**CLUSTERING PERILAKU PENGGUNA WEBSITE BERDASARKAN AKTIVITAS BROWSING MENGGUNAKAN ALGORITMA K-MEANS**

**Reyza Abdillah 1\*, Rangga Syaputra 2, Elkin Rilvani 3.**

1,2,3,4,5Program Studi Teknik Informatika, Fakultas Teknik, Universitas Pelita Bangsa, Cikarang.

E-mail: [getmorelicks@gmail.com](mailto:getmorelicks@gmail.com), [Ranggasyaputra215@gmail.com](mailto:Ranggasyaputra215@gmail.com), elkin.rilvani@pelitabangsa.ac.id

|  |  |
| --- | --- |
| **ABSTRACT**  *This study aims to analyze website visitor behavior by applying the K-Means algorithm for clustering. The research focuses on user segmentation based on their browsing activities, such as the number of pages visited, time spent on the website, and bounce rate. The methodology employs an unsupervised learning approach, where K-Means is used to group users into three clusters: active visitors, regular visitors, and quick-exit (bounce) visitors. Data preprocessing, including missing value imputation and normalization, was conducted to ensure dataset quality and consistency. The results demonstrate that K-Means with K=3 provides an optimal clustering solution, achieving a Silhouette Score of 0.627. These findings highlight distinct user behaviors that can aid in improving website optimization and targeted marketing strategies. The study concludes that unsupervised learning techniques like K-Means can effectively segment website visitors without requiring labeled data, offering valuable insights into user engagement and behavioral patterns.*  ***Keywords: Website User Behavior, Clustering, K-Means, Visitor Segmentation, Unsupervised Learning.*** ABSTRAK Penelitian ini bertujuan untuk menganalisis perilaku pengunjung website dengan menerapkan algoritma K-Means untuk clustering. Penelitian ini fokus pada segmentasi pengguna berdasarkan aktivitas browsing mereka, seperti jumlah halaman yang dikunjungi, waktu yang dihabiskan di website, dan tingkat bounce. Metodologi yang digunakan melibatkan pendekatan **unsupervised learning**, di mana K-Means diterapkan untuk mengelompokkan pengguna menjadi tiga cluster: pengunjung aktif, pengunjung biasa, dan pengunjung yang cepat keluar (bounce). Pra-pemrosesan data, termasuk imputasi nilai yang hilang dan normalisasi, dilakukan untuk memastikan kualitas dan konsistensi dataset. Hasil penelitian menunjukkan bahwa K-Means dengan **K=3** memberikan solusi clustering yang optimal, dengan **Silhouette Score** sebesar 0.627. Temuan ini menyoroti perilaku pengguna yang berbeda yang dapat membantu dalam meningkatkan optimasi website dan strategi pemasaran yang lebih terarah. Penelitian ini menyimpulkan bahwa teknik unsupervised learning seperti K-Means dapat secara efektif melakukan segmentasi pengunjung website tanpa memerlukan data berlabel, memberikan wawasan berharga tentang keterlibatan dan pola perilaku pengguna.  ***Kata Kunci: Perilaku Pengguna Website, Clustering, K-Means, Segmentasi Pengunjung, Unsupervised Learning.*** | **Article History**  Received: Agustus 2025  Reviewed: Agustus 2025  Published: Agustus 2025  Plagiarism Checker No 235  Prefix DOI : 10.8734/Kohesi.v1i2.365  **Copyright : Author**  **Publish by : Kohesi**    Creative Commons License  This work is licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/) |

# PENDAHULUAN

Perilaku pengguna website memiliki peran yang sangat penting dalam meningkatkan pengalaman pengguna serta konversi pada situs e-commerce. Dengan berkembangnya teknologi dan semakin kompleksnya perilaku pengguna, penting bagi pengelola website untuk memahami bagaimana pengunjung berinteraksi dengan situs yang mereka kunjungi. Pengukuran aktivitas pengguna, seperti jumlah halaman yang dikunjungi, lama waktu kunjungan, serta tingkat bounce rate, memberikan gambaran yang lebih jelas mengenai pola interaksi dan kepuasan pengunjung terhadap website tersebut. Pemahaman yang baik mengenai perilaku ini memungkinkan pengelola situs untuk meningkatkan pengalaman pengguna dan, pada akhirnya, meningkatkan hasil konversi yang diinginkan.

