Volume 3 Nomor 1
Maret 2025

Studi Klasterisasi Data LSM dengan Algoritma K-Means Menggunakan RapidMiner

NGO Data Clustering Study with K-Means Algorithm Using RapidMiner

Adel Nayyan Amiva[1], Hasbi Firmansyah[2]
[1,2] Universitas Pancasakti Tegal
Corresponding author : adelnayanamiva@gmail.com

Abstrak

Penelitian ini berfokus pada pengorganiz data dari Lembaga Swadaya Masyarakat (LSM) dengan menggunakan teknik pengelompokan *K-Means*, yang diimplementasikan melalui perangkat lunak *RapidMiner*. Metode ini berguna untuk menemukan pola dan hubungan dalam kumpulan data ekstensif yang tidak memiliki kategori yang ditetapkan. *RapidMiner* memungkinkan analisis data yang efisien tanpa memerlukan keterampilan pemrograman tingkat lanjut, sedangkan algoritma *K-Means* mengelompokkan data dengan mengidentifikasi kesamaan karakteristik menggunakan jarak *Euclidean*. Kualitas klaster yang terbentuk dievaluasi dengan Indeks *Davies-Bouldin* (DBI), yang mengukur efektivitas klaster. Temuan penelitian ini diharapkan dapat menawarkan wawasan berharga dalam mengelola LSM, bersama dengan menunjukkan penggunaan teknik pengelompokan yang berlaku di bidang lain.

Kata Kunci : data mining, klasterisasi, K-Means, RapidMiner, Davies-Bouldin Index

PENDAHULUAN

Lembaga Swadaya Masyarakat (LSM) adalah salah satu komponen kunci dalam penerapan tata kelola yang baik. Di berbagai negara, LSM berperan signifikan sebagai mitra pemerintah dalam menyediakan layanan publik yang memadai untuk masyarakat yang memerlukan, termasuk mereka yang membutuhkan bantuan, terutama korban konflik bersenjata, kelaparan, dan bencana alam. Terkadang organisasi tersebut juga dikenal sebagai lembaga bantuan. LSM memiliki peran penting dalam pembangunan di sektor sosial, ekonomi, dan lingkungan, sering kali mereka harus mengelola data yang besar dan rumit. Pengelolaan data ini menjadi tantangan, terutama dalam menemukan pola, hubungan, atau segmen yang bisa membantu mengambil keputusan strategis. Maka dari itu, diperlukan pendekatan yang efektif untuk mengeksplorasi dan menganalisis data yang dikelola oleh LSM guna meningkatkan efisiensi operasi dan dampak dari program yang dilaksanakan.

Salah satu cara untuk menangani informasi ini adalah melalui metode pengelompokan. Pengelompokan adalah teknik yang digunakan dalam penambangan data yang berfokus pada pengorganisasian item dalam kumpulan data sesuai dengan fitur umumnya, tanpa bergantung pada label atau klasifikasi yang ditetapkan. Dari sekian banyak algoritma pengelompokan yang ada, *K-Means* menonjol sebagai pilihan yang disukai karena keterusterangan dan efektivitasnya dalam memproses data yang luas. Algoritma ini berfungsi dengan mengelompokkan data ke dalam kluster berdasarkan

kedekatan titik data dengan pusat klaster, menghasilkan grup yang serupa dalam setiap kluster tetapi berbeda dari yang ada di kluster lain.

Penelitian ini menekankan penggunaan algoritma *K-Means* untuk mengkategorikan data LSM melalui alat analisis *RapidMiner*. *RapidMiner* berfungsi sebagai *platform* analitik yang menawarkan antarmuka langsung di samping fungsionalitas canggih yang membantu dalam proses penambangan data. Dalam penelitian ini, *RapidMiner* digunakan untuk meningkatkan investigasi data, memilih parameter, dan mengevaluasi hasil pengelompokan. Diperkirakan bahwa hasil pengelompokan akan memberikan perspektif baru yang membantu manajer LSM dalam membuat pilihan strategis. Secara teoritis, penelitian ini didasarkan pada konsep pengelompokan dasar, terutama mengenai algoritma *K-Means*, yang mencakup pemahaman jarak *Euclidean* dan proses berulang untuk mengidentifikasi titik pusat yang stabil. Selain itu, evaluasi hasil pengelompokan dilakukan dengan menggunakan metrik seperti Indeks *Davies-Bouldin* (DBI) untuk menentukan keakuratan klaster yang dibuat. Oleh karena itu, penelitian ini tidak hanya memperdalam pemahaman penerapan algoritma *K-Means* pada data LSM tetapi juga menampilkan contoh aplikasi yang dapat dimanfaatkan dalam berbagai pengaturan lainnya.

