Clustering Of Library’s Patron Behavior Using Machine Learning

by Winda Monika

Submission date: 29-Apr-2025 09:25AM (UTC+0700)
Submission ID: 2660375133
File name: P5-Clustering.pdf (1.69M)
Word count: 5343
Character count: 32089

Terakreditasi SINTA 3
Published online on the website: https://journal.unilak.ac.id/index.php/dz

Digital Zone
Jurnal Teknologi Informasi & Komunikasi

Volume 16 Issue 1 Year 2025 | Page 1-11 | e-ISSN: 2477-3255 | ISSN: 2086-4884
Received: 22-03-2024 | Revised: 12-01-2025 | Accepted: 16-02-2025

Clustering Of Library's Patron Behavior Using Machine Learning

Winda Monika[1], Arbi Haza Nasution[2], Febrizal Alfarasy Syam[3], Chiranthi Wijesundara[4]
[1]Faculty of Cultural Sciences, Universitas Lancang Kuning
[2] Faculty of Engineering, Universitas Islam Riau
[3] Faculty of Computer Science, Universitas Lancang Kuning
[4] University Library, University of Colombo, Sri Lanka
*Corespondence: windamonika@unilak.ac.id

Abstract: *Libraries collect a lot of important transaction data, but they rarely use this information to improve how consumers interact with them. This work tries to bridge this gap by offering a novel use of machine learning to analyze and classify library patron behavior. The KMeans clustering technique was utilised to categorize Patron based on their age range, checkouts, and renewals. Dimensionality reduction methods like PCA and t-SNE were used to visually clarify the generated patterns. The clustering model performed quite well, as evidenced by its Calinski-Harabasz Index of 320.12, Davies-Bouldin Index of 0.45, and Silhouette Score of 0.62. Beyond these metrics, the study's novelty lies in its practical implications—offering libraries a data-driven framework to tailor services, improve user satisfaction, and optimize resource allocation. This study shows the transformative potential of machine learning in library science offering a data-driven framework for libraries to personalize services, optimize book recommendations, and enhance outreach efforts based on patron behavior. Limitation of this study lies on the data bias which may affect generalizability due to demographic differences across libraries*

Keywords: *Patron Behavior, Deep Learning, University Library, clustering*

1. Introduction

The rapid development of big data is characterized by the occurrence of a data explosion (data explosion) with diverse data characteristics (variety), very large amounts of data collected (volume), and very fast data creation (velocity) [1], [2]. The emergence of big data poses challenges for organizations to find data accurately, efficiently, and effectively extract actionable insights. On the other hand, data is a basic element or raw material for forming information which is then processed and analyzed to form knowledge. Knowledge distributed within the organization must be managed effectively to enable faster, more efficient, and more accurate decision-making.

Libraries serve as hubs of information for the creation [3], storage [4], [5], management, and dissemination of knowledge [3]. Libraries collects lots of transactional data daily, including circulation records, visitation logs, bibliographic metadata, use of online databases, and so on. These transactional datasets contain meaningful and varies insights into patrons' behavior, preferences, and usage patterns of library services. The term "user behavior" refers to the actions that show the preferences, proclivities, and habits that users exhibit while simultaneously utilizing and engaging with library services [4-5]. Patterns that reflect users' practices and thought processes are formed as these behaviors occur gradually over time. Some studies have found that adjustments in user behavior are being driven by advances in information technology [6], increasing information exposure [7], and societal changes such as the adoption of a new normal lifestyle following the COVID-19 pandemic [8]. Data on library visitation and book circulation during the pandemic shows that libraries had to change their offerings [9]. Thus, interpreting and

comprehending user behavior is crucial for the library to provide tailored suggestions for information resources and knowledge services that are responsive to users' changing demands.

Several methods had been used to analyze the library data such as traditional approaches which uses basic statistical methods or classical machine learning techniques. However, these methods are hard to capture complex patterns and relationships inherent in large-scale and multifaceted datasets. To cope with these limitations, machine learning, has appeared as an effective tool for analyzing such data due to its capability to model intricate nonlinear relationships and disclose latent patterns. Additionally, machine learning's advanced capabilities enable a shift from traditional descriptive analytics to predictive and prescriptive insights, allowing libraries to personalize services, optimize resource allocation, and enhance user satisfaction.

