

Identifikasi Peluang Usaha Melalui Segmentasi Konsumen Menggunakan Pca Dengan Pendekatan Gaussian Mixture Model Dan K-Means

Meriani Silvia Tambunan¹, Susandri², Muhammad Rosyadi³

^{1,2,3}Magister Ilmu Komputer, Universitas Lancang Kuning
^{1,2,3}Jl. Yos Sudarso KM. 8 Rumbai, Pekanbaru, Riau, telp. 0811 753 2015
e-mail: [1merianisilvia1@gmail.com](mailto:merianisilvia1@gmail.com), [2susandri@unilak.ac.id](mailto:susandri@unilak.ac.id),
[3muhammadrosyadi26@gmail.com](mailto:muhammadrosyadi26@gmail.com)

Abstrak

Dalam era digital, data konsumen menjadi aset strategis dalam menganalisis tren dan mengidentifikasi peluang usaha. Penelitian ini bertujuan untuk mengidentifikasi segmentasi pelanggan berdasarkan perilaku belanja dengan menggunakan metode Principal Component Analysis (PCA) dan algoritma kluster Gaussian Mixture Model (GMM) dan algoritma K-Means. Dataset yang digunakan, yaitu "shopping_trends_updated.csv", memuat informasi terkait perilaku konsumen. Setelah melalui tahap pra-proses dan transformasi data, PCA diterapkan untuk mereduksi dimensi fitur, dengan memilih dua komponen utama. Hasil PCA menunjukkan bahwa dua komponen utama hanya mampu menjelaskan 23,59% variansi data, yang mengindikasikan kompleksitas fitur yang tinggi. Selanjutnya, GMM dibandingkan dengan K-Means untuk proses klusterisasi. Evaluasi menggunakan Silhouette Score menunjukkan bahwa kedua metode menghasilkan hasil yang sebanding: 0,4246 untuk K-Means dan 0,4202 untuk GMM. Meskipun skornya serupa, GMM lebih fleksibel dalam menangani bentuk kluster yang tidak berbentuk bulat. Temuan ini menunjukkan bahwa pendekatan PCA + GMM dapat digunakan sebagai metode awal untuk mengeksplorasi peluang usaha berbasis data perilaku konsumen. Studi ini juga membuka ruang untuk pengembangan metode lanjutan berbasis deep learning atau integrasi data multikanal guna memperoleh segmentasi yang lebih akurat dan aplikatif.

Kata kunci: PCA, GMM, Segmentasi Konsumen, Peluang Usaha, K-Means.

Abstract

In the digital era, consumer data has become a strategic asset for analyzing trends and identifying business opportunities. This study aims to identify customer segmentation based on shopping behavior using Principal Component Analysis (PCA) and the Gaussian Mixture Model (GMM) clustering algorithm and K-Means algorithm. The dataset used, "shopping_trends_updated.csv", contains detailed information about consumer behavior. After undergoing preprocessing and data transformation stages, PCA was applied to reduce the dimensionality of the features, selecting two principal components. The PCA results showed that the two components explained only 23.59% of the data variance, indicating high feature complexity. Subsequently, GMM was compared to K-Means for clustering. Evaluation using the Silhouette Score showed comparable results: 0.4246 for K-Means and 0.4202 for GMM. Although the scores were similar, GMM proved more flexible in handling non-spherical cluster shapes. These findings suggest that the PCA + GMM approach can serve as an initial method for exploring business opportunities based on consumer behavior data. This study also opens possibilities for further development using deep learning-based techniques or integrating multi-channel data sources to achieve more accurate and actionable segmentation results.

Keywords: PCA, GMM, Consumer Segmentation, Business Opportunity, K-Means.

1. PENDAHULUAN

Perkembangan teknologi informasi yang pesat dalam era digital telah menyebabkan akumulasi data dalam jumlah besar, khususnya dalam sektor perdagangan dan pemasaran. Data konsumen kini menjadi aset strategis untuk memahami perilaku pasar dan mengidentifikasi peluang usaha baru [1]. Meningkatnya volume data digital yang dihasilkan dari interaksi pengguna di berbagai platform, seperti media sosial, e-commerce, dan aplikasi berbasis internet,

memberikan peluang baru bagi perusahaan untuk mengimplementasikan strategi pemasaran yang lebih personal [2]. Analisis data ini menjadi elemen krusial karena kemampuannya dalam memproses volume data yang besar dan menghasilkan wawasan mendalam mengenai perilaku pelanggan.

