



Analysis of Preference Clustering and Strategies for Utilizing Sources and Media in Accessing Academic References by College Students Using the K-Means and DBSCAN Algorithms

Putri Varila^{1*}, Rizka Ariyanti², and Nayaka Adyatma Nugroho³, Saniatun Hazanah⁴, Muhammad Fajar Kurniawan⁵, Najwa Andity Khairani⁶

^{1,2,3}UPN "Veteran" Jawa Timur, ⁴Al Azhar University, ⁵Universitas Gajah Mada, ⁶Institut Teknologi Sepuluh Nopember

¹25083010075@student.upnjatim.ac.id, ²25083010057@student.upnjatim.ac.id, ³25083010059@student.upnjatim.ac.id, ⁴saniatunhazanah@gmail.com, ⁵muhammadfajarkurniawan2007@mail.ugm.ac.id, ⁶5002251040@student.its.ac.id

ABSTRACT

Abstract— The increasing availability of academic information sources has led to diverse preferences and strategies among university students in searching for and utilizing academic references. This study aims to identify patterns in students' academic reference-seeking behavior by applying Principal Component Analysis (PCA) and comparing the clustering performance of K-Means and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). Data were collected through an online questionnaire involving 150 university students, with ten numerical variables used in the analysis. The data were preprocessed, standardized using StandardScaler, and reduced using PCA before clustering. The results showed that PCA reduced the ten variables to six principal components while retaining 83.53% of the cumulative variance. K-Means produced two clusters containing 123 and 27 respondents, achieving a Silhouette Score of 0.4543, a Calinski-Harabasz Index (CHI) of 52.2985, and a Davies-Bouldin Index (DBI) of 1.4225. Meanwhile, DBSCAN, with $\text{eps} = 2.3$ and $\text{min_samples} = 2$, identified two clusters consisting of 137 and 2 respondents, along with 11 noise points. DBSCAN achieved a Silhouette Score of 0.5469, a CHI of 15.3191, and a DBI of 0.4818. Therefore, DBSCAN was selected for further analysis to identify differences in students' academic reference-seeking behaviors, particularly in source verification, reference usage, and manual and AI-assisted searching.

Keywords: Principal Component Analysis, K-Means, DBSCAN, Clustering, Academic References.

I. INTRODUCTION

In today's era of digital transformation, access to academic resources has expanded dramatically. Students no longer rely solely on physical libraries and printed books, instead, they have broad access to various digital media, such as online scientific journals, institutional repositories, international databases, and even artificial intelligence (AI)-based search engines. While the availability of these diverse sources and media offers convenience and saves time, it also creates a new challenge in the form of information overload. This situation requires students to develop effective search strategies and the ability to select sources that are truly relevant and credible.

In everyday academic practice, behavioral patterns and preferences regarding the search for and use of literature vary widely. Some students tend to rely on online journals as their primary reference, using Google Scholar as their dominant search tool [1], while others still prefer printed books due to concerns about the credibility of the material [2]. From a technical management perspective, the use of reference management software such as Mendeley and Zotero is increasingly encouraged to organize

* Corresponding author.

E-mail address: 25083010075@student.upnjatim.ac.id

<http://doi.org/10.33005/jasid.v2i2.89>

citations and prevent the risk of plagiarism [3]. Nevertheless, students still frequently face various structural barriers such as limited access to paywalled journals and technical issues with e-libraries in meeting their scientific reference needs [4]. Additionally, variations across academic disciplines influence information-search patterns, with students in certain fields tending toward a linear-efficient approach, while those in other fields adopt a more exploratory approach [5].

