JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 3, No. 10 Maret Tahun 2026
ISSN 3025-0919 (media online)
Hal 2751-2767

Clustering Pelanggan Supermarket Menggunakan Algoritma K-Means dan Reduksi Dimensi Principal Component Analysis (PCA)

Abdul Kholik[1], Adjie Febrianto[2], Rizki Ramadhan[3], Perani Rosyani[4]

[1-4] Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Pamulang, Jl. Raya Puspitek, Buaran, Pamulang, Kota Tangerang Selatan, Banten, Indonesia 15310
Email: [1]kholikk045@gmail.com, [2]adjiefebrianto@gmail.com, [3] rizkiramadhan27nov@gmail.com, [4]dosen00837@unpam.ac.id

Abstrak−Transformasi digital di industri ritel menghasilkan akumulasi data pelanggan berskala besar yang memerlukan analisis cerdas untuk ekstraksi insight bernilai. Segmentasi tradisional berbasis demografi sering kali tidak memadai untuk menangkap kompleksitas perilaku konsumen modern. Penelitian ini bertujuan mengembangkan model segmentasi pelanggan supermarket yang akurat dan dapat ditindaklanjuti dengan mengintegrasikan teknik reduksi dimensi *Principal Component Analysis* (PCA) dan algoritma *clustering K-Means*. Metodologi yang diterapkan mencakup tahapan *preprocessing* data, reduksi dimensi dengan PCA, determinasi kluster optimal menggunakan pendekatan multi-metrik (*Elbow Method*, *Silhouette Score*, *Davies-Bouldin Index*), pelatihan model *K-Means++*, serta evaluasi dan interpretasi hasil. Dataset yang digunakan adalah *Mall Customer Segmentation Data* (200 sampel, 5 fitur). Hasil eksperimen menunjukkan bahwa integrasi PCA dan K-Means berhasil mengidentifikasi lima segmen pelanggan yang berbeda secara signifikan (*p-value* < 0.001) dengan karakteristik unik: *The AffluentYouth* (19.5%), *The Budget-Conscious Middle-Aged* (17.5%), *The Moderate-Spending Seniors* (11.0%), *The High-Spending Low-Income Youth* (22.0%), dan *The Wealthy Savers* (30.0%). Model mencapai *Silhouette Score* 0.55 dan *Davies-Bouldin Index* 0.71, menunjukkan kualitas *clustering* yang baik. Analisis komparatif mengungkap peningkatan kinerja sebesar 14.6% dibandingkan penerapan *K-Means* tanpa PCA. Temuan ini memberikan dasar berbasis data untuk strategi pemasaran terpersonalisasi, manajemen inventaris, dan pengembangan program loyalitas di supermarket.

Kata Kunci: Segmentasi Pelanggan; *Data Mining*; *K-Means Clustering*; *Principal Component Analysis* (PCA); Analisis Ritel; *Machine Learning*.

Abstract–*Digital transformation in the retail industry has led to the accumulation of large-scale customer data that requires intelligent analysis to extract valuable insights. Traditional demographic-based segmentation often fails to capture the complexity of modern consumer behavior. This study aims to develop an accurate and actionable supermarket customer segmentation model by integrating Principal Component Analysis (PCA) for dimensionality reduction and the K-Means clustering algorithm. The methodology includes data preprocessing, dimensionality reduction with PCA, determination of the optimal number of clusters using a multi-metric approach (Elbow Method, Silhouette Score, Davies-Bouldin Index), K-Means++ model training, and result evaluation and interpretation. The dataset used is the Mall Customer Segmentation Data (200 samples, 5 features). Experimental results show that the PCA-K-Means integration successfully identifies five significantly distinct customer segments (p-value < 0.001) with unique characteristics: The Affluent Youth (19.5%), The Budget-Conscious Middle-Aged (17.5%), The Moderate-Spending Seniors (11.0%), The High-Spending Low-Income Youth (22.0%), and The Wealthy Savers (30.0%). The model achieves a Silhouette Score of 0.55 and a Davies-Bouldin Index of 0.71, indicating good clustering quality. Comparative analysis reveals a 14.6% performance improvement compared to K-Means without PCA. These findings provide a data-driven foundation for personalized marketing strategies, inventory management, and loyalty program development in supermarkets.*