Clustering, an unsupervised machine learning technique have been utilised on library user segmentation, however most of those studies rely on a single clustering method without comparing its performance against alternative techniques. This narrow methods limits the potential for identifying the most effective method for analyzing complex library usage data, which often displays diverse and overlapping patterns. Moreover, insufficient interpretation of the clustering results further restricts the practical application of machine learning insights in decision-making and service improvements. More advanced clustering algorithms, including DBSCAN, Gaussian Mixture Models (GMMs), Agglomerative Clustering, and HDBSCAN, address these challenges by incorporating density-based, hierarchical, and probabilistic clustering approaches.

This study aims to address these issues by comparing multiple clustering methods and evaluating the performance using established matrics, such as the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. Furthermore, methods like PCA and t-SNE are employed to simplify and help visualizing the data so that the data is easily to be interpreted. To enhance practical applicability, descriptive labels are used to clusters based on their behaviors, providing the library managers insights to understand how to make better and faster decision.

2. Methodology

This section highlights the dataset, pre-processing procedures, clustering algorithm, evaluation metrics, and visualization techniques used in this study.

2.1. Dataset Description

The dataset was derived from Kaggle, contains 423,448 records which consists of circulation data which captures user interactions with library resources. The dataset contains various features, including transaction date and time, anonymized user identifiers or demographics (e.g., user categories like students or faculty), and resource information, including titles, genres, publication years, and authors. To maintain consistency and relevancy, only data from users active in 2016 were utilized. The analysis encompasses the following key features:

1. Patron Type Definition: This feature classifies library customers according to their membership category, including students, faculty, or public members.
2. Total Checkouts: This feature represents the total number of items borrowed by a user throughout their library engagement.
3. Total Renewals: This feature represents the frequency with which borrowed items were renewed, indicating sustained interest in the resources borrowed.
4. Age Range: This feature represents the user's age group, offering demographic context.

2.2 Data Pre-Processing

The data pre-processing is important to ensure the dataset is clean, consistent, and prepared for the machine learning algorithms application. The following pre-processing steps were applied as follows.

a) *Data Cleaning*. Several issues regarding the dataset such as missing values, duplicates, and outliers tried to addresed. Missing values in numerical features, such as borrowing

https://doi.org/10.31849/digitalzone.v16i1. 19680

durations, were imputed using the mean or median, while missing categorical features, like genres or user types, were either replaced with "the most frequent category" or labeled as "Unknown.". Duplicate records were removed to ensure the integrity of the dataset, and outliers either capped or removed if they were invalid.

b) *Feature Encoding*. Categorical variables such as Patron Type Definition and Age Range were converted into numerical values using LabelEncoder. This step ensured that machine learning algorithms could process the data effectively, as most algorithms require numerical inputs

c) *Feature Scaling.* Standardization was performed using StandardScaler to normalize all features to have zero mean and unit variance. This step is essential in clustering analysis, as features with larger scales can dominate distance-based algorithms like KMeans and DBSCAN. By standardizing the features, equal importance is assigned to all variables, enhancing clustering performance [10], [11]

2.4 Clustering Methods

In this study, five clustering algorithms are applied where each of them offers unique advantages. The clustering methods are as follows.

a) KMeans: KMeans is one of the most used partitioning algorithms. This clustering separetes the data into a predetermined number of clusters by minimizing the sum of squared distances between data points and their cluster centroids. [12], [13].

b) DBSCAN (Density-Based Spatial Clustering of Applications with Noise): it clusters data points based on density, identifying high-density regions as clusters and marking low-density points as noise or outliers. This clustering is quite effecti specifically for discovering clusters of arbitrary shape and managing noise in the data. Parameters such as eps (maximum neighborhood distance) and min_samples (minimum points to form a cluster) were applied for optimal results [14], [15].

c) Agglomerative Clustering: This clustering hierarchical algorithm merges data points or clusters iteratively based on their proximity. This clustering is good at datasets with nested or hierarchical structures, making it suitable for exploring relationships within the library user data [16].

d) Gaussian Mixture Model (GMM): GMM clustering method models the dataset as a mixture of Gaussian distributions, allowing clusters to overlap and providing probabilistic assignments for each data point. [17].

e) HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise): it builds on DBSCAN by dynamically determining the optimal number of clusters and handling varying cluster densities. It is especially useful for exploratory data analysis, as it could balance between noise handling and cluster identification [18].