This clustering process begins by applying the K-Means algorithm. According to Hasan (2024), K-Means is a partition-based clustering algorithm that aims to divide a set of data points into k clusters based on the similarity of their characteristics, specifically by minimizing the variance within each cluster through the calculation of the shortest distance between data points and their cluster centers (centroids) [6]. This approach is highly efficient for segmenting students into dominant behavioral segments based on their media preferences and browsing habits. However, given that human behavioral data often exhibits irregular distributions and contains unique patterns, the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm was also applied to address these limitations. According to Mahardika (2024), DBSCAN is a non-parametric clustering method that forms clusters based on the density of data points in a spatial space; this algorithm automatically identifies high-density areas separated by low-density areas and is capable of isolating anomalous data as noise without having to specify the number of clusters at the outset [7]. This algorithm is capable of forming clusters with flexible patterns while simultaneously identifying outliers or anomalies.

Based on this background, this paper focuses on examining and analyzing the clustering of preferences and strategies for utilizing academic reference sources and media through the application of the K-Means and DBSCAN algorithms. The results of this analysis are expected to provide a comprehensive understanding of the cluster map of academic literacy and reference behaviors, while also serving as a strategic foundation for designing more targeted academic and digital literacy services.

II. METHODOLOGY

This study was conducted in several stages, namely data collection, preprocessing, standardization, dimensionality reduction using Principal Component Analysis (PCA), clustering using K-Means and DBSCAN, and evaluation of the clustering results. The research workflow is presented in Figure 1.

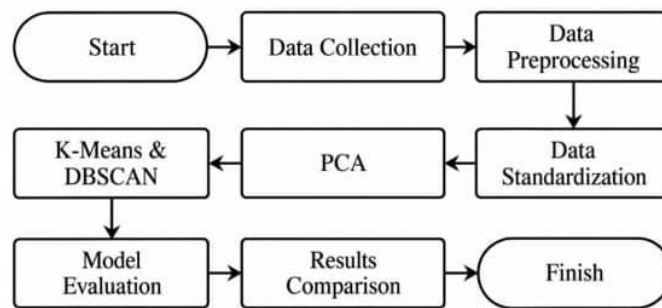


Figure 1. Research Flowchart

A. Data Collection

The research data were obtained by distributing an online questionnaire via Google Forms to students to determine their preferences and strategies for accessing academic references. Ethical considerations regarding participant involvement, consent, and data privacy were strictly maintained throughout the study. Informed consent was explicitly obtained from all respondents on the introductory page of the online questionnaire, which detailed the study's scope, objectives, and the voluntary nature of participation. Although basic personal identifiers (such as names or institutional emails) were collected for administrative and verification purposes, all data were strictly de-identified prior to analysis to ensure participant anonymity and confidentiality. Furthermore, all individual responses were stored securely in a password-protected environment and processed exclusively for academic research purposes. The dataset used consisted of 150 respondents with 10 numerical attributes covering source verification, free reference sources,

source completeness, monthly reference usage, searching and comparison activities, library usage frequency, academic reference activities, and the time required to search for references manually and with AI assistance. All 10 attributes were used in the analysis, as presented in Table 1.

Table 1. Research Variables

ID	Attribute	Description	Data Type
1	Source_Verification	Number of checks or verifications performed on reference sources.	Numeric
2	Free_Sources	The number or rate of utilization of reference sources available for free.	Numeric
3	Source_Completeness	The number or level of completeness of the reference sources used.	Numeric
4	Monthly_Sources	The number of reference sources used in a single search.	Numeric
5	Search_By_Sources	The amount of time spent comparing multiple reference sources.	Numeric
6	Comparison_Time	The amount of time spent comparing multiple reference sources.	Numeric
7	Library_Frequency	The frequency of library visits or use to search for references.	Numeric
8	Activity_By_References	The number of references used in an academic activity.	Numeric
9	Manual_Time	The time taken to search for references manually without AI assistance.	Numeric
10	AI_Time	The time taken to search for references with the assistance of Artificial Intelligence (AI).	Numeric

B. Data Preprocessing

The preprocessing stage involved checking for missing values, duplicate data, and data types, followed by the selection of ten numerical variables. Next, the data were standardized using StandardScaler, a standardization technique that transforms the data so that each variable has a mean of 0 and a standard deviation of 1. This standardization ensures that all variables are on a comparable scale, preventing any single variable from dominating distance calculations due to differences in value ranges.

