EJECTS: Journal Computer, Technology and Informations System
Vol.4 No.2 Maret 2025, P-ISSN: 2988-5477 E-ISSN: 2988-4446
LPPM Universitas Darwan Ali

Segmentasi Pengguna Media Sosial dengan K-Means Clustering

Muhammad Risfan Nurdin

Program Studi : Sistem Informasi, Universitas Darwan Ali
Email : mrisfannurdin@gmail.com

ABSTRACT— This study aims to identify groups of social media users based on demographic and behavioral attributes such as age and daily usage duration. The segmentation process was carried out using the K-Means Clustering algorithm, executed within the RapidMiner platform. The dataset was sourced from Kaggle and underwent preprocessing stages to ensure the quality and suitability of the analysis. Using an unsupervised learning approach, the study successfully formed three user segments: active, moderate, and passive. Each group has distinct characteristics that can serve as a reference for developing personalized content strategies and enhancing user engagement more effectively.

Keywords— K-Means Clustering, Social Media, User Segmentation, RapidMiner, Unsupervised Learning.

ABSTRAK— Penelitian ini bertujuan untuk mengidentifikasi kelompok pengguna media sosial berdasarkan atribut demografis dan perilaku, seperti usia dan durasi penggunaan harian. Proses segmentasi dilakukan menggunakan algoritma K-Means Clustering yang dijalankan dalam platform RapidMiner. Data yang digunakan berasal dari Kaggle dan telah melalui tahapan praproses untuk memastikan kualitas dan kesesuaian analisis. Dengan pendekatan unsupervised learning, penelitian ini berhasil membentuk tiga segmen pengguna, yaitu aktif, sedang, dan pasif. Masing-masing kelompok memiliki ciri khas tersendiri yang dapat dijadikan acuan dalam pengembangan strategi personalisasi konten dan peningkatan keterlibatan pengguna secara lebih tepat sasaran.

Kata kunci— K-Means Clustering, Media Sosial, Segmentasi Pengguna, RapidMiner, Unsupervised Learning.

I. PENDAHULUAN

Media sosial telah menjadi bagian integral dari kehidupan masyarakat modern, mencakup berbagai aspek mulai dari komunikasi pribadi, hiburan, hingga strategi pemasaran digital. Platform seperti Instagram, Twitter, dan TikTok menjadi pusat interaksi sosial yang berlangsung hampir setiap saat. Pertumbuhan pengguna yang sangat pesat mendorong perlunya analisis mendalam terhadap perilaku dan pola penggunaan yang terbentuk di dalamnya.

Dalam menghadapi dinamika ini, segmentasi pengguna menjadi pendekatan strategis yang banyak digunakan untuk memahami karakteristik kelompok pengguna. Segmentasi berbasis data mining memungkinkan proses identifikasi kelompok pengguna yang homogen berdasarkan variabel-variabel perilaku atau demografi tanpa memerlukan label awal. Oleh karena itu, penelitian ini diarahkan untuk menerapkan segmentasi pengguna media sosial berdasarkan dua atribut utama: usia dan durasi penggunaan harian.

Berdasarkan data dari platform Kaggle, diperoleh dataset yang merekam informasi terkait pengguna media sosial, termasuk usia dan rata-rata waktu penggunaan harian. Dataset ini dipilih karena kesederhanaannya namun tetap relevan dalam merepresentasikan perilaku digital secara umum. Ukuran data yang cukup besar dan bebas label menjadikan pendekatan unsupervised learning, seperti K-Means Clustering, sangat sesuai untuk diterapkan dalam penelitian ini.

Permasalahan yang diangkat dalam penelitian ini adalah bagaimana mengelompokkan pengguna media sosial ke dalam klaster perilaku yang bermakna hanya dengan menggunakan dua atribut. Sering kali, analisis perilaku digital membutuhkan banyak variabel, namun pendekatan minimalis berbasis dua atribut utama diharapkan tetap mampu memberikan segmentasi yang informatif. Tantangannya adalah bagaimana memastikan kualitas klaster tetap tinggi meskipun input variabel dibatasi.

Beberapa penelitian sebelumnya menunjukkan bahwa K-Means Clustering cukup efektif dalam mengidentifikasi segmen pengguna digital [1]. Metode ini telah diterapkan untuk mengelompokkan pelanggan e-commerce berdasarkan pola pembelian [1]. Studi lainnya berhasil mengelompokkan akun Instagram berdasarkan engagement rate [2], serta menyoroti pentingnya atribut usia dan waktu penggunaan sebagai penentu perilaku digital [3].

Penelitian-penelitian terdahulu banyak memanfaatkan atribut-atribut kompleks atau kombinasi data perilaku dan demografi secara luas. Namun, belum banyak yang secara spesifik menguji kekuatan dua atribut saja, yakni usia dan waktu penggunaan, dalam membentuk struktur segmentasi yang representatif. Penelitian ini menawarkan pendekatan yang lebih sederhana namun fokus, serta menggunakan algoritma K-Means sebagai solusi teknis untuk memecahkan tantangan segmentasi minimalis tersebut.

Tema ini diangkat karena adanya kebutuhan yang terus meningkat terhadap pemahaman mendalam atas

perilaku pengguna, terutama di era digital yang sangat kompetitif. Organisasi dan pengembang aplikasi membutuhkan strategi personalisasi yang lebih akurat namun tidak memerlukan banyak data sensitif. Oleh sebab itu, dengan menggunakan RapidMiner sebagai alat bantu analisis, penelitian ini diharapkan dapat memberikan kontribusi terhadap pengembangan strategi digital berbasis segmentasi perilaku aktual pengguna [4][5].

II. METODOLOGI PENELITIAN

Penelitian ini dilakukan menggunakan pendekatan kuantitatif eksploratif dengan metode K-Means Clustering. Pendekatan ini dipilih karena mampu mengelompokkan data yang tidak memiliki label berdasarkan kemiripan nilai antar data (distance-based clustering). Seluruh proses analisis dilakukan menggunakan platform RapidMiner, yang memungkinkan pengguna menjalankan alur analisis data secara visual dan interaktif, tanpa memerlukan banyak kode pemrograman[9].

1. Jenis Penelitian

Jenis penelitian ini termasuk dalam kategori data mining dengan pendekatan unsupervised learning, di mana algoritma pembelajaran mesin digunakan untuk mengelompokkan data tanpa label atau kategori yang telah ditentukan sebelumnya. Pendekatan unsupervised learning sangat sesuai untuk menggali pola tersembunyi dalam data numerik, terutama ketika belum tersedia struktur klasifikasi yang eksplisit[2]. Metode ini sangat cocok digunakan dalam eksplorasi awal terhadap perilaku pengguna media sosial karena dapat mengelompokkan data berdasarkan kemiripan karakteristiknya.