METODE

1. **Data Mining**

   Data Mining mengacu pada teknik menemukan pola, wawasan, atau informasi berguna dari kumpulan data yang luas dengan menggunakan metode, strategi, algoritma, atau teknologi spesifik yang berbeda. Proses ini mencakup pemanfaatan berbagai pendekatan, termasuk kecerdasan buatan, pembelajaran mesin, dan metode statistik, untuk menganalisis data dengan cara yang menghasilkan pengetahuan yang dapat ditindaklanjuti.

   Data Mining termasuk dalam proses Penemuan Pengetahuan dalam *Database* yang lebih luas, di mana metode atau algoritma yang dipilih secara signifikan ditentukan oleh tujuan analisis dan sifat data. Ini mengacu pada berbagai bidang dan metodologi yang sudah ada, menunjukkan bahwa penambangan data bukanlah konsep yang sepenuhnya baru, tetapi kemajuan dari bidang lain seperti statistik, kecerdasan buatan, dan sistem manajemen data. Tujuan utama penambangan data adalah untuk mengungkap pola atau koneksi tersembunyi yang dapat menawarkan perspektif baru dari informasi yang disimpan dalam database yang luas.

2. **K-Means**

   K-Means adalah teknik pembelajaran tanpa pengawasan yang digunakan untuk pengelompokan data. Metode ini mengatur data ke dalam berbagai kelompok dengan memanfaatkan jarak *Euclidean*, yang membantu mengukur seberapa dekat titik data satu sama lain. *K-Means* sangat berharga untuk mengatur

kumpulan data yang luas untuk mengungkap pola yang mungkin tidak langsung terlihat. Ini sering digunakan dalam analisis data penjualan untuk mendapatkan wawasan tentang kelompok pelanggan yang berbeda. Karena efektivitasnya dalam menangani kumpulan data besar dan mengungkap pola yang mendasarinya, *K-Means* adalah salah satu metode pengelompokan yang paling banyak digunakan. Dalam penelitian ini, *K-Means* diterapkan untuk menganalisis dan mengkategorikan data dari LSM, yang bertujuan untuk menawarkan sudut pandang strategis yang membantu manajemen organisasi LSM.

$$D(i,j) = \sqrt{(X1i - X1j)^2 + (X2i - X2j)^2 + \ldots + (Xki - Xkj)^2}$$

Dimana :

D(i,j) = Jarak data ke i ke pusat cluster j
Xki = Data ke i pada atribut data ke k
Xkj = Titik pusat ke j pada atribut ke k

gambar 1.Rumus K-means

2. Clustering

Clustering adalah teknik analisis data yang digunakan untuk mengatasi masalah pengelompokan. Tujuan dari pendekatan ini adalah untuk mengelompokkan objek, seperti orang, kejadian, atau benda, ke dalam kategori tertentu berdasarkan kesamaan yang dimiliki atribut-atribut tersebut. *Clustering* juga sering disebut sebagai segmentasi, karena metode ini dapat mengidentifikasi kelompok-kelompok dalam suatu dataset dengan cara memisahkan data ke dalam beberapa kategori berdasarkan sifat yang ada. Hubungan antar anggota dalam satu kelompok (kluster) biasanya sangat erat, sedangkan hubungan antara anggota dari kelompok yang berbeda umumnya kurang kuat.

