

Laboratory Assistant Performance Analysis At STMIK Widya Cipta Dharma Using K-Means

Alisa Nuraini¹, Heny Pratiwi^{2*}, Ahmad Fajri³

¹Department of Informatics Engineering, STMIK Widya Cipta Dharma, Indonesia

^{2,3}Department of Information Systems, STMIK Widya Cipta Dharma, Indonesia

* Corresponding author:

Email: henypratiwi@wicida.ac.id

Abstract.

Laboratory assistants play a crucial role in supporting computer science practicum learning; however, conventional evaluation methods based on aggregate average scores are unable to identify hidden performance patterns and characteristics. This study aims to analyze the patterns and characteristics of laboratory assistant performance at STMIK Widya Cipta Dharma using an Educational Data Mining approach based on the K-Means clustering algorithm. This study employed a quantitative research method using questionnaire data collected from 178 students with ten performance indicators measured using a five-point Likert scale. The analysis process consisted of several stages, including data collection, data preprocessing through data cleaning and attribute selection, instrument testing using Pearson Product Moment validity and Cronbach's Alpha reliability analysis, Z-Score normalization to standardize indicator values, K-Means clustering using Euclidean distance calculations, evaluation of optimal cluster numbers using the Elbow Method, Silhouette Score, and Davies-Bouldin Index (DBI), and Principal Component Analysis (PCA) visualization for cluster interpretation. The results showed that the instrument was valid and reliable, and the K-Means algorithm produced three performance clusters consisting of Low Performance (20.8%), Medium Performance (55.1%), and High Performance (24.1%) groups.

Keywords: Laboratory Assistant, Performance Evaluation, K-Means Clustering, Educational Data Mining and Student Evaluation.

I. INTRODUCTION

Laboratory assistants play an important role in supporting practical learning activities in higher education, especially in computer science programs. Their performance influences students' learning experiences and the effectiveness of laboratory activities. Therefore, evaluating laboratory assistant performance is necessary to improve practical learning quality[1].

At STMIK Widya Cipta Dharma, evaluation is conducted through student assessments covering technical ability, communication, professionalism, and learning support. However, conventional evaluations based only on average scores cannot identify hidden performance patterns, differences among assistants, or specific improvement areas[2].

Data mining techniques, particularly K-Means clustering, provide an approach to analyze evaluation data by grouping objects based on similarity characteristics. Previous studies have applied clustering for performance analysis, but many focus only on group formation without considering instrument validity, reliability, and detailed interpretation of cluster characteristics.

Previous studies in educational data mining have mainly focused on student performance prediction and academic clustering. However, the application of clustering approaches for laboratory assistant performance evaluation remains limited, particularly studies that integrate questionnaire validation, clustering validation, and interpretation of cluster characteristics. Therefore, this study proposes an integrated framework combining measurement evaluation and clustering analysis to identify laboratory assistant performance patterns.

This research proposes an integrated framework combining questionnaire validation, reliability testing, data preprocessing, K-Means clustering, cluster validation, and performance characteristic analysis. The objectives of this study are to evaluate the measurement instrument, identify laboratory assistant performance groups, and analyze the characteristics of each cluster to provide data-driven information for performance improvement strategies[3].

II. METHODS

2.1 Research Design

This study employed a quantitative research approach combined with an unsupervised machine learning method to analyze laboratory assistant performance patterns. The research was conducted to identify groups of laboratory assistants based on similarities in student evaluation responses using the K-Means clustering algorithm[4].

The research process consisted of several sequential stages to transform questionnaire-based evaluation data into laboratory assistant performance clusters. First, student evaluation data were collected from laboratory assistant assessments. The collected dataset consisted of 178 responses measured using ten performance indicators. The dataset was then cleaned by checking incomplete responses and duplicate records to ensure data consistency. After feature selection, the ten performance indicators used as clustering variables were evaluated through validity and reliability testing to ensure that the questionnaire items were appropriate for analysis. Furthermore, the validated data were normalized using Z-score normalization before being processed using the K-Means clustering algorithm. The optimal number of clusters was determined using the Elbow Method, Silhouette Score, and Davies-Bouldin Index. Finally, cluster validation, PCA visualization, and cluster characteristic analysis were performed to interpret laboratory assistant performance patterns[5].