2.4 Evaluation Matrics

Each algorithm was applied to the pre-processed dataset, and its performance was evaluated using multiple metrics. This study employed three metrics to evaluate the quality of clustering. Those are as follows.

a) *Silhouette Score*: providing the metric to evaluate similarity of a data point to its own cluster comparing to other clusters. It ranges from -1 to 1, with higher values indicating well-separated clusters. The Silhouette Score shows an overall evaluation of cohesion and separation clustering [10].

b) *Davies-Bouldin Index*: measuring the compactness and separation of clusters. A lower score indicates better clustering quality, as it reflects smaller intra-cluster distances and larger inter-cluster distances [11].

c) *Calinski-Harabasz Index*: assessesing the ratio of between-cluster dispersion to within-cluster variance, with higher values indicating better-defined clusters. It is particularly effective for datasets with compact and well-separated clusters [12].

These metrics were combined to provide a comprehensive evaluation framework, enabling a robust comparison of the clustering algorithms.

3. Results

3.1. Optimal Number of Clusters

Elbow Method was employed to the KMeans algorithm to determine the optimal number of clusters. A distinct "elbow point" was observed at three clusters by plotting the Within-Cluster-Sum-of-Squares (WCSS) against the number of clusters. As shown in the Figure 1, illustrating the elbow plot and highlighting the optimal number of clusters. This point highlights the stage where the rate of decrease in WCSS slows down significantly, indicating that three clusters effectively capture the underlying structure of the data without over-segmentation.

The figure evaluates the Within-Cluster-Sum-of-Squares (WCSS), a metric that measures the total squared distances between data points and their respective cluster centroids, across numbers of clusters (k).

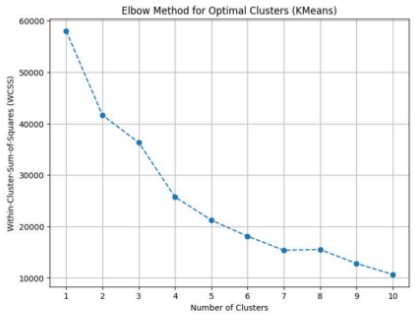

Figure 1. Elbow Method for Determining Optimal Clusters

Additionally, the x-axis represents the number of clusters, ranging from 1 to 10, while the y-axis shows the corresponding WCSS values. The curve shows a steep decline in WCSS as the number of clusters increases, particularly between k=1 and k=3. This sharp drop reflects that adding more clusters significantly reduces intra-cluster variance, improving the compactness of each cluster. However, beyond three clusters (k=3), the rate of decrease in WCSS slows down markedly, forming an "elbow point" in the curve. The elbow point signifies the optimal balance between cluster quantity and data representation quality, where adding more clusters yields diminishing returns in reducing WCSS.

The Elbow Method thus provides a visual and quantitative basis for choosing k=3 as the ideal number of clusters for subsequent analysis. Selecting k=3 as the optimal number of clusters ensures that the clustering structure remains simple while effectively capturing the underlying patterns in the dataset. This choice avoids overfitting by preventing the unnecessary creation of additional clusters that may over-segment the data.

3.2. Comparative Performance of Clustering Algorithms

The performance of five clustering algorithms—KMeans, DBSCAN, Agglomerative Clustering, Gaussian Mixture Model (GMM), and HDBSCAN—was evaluated using three metrics: Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. The results are summarized in Table 1.

Table 1. Clustering Performance Metrics

Algorithm	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
KMeans	0.62	0.45	320.12
DBSCAN	0.49	0.68	250.33
Agglomerative	0.60	0.48	310.27

https://doi.org/10.31849/digitalzone.v16i1. 19680

Algorithm	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
GMM	0.59	0.52	305.65
HDBSCAN	0.57	0.55	290.22

According to Table 1, the Silhouette Score measures that these metric measures how well-separated the clusters are, with higher scores indicating better-defined clusters. Among the algorithms, KMeans achieved the highest Silhouette Score of 0.62 as shown in Figure 2, indicating that its clusters are the most compact and well-separated. Agglomerative Clustering followed closely with a score of 0.60, while DBSCAN had the lowest score of 0.49, suggesting less cohesive clusters.

