Available online at www.coins-indocompt.org

COINS JOURNAL
Computing and Information System Journal

COmputing and INformation Systems Journal

IndoCompt PUBLISHER

E-ISSN: 3109-3248

Vol. 2, No. 1, April, 2026

Clustering of Lecturer Performance Based on Internal Data and Questionnaire Using the X-Means Algorithm

Moh. Nur Ali Afendi[a*], Rosita Yanuarti[b], Qurrota A'yun[c]

Muhammadiyah University of Jember, Jember 68121, Indonesia

mochali916@gmail.com

Received 27 Pebruary 2026; Revised 03 Mart 2026; Accepted 18 Mart 2026; Published 01 April 2026

Abstract

Lecturer performance evaluation is essential for improving the quality of higher education. In the Informatics Engineering Study Program at Universitas Muhammadiyah Jember, assessments are still conducted manually, leading to inefficiency and subjectivity. This study aims to cluster lecturer performance based on internal data—credit load, publications, attendance, and students' average grades—and student perceptions collected through questionnaires, using the X-Means Clustering algorithm. X-Means was chosen because it can automatically determine the optimal number of clusters using the Bayesian Information Criterion (BIC) and log-likelihood. The research involved data collection, cleaning, normalization with the Min-Max method, implementation of X-Means, and visualization using Principal Component Analysis (PCA). The results identified five distinct clusters: high-performing lecturers, lecturers with high publication output, research-active lecturers with low student grades, low-performing lecturers, and stable-performing lecturers. These findings provide an objective view of lecturer performance variations and can serve as a basis for developing effective strategies for lecturer improvement and continuous performance evaluation.

Keywords: Lecturer Performance, Internal Data, Student Perception, X-Means, Clustering

1. Introduction

Lecturers are key human resources in higher education, responsible for advancing knowledge through teaching, research, and community service. Therefore, evaluating lecturer performance is essential for improving institutional quality. In the Informatics Engineering Study Program at Universitas Muhammadiyah Jember, lecturer performance assessment is still carried out manually through student surveys, administrative reports, and other documents. Such manual processes tend to be subjective, inefficient, and unable to capture the complexity of lecturer performance. To support data-driven decision-making, internal data such as teaching credit load, scientific publications, attendance, and students' average grades can be utilized to provide a more objective performance overview. Student questionnaires further complement this information by capturing students' perceptions of teaching clarity, classroom management, and instructional interaction. Clustering is an effective approach for analyzing these multidimensional data, as it groups objects based on similarity without requiring predefined labels. Among various clustering methods, the X-Means algorithm offers advantages in automatically determining the optimal number of clusters using the Bayesian Information Criterion (BIC), making it more adaptive to diverse and complex datasets. Although X-Means has been applied in several educational studies, its use in clustering lecturer performance based on integrated internal data and student perceptions remains limited. This study therefore applies the X-Means algorithm to cluster lecturer performance in the Informatics Engineering Study Program. The resulting clusters are expected to provide an objective understanding of performance variations and support continuous quality improvement strategies

2. Methode

The research flow diagram illustrates the steps carried out in this study. The research flow diagram is presented in the following figure.

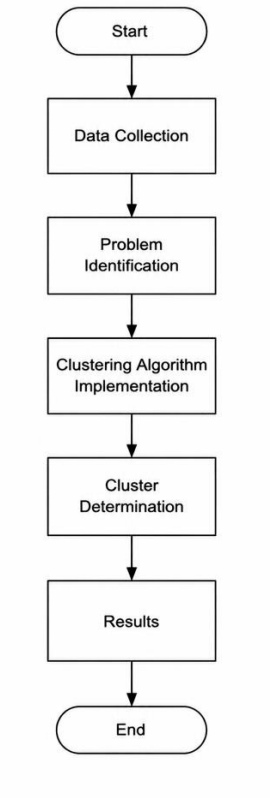

Figure 1. Research Stages

2.1. Data Collection

The data collection stage is essential to ensure the quality and validity of the analysis results. This study uses internal data from the Informatics Engineering Study Program at Universitas Muhammadiyah Jember related to lecturer performance, obtained from official institutional documents and student questionnaires. The collected data include quantitative performance indicators such as students' average grades, lecturer attendance, teaching credit load, number of scientific publications, and student perceptions through questionnaires. Data were gathered from institutional documents, reports, and internal campus databases in Excel format, as well as through questionnaires distributed to Informatics Engineering students.

2.2. Problem Identification

Identifying the objectives and defining the specific problems that will be addressed through the clustering process.

2.3. Processing Data

The data processing stage was carried out after collecting data from various internal sources within the Informatics Engineering Study Program at the University of Muhammadiyah Jember. The quantitative data used relate to lecturers' activities and performance within one academic period, including lecturer attendance, the number of credit hours taught, scientific publications from 2024, students' average final grades, and student perception questionnaires assessed using a 1–5 Likert scale. The data then underwent a cleaning process to remove duplicates, missing values, and illogical outliers, followed by normalization to ensure that all variables were on a comparable scale before being used in the clustering process.