2. Pengumpulan dan Pemilihan Data

Data diperoleh dari situs berbagi dataset Kaggle, dalam format file CSV, dan berisi data pengguna media sosial. Atribut yang digunakan dalam penelitian ini adalah age (usia) dan time_spent (durasi penggunaan media sosial per hari). Kedua atribut ini dipilih karena dianggap dapat mewakili karakteristik demografis dan perilaku pengguna secara langsung. Atribut berbasis perilaku dan demografis memiliki relevansi tinggi dalam proses segmentasi karena mampu menunjukkan kecenderungan keterlibatan pengguna terhadap suatu platform secara lebih akurat[5].

3. Proses Praproses dan Normalisasi

Langkah pertama pada tahap praproses adalah mengevaluasi kelengkapan dan validitas data. Setelah memastikan bahwa data bersih dan konsisten, proses dilanjutkan dengan seleksi atribut. Kemudian, dilakukan normalisasi data menggunakan metode Min-Max Scaling agar seluruh nilai berada pada skala 0 hingga 1. Normalisasi sangat penting untuk menghindari dominasi atribut tertentu dalam perhitungan jarak Euclidean, yang menjadi dasar pengelompokan dalam algoritma K-Means[9].

4. Penjelasan Algoritma K-Means

K-Means Clustering merupakan salah satu metode partisi yang populer dalam dunia data mining. Algoritma ini bekerja dengan membagi data ke dalam sejumlah K klaster berdasarkan kedekatan jarak dengan pusat klaster (centroid). Proses diawali dengan pemilihan centroid awal secara acak, kemudian setiap data akan diklasifikasikan ke dalam klaster berdasarkan jarak terdekat. Setelah semua data dikelompokkan, posisi centroid akan dihitung ulang berdasarkan rata-rata titik-titik data dalam klaster tersebut. Langkah ini diulang hingga tidak terjadi perubahan posisi centroid secara signifikan, atau telah mencapai iterasi maksimum [4].

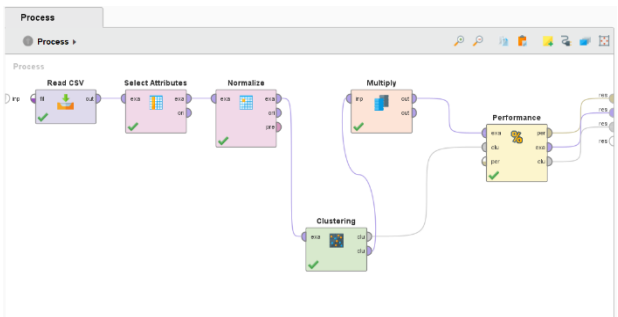
Kelebihan utama dari algoritma ini antara lain efisiensi dalam menangani dataset berukuran besar, kemudahan interpretasi hasil, serta waktu pemrosesan yang relatif cepat. Selain itu, K-Means juga cocok digunakan ketika jumlah klaster yang diinginkan telah diketahui sebelumnya. Akan tetapi, kelemahan metode ini mencakup ketergantungan pada nilai inisialisasi centroid awal dan keharusan menentukan jumlah klaster secara manual [6].

Dalam penelitian ini, penulis menentukan jumlah klaster sebanyak tiga (K=3), yang merujuk pada studi [7] serta konteks segmentasi pengguna aktif, sedang, dan pasif. Meskipun metode Elbow dapat digunakan untuk memvalidasi jumlah K yang optimal, pada penelitian ini K=3 telah dianggap representatif dalam konteks perilaku pengguna media sosial. Untuk mengevaluasi hasil segmentasi tersebut, dilakukan perhitungan metrik performa klaster menggunakan operator Performance, sehingga analisis tidak hanya berhenti pada visualisasi klaster, tetapi juga didukung dengan indikator evaluatif berbasis statistik.

5. Implementasi di RapidMiner

RapidMiner digunakan sebagai alat bantu utama dalam seluruh proses pengolahan dan analisis data. Platform ini menyediakan berbagai operator yang dapat disusun secara modular untuk membentuk alur kerja (workflow) analisis. Tahapan implementasi meliputi:

- Import dataset dengan operator Read CSV
- Seleksi kolom age dan time_spent menggunakan Select Attributes
- Normalisasi data numerik menggunakan operator Normalize
- Penerapan algoritma K-Means dengan jumlah klaster K=3
- Penggandaan output hasil klasterisasi menggunakan operator Multiply
- Evaluasi performa hasil segmentasi menggunakan operator Performance



Gambar 1. Workflow Proses Segmentasi di RapidMiner

Workflow pada Gambar 1 memperlihatkan bagaimana tahapan segmentasi pengguna media sosial

dilakukan secara modular. Penggunaan diagram ini sangat berguna untuk menunjukkan hubungan logis antar tahapan analisis, sekaligus mendemonstrasikan bagaimana setiap operator dihubungkan dalam alur kerja RapidMiner secara sistematis. Hasil dari operator K-Means kemudian ditampilkan dalam bentuk tabel centroid, scatter plot berwarna berdasarkan klaster, dan label klaster untuk masing-masing baris data. Setelah itu dimasukkan ke operator Multiply agar hasil dari proses clustering dapat digunakan secara paralel—baik untuk evaluasi maupun visualisasi. Sementara itu, operator Performance digunakan untuk menghitung metrik kualitas klaster seperti Average Within-Centroid Distance dan Davies-Bouldin Index, yang penting untuk menilai seberapa baik klaster yang terbentuk. Dengan struktur ini, proses analisis tidak hanya menghasilkan pembagian kelompok pengguna, tetapi juga disertai evaluasi kuantitatif yang memperkuat validitas hasil segmentasi.

6. Visualisasi dan Interpretasi

Hasil segmentasi divisualisasikan dalam bentuk scatter plot dua dimensi, yang menampilkan distribusi pengguna berdasarkan usia dan waktu penggunaan. Setiap titik mewakili satu pengguna, dengan warna berbeda untuk setiap klaster. Visualisasi ini sangat membantu dalam menilai pemisahan antar kelompok. Selain itu, nilai centroid pada setiap klaster digunakan untuk memahami karakteristik rata-rata dari masing-masing kelompok.