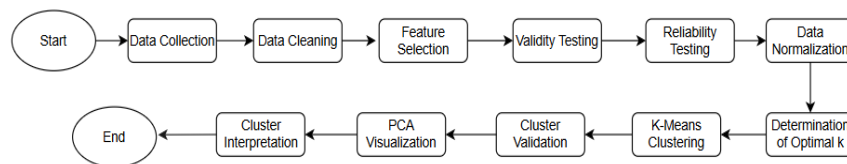


Fig. 1. Research Workflow

As shown in Figure 1, the research process started with collecting laboratory assistant evaluation data. The collected data were cleaned to remove inconsistent records, followed by feature selection to determine relevant performance indicators used as clustering variables. The selected indicators were evaluated through validity and reliability testing before normalization. The normalized dataset was then processed using K-Means clustering after determining the optimal number of clusters. Finally, cluster validation, PCA visualization, and cluster interpretation were performed to analyze laboratory assistant performance patterns[6].

The proposed framework aims to overcome limitations of conventional evaluation methods that only rely on average scores by identifying hidden performance patterns and characteristics among laboratory assistants

2.2 Data Collection and Questionnaire Instrument

The dataset used in this study was obtained from questionnaire responses collected from students of STMIK Widya Cipta Dharma who had direct interaction with laboratory assistants during practicum activities. A total of 178 valid responses were used for the analysis process.

The indicators, measured using a five-point Likert scale, represented several performance aspects, including technical competence, communication ability, professionalism, and learning support.

The questionnaire indicators used in this research are presented in Table 1.

Table 1. Questionnaire Indicators

Code	Indicator
Q1	Technical mastery
Q2	Punctuality
Q3	Communication clarity
Q4	Patience
Q5	Proactiveness
Q6	Additional tips and guidance
Q7	Composure under pressure
Q8	Professionalism and fairness

Q9	Grading objectivity
Q10	Time management

2.3. Data Preprocessing

Data preprocessing was performed before applying the clustering algorithm to improve data quality and ensure that the dataset was suitable for analysis. The preprocessing stage consisted of data cleaning and data preparation[7],[8].

First, missing values and duplicate responses were examined to prevent inaccurate results caused by incomplete or repeated data. After the cleaning process, the valid questionnaire responses were retained for subsequent analysis. The ten questionnaire indicators were selected as clustering attributes, while irrelevant attributes such as respondent identification information were excluded.

Furthermore, the questionnaire scores were transformed into a numerical format based on the Likert scale responses. Since K-Means clustering uses distance-based calculations, the data were standardized using Z-score normalization to ensure that all indicators contributed equally during the clustering process[9].

The Z-score normalization formula is defined as:

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

where x represents the original value, μ represents the mean value, and σ represents the standard deviation. In this formula, x represents the original questionnaire score obtained from each respondent, μ represents the average value calculated from all respondents, and σ represents the standard deviation that describes the distribution of the data.

In this study, Z-Score normalization was performed using the StandardScaler function from the Scikit-learn library in Python. The transformation converts each indicator value into a standardized value with a mean of 0 and a standard deviation of 1.

The mean value and standard deviation of each indicator were calculated based on 178 valid respondents before the standardization process. The mean value of the Q1 indicator (Technical Mastery) was calculated using the following formula:

$$\mu = \frac{\sum x}{n} \quad (2)$$

where:

1. μ = mean value of the indicator
2. $\sum x$ = total score of all respondents
3. n = number of respondents

For the Q1 indicator, the total score from 178 respondents was 699. The mean calculation was performed as follows:

$$\mu = \frac{699}{178} \quad (3)$$

$$\mu = 3.93 \quad (4)$$

The result indicates that the average score of the Q1 indicator from all respondents was 3.93. This mean value was used together with the standard deviation value in the Z-Score normalization process.

To illustrate the Z-Score normalization process, an example calculation was performed using the Q1 indicator (Technical Mastery) from the dataset consisting of 178 valid respondents.

For example, one respondent obtained an original Q1 score of 4. Based on the calculation of 178 respondents, the mean value of Q1 was 3.93, and the standard deviation was 0.714 :

$$Z = \frac{x - \mu}{\sigma} \quad (5)$$

$$Z = \frac{4 - (3.93)}{(0.714)} \quad (6)$$

$$Z = \frac{0.07}{0.71} \quad (7)$$

$$Z = 0.098 \quad (8)$$

$$Z \approx 0.10 \quad (9)$$

The calculation result shows that the respondent's Q1 value obtained a Z-Score of 0.098 or approximately 0.10. This indicates that the respondent's technical mastery score is 0.10 standard deviations above the average value of all respondents. The same normalization process was applied to all ten questionnaire indicators (Q1–Q10) using 178 valid respondents before being used in the K-Means clustering process. **2.4.**

Validity and Reliability Testing

Before performing clustering analysis, the questionnaire instrument was evaluated to ensure the validity and consistency of the measurement items

The validity test was conducted using Pearson Product Moment correlation through the corrected item-total correlation approach. Each questionnaire item was correlated with the total score excluding the corresponding item. An item was considered valid when the calculated correlation coefficient (r-count) exceeded the critical value (r-table). Based on 178 respondents and a significance level of 5% ($\alpha = 0.05$), the degree of freedom was calculated as $df = n - 2 = 176$, resulting in an r-table value of 0.147[10].

