Volume 3 Nomor 2
September 2025

Penerapan K-Means Clustering untuk Segmentasi Dataset ulasan wisata HolidayIQ

Application of K-Means Clustering for Segmentation of HolidayIQ travel review dataset

Yoga Adhi Pralampitho[1], Hasbi Firmansyah[2], Amat Damuri[3], Iwan Mulyana[4]
[1,2] Universitas Pancasakti Tegal, Kota Tegal
[3,4] STMIK Al-Muslim Bekasi, Bekasi
Corresponding author : adhiyhoga64@gmail.com

Abstrak

Penelitian ini mengeksplorasi penggunaan algoritma K-means untuk membagi data ulasan tentang pariwisata yang diambil dari platform HolidayIQ. Dataset ini mencakup informasi mengenai minat wisatawan dalam enam kategori berbeda: alam, budaya, relaksasi, petualangan, kehidupan malam, dan belanja. Tujuan utama dari penelitian ini adalah untuk mengelompokkan wisatawan yang memiliki ciri-ciri yang sama, sehingga pola preferensi mereka dapat dianalisis lebih mendalam. Proses pengelompokan dilakukan dengan menerapkan algoritma K-means, dengan penekanan pada identifikasi jumlah kluster yang ideal serta menganalisis karakteristik setiap kluster berdasarkan atribut yang relevan. Hasil penelitian diharapkan dapat memberikan wawasan yang lebih komprehensif bagi pihak pengelola destinasi wisata dalam merancang strategi pemasaran dan pengembangan layanan yang lebih tepat sasaran, serta menjadi acuan dalam pengambilan keputusan berbasis data.

Kata Kunci : K-Means, Clustering, RapidMiner, HolidayIQ, Segmentasi Dataset.

PENDAHULUAN

Perkembangan pesat teknologi informasi telah mengubah cara wisatawan menemukan dan berbagi informasi tentang tujuan perjalanan mereka. Digitalisasi mendorong peralihan perilaku wisatawan dari penggunaan brosur atau rekomendasi konvensional menuju pencarian informasi berbasis internet. Dengan akses yang luas dan cepat, wisatawan kini dapat mengevaluasi destinasi secara lebih objektif melalui berbagai sumber digital. Salah satu kontribusi terpenting teknologi pada sektor pariwisata adalah munculnya ulasan online yang dipublikasikan melalui platform seperti HolidayIQ. Platform ini menyediakan ruang bagi wisatawan untuk berbagi pengalaman pribadi setelah mengunjungi suatu destinasi wisata. Informasi yang dipublikasikan bersifat langsung dan berasal dari pengalaman nyata, sehingga lebih dipercaya dibandingkan materi promosi resmi.

Ulasan tersebut mencerminkan pengalaman, kepuasan, dan preferensi wisatawan mengenai berbagai aspek destinasi seperti fasilitas, pemandangan, pelayanan, maupun harga. Setiap ulasan dapat memberikan perspektif yang berbeda dan memperkaya pemahaman terhadap kualitas suatu destinasi. Selain itu, ulasan juga dapat berfungsi sebagai sumber evaluasi bagi pengelola wisata untuk meningkatkan layanan (Al-Fahmi, 2023).

Seiring dengan meningkatnya jumlah ulasan yang tersedia setiap hari, analisis data ini menjadi semakin penting untuk menghasilkan wawasan yang dapat membantu pengelola destinasi, operator pariwisata, dan wisatawan sendiri dalam mengambil keputusan. Pengguna dapat menentukan pilihan destinasi terbaik berdasarkan kebutuhan dan minat tertentu, sedangkan penyedia layanan dapat menyesuaikan strategi pengembangan berdasarkan preferensi pengunjung. Namun, seiring bertambahnya jumlah dan variasi tinjauan online, analisis manual menjadi tidak efisien dan sulit dilakukan. Penumpukan data dalam jumlah besar menyebabkan proses analisis membutuhkan pendekatan komputasi yang lebih terstruktur dan otomatis. Tanpa dukungan teknologi, informasi berharga dalam data tersebut berpotensi tidak dimanfaatkan secara optimal (Hailong, 2021).