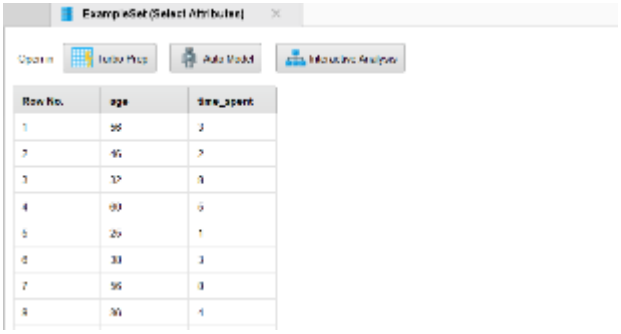
Interpretasi hasil melalui centroid sangat krusial untuk memahami kecenderungan umum dalam tiap segmen. Misalnya, klaster dengan centroid bernilai time_spent tinggi dan age rendah dapat diinterpretasikan sebagai pengguna aktif berusia muda, sedangkan klaster dengan nilai sebaliknya dapat digolongkan sebagai pengguna pasif[11]. Penafsiran seperti ini memberikan wawasan yang dapat dimanfaatkan dalam strategi pemasaran digital dan pengembangan konten. Struktur metodologi dalam penelitian ini telah disusun secara sistematis dan mengacu pada kaidah ilmiah yang menekankan pentingnya kejelasan tahapan dalam penelitian teknologi informasi agar hasilnya dapat dipertanggungjawabkan secara akademis[12].

III. HASIL DAN PEMBAHASAN

Tahapan analisis diawali dengan pembacaan dataset menggunakan operator Read CSV di RapidMiner. Operator ini mengimpor data mentah dari file CSV, yang menjadi dasar dalam proses selanjutnya. Dataset ini berisi informasi pengguna media sosial, namun penelitian hanya memfokuskan pada dua atribut yaitu age dan time_spent yang paling relevan dalam menggambarkan perilaku pengguna [5].

Gambar 2. Result Read CSV

Operator Read CSV digunakan untuk mengimpor dataset mentah dari file CSV, yang kemudian dihubungkan dengan operator Select Attributes. Operator ini berfungsi untuk memilih hanya dua kolom utama (age dan time_spent) yang akan dianalisis. Langkah ini penting untuk menyaring data yang relevan dan menyederhanakan proses selanjutnya. Pemilihan atribut yang sesuai dapat membantu menghindari kebisingan data dan menghasilkan klaster yang lebih bermakna, seperti yang dijelaskan dalam [2] dan diperkuat pula oleh studi lain dalam konteks media sosial. Selain itu, data sosial media bersifat dinamis dan temporal, sehingga pendekatan segmentasi perlu mempertimbangkan fleksibilitas terhadap perubahan perilaku pengguna dari waktu ke waktu [8].



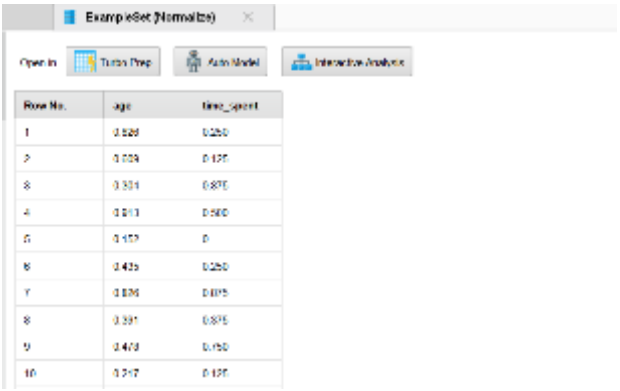
Gambar 3. Result Select Attributes

Setelah data berhasil dimuat, langkah berikutnya ditunjukkan pada Gambar 3, yaitu pemilihan atribut menggunakan operator Select Attributes. Operator ini memastikan bahwa hanya kolom yang relevan dalam hal ini *age* dan *time_spent* yang dilibatkan dalam analisis klasterisasi, sehingga mengurangi kompleksitas data. Setelah atribut ditentukan, dilakukan proses normalisasi menggunakan operator Normalize dengan metode Min-Max Scaling. Transformasi nilai ini dilakukan dengan menggunakan rumus normalisasi

$$x' = \frac{x - x\,min}{x\,max - x\,min}$$

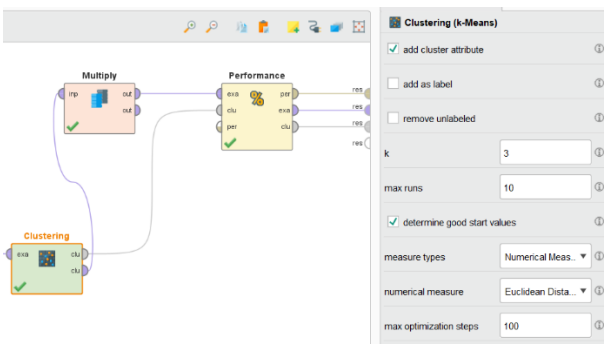
Dalam hal ini, di mana x adalah nilai asli suatu atribut, $x\,min$ adalah nilai terkecil, dan $x\,max$ adalah nilai terbesar dalam kolom tersebut. Rumus ini akan menghasilkan nilai baru x' yang berada dalam rentang 0

hingga 1, yang kemudian digunakan sebagai input dalam algoritma clustering. Misalnya, jika atribut *age* memiliki nilai minimum 15 dan maksimum 60, maka seseorang yang berusia 45 tahun akan memiliki nilai normalisasi sebesar $x' = \frac{45-15}{60-15} \approx 0{,}666$. Dengan skala yang telah diseragamkan ini, algoritma K-Means dapat menghitung jarak antar data dengan lebih adil dan akurat, tanpa adanya dominasi dari atribut yang memiliki skala numerik lebih besar. Proses normalisasi ini sangat penting untuk memastikan bahwa semua atribut memiliki kontribusi yang seimbang dalam proses pengelompokan [9].