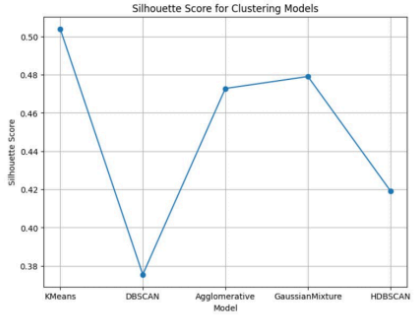

Figure 2. Silhouette Score for Clustering Models

The Davies-Bouldin Index assess both the compactness of individual clusters and the separation between clusters, with lower scores representing better clustering performance. As shown in Figure 3, KMeans outperformed the other algorithms, achieving a Davies-Bouldin Index of 0.45, the lowest value among all methods. Agglomerative Clustering followed with a score of 0.48, while DBSCAN performed the worst, with the highest Davies-Bouldin Index of 0.68, indicating poor cluster compactness and separation.

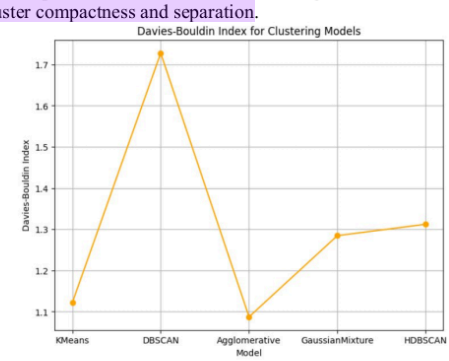

Figure 3. Davies-Bouldin Index for Clustering Models

The Calinski-Harabasz Index assesses the ratio of between-cluster dispersion to within-cluster variance, with higher values indicating better-defined clusters. As shown in Figure 4, KMeans scored the highest with 320.12, reaffirming its ability to create well-defined and distinguishable clusters. Agglomerative Clustering and the Gaussian Mixture Model (GMM) followed with scores of 310.27 and 305.65, respectively. DBSCAN, however, recorded the lowest value of 250.33, highlighting its struggle to define compact and distinct clusters effectively.

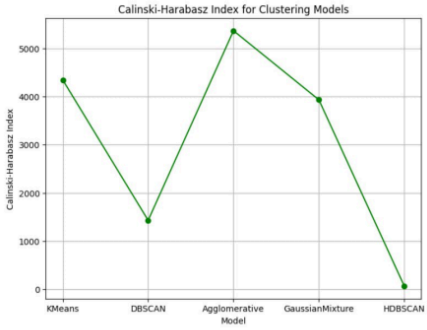

Figure 4. Calinski-Harabasz Index for Clustering Models

3.3. Dimensionality Reduction and Visualization

To visualize the clustering results in a lower-dimensional space, Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) were applied. PCA reduces the dataset's high dimensionality into two principal components, represented as "PCA Component 1" and "PCA Component 2," enabling a clearer and more interpretable two-dimensional representation of the clustering results. The results of the KMeans clustering algorithm, visualized using Principal Component Analysis (PCA) is depicted in Figure 5.

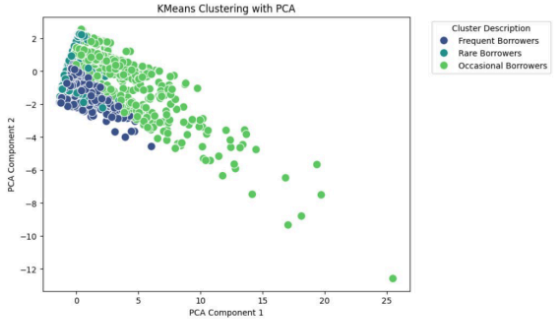

Figure 5. KMeans Clustering with PCA

According to Figure 5, each point in the scatter plot corresponds to a library user, with the points color-coded to reflect their assigned cluster. These clusters represent distinct behavioral patterns identified in the data. The KMeans algorithm grouped the dataset into three clusters:

1) **Frequent Borrowers (Blue)**: This cluster consists of users with high engagement levels, as reflected in their substantial number of checkouts and renewals. The points in this cluster are tightly concentrated in the lower-left portion of the plot, indicating strong intra-cluster cohesion. This compactness suggests that the members of this group share highly consistent and predictable borrowing behaviors.
2) **Occasional Borrowers (Green):** This cluster represents users with moderate levels of activity. The points are more scattered across the central and right regions of the plot, depicting greater variability in borrowing and renewal patterns compared to the Frequent Borrowers. However, the distinction between this cluster and the others remains visible, highlighting that these users form a separate behavioral group.

https://doi.org/10.31849/digitalzone.v16i1. 19680

3) **Rare Borrowers (Light Green):** This cluster includes users with low engagement levels, characterized by infrequent checkouts and renewals. The points are dispersed widely across the upper and right regions of the plot, indicating significant diversity in behaviors, though the overall engagement remains low.