Di tengah persaingan pasar yang semakin ketat, perusahaan menghadapi tantangan besar karena belum bisa memahami konsumen secara mendalam. Masalah ini diperparah oleh volume data konsumen yang sangat besar dan rumit, yang jika tidak dianalisis dengan baik, tidak akan memberikan nilai apa pun. Muhammad Faizal Fazri et al. menjelaskan bahwa solusi untuk tantangan ini adalah dengan memanfaatkan analisis big data. Jurnal tersebut menyoroti bahwa analisis big data menjadi aset strategis untuk mendapatkan wawasan tentang "perilaku pelanggan, preferensi, dan tren" serta pengambilan keputusan secara real-time, yang pada akhirnya memberikan keunggulan kompetitif yang berkelanjutan dan memungkinkan perusahaan untuk merespons perubahan pasar dengan lebih cepat dan akurat [3]. Salah satu tantangan utama dalam analisis data konsumen adalah tingginya dimensi dan kompleksitas fitur yang berasal dari data demografi serta perilaku belanja. Data ini sering kali mencakup berbagai variabel, mulai dari usia, lokasi, hingga riwayat pembelian dan interaksi di media sosial. Tanpa pendekatan yang tepat, volume data yang sangat besar dan rumit ini bisa menyulitkan proses analisis dan menyembunyikan pola-pola penting. Metode reduksi dimensi seperti Principal Component Analysis (PCA) telah terbukti efektif untuk menyederhanakan representasi data dan membantu visualisasi pola laten yang tersembunyi. [4]. Di sisi lain, algoritma klasterisasi seperti Gaussian Mixture Model (GMM) mampu menangani distribusi data yang kompleks secara probabilistik [5].

Kombinasi PCA dengan algoritma klasterisasi seperti GMM dan K-Means dapat membuka peluang dalam mengelompokkan konsumen ke dalam segmen-segmen yang bermakna secara bisnis. Hal ini menguatkan relevansi penggunaan kombinasi PCA pada GMM dan K-Means dalam penelitian ini, karena hasil segmentasi sangat dipengaruhi oleh pemilihan algoritma klustering dan tujuan penelitian untuk menemukan pola segmentasi konsumen, dimana setiap metode dapat memberikan performa yang berbeda pada data e-commerce.

Principal Component Analysis (PCA) adalah metode yang digunakan sebagai tahap pra-pemrosesan data. Tujuannya adalah untuk mereduksi dimensi data yang kompleks menjadi lebih sederhana, tanpa kehilangan informasi penting. Dengan kata lain, PCA menemukan pola dalam data dan mengidentifikasi fitur-fitur yang paling dominan, kemudian mengeliminasi faktor-faktor yang kurang relevan. Dalam penelitian ini, PCA mereduksi data menjadi dua komponen utama (PCA1 dan PCA2) yang kemudian menjadi input untuk proses klasterisasi. Hasil ini membantu menyederhanakan kompleksitas data sehingga analisis lebih mudah dilakukan [6]. PCA juga berfungsi sebagai tahap pra-pemrosesan data untuk menyederhanakan dimensi atribut penjualan, sehingga proses klasterisasi dengan algoritma K-Means menjadi lebih efisien dan terstruktur. Secara keseluruhan, kedua jurnal tersebut menegaskan bahwa PCA adalah alat yang efektif untuk menyederhanakan data yang kompleks, yang pada akhirnya mempermudah proses analisis dan visualisasi data [7].

GMM merupakan model probabilistik yang mengasumsikan bahwa semua titik data dihasilkan dari campuran beberapa distribusi Gaussian (normal) dengan parameter yang tidak diketahui. Setiap distribusi Gaussian dalam campuran ini mewakili sebuah klaster yang berbeda. Pendekatan ini memungkinkan segmentasi yang lebih fleksibel dan akurat, karena GMM dapat memodelkan klaster dengan bentuk dan ukuran yang bervariasi, tidak seperti metode lain yang memaksakan batas yang kaku. Jurnal tersebut menjelaskan bahwa GMM dapat mengakomodasi variabilitas yang melekat pada data pelanggan, memberikan pemahaman yang lebih bernuansa tentang segmen pelanggan [8].