The reliability of the questionnaire instrument was evaluated using Cronbach's Alpha coefficient. Cronbach's Alpha measures the internal consistency among questionnaire items. A higher alpha coefficient indicates stronger reliability of the instrument. In this study, Cronbach's Alpha values above 0.90 were categorized as excellent reliability, indicating high internal consistency among questionnaire items[11].

2.5 K-Means Clustering Algorithm

After preprocessing and normalization, the dataset was analyzed using the K-Means clustering algorithm. K-Means is an unsupervised learning method that groups data objects into clusters based on similarity between observations[12].

The clustering process begins by determining the number of clusters (k), initializing cluster centroids, calculating the distance between each data point and centroid using Euclidean distance, and assigning each observation to the nearest cluster. The centroid position is updated iteratively until the clustering process reaches convergence.

The Euclidean distance calculation is defined as:

$$d(x,y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (10)$$

where:

1. $d(x,y)$ = distance between respondent data and centroid
2. x_i = Z-Score value of respondent data on indicator i
3. y_i = centroid value of cluster on indicator i
4. n = number of indicators used in clustering (10 indicators)[13].

The value of x represents the standardized performance evaluation values of each respondent after the Z-Score normalization process, while the value of y represents the centroid value generated from the K-Means clustering process. The centroid value represents the center point of each cluster calculated from the average standardized values of observations belonging to that cluster.

The Euclidean distance calculation uses the standardized values obtained from the Z-Score normalization process. Each respondent is represented by ten standardized attributes (Q1–Q10), which are compared with the centroid values generated by the K-Means algorithm.

The respondent data were obtained from the standardized dataset after the Z-Score normalization process using 178 valid respondents. The Euclidean distance calculation uses the standardized values obtained from the Z-Score normalization process. Each respondent is represented by ten standardized attributes (Q1–Q10), which are compared with the centroid values generated by the K-Means algorithm.

For example, one respondent from the 178 valid respondents obtained the following standardized values:

The respondent data were obtained from the standardized dataset after the Z-Score normalization process using 178 valid respondents.

$$x = (x_1, x_2, x_3, \dots, x_{10}) \quad (11)$$

$$y = (y_1, y_2, y_3, \dots, y_{10}) \quad (12)$$

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (13)$$

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2 + (x_4 - y_4)^2 + (x_5 - y_5)^2 + (x_6 - y_6)^2 + (x_7 - y_7)^2 + (x_8 - y_8)^2 + (x_9 - y_9)^2 + (x_{10} - y_{10})^2} \quad (14)$$

$$x = (x_1, x_2, x_3, \dots, x_{10})$$

$$x = (-2.993, -2.992, 0.139, 0.090, 0.194, 1.383, -2.228, -1.897, -2.073, 1.168) \quad (15)$$

$$y = (y_1, y_2, y_3, \dots, y_{10}) \quad (16)$$

$$y = (-1.204, -1.032, -1.110, -1.205, -1.191, -0.843, -1.268, -1.247, -1.082, -1.036) \quad (17)$$

$$d(x, y) = \sqrt{((x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2 + (x_4 - y_4)^2 + (x_5 - y_5)^2) + ((x_6 - y_6)^2 + (x_7 - y_7)^2 + (x_8 - y_8)^2 + (x_9 - y_9)^2 + (x_{10} - y_{10})^2)} \quad (18)$$

$$d(x, y) = \sqrt{(24.371)} \quad (19)$$

$$d(x, y) = 4.936 \quad (20)$$

The obtained distance value indicates the similarity level between the respondent data point and Cluster 1 centroid. A smaller distance value indicates that the respondent has a higher similarity with that cluster. The same calculation was performed for all cluster centroids, and each respondent was assigned to the cluster with the smallest distance value.

The K-Means algorithm was implemented using Python with the Scikit-learn library. Multiple centroid initialization was applied to improve clustering stability

2.6. Cluster Validation

The optimal number of clusters was determined before implementing the final K-Means model. Two evaluation approaches were used: the Elbow Method and Silhouette Score[14].

The Elbow Method evaluates the Within-Cluster Sum of Squares (WCSS) value for different numbers of clusters. The optimal cluster number is selected based on the point where additional clusters provide limited improvement[15].