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

z : Standardized value

x : Original value

μ : Mean of the variable

σ : Standard Deviation of the variable

C. Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction method used to transform

a set of variables into a smaller number of principal components while retaining most of the variance in the data. According to Marukatat [8], PCA constructs principal components based on the directions of maximum variance in the data. In this study, PCA began with the calculation of the covariance matrix to describe the relationships between variables, followed by the computation of eigenvalues and eigenvectors. Eigenvalues indicate the amount of variance explained by each principal component, while eigenvectors determine the direction of each component. The eigenvalues were then arranged in descending order to determine the number of components based on cumulative explained variance, targeting a cumulative threshold range of 75% to 87% [9]. Retaining a subset of principal components within this target range ensures that the core variability and underlying information of the original variables are preserved while effectively eliminating redundant dimensions and potential noise. This dimensionality reduction directly supports the subsequent K-Means and DBSCAN clustering analyses; by transforming correlated original variables into orthogonal principal components, it mitigates the 'curse of dimensionality' and prevents variables with larger scales from disproportionately biasing distance or density calculations, thereby enabling more robust, stable, and accurate cluster formations [10] as expressed in Equation (2).

$$CEV_k = \frac{\sum_{i=1}^k \lambda_i}{\sum_{j=1}^p \lambda_j} \quad (2)$$

- CEV_k : Cumulative proportion of data variance explained by the first k principal components.
 λ_i : Eigenvalue of the i -th principal component.
 λ_j : Eigenvalue of the j -th principal component.
 k : Number of selected principal components.
 p : Total number of principal components.
 $\sum$: Summation operator.

D. Clustering

The clustering process was performed to group student data based on the similarity of their characteristics in accessing academic references. This study used two algorithms, K-Means and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). Both algorithms were applied to the data reduced using PCA to compare the resulting clustering patterns.

1. K-Means

K-Means is a clustering algorithm that groups data based on proximity to cluster centers, or centroids. According to Ikotun et al. [11], the K-Means algorithm works by determining an initial number of centroids, assigning each data point to the nearest centroid, and then iteratively updating the centroid positions until convergence is reached. In this study, the number of clusters was determined by testing several values of k using three evaluation metrics: the Silhouette Score, the Davies-Bouldin Index (DBI), and the Calinski-Harabasz Index (CHI). The optimal number of clusters was selected based on these evaluation results and then used in the final K-Means model [12].

$$d(x_i, \mu_j) = \sqrt{\sum_{r=1}^p (x_{ir} - \mu_{jr})^2} \quad (3)$$

- $d(x_i, \mu_j)$: Euclidean distance between the i -th data point and the centroid of the j -th cluster.
 x_{ir} : Value of the r -th variables of the i -th data point.
 μ_{jr} : Value of the r -th variables of the centroid of the j -th cluster.
 p : Number of principal components used in clustering.
 $\sum$: Summation operator

2. DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a clustering algorithm that forms clusters based on the density of data points. According to Bushra and Yi [13], DBSCAN is capable of identifying clusters of various shapes and distinguishing data points that do not belong to a cluster as noise. This algorithm uses two main parameters: ϵ , which determines the neighborhood radius of a point, and $\min_samples$, which determines the minimum number of points required to form a dense region. In this study, the parameters were determined using a K-Distance Plot and by testing several combinations of ' ϵ ' and ' $\min_samples$ ' using the PCA-transformed data [14]. The formation of point neighborhoods in DBSCAN is expressed by Equation (4).

$$N_\epsilon(x_i) = \{x_j \mid d(x_i, x_j) \leq \epsilon\} \quad (4)$$

$N_\epsilon(x_i)$: Set of data points within the epsilon neighborhood of x_i .
 x_i : Reference data point
 x_j : Data point within the dataset
 ϵ : Neighborhood radius (eps).
 $d(x_i, x_j)$: Distance between two data points.

