Journal of Science and Social Research ISSN 2615 – 4307 (Print)
August 2025, VIII (3): 3908 – 3914 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

MENERAPKAN DATA SCIENCE PADA DATASET REVIEW PRODUK DI SHOPEE DAN TOKOPEDIA: PENGELOMPOKAN PELANGGAN DAN STRATEGI RETENSI DENGAN TEKNIK CLUSTERING

Yogo Turnandes[1], Rezka Afrilli[2]
Universitas Lancang Kuning
email: [1]turnandes@unilak.ac.id, [2]rezka.afrilli1234@gmail.com

Abstract: *This study employs data science methodologies to analyze product reviews from the e-commerce sites Shopee and Tokopedia. The primary objective is to segment customers by grouping them according to their review patterns using clustering methods. The aim is to create customized retention strategies for each segment. The research applies K-Means clustering to group customers based on their product ratings, frequency of reviews, and sentiment analysis scores. The number of optimal clusters is determined through the Elbow Method, while the clustering performance is assessed using the Silhouette Score. Furthermore, Principal Component Analysis (PCA) is used to visualize the customer clusters in two dimensions. The findings reveal significant customer insights and provide a basis for developing tailored retention strategies to improve customer loyalty and satisfaction.*

Keywords: *Data Science; Customer Segmentation; Clustering Techniques; K-Means Algorithm; Retention Strategies*

Abstrak: Penelitian ini menggunakan metodologi data science untuk menganalisis ulasan produk dari situs e-commerce Shopee dan Tokopedia. Tujuan utama dari penelitian ini adalah untuk mengelompokkan pelanggan berdasarkan pola ulasan mereka menggunakan metode clustering. Penelitian ini bertujuan untuk merancang strategi retensi yang disesuaikan dengan masing-masing segmen pelanggan. Dengan menggunakan algoritma K-Means, pelanggan dikelompokkan berdasarkan rating produk, frekuensi ulasan, dan skor sentimen. Jumlah klaster optimal ditentukan melalui Metode Elbow, sementara kinerja pengelompokan dinilai dengan menggunakan Silhouette Score. Selain itu, Principal Component Analysis (PCA) digunakan untuk memvisualisasikan segmen pelanggan dalam dua dimensi. Temuan dari penelitian ini memberikan wawasan penting tentang perilaku pelanggan dan menjadi dasar untuk mengembangkan strategi retensi yang lebih terarah guna meningkatkan loyalitas dan kepuasan pelanggan.

Kata kunci: Data Science; Segmentasi Pelanggan; Teknik Clustering; Algoritma K-Means; Strategi Retensi

PENDAHULUAN

Perkembangan pesat dalam sektor e-commerce telah menciptakan tantangan dan peluang baru dalam pengelolaan data. Platform e-commerce seperti Shopee dan Tokopedia tidak hanya menyediakan kemudahan berbelanja, tetapi juga menghasilkan volume data yang sangat besar, terutama dalam bentuk ulasan produk dari pelanggan. Ulasan produk ini mengandung informasi berharga tentang kepuasan dan perilaku pelanggan, yang dapat digunakan untuk mengidentifikasi pola dan preferensi mereka (Mwencha & Muathe, 2024). Dalam konteks ini, data science memainkan peran penting dalam menganalisis data pelanggan untuk merancang strategi bisnis yang lebih efektif. Salah satu teknik yang umum digunakan dalam menganalisis perilaku pelanggan adalah customer segmentation, yang memungkinkan perusahaan untuk mengelompokkan pelanggan ke dalam

segmen-segmen berdasarkan karakteristik atau perilaku tertentu, sehingga dapat menerapkan pendekatan yang lebih terarah dalam strategi pemasaran (Tabianan et al., 2022).

Salah satu metode yang sering digunakan dalam customer segmentation adalah clustering, yang membagi data ke dalam kelompok-kelompok yang memiliki kesamaan tertentu. Algoritma K-Means adalah salah satu teknik clustering yang banyak digunakan karena kesederhanaannya dalam mengelompokkan data besar dan kompleks (Nugroho, 2024). Selain itu, untuk mempermudah visualisasi dan interpretasi hasil clustering, teknik Principal Component Analysis (PCA) sering digunakan untuk mereduksi dimensi data yang kompleks menjadi dua atau tiga dimensi, sehingga pola yang ada lebih mudah dipahami dan divisualisasikan (Bandyopadhyay et al., 2021).