Gambar 4. Result Normalize

Selanjutnya, data yang telah disaring kemudian dinormalisasi seperti diperlihatkan pada Gambar 4. Proses ini dilakukan dengan operator Normalize agar setiap nilai atribut berada dalam skala yang seragam, yaitu antara 0 hingga 1, guna memastikan kesetaraan bobot antar fitur dalam perhitungan jarak. Operator Normalize secara otomatis menghitung nilai minimum dan maksimum untuk setiap atribut dan melakukan transformasi terhadap seluruh nilai data. Dengan skala yang seimbang, algoritma clustering akan mempertimbangkan semua atribut secara adil saat menghitung jarak antar titik data. Langkah berikutnya adalah penerapan algoritma K-Means, yang memerlukan input berupa data yang telah dinormalisasi dan parameter jumlah klaster (k) yang ingin dibentuk. Dalam penelitian ini, nilai k ditetapkan sebanyak tiga (K=3), sesuai dengan tiga kategori pengguna yang dihipotesiskan: aktif, sedang, dan pasif. Proses ini menggunakan parameter default iterasi maksimum 100 dan Euclidean Distance sebagai metode pengukuran jarak, seperti yang umum digunakan dalam implementasi K-Means [4].



Gambar 5. Parameter Clustering (k-Means)

Setelah data dinormalisasi, Gambar 5 memperlihatkan skema proses sekaligus pengaturan parameter dari algoritma K-Means Clustering di RapidMiner. Di bagian kanan gambar tampak konfigurasi lengkap yang digunakan dalam penelitian ini, di mana beberapa parameter kunci ditentukan secara eksplisit untuk menjamin hasil segmentasi yang optimal. Parameter "add cluster attribute" dicentang (true) agar setiap baris data dalam output akhir memiliki atribut tambahan berupa label klaster yang diperoleh, yang sangat penting untuk keperluan analisis dan visualisasi hasil. Kemudian, nilai k ditetapkan sebesar 3, sesuai dengan hipotesis awal bahwa terdapat tiga kategori pengguna media sosial, yakni aktif, sedang, dan pasif. Penggunaan nilai k ini menjadi dasar dari jumlah klaster yang akan dibentuk oleh algoritma.

Selanjutnya, parameter "max runs" diset ke angka 10, yang berarti algoritma akan menjalankan proses pengelompokan sebanyak 10 kali dengan inisialisasi centroid yang berbeda-beda, lalu memilih hasil terbaik dari semua percobaan tersebut. Pengaturan ini berguna untuk menghindari jebakan minimum lokal, di mana hasil akhir clustering bisa sangat dipengaruhi oleh pemilihan centroid awal. Untuk mengatasi masalah ini, opsi "determine good start values" juga diaktifkan, yang berfungsi agar RapidMiner menggunakan metode inisialisasi centroid yang lebih strategis ketimbang sekadar acak murni. Sedangkan pada bagian "measure types" dipilih Numerical Measures, karena seluruh atribut yang digunakan yaitu age dan time_spent merupakan data numerik. Metode pengukuran jarak yang digunakan adalah Euclidean Distance, yang merupakan metode standar dalam algoritma K-Means karena menghitung jarak langsung antara titik data dan centroid dalam ruang vektor. Akhirnya, parameter "max optimization steps" ditetapkan sebanyak 100, yang artinya proses iterasi pembaruan centroid akan dilakukan hingga maksimal 100 langkah atau hingga konvergen lebih cepat. Seluruh konfigurasi ini dirancang untuk memastikan bahwa algoritma K-Means dapat bekerja optimal dalam membentuk segmentasi yang akurat dan stabil.

Centroid berfungsi sebagai representasi pusat dari suatu klaster, dan menjadi acuan utama dalam proses pengelompokan data. Dalam algoritma K-Means, penentuan klaster dilakukan dengan mengukur seberapa dekat suatu data ke masing-masing centroid yang telah ditentukan. Pengukuran ini menggunakan rumus Euclidean Distance, yaitu:

$$d(x, c) = \sqrt{(x_1 - c_1)^2 + (x_2 - c_2)^2}$$

Dalam konteks K-Means Clustering, rumus ini esensial untuk mengukur seberapa dekat suatu titik data dengan pusat klaster atau centroid. Variabel $x = (x_1, x_2)$ merepresentasikan sebuah vektor data spesifik, yang dalam contoh ini terdiri dari dua atribut seperti usia (age) dan durasi penggunaan (time_spent). Sementara itu, $c = (c_1, c_2)$ adalah koordinat dari centroid sebuah klaster, yang juga memiliki dua atribut yang sama. Proses perhitungan melibatkan pengurangan nilai atribut yang bersesuaian antara titik data dan centroid,

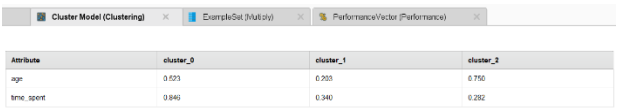
mengkuadratkan hasilnya, menjumlahkan kuadrat dari semua selisih atribut, dan terakhir mengambil akar kuadrat dari jumlah tersebut. Intinya adalah, semakin kecil nilai jarak $d(x, c)$ yang diperoleh dari perhitungan ini, semakin dekat posisi titik data tersebut dengan centroid klaster yang sedang dipertimbangkan. Logika ini fundamental dalam algoritma K-Means, karena menjadi dasar penentuan klaster: setiap data akan secara otomatis dikelompokkan ke dalam klaster yang memiliki centroid paling dekat dengannya. Sebagai contoh konkret, jika ada data pengguna dengan nilai usia 0.5 dan durasi penggunaan 0.6, serta sebuah centroid klaster berada pada koordinat (0.4, 0.7), maka jarak Euclidean antara kedua titik ini akan dihitung berdasarkan rumus tersebut untuk menentukan kedekatan mereka.

$$d = \sqrt{(0.5 - 0.4)^2 + (0.6 - 0.7)^2} = \sqrt{0.01 + 0.01} = \sqrt{0.02} \approx 0.141$$

Setelah proses pengelompokan selesai, posisi centroid akan diperbarui menggunakan rata-rata dari seluruh data yang berada di dalam klaster tersebut. Pembaruan ini dilakukan dengan rumus:

$$C_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$$

Dalam formula ini, C_j adalah centroid baru untuk klaster ke-j, di mana $|c_j|$ merujuk pada jumlah total data yang telah ditetapkan ke dalam klaster ke-j pada iterasi saat ini, dan $\sum_{x_i \in C_j} x_i$ melambangkan penjumlahan semua titik data individual (x_i) yang kini menjadi anggota dari klaster tersebut. Secara substansial, rumus ini menghitung posisi centroid baru sebagai nilai rata-rata dari semua data poin yang berada di dalam klaster yang bersangkutan. Proses penghitungan ulang centroid ini diulang secara iteratif, di mana setelah setiap pembaruan, titik-titik data akan kembali diukur jaraknya ke centroid baru dan ditetapkan ke klaster terdekatnya. Iterasi ini akan terus berlanjut hingga tercapai kondisi konvergensi, di mana posisi centroid tidak lagi berpindah secara signifikan, atau sampai jumlah iterasi maksimum yang telah ditentukan sebelumnya terpenuhi [7]. Stabilitas yang dicapai melalui proses iteratif ini sangat penting untuk memastikan bahwa hasil pembentukan klaster benar-benar stabil dan mampu merepresentasikan pola tersembunyi dalam data secara akurat.