K-Means adalah salah satu algoritma klustering non-hierarki yang sangat populer dan sering digunakan dalam bidang data mining. Algoritma ini dirancang untuk mengelompokkan objek data yang memiliki karakteristik serupa ke dalam satu klaster. Tujuan utamanya adalah untuk memastikan bahwa data di dalam satu klaster memiliki variasi yang sangat kecil (homogen),

sementara data di antara klaster yang berbeda memiliki variasi yang signifikan (heterogen). K-Means bekerja dengan mempartisi dataset menjadi sejumlah klaster (K) yang telah ditentukan sebelumnya. Proses ini bersifat iteratif, di mana algoritma akan terus menyesuaikan pusat klaster (centroid) hingga posisi mereka tidak lagi berubah [9]. Pada ranah online retail atau e-commerce K-Means digunakan untuk menganalisis data transaksi penjualan dengan fokus pada variabel seperti kuantitas dan harga satuan produk dengan mengelompokkan pelanggan ke dalam klaster-klaster berdasarkan kesamaan perilaku dan karakteristik mereka [10]

Melalui segmentasi, perusahaan dapat mengidentifikasi kelompok-kelompok konsumen yang homogen untuk merancang produk serta layanan yang lebih personal dan tepat sasaran, sehingga mampu meningkatkan keuntungan. Dengan menganalisis data yang dikumpulkan dari interaksi digital, bisnis dapat menentukan segmen pelanggan yang paling menguntungkan, yang memungkinkan mereka mengoptimalkan efektivitas kampanye pemasaran dengan menawarkan produk yang tepat kepada audiens yang tepat.

Terdapat beberapa penelitian terdahulu yang membahas tentang PCA dengan menggunakan Algoritma GMM dan K-Means. Menurut penelitian karya Andrean Maulana Artiarno dkk. (2025) berjudul “K-Means Clustering untuk Segmentasi Pelanggan: Mengungkap Pola Pembelian Strategi Pemasaran pada Sektor Ritel” membahas penerapan algoritma K-Means untuk mengelompokkan pelanggan ritel berdasarkan data transaksi sintetis dari Kaggle. Dengan menggunakan variabel usia, jenis kelamin, kategori produk, lokasi, jumlah pembelian, dan frekuensi transaksi, penelitian ini melalui tahapan preprocessing, reduksi dimensi dengan PCA, serta evaluasi klaster menggunakan elbow method, silhouette score, Calinski-Harabasz, dan Davies-Bouldin. Hasil analisis menghasilkan lima segmen pelanggan dengan karakteristik yang berbeda, seperti pelanggan remaja dengan pembelian periodik bernilai kecil hingga pelanggan dewasa bernilai tinggi namun jarang bertransaksi. Temuan ini menjadi dasar rekomendasi strategi pemasaran berbasis data, seperti program loyalitas, promosi musiman, hingga pendekatan personalisasi digital, sehingga mampu membantu perusahaan ritel meningkatkan efektivitas promosi dan retensi pelanggan [11].

Penelitian lainnya yang membahas tentang PCA dan GMM berjudul “Customer Segmentation and Analysis Based on Gaussian Mixture Model Algorithm” karya Eka Angga Laksana dan Marchel Maulana Fahrezi membahas penerapan algoritma Gaussian Mixture Model (GMM) dalam segmentasi pelanggan. Penelitian ini berangkat dari tantangan era digital, di mana perilaku konsumen online berubah cepat sehingga perusahaan perlu memahami preferensi pelanggan secara mendalam untuk merancang strategi pemasaran yang responsif dan relevan. Dengan menggunakan data pengeluaran konsumen, GMM digunakan untuk mereduksi dan mengelompokkan data ke dalam klaster yang merepresentasikan kecenderungan serta preferensi belanja. Hasil eksperimen menunjukkan terdapat empat klaster optimal yang dapat dijadikan dasar dalam merumuskan strategi pemasaran, seperti promosi gaya hidup sehat, peningkatan penjualan produk tertentu, serta edukasi bagi kelompok pelanggan dengan daya beli rendah. Penelitian ini juga menekankan pentingnya data preprocessing dan feature selection agar hasil segmentasi lebih akurat dan bermanfaat bagi perusahaan dalam menghadapi persaingan bisnis modern [12].

