 ISSN 2797-7889

Implementation of Data Mining for Mapping Student Visiting Interest Based on Study Programs as an Effort to Optimize Library Services at STMIK Methodist Binjai

Anzas Ibezato Zalukhu[1], Adil Priman Hati Hulu[2]
[1]Information Systems, STMIK Methodist Binjai, Binjai, Indonesia, [2]Informatics Engineering, STMIK Methodist Binjai, Binjai, Indonesia
Email: anzaszalukhu@stmikmethodistbinjai.ac.id[1], adilprimanhatihulu@gmail.com[2]

This study successfully implemented the K-Means Clustering algorithm to map students' library visit interest patterns at the STMIK Methodist Binjai Library. The clustering results were empirically validated using a Silhouette Coefficient score of 0.8304, with the optimal number of clusters determined as k = 3 through the Elbow Method. The clustering process identified three strategic profiles: a Low-Interest Cluster consisting of 117 students with an average of 1.50 visits, a Moderate-Interest Cluster comprising 19 students with an average of 9.32 visits, and a High-Interest Cluster including 8 students with an average of 21.00 visits. The data analysis revealed significant disparities in visit interest across academic programs, where students from the Informatics Engineering program demonstrated higher levels of engagement compared to those from the Information Systems program, which was predominantly characterized by low visit frequency. These findings provide a scientific foundation for library management to formulate segmented service optimization policies, including retention programs for active users and personalized literacy stimulation strategies to enhance student engagement in academic programs with lower visit intensity.
Keywords: Data Mining, K-Means Clustering, Visit Interest, Library, Service Optimization

Corresponding Author:
Anzas Ibezato Zalukhu
STMIK Methodist Binjai
Jl. Jenderal Gatot Subroto, Bandar Senembah, Kec. Binjai Bar., Kota Binjai, Sumatera Utara 20716
anzaszalukhu@stmikmethodistbinjai.ac.id

1. Introduction

University libraries serve as central information resource hubs that play a vital role in supporting the implementation of the Tri Dharma of Higher Education. In the era of digital transformation, libraries are required to provide adaptive, technology-based information services [1], [2]. Efforts to improve library services at STMIK Methodist Binjai are aligned with Law Number 43 of 2007 on Libraries, which emphasizes the function of libraries as facilities for education, research, and information dissemination aimed at enhancing the quality of human resources [3].

One of the key indicators of successful library management is a high level of student visits and the utilization of library collections [4]. However, in practice, library management often faces challenges in accurately mapping visitor characteristics [5], [6]. Visitor log data collected daily through the digital attendance system tend to accumulate merely as administrative archives without further processing [7]. As a result, literacy promotion policies and collection acquisition strategies are often generic and insufficiently targeted, as they are not grounded in user behavior data (data-driven decision making).

Information gaps regarding visit interest across academic programs have become a critical issue. Differences in curricula and academic workload among programs, such as Information Systems and Informatics Engineering, are predicted to result in distinct library visitation behavior patterns. Without

 ISSN 2797-7889

systematic mapping, library administrators face difficulties in identifying academic programs with low engagement levels, thereby limiting the development of specific and effective intervention strategies [8].

Data mining is a systematic process for extracting meaningful patterns, relationships, and knowledge from large datasets by employing statistical techniques, artificial intelligence, and machine learning approaches [9], [10], [11]. Data mining provides a solution through clustering techniques to uncover hidden knowledge from visitor log data [12]. The K-Means Clustering algorithm is an unsupervised learning method that is highly effective in grouping data based on similarity characteristics without requiring prior labels. By applying this algorithm, students can be categorized into several engagement segments (such as highly active, active, and infrequent visitors) based on their academic program affiliation. [11], [13], [14].

This study aims to implement data mining using the K-Means algorithm to map student visit interest patterns at the STMIK Methodist Binjai Library. The resulting mapping is expected to make a tangible contribution to library management in optimizing services, including adjusting service hours, expanding program-specific reference collections, and designing more effective literacy ambassador programs. Ultimately, the STMIK Methodist Binjai Library is expected to transform into a more responsive and high-quality academic support unit.