Gambar 6. Centroid Table

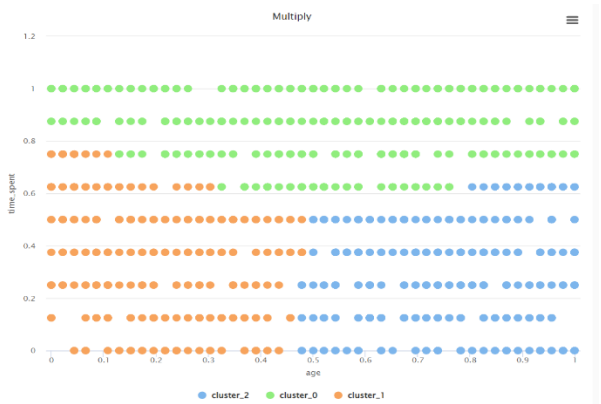
Setelah pengaturan parameter diselesaikan dan algoritma dijalankan, hasil perhitungan centroid ditampilkan pada Gambar 6. Setiap nilai pada tabel merupakan rata-rata dari atribut age dan time_spent dalam masing-masing klaster, yang mencerminkan karakteristik umum setiap kelompok pengguna. Gambar tabel di atas menunjukkan hasil perhitungan centroid untuk masing-masing klaster hasil algoritma K-Means. Setiap nilai merepresentasikan rata-rata dari atribut age dan time_spent pada masing-masing klaster. Karena data telah melalui proses normalisasi Min-Max, seluruh nilai berada dalam skala 0 hingga 1 sehingga interpretasi difokuskan pada kecenderungan relatif, bukan nilai absolut.

Klaster 0 memiliki nilai rata-rata usia sebesar 0.750 dan durasi penggunaan sebesar 0.282. Ini mengindikasikan bahwa kelompok ini terdiri dari pengguna berusia relatif lebih tua yang hanya menggunakan media sosial dalam durasi yang pendek. Berdasarkan karakteristik tersebut, klaster ini dapat diklasifikasikan sebagai segmen pengguna pasif yang keterlibatannya tergolong rendah dalam aktivitas media sosial.

Berbeda dari itu, klaster 1 menunjukkan nilai age sebesar 0.203 dan time_spent sebesar 0.340. Nilai usia yang rendah menggambarkan dominasi pengguna muda, sedangkan durasi penggunaan yang sedang menandakan mereka cukup aktif tetapi tidak intensif. Klaster ini menggambarkan segmen pengguna dengan tingkat keterlibatan menengah, atau pengguna sedang.

Sementara itu, klaster 2 memiliki nilai age sebesar 0.523 dan time_spent tertinggi yaitu 0.846. Ini menunjukkan bahwa kelompok ini terdiri dari pengguna berusia menengah namun sangat aktif dalam menggunakan media sosial. Dengan kata lain, klaster ini dapat diinterpretasikan sebagai kelompok pengguna aktif yang menunjukkan frekuensi penggunaan tinggi terlepas dari faktor usia.

Analisis terhadap nilai centroid ini memungkinkan pemahaman yang lebih dalam mengenai kecenderungan perilaku pengguna di setiap kelompok. Ketiga klaster memperlihatkan perbedaan signifikan dalam dimensi usia dan waktu penggunaan, yang keduanya berkontribusi langsung terhadap tingkat keterlibatan pengguna. Hasil segmentasi ini dapat dimanfaatkan lebih lanjut untuk keperluan strategi pemasaran digital atau pengembangan fitur yang disesuaikan dengan profil tiap kelompok pengguna.



Gambar 7. Visualisasi Scatter Plot

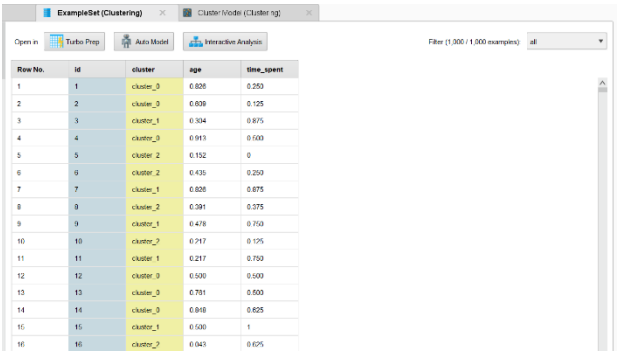
Untuk memberikan gambaran visual mengenai hasil klasterisasi, Gambar 7 menampilkan scatter plot dari pengguna media sosial berdasarkan dua atribut utama. Setiap titik mewakili satu individu, dan warna berbeda menunjukkan klaster hasil pengelompokan yang

memudahkan interpretasi secara visual. Scatter plot hasil clustering pada Gambar menunjukkan distribusi pengguna media sosial berdasarkan dua atribut utama, yaitu usia (age) pada sumbu X dan durasi penggunaan harian (time_spent) pada sumbu Y. Setiap titik dalam grafik mewakili satu individu dalam dataset, dengan warna yang berbeda menunjukkan klaster hasil pengelompokan oleh algoritma K-Means. Visualisasi ini memberikan gambaran intuitif mengenai struktur kelompok pengguna berdasarkan kedekatan perilaku, memudahkan dalam menilai keberhasilan proses segmentasi secara visual.

Dari pola persebaran titik, terlihat bahwa klaster dengan dominasi warna hijau (cluster_0) terletak di bagian atas grafik dengan rentang usia sedang hingga tinggi, namun dengan nilai time_spent tertinggi di antara klaster lainnya. Ini menunjukkan kelompok pengguna yang aktif di media sosial meskipun tidak termasuk kelompok usia paling muda. Sementara itu, klaster dengan warna biru (cluster_2) terkonsentrasi pada area dengan nilai usia tinggi dan durasi penggunaan rendah, mengindikasikan pengguna pasif yang lebih jarang berinteraksi di media sosial.