Oleh karena itu, diperlukan metode untuk mengelompokkan data tersebut secara efektif dan mengidentifikasi preferensi wisatawan dengan lebih jelas. Salah satu teknik analisis yang umum digunakan adalah clustering, yaitu proses yang memungkinkan data dikelompokkan berdasarkan kesamaan atribut tertentu. Teknik ini dapat mengungkap pola tersembunyi yang tidak dapat terlihat melalui perhitungan sederhana.

Dengan memahami preferensi wisatawan melalui segmentasi hasil clustering, diharapkan pengelompokkan ini dapat memberikan manfaat langsung seperti mendukung strategi pemasaran yang lebih tepat sasaran, mempersonalisasi pengalaman wisata, meningkatkan kualitas layanan destinasi, dan mendukung pengambilan keputusan berbasis data dalam pengembangan sektor pariwisata secara keseluruhan.

METODE

A. Data Mining

Salah satu teknik dalam penelitian ini meliputi penerapan teknik data mining untuk menganalisis kumpulan data yang tersedia. Penambangan data adalah proses yang menggabungkan metode statistik, pembelajaran mesin, dan manajemen basis data untuk mengekstrak informasi dan pola yang berarti dari sejumlah besar data (Alasali, 2024). Proses penelitian diawali dengan pengumpulan data dari sumber yang relevan, dilanjutkan dengan langkah pemrosesan yang mengorganisasikan dan menyiapkan data agar layak untuk dianalisis. Kemudian menerapkan algoritma tertentu, seperti K-means clustering, untuk menemukan pola dan kelompok dalam data. Analisis ini dilakukan dengan menggunakan perangkat lunak seperti machine learning (Xiaoling, 2023). Hasil yang diperoleh dievaluasi berdasarkan metrik tertentu untuk memastikan keakuratan dan validitas analisis dalam memenuhi tujuan penelitian.

B. Clustering

Pada tahap ini, clustering adalah sebuah metode analisis data yang berkonsentrasi pada pengelompokan data yang memiliki karakteristik yang sama. Metode ini berguna

untuk mengungkap pola tersembunyi dalam data tanpa perlu adanya label atau kategori yang sudah ditentukan sebelumnya (Maria Ulfah, 2022). Oleh karena itu, teknik ini termasuk dalam metode pembelajaran tanpa pengawasan. Berbeda dengan klasifikasi yang menetapkan pengelompokan berdasarkan data yang sudah dilabeli, clustering menciptakan kategori baru berdasarkan pola yang secara otomatis diidentifikasi dari data.

C. Algoritma K-Means

Algoritma K-Means adalah teknik yang sangat efektif untuk menyederhanakan informasi kompleks menjadi kategori yang berarti. Metode ini mendukung proses pengambilan keputusan berdasarkan data dengan cara mengelompokkan objek yang memiliki karakteristik serupa. Salah satu manfaat dari algoritma ini adalah kemampuannya untuk dengan cepat menangani data dalam jumlah besar dan memiliki dasar matematis yang sederhana, membuatnya ideal untuk berbagai penggunaan dalam praktik, seperti segmentasi pasar, analisis perilaku, dan pengelompokan wilayah geografis. Namun, keberhasilannya sangat bergantung pada beberapa faktor, seperti jumlah kluster yang ditentukan (k) dan sensitivitas terhadap penempatan awal pusat kluster. Karena fleksibilitasnya, K-means menjadi salah satu algoritma yang paling sering dipakai dalam eksplorasi data untuk mengidentifikasi pola tersembunyi dan menghasilkan analisis yang cepat diterapkan di berbagai bidang seperti bisnis, penelitian, dan teknologi. Berikut adalah diagram yang menunjukkan rumus jarak Euclidean Distance.

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2} \quad (1)$$

Keterangan:
d : jarak determinan (jarak euclidean) x : pusat cluster
y : informasi data n : total data
i : indeks data ke-

HASIL DAN PEMBAHASAN

A. Data Collection

Kumpulan data yang digunakan dalam studi ini diperoleh dari UCI Machine Learning Repository, yaitu dataset Buddymove HolidayIQ, yang berisi informasi tentang ulasan wisatawan berdasarkan aktivitas dan minat perjalanan mereka. Dataset ini mencakup data ulasan yang telah diunggah oleh pengguna dan disajikan dalam bentuk tabel dengan atribut yang terstruktur dan terpisah secara jelas, seperti User ID, Sports, Religious, Nature, Theatre, Shopping, dan Picnic. Setiap atribut menggambarkan frekuensi atau tingkat ketertarikan pengguna terhadap kategori aktivitas tertentu, sehingga memungkinkan analisis lebih mendalam terkait pola preferensi wisatawan.