The Silhouette Score was used to measure cluster separation quality by evaluating the similarity of each data point within its own cluster compared with other clusters. A higher silhouette value indicates better clustering performance[16].

In addition, the Davies-Bouldin Index (DBI) was considered to evaluate cluster compactness and separation, where lower DBI values indicate better clustering quality[17].

2.7. PCA Visualization

Principal Component Analysis (PCA) was applied to visualize the clustering results in a two-dimensional space. PCA reduces the dimensionality of the dataset while preserving important variance information[18].

The PCA visualization was used to observe the distribution patterns of the generated clusters and provide a clearer interpretation of differences between laboratory assistant performance groups[19]

Data preprocessing, clustering analysis, and visualization were performed using Python programming with relevant machine learning libraries. The questionnaire validity and reliability tests were conducted using IBM SPSS Statistics before the clustering process[20].

III. RESULT AND DISCUSSION

3.1. Data Cleaning Results

The data preprocessing stage was conducted to ensure the quality and consistency of the questionnaire dataset before clustering analysis. The dataset consisted of ten performance indicators collected from student evaluation responses. The data inspection process included checking missing values and duplicate records.

The analysis showed that all questionnaire indicators contained complete responses from 178 respondents. No missing values were identified in the dataset. Furthermore, duplicate checking was performed to ensure that each response represented a unique evaluation record. After the cleaning process, 178 valid responses were retained and used for further analysis, as summarized in Table 2.

Table 2. Dataset Cleaning Results

Process	Result
Total responses	178
Missing values	0
Duplicate records	0
Final dataset	178

3.2. Validity and Reliability test Results

The validity test was conducted using the Pearson Product Moment correlation method through corrected item-total correlation. Each questionnaire item was correlated with the total score of the remaining indicators excluding the tested item. In this study, 178 valid responses were analyzed. The degree of freedom was determined using $df=N-2$ because Pearson correlation evaluates the relationship between two variables, namely the item score (X) and the total score (Y). Therefore, the degree of freedom was calculated as 176. Using a significance level of 5% ($\alpha=0.05$), the r-table value obtained was 0.147.

$$df = N - 2 \quad (21)$$

$$df = 178 - 2 = 176 \quad (22)$$

$$\alpha = 0.05 \quad (23)$$

To illustrate the validity testing process, the calculation of the Q1 indicator (Technical Mastery) was used as an example. The Pearson Product Moment correlation was calculated by correlating the score of each questionnaire item with the total score of respondents. In this process, the Q1 score acted as the X variable, while the total questionnaire score acted as the Y variable. The correlation coefficient was calculated based on the responses from 178 valid respondents.

$$r_{(xy)} = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}} \quad (24)$$

where:

1. $r_{(xy)}$ =Pearson correlation coefficient
2. n=number of respondents
3. x=score of questionnaire item
4. y=total score of questionnaire

$$X = Q1 \quad (25)$$

$$Y = Q1 + Q2 + Q3 + Q4 + Q5 + Q6 + Q7 + Q8 + Q9 + Q10 \quad (26)$$

Based on the SPSS corrected item-total correlation output, Q1 obtained a correlation coefficient of 0.7098.

$$r_{count,Q1} > r_{table}(27)$$

$$r_{count,Q1} = 0.7098 > r_{table} = 0.147(28)$$

Therefore, Q1 was declared valid because the calculated correlation coefficient exceeded the critical r-table value

Table 3. Validity Testing Results

Item	r-count	Result
Q1	0.7098	Valid
Q2	0.6247	Valid
Q3	0.7276	Valid
Q4	0.7666	Valid
Q5	0.7268	Valid
Q6	0.5693	Valid
Q7	0.7483	Valid
Q8	0.7646	Valid
Q9	0.7199	Valid
Q10	0.6277	Valid

The overall Cronbach's Alpha for the ten-item instrument was 0.918, which falls into the excellent reliability category, well above the commonly accepted minimum threshold of 0.60-0.70 for social science

instruments. This result confirms that the questionnaire produced consistent and dependable measurements, providing a solid basis for the subsequent clustering analysis.

3.3. Determination of Optimal Cluster Number

The optimal number of clusters was evaluated before implementing the final K-Means clustering model. The evaluation was performed using the Elbow Method, Silhouette Score, and Davies-Bouldin Index (DBI). The number of clusters was tested from $k=2$ to $k=8$ to compare different grouping possibilities and identify the most appropriate clustering structure.