Sebaliknya, klaster berwarna oranye (cluster_1) mendominasi area bawah kiri grafik, yang menunjukkan pengguna berusia muda dengan durasi penggunaan media sosial yang tergolong sedang. Kombinasi antara usia dan aktivitas ini menjadikan cluster_1 sebagai representasi pengguna sedang yang tidak terlalu intensif namun tetap terlibat dalam penggunaan media sosial. Kecenderungan distribusi yang demikian memperlihatkan adanya konsistensi segmentasi berdasarkan dua atribut sederhana namun efektif dalam membedakan pola keterlibatan pengguna.

Visualisasi seperti ini penting untuk menganalisis struktur segmentasi tanpa harus merujuk langsung pada tabel numerik. Pola pemisahan yang cukup tegas antar klaster membuktikan bahwa atribut age dan time_spent mampu memberikan dasar yang kuat untuk membentuk segmentasi perilaku pengguna yang bermakna. Hal ini sekaligus menegaskan bahwa metode clustering berbasis jarak, seperti K-Means, mampu bekerja optimal bahkan dalam ruang fitur berdimensi rendah jika atributnya dipilih secara tepat.



Gambar 8. Result Clustering

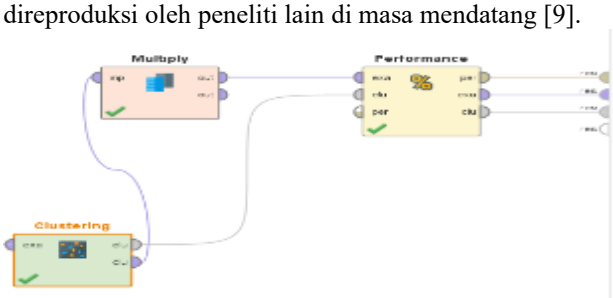
Gambar ini menunjukkan hasil akhir dari proses clustering menggunakan algoritma K-Means di RapidMiner. Setiap baris pada tabel mewakili satu individu dalam dataset yang telah secara otomatis dikelompokkan ke dalam klaster tertentu berdasarkan nilai atribut age dan time_spent yang telah dinormalisasi sebelumnya. Penambahan label klaster (cluster_0, cluster_1, cluster_2) memungkinkan pengguna untuk secara langsung mengetahui hasil segmentasi tanpa perlu analisis manual yang rumit.

Berbeda dengan tabel centroid yang menyajikan ringkasan nilai rata-rata tiap atribut per klaster, tabel ini menampilkan data mentah per individu, sehingga sangat berguna untuk analisis mendetail (granular). Melalui tampilan ini, peneliti dapat menelusuri siapa saja yang tergolong dalam suatu klaster dan melihat distribusi nilai mereka secara individual. Pendekatan ini memudahkan dalam identifikasi outlier, pengambilan sampel per klaster, serta perancangan strategi yang disesuaikan dengan profil pengguna dalam tiap kelompok.

Jumlah anggota masing-masing klaster juga dapat dihitung dengan mudah dari tabel ini, lalu divisualisasikan kembali ke dalam bentuk diagram batang atau pie chart untuk menunjukkan proporsi setiap segmen. Analisis ini mendukung gagasan bahwa atribut age dan time_spent dapat menjadi dasar segmentasi yang sederhana namun bermakna dalam mengelompokkan pengguna media sosial berdasarkan perilaku aktual mereka. Pendekatan berbasis data numerik seperti ini telah terbukti efektif dalam berbagai penelitian sebelumnya, terutama ketika tujuan segmentasi adalah untuk keperluan desain layanan yang adaptif [7].

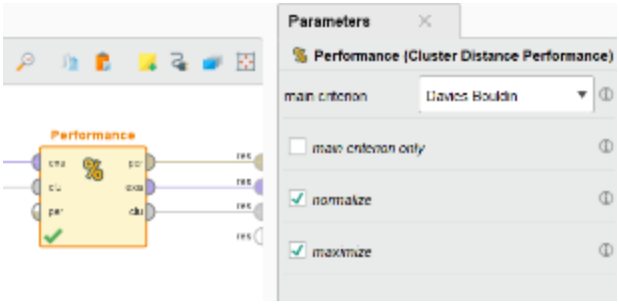
Model K-Means sendiri bekerja dengan prinsip pembaruan posisi centroid secara iteratif, di mana posisi baru ditentukan dari rata-rata data yang termasuk dalam masing-masing klaster hingga mencapai titik konvergensi. Tidak seperti metode supervised learning yang membutuhkan label awal, K-Means termasuk dalam kelompok unsupervised learning yang sangat cocok untuk data eksploratif seperti ini. Oleh karena itu, segmentasi yang dihasilkan tidak dipengaruhi oleh asumsi awal melainkan benar-benar terbentuk dari struktur data internal [2].

Keunggulan tambahan dari penggunaan RapidMiner dalam penelitian ini adalah tersedianya antarmuka visual interaktif yang memungkinkan implementasi algoritma machine learning tanpa perlu pengkodean manual. Hal ini sangat membantu dalam meminimalkan kesalahan prosedural dan mempercepat iterasi eksperimen serta validasi hasil. Kombinasi antara algoritma K-Means dan platform RapidMiner menjadikan proses segmentasi berjalan lebih efisien, transparan, serta mudah direproduksi oleh peneliti lain di masa mendatang [9].



Gambar 9. Operator Multiply dan Performance

Untuk mendukung proses evaluasi performa klasterisasi secara menyeluruh, digunakan dua operator tambahan dalam workflow RapidMiner, yaitu Multiply dan Cluster Distance Performance seperti yang terlihat pada gambar 9. Operator Multiply memiliki peran penting dalam konteks pemrosesan data yang bersifat paralel. Fungsinya adalah menggandakan (menyalin) hasil keluaran dari algoritma K-Means, yaitu data yang telah diberi label klaster. Salinan pertama dikirim langsung ke operator Performance untuk dilakukan pengukuran kualitas klaster, sementara salinan lainnya diteruskan sebagai output akhir untuk keperluan interpretasi visual dan lanjutan. Tanpa penggunaan Multiply, hasil dari K-Means hanya dapat diteruskan ke satu jalur saja, sehingga proses evaluasi dan visualisasi tidak dapat berjalan bersamaan dalam satu proses otomatis. Satu jalur dialirkan ke proses visualisasi seperti scatter plot dan centroid table, sementara jalur lainnya diteruskan ke proses evaluasi performa. Dengan demikian, pengguna dapat mengevaluasi hasil klasterisasi secara paralel tanpa kehilangan data pada cabang proses manapun.