Berdasarkan kajian tersebut, dapat disimpulkan bahwa masih terdapat celah penelitian dalam mengintegrasikan metode PCA dan GMM untuk menganalisis data konsumen yang bersifat campuran. Penelitian ini hadir untuk menjawab celah tersebut dengan menguji efektivitas kombinasi PCA dan GMM, sekaligus membandingkannya dengan PCA + K-Means, dalam mengidentifikasi peluang usaha melalui segmentasi konsumen berbasis perilaku belanja.

2. METODE PENELITIAN

Penelitian ini bertujuan untuk mengidentifikasi segmen konsumen sebagai dasar eksplorasi peluang usaha, dengan mengombinasikan teknik reduksi dimensi Principal Component Analysis (PCA) dan algoritma clustering Gaussian Mixture Model (GMM). Proses penelitian dilakukan melalui beberapa tahapan sebagai berikut:

2.1 Dataset

Dataset yang digunakan adalah “shopping_trends_updated.csv” yang tersedia secara publik melalui platform Kaggle. Dataset ini memuat sebanyak 3900 baris dan 17 kolom, yang mencerminkan informasi demografis (seperti usia dan jenis kelamin), serta perilaku belanja konsumen seperti kategori produk, frekuensi pembelian, dan metode pembayaran. Dataset ini mencerminkan data campuran (numerik dan kategorikal), sehingga memerlukan pendekatan preprocessing yang tepat.

2.2 Pra-pemrosesan Data

Pra-pemrosesan data mencakup tahapan sebagai berikut:

- a. Penanganan Missing Value: Data numerik yang hilang diisi dengan rata-rata (mean), sedangkan data kategorikal diisi dengan modus (nilai terbanyak).
- b. Encoding Fitur Kategorikal: Seluruh fitur kategorikal dikonversi ke bentuk numerik menggunakan metode OneHotEncoding.
- c. Standarisasi Fitur Numerik: Fitur numerik distandardisasi dengan StandardScaler agar memiliki rata-rata nol dan deviasi standar satu, guna menghindari bias dalam pemrosesan PCA dan clustering.

2.3 Reduksi Dimensi dengan PCA

Principal Component Analysis (PCA) digunakan untuk mereduksi dimensi data hasil pra-pemrosesan. Dua komponen utama dipilih berdasarkan nilai proporsi variansi kumulatif tertinggi, yaitu sebesar 23.59%. Reduksi ini bertujuan untuk mempermudah proses visualisasi serta meningkatkan efisiensi komputasi pada tahap clustering.

2.4 Clustering

Dua metode klusterisasi diterapkan pada data hasil reduksi dimensi, yakni:

- 1) K-Means: Diimplementasikan dengan `KMeans(n_clusters=4, init='k-means++', random_state=42)`. Algoritma ini digunakan sebagai baseline untuk membandingkan hasil dengan GMM.
- 2) Gaussian Mixture Model (GMM): Diimplementasikan dengan `GaussianMixture(n_components=4, random_state=42)`, yang memungkinkan pembentukan klaster dengan distribusi elips atau tidak simetris.

2.5 Evaluasi Model

Evaluasi kualitas segmentasi dilakukan menggunakan metrik Silhouette Score, yang mengukur sejauh mana objek berada di dalam klasternya dibandingkan dengan klaster lain. Skor bernilai antara -1 hingga 1, di mana nilai lebih tinggi menunjukkan segmentasi yang lebih baik. Metrik ini dipilih karena sesuai untuk mengevaluasi hasil clustering pada data berdimensi rendah yang dihasilkan oleh PCA.