The Elbow Method evaluates the Within-Cluster Sum of Squares (WCSS) value for each tested cluster number. WCSS measures the total squared distance between each data point and its assigned cluster centroid. The WCSS calculation is defined as:

$$WCSS = \sum_{j=1}^k \sum_{x_i \in C_j} \|x_i - \mu_j\|^2 \quad (29)$$

Where:

1. $xWCSS$ = Within-Cluster Sum of Squares
2. K = number of clusters
3. x_i = respondent data point
4. C_j = the j -th cluster
5. μ_j = centroid value of cluster j -th cluster

The WCSS values were calculated by applying the K-Means algorithm repeatedly using different cluster numbers ranging from $k=2$ to $k=8$. For each value of k , the algorithm generated different centroid positions and calculated the total distance between respondents and their corresponding centroids. For each tested value of k , the K-Means algorithm initialized the centroid positions and assigned each respondent to the nearest cluster based on Euclidean distance. After the clustering process reached convergence, the total squared distance between each respondent and its assigned centroid was calculated as the WCSS value. For example, the WCSS value at $k=2$ represents the total squared distance between observations and their assigned centroid when the dataset is divided into two clusters. The resulting WCSS values were visualized using the Elbow Method to observe the reduction pattern of within-cluster variation.

The optimal cluster number was then determined based on the Elbow Method and supported by clustering evaluation metrics. The Silhouette Score evaluation was performed for cluster values ranging from $k=2$ to $k=8$. The Silhouette Score measures the similarity of each observation within its assigned cluster compared with observations in other clusters. A higher Silhouette Score indicates better-defined cluster separation.

The Silhouette Score measures the similarity of each observation within its assigned cluster compared with other clusters. A higher Silhouette Score indicates better separation between clusters.

two clusters. The same calculation process was repeated for $k=3$ until $k=8$ to observe the reduction pattern of within-cluster variation. The cluster range from $k=2$ to $k=8$ was selected as an evaluation range to identify the most appropriate number of clusters. The value of $k=2$ represents the minimum meaningful clustering scenario, while higher cluster numbers were tested to observe whether

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (30)$$

where:

1. $a(i)$ = average distance between observation i and other observations in the same cluster
2. $b(i)$ = minimum average distance between observation i and observations in other clusters

The Silhouette Score evaluation showed variations among different cluster numbers from $k=2$ to $k=8$. Although $k=2$ produced a higher Silhouette Score, $k=3$ was selected because it provided better interpretability by representing Low Performance, Medium Performance, and High Performance categories aligned with the research objective. The final clustering result obtained a Silhouette Score of 0.2667 and a Davies-Bouldin Index of 1.244, indicating the quality of cluster separation and compactness.

$$DBI = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right) \quad (31)$$

Where:

1. DBI= Davies-Bouldin Index value
2. k = number of clusters
3. S_i = average distance between observations and the centroid of cluster C_i
4. S_j = average distance between observations and the centroid of cluster j
5. M_{ij} = distance between the centroids of cluster ii and cluster j
6. $\text{Max}_{j \neq i}$ = maximum similarity ratio between cluster ii and other clusters

In addition, the Davies-Bouldin Index (DBI) was considered to evaluate cluster compactness and separation, where lower DBI values indicate better clustering quality.

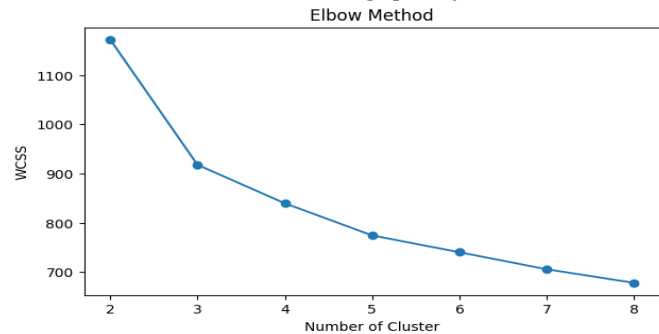


Fig. 2. Elbow Method Evaluation

Based on the Elbow Method graph, the WCSS value decreases significantly until k=3 and becomes less significant afterward. Therefore, k=3 was selected as the optimal number of clusters for identifying laboratory assistant performance patterns.

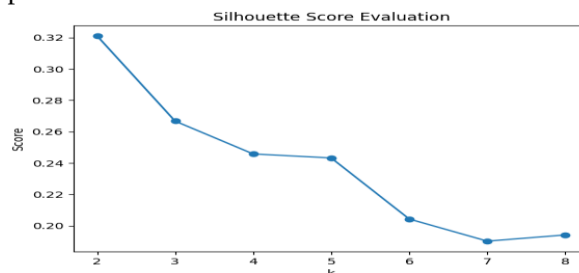


Fig. 3. Silhouette Score Evaluation

Based on the Silhouette Score evaluation graph, the values vary across different cluster numbers from k=2 to k=8. Although k=2 achieved the highest score, k=3 was selected because it provides meaningful performance categories: Low Performance, Medium Performance, and High Performance, which align with the research objective.