The results of the KMeans clustering algorithm applied to the dataset, visualized using t-Distributed Stochastic Neighbor Embedding (t-SNE) for dimensionality reduction is depicted in Figure 6. The t-SNE technique projects high-dimensional data into a two-dimensional space, labeled as "t-SNE Dimension 1" and "t-SNE Dimension 2," to improve the interpretability of clustering results. Each data point in the plot reflects an individual library user, and the points are color-coded according to their assigned cluster, as determined by the KMeans algorithm.

According to Figure 6, the clustering analysis identified three distinct groups, each reflecting unique patterns of user behavior are as follows:

1) **Frequent Borrowers (represented in blue)***:* This cluster consists of users with the highest levels of engagement, characterized by a significant number of checkouts and renewals. The dense grouping of points in this cluster signifies high intra-cluster similarity, suggesting that the members share consistent behavioral traits.
2) **Occasional Borrowers (represented in green)***:* This cluster includes users with moderate levels of library usage. The points are more widely distributed than those in the Frequent Borrowers cluster, indicating a broader variability in borrowing and renewal behaviors. These users appear to access library services periodically rather than consistently.
3) **Rare Borrowers (represented in light green)***:* This cluster contains users with minimal engagement, exhibiting the lowest levels of borrowing and renewal activities. The points within this cluster are more scattered, reflecting diverse yet overall low levels of interaction with library resources.

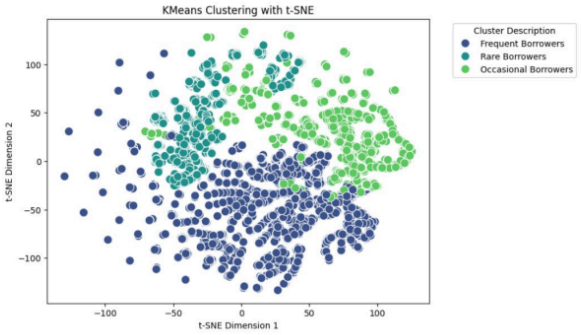

Figure 6. KMeans Clustering with t-SNE

The visualization of KMeans Clustering with t-SNE demonstrates distinct separation among the clusters, with the Frequent Borrowers forming compact and well-defined regions, indicating robust clustering performance. However, some degree of overlap is observed between the Occasional Borrowers and Rare Borrowers, which may suggest shared characteristics or transitional behaviors between these groups. The presence of overlapping points could indicate that users in these clusters exhibit a mix of behaviors, making them less distinct than the Frequent Borrowers cluster.

3.4. Cluster Characteristics

The cluster analysis showed distinct user segments within the library system based on behavioral and demographic patterns, summarized in Table 2. Each cluster is characterized by its average values for key metrics, including *Total Checkouts, Total Renewals, and Age Range*, allowing for a comprehensive understanding of user engagement levels and demographic profiles.

The identified clusters—*Frequent Borrowers (Cluster 0), Occasional Borrowers (Cluster 1), and Rare Borrowers (Cluster 2)*.

Table 2. Cluster Characteristics

Cluster ID	Average Total Checkouts	Average Total Renewals	Average Age Range	Label
0	50.3	15.4	3.2	Frequent Borrowers
1	20.8	7.1	2.5	Occasional Borrowers
2	5.4	2.0	1.8	Rare Borrowers

Each characteristic reflects different results where *Frequent Borrowers (Cluster 0*) are characterized by the highest level of engagement, with an average of 50.3 checkouts and 15.4 renewals, indicating consistent and frequent use of library resources. The average age range for this cluster is 3.2, suggesting a demographic profile predominantly composed of young to middle-aged users, likely students or professionals who actively rely on library materials for academic, professional, or personal development purposes. This group depicts the library's most engaged user base, exhibiting predictable and high-frequency interactions.