Proses *clustering* melibatkan pembagian sejumlah data menjadi beberapa kelompok dengan mengukur kesamaan atribut di antara data tersebut. Objek-objek yang memiliki karakteristik serupa akan dikelompokkan dalam kluster yang sama, sedangkan objek-objek yang memiliki karakteristik berbeda akan digolongkan ke kluster lainnya. Hasilnya adalah pengelompokan yang menampilkan hubungan yang lebih dekat di dalam kluster serta perbedaan yang jelas antara kluster-kluster yang ada.

3. Rapidminer

RapidMiner adalah aplikasi yang dibuat oleh Dr. Markus Hofmann dari *Institute of Technology Blanchardstown* dan Ralf Klinkenberg dari *rapid-i.com*. Program ini memiliki desain intuitif yang memfasilitasi interaksi pengguna. Ini adalah perangkat lunak sumber terbuka yang dibuat di *Java* dan dilisensikan di bawah Lisensi Publik GNU, membuatnya kompatibel dengan semua sistem operasi. Pengguna dapat menggunakan *RapidMiner* tanpa memerlukan keterampilan pemrograman yang ekstensif, karena semua fungsi yang diperlukan sudah

disertakan. Fokus perangkat lunak ini adalah pada pemrosesan data. Ini mencakup berbagai model komprehensif, seperti Model *Bayesian*, Pemodelan, Pohon Induksi, Jaringan Saraf, dan lain-lain. *RapidMiner* menawarkan berbagai teknik termasuk klasifikasi, pengelompokan, asosiasi, dan banyak lagi. Jika pengguna tidak dapat menemukan algoritma tertentu di Weka, mereka memiliki opsi untuk memasukkan modul tambahan. Karena Weka adalah *open-source*, siapa pun diizinkan untuk berkontribusi pada kemajuan perangkat lunak ini.

HASIL DAN PEMBAHASAN

1. Pemilihan Data

Pemilihan sumber data dilakukan pada website kaggle.com dengan nama *Clustering & PCA Assignment* yang memiliki fitur sebagai berikut:

a. *Title of the nation,*
b. *Number of deaths among infants under five years per 1000 live births,*
c. *Trade of commodities and services; The percentage of goods and services exported relative to the overall GDP;*
d. *Trade of commodities and services, expressed as a percentage of the overall GDP;*
e. *Average earnings per individual;*
f. *The calculation of the yearly increase in total GDP;*
g. *Average lifespan of a newborn if current death rates continue as they are;*
h. *The total number of children a woman would have if the existing age-specific fertility rates persist.*

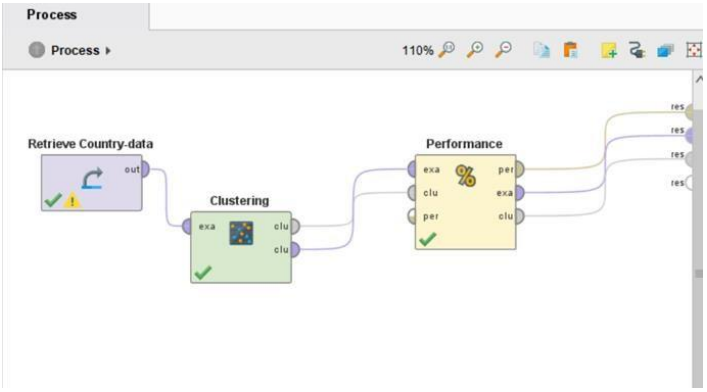
Data yang diperoleh sebanyak 167 data, tapi yang ditampilkan pada tabel hanya beberapa data saja.

country	child mort	exports	health	imports	income	inflation	life expec	total fer	gdpp
Afghanistan	90.2	10	7.58	44.9	1610	9.44	56.2	5.82	553
Albania	16.6	28	6.55	48.6	9930	4.49	76.3	1.65	4090
Algeria	27.3	38.4	4.17	31.4	12900	16.1	76.5	2.89	4460
Angola	119	62.3	2.85	42.9	5900	22.4	60.1	6.16	3530
Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200
Argentina	14.5	18.9	8.1	16	18700	20.9	75.8	2.37	10300
Armenia	18.1	20.8	4.4	45.3	6700	7.77	73.3	1.69	3220
Australia	4.8	19.8	8.73	20.9	41400	1.16	82	1.93	51900
Austria	4.3	51.3	11	47.8	43200	0.873	80.5	1.44	46900
Azerbaijan	39.2	54.3	5.88	20.7	16000	13.8	69.1	1.92	5840