The evaluation results indicate that three clusters provide a more interpretable grouping structure for laboratory assistant performance analysis.

3.4. K-Means Clustering Results

The K-Means clustering algorithm was applied using three clusters. The clustering results divided 178 laboratory assistant evaluation records into three performance groups, which are detailed in Table 6.

The cluster labels (Low Performance, Medium Performance, and High Performance) were assigned based on the average centroid values of each cluster, where lower centroid values represented lower performance levels and higher centroid values represented higher performance levels.

Table 4. K-Means Clustering Results

Cluster	Number Of Data	Percentage	Category
Cluster 0	98	55.1%	Medium Performance
Cluster 1	37	20.8%	Low Performance
Cluster 2	43	24.1%	High Performance

The clustering results show that the generated groups have different performance profiles. Further analysis of each cluster characteristic is presented in the following section based on the average score of each performance indicator.

These clustering results indicate that laboratory assistant performance showed variations across the generated groups, with differences observed in several evaluation dimensions.

The characteristics of each cluster were analyzed based on the average score of each performance indicator to identify indicators distinguishing laboratory assistant performance groups, as presented in Table 7.

Table 5. Cluster Performance Characteristics Based on Evaluation Indicators

Indicator	Low Performance	Medium Performance	High Performance
Technical mastery	2.676	3.867	4.628
Punctuality	2.892	3.878	4.767
Communication clarity	2.838	3.816	4.884
Patience	2.865	3.929	4.814
Proactiveness	2.919	3.918	4.791
Additional tips and guidance	2.568	3.449	4.674
Composure under pressure	2.946	4.000	4.744
Professionalism and fairness	2.622	3.827	4.814
Grading objectivity	2.784	3.898	4.837
Time management	2.865	4.010	4.674

The cluster labels were determined based on the centroid values generated by the K-Means algorithm. The centroid represents the average standardized performance value of respondents across ten performance indicators, which were used to classify clusters into Low Performance, Medium Performance, and High Performance.

The centroid values presented in Table 5 were used as the basis for determining the performance categories of each cluster.

$$CM_j = \frac{\sum_{i=1}^n C_{ij}}{n} \quad (32)$$

$$CM_{High} = \frac{4.6279 + 4.7674 + 4.8837 + 4.8140 + 4.7442 + 4.3721 + 4.7674 + 4.8140 + 4.8372 + 4.6977}{10} \quad (33)$$

$$CM_{High} = 4.733 \quad (34)$$

Where:

1. C_{ij} = centroid value of indicator i in cluster j
2. n = number of indicators used in clustering (10 indicators)
3. CM_j = mean centroid value of cluster j

The same calculation procedure was applied to the Low Performance and Medium Performance clusters.

The calculated cluster mean values were 2.797 for Low Performance, 3.851 for Medium Performance, and 4.733 for High Performance.

These values indicate that the High Performance cluster had the highest average centroid value across all indicators, while the Low Performance cluster showed the lowest average centroid value.

The cluster with the lowest centroid mean value (2.797) was categorized as Low Performance, while the highest centroid mean value (4.733) was categorized as High Performance. The intermediate centroid value (3.851) was categorized as Medium Performance.

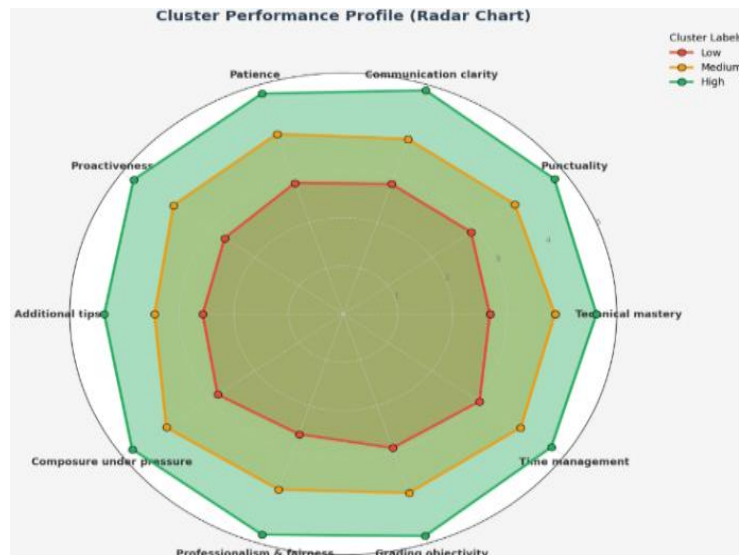


Fig. 4. Cluster Performance Profile (Radar Chart)

The visualization shows that the High Performance cluster consistently achieved higher scores across all indicators compared with other clusters. Meanwhile, the Low Performance cluster demonstrated lower scores, particularly in additional tips and guidance, professionalism and fairness, and technical mastery. The Medium Performance cluster showed intermediate characteristics between the two groups.