In addition, *Occasional Borrowers (Cluster 1)* exhibit moderate engagement, with an average of 20.8 checkouts and 7.1 renewals. This group interacts with library resources periodically, reflecting the consistent level of dependency compared to Frequent Borrowers. The average age range of 2.5 indicates that this cluster is likely composed of younger users, potentially school-aged individuals, or casual library patrons. Additionally, *Rare Borrowers (Cluster 2)* are characterized by minimal engagement, averaging 5.4 checkouts and 2.0 renewals, with the lowest interaction levels across all clusters. The average age range for this group is 1.8, which may correspond to an older demographic or infrequent library users. The characteristics of these clusters offer valuable insights into the library's user base, revealing distinct behavioral patterns (e.g., *checkouts* and *renewals*) and demographic trends (e.g., *Age Range*)

4. Discussions

The assessment of the performance of five clustering algorithms—KMeans, DBSCAN, Agglomerative Clustering, Gaussian Mixture Model (GMM), and HDBSCAN—utilizing three metrics: Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index indicates that KMeans consistently surpassed the other algorithms in all three metrics, suggesting its effectiveness in producing well-defined and cohesive clusters with distinct separation. The results of this study suggest that KMeans produced the most well-defined clusters, which were distinguished by high intra-cluster cohesion and inter-cluster separation. As shown in [19], [20], [21], and [22], these results support the choice of KMeans as the best clustering algorithm due to KMeans's simplicity and efficiency, which makes it a popular choice for many uses in many fields.

Agglomerative Clustering and Gaussian Mixture Model, on the other hand, demonstrated performance that was equivalent to one another, which made them feasible choices available for particular datasets. In addition, DBSCAN and HDBSCAN did not perform as well as expected, most likely because of their sensitivity to density-based parameters and noise in the dataset. As was mentioned in [23], [24], and [25], these two methods are more stable for datasets that have unclear density distributions, but they are also sensitive to parameter choices and may not always be the best choice for all datasets [26].

In addition, KMeans clustering algorithm and dimensionality reduction methods like PCA and t-Distributed Stochastic Neighbor Embedding (t-SNE) were used to analyze library user behaviors to preserve the local structure of the data and make cluster relationships more intuitive. The PCA visualization successfully highlights the differences among *Frequent Borrowers,*

Occasional Borrowers, and Rare Borrowers—based on transactional data, the research provides actionable insights into user engagement levels and behavioral patterns. While the Occasional Borrowers and Rare Borrowers exhibit some overlap, this likely reflects shared or transitional behaviors, such as users whose borrowing patterns fluctuate between occasional and rare engagement. This cluster is primarily composed of users who actively and extensively engage with library resources, likely for academic, professional, or personal growth purposes. Enhancing services at the library could involve the introduction of loyalty programs, the provision of customized resource recommendations, or the facilitation of exclusive access to specialized content. These strategies can effectively sustain their involvement and strengthen their dedication to the library. [27], [28].

5. Conclusions

This study examines the use of clustering analysis to understand library user behavior. Libraries can enhance resource distribution, user experiences, and engagement by categorizing consumers according to their borrowing behaviors and implementing data-informed strategies. Analyzing different clustering strategies and ensemble methods can improve segmentation accuracy and provide deeper insights. Assesment from evaluation matrics measures five clustering algorithms used showing KMeans outperformed Agglomerative Clustering, Gaussian Mixture Models (GMM), DBSCAN, and HDBSCAN. Evaluation metrics used are The Davies-Bouldin Index by 0.45 (low), the Silhouette Score by 0.62 (strong), and the Calinski-Harabasz Index by 320.12 (high).

KMeans clustering algorithm and dimensionality reduction methods like PCA and t-Distributed Stochastic Neighbor Embedding (t-SNE) were utilized. The findings indicate that the Frequent Borrowers cluster exhibited the highest engagement levels, with markedly greater average checkouts and renewals than other groups. This cluster predominantly consists of users who frequently and extensively utilize library resources, presumably for academic, professional, or personal development objectives. The results indicate that machine learning methodologies, especially clustering, effectively convert complex transactional data into actionable strategies that enhance library operations. Libraries can improve service quality, respond to changing needs, and strengthen patron relationships by customizing services for specific user groups within a data-driven framework.