Tabel 1.dataset LSM

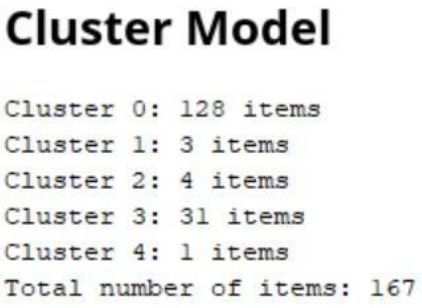
2. Preprocessing

Setelah menyelesaikan proses pemilihan data, langkah berikutnya adalah melakukan persiapan data. Tujuan dari persiapan data adalah untuk menghapus atribut yang tidak dibutuhkan, mengatasi data yang kosong atau hilang, mengeliminasi data yang tidak konsisten, serta menambahkan *ID* ke dalam dataset yang akan digunakan. Dalam tahap persiapan data ini, penulis menerapkan dua jenis *segmentasi*, yaitu *k-means* dan *cluster distance performance*.



gambar 2.Preprocessing Data

3. Pemilihan Atribut



gambar 3.Cluster Model

Pada *Cluster* model yang dihasilkan dari penghitungan *Rapidminer* yaitu klaster 0: 128 *items*, klaster 1: 3 *items*, klaster 2: 4 *items*, klaster 3: *31 items*, klaster 4: *items*, maka data keseluruhan yang telah dihitung yaitu: 167

4. Jarak Centroid

Berikut merupakan jarak antar centroid cluster dari data LSM menggunakan algoritma *K-means* bisa dilihat pada **gambar4.**

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
child_mort	47.390	3.500	8.175	8.806	9
exports	35.866	92.900	102.950	49.084	62.300
health	6.332	9.583	3.272	9.164	1.810
imports	45.922	74.600	74	45.455	23.800
income	8569.242	69833.333	71375	36977.419	125000
inflation	8.958	3.296	10.088	3.089	6.980
life_expec	67.873	81.500	78.625	79.242	79.500
total_fer	3.263	1.700	1.768	1.947	2.070
gdpp	4438.391	89133.333	38850	35606.452	70300

gambar 4.Jarak Centroid

5. Implementasi Algoritma K-Means Dengan Rapidminer

Langkah pertama yang dilakukan adalah memasukan data ke aplikasi *Rapidminer*. Lalu dilakukan *preparation* dengan mengambil data yang hanya

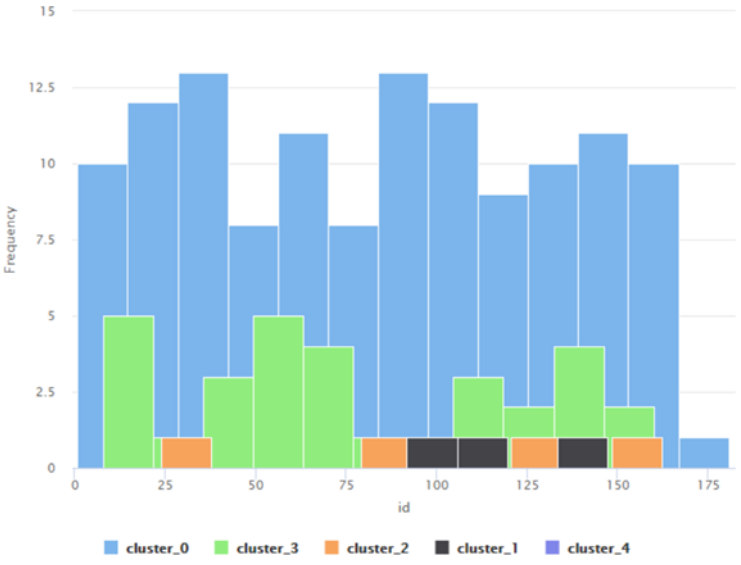
bersifat *numeric*. Setelah dilakukan *preprocessing* data, dicari iterasi yang telah dihitung oleh aplikasi. Yang didapat hasilnya sebagai berikut

PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: -93164821.393
Avg. within centroid distance_cluster_0: -63019828.352
Avg. within centroid distance_cluster_1: -401703628.770
Avg. within centroid distance_cluster_2: -94411119.883
Avg. within centroid distance_cluster_3: -190620379.926
Avg. within centroid distance_cluster_4: -0.000
```

gambar 5.Performance Vektor

6. Analisis Rekap Cluster



gambar 6.Plot Grafik

Uraian diagram di atas menunjukan nilai *clustering* untuk k=5 dihasil pengelompokan dengan klaster 0-4 memperoleh jumlah 167 item. Dapat dilihat pada **Gambar 4**.

7. Uji Validasi

Uji validasi yang diterapkan adalah *Davies Bouldin Index* (DBI). Dari hasil uji DBI, kita bisa mengetahui seberapa efektif proses pengelompokan yang telah dilakukan; semakin rendah nilai yang diperoleh, semakin baik kualitas pengelompokan tersebut. Pengujian ini dilakukan satu kali. Setelah menjalani pengujian dengan pengaturan k=5, nilai yang didapatkan untuk DBI atau *Davies Bouldin Index* dapat dilihat pada **Gambar 7**.

gambar 7.Davies Bouldin

KESIMPULAN

Penelitian ini menjelaskan bagaimana algoritma *K-Means* digunakan untuk menyortir data LSM dengan bantuan perangkat lunak *RapidMiner*. Tahap eksplorasi data dipandang penting untuk mengidentifikasi pola dan wawasan yang berguna dari kumpulan data besar, yang dapat meningkatkan efisiensi operasional dan membantu dalam perencanaan strategis dalam sektor LSM. Metode *K-Means* telah terbukti secara efektif mengatur data berdasarkan sifat bersama, sementara *RapidMiner* menawarkan antarmuka yang mudah digunakan yang memungkinkan analisis dilakukan tanpa memerlukan keahlian pemrograman tingkat lanjut. Menilai hasil pengelompokan dengan *Indeks Davies-Bouldin* menunjukkan bahwa teknik ini dapat menghasilkan klaster berkualitas tinggi. Studi ini menambah pengetahuan teoritis dan praktis tentang pengelompokan data, yang juga dapat digunakan di bidang lain untuk meningkatkan analisis data.

DAFTAR PUSTAKA

Hartini, T., Purnamasari, A. I., Bahtiar, A., & Kaslani, K. (2025). IMPLEMENTASI ALGORITMA K-MEANS CLUSTERING UNTUK KLASIFIKASI DIAGNOSIS PENYAKIT LAMBUNG. JATI (Jurnal Mahasiswa Teknik Informatika), 9(1), 76-83.

Romly, M. Z., & Fatah, Z. (2024). ANALISIS POLA PEMBELIAN PRODUK KEBERSIHAN DI SWALAYAN SALAFIYAH SYAFI'IYAH MENGGUNAKAN ALGORITMA K-MEANS. Jurnal Riset Sistem Informasi, 1(4), 86-93.

Setyaningtyas, S., Nugroho, B. I., & Arif, Z. (2022). TINJAUAN PUSTAKA SISTEMATIS PADA DATA MINING: STUDI KASUS ALGORITMA K-MEANS CLUSTERING. J Teknoif Teknik Informatika, 10, 52-61.

Sudarsono, B. G., Leo, M. I., Santoso, A., & Hendrawan, F. (2021). Analisis Data Mining Data Netflix Menggunakan Aplikasi Rapid Miner. JBASE-Journal of Business and Audit Information Systems, 4(1).

Yusuf, R. R. (2021). Globalisasi dan Akuntabilitas Lembaga Swadaya Masyarakat (LSM). Jurnal Pembangunan dan Administrasi Publik.