786404989399143	0.2%	0.11%

Bedasarkan data tabel 2. hasil DBI pemilihan *cluster* terbaik dari Normalisasi L1 dan Normalisasi L2 adalah pada *cluster* K=3 yang cenderung hasilnya stabil dan hampir sama efisiensinya.

V. KESIMPULAN

Dari hasil percobaan data tentang titik *hotspot* untuk metode *clustering* dari data Satelit MODIS, ditarik kesimpulan bahwa praproses dalam bentuk normalisasi ternyata memang bisa meningkatkan efisiensi algoritma yakni K-Means untuk proses *clustering* titik api potensi kebakaran hutan.

Proses *clustering* yang paling penulis sarankan berdasar data DBI adalah pada pemilihan *cluster* K=3. Karena terlihat baik normalisasi L1 dan L2 hasil efisiensinya stabil. Berdasarkan grafik *elbow* proses uji *clustering* disarankan adalah di bawah K=5 karena setelah K=5 cenderung hasil sudah stabil.

Hasil pengujian pada penelitian ini bisa dijadikan salah satu acuan apabila akan dipakai untuk pembuatan aplikasi yang lebih sempurna bagi para pemangku kebijakan terutama dalam penanggulangan bencana kebakaran hutan dan lahan.

DAFTAR PUSTAKA

[1] A. C. Rahayu, "COP26 Sepakati Jaga Pemanasan Global Tidak Lebih dari 1,5 Derajat Celcius," *kontan.co.id*, 2021. https://internasional.kontan.co.id/news/cop26-sepakati-jaga-pemanasan-global-tidak-lebih-dari-15-derajat-celcius (accessed Dec. 25, 2021).

[2] Rodney J. Keenan, "Climate Change Impacts and Adaptation in Forest Management: a review," *Annals of Forest Science 72*, pp. 145–167, 2015, Accessed: Dec. 26, 2021. [Online]. Available: https://link.springer.com/article/10.1007/s13595-014-0446-5

[3] K. A. DS, P. Sofan, S. Suwarsono, I. Prasasti, and F. Yulianto, *Evaluasi Hasil Estimasi Suhu Udara dari Data Satelit NOAA-18 AVHRR di Pulau Sumatera, Kalimantan dan Jawa*. JAKARTA TIMUR: LAPAN, 2015. Accessed: Dec. 24, 2021. [Online]. Available: https://onesearch.id/Record/IOS4589.slims-4617

[4] "About MODIS (Moderate Resolution Imaging Spectroradiometer)," *MODIS (Moderate Resolution Imaging Spectroradiometer)*. https://modis.gsfc.nasa.gov/about/ (accessed Dec. 25, 2021).

[5] M. Armani and I. D. Wedhaswary, "Data Terkini Titik Panas di Indonesia dan Wilayah Asia Tenggara," *kompas.com*, 2019. Accessed: Dec. 25, 2021. [Online]. Available: https://www.kompas.com/tren/read/2019/09/07/084506865/data-terkini-titik-panas-di-indonesia-dan-wilayah-asia-tenggara?page=all

[6] T. R. A. Sukamto Sukamto Ibnu Daqiqil Id, "Penentuan Daerah Rawan Titik Api Di Provinsi Riau Menggunakan Clustering Algoritma K-Means," *JUITA*, vol. 6, no. 2, pp. 137–148, 2018, [Online]. Available: http://jurnalnasional.ump.ac.id/index.php/JUITA/article/view/3172

[7] A. Athifaturrofifah, R. Goejantoro, and D. Yuniarti, "Perbandingan Pengelompokan K-Means dan K-Medoids Pada Data Potensi Kebakaran Hutan/Lahan Berdasarkan Persebaran Titik Panas," *EKSPONENSIAL*, vol. 10, no. 2, pp. 143–152, 2020, [Online]. Available: http://jurnal.fmipa.unmul.ac.id/index.php/exponensial/article/view/572

[8] U. K. Krisman Pratama Simanjuntak, "Pengelompokkan Titik Api di Provinsi Jambi dengan Algoritma Agglomerative Hierarchical Clustering," *MALCOM*, vol. 1, no. 1, pp. 7–16, 2021, [Online]. Available: https://journal.irpi.or.id/index.php/malcom/article/view/6/9

[9] P. Dangeti, *Statistics for Machine Learning: Techniques for exploring supervised, unsupervised, and reinforcement learning models with Python and RJuly 2017*. Packt Publishing, 2017.

[10] S. Nawrin, M. R. Rahman, and S. Akhter, "Exploreing K-Means with Internal Validity Indexes for Data Clustering in Traffic Management System," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 3, 2017, doi: 10.14569/IJACSA.2017.080337.

[11] A. Bates and J. Kalita, “Counting Clusters in Twitter Posts,” in *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies*, 2016, pp. 1–9.