E. Model Evaluation

An evaluation was conducted to measure the quality of clustering produced by the K-Means and DBSCAN algorithms. This study used three evaluation metrics: the Silhouette Score, the Davies-Bouldin Index (DBI), and the Calinski-Harabasz Index (CHI). These three metrics were used to assess the compactness of data within a cluster as well as the degree of separation between clusters. According to Ikotun et al. [15], the use of cluster validity indices can help assess the quality of clustering based on the characteristics of the resulting cluster structure. The Silhouette Score measures how well data points fit within their assigned clusters compared to other clusters. Its value ranges from -1 to 1, with values closer to 1 indicating better clustering quality. The calculation of the Silhouette Score is shown in Equation (5).

$$s(i) = \frac{b(i) - a(i)}{\max \{a(i), b(i)\}} \quad (5)$$

$s(i)$: Silhouette Score of the i -th data point.
 $a(i)$: Average distance between the data point and other points within the same cluster.
 $b(i)$: Minimum average distance between the data point and points in another cluster.
 $\max$: Maximum value between $a(i)$ and $b(i)$.

The Davies-Bouldin Index (DBI) evaluates the quality of clustering based on the ratio of the spread of data within clusters to the distance between cluster centers. A lower DBI value indicates better clustering quality.

$$DBI = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right) \quad (6)$$

k : Number of clusters.
 S_i and S_j : Average distances of data points within their respective clusters to their centroids.
 M_{ij} : Distance between the centroids of clusters i and j .
 $\max$: Maximum value of the ratio for all other clusters ($j \neq i$).
 Σ : Summation operator

Meanwhile, the Calinski-Harabasz Index (CHI) measures the ratio of inter-cluster variation to intra-cluster variation. A higher CHI value indicates better cluster separation.

$$CHI = \frac{B_k / (k - 1)}{W_k / (n - k)} \quad (7)$$

B_k : Between-cluster sum of squares (between-cluster dispersion).
 W_k : Within-cluster sum of squares (within-cluster dispersion).
 k : Number of clusters.
 n : Number of evaluated data points.

III. RESULTS AND DISCUSSION

This study compared the K-Means and DBSCAN algorithms for clustering academic reference-seeking behaviors among 150 respondents based on ten numerical variables. Principal Component Analysis (PCA) was applied for dimensionality reduction, and the clustering results were evaluated using the Silhouette Score, Calinski-Harabasz Index (CHI), and Davies-Bouldin Index (DBI).

A. Descriptive Statistics

Descriptive statistical analysis was conducted to provide an overview of the data characteristics of 150 respondents across the research variables. The analysis included the number of observations, mean, standard deviation, minimum value, quartiles, and maximum value to identify data patterns and variability before clustering. The descriptive statistics are presented in Table 2.

Table 2. Descriptive Statistics

	count	mean	std	min	25%	50%	75%	max
Source_Verification	150	4,07	2,61	1	3	3	5	15
Free_Sources	150	3,77	2,94	1	2	3	4	20
Source_Completeness	150	3,88	2,73	1	2	3	5	20
Monthly_Sources	150	12,47	14,52	1	4,25	9	15	120
Search_By_Sources	150	5,07	3,24	2	3	4	5	15
Comparison_Time	150	32,01	27,71	1	15	30	43,75	180
Library_Frequency	150	4,15	5,61	0	1	2	5	30
Activity_By_References	150	5,71	4,58	1	3	5	6	35
Manual_Time	150	48,1	64,45	2	20	30	60	480
AI_Time	150	20,75	18,52	1	10	15	30	120

The descriptive statistics reveal variations in data dispersion across the research variables, as indicated by their standard deviations, quartile ranges, and minimum and maximum values. Some variables exhibit wider distributions and maximum values substantially above their upper quartiles, indicating considerable variation in respondents' reference-searching behaviors. These differences in scale may cause variables with larger numerical ranges to dominate distance calculations, potentially resulting in clusters that are more influenced by scale differences than by the underlying patterns in the data. Therefore, data standardization is necessary before clustering to reduce this imbalance and ensure that each variable contributes more proportionally to the cluster formation process.

B. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was used to reduce ten numerical variables to a smaller set of principal components. The number of components was determined based on a cumulative explained variance threshold ranging from 75% to 87%. The calculated eigenvalues and explained variance percentages are presented in Table 3.

Table 3. PCA Result

Component	Eigenvalue	Explained Variance (%)	Cumulative Variance (%)
PC1	3,35	33,30	33,30
PC2	1,54	15,37	48,67
PC3	1,07	10,64	59,30
PC4	0,89	8,86	68,16
PC5	0,80	7,95	76,11
PC6	0,75	7,42	83,53

Based on Table 3, six principal components were selected because they accounted for approximately 83.53% of the data variance. This cumulative variance falls well within the recommended threshold range of 75% to 87%, confirming that the six selected components successfully capture the vast majority of the essential information from the ten original variables while filtering out minor noise and redundancy. By reducing the variables space from ten dimensions to six orthogonal principal components, the analysis effectively mitigates the 'curse of dimensionality'. This reduction ensures that subsequent distance and density calculations in K-Means and DBSCAN such as Euclidean distance metrics and neighborhood formations are not distorted by multicollinearity or disproportionate variables scaling, thereby producing more stable and reliable cluster structures. Thus, the data dimensions were reduced from ten to six new variables without reducing the number of 150 respondents. These six components were then used as input in the clustering process using the K-Means and DBSCAN algorithms.

C. K-Means Clustering Result

Clustering using the K-Means algorithm was performed by testing the number of clusters (K) from 2 to 8. The optimal number of clusters was determined based on three evaluation metrics: the Silhouette Score, the Calinski-Harabasz Index, and the Davies-Bouldin Index. The evaluation results for each value of K are shown in Table 4.

Table 4. Evaluation to Determine the Optimal K

K	Silhouette Score	CHI	DBI
2	0,4543	52,2985	1,4225
3	0,2526	43,8296	1,4888
4	0,2326	39,8987	1,5234
5	0,2382	38,6448	1,3892
6	0,2442	37,2026	1,1489
7	0,2410	36,8371	1,0873
8	0,2448	36,3839	1,1363

Based on Table 4, $K = 2$ yields the highest Silhouette Score of 0.4543 and the highest CHI of 52.2985, while the lowest DBI of 1.0873 is obtained at $K = 7$. However, $K = 7$ results in two clusters, each containing only one member. Therefore, $K = 2$ was selected because it outperformed the other values on two evaluation metrics and produced a distribution of cluster members that was more amenable to analysis.

Applying the K-Means algorithm with ($K = 2$) resulted in two clusters. The clustering results are visualized using the first two principal components from PCA, as these components allow the cluster distributions to be represented in a two-dimensional plot, as shown in Figure 2.

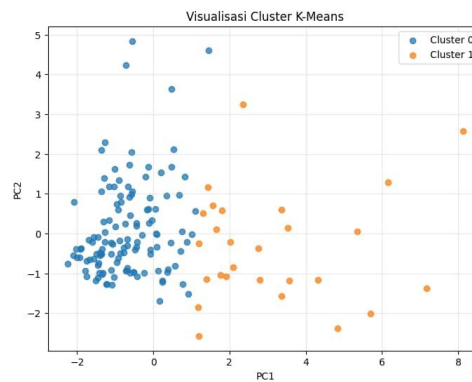


Figure 2. K-Means Clustering Visualization

Based on Figure 2, the K-Means visualization results show the formation of two clusters with distinct distribution patterns. Cluster 0, consisting of 123 respondents (82%), tends to be concentrated at lower PC1 values, while Cluster 1, with 27 respondents (18%), has a broader distribution toward positive PC1 values. Although some points appear close together, the two clusters exhibit distinct clustering patterns in the two-dimensional visualization.