Gambar 10. Parameter Performance

Konfigurasi evaluasi performa segmentasi yang ditampilkan pada Gambar 10 menunjukkan penggunaan operator Performance dengan tipe Cluster Distance Performance dalam platform RapidMiner. Dalam pengaturan ini, metrik utama yang digunakan untuk mengevaluasi hasil segmentasi adalah Davies-Bouldin Index (DBI), yang dipilih melalui menu main criterion. DBI merupakan salah satu metrik paling umum dan kredibel dalam mengevaluasi hasil klasterisasi, terutama untuk metode berbasis jarak seperti K-Means. DBI mengukur seberapa baik pemisahan antar klaster (separation) dan seberapa rapat penyebaran data di dalam masing-masing klaster (compactness). Nilai DBI yang semakin kecil mengindikasikan bahwa klaster yang terbentuk semakin optimal—artinya, anggota klaster saling berdekatan satu sama lain, sementara antar klaster saling berjauhan.

Yang menarik dalam konfigurasi ini adalah dua parameter tambahan yaitu normalize dan maximize yang dicentang. Parameter normalize digunakan untuk melakukan penskalaan otomatis terhadap nilai-nilai yang diukur, sehingga semua metrik evaluasi disetarakan dalam satuan atau skala yang sama. Ini sangat penting dalam konteks evaluasi performa berbasis jarak karena tanpa normalisasi, atribut yang memiliki skala besar dapat mendominasi hasil evaluasi dan menyebabkan interpretasi yang bias. Sementara itu, parameter maximize digunakan untuk memastikan bahwa nilai evaluasi akan dicari dalam bentuk terbaiknya, yaitu dengan mengoptimalkan metrik ke arah hasil yang dianggap paling ideal—dalam hal DBI, artinya nilai seminimal mungkin. Jika maximize tidak dicentang, beberapa algoritma mungkin akan mencari nilai yang besar sebagai target performa, padahal untuk DBI justru semakin kecil semakin baik.

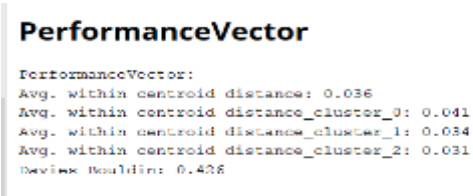
Salah satu metrik utama yang dihasilkan dari operator ini adalah Davies-Bouldin Index (DBI). DBI digunakan untuk mengukur kualitas pemisahan antar klaster (separation) dan kekompakan internal klaster (compactness). Semakin rendah nilai DBI, semakin baik segmentasi data yang terbentuk. Rumus Davies-Bouldin secara umum ditulis sebagai:

$$DB = \frac{1}{n}\sum_{i=1}^{n} \underset{j \neq i}{max} \left(\frac{S_i + S_j}{dij}\right)$$

Dalam rumus tersebut, S_i adalah rata-rata jarak dari setiap anggota dalam klaster ke-i ke pusat klasternya (centroid), sedangkan dij adalah jarak antara dua centroid dari klaster ke-i dan ke-j. Nilai maksimum diambil karena DBI mencari rasio terburuk antar klaster. Semakin kecil nilai $\frac{S_i+S_j}{dij}$, semakin baik pemisahan antar klaster tersebut. Karena DBI merupakan rata-rata dari nilai maksimum ini untuk setiap klaster, maka hasil akhirnya merepresentasikan kualitas keseluruhan dari segmentasi. Contoh sederhana: misalkan ada tiga klaster dengan nilai S_0 = 0.08, S_1 = 0.07, dan S_2 = 0.09. Jarak antar centroid misalnya d_{01} = 0.5, d_{02} = 0.6, dan d_{12} = 0.45. Maka misalnya untuk klaster 0, perhitungan rasio ke klaster lain bisa didapatkan sebagai:

$$R_{01} = \frac{0.08 + 0.07}{0.5} = 0.3, \quad R_{02} = \frac{0.08 + 0.09}{0.6} \approx 0.283$$

Nilai maksimum dari dua rasio tersebut adalah 0.3. Jika proses serupa dilakukan untuk klaster 1 dan 2, dan hasil akhirnya rata-rata dari ketiganya adalah misalnya 0.28, maka DBI ≈ 0.28.



Gambar 11. Hasil Evaluasi Performa Clustering (PerformanceVector)

Gambar 11 menampilkan hasil evaluasi performa segmentasi menggunakan operator Cluster Distance Performance di RapidMiner. Evaluasi ini menghasilkan nilai Davies-Bouldin Index (DBI) sebesar 0.426, yang menandakan bahwa hasil klasterisasi cukup baik. Nilai DBI ini menunjukkan bahwa pemisahan antar klaster cukup jelas dan kompak secara internal, karena semakin kecil nilai DBI maka semakin optimal kualitas segmentasi yang terbentuk. Dengan demikian, klaster yang terbentuk tidak tumpang tindih dan anggota dalam masing-masing klaster memiliki kemiripan karakteristik yang tinggi.

Selain nilai DBI, gambar ini juga menyajikan informasi tentang average within centroid distance, yaitu rata-rata jarak anggota klaster terhadap centroid-nya. Nilainya adalah 0.041 untuk klaster 0, 0.034 untuk klaster 1, dan 0.031 untuk klaster 2, dengan rata-rata keseluruhan sebesar 0.036. Nilai ini menunjukkan bahwa masing-masing klaster memiliki sebaran internal yang rendah, yang artinya data dalam klaster tersebut cukup rapat dan tidak menyebar terlalu jauh dari pusatnya. Ini merupakan indikator penting bahwa proses klasterisasi telah berjalan efektif dan menghasilkan kelompok yang homogen secara internal.

Secara keseluruhan, visualisasi hasil evaluasi ini menunjukkan bahwa segmentasi menggunakan algoritma K-Means pada atribut usia dan durasi penggunaan media sosial berhasil membentuk klaster yang terdefinisi dengan baik. Informasi ini sangat penting untuk memberikan dasar dalam pengambilan keputusan lanjutan, seperti strategi pemasaran, pengembangan fitur personalisasi, maupun analisis perilaku pengguna secara lebih mendalam. Hasil ini juga menunjukkan bahwa dua atribut sederhana sudah cukup untuk menghasilkan segmentasi yang bermakna dalam konteks perilaku pengguna media sosial.