2. Literature Review and Problem Statement

Several previous studies have demonstrated the effectiveness of data mining techniques in the context of library management. A prior study successfully mapped the condition of libraries in Indonesia in 2023 into six clusters using the K-Means Clustering algorithm [15]. Another study implemented the K-Means algorithm to classify visitors and borrowers at the ITN Malang Library into three frequency-based categories, namely very frequent, frequent, and infrequent users [16]. In addition, research conducted at the Faculty of Engineering Library of Universitas Negeri Gorontalo applied the K-Means Clustering method to classify students' reading interests based on borrowing frequency, resulting in strategic recommendations for targeted collection development [17].

However, a significant challenge faced by the STMIK Methodist Binjai Library is the high volume of daily visitor log data that has not yet been optimally managed for policy-making purposes. At present, visit-level analysis is conducted only in a general manner without considering academic program affiliation, causing literacy development programs to often fail to target the student groups most in need. There are indications of participation disparities across academic programs which, if left unaddressed, may hinder the achievement of library service quality standards. Therefore, a data-driven mapping approach is required to precisely identify clusters of student engagement.

In line with the proven effectiveness of data mining techniques in mapping national library conditions and optimizing services across various institutions, this study applies the K-Means Clustering algorithm to classify student engagement into low, moderate, and high categories. Consistent with these findings, the same algorithm is employed to categorize the engagement levels of students at STMIK Methodist Binjai into low, moderate, and high groups based on academic program affiliation. The primary focus of this research is to extract visitor log data into a scientifically valid distribution map of visit interest. The results are expected to integrate administrative data with the strategic needs of library management in order to optimize services and segment the acquisition of reference collections.

 ISSN 2797-7889

3. Method

This study adopts the international standard Cross-Industry Standard Process for Data Mining (CRISP-DM). This methodology was selected due to its highly structured framework for transforming visitor log data into strategic information.