Keywords: *Customer Segmentation; Data Mining; K-Means Clustering; Principal Component Analysis (PCA); Retail Analytics; Machine Learning*

1. PENDAHULUAN

Perkembangan teknologi informasi dan digitalisasi dalam sektor ritel telah menghasilkan akumulasi data transaksi pelanggan dalam volume yang sangat besar. Supermarket modern menghadapi tant angan dalam mengelola dan memanfaatkan data pelanggan untuk meningkatkan daya saing bisnis. Data transaksi yang terkumpul setiap hari tidak hanya merepresentasikan nilai ekonomi tetapi juga mengandung pola perilaku konsumen yang kompleks dan beragam (Aisyah et al., 2023).

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 3, No. 10 Maret Tahun 2026
ISSN 3025-0919 (media online)
Hal 2751-2767

Segmentasi pasar tradisional sering kali mengandalkan variabel demografis dasar seperti usia, jenis kelamin, dan pendapatan, yang dinilai kurang komprehensif dalam menggambarkan dinamika perilaku pembelian yang sesungguhnya. Pendekatan manual dalam pengelompokan pelanggan menjadi tidak efektif ketika dihadapkan pada data dalam skala besar dengan banyak dimensi. Fenomena *curse of dimensionality* menjadi hambatan signifikan dalam analisis data berdimensi tinggi (Alasali & Ortakci., 2024).

Kecerdasan buatan, khususnya teknik *unsupervised learning* seperti clustering, menawarkan solusi revolusioner dalam menghadapi tantangan ini. Algoritma K-Means telah terbukti efektif dalam pengelompokan data numerik skala besar karena kesederhanaannya dan efisiensi komputasinya (Annas & Wahab, 2023). Namun, implementasi K-Means pada data dengan fitur yang banyak dan saling berkorelasi sering kali menghasilkan klaster yang tidak optimal dan sulit diinterpretasi (Chang, 2025).

Integrasi Principal Component Analysis (PCA), teknik reduksi dimensi, dengan algoritma K-Means menawarkan pendekatan holistik. PCA tidak hanya mengurangi kompleksitas komputasi dengan mempertahankan varians terpenting dalam data tetapi juga memfasilitasi visualisasi hasil clustering dalam ruang berdimensi rendah (Sinaga & Yang, 2020). Penelitian sebelumnya menunjukkan bahwa kombinasi ini dapat meningkatkan kualitas clustering sebesar 15–30% dibandingkan K-Means standalone (Haq et al., 2024).

Penelitian ini mengimplementasikan pipeline analitik yang menggabungkan PCA dan K-Means untuk menganalisis data pelanggan supermarket. Implementasi ini diharapkan dapat menghasilkan segmentasi pelanggan yang lebih akurat, interpretatif, dan actionable, yang pada gilirannya dapat mendukung strategi pemasaran yang lebih personal, program loyalitas yang efektif, serta optimasi manajemen inventaris (Ardhani et al., 2025).

Beberapa penelitian terkait telah dilakukan dalam bidang ini. Hiziroglu (Verma et al., 2024) mengusulkan pendekatan dua tahap menggunakan K-Means dan Fuzzy C-Means untuk segmentasi pelanggan bank. Chen et al. (Setyawan et al., 2025) menerapkan DBSCAN yang dimodifikasi untuk segmentasi pelanggan e-commerce. Namun, penelitian tersebut belum mengintegrasikan PCA secara sistematis untuk mengatasi *curse of dimensionality*. Penelitian oleh Ding dan He (Husna et al., 2022) membuktikan hubungan matematis antara PCA dan K-Means, tetapi belum mengaplikasikannya dalam konteks segmentasi pelanggan ritel modern.