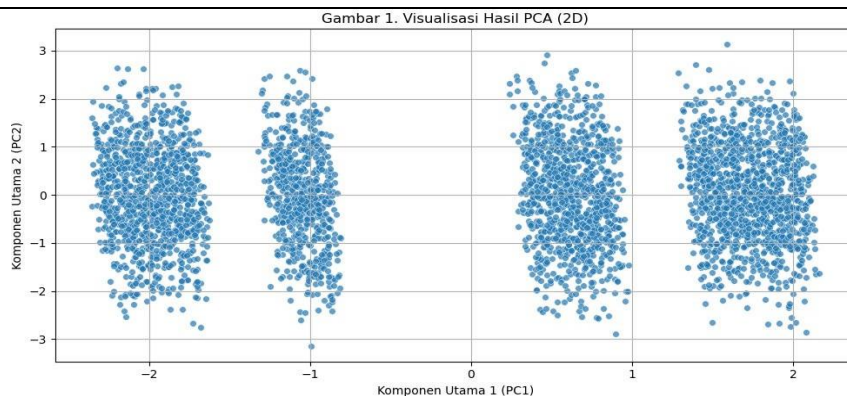
3. HASIL DAN PEMBAHASAN

3.1 Hasil Reduksi Dimensi dengan PCA

Setelah dilakukan pra-pemrosesan, PCA diterapkan untuk mereduksi dimensi dari 58 fitur hasil encoding menjadi dua komponen utama. Hasil PCA menunjukkan bahwa:

- Komponen pertama (PC1) menjelaskan 13.14% variansi data.
- Komponen kedua (PC2) menjelaskan 10.45% variansi data.

Untuk memvisualisasikan data berdimensi tinggi dan memudahkan proses klusterisasi, dilakukan reduksi dimensi menggunakan Principal Component Analysis (PCA) dengan dua komponen utama. Variansi total yang dapat dijelaskan oleh dua komponen utama PCA mencapai 23.59%, yang dianggap cukup representatif dalam menggambarkan struktur data secara umum.



Gambar 1. Visualisasi Hasil PCA (2D)

3.2 Hasil Clustering dan Evaluasi Silhouette Score

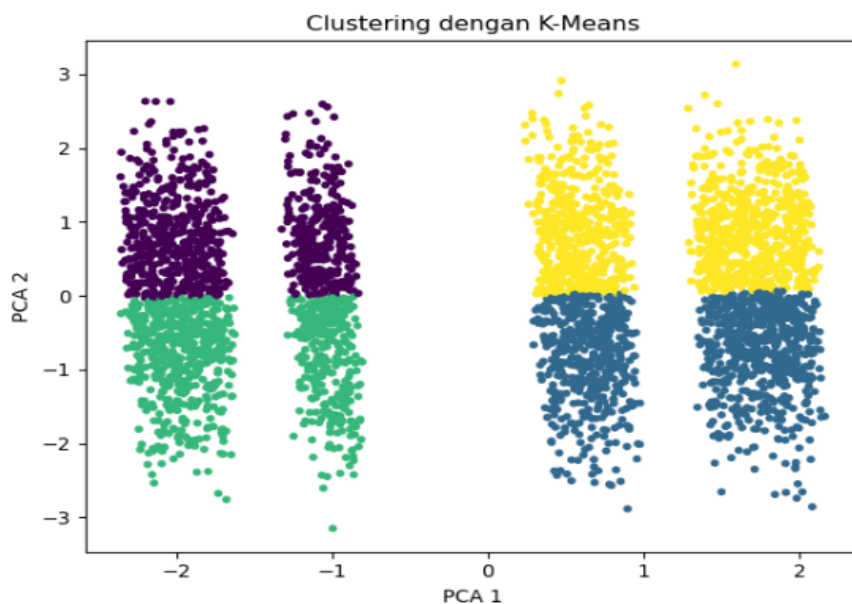
Dua algoritma clustering diterapkan pada data hasil PCA: K-Means dan Gaussian Mixture Model (GMM), masing-masing dengan jumlah kluster sebanyak 4.