Research Stages

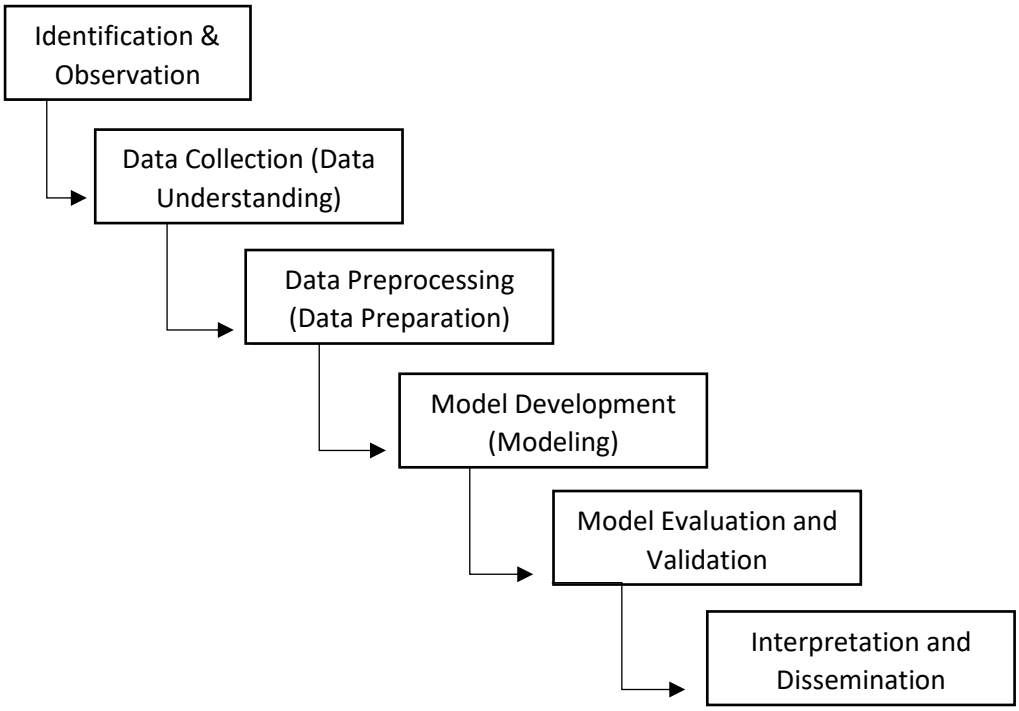

Figure 1. Research Stages

1. Identification & Observation (Business Understanding)
   The initiation stage was carried out by identifying managerial problems within the STMIK Methodist Binjai Library unit. The primary focus of this phase was to formulate strategic needs for service optimization through mapping student visitation behavior patterns, which have previously been heterogeneous and unstructured.
2. Data Collection (Data Understanding)
   Primary data were extracted from the library visitor log repository database and collected over the period from October 2025 to January 2026. An initial exploratory process was conducted to understand the metadata structure, which included student identity (Student ID), academic affiliation (Study Program), and temporal variables in the form of visit timestamps. This stage aimed to identify the initial data distribution and detect potential anomalies within the dataset.
3. Data Pre-processing (Data Preparation)
   The data pre-processing stage at the STMIK Methodist Binjai Library was conducted systematically through data cleaning to eliminate redundant attendance logs and feature engineering to transform transactional data into student visit frequency variables by study program. To ensure the accuracy of visit interest mapping, Z-Score Standardization was implemented to normalize data scales and minimize distance calculation bias in the K-Means algorithm. Through this standardization process, the clustering of study programs into low, moderate, and high visit interest categories became more objective, thereby providing a valid foundation for management in formulating targeted library service optimization strategies

 ISSN 2797-7889

4. Model Development (Modeling)
   Modeling was implemented by applying the K-Means Clustering algorithm, an unsupervised learning method that partitions objects into k clusters based on similarities in student visit frequency characteristics. An iterative process was performed to minimize the Sum of Squared Error (SSE) by calculating the Euclidean distance between data points and their respective cluster centroids, resulting in an accurate mapping of student visit interest based on study programs. Through this mathematical approach, the STMIK Methodist Binjai Library can objectively identify visitor behavior segmentation, which serves as a scientific basis for formulating more personalized and targeted library service optimization strategies for each academic program. The K-Means Clustering formula is presented as follows:

$$d(x,c) = \sqrt{\sum_{i=1}^{n}(x_i - C_i)^2} \qquad (1)$$

   Where:
   d(x,c) = the distance between data point x and centroid c.
   n = the number of dimensions or features
   x_1 = the value of the data point in the i-th dimension.
   c_1 = the value of the centroid in the i-th dimension.

5. Model Evaluation and Validation (Evaluation)
   At the evaluation stage, the clustering results were validated using the Silhouette Coefficient to assess the strength of data cohesion within each cluster, while also ensuring that the mapping of student visit interest based on academic programs was optimally structured. Subsequently, the results were interpreted to assign labels to each cluster, such as low, moderate, and high categories, to facilitate the identification of visitor behavior characteristics at the STMIK Methodist Binjai Library. This validation and labeling process represents a crucial step in the data mining implementation, as it provides accurate and measurable information that serves as a foundation for management in formulating more strategic library service optimization programs tailored to the needs of each academic program.
6. Interpretation and Dissemination (Deployment)
   The final stage involves interpreting the profiles of each formed cluster, where the results of visit interest mapping based on academic programs are subsequently transformed into strategic recommendations for the management of the STMIK Methodist Binjai Library in order to personalize services and optimize the distribution of literature collections. This data-to-knowledge transformation process enables management to implement more specific policy interventions, such as expanding reference collections for academic programs with high visit interest and designing literacy promotion programs for student groups within the low-interest cluster. Consequently, the implementation of data mining makes a tangible contribution to improving operational efficiency and enhancing the quality of library services that are more adaptive to the needs of the academic community.

4. Results and Discussion

The initial stage of the study was conducted on 525 library visit log records from the STMIK Methodist Binjai Library collected during the period from October to December 2025. The data cleaning process included

the removal of identity redundancies and the standardization of study program name formatting. Subsequently, through the feature engineering stage, transactional data were transformed into visit frequency variables per student, resulting in 137 unique student records. To minimize bias arising from differences in data scales, normalization was performed using Z-Score Standardization before the data were processed using the K-Means algorithm.

K-Means Clustering Method

Manual computation was performed as a validation step for the algorithm implemented in the system. The calculations were conducted using a sample of student visit frequency data, following the stages outlined below:

1. Data Initialization and Initial Centroid Determination
   The process began with the selection of a sample of student visit frequency data (x) that represents the variance of the dataset, with the selected subjects including.
   a. x_1 (Adelia) : 1 (Low-Interest Visit Cluster)
   b. x_2 (Abet Nego) : 9 (Moderate-Interest Visit Cluster)
   c. x_3 (Bimo Irfan) : 22 (High-Interest Visit Cluster)
   The initial cluster centroids were determined based on an estimation of the initial data distribution namely C_1 = 1,0; C_2 = 8,0; dan C_3= 20,0.