Analisis perilaku pengguna seringkali dilakukan dengan teknik clustering, di mana pengunjung website dikelompokkan berdasarkan kesamaan perilaku mereka. Salah satu algoritma yang paling banyak digunakan dalam clustering adalah K-Means, yang mampu mengelompokkan data secara otomatis ke dalam beberapa cluster berdasarkan kedekatan antar data. Algoritma ini memiliki kelebihan dalam hal kecepatan dan efisiensi dalam mengolah data dalam jumlah besar. Penggunaan K-Means dalam analisis perilaku pengguna memberikan wawasan mengenai segmentasi pengguna yang berbeda, seperti pengunjung yang sangat aktif, pengunjung biasa, dan pengunjung yang cenderung keluar cepat (bounce).

Berbagai penelitian telah dilakukan untuk mengaplikasikan algoritma K-Means dalam analisis perilaku pengguna. Chhabra et al. [1] menggunakan K-Means dalam konteks imputasi nilai yang hilang pada data pengguna, dengan menggabungkan K-Means dan aturan asosiasi untuk memperbaiki masalah ketidaklengkapan data dalam dataset besar. Hal serupa juga dilakukan oleh Kumar et al. [2], yang membandingkan K-Means dengan Hierarchical Clustering dalam segmentasi pelanggan e-commerce, dimana mereka menemukan bahwa K-Means lebih efisien dalam mengelompokkan pelanggan berdasarkan data perilaku seperti waktu kunjungan dan interaksi dengan produk. Li et al. [3] memfokuskan penelitian mereka pada bagaimana analisis perilaku pengguna dapat meningkatkan rekomendasi produk pada e-commerce dengan bantuan K-Means, yang secara efektif membagi pengguna ke dalam kelompok berdasarkan interaksi mereka dengan produk.

Necula [4] melakukan penelitian tentang pengaruh waktu yang dihabiskan pengguna untuk membaca informasi produk, yang memiliki dampak besar pada keputusan pembelian. Kembali ke analisis perilaku, banyak penelitian sebelumnya mengandalkan teknik supervised learning, yang mengharuskan adanya label dalam dataset, seperti pembelian atau tidak. Pendekatan ini membatasi fleksibilitas analisis pada data tanpa label. Sebagai solusi, penelitian ini mengusulkan penggunaan algoritma unsupervised learning, yaitu K-Means, yang mampu mengelompokkan pengguna tanpa memerlukan label yang sudah ada sebelumnya, serta lebih mudah diterapkan pada dataset yang besar dan tidak terstruktur.

Penelitian lain juga menunjukkan pentingnya normalisasi dalam meningkatkan akurasi algoritma clustering. Virmani et al. [5] meneliti teknik normalization dalam K-Means untuk memastikan bahwa setiap fitur dalam dataset memiliki pengaruh yang setara dalam pembentukan cluster, menghindari dominasi oleh fitur yang memiliki skala lebih besar. Penelitian ini mengusulkan penggunaan metode normalization-based K-Means untuk meningkatkan kinerja clustering pada dataset dengan berbagai skala fitur.

Sebagian besar penelitian tentang perilaku pengguna menggunakan teknik supervised learning, yang mengandalkan label data seperti status pembelian untuk mengklasifikasikan pengguna. Pendekatan ini mengharuskan adanya data berlabel yang seringkali tidak selalu tersedia pada data perilaku pengguna. Oleh karena itu, penting untuk menggunakan pendekatan unsupervised learning seperti K-Means, yang memungkinkan segmentasi pengguna tanpa memerlukan label yang sudah ada. Penelitian ini mengisi celah tersebut dengan mengaplikasikan K-Means pada data perilaku pengguna, yang memungkinkan segmentasi pengunjung berdasarkan interaksi mereka dengan situs web, meskipun tanpa informasi pembelian atau data berlabel lainnya.

Penelitian ini bertujuan untuk menganalisis perilaku pengunjung website menggunakan algoritma K-Means. Fokus utama penelitian ini adalah untuk mengidentifikasi segmentasi pengunjung berdasarkan jumlah halaman yang dikunjungi, lama kunjungan, dan bounce rate. Segmentasi ini akan memberikan wawasan lebih dalam mengenai jenis-jenis pengunjung dan perilaku mereka selama mengakses website, yang nantinya dapat digunakan untuk meningkatkan strategi desain dan pemasaran pada website.