Struktur data yang terorganisir ini menjadi landasan penting dalam proses analisis dan pengelompokan menggunakan algoritma clustering. Berikut adalah beberapa informasi yang akan diperlihatkan pada tabel data awal untuk memberikan gambaran umum mengenai isi dataset tersebut.

Gambar 1. Dataset Buddymove Holidayiq

Row No.	id	cluster	Sports	Religious	Nature	Theatre	Shopping	Picnic
1	1	cluster_3	2	77	79	69	68	95
2	2	cluster_3	2	62	76	76	69	68
3	3	cluster_3	2	50	97	87	50	75
4	4	cluster_3	2	68	77	95	76	61
5	5	cluster_1	2	98	54	59	95	86
6	6	cluster_3	3	52	109	93	52	76
7	7	cluster_3	3	64	85	82	73	69
8	8	cluster_3	3	54	107	92	54	76
9	9	cluster_3	3	64	108	64	54	93
10	10	cluster_1	3	86	76	74	74	103
11	11	cluster_1	3	107	54	64	103	94
12	12	cluster_1	3	103	60	63	102	93
13	13	cluster_3	3	64	82	82	75	69
14	14	cluster_1	3	93	54	74	103	69
15	15	cluster_3	3	63	82	81	78	69
16	16	cluster_3	3	82	79	75	75	82
17	17	cluster_3	5	59	131	103	54	86
18	18	cluster_3	5	56	124	108	56	85
19	19	cluster_3	4	85	67	111	65	72
20	20	cluster_1	5	114	83	65	114	102