3.5.1. Low Performance Cluster

The Low Performance cluster consisted of 37 evaluation records (20.8%). This cluster showed the lowest average scores across all performance indicators. The lowest values were observed in additional tips and guidance (2.568), professionalism and fairness (2.622), and technical mastery (2.676). These findings indicate that improvement programs for this group should focus on strengthening technical capability, professional behavior, and student learning support.

3.5.2. Medium Performance Cluster

The Medium Performance cluster represented the largest group with 98 evaluation records (55.1%). This group demonstrated relatively stable performance across indicators. The highest scores were observed in time management (4.010) and composure under pressure (4.000). However, additional learning guidance remained an area requiring improvement.

3.5.3. High Performance Cluster

The High Performance cluster consisted of 43 evaluation records (24.1%). This cluster demonstrated the highest scores in communication clarity (4.884), grading objectivity (4.837), and professionalism and fairness (4.814). These characteristics indicate that effective communication, fair assessment practices, and professional behavior were among the indicators showing higher scores in the High Performance cluster.

The characteristics identified in these clusters indicate that technical competence should be complemented by communication skills and fair assessment practices, as these indicators showed higher scores in the High Performance cluster. Furthermore, the consistent appearance of "additional tips and guidance" as a low-scoring indicator in both Low and Medium Performance clusters suggests an area requiring improvement in proactive student support and targeted training interventions.

3.6. PCA Visualization

Table 6. PCA Explained Variance

Component	Explained Variance (%)
PC1	58.49
PC2	8.88
Total Variance	67.37

Principal Component Analysis (PCA) was applied to visualize the distribution of the generated clusters in a two-dimensional space. The first principal component explained 58.49% of the total variance, while the second principal component explained 8.88%. Together, both components represented 67.37% of the data variance.

The PCA visualization illustrates the distribution pattern of the three K-Means clusters based on laboratory assistant performance indicators, as shown in Figure 5. The visualization shows the distribution pattern of the generated clusters, providing additional insight into the separation characteristics and performance differences among the Low Performance, Medium Performance, and High Performance clusters.

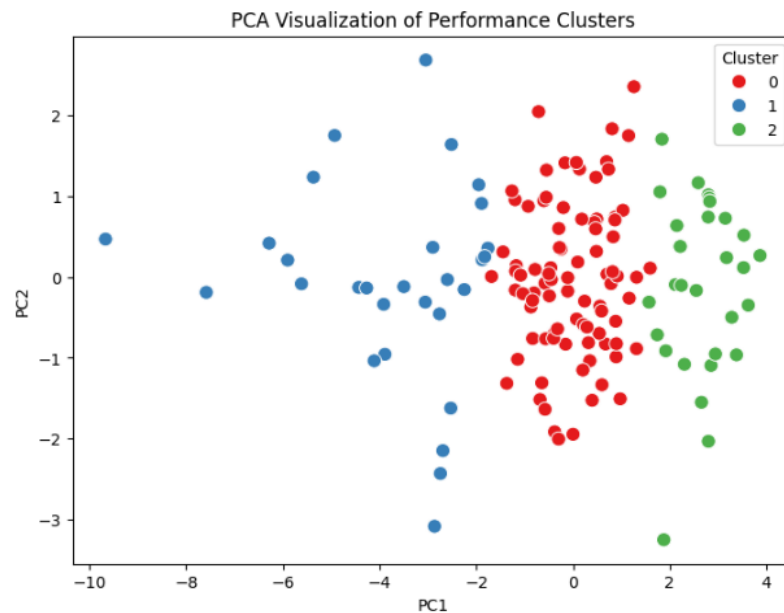


Fig. 5. PCA Visualization of Laboratory Assistant Performance Clusters

The PCA visualization provides a two-dimensional representation of the clustering results and provides additional visualization of performance differences among the three laboratory assistant groups

3.7. Research Limitations

This study has several limitations. First, the evaluation data were obtained from student perceptions, which may contain subjective judgments. Second, the dataset was collected from a single institution, therefore the findings may not fully represent laboratory assistant performance in other higher education environments. Future studies may involve multiple institutions and integrate additional objective indicators such as practicum outcomes or laboratory activity records.