D. DBSCAN Clustering Result

DBSCAN clustering was performed by determining the epsilon (ϵ) and minimum samples (min_samples) parameters. Candidate epsilon values were identified using a K-Distance Plot and evaluated through several parameter combinations. The evaluation considered the Silhouette Score, Davies-Bouldin Index (DBI), Calinski-Harabasz Index (CHI), number of clusters, and number of noise points. The results of the parameter evaluation are presented in Table 5.

Table 5. Epsilon Evaluation Test for DBSCAN

Epsilon	Min Samples	Number of Clusters	Noise	Silhouette Score	CHI	DBI
2,1	2	4	14	0,4547	15,5473	0,5834
2,2	2	3	14	0,4928	16,1362	0,5113
2,3	2	2	11	0,5469	15,3191	0,4818
2,4	2	2	10	0,5374	14,4719	0,4898
2,5	2	2	10	0,5374	14,4719	0,4898
2,6	2	2	10	0,5374	14,4719	0,4898
2,7	2	2	10	0,5374	14,4719	0,4898

Based on Table 5, the configuration ($\text{eps} = 2.3$) and ($\text{min_samples} = 2$) was selected because it achieved the highest Silhouette Score of 0.5469 and the lowest Davies-Bouldin Index of 0.4818 among the tested configurations. Although ($\text{eps} = 2.2$) yielded the highest Calinski-Harabasz Index and produced three clusters, it also resulted in a higher number of noise points. Therefore, ($\text{eps} = 2.3$) was selected based on the majority of evaluation metrics and its lower number of noise points (11 respondents).

Applying the DBSCAN algorithm with ($\text{eps} = 2.3$) and ($\text{min_samples} = 2$) resulted in two clusters and a set of noise points. Although clustering was performed using six principal components from PCA, the results are visualized using the first two principal components to represent the cluster distribution in a two-dimensional plot, as shown in Figure 3.

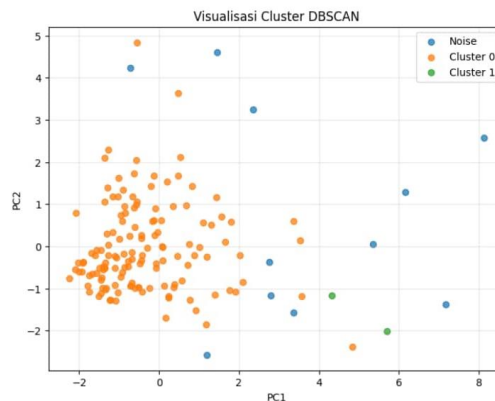


Figure 3. DBSCAN Clustering Visualization

Based on Figure 3, the DBSCAN visualization shows that the majority of respondents belong to Cluster 0, consisting of 137 respondents (91.33%), concentrated mainly in the left and central-left regions of the plot. Cluster 1 consists of 2 respondents (1.33%), located separately from the main group in the lower-right region. Additionally, 11 respondents (7.33%) were classified as noise and are scattered across various regions of the plot. These results illustrate that DBSCAN forms one dominant cluster, one small cluster, and identifies several data points that do not meet the density criteria for inclusion in a cluster. However, the DBSCAN results show a highly imbalanced cluster distribution. This imbalance indicates that the clustering results primarily describe the dominant pattern within the sample, while Cluster 1 is treated as an exploratory finding reflecting the characteristics of only two respondents and requiring further validation before it can be

considered representative of a broader student behavioral pattern.

E. Model Evaluation

The performance of K-Means and DBSCAN was evaluated using the Silhouette Score, Calinski-Harabasz Index (CHI), and Davies-Bouldin Index (DBI). The evaluation results for both clustering methods are presented in Table 6.