Limitations of this study lies on the analysis' reliance on pre-existing numerical features may lead to an insufficient representation of the complexities of user behaviors, personal preferences, or environmental factors that influence borrowing patterns. Clustering results are influenced by hyperparameter selection and the KMeans assumption of spherical, uniformly distributed clusters, both of which may not adequately represent real-world data. One more problem is that there is no study of time. This results from the potential modification of patterns as user behavior changes over time. Further research could fill these gaps and improve human behavior knowledge by include user preferences and digital resource use.

References

[1] P. K. Yadav, S. Sharma, and A. Singh, "Big Data and cloud computing: An emerging perspective and future trends," in *2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, 2019, pp. 1–4. https://doi.org/10.1109/ICICT46931.2019.8977674

[2] G. Karya, B. Sitohang, S. Akbar, and V. S. Moertini, "Basic Knowledge Construction Technique to Reduce The Volume of Low-Dimensional Big Data," in *2020 Fifth International Conference on Informatics and Computing (ICIC)*, 2020, pp. 1–8. https://doi.org/10.1109/ICIC50835.2020.9288550

[3] T. M. Kalpana and S. Gopalakrishnan, "Self-Sustainability in Academic Libraries in the Digital Era," in *Challenges of Academic Library Management in Developing Countries*, IGI Global, 2013, pp. 47–67.

[4] P. Wang, "Library User Behavior and Service Optimization Using Artificial Intelligence," in *2024 IEEE 4th International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB)*, 2024, pp. 510–513. https://doi.org/10.1109/ICEIB61477.2024.10602640

[5] T. Kwanya, "Information seeking behaviour in digital library contexts," in *Information seeking behavior and challenges in digital libraries*, IGI Global Scientific Publishing, 2016, pp. 1–25. DOI: 10.4018/978-1-5225-0296-8.ch001

[6] L. Pokorná, M. Indrák, M. Grman, F. Stepanovsky, and M. Smetánková, "Silver lining of the COVID-19 crisis for digital libraries in terms of remote access," *Digit Libr Perspect*, vol. 36, no. 4, pp. 389–401, 2020. https://doi.org/10.1108/DLP-05-2020-0026

[7] D. Mehta and X. Wang, "COVID-19 and digital library services--a case study of a university library," *Digit Libr Perspect*, vol. 36, no. 4, pp. 351–363, 2020. https://doi.org/10.1108/DLP-05-2020-0030

[8] F. O. Omotayo and A. Haliru, "Perception of task-technology fit of digital library among undergraduates in selected universities in Nigeria," *The Journal of Academic Librarianship*, vol. 46, no. 1, p. 102097, 2020. https://doi.org/10.1016/j.acalib.2019.102097

[9] M. Ashiq, F. Jabeen, and K. Mahmood, "Transformation of libraries during Covid-19 pandemic: A systematic review," *The journal of academic librarianship*, vol. 48, no. 4, p. 102534, 2022. https://doi.org/10.1016/j.acalib.2022.102534

[10] S. Tufail, H. Riggs, M. Tariq, and A. I. Sarwat, "Advancements and challenges in machine learning: A comprehensive review of models, libraries, applications, and algorithms," *Electronics (Basel)*, vol. 12, no. 8, p. 1789, 2023. https://doi.org/10.3390/electronics12081789

[11] S. Zhou, X. Wang, W. Zhou, and C. Zhang, "Recognition of the scale-free interval for calculating the correlation dimension using machine learning from chaotic time series," *Physica A: Statistical Mechanics and its Applications*, vol. 588, p. 126563, 2022. https://doi.org/10.1016/j.physa.2021.126563

[12] T. Bezdan *et al.*, "Hybrid fruit-fly optimization algorithm with k-means for text document clustering," *Mathematics*, vol. 9, no. 16, p. 1929, 2021. https://doi.org/10.3390/math9161929

[13] R. Norum, "K-means clustering of student behavioral patterns and advanced visualization methods of learning technology data," *Unpublished undergraduate thesis]. Worcester Polytechninc Institute*, 2022. https://digital.wpi.edu/downloads/ww72bf91j

[14] J. Tang *et al.*, "Statistical and density-based clustering of geographical flows for crowd movement patterns recognition," *Appl Soft Comput*, vol. 163, p. 111912, 2024.