Tabel 1 berikut menampilkan nilai Silhouette Score sebagai metrik evaluasi untuk masing-masing model:

Tabel 1 Perbandingan Skor Silhouette untuk K-Means dan GMM

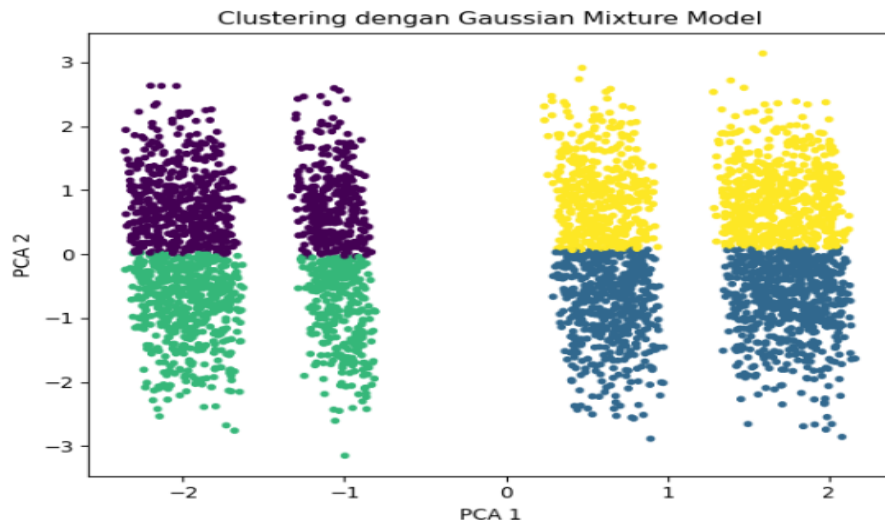
Algoritma Clustering	Silhouette Score
K-Means (k=4)	0.4246
GMM (n=4)	0.4202

Nilai Silhouette Score yang diperoleh dari kedua metode menunjukkan performa clustering yang relatif baik dan setara, meskipun secara umum K-Means menghasilkan kluster yang lebih terpisah secara geometris, sedangkan GMM lebih fleksibel dalam bentuk distribusi kluster.



Gambar 2. Hasil Klasterisasi dengan K-Means (PCA 2D)

Visualisasi ini menunjukkan hasil segmentasi konsumen menggunakan algoritma K-Means. Empat kluster divisualisasikan dengan warna berbeda. Distribusi terlihat simetris secara horizontal, mencerminkan kemampuan K-Means dalam membentuk kluster yang seimbang berdasarkan jarak. Silhouette Score K-Means sebesar 0.4246 mengindikasikan pemisahan kluster yang cukup baik.



Gambar 3. Hasil Klasterisasi dengan GMM (PCA 2D)

Visualisasi ini menampilkan hasil klasterisasi menggunakan algoritma Gaussian Mixture Model. Dengan pendekatan probabilistik, GMM memberikan fleksibilitas dalam mendeteksi bentuk klaster yang tidak harus simetris. Meski hasilnya secara visual menyerupai K-Means, GMM sedikit tertinggal dari sisi performa dengan Silhouette Score sebesar 0.4202.

3.3 Hasil Clustering dan Evaluasi Silhouette Score

- **Efektivitas PCA:** Dengan hanya menjelaskan 23.59% variansi data, PCA terbukti tetap membantu dalam menyederhanakan kompleksitas fitur dan memungkinkan visualisasi klaster yang cukup jelas.
- **GMM vs. K-Means:** Meskipun K-Means memiliki Silhouette Score yang sedikit lebih tinggi, GMM memiliki fleksibilitas lebih tinggi dalam mendeteksi bentuk klaster yang tidak berbentuk bulat sempurna.
- **Implikasi Bisnis:** Identifikasi segmen berdasarkan perilaku konsumen dapat memberikan wawasan bagi pelaku usaha dalam merancang strategi pemasaran yang lebih personal, serta menemukan ceruk pasar berdasarkan pola konsumsi tertentu.

4. KESIMPULAN

Kombinasi Principal Component Analysis (PCA) dengan Gaussian Mixture Model (GMM) dan K-Means dapat dimanfaatkan untuk mengidentifikasi segmentasi konsumen sebagai dasar dalam menemukan peluang usaha. PCA membantu mereduksi dimensi data sehingga proses klasterisasi menjadi lebih efisien dan mudah divisualisasikan, meskipun hanya 23,59% variansi data yang dapat dijelaskan oleh dua komponen utama. Hasil evaluasi menunjukkan bahwa metode K-Means dan GMM memiliki performa yang sebanding dengan Silhouette Score masing-masing sebesar 0,4246 dan 0,4202. Namun, GMM lebih unggul karena mampu menangani distribusi data yang tidak berbentuk bulat serta memberikan pendekatan probabilistik dalam menentukan keanggotaan klaster.