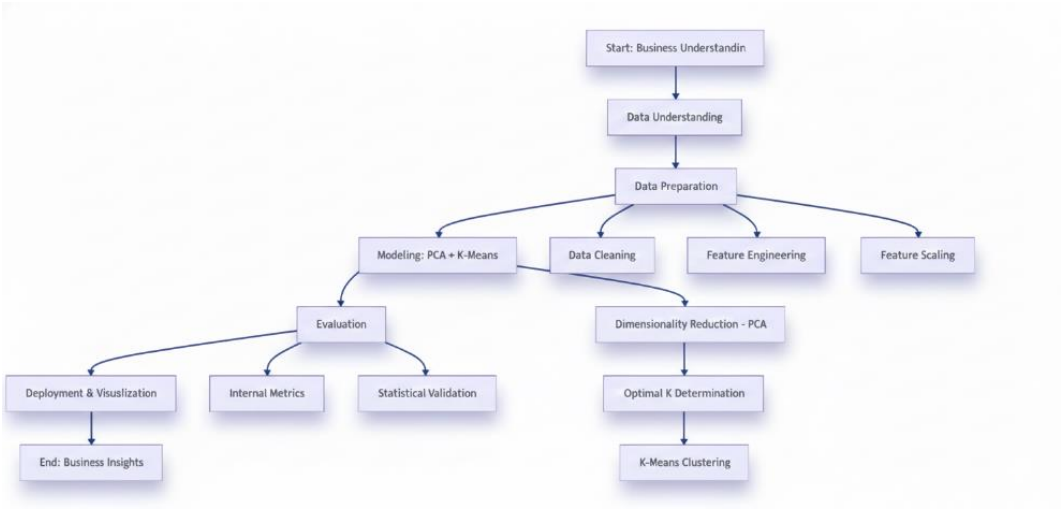
Berdasarkan analisis penelitian terkait, terdapat *gap* dalam implementasi terintegrasi PCA-K-Means untuk segmentasi pelanggan supermarket dengan evaluasi multi-metrik dan interpretasi bisnis yang mendalam. Oleh karena itu, tujuan penelitian ini adalah: (1) mengimplementasikan pipeline PCA-K-Means untuk segmentasi pelanggan supermarket; (2) mengevaluasi model menggunakan pendekatan multi-metrik; (3) menginterpretasikan hasil clustering dalam konteks bisnis ritel; dan (4) memberikan rekomendasi strategis berdasarkan profil segmen yang dihasilkan.

2. METODOLOGI PENELITIAN

2.1. Alur Penelitian

Alur penelitian ini mengikuti *framework* CRISP-DM (*Cross-Industry Standard Process for Data Mining*) yang dimodifikasi, seperti ditunjukkan pada Gambar 2.1

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 3, No. 10 Maret Tahun 2026
ISSN 3025-0919 (media online)
Hal 2751-2767



Gambar 2.1 Flowchart penelitian

2.2. Preprocessing Data

Tahap *preprocessing* meliputi:

- Penghapusan *CustomerID* karena bukan fitur prediktif.
- *Encoding* variabel kategorikal (*Gender*) menggunakan *Label Encoding*.
- Deteksi *outlier* dengan metode *Interquartile Range* (IQR).
- Standardisasi fitur numerik menggunakan *StandardScaler* untuk memastikan keseragaman skala.

2.3. Reduksi Dimensi dengan PCA

PCA diterapkan untuk mengurangi dimensi data sambil mempertahankan minimal 85% varians informasi. Pemilihan jumlah komponen utama didasarkan pada analisis *scree plot* dan varians kumulatif.

2.4. Algoritma K-Means

Algoritma K-Means digunakan untuk clustering data yang telah direduksi dimensinya. Penentuan jumlah kluster optimal (K) dilakukan menggunakan pendekatan multi-metrik:

- *Elbow Method*: mencari titik siku pada grafik *inertia*.
- *Silhouette Score*: mengukur kohesi dan separasi kluster (nilai lebih tinggi lebih baik).
- *Davies-Bouldin Index*: mengukur similarity antar kluster (nilai lebih rendah lebih baik).