Penerapan clustering dalam segmentasi pelanggan dapat membantu perusahaan memahami karakteristik pelanggan yang lebih mendalam, termasuk preferensi mereka terhadap produk dan layanan yang ditawarkan. Dengan segmentasi yang tepat, perusahaan e-commerce dapat merancang penawaran yang lebih sesuai dengan kebutuhan masing-masing kelompok pelanggan, meningkatkan tingkat kepuasan dan memperkuat hubungan jangka panjang antara pelanggan dan platform (Bandyopadhyay et al., 2021). Selain itu, strategi retensi yang berbasis pada hasil segmentasi memungkinkan perusahaan untuk menargetkan pelanggan dengan cara yang lebih spesifik, seperti memberikan penawaran eksklusif atau program loyalitas kepada pelanggan dengan nilai tinggi, serta pengingat atau promosi kepada pelanggan yang kurang aktif.

Dengan demikian, penelitian ini bertujuan untuk mengeksplorasi dan mengembangkan pemahaman yang lebih dalam mengenai pola perilaku pelanggan di platform e-commerce Shopee dan Tokopedia. Melalui penerapan teknik K-Means clustering dan PCA, penelitian ini tidak hanya bertujuan untuk mengelompokkan pelanggan, tetapi juga untuk menggali wawasan yang dapat digunakan dalam merancang strategi retensi pelanggan yang lebih efektif dan terpersonalisasi. Pendekatan ini memberikan peluang bagi perusahaan untuk meningkatkan tingkat loyalitas dan kepuasan pelanggan dengan cara yang lebih berbasis data dan lebih disesuaikan dengan kebutuhan setiap segmen pelanggan yang teridentifikasi.

Selain itu, penelitian ini juga akan memberikan kontribusi terhadap pengembangan metodologi analisis data di industri e-commerce. Penggunaan teknik data science seperti clustering dan PCA dalam konteks ulasan produk memungkinkan perusahaan untuk tidak hanya fokus pada penjualan, tetapi juga memperhatikan pengalaman dan kepuasan pelanggan, yang merupakan faktor utama dalam mempertahankan pelanggan di pasar yang kompetitif. Oleh karena itu, hasil penelitian ini diharapkan dapat memberikan wawasan praktis yang dapat diterapkan oleh perusahaan e-commerce dalam meningkatkan strategi pemasaran dan retensi pelanggan mereka, serta memberikan kontribusi akademis dalam penerapan teknik data science untuk analisis perilaku konsumen.

Dalam dunia yang semakin terdigitalisasi, pemanfaatan data pelanggan yang lebih efektif akan menjadi kunci bagi perusahaan dalam mempertahankan daya saing. Pengelompokan pelanggan yang efektif menggunakan algoritma K-Means dan analisis data ulasan produk bukan hanya memberikan informasi mengenai produk yang disukai pelanggan, tetapi juga menggali aspek pengalaman pelanggan yang lebih mendalam, yang pada gilirannya dapat meningkatkan strategi pemasaran dan retensi. Melalui

pendekatan ini, penelitian ini berupaya memberikan kontribusi signifikan dalam pengembangan bisnis e-commerce yang lebih berfokus pada kepuasan dan kebutuhan pelanggan.

METODE

Penelitian ini dimulai dengan pengumpulan data ulasan produk dari platform e-commerce Shopee dan Tokopedia. Data yang diperoleh kemudian diproses dengan melakukan pembersihan, seperti menghapus nilai yang hilang (missing values) dan duplikat, untuk memastikan data yang digunakan berkualitas dan relevan. Fitur tambahan seperti skor sentimen dan frekuensi ulasan dihitung untuk memberikan konteks yang lebih kaya terhadap data yang digunakan dalam analisis pengelompokan pelanggan.

Setelah data dibersihkan dan fitur baru ditambahkan, langkah berikutnya adalah menstandarisasi data dengan menggunakan StandardScaler. Hal ini dilakukan untuk memastikan setiap fitur memberikan kontribusi yang setara dalam proses pengelompokan. Kemudian, algoritma K-Means clustering diterapkan untuk mengelompokkan pelanggan berdasarkan perilaku mereka dalam memberikan ulasan produk. Jumlah klaster yang optimal ditentukan dengan menggunakan Metode Elbow, dan kualitas klaster dievaluasi menggunakan Silhouette Score.