Penelitian sebelumnya banyak yang menggunakan teknik supervised learning, yang memerlukan label dalam data untuk melakukan analisis. Pendekatan ini membatasi fleksibilitas aplikasi algoritma pada data yang tidak memiliki label atau informasi yang sudah diketahui. Oleh karena itu, penelitian ini menggunakan pendekatan unsupervised learning, yaitu K-Means, untuk mengklasifikasikan pengunjung website tanpa menggunakan label yang sudah ada, memberikan cara yang lebih alami dan efisien untuk mengidentifikasi pola perilaku pengunjung. Dengan mengaplikasikan K-Means pada data yang tidak berlabel, penelitian ini membuka peluang untuk segmentasi pengguna yang lebih luas dan lebih mudah diterapkan pada berbagai jenis data pengguna yang ada di website.

1. **METODE PENELITIAN**

Penelitian ini menggunakan pendekatan unsupervised learning untuk menganalisis perilaku pengunjung website dengan algoritma K-Means. Algoritma K-Means dipilih karena kemampuannya dalam mengelompokkan data yang tidak berlabel (unlabeled data) dan efisiensinya dalam menangani dataset besar. Proses penelitian dimulai dengan pengumpulan data dari dataset Online Shoppers Intention, dilanjutkan dengan pra-pemrosesan data, penerapan algoritma K-Means, serta evaluasi hasil clustering yang terbentuk. Berikut adalah langkah-langkah yang diambil untuk menyelesaikan penelitian ini.

1. **Pengumpulan Data**

Dataset yang digunakan dalam penelitian ini adalah **Online Shoppers Intention** yang berisi data mengenai perilaku pengguna website, termasuk jumlah halaman yang dikunjungi (PageValues), lama kunjungan (ExitRates, BounceRates), dan waktu yang dihabiskan pada halaman produk (ProductRelated\_Duration). Dataset ini diunduh melalui Kaggle dan digunakan untuk analisis perilaku pengunjung website.

1. **Pra-Pemrosesan Data**

Langkah pertama dalam pemrosesan data adalah penanganan nilai yang hilang. Dataset yang digunakan memiliki beberapa nilai yang kosong, yang dapat mempengaruhi hasil analisis. Oleh karena itu, nilai yang hilang digantikan dengan nilai rata-rata dari masing-masing kolom menggunakan metode imputasi. Setelah itu, dilakukan normalisasi data menggunakan StandardScaler dari pustaka scikit-learn. Proses normalisasi ini penting karena algoritma K-Means sensitif terhadap skala data. Tanpa normalisasi, fitur dengan skala lebih besar (seperti PageValues) bisa mendominasi proses clustering.

1. **Penerapan Algoritma K-Means**

K-Means digunakan untuk mengelompokkan pengguna website berdasarkan aktivitas mereka selama mengunjungi situs. Algoritma ini memulai dengan memilih K centroid acak, yang kemudian digunakan untuk mengelompokkan data ke dalam cluster berdasarkan kedekatannya dengan centroid tersebut. Setelah setiap data dikelompokkan, posisi centroid dihitung ulang berdasarkan rata-rata posisi data dalam cluster tersebut. Proses ini diulang hingga posisi centroid stabil (konvergen). Dalam penelitian ini, K=3 dipilih sebagai jumlah cluster yang optimal, dengan tujuan mengelompokkan pengguna menjadi tiga kategori: pengunjung aktif, pengunjung biasa, dan pengunjung bounce.

1. **Evaluasi Hasil Clustering**

Untuk mengevaluasi kualitas clustering, digunakan Silhouette Score, yang mengukur seberapa baik setiap data dikelompokkan dalam cluster yang benar. Nilai Silhouette Score berkisar antara -1 dan 1, dengan nilai yang lebih tinggi menunjukkan pemisahan cluster yang lebih baik. Selain itu, Principal Component Analysis (PCA) digunakan untuk mereduksi dimensi data dari empat fitur menjadi dua dimensi, yang memungkinkan hasil clustering divisualisasikan dalam bentuk scatter plot. Visualisasi ini memberikan gambaran yang jelas tentang seberapa baik cluster terpisah satu sama lain.

* 1. **Research Methods**

Penelitian ini menggunakan metode **unsupervised learning** dengan algoritma **K-Means** untuk clustering. K-Means dipilih karena kemampuannya dalam mengelompokkan data yang tidak berlabel, serta efisiensinya dalam memproses data besar yang sering kali ditemukan dalam analisis perilaku pengguna. Penelitian sebelumnya oleh Virmani et al. [5] dan Kumar et al. [2] menunjukkan bahwa K-Means adalah salah satu algoritma yang efektif untuk clustering, terutama dalam analisis data perilaku pengguna. Dalam penelitian ini, kami mengadaptasi metode K-Means yang telah terbukti efektif pada data berlabel dan mengaplikasikannya pada data tanpa label untuk mengelompokkan pengunjung berdasarkan aktivitas mereka.