IV. KESIMPULAN

Penelitian ini berhasil menunjukkan bahwa algoritma K-Means Clustering dapat diterapkan secara efektif untuk melakukan segmentasi pengguna media sosial berdasarkan dua atribut utama, yaitu usia (age) dan durasi penggunaan harian (time_spent). Meskipun atribut yang digunakan tergolong sederhana, hasil klasterisasi yang diperoleh mampu membagi pengguna menjadi tiga kelompok yang berbeda secara bermakna, yakni pengguna aktif, pengguna sedang, dan pengguna pasif. Ketiga klaster ini menggambarkan variasi perilaku pengguna yang dapat diamati secara jelas melalui pola interaksi mereka terhadap media sosial.

Seluruh proses dilakukan secara sistematis menggunakan platform RapidMiner, yang memfasilitasi tahapan mulai dari pemilihan atribut, normalisasi data, penerapan algoritma K-Means, hingga evaluasi hasil klasterisasi. Evaluasi performa dilakukan dengan menggunakan metrik Davies-Bouldin Index (DBI) dan average within centroid distance. Hasil evaluasi menghasilkan nilai DBI sebesar 0.426, yang menunjukkan kualitas pemisahan antar klaster yang cukup optimal. Selain itu, nilai average within centroid distance yang rendah di masing-masing klaster (0.041, 0.034, dan 0.031) menunjukkan bahwa sebaran data dalam setiap kelompok cukup kompak, sehingga memperkuat validitas hasil segmentasi.

Visualisasi hasil dalam bentuk scatter plot dan tabel centroid semakin memperjelas perbedaan antar klaster yang terbentuk. Tabel centroid memberikan informasi kuantitatif mengenai rata-rata nilai setiap atribut dalam masing-masing klaster, sedangkan scatter plot memberikan representasi visual terhadap pola pemisahan antar kelompok. Keduanya berperan penting dalam mendukung proses interpretasi hasil dan pemahaman terhadap karakteristik pengguna di setiap segmen.

Secara keseluruhan, penerapan K-Means dalam konteks ini membuktikan bahwa segmentasi yang efektif dapat diperoleh bahkan dengan atribut terbatas, selama dilakukan dengan pendekatan yang tepat. Temuan dari penelitian ini memiliki nilai praktis yang tinggi dalam pengembangan strategi digital, personalisasi layanan, dan kampanye pemasaran yang lebih relevan terhadap perilaku aktual pengguna. Di tengah meningkatnya kebutuhan akan pendekatan berbasis data, metode klasterisasi seperti K-Means menjadi salah satu solusi penting untuk memahami struktur dan dinamika perilaku pengguna secara objektif dan berkelanjutan.

V. REFERENSI

[1] T. C. Saputra, S. M. Fadhilah, S. U. Mangkuto, and J. Heikal, “Segmentation, targeting and positioning analysis using k-means clustering model: A case study of the laptop market in Indonesia,” 2024. [Online]. Available: www.ijafibs.pelnus.ac.id

[2] A. Riadi and I. Prayudi, “Cyberbullying Analysis on Instagram Using K-Means Clustering,” 2022.

[3] K. Tabianan, S. Velu, and V. Ravi, “K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data,” *Sustainability (Switzerland)*, vol. 14, no. 12, Jun. 2022, doi: 10.3390/su14127243.

[4] I. J. Cruickshank and K. M. Carley, “Characterizing Communities of Hashtag Usage on Twitter During the 2020 COVID-19 Pandemic by Multi-view Clustering,” Aug. 2020, [Online]. Available: http://arxiv.org/abs/2008.01139

[5] J. Banjarnahor, J. P. Hutagalung, and F. J. W. Sitorus, “Analyzing Consumer Shopping Interest via Social Media Ads with K-Means and C4.5 Algorithm,” *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 13, no. 3, pp. 416–421, Nov. 2024, doi: 10.32736/sisfokom.v13i3.2228.

[6] M. Ilhan Mansiz and Z. Fatah, “Pengelompokan Pengguna Media Sosial Berdasarkan Pola Interaksi Menggunakan K-Means,” pp. 388–397, Nov. 2024, doi: 10.59435/gjmi.v2i11.1100.

[7] R. Y. Daulay, R. A. Passalaras, and J. Heikal, “Customer Segmentation Using K-Means Clustering with SPSS Program in a Case Study of Consumer Interest in Current Coffee Shop,” *BUDGETING : Journal of Business, Management and Accounting*, vol. 5, no. 2, pp. 721–740, Apr. 2024, doi: 10.31539/budgeting.v5i2.9288.

[8] C. G. Lengari and I. Puspitasari, “Identifying Twitter Topics Using K-Means Clustering and Association Rule Mining for Improved Insights,” *Indonesian Journal of Artificial Intelligence and Data Mining*

(IJAIDM), vol. 8, no. 1, pp. 67–75, 2025, doi: 10.24014/ijaidm.v8i1.31720.

[9] X. Huang, M. J. Paul, R. Burke, F. Dernoncourt, and M. Dredze, "User Factor Adaptation for User Embedding via Multitask Learning," Feb. 2021, [Online]. Available: http://arxiv.org/abs/2102.11103

[10] I. YUNITA, P. R. Ali, M. A. Kartawidjaja, and R. Sukwadi, "Segmentasi Pelanggan Menggunakan K-Means Clustering: Menganalisis Metrik RFM untuk Strategi Pemasaran," *Jurnal Media Teknik dan Sistem Industri*, vol. 9, no. 1, p. 58, Mar. 2025, doi: 10.35194/jmtsi.v9i1.4452.

[11] R. W. Sembiring Brahmana, F. A. Mohammed, and K. Chairuang, "Customer Segmentation Based on RFM Model Using K-Means, K-Medoids, and DBSCAN Methods," *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, vol. 11, no. 1, p. 32, Apr. 2020, doi: 10.24843/lkjiti.2020.v11.i01.p04.

[12] J. Chitra and J. Heikal, "Customer segmentation using the K-Means Clustering algorithm in Foreign Banks in Indonesia," 2024.

[13] A. Gupta, A. Tiwari, and C. Sanwal, "Social Media Platform Using K-Mean Clustering," Apr. 2021.

[14] F. Hasan, K. S. Xu, J. R. Foulds, and S. Pan, "Learning User Embeddings from Temporal Social Media Data: A Survey," May 2021, [Online]. Available: http://arxiv.org/abs/2105.07996

[15] Sujeet Kumar Sahani and Dr. Sonam Singh, "Analysing social media with an Improved K-Means Clustering Algorithm," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 10, no. 4, pp. 31–38, Jul. 2024, doi: 10.32628/cseit24104106.