2. Euclidean Distance Calculation
   The proximity of each data object to the cluster centroid was measured using the one-dimensional Euclidean Distance metric, as expressed by the following equation:
   $$d(x, C) = \sqrt{(x - C)^2} \quad (2)$$
   The results of the distance calculations in the initial iteration are as follows:
   a. Object x_1 (1 visit)
      The calculated distances are d(X_1, C_1) = 0, d(X_1, C_2) = 7, and d(X_1, C_3)=19. The minimum distance occurs at C_1 thereby assigning this object to Cluster 1
   b. Object x_2 (9 visits)
      The calculated distances are d(X_2, C_1) = 8, d(X_2, C_2) = 1, and d(X_2, C_3) = 11. The minimum distance occurs at C_2, thus placing this object in Cluster 2.
   c. Object x_3 (22 visits)
      The calculated distances are d(X_3, C_1) = 21, d(X_3, C_2) = 14, dan d(X_3, C_3) = 2. he minimum distance occurs at C_3 assigning this object to Cluster 3.

3. Updating Cluster Centroids
   After the entire dataset, consisting of 137 student entities, was allocated to their respective clusters, the centroid values were updated based on the mean of all cluster members until convergence was achieved. The calculation was performed using the following equation:
   $$C_j = \frac{1}{n}\sum_{i=1}^{n} x_1 \quad (3)$$

Through an iterative process within the system, the final cluster centroids were obtained and used as parameters for mapping, namely C_1 = 1,57 (low), C_2 = 9,37 (moderate), dan C_3 = 21,00 (high).

Table 1. Final Distance Matrix and Cluster Assignment

Student Name	Frequency (x)	Distance to C_1	Distance to C_2	Distance to C_3	Selected Cluster
Adelia	1	0.57	8.37	20.00	Cluster 1
Abet Nego	9	7.43	0.37	12.00	Cluster 2
Bimo Irfan	22	20.43	12.63	1.00	Cluster 3

 ISSN 2797-7889

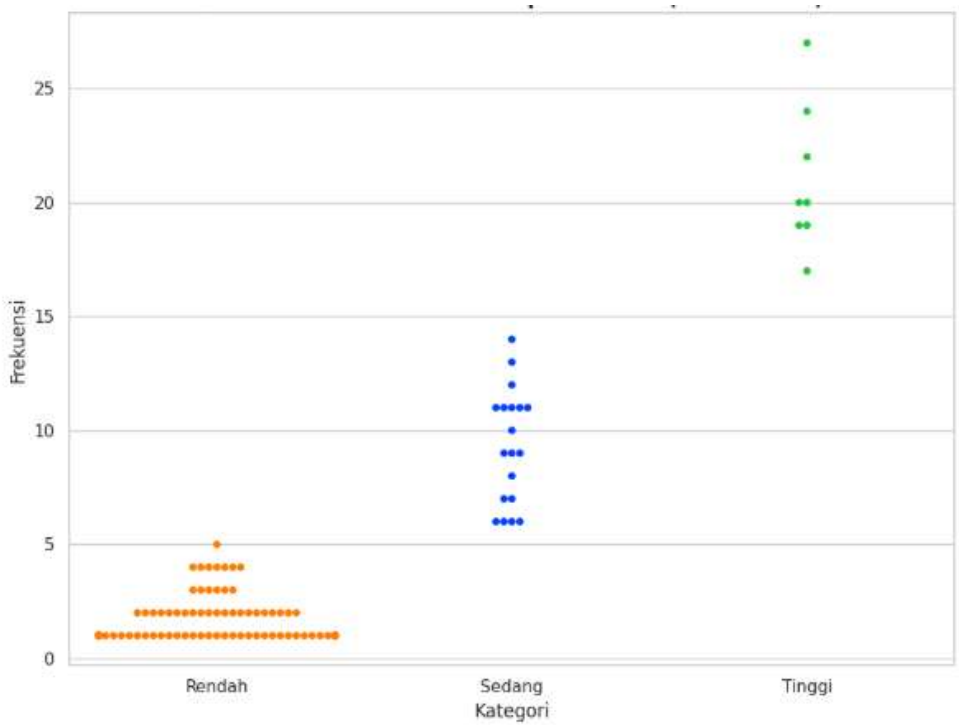

Figure 2. Distribution of Individual Student Visit Frequencies per Cluster

Figure 2 illustrates the data density distribution, showing that the low-category cluster contains the largest number of students.