K-Means beroperasi dengan menginisialisasi beberapa centroid secara acak, yang kemudian digunakan untuk mengelompokkan data berdasarkan kedekatannya dengan centroid tersebut. Setiap iterasi algoritma menghitung ulang posisi centroid berdasarkan rata-rata posisi data dalam cluster tersebut. Dalam penelitian ini, kami menguji dengan **K=3** cluster, karena tiga kategori pengunjung (aktif, biasa, bounce) cukup representatif untuk tujuan analisis. Proses ini diulang hingga algoritma mencapai konvergensi, yaitu saat posisi centroid tidak lagi berubah secara signifikan.

Fungsi objektif dalam algoritma K-Means adalah untuk meminimalkan jarak antar data dalam satu cluster terhadap centroid cluster tersebut. Fungsi objektif ini dapat dituliskan dengan rumus matematis sebagai berikut:

(1)

Dimana:

* J adalah fungsi objektif yang ingin diminimalkan.
* K adalah jumlah cluster.
* ​ adalah cluster ke-i.
* ​ adalah data yang berada dalam cluster ​.
* ​ adalah centroid dari cluster ​.
* adalah jarak Euclidean antara data dan centroid ​​.

Dalam proses K-Means, algoritma ini terus mengupdate posisi centroid dengan tujuan meminimalkan jarak antar data dalam cluster.

Metode yang digunakan dalam penelitian ini mengacu pada teknik yang telah diterapkan dalam penelitian sebelumnya, seperti yang dijelaskan oleh Kumar et al. [2] dan Virmani et al. [5]. Dengan mengadaptasi K-Means pada data perilaku pengguna yang tidak berlabel, penelitian ini memberikan pendekatan baru dalam mengelompokkan pengguna berdasarkan aktivitas browsing mereka. Pendekatan ini lebih fleksibel dan tidak memerlukan label yang biasanya sulit didapatkan dalam banyak kasus e-commerce atau situs web yang tidak memiliki data pembelian eksplisit.

1. **HASIL DAN PEMBAHASAN**

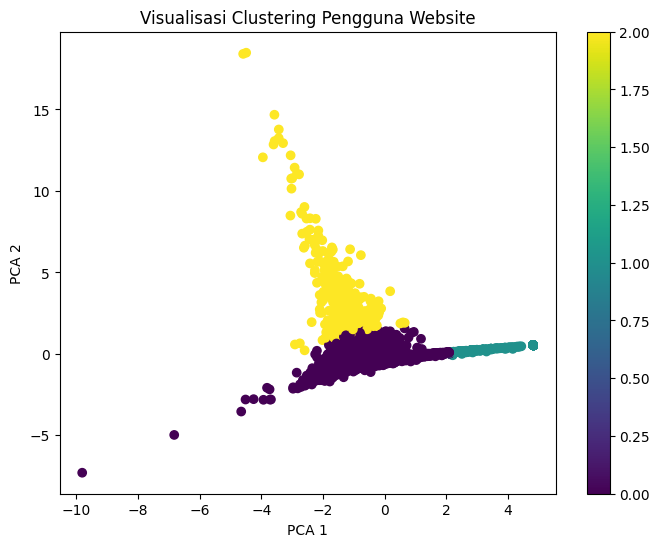
Penelitian ini bertujuan untuk mengelompokkan pengunjung website berdasarkan aktivitas browsing mereka menggunakan algoritma K-Means. Hasil clustering akan membantu dalam memahami perilaku pengguna yang berbeda, yang pada akhirnya dapat digunakan untuk meningkatkan pengalaman pengguna dan optimasi strategi pemasaran. Dalam bagian ini, hasil clustering yang diperoleh akan dibahas secara mendalam, termasuk evaluasi kualitas clustering dan analisis hasil segmentasi pengunjung.