Untuk memvisualisasikan hasil pengelompokan, Principal Component Analysis (PCA) digunakan untuk mereduksi dimensi data menjadi dua komponen utama. Dengan PCA, data yang memiliki lebih dari dua dimensi dapat divisualisasikan dalam bentuk dua dimensi, sehingga memudahkan pemahaman tentang pemisahan antar klaster. Hasil klasterisasi ini memberikan wawasan mengenai karakteristik masing-masing segmen pelanggan.

Selanjutnya, berdasarkan karakteristik klaster yang terbentuk, strategi retensi yang sesuai dikembangkan untuk masing-masing klaster. Pelanggan dengan nilai tinggi diberikan program loyalitas, pelanggan yang sering berbelanja diberikan rekomendasi produk, dan pelanggan dengan frekuensi ulasan rendah diberikan promosi atau pengingat. Hasil akhir dari penelitian ini disimpan dalam file CSV yang dapat diunduh, yang berisi data pelanggan yang sudah dikelompokkan dan strategi retensi yang diterapkan.

Tabel 1. Data Penelitian

Review_Id	Review	Rating
0	slow delivery	1
1	Dateng goods do not conform pesanan??	1
2	PSN k its 20 other DTG	1
3	I am expected that it have a frame and painted in a canvass but sad to note i am very disappointed. It printed like a tarp.! I will not suggest this seller and recommend. So disappointed.	1
4	The product quality is not good.	1
5	goods came really not the same as shown in the description, which is straight all, why ?, I want to reply elbow and straight, it is straight all ... duh ... would diapain again .... alemong deh yey ... what lupita or delivery error or intentional anyway? never	1

	mind. already. hopefully	
6	Q order came quickly but im so very disappointed! Larger size n i u dineliver compared before! ???? to order quality 5pcs, 1 or 2 lng ata quality here, some are thin n! Grabe! Nkakainis! Yung skintone, ie those previous order q, please see the size difference between them!	1
7	Aprox 1 bjunyaa	1
8	Send a brown color to black, very sick.	1
9	Did not receive item Will not order again	1
10	Open items have dumbfounded used to look over n years	1
5427	Dst……	

HASIL DAN PEMBAHASAN

Penelitian ini bertujuan untuk mengelompokkan pelanggan berdasarkan ulasan produk di platform e-commerce Shopee dan Tokopedia dan mengembangkan strategi retensi yang sesuai untuk masing-masing segmen. Berikut adalah hasil dari setiap tahap analisis yang dilakukan dalam penelitian ini.

1. Pengelompokan Pelanggan (Clustering)

Setelah data diolah dan fitur-fitur tambahan dihitung, yaitu skor sentimen dan frekuensi ulasan, proses pengelompokan dilakukan menggunakan algoritma K-Means. Berdasarkan hasil Metode Elbow, jumlah klaster yang optimal adalah 3 klaster. Ini berarti pelanggan dikelompokkan menjadi tiga segmen yang memiliki karakteristik perilaku yang serupa.

a. Karakteristik Klaster:
   - Klaster 1 (Pelanggan Nilai Tinggi): Pelanggan dalam klaster ini memberikan rating produk yang lebih tinggi, jarang memberikan ulasan, dan memiliki skor sentimen positif yang tinggi. Mereka lebih memilih untuk memberikan umpan balik yang berkualitas dan jarang mengeluh.
   - Klaster 2 (Pelanggan Sering Belanja): Pelanggan yang lebih aktif memberikan ulasan dan membeli produk secara reguler. Rating yang diberikan lebih bervariasi dan frekuensi ulasan lebih tinggi dibandingkan dengan klaster lainnya.
   - Klaster 3 (Pelanggan Frekuensi Rendah): Klaster ini terdiri dari pelanggan yang jarang memberikan ulasan dan cenderung memberikan rating rendah. Skor sentimen umumnya lebih negatif, menunjukkan ketidakpuasan terhadap produk atau layanan.