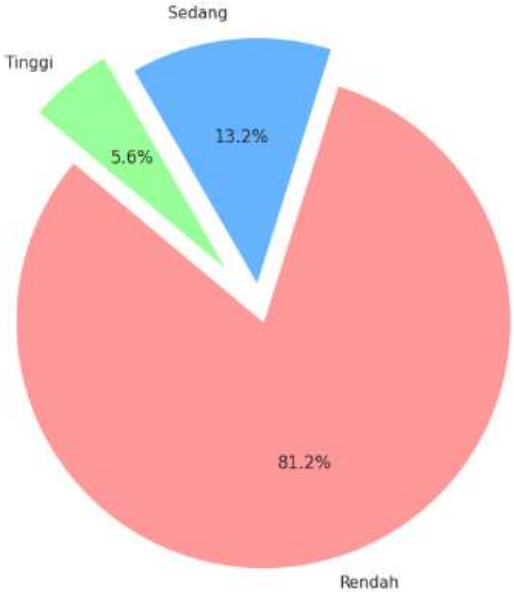

Figure 3. Percentage of Students According to Visit Interest Category

Figure 3. Visualization showing the predominance of students in the low visit interest category.

 ISSN 2797-7889

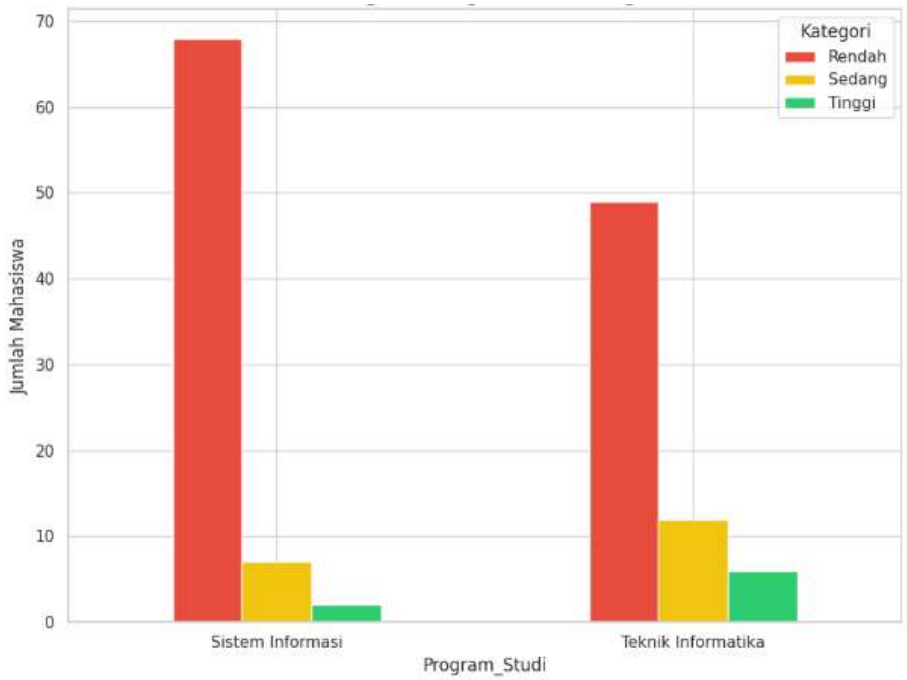

Figure 4. Comparison of Student Visit Interest across Academic Programs

Figure 4 highlights the significant differences between the Informatics Engineering and Information Systems programs.