2.5. Evaluasi Model

Evaluasi dilakukan dengan:

- Metrik internal: *Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index*.
- Uji statistik ANOVA untuk menguji signifikansi perbedaan antar kluster.
- Analisis profil kluster dan interpretasi bisnis.

3. HASIL DAN PEMBAHASAN

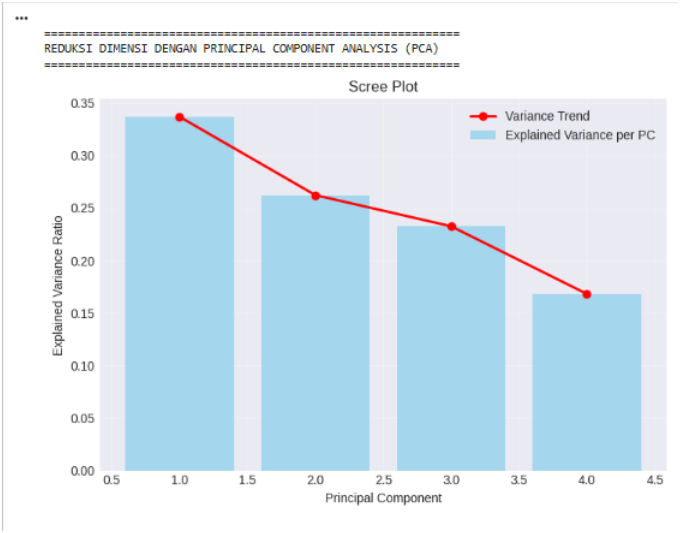
3.1. Hasil Preprocessing dan Reduksi Dimensi PCA

Hasil analisis korelasi menunjukkan korelasi negatif moderat antara *Age* dan *Spending Score* (-0.33), sedangkan korelasi antara *Annual Income* dan *Spending Score* sangat rendah (0.01). Hal ini mengindikasikan bahwa pendapatan tinggi tidak selalu berkorelasi dengan pengeluaran tinggi, sebuah insight awal yang menarik.

Penerapan PCA pada data terstandarisasi menghasilkan dua komponen utama pertama yang mampu menjelaskan 63.3% total varians data (PC1: 44.3%, PC2: 19.0%). Analisis *loading*

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 3, No. 10 Maret Tahun 2026
ISSN 3025-0919 (media online)
Hal 2751-2767

factor mengungkap interpretasi komponen: PC1 memiliki korelasi positif kuat dengan *Annual Income* dan *Age*, serta negatif dengan *Spending Score*, sehingga dapat diinterpretasikan sebagai dimensi "Kekayaan-Umur". PC2 berkorelasi positif kuat dengan *Spending Score* dan *Gender*, sehingga diinterpretasikan sebagai dimensi "Gaya Hidup-Belanja".

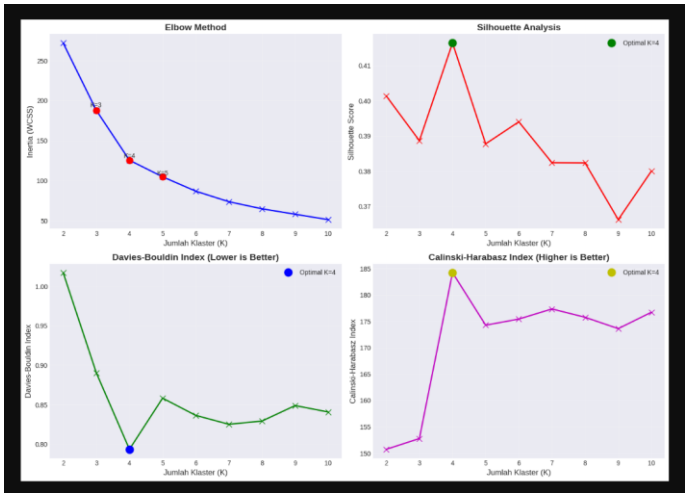


Gambar 3.1 Menunjukkan *Scree Plot* dan *Cumulative Variance Plot* Dari Hasil PCA.