Table 1. Lecturer Performance Evaluation Questionnaire

No	Question
1	Able to create an engaging classroom atmosphere, encourage student participation, and motivate learning
2	Able to deliver the material clearly
3	The lecturer's teaching method encourages students to actively ask questions and participate in discussions (two-way communication)

4	Open and tolerant toward other people's opinions
5	Able to optimize teaching time according to the निर्धारित schedule and regulations
6	Teaches according to the syllabus provided to the students
7	The lecturer gives students opportunities to ask questions and provides clear answers
8	The lecturer reviews and returns corrected quiz/assignment results to students
9	Exam questions are relevant to the course material taught
10	The lecturer gives quizzes or homework assignments as part of learning evaluation
11	Lecturer attendance in class meetings (including replacement classes)
12	The lecturer attends classes on time
13	The lecturer informs students in case of delays, postponements, or rescheduling of classes
14	The lecturer explains the Learning and Teaching Activity Plan (RKBM) or Course Outline (GBPP)
15	The lecturer informs students about the assessment components and their percentages

2.4. X-Means

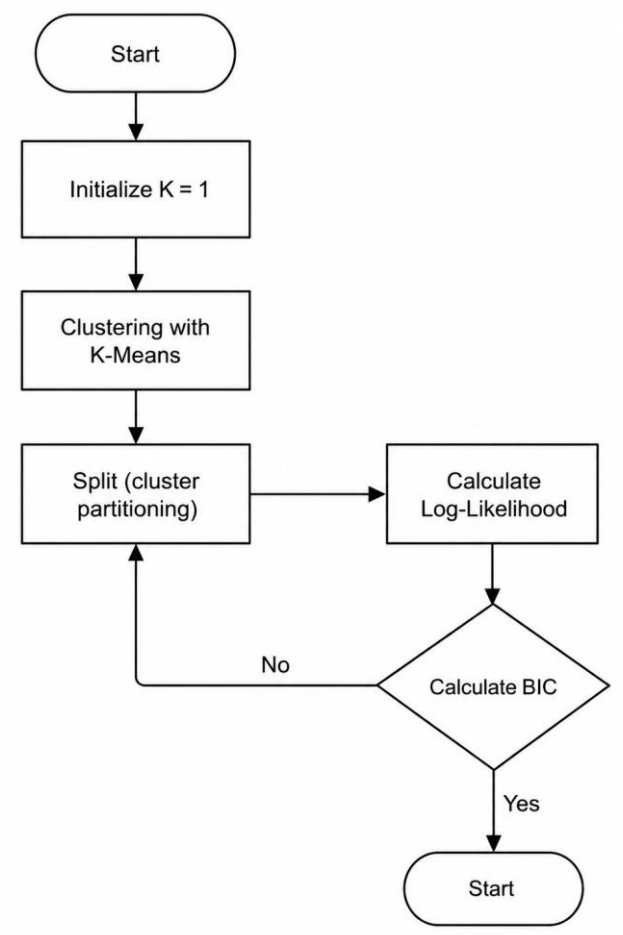

Figure 2. Flowchart of the Clustering Algorithm Process

The X-Means algorithm is an extension of the K-Means algorithm that has the main advantage of automatically determining the number of clusters. While K-Means requires the number of clusters (k) to be defined in advance, X-Means expands and optimizes the value of k based on statistical evaluation.

1. Normalisasi

In this study, which focuses on clustering lecturer performance at Universitas Muhammadiyah Jember, normalization is a crucial step. The data consist of various types of variables, such as students' average grades, the number of credit hours taught, the number of classes handled, and lecturer attendance, each of which has a different value range. By applying normalization beforehand, the clustering process using the X-Means algorithm can be performed more effectively, as each feature contributes proportionally to the cluster formation. This helps produce more accurate and meaningful lecturer groups based on their performance.

The following is the Min–Max normalization formula:

$$X^1 = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (3)$$

Notes:

X = original value
X_{min} = minimum value in the column
X_{max} = maximum value in the column
X^1 = normalized value.

2. Centroid

Centroids are very helpful in determining the center of each cluster based on data such as students' average grades, the number of credit hours taught, the number of classes handled, and lecturer attendance. The X-Means algorithm uses centroids as a reference to minimize the distance between data points and the cluster centers, allowing each lecturer to be grouped more accurately according to their performance characteristics.

The following is the formula for calculating the centroid value:

$$\boldsymbol{\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}} \quad \textbf{(4)}$$

Notes:

$\boldsymbol{\mu_j}$ = Centroid value for the j-th feature.
$\boldsymbol{n}$ = Total number of data points in the cluster
$\boldsymbol{x_{ij}}$ = Value of the i-th data point for the j-th feature
$\sum$ = The sum of all data points within the cluster

3. Euclidean

Euclidean distance is used to measure the similarity between lecturer performance data, which consist of attributes such as students' average grades, the number of credit hours taught, the number of classes handled, and lecturer attendance. This distance value serves as the basis for the X-Means algorithm in forming clusters of lecturers with similar performance characteristics. Therefore, the use of Euclidean distance is highly relevant for producing objective and accurate lecturer groupings based on the available internal data.