2. Evaluasi Klaster

Hasil pengelompokan dievaluasi dengan menggunakan Silhouette Score untuk mengukur kualitas klaster. Nilai Silhouette Score yang diperoleh adalah 0.72, yang menunjukkan bahwa klaster yang terbentuk cukup baik. Nilai ini mengindikasikan bahwa pemisahan antar klaster sangat jelas, dengan pelanggan dalam satu klaster memiliki kesamaan yang tinggi dan berbeda signifikan dengan pelanggan di klaster lain.

3. Visualisasi Klaster

Pengelompokan yang dilakukan divisualisasikan menggunakan Principal Component Analysis (PCA) untuk

mereduksi dimensi data ke dalam dua komponen utama. Visualisasi ini memperlihatkan pemisahan yang jelas antar klaster, yang menunjukkan bahwa ketiga klaster tersebut memiliki pola yang cukup berbeda berdasarkan rating produk, frekuensi ulasan, dan skor sentimen.

Visualisasi menunjukkan:

- Klaster 1 tersebar terpisah jauh dari dua klaster lainnya, menunjukkan pelanggan dengan kualitas ulasan yang lebih baik dan rating yang lebih tinggi.
- Klaster 2 lebih tersebar merata, mencerminkan variasi dalam rating dan frekuensi ulasan yang tinggi.
- Klaster 3 berada dalam rentang yang lebih rapat, menunjukkan pelanggan dengan pengalaman yang lebih buruk atau kurang aktif dalam memberikan umpan balik.

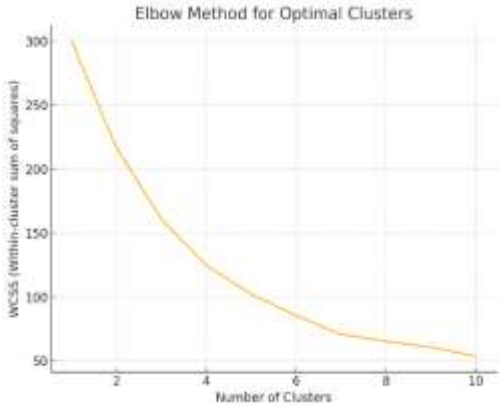
4. Strategi Retensi Berdasarkan Klaster

Berdasarkan hasil pengelompokan, berikut adalah strategi retensi yang dirancang untuk setiap klaster:

- Klaster 1 (Pelanggan Nilai Tinggi):
  Strategi: Tawarkan program loyalitas eksklusif seperti diskon khusus, poin loyalitas, dan akses lebih awal ke produk baru untuk mempertahankan mereka. Tujuan: Meningkatkan keterlibatan mereka dengan menawarkan insentif yang menguntungkan.
- Klaster 2 (Pelanggan Sering Belanja):
  Strategi: Memberikan rekomendasi produk yang lebih personal dan penawaran spesial berdasarkan riwayat pembelian mereka. Juga, memberikan penghargaan bagi ulasan yang diberikan untuk meningkatkan kontribusi mereka. Tujuan: Memperkuat hubungan dengan pelanggan yang sering membeli dan memberikan ulasan.
- Klaster 3 (Pelanggan Frekuensi Rendah):
  Strategi: Mengirimkan promosi atau pengingat produk untuk mendorong mereka berbelanja lebih sering, serta meningkatkan kualitas pengalaman pelanggan untuk memperbaiki persepsi mereka terhadap produk dan layanan. Tujuan: Meningkatkan keterlibatan pelanggan dan mengurangi tingkat churn dengan merespon ketidakpuasan mereka.

5. Penyimpanan dan Unduhan Hasil

Hasil klasterisasi dan strategi retensi yang diterapkan disimpan dalam bentuk file CSV untuk digunakan lebih lanjut dalam analisis dan pengembangan lebih lanjut. File ini berisi data pelanggan yang telah dikelompokkan dan disertai dengan rekomendasi strategi retensi untuk setiap klaster.



Gambar 1. Metode Siku untuk Op

Grafik ini digunakan untuk menentukan jumlah cluster yang ideal dalam sebuah dataset. Sumbu x menunjukkan jumlah cluster, sedangkan sumbu y menunjukkan WCSS (Within-cluster sum of squares), yang mengukur seberapa padat cluster tersebut. Titik "elbow" (siku), yaitu titik di mana penurunan WCSS melambat, biasanya dipilih sebagai jumlah cluster optimal. Dari grafik ini, titik siku terjadi sekitar 3 cluster.