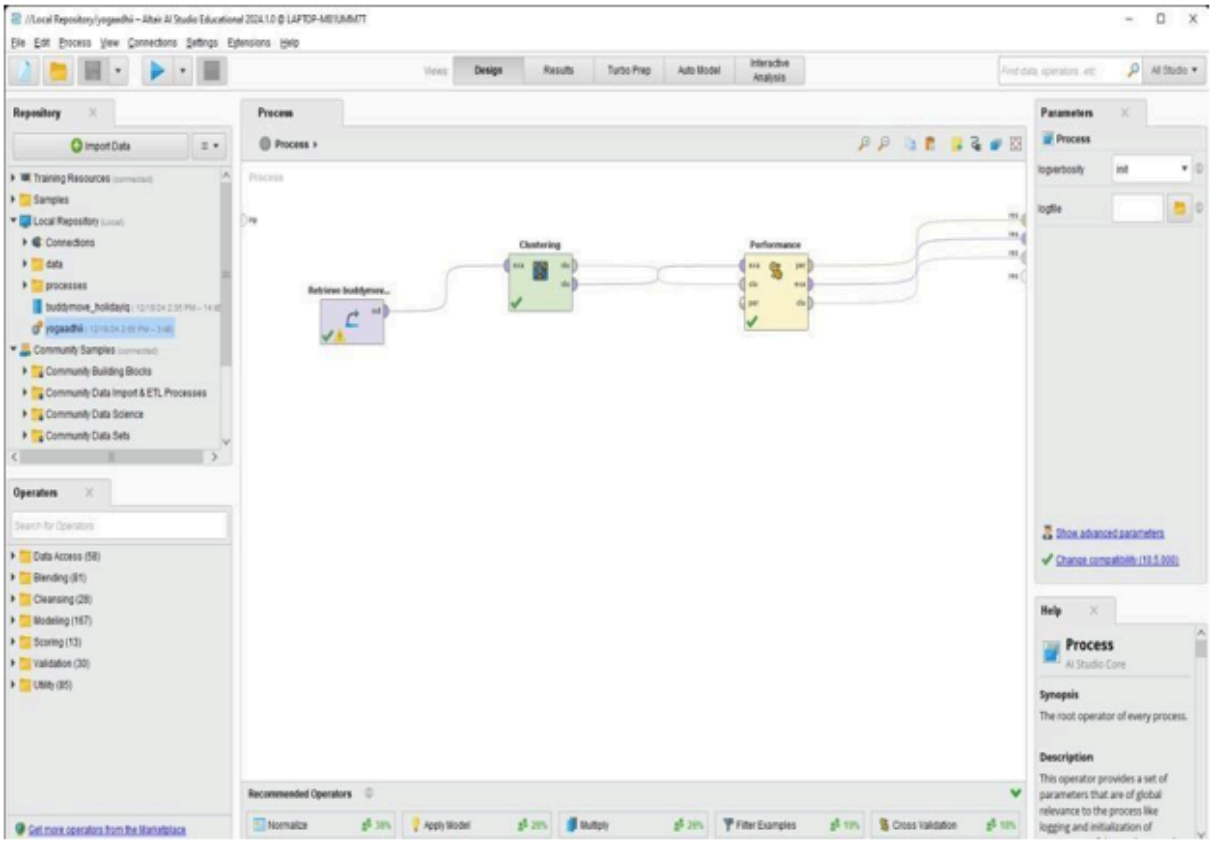
B. Procesing Data

Serangkaian langkah yang dilakukan untuk mengonversi data awal menjadi informasi yang bermanfaat merupakan proses penting dalam analisis data. Proses ini mencakup beberapa tahap utama, dimulai dari pengumpulan data dari sumber yang relevan, dilanjutkan dengan proses pembersihan, seleksi, dan transformasi data untuk memastikan bahwa data berada dalam kondisi yang layak untuk dianalisis. Setelah data terstruktur dan bebas dari nilai yang tidak konsisten atau tidak lengkap, tahapan berikutnya adalah melakukan analisis menggunakan algoritma yang sesuai dengan tujuan penelitian. Setiap tahap dalam alur pengolahan data saling terkait dan berperan penting dalam memastikan bahwa data yang digunakan berkualitas, serta menghasilkan keluaran analisis yang akurat, valid, dan dapat diinterpretasikan dengan baik.

Dalam penelitian ini, proses pemodelan dilakukan menggunakan perangkat lunak RapidMiner, yang menyajikan alur kerja analisis data secara visual melalui struktur process workflow. Pendekatan ini memudahkan peneliti dalam memantau setiap tahapan yang dilakukan, mulai dari input data, pengolahan, hingga evaluasi hasil. Selain

itu, visualisasi pemodelan memungkinkan proses pengujian dan penyesuaian model dilakukan secara sistematis untuk memperoleh performa algoritma terbaik. Dengan demikian, RapidMiner tidak hanya berfungsi sebagai alat pemrosesan data, tetapi juga sebagai media pembelajaran yang memperjelas hubungan antartahap dalam proses analisis. Pemodelan tersebut dapat dilihat pada gambar berikut.

Gambar 2. Procesing Data



C. Cluster Model

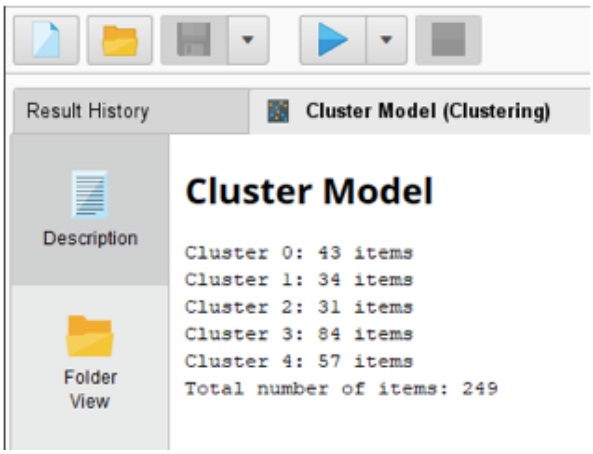
Pada bagian Cluster Model, kita dapat melihat tabel distribusi jumlah anggota pada setiap cluster yang menunjukkan karakteristik unik dari tiap kelompok hasil pengelompokan. Informasi ini mencerminkan bagaimana data terbagi berdasarkan kesamaan atribut, sehingga masing-masing cluster memiliki profil khusus yang membedakannya dari cluster lainnya. Menganalisis pola distribusi ini dapat memberikan wawasan yang lebih mendalam mengenai perbedaan serta persamaan antar cluster pada berbagai konteks analisis, seperti preferensi wisata, pola perilaku pengguna, atau kecenderungan keputusan tertentu.