Setelah menerapkan algoritma K-Means dengan K=3 cluster, hasil clustering menunjukkan adanya pemisahan pengunjung menjadi tiga kelompok yang berbeda berdasarkan perilaku mereka di website. Pengunjung dalam masing-masing cluster memiliki karakteristik yang cukup berbeda satu sama lain, terutama dalam hal jumlah halaman yang dikunjungi, lama kunjungan, dan bounce rate. Untuk mengevaluasi seberapa baik pemisahan antara cluster, dihitung Silhouette Score, yang merupakan ukuran dari seberapa baik data dalam satu cluster terpisah dari cluster lainnya.

Silhouette Score yang dihitung untuk K=3 adalah 0.627. Nilai ini menunjukkan bahwa clustering yang terbentuk cukup baik, karena nilai Silhouette Score yang lebih tinggi mendekati 1 menandakan bahwa cluster yang terbentuk terpisah dengan jelas dan data dalam setiap cluster lebih homogen. Dalam konteks penelitian ini, nilai 0.627 mengindikasikan bahwa hasil clustering memiliki pemisahan yang cukup baik antar cluster.

Untuk memvisualisasikan hasil clustering dan untuk memberikan gambaran yang lebih jelas tentang bagaimana cluster terdistribusi, Principal Component Analysis (PCA) digunakan untuk mereduksi dimensi data dari empat fitur (PageValues, ExitRates, BounceRates, ProductRelated\_Duration) menjadi dua dimensi. Visualisasi ini memungkinkan kita untuk melihat pemisahan antar cluster dalam ruang dua dimensi, yang seringkali lebih mudah untuk dianalisis dan dipahami.

Gambar 1 di bawah ini menunjukkan hasil **PCA** yang memvisualisasikan hasil clustering pengunjung website:



**Gambar 1. Visualisasi Clustering Pengguna Website dengan PCA**

Pada **Gambar 1**, kita dapat melihat tiga cluster yang terpisah dengan jelas berdasarkan warna yang berbeda. Setiap warna mewakili satu cluster, dengan cluster pertama (warna ungu) menunjukkan pengunjung yang lebih sedikit berinteraksi dengan situs, cluster kedua (warna kuning) menunjukkan pengunjung yang lebih aktif, dan cluster ketiga (warna cyan) mencerminkan pengunjung yang keluar cepat (bounce). Visualisasi ini memberikan indikasi yang jelas bahwa pengunjung dengan perilaku serupa dikelompokkan dengan baik oleh algoritma K-Means. Setelah dilakukan clustering, hasil analisis rata-rata fitur untuk masing-masing cluster menunjukkan bahwa ketiga cluster memiliki karakteristik perilaku yang berbeda. Berikut adalah rata-rata nilai untuk setiap fitur dalam masing-masing cluster:

**Tabel 1. Rata-Rata Fitur untuk Setiap Cluster**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cluster** | **PageValues** | **ExitRates** | **BounceRates** | **ProductRelated\_Duration** |
| 0 | 2.387881 | 0.032459 | 0.009807 | 1285.841440 |
| 1 | 0.000000 | 0.186944 | 0.179136 | 47.201124 |
| 2 | 64.915559 | 0.014617 | 0.002990 | 1327.122523 |

Dari Tabel 1, dapat dilihat bahwa Cluster 0 menunjukkan pengunjung yang memiliki PageValues rendah, ExitRates rendah, dan ProductRelated\_Duration panjang. Ini menunjukkan bahwa pengunjung ini menghabiskan waktu yang cukup lama di situs, mungkin untuk melihat lebih banyak produk, tetapi tidak banyak berinteraksi lebih jauh (mungkin hanya melihat produk tanpa melakukan pembelian atau tindakan lain).

Cluster 1 menunjukkan pengunjung dengan PageValues sangat rendah, ExitRates tinggi, dan BounceRates tinggi. Ini menunjukkan bahwa pengunjung di cluster ini cenderung meninggalkan situs setelah melihat sedikit halaman atau hanya melakukan kunjungan singkat tanpa banyak berinteraksi. Mereka mungkin adalah pengunjung yang "bounce" dan tidak menemukan konten yang menarik untuk mereka.