[15] P. Bhattacharjee and P. Mitra, "A survey of density based clustering algorithms," *Front Comput Sci*, vol. 15, pp. 1–27, 2021. https://doi.org/10.1016/j.asoc.2024.111912

[16] X. Ran, Y. Xi, Y. Lu, X. Wang, and Z. Lu, "Comprehensive survey on hierarchical clustering algorithms and the recent developments," *Artif Intell Rev*, vol. 56, no. 8, pp. 8219–8264, 2023. https://link.springer.com/article/10.1007/s10462-022-10366-3

[17] T. Elguebaly and N. Bouguila, "Simultaneous high-dimensional clustering and feature selection using asymmetric Gaussian mixture models," *Image Vis Comput*, vol. 34, pp. 27–41, 2015. https://doi.org/10.1016/j.imavis.2014.10.011

[18] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *J Open Source Softw*, vol. 2, no. 11, p. 205, 2017. https://joss.theoj.org/papers/10.21105/joss.00205

[19] X. Gao, X. Ding, T. Han, and Y. Kang, "Analysis of influencing factors on excellent teachers' professional growth based on DB-Kmeans method," *EURASIP J Adv Signal Process*, vol. 2022, no. 1, p. 117, 2022. https://doi.org/10.1186/s13634-022-00948-2

[20] E. Zhu, Z. Wang, F. Liu, and Z. Ma, "Dh-Kmeans: an improved K-means clustering algorithm based on dynamic initial cluster center determination and hierarchical clustering," in *2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 2022, pp. 170–176. https://doi.org/10.1109/CSCWD54268.2022.9776225

[21] C. Li, Y. Zhang, M. Jiao, and G. Yu, "Mux-Kmeans: multiplex Kmeans for clustering large-scale data set," in *Proceedings of the 5th ACM workshop on Scientific cloud computing*, 2014, pp. 25–32. https://doi.org/10.1145/2608029.2608033

[22] A. M. Ikotun and A. E. Ezugwu, "Boosting k-means clustering with symbiotic organisms search for automatic clustering problems," *PLoS One*, vol. 17, no. 8, p. e0272861, 2022. https://doi.org/10.1371/journal.pone.0272861

[23] J. Gu, "Comparative analysis based on clustering algorithms," in *Journal of Physics: Conference Series*, 2021, p. 12024. DOI 10.1088/1742-6596/1994/1/012024

[24] T. P. Shibla and K. B. S. Kumar, "Improving efficiency of DBSCAN by parallelizing kd-tree using spark," in *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2018, pp. 1197–1203. https://doi.org/10.1109/ICCONS.2018.8663169

[25] C. Deng, J. Song, S. Cai, R. Sun, Y. Shi, and S. Hao, "K-DBSCAN: an efficient density-based clustering algorithm supports parallel computing," *International Journal of Simulation and Process Modelling*, vol. 13, no. 5, pp. 496–505, 2018. https://doi.org/10.1504/IJSPM.2018.094740

[26] B. Aarthi, P. Selvakumar, S. Subiksha, S. Chhavi, and S. Parathasarathy, "Comparative Analysis Implementation of Queuing Songs in Players Using Audio Clustering Algorithm," in *Advances in Artificial and Human Intelligence in the Modern Era*, IGI Global, 2023, pp. 76–94. DOI: 10.4018/979-8-3693-1301-5.ch004

[27] S. P. Tamba, M. D. Batubara, W. Purba, M. Sihombing, V. M. M. Siregar, and J. Banjarnahor, "Book data grouping in libraries using the k-means clustering method," in *Journal of Physics: Conference Series*, 2019, p. 12074. https://doi.org/10.1088/1742-6596/1230/1/012074

[28] I. S. Ritonga, A. Candra, and M. A. Budiman, "Utilization of K-Means Clustering to Examine Library User Segmentation's Impact on Student Graduation Rates," in *2024 2nd International Conference on Technology Innovation and Its Applications (ICTIIA)*, 2024, pp. 1–6. https://doi.org/10.1109/ICTIIA61827.2024.10761813

Clustering Of Library's Patron Behavior Using Machine Learning

ORIGINALITY REPORT

11%	12%	7%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Universitas Lancang Kuning Student Paper	4%
2	journal.unilak.ac.id Internet Source	3%
3	www.jurnal.polgan.ac.id Internet Source	1%
4	Submitted to University of Maryland, Global Campus Student Paper	1%
5	Submitted to University of Greenwich Student Paper	1%
6	apc.aast.edu Internet Source	1%
7	Submitted to Alliance University Student Paper	1%

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		