Pemahaman yang lebih detail mengenai dinamika dalam setiap cluster menjadi penting untuk meningkatkan performa model sekaligus memberikan gambaran yang lebih komprehensif mengenai struktur data secara keseluruhan. Dengan meninjau komposisi anggota di setiap cluster, peneliti dapat menilai apakah hasil pengelompokan

sudah menggambarkan segmentasi yang realistis, atau justru masih memerlukan penyesuaian metode.

Selain itu, visualisasi jumlah anggota pada setiap cluster membantu mempermudah interpretasi hasil pengelompokan karena penyajian data secara grafis dapat memperjelas pola distribusi yang mungkin tidak terlihat hanya melalui angka dalam tabel. Visualisasi tersebut juga memungkinkan peneliti untuk mengidentifikasi potensi masalah, seperti ketidakseimbangan jumlah anggota pada antar cluster, dominasi cluster tertentu yang terlalu besar, atau cluster dengan jumlah sangat sedikit yang mungkin mengindikasikan outlier. Melalui evaluasi ini, peneliti dapat melakukan perbaikan model, baik melalui penyesuaian parameter algoritma maupun menentukan jumlah cluster yang lebih optimal agar sesuai dengan pola sebenarnya yang terdapat dalam data.

Gambar 3. Procesing Data



Pada ilustrasi di atas, kita dapat menyimpulkan bahwa Cluster 0 terdiri dari 43 anggota, Cluster 1 terdapat 34 anggota, Cluster 2 memiliki 31 anggota, Cluster 3 mengumpulkan 84 anggota, dan Cluster 4 memiliki 57 anggota. Dengan demikian, jumlah total dalam Cluster Model mencapai 249 anggota, atau keseluruhan Dataset pada Buddymove Holidayiq.

D. Centroid Table

Dalam Centroid Table ini, dapat dipastikan bahwa informasi yang tersedia berasal dari hasil pengolahan model clustering menggunakan RapidMiner. Informasi yang disajikan pada tabel centroid menunjukkan nilai center atau mean dari setiap cluster berdasarkan atribut-atribut yang digunakan dalam proses pengelompokan. Nilai rata-rata tersebut memberikan gambaran karakteristik utama atau profil umum dari setiap cluster, sehingga memudahkan untuk mengidentifikasi perbedaan signifikan antar kelompok. Tabel ini merangkum kecenderungan atribut yang dimiliki anggota masing-masing cluster, dan menjadi dasar analisis untuk memahami pola yang terbentuk.

Dengan demikian, centroid table berperan penting dalam proses interpretasi hasil clustering, seperti yang dapat dilihat pada bagian berikut:

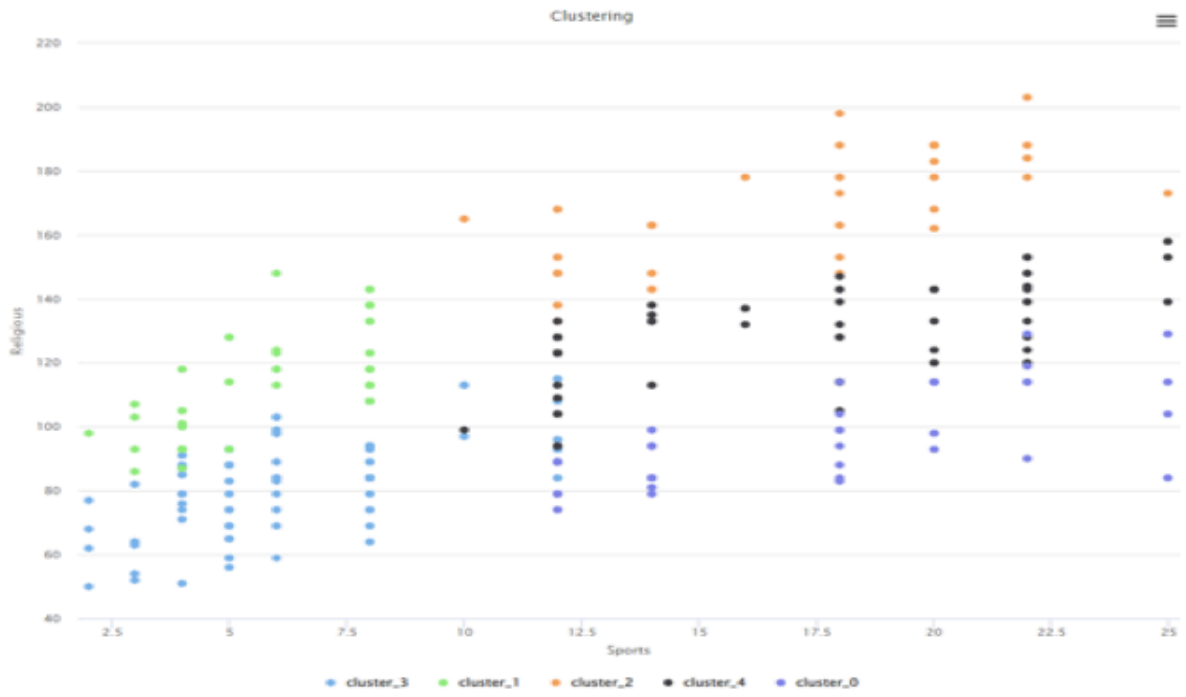
Gambar 4. Centroid Table

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
Sports	17.674	5.676	17.129	6.167	17.246
Religious	96.860	112.265	168.387	81.298	128.140
Nature	195.721	74.441	85.645	110.821	142
Theatre	132.581	80.882	112.871	104.750	144.368
Shopping	90.302	111.882	191.290	79.798	135.561
Picnic	147.442	105.441	141.839	91.750	139.491

Setiap baris dalam tabel mewakili pusat cluster, dan kolom mewakili atribut kumpulan data. Nilai-nilai ini memberikan gambaran umum tentang karakteristik cluster, termasuk atribut utama dan nilai rata-rata unik dalam setiap cluster table.

E. Visualization

Gambar 5. Visualizations



Visualisasi RapidMiner pada tampilan Scatter/Bubble berfungsi sebagai representasi grafis yang menampilkan hasil analisis data secara lebih intuitif dan mudah dipahami. Melalui visualisasi ini, pola distribusi data, struktur pembentukan cluster, serta hubungan antar atribut dapat diamati dengan jelas dalam bentuk titik atau gelembung yang diposisikan berdasarkan karakteristik nilai setiap variabel. Warna, ukuran, dan posisi setiap titik atau gelembung menggambarkan perbedaan antar

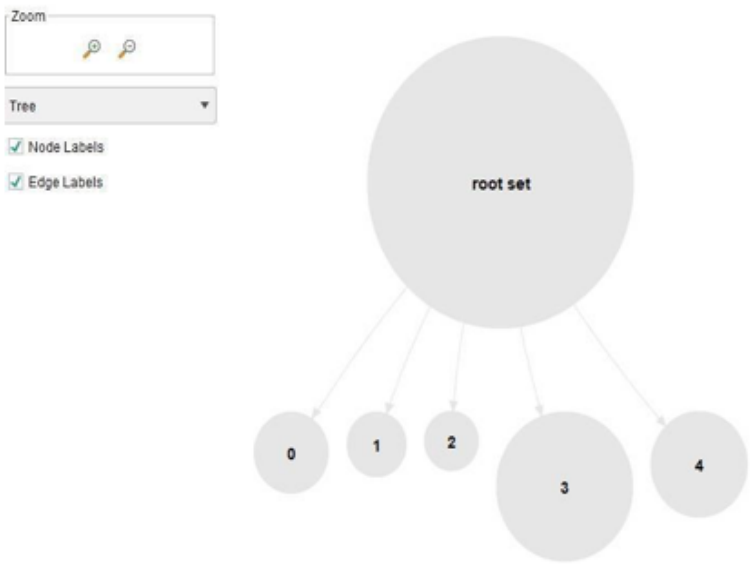
kelompok data, sehingga memudahkan peneliti dalam mengidentifikasi kecenderungan, anomali, maupun kedekatan antar objek dalam suatu cluster.

Selain itu, visualisasi Scatter/Bubble juga membantu mengevaluasi efektivitas algoritma pengelompokan yang digunakan, misalnya apakah cluster terbentuk dengan pemisahan yang tegas atau masih saling tumpang tindih. Dengan demikian, tampilan grafis ini tidak hanya memperjelas interpretasi hasil analisis, tetapi juga menjadi alat penting dalam pengambilan keputusan untuk penyesuaian parameter model, pemilihan algoritma yang lebih optimal, atau peningkatan kualitas preprocessing data. Visualisasi yang informatif ini pada akhirnya memberikan gambaran komprehensif mengenai struktur data dan mendukung penyampaian hasil penelitian secara lebih komunikatif dan profesional.