UCAPAN TERIMA KASIH

Terima kasih penulis sampaikan kepada pihak institusi yang telah memberikan fasilitas dan dukungan, para dosen pembimbing serta rekan peneliti yang telah membantu melalui arahan dan diskusi konstruktif, serta keluarga dan sahabat yang senantiasa memberikan doa dan dukungan moril sehingga penelitian berjudul "*Identifikasi Peluang Usaha melalui Segmentasi Konsumen Menggunakan PCA dengan Pendekatan Gaussian Mixture Model dan K-Means*" dapat diselesaikan dengan baik.

DAFTAR PUSTAKA

- K. Tabianan, S. Velu, and V. Ravi, "K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data," *Sustainability*, vol. 14, no. 7243, pp. 1–15, Jun. 2022.
- [2] I. H. Na, Y. I. Jae, and P. I. Hwa, "Personalized Digital Marketing Strategies: A Data-Driven Approach Using Marketing Analytics," *Journal of Management and Informatics (JMI)*, vol. 4, no. 1, pp. 668–686, Apr. 2025.
- [3] T. Hariguna and S. C. Chen, "Customer Segmentation and Targeted Retail Pricing in Digital Advertising using Gaussian Mixture Models for Maximizing Gross Income," *Journal of Digital Marketing and Digital Currency*, vol. 1, no. 2, pp. 183–203, 2024.
- [4] T. Wahyudi and T. Silfia, "Implementation of Data Mining Using K-Means Clustering Method to Determine Sales Strategy in S&R Baby Store," *Journal of Applied Engineering and Technological Science*, vol. 4, no. 1, pp. 93–103, 2022.
- [5] E. F. L. Awalina and W. I. Rahayu, "Optimalisasi Strategi Pemasaran dengan Segmentasi Pelanggan Menggunakan Penerapan K-Means Clustering pada Transaksi Online Retail," *Jurnal Teknologi dan Informasi (JATI)*, vol. 13, no. 2, pp. 122–130, Sep. 2023.
- [6] W. R. Nurputra, W. Witanti, and A. Komarudin, "Principal Component Analysis (PCA) Untuk Meningkatkan Hasil Klasterisasi Penjualan Video Game Menggunakan Algoritma K-Means," *Jurnal Locus: Penelitian & Pengabdian*, vol. 4, no. 8, pp. 4683–4685, Jul. 2025.
- [7] S. Mulyaningsih and J. Heikal, "K-Means Clustering Using Principal Component Analysis (PCA) Indonesia Multi-Finance Industry Performance Before and During Covid-19," *Asia-Pacific Management and Business Application*, vol. 11, no. 2, pp. 131–142, 2022.
- [8] M. F. Fazri, T. Ramadhan, D. Apriliasari, D. Julianingsih, and A. Fitzroy, "Leveraging Big Data Analytics for Strategic Marketing Optimization: Insights and Impacts," *Journal of Computer Science and Technology Application (CORISINTA)*, vol. 1, no. 2, pp. 144–153, 2024.
- [9] A. Fitzroy, D. Julianingsih, and M. F. Fazri, "Leveraging Big Data Analytics for Strategic Marketing Optimization: Case Study in Digital Business," *Journal of Information Systems Research*, vol. 5, no. 1, pp. 55–66, 2024.
- [10] E. F. L. Awalina and W. I. Rahayu, "Optimalisasi Strategi Pemasaran dengan Segmentasi Pelanggan Menggunakan K-Means Clustering pada Transaksi E-Retail," *Jurnal Teknologi dan Sistem Informasi Bisnis*, vol. 12, no. 3, pp. 211–220, 2023.
- [11] A. M. Artiarno, P. Setiaji, and F. Nugraha, "K-Means Clustering untuk Segmentasi Pelanggan: Mengungkap Pola Pembelian Strategi Pemasaran pada Sektor Ritel," *Edumatic: Jurnal Pendidikan Informatika*, vol. 9, no. 2, pp. 442–451, 2025.
- [12] E. A. Laksana and M. M. Fahrezi, "Customer Segmentation and Analysis Based on Gaussian Mixture Model Algorithm," *Proceedings/Widyatama University, Bandung, Indonesia*, 2025.