IV CONCLUSION

This study demonstrates that K-Means clustering can be applied as an analytical approach to identify laboratory assistant performance patterns beyond conventional average-based evaluation. The clustering process generated three performance groups consisting of Low Performance (20.8%), Medium Performance (55.1%), and High Performance (24.1%). The identified clusters provide information regarding different performance characteristics, communication clarity, grading objectivity, and professionalism were indicators with higher scores in the High Performance group.. These findings can support targeted improvement programs and provide data-driven information for laboratory assistant management.

V. ACKNOWLEDGMENTS

The authors would like to thank STMIK Widya Cipta Dharma for providing support and facilities during this research. The authors also express their appreciation to all students who participated in the questionnaire evaluation process and contributed valuable data for this study.

REFERENCES

- [1] M. Gagolewski, M. Bartoszek, and A. Cena, "Are cluster validity measures (in)valid?," *Inf. Sci. (Ny)*, vol. 581, pp. 620–636, 2021.
- [2] I. Arfyanti, "Implementasi Data Mining dengan Menerapkan Algoritma K-Means Clustering untuk Memberikan

- Rekomendasi Jurusan Kuliah Bagi Mahasiswa Baru,” *J. Media Inform. Budidarma*, 2024.
- [3] M. F. Azmi, “K-Means Clustering with Elbow Method and Validity Indices for Classifying Student Academic Achievement,” *J. Tek. Inform.*, 2026.
- [4] S. P. Ranguti and M. Fakhriza, “Implementasi Algoritma K-Means Clustering untuk Penilaian Kinerja Anggota Organisasi,” 2025.
- [5] I. Darmayanti, “Improving K-Means Clustering Accuracy for Academic Success Investigation with Extreme Gradient Boosting Algorithm,” 2024.
- [6] E. Kartikadarma, “Educational Data Mining to Improve Decision Support on the Ratio of Students and Study Groups in Indonesia Using K-Means Method,” *Elem. Educ. Online*, 2021.
- [7] Y. Elda, “Klasterisasi Penempatan Siswa yang Optimal untuk Meningkatkan Nilai Rata-Rata Kelas Menggunakan K-Means,” *J. Inf. dan Teknol.*, 2021.
- [8] N. Yati, “Performance Analysis of The K-Means Algorithm on Intensity of Science Laboratory Use and Science Skills,” 2024.
- [9] N. Sitepu, “K-Means Clustering of Student Mid-Term and Final Exam Data,” *J. Komput. Indones.*, 2024.
- [10] A. Setiawan, “Penerapan Algoritma Clustering K-Means Data Mining dalam Pengelompokan Mahasiswa Penerima Beasiswa,” *J. Comput. Syst. Informatics*, 2024.
- [11] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, “K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data,” *Inf. Sci. (Ny)*, vol. 622, pp. 178–210, 2023.
- [12] C. Romero and S. Ventura, “Educational data mining and learning analytics: An updated survey,” *WIREs Data Min. Knowl. Discov.*, 2024.
- [13] E. Schubert, “Stop using the elbow criterion for k-means and how to choose the number of clusters instead,” *ACM SIGKDD Explor. Newsl.*, 2023.
- [14] B. Albreiki, N. Zaki, and H. Alashwal, “A systematic literature review of student performance prediction using machine learning techniques,” *Educ. Sci.*, 2021.
- [15] I. T. Utami, “K-Means cluster count optimization with Silhouette Index validation and Davies Bouldin Index,” *BAREKENG*, 2023.
- [16] A. R. Almasri, “Educational data mining for predicting student performance: A systematic review,” *Educ. Inf. Technol.*, 2022.
- [17] M. Auva, “Klasifikasi Kepuasan Mahasiswa Terhadap Evaluasi Kinerja Asisten Laboratorium di STMIK Kaputama Menggunakan Metode K-Means,” *PRINSIP*, 2025.
- [18] M. Wu, “K-Means Clustering-Based Feature Generation for Student Performance Prediction,” *ICCK Trans. Educ. Data Min.*, 2026.
- [19] R. Liu, “Data Analysis of Educational Evaluation Using K-Means Clustering Method,” *Comput. Intell. Neurosci.*, 2022.
- [20] B. Harpad, M. Fahmi, P. Pahrudin, and R. Andrea, “Pengelompokan Siswa Layak Penerima Beasiswa dengan Menerapkan Algoritma K-Means Clustering Data Mining,” 2024.