F. Graph

Pada grafik di bawah ini ditampilkan model yang dihasilkan dari data yang telah diproses, yang kemudian membentuk baris-baris (rows) pengetahuan sebagai representasi dari pola dan informasi yang berhasil diekstraksi. Setiap row menggambarkan hasil transformasi data mentah menjadi pengetahuan yang lebih terstruktur, sehingga memungkinkan peneliti untuk memahami hubungan antar variabel, kecenderungan tertentu, serta karakteristik utama yang muncul dari proses analisis. Melalui visualisasi tersebut, wawasan baru dapat diperoleh secara lebih jelas dan sistematis, karena model tidak hanya menampilkan data dalam bentuk angka, tetapi mengubahnya menjadi informasi yang dapat diinterpretasikan dan digunakan sebagai dasar dalam pengambilan keputusan analitis atau strategis. Dengan demikian, grafik ini menjadi elemen penting dalam menjembatani data mentah menuju pengetahuan yang memiliki nilai dan manfaat nyata.

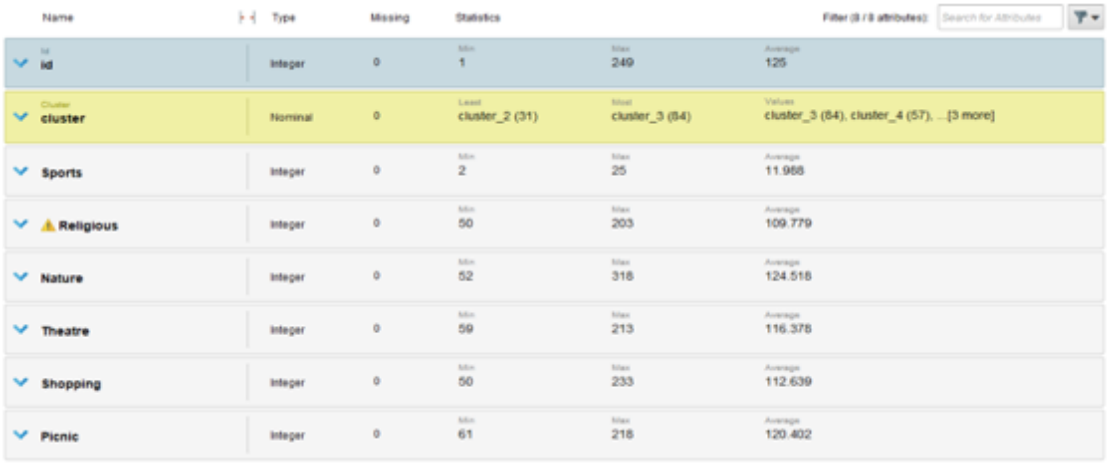
Gambar 6. Graph



G. Statistics

Pengguna dapat melihat penyebaran atribut data pada setiap Cluster dan memiliki opsi untuk memilih atribut mana yang ingin ditampilkan atau dianalisis lebih lanjut.

Gambar 7. Statistic Table



Name	Type	Missing	Statistics		
id	Integer	0	Min 1	Max 249	Average 125
cluster	Nominal	0	Least cluster_2 (31)	Most cluster_3 (64)	Values cluster_3 (64), cluster_4 (57), ... [3 more]
Sports	Integer	0	Min 2	Max 25	Average 11.988
Religious	Integer	0	Min 50	Max 203	Average 109.779
Nature	Integer	0	Min 52	Max 318	Average 124.518
Theatre	Integer	0	Min 59	Max 213	Average 116.378
Shopping	Integer	0	Min 50	Max 233	Average 112.639
Picnic	Integer	0	Min 61	Max 218	Average 120.402

H. Davies Bouldin

Pengelompokan yang menerapkan algoritma k means dalam penelitian ini memberikan hasil yang memuaskan, berdasarkan evaluasi yang dilakukan menggunakan indeks Davies Bouldin (Yolandari, 2025) dan nilai berdasarkan pada gambar berikut.