Table 6. Model Evaluation using CHI, DBI, and Silhouette Score

Method	Number of Clusters	Silhouette Score	CHI	DBI
K-Means	2	0,4543	52,2985	1,4225
DBSCAN	2	0,5469	15,3191	0,4818

Based on Table 6, DBSCAN produced a higher Silhouette Score (0.5469) and a lower Davies–Bouldin Index (0.4818) than K-Means, although the Calinski–Harabasz Index indicated better performance for K-Means. In the DBSCAN evaluation, 11 respondents categorized as noise were excluded from the internal validation metric calculations, so the evaluation included only the 139 respondents who were assigned to clusters. Consequently, the DBSCAN metric values reflect the quality of clustering for the data that met the density criteria, rather than for all 150 respondents. Based on its superior performance on two evaluation metrics and its ability to identify noise, DBSCAN was selected as the primary algorithm for the exploratory analysis in this study. Thus, the choice of DBSCAN is based on the evaluation criteria and the objectives of the analysis, not on the assumption that DBSCAN is superior to K-Means in all respects. The characteristics of the resulting clusters were then examined based on the study variables, as presented in Table 7.

Table 7. DBSCAN Cluster Characteristics Identification

Variable	Cluster 0	Cluster 1	Noise
Source_Verification	3,67	10,00	7,91
Free_Sources	3,44	15,00	5,91
Source_Completeness	3,53	10,00	7,18
Monthly_Sources	10,80	17,50	32,36
Search_By_Sources	4,64	12,50	9,00
Comparison_Time	31,69	20,00	38,18
Library_Frequency	3,58	2,50	11,55
Activity_By_References	5,05	12,50	12,64
Manual_Time	38,08	60,00	176,36
AI_Time	19,83	7,50	34,64

Based on Table 7, Cluster 0 (137 respondents) shows higher average values for Comparison_Time (31.69), Library_Frequency (3.58), and AI_Time (19.83) than Cluster 1, indicating longer reference comparison and AI-assisted searching times, as well as more frequent library use. Meanwhile, Cluster 1 (2 respondents) shows higher values for Source_Verification

(10.00), Free_Sources (15.00), Source_Completeness (10.00), and Manual_Time (60.00) than Cluster 0. However, these characteristics are limited to two respondents and cannot be generalized. The 11 observations classified as noise exhibit distinct characteristics from those of Cluster 0 and Cluster 1, with some average values higher than those of both clusters. These observations are treated separately because they do not meet DBSCAN's specified density criteria for assignment to either cluster. These findings describe behavioral differences without implying causal relationships.

IV. CONCLUSION

The clustering analysis of academic reference-seeking preferences and strategies among 150 respondents from various universities in Indonesia was conducted using K-Means and DBSCAN after reducing 10 numerical variables to six principal components through Principal Component Analysis (PCA), which retained 83.53% of the cumulative variance. DBSCAN was selected for further analysis because it achieved a higher Silhouette Score (0.5469) and a lower Davies–Bouldin Index (0.4818) than K-Means, although its Calinski–Harabasz Index was lower. Using an *eps* value of 2.3 and *min_samples* of 2, DBSCAN identified Cluster 0, comprising 137 respondents and characterized by higher AI usage time, library usage frequency, and source comparison time, and Cluster 1, comprising 2 respondents and characterized by higher source verification, use of free and diverse references, and manual search time. However, these characteristics are not generalized to the broader student population due to the very small cluster size. The remaining 11 respondents were classified as noise, exhibiting distinct patterns.

These clustering results can serve as a basis for universities to develop information literacy strategies tailored to students' reference-seeking behaviors. Such initiatives may include strengthening students' ability to verify source credibility, improving the use of libraries and digital references, and promoting the appropriate use of AI in academic information searches. Furthermore, these efforts can help universities provide more targeted academic support, enabling students to identify reliable sources and improve the effectiveness of their academic reference-seeking processes.