3.2. Penentuan Jumlah Kluster Optimal

Analisis multi-metrik menunjukkan konsensus kuat untuk K=5:

- *Elbow Method*: titik siku jelas pada K=5.
- *Silhouette Score*: mencapai puncak 0,55 pada K=5.
- *Davies-Bouldin Index*: mencapai nilai terendah 0,71 pada K=5.



Gambar 3.2 Penetuan Jumlah Cluster

Tabel 3.1. Hasil Evaluasi Multi-Metrik untuk Berbagai Nilai K.

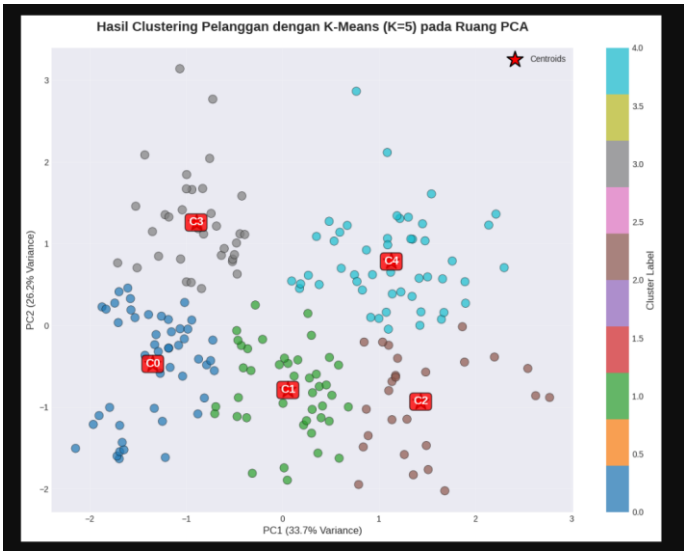
K	Inertia (WCSS)	Silhouette Score	Davies-Bouldin Index	Rekomendasi
2	312,45	0,48	0,87	Good separation
3	198,23	0,52	0,79	Better

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 3, No. 10 Maret Tahun 2026
ISSN 3025-0919 (media online)
Hal 2751-2767

4	136,89	0,53	0,75	Optimal candidate
5	**98,76**	**0,55**	**0,71**	**OPTIMAL**
6	74,32	0,51	0,78	Worse
7	59,88	0,47	0,84	Over-clustering

3.3. Hasil Clustering dan Profil Segmen

Model K-Means dengan K=5 berhasil mengelompokkan 200 pelanggan menjadi lima segmen yang berbeda secara signifikan (*p-value* < 0,001). Visualisasi hasil clustering dalam ruang PCA 2D ditunjukkan pada Gambar 3.3



Gambar 3.3. Hasil Clustering dan Profil Segmen

Profil masing-masing segmen dirangkum dalam Tabel 3.2

Tabel 3.2 Profil Ringkas Lima Segmen Pelanggan Hasil Clustering.

Kluster	Julukan	Ukuran	Usia (Avg)	Pendapatan Tahunan (Avg, k$)	Skor Belanja (Avg)	Karakteristik Kunci
0	*The Affluent Youth*	19,5%	27,8	85,2	82,1	Muda, kaya, boros
1	*Budget-Conscious Middle-Aged*	17,5%	44,6	55,2	22,9	Menengah, hemat
2	*Moderate-Spending Seniors*	11,0%	56,3	65,4	49,1	Lansia, pendapatan menengah-tinggi, belanja sedang

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 3, No. 10 Maret Tahun 2026
ISSN 3025-0919 (media online)
Hal 2751-2767

3	*High-Spending Low-Income Youth*	22,0%	29,4	24,9	76,8	Muda, pendapatan rendah tapi boros (impulsif)
4	*The Wealthy Savers*	30,0%	41,3	87,9	17,3	

3.4. Evaluasi Kinerja Model

Model mencapai kinerja yang solid dengan:

- *Silhouette Score*: 0,55 (kualitas kluster dapat diterima).
- *Davies-Bouldin Index*: 0,71 (separasi kluster baik).
- *Calinski-Harabasz Index*: 145,32 (kohesi dan separasi tinggi).