Gambar 8. Hasil Davies Bouldin

KESIMPULAN

Dalam kesimpulan ini, kita dapat memahami mengenai dataset dari Buddymove HolidayIQ yang memanfaatkan Algoritma K-Means untuk mengelompokkan ulasan wisata dari platform HolidayIQ. Dataset ini mencakup informasi mengenai keinginan para pelancong yang terbagi dalam enam kategori, yaitu enam variabel utama (SPORT, RELIGIOUS, NATURE, THEATER, SHOPPING, dan PICNIC) yang diambil dari Repositori UCI untuk analisis lebih lanjut. Dengan penerapan algoritma K-Means, proses pengelompokan data dilakukan menjadi enam kluster selama fase pemrosesan. Hasil

dari evaluasi pengelompokan berdasarkan indeks Davies-Bouldin memberikan wawasan yang lebih mendalam mengenai kontribusi penelitian ini untuk pengembangan sistem pembelajaran yang adaptif. Ini dapat memfasilitasi inovasi dalam pendidikan yang didasarkan pada teknologi.

DAFTAR PUSTAKA

Alasali, T., Ortakci, Y. (2024). Clustering Techniques in Data Mining: A Survey of Methods, Challenges, and Applications. Journal of Computer Science 9(1), 32-50. 10.53070/bbd.1421527

Al-Fahmi, B. M. (2023). "Penerapan Metode Data Mining dengan Algoritma K-Means untuk Klasterisasi Destinasi Wisata." Teknosi: Jurnal Teknologi dan Sistem Informasi. 9(2), 141-149. https://doi.org/10.25077/TEKNOSI.v9i2.2023.141-149

Astiti, S., Harman, R., & Darmansah, D. (2024). Pengelompokan Destinasi Wisata di Batam Berdasarkan Daya Tarik dan Fasilitas Menggunakan Metode K-Means Clustering. Kesatria: Jurnal Penerapan Sistem Informasi (Komputer dan Manajemen), 5(4), 2005-2012.

Atmadja, B. R. (2022). Analisis Sentimen Bahasa Indonesia Pada Tempat Wisata Di Kabupaten Sukabumi Dengan Naive Bayes Classifier. Elkom: Jurnal Elektronika dan Komputer, 15(2), 371-382.

Habiballoh, H., Faqih, A., & Suprapti, T. (2024). Implementasi Algoritma K-Means Dalam Mengelompokan Kabupaten/Kota Di Jawa Barat Berdasarkan Jenis Dan Jumlah Potensi Objek Daya Tarik Wisata. Jurnal Informatika dan Teknik Elektro Terapan, 12(2).

Hailong Chen, Yi Liu and Kaiqi Chen. (2021). Big Data in Tourism: General Issues and Challenges. Journal of Tourism & Hospitality, Vol. 10, 1-6.

Lusianah, N., Purnamasari, A. I., & Nurhakim, B. (2023). Implementasi Algoritma K-Means Dalam Pengelompokan Jumlah Wisatawan Akomodasi Di Jawa Barat. Jurnal Ekonomi, Bisnis dan Manajemen, 2(1), 254-268.

Maria Ulfah. (2022). Penerapan Data Mining Clustering Menggunakan Metode K-Means Dalam Pengelompokan Buku Perpustakaan Politeknik Negeri Balikpapan. Jurnal Fidelity : Jurnal Teknik Elektro, Vol. 4, No. 3, 62-68

Ritonga, A. S., & Muhandhis, I. (2021). Teknik Data Mining Untuk Mengklasifikasikan Data Ulasan Destinasi Wisata Menggunakan Reduksi Data Principal Component Analysis (Pca). Jurnal Ilmiah Edutic: Pendidikan dan Informatika, 7(2), 124-133.

Xiaoling Shu, Yiwan Ye. (2023). Knowledge Discovery: Methods from data mining and machine learning. Journal Social Science Research. Vol. 110, https://doi.org/10.1016/j.ssresearch.2022.102817

Yolandari, NA., et. al. (2025). Analisis Perbandingan K-Means Dan Dbscan Dalam Pengelompokan Data Travel Review Ratings Menggunakan Evaluasi Silhouette Index Dan Davies-Bouldin Index. Jurnal Informatika dan Teknik Elektro Terapan. 13(3), 78-84. https://doi.org/10.23960/jitet.v13i3.6884