REFERENCES

- [1] A. A. Bushra and G. Yi, "Comparative analysis review of pioneering DBSCAN and successive density-based clustering algorithms," *IEEE Access*, vol. 9, pp. 87918–87935, 2021. <https://doi.org/10.1109/ACCESS.2021.3089036>
- [2] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, analysis of variants, and advances in the era of big data," *Information Sciences*, vol. 622, pp. 178–210, 2023. <https://doi.org/10.1016/j.ins.2022.11.139>
- [3] A. M. Ikotun, F. Habyarimana, and A. E. Ezugwu, "Cluster validity indices for automatic clustering: A comprehensive review," *Heliyon*, vol. 11, no. 2, Art. o.e41953, 2025. <https://doi.org/10.1016/j.heliyon.2025.e41953>
- [4] Asiah, N., Jatmiko, A., Erfayliana, Y., & Desky, H. (2025). Training in Application-Based Scientific Reference Management (Mendeley & Zotero) to Improve the Quality of Student Academic Writing. *Dedication: Journal of Community Service*, 9(2), 279–286. <https://doi.org/10.31537/dedication.v9i2.2591>
- [5] Azalia, C. M., Alifya, S., Lestari, A., Nadzif, M. R., Mulviani, A. D., & Nugraha, J. T. (2025). Students' Preferences for Accessing Information from the Internet Compared to Books. *EduTeach: Journal of Education & Learning Technology*, 6(2), 94–104 <https://ejurnal.umri.ac.id/index.php/eduteach/>
- [6] Hasan, B. M. S., & Abdulazeez, A. M. (2021). A review of principal component analysis algorithm for dimensionality reduction. *Journal of Soft Computing and Data Mining*, 2(1), 20–30. <https://doi.org/10.30880/jscdm.2021.02.01.003>
- [7] Hasan, Y. (2024). Measurement of the Silhouette Score and Davies-Bouldin Index in K-Means and DBSCAN Clustering Results. *Journal of Informatics and Applied Electrical Engineering (JIETET)*, 12(2).
- [8] Hastuti, U. R., & Ningrum, A. K. (2025). Trends in Information Source Usage Among UIN SAIKU Students. *Pustakaloka: Journal of Information and Library Studies*, 17(1), 1–16.
- [9] Ikotun, A. M., Habyarimana, F., & Ezugwu, A. E. (2025). Cluster validity indices for automatic clustering: A comprehensive review. *Heliyon*, 11(2), e41953. <https://doi.org/10.1016/j.heliyon.2025.e41953>
- [10] Mahardika, G. (2024). Clustering IoT Sensor Data Using the DBSCAN Algorithm. *Journal of Data World*, 1(1).
- [11] Mbonigaba, C., Vasuki, M., Sujatha, S., & Kumar, A. D. (2025). Principal Component Analysis for Simplifying Multivariate Financial Data in Portfolio Risk Analysis. *Brainae Journal of Business, Sciences and Technology*, 9(2), 171–179. <https://doi.org/10.5281/zenodo.14880988>
- [12] Mubarok, A., Khairunnisa, A. H., Murtanto, A. J., Putrie, A. C. N., Sjawaloedin, J. B. A., Simanjuntak, D. J. K., Danadhyaksa, H., Pratama, I. R., Basyary, M. F., Raditya, R., Harits, R. R., & Sumargi, W. A. (2026). Analysis of Limitations in Access to Academic Literature for Political Science Students at UPNVJ from the Perspective of the Right to Education. Project Article for a Required Course in the UPN Veteran Jakarta Curriculum
- [13] Nalole, S. R., Runtuwene, A., & Rawis, S. M. (2026). Information-Seeking Behavior of Sam Ratulangi University Students Toward Online Journals. *Media Informasi*, 35(1), 1–10. <https://doi.org/10.22146/mi.v35i1.26392>

- [14] S. Marukatat, "Tutorial on PCA, Approximate PCA, and Approximate Kernel PCA," *Artificial Intelligence Review*, vol. 56, no. 6, pp. 5445–5477, 2023. <https://doi.org/10.1007/s10462-022-10297-z>
- [15] Wahyuni, T., Faisal, M., Kuba, M. S. A. S., Syatriana, E., & Al Haikal, M. (2026). Climate-Driven Disease Outbreak Anomaly Detection Using PCA, DBSCAN, and Isolation Forest. *Joutica*, 11(02), 175-192.