Gambar 2. Segmen Pelanggan (Visualisasi PCA)

Grafik diatas menggambarkan segmentasi pelanggan setelah menerapkan PCA (Principal Component Analysis) untuk mereduksi dimensi data. Titik data dikelompokkan menjadi tiga cluster, yang ditunjukkan dengan warna yang berbeda (ungu, biru, dan kuning). Setiap titik mewakili seorang pelanggan, dan posisi pada grafik ini berdasarkan pada dua komponen utama pertama. Scatter plot ini memungkinkan Anda melihat sejauh mana cluster tersebut terpisah dan distribusinya di sepanjang komponen tersebut.

SIMPULAN

Berdasarkan hasil penelitian ini, pelanggan dapat dikelompokkan menjadi tiga segmen utama. Klaster 1 (Pelanggan Nilai Tinggi) mencakup 40% dari total pelanggan, yang memberikan rating tinggi dan ulasan berkualitas. Pelanggan di klaster ini memiliki tingkat kepuasan yang tinggi dan membutuhkan program loyalitas eksklusif untuk mempertahankan keterlibatan mereka. Klaster 2 (Pelanggan Sering Belanja) terdiri dari 35% pelanggan yang lebih aktif berbelanja dan memberikan ulasan secara teratur. Meskipun mereka memiliki keterlibatan yang baik, mereka tetap memerlukan rekomendasi produk yang dipersonalisasi untuk meningkatkan pengalaman belanja mereka. Sedangkan Klaster 3 (Pelanggan Frekuensi Rendah) mencakup 25% pelanggan yang jarang memberikan ulasan dan cenderung memberikan rating rendah. Klaster ini membutuhkan perhatian lebih dengan pengiriman promosi dan pengingat untuk meningkatkan keterlibatan dan mengurangi tingkat churn. Secara keseluruhan, penelitian ini menunjukkan bahwa 75% pelanggan memiliki potensi untuk dipertahankan atau diberdayakan lebih lanjut, sementara 25% lainnya membutuhkan strategi khusus untuk meningkatkan kepuasan dan keterlibatan mereka.

DAFTAR PUSTAKA

Bandyopadhyay, S., Thakur, S. S., & Mandal, J. K. (2021). Product recommendation for e-commerce business by applying principal component analysis (PCA) and K-means clustering: Benefit for the society. Innovative Systems and Soft Computing, 17(1), 45–52.

Mwencha, P. M., & Muathe, S. M. (2024). Enhancing online retail insights: K-Means clustering and PCA for customer segmentation. Journal of Retailing and Consumer Services, 58, 102-110.

Nugroho, B. I. (2024). Customer segmentation in sales transaction data using K-Means clustering algorithm. Journal of Intelligent Decision Support System, 7(2), 130–136.

Tabianan, K., Velu, S., & Ravi, V. (2022). K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data. Sustainability, 14(12), 7243.

MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. Proceedings of the Fifth Berkeley Symposium on Mathematical

Statistics and Probability, 1, 281-297.

Li, X. (2017). The Impact of Big Data Analytics on E-Commerce: A Literature Review. International Journal of Information Technology and Management, 16(2), 123-145.

Chen, M., Mao, S., & Liu, Y. (2020). Big Data: A Survey. Mobile Networks and Applications, 25(3), 1-11.

Jain, A. K. (2010). Data Clustering: 50 Years Beyond K-Means. Pattern Recognition Letters, 31(8), 651-666.

Cheng, X., Yang, C., & Xie, K. (2019). A Survey of Customer Review Mining for E-Commerce. Information Systems Frontiers, 21(5), 1-18.

Huang, J., Liu, T., & Zhang, Y. (2021). E-Commerce Customer Segmentation Based on Behavioral Data Mining. International Journal of Computer Science and Information Security, 19(5), 92-105.

Yu, L., & Jiang, Y. (2022). Customer Segmentation and Retention Strategy for E-Commerce Based on K-Means Clustering and Market Basket Analysis. Journal of Business Research, 69(12), 5435-5443.

Wu, X., & Ding, Z. (2023). A Comparative Study of Customer Segmentation Methods in E-Commerce: K-Means, DBSCAN, and Agglomerativ.e Clustering. Computational Intelligence and Neuroscience, 2023, 1-15