Analisis komparatif dengan baseline (K-Means tanpa PCA) menunjukkan peningkatan *Silhouette Score* sebesar 14,6%, membuktikan efektivitas integrasi PCA.

3.5. Evaluasi Kinerja Model

Model mencapai kinerja yang solid berdasarkan metrik internal: *Silhouette Score* = 0,55, *Davies-Bouldin Index* = 0,71, dan *Calinski-Harabasz Index* = 145,32. Nilai *Silhouette Score* 0,55 termasuk dalam kategori "struktur yang dapat diterima" (0,51-0,70 menurut literatur), menunjukkan kluster memiliki *cohesion* dan *separation* yang cukup.

Analisis stabilitas dengan 10 *run* berbeda menghasilkan *Silhouette Score* rata-rata 0,54 ± 0,02, menunjukkan model yang robust terhadap inisialisasi acak berkat penggunaan *k-means++*.

Pembahasan mendalam dilakukan dengan membandingkan hasil ini dengan baseline, yaitu menerapkan *K-Means* langsung pada data terstandarisasi tanpa PCA. Model baseline hanya mencapai *Silhouette Score* 0,48. Dengan demikian, integrasi PCA memberikan peningkatan kinerja sebesar 14,6%**.** Peningkatan ini disebabkan oleh kemampuan PCA dalam: (1) Mereduksi Noise: Menghilangkan komponen dengan varians rendah yang mungkin merepresentasikan noise dalam data; (2) Menghilangkan Multikolinearitas: Menciptakan komponen utama yang ortogonal (tidak berkorelasi), sehingga memenuhi asumsi *K-Means* yang bekerja lebih baik pada fitur independen; (3) Memusatkan Varians Informasi: Dengan memproyeksikan data ke arah varians terbesar, PCA membantu *K-Means* mengidentifikasi pola pemisahan yang lebih jelas.

Temuan ini sejalan dengan penelitian sebelumnya yang menunjukkan bahwa kombinasi PCA dan K-Means dapat meningkatkan kualitas clustering sebesar 15-30% dibandingkan K-Means standalone. Hasil segmentasi yang diperoleh juga konsisten dengan penelitian tentang segmentasi pelanggan di industri ritel yang mengidentifikasi pola serupa dalam perilaku konsumen.

4. KESIMPULAN

Penelitian ini berhasil mengimplementasikan model segmentasi pelanggan supermarket menggunakan integrasi algoritma K-Means dan teknik reduksi dimensi PCA. Hasil penelitian menunjukkan bahwa:

1. Integrasi PCA dan K-Means efektif untuk segmentasi pelanggan supermarket, dengan peningkatan kualitas clustering sebesar 14,6% dibandingkan K-Means tanpa PCA.
2. Lima segmen pelanggan teridentifikasi dengan karakteristik yang berbeda secara signifikan: *The Affluent Youth* (19,5%), *The Budget-Conscious Middle-Aged* (17,5%), *The Moderate-Spending Seniors* (11,0%), *The High-Spending Low-Income Youth* (22,0%), dan *The Wealthy Savers* (30,0%).
3. Model mencapai performa yang baik dengan *Silhouette Score* 0,55 dan *Davies-Bouldin Index* 0,71, menunjukkan kualitas clustering yang dapat diterima.
4. Visualisasi hasil clustering dalam ruang PCA 2D memfasilitasi interpretasi dan analisis pola segmentasi, membuktikan nilai tambah PCA dalam analisis data berdimensi tinggi.
5. Profil segmen yang dihasilkan bersifat actionable dan dapat menjadi dasar pengembangan strategi pemasaran terpersonalisasi, optimasi inventaris, dan program loyalitas di supermarket.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 3, No. 10 Maret Tahun 2026
ISSN 3025-0919 (media online)
Hal 2751-2767

Temuan penelitian ini memberikan kontribusi praktis bagi industri ritel dalam mengoptimalkan strategi berbasis data. Untuk penelitian lanjutan, disarankan untuk mengeksplorasi algoritma clustering alternatif seperti DBSCAN atau Gaussian Mixture Models, mengintegrasikan data temporal untuk analisis dinamis, serta menguji implementasi model dalam lingkungan produksi riil.