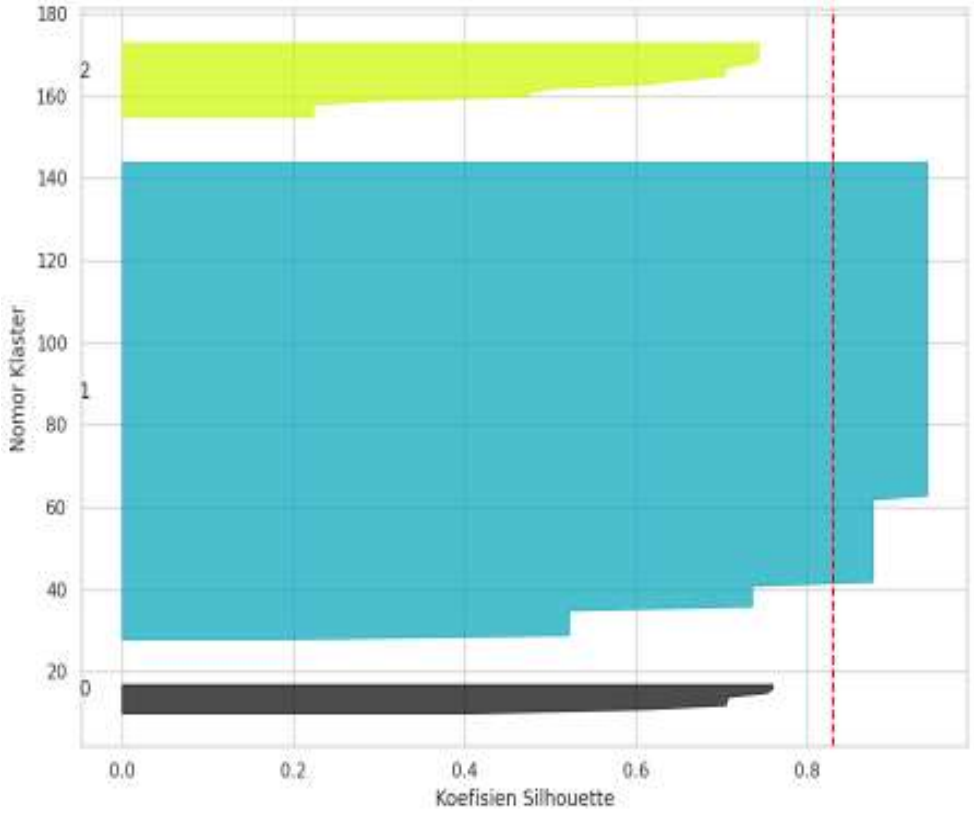

Figure 5. Silhouette Coefficient Analysis Results for Cluster Strength Validation

Figure 5 illustrates the Silhouette Score (0.8304). A value approaching 1.0 indicates that the clustering is very strong, with well-separated clusters and high cohesion within each cluster.

5. Conclusion

Based on the analysis results, this study concludes that the implementation of the K-Means Clustering algorithm at the STMIK Methodist Binjai Library is proven to be effective with a high level of validity, as confirmed by a Silhouette Coefficient score of 0.8304 and the determination of K=3K = 3K=3 as the most optimal number of clusters through the Elbow Method. The visitor segmentation successfully identified three main profiles: the Low Cluster (117 students) with an average of 1.50 visits, the Moderate Cluster (19 students) with an average of 9.32 visits, and the High Cluster (8 students) with an average of 21.00 visits. The findings indicate significant disparities between academic programs, where Informatics Engineering has a higher proportion of active students compared to Information Systems, whose majority of students (68 individuals) remain in the low visit interest category. Therefore, future service optimization strategies should focus on literacy stimulation for this program while providing recognition and incentives for active visitors.

6. References

[1] A. Yasir, "SISTEM INFORMASI PERPUSTAKAAN BERBASIS WEB PADA PERPUSTAKAAN UNIVERSITAS DHARMAWANGSA," *Djtechno: Jurnal Teknologi Informasi*, vol. 1, no. 2, pp. 36–40, Dec. 2020, doi: 10.46576/DJTECHNO.V1I2.970.

[2] D. Nafila and V. Riyanto, "Penerapan Algoritma K-Means Clustering Pada Pola Kunjungan Perpustakaan menggunakan Soft System Methodology," *Jurnal Ticom: Technology of Information and Communication*, vol. 13, no. 1, pp. 30–37, Sep. 2024, doi: 10.70309/TICOM.V13I1.126.

[3] Indonesia, *Undang-Undang Nomor 43 Tahun 2007 tentang Perpustakaan*. 2007.

[4] R. Mahardika Sari, I. Gusti Ayu Karnasih, I. Fakhriza, and P. Kemenkes Malang, "FAKTOR-FAKTOR RENDAHNYA KUNJUNGAN MAHASISWA KE PERPUSTAKAAN," *LEARNING : Jurnal Inovasi Penelitian Pendidikan dan Pembelajaran*, vol. 5, no. 2, pp. 599–608, May 2025, doi: 10.51878/LEARNING.V5I2.4867.

[5] S. Hidayati and U. Suciati, "Memahami karakteristik Pemustaka dalam layanan Perpustakaan," *Media Informasi*, vol. 29, no. 1, pp. 128–141, Jun. 2020, doi: 10.22146/MI.V29I1.4014.