Cluster 2 menunjukkan pengunjung dengan PageValues sangat tinggi dan ProductRelated\_Duration panjang, yang menunjukkan bahwa mereka lebih banyak menghabiskan waktu di situs dan melihat lebih banyak produk. Pengunjung di cluster ini kemungkinan besar lebih tertarik pada produk yang ditampilkan dan berinteraksi lebih lama, yang membuat mereka lebih berpotensi menjadi pelanggan atau pengguna yang lebih setia. Sebagai langkah evaluasi tambahan, penelitian ini juga menguji hasil clustering dengan jumlah cluster yang berbeda, yaitu K=4 dan K=5. Hasil Silhouette Score untuk K=4 dan K=5 adalah sebagai berikut:

1. **Silhouette Score untuk K=4**: 0.6067
2. **Silhouette Score untuk K=5**: 0.4657

Dari hasil ini, terlihat bahwa K=3 menghasilkan Silhouette Score tertinggi, yang menunjukkan bahwa K=3 adalah pilihan yang optimal untuk jumlah cluster pada dataset ini. Dengan K=4 dan K=5, kualitas clustering menurun, yang berarti penambahan jumlah cluster tidak memberikan pemisahan yang lebih baik antar kelompok pengunjung. Oleh karena itu, jumlah cluster yang terbaik untuk data ini adalah K=3, yang memberikan hasil yang lebih terstruktur dan dapat diinterpretasikan dengan baik.

Berdasarkan hasil yang diperoleh, dapat disimpulkan bahwa algoritma K-Means berhasil mengelompokkan pengunjung website menjadi tiga cluster yang berbeda berdasarkan perilaku mereka. Pengelompokan ini memberikan wawasan yang berharga tentang jenis-jenis pengunjung yang ada di situs, yang dapat digunakan untuk meningkatkan pengalaman pengguna dan mengoptimalkan strategi pemasaran. Pengunjung yang lebih aktif (Cluster 2) dapat diberi perhatian lebih dalam hal promosi produk, sementara pengunjung yang lebih sedikit berinteraksi (Cluster 1) mungkin memerlukan penyesuaian dalam desain atau strategi untuk mengurangi tingkat bounce.

1. **KESIMPULAN**

Penelitian ini berhasil mengidentifikasi segmentasi pengunjung website berdasarkan perilaku browsing mereka menggunakan algoritma K-Means. Dengan menggunakan dataset Online Shoppers Intention, algoritma ini berhasil mengelompokkan pengunjung ke dalam tiga cluster yang berbeda, yaitu pengunjung aktif, pengunjung biasa, dan pengunjung yang cepat keluar (bounce). Hasil evaluasi menggunakan Silhouette Score menunjukkan bahwa K-Means dengan K=3 adalah jumlah cluster yang optimal, memberikan pemisahan yang jelas antar kelompok pengunjung dengan skor 0.627. Visualisasi hasil clustering dengan menggunakan Principal Component Analysis (PCA) juga memperlihatkan pemisahan yang baik antara ketiga cluster. Berdasarkan analisis, ditemukan bahwa pengunjung di cluster pertama memiliki sedikit interaksi dengan situs, sementara pengunjung di cluster kedua lebih aktif dan menghabiskan lebih banyak waktu di halaman produk. Cluster ketiga menunjukkan pengunjung dengan interaksi moderat. Dengan demikian, penelitian ini menyarankan bahwa untuk meningkatkan pengalaman pengguna, pengelola situs dapat menyesuaikan desain dan strategi pemasaran berdasarkan segmentasi perilaku pengguna yang telah diidentifikasi. Pendekatan unsupervised learning yang digunakan dalam penelitian ini juga menunjukkan potensi yang besar dalam segmentasi pengguna tanpa memerlukan data berlabel, membuka peluang baru untuk analisis perilaku pengguna di berbagai jenis website.

**DAFTAR PUSTAKA**

[1] G. Chhabra, V. Vashisht, and J. Ranjan, "Missing value imputation using hybrid k-means and association rules," in *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, 2018, pp. 501-509, IEEE.

[2] S. Kumar, R. Rani, S. K. Pippal, and R. Agrawal, "Customer segmentation in e-commerce: K-means vs hierarchical clustering," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 23, no. 1, pp. 119-128, 2025.

[3] Y. Li, H. Sun, W. Chen, X. Zhang, J. Wang, and L. Zhao, "The Role of User Behavior Analytics in Enhancing Personalized E-commerce Recommendations," *The Role of User Behavior Analytics in Enhancing Personalized E-commerce Recommendations*, January 2025, DOI:10.13140/RG.2.2.15061.95207.

[4] S. C. Necula, "Exploring the impact of time spent reading product information on e-commerce websites: A machine learning approach to analyze consumer behavior," *Behavioral Sciences*, vol. 13, no. 6, pp. 439, 2023.

[5] D. Virmani, S. Taneja, and G. Malhotra, "Normalization based k means clustering algorithm," *arXiv preprint arXiv:1503.00900*, 2015.