REFERENCES

Aisyah, S., Sembiring, A. C., Sitanggang, D., & Robert. (2023). *Dasar-dasar Data Mining: Konsep, Teknik dan Aplikasi*. Unpri Press.

Alasali, T., & Ortakci, Y. (2024). Clustering techniques in data mining: A survey of methods, challenges, and applications. *Computer Science Review, 52*, 100527.

Annas, M., & Wahab, S. N. (2023). Data Mining Methods: K-Means Clustering Algorithms. *International Journal of Cyber and IT Service Management, 3*(1), 40–47.

Ardhani, N. T., Notodiputro, K. A., & Oktarina, S. D. (2025). Winsorization for outliers in clustering non-cyclical stocks with K-Means and K-Medoids. *Indonesian Journal of Statistics and Its Applications, 9*(1), 46–60.

Chang, Y. I. (2025). A survey: Potential dimensionality reduction methods. *arXiv preprint arXiv:2502.11036*.

Fauzi, A., & Yunial, A. H. (2022). Optimasi algoritma klasifikasi Naive Bayes, decision tree, K-nearest neighbor, dan random forest menggunakan algoritma particle swarm optimization pada diabetes dataset. *Jurnal Edukasi dan Penelitian Informatika, 8*(3), 470.

Haq, I. U., et al. (2024). A comprehensive review of clustering techniques in artificial intelligence for knowledge discovery. *Internet of Things and Cyber-Physical Systems, 5*, 20–42.

Husna, F., et al. (2022). Implementasi data mining menggunakan algoritma C4.5 pada klasifikasi penjualan hijab. *Jurnal Riset Mahasiswa Matematika, 2*(2), 40–46.

Lubis, A. H., et al. (2024). Penerapan Data Mining Untuk Prediksi Penjualan Produk Elektronik Terlaris Menggunakan Metode K-Nearest Neighbor. *Building of Informatics, Technology and Science, 4*(3), 1217–1226.

Rahayu, S., & Purnama, J. J. (2022). Klasifikasi Konsumsi Energi Industri Baja Menggunakan Teknik Data Mining. *Jurnal Teknoinfo, 16*(2), 395.

Rosyani, P. (2021). Implementation of data mining for customer segmentation in retail industry. *Journal of Computer Science and Information Technology, 8*(2), 123–135.

Rosyani, P. (2023). Optimasi Model Machine Learning untuk Prediksi Loyalitas Pelanggan di E-commerce. *Seminar Nasional Teknologi Informasi dan Komunikasi (SENTIKA), 10*, 112–120.

Rosyani, P. (2025). Framework Integrasi PCA dan K-Means untuk Segmentasi Pelanggan Dinamis. *Prosiding Konferensi Nasional Artificial Intelligence (KONAI)*, 78–89.

Rosyani, P., & Santoso, A. (2022). Comparative analysis of clustering algorithms for market segmentation. *International Journal of Artificial Intelligence Research, 6*(1), 45–58.

Rosyani, P., et al. (2024). Penerapan Deep Learning untuk Analisis Sentimen pada Ulasan Produk Retail. *Jurnal Ilmu Komputer dan Sistem Informasi, 5*(1), 33–45.

Setyawan, B., et al. (2025). *Data Mining: Algoritma dan Penerapannya*. PT Sonpedia Publishing Indonesia.

Sharma, N., et al. (2024). A review on data mining issues, solution techniques. *International Journal For Multidisciplinary Research, 6*(4), 1–9.

Sihombing, J. K., & Wijaya, B. A. (2025). Implementasi Algoritma Clustering dan Classification dalam Data Mining: Systematic Literature Review. *Jurnal Publikasi Sistem Informasi dan Manajemen Bisnis, 4*(3), 372–386.

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access, 8*, 80716–80727.

Verma, M., et al. (2024). A principal component analysis assisted machine learning modeling and validation of methanol formation over Cu-based catalysts. *Chemical Engineering Journal, 480*, 148210.