[6] V. A. kumar and M.Chidambaram, "Personalization and User Behavior Analysis in Digital Libraries: A Systematic Review," *Academic Research Journal of Science and Technology (ARJST)*, vol. 2, no. 02, pp. 37–43, Aug. 2025, doi: 10.63300/ARJST0202202505.

[7] Y. Gao and W. Gao, "User profiling in university libraries by combining multi-perspective clustering algorithm and reader behavior analysis," *Nonlinear Engineering*, vol. 14, no. 1, pp. 1–12, Jan. 2025, doi: 10.1515/NLENG-2025-0175/XML.

[8] O. Durodolu, J. Onaade Ojo, and T. Abosede Ajayi, "Evaluating User Perceptions and the Effectiveness of Gamification Elements in Enhancing Engagement with Library Resources," *The Journal of Information and Documentation Studies*, vol. 0, no. 23, pp. 21–34, Jun. 2025, doi: 10.26650/BBA.2025.23.1633527.

[9] C. S. Octiva, T. I. Fajri, E. B. Sulistiarini, S. Suharjo, and U. W. Nuryanto, "Penggunaan Teknik Data Mining untuk Analisis Perilaku Pengguna pada Media Sosial," *Jurnal Minfo Polgan*, vol. 13, no. 1, pp. 1074–1078, Jul. 2024, doi: 10.33395/JMP.V13I1.13936.

[10] A. I. Zalukhu, M. Iqbal, and D. Nasution, "ANALISIS DATA MINING DALAM PENGELOLAAN PERSEDIAAN STOK DENGAN ALGORITMA RANDOM FOREST DAN APRIORI (STUDI KASUS: TOKO CERIA BABYSHOP)," *JOURNAL OF SCIENCE AND SOCIAL RESEARCH*, vol. 8, no. 3, pp. 3396–3405, Jun. 2025, doi: 10.54314/JSSR.V8I3.3544.

[11] Z. Sitorus, I. Syahputra, C. I. Angkat, and D. Sartika, "Implementation of K-Means Clustering for Inventory Projection," *International Journal of Science, Technology & Management*, vol. 5, no. 3, pp. 673–678, May 2024, doi: 10.46729/IJSTM.V5I3.856.

[12] N. K. Sari and Y. Hendriyani, "Clustering Data Pengunjung UPT Perpustakaan, Penerbitan dan Percetakan Universitas Negeri Padang Menggunakan Algoritma K-Means," *Jurnal Pendidikan Tambusai*, vol. 7, no. 3, pp. 29913–29923, Dec. 2023, doi: 10.31004/JPTAM.V7I3.11826.

[13] J. Hutagalung, Y. Hendro Syahputra, Z. Pertiwi Tanjung, S. Triguna Dharma, and J. I. Pintu Air, "Pemetaan Siswa Kelas Unggulan Menggunakan Algoritma K-Means Clustering," *Hal AH Nasution*, vol. 9, no. 1, 2022, [Online]. Available: http://jurnal.mdp.ac.id

[14] A. Falakhi, "Pengolahan Data Pelanggan Dengan Tenik Clustering K-Means Di Aplikasi Weka," *Journal Computer Science and Information Systems : J-Cosys*, vol. 3, no. 2, pp. 54–60, Jul. 2023, doi: 10.53514/JCO.V3I2.394.

[15] S. Amalia *et al.*, "IMPLEMENTASI ALGORITMA K-MEANS CLUSTERING DALAM PEMETAAN KONDISI PERPUSTAKAAN DI INDONESIA BERDASARKAN PROVINSI TAHUN 2023," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 4, pp. 6252–6259, May 2025, doi: 10.36040/JATI.V9I4.14052.

[16] S. Angelika, G. Daely, A. E. Sanjaya, and A. Wijaya, "Analisis Pola Kepuasan Pengunjung Amanzi Waterpark Palembang Menggunakan Algoritma K-Means Clustering," *Jurnal Ilmu Komputer Dan Informatika*, vol. 2, no. 3, pp. 52–60, 2026, [Online]. Available: https://jurnal.globalscients.com/index.php/jiki.930

[17] I. R. Padiku and A. Lahinta, "Penerapan Clustering K-Means Untuk Mendukung Pengelolaan Koleksi Pada Perpustakaan Fakultas Teknik Universitas Negeri Gorontalo," *Jurnal Teknik*, vol. 20, no. 1, pp. 54–62, Jul. 2022, doi: 10.37031/JT.V20I1.206.