The following is the formula for calculating the Euclidean distance

$$d(x_i, \mu) = \sqrt{\sum_{j=1}^{d} (x_{ij} - \mu_j)^2} \quad (5)$$

Notes:

$d(x_i, \mu)$ = Euclidean distance
x_{ij} = Value of the i-th data point for the j-th feature
μ_j = Value of the j-th feature of the centroid.
d = Number of features

4. Sum of Squared Error (SSE)

SSE plays an important role in evaluating the quality of the clustering results. The data used—such as students' average grades, the number of credit hours taught, the number of classes handled, and lecturer attendance—are grouped based on their similarity. SSE helps determine whether lecturers assigned to the same cluster truly share similar performance characteristics. A smaller SSE value indicates that the X-Means algorithm has performed effective clustering and that the resulting clusters are representative of the internal data used.

The following is the formula for calculating the Sum of Squared Error (SSE):

$$\boldsymbol{SSE = \sum_{i=1}^{n}\sum_{j=1}^{d}(x_{ij} - \mu_j)^2} \quad \textbf{(6)}$$

Notes:
$\boldsymbol{x_i}$ = the i-th data point
$\boldsymbol{\mu}$ = centroid (mean) of the cluster
$\boldsymbol{j}$ = feature index
$\boldsymbol{d}$ = number of features
$\boldsymbol{n}$ = number of data points in the cluster

5. Log-likelihood

Log-likelihood, as part of the evaluation process, can assess whether cluster splitting truly improves the model's fit to the underlying data patterns. An increase in the log-likelihood value after cluster division indicates that the split enhances the understanding of lecturer performance structures. Therefore, the use of log-likelihood not only improves clustering accuracy but also strengthens the validity of this study.

The following is the formula for calculating the log-likelihood value:

$$\boldsymbol{L = -\frac{n \cdot d}{2} \cdot \ln(2\pi) - \frac{n \cdot d}{2} \cdot \ln(\sigma^2) - \frac{1}{2\sigma^2} \cdot SSE} \quad \textbf{(7)}$$

Notes:
$\boldsymbol{n}$ = number of data points in the cluster
$\boldsymbol{\pi}$ = konstanta Pi
$\boldsymbol{\sigma^2}$ = variance of the data within the cluster
$\boldsymbol{d}$ = number of features

6. Bayesian Information Criterion (BIC)

The Bayesian Information Criterion (BIC) plays a crucial role in this study. The lecturer performance data used consist of multiple aspects, such as students' average grades, teaching credit load, number of classes handled, and lecturer attendance. Due to the complexity and multidimensional nature of the data, the X-Means algorithm is employed to automatically determine the most optimal number of clusters based on the information contained in the data. With the assistance of BIC, it can be ensured that the resulting clustering is not overly simplistic and better represents the underlying data structure.

The following is the formula for the Bayesian Information Criterion (BIC):

$BIC = L - \frac{p}{2} \cdot \ln(n)$	(8)
$p = K(d+1)$	(9)
Notes: $\boldsymbol{L}$ **= log-likelihood value of the model (measuring how well the model fits the data)** $\boldsymbol{p}$ **= number of parameters in the model** $\boldsymbol{n}$ **= total number of data points** $\boldsymbol{d}$ **= number of features**	

2.5. PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA is not only useful for dimensionality reduction, but also helps simplify complex data to make it easier to analyze. In the context of big data, this technique reduces computational complexity without removing important information, allowing researchers to focus on identifying the main patterns within the data.

1. Covariance

Covariance plays a crucial role in PCA because it forms the basis for constructing a new data structure. Without covariance, it would be difficult to understand how variables are related and move together. The covariance matrix helps identify the directions that contain the most information, from which the principal components are derived. Covariance can be viewed as an initial map that guides PCA in finding the most efficient paths for data simplification. As a result, complex data can be transformed into a simpler, more structured, and still informative form.

The following is the formula for the Covariance:

$Cov(X,Y) = \frac{1}{n-1} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$	(10)
Notes: X_i, Y_i **= the value of the i-th data point** $X, \bar{Y}$ **= the mean of each variable** $\boldsymbol{n}$ **= total number of data points**	

1. Eigenvalue & Eigenvector

Eigenvalues and eigenvectors are two complementary concepts in PCA. Eigenvalues represent the amount of information or data variance captured, while eigenvectors indicate the new directions in which this variance is most clearly observed. Together, they work to simplify the data: eigenvalues determine the priority of the most important principal components, while eigenvectors define how the data are projected so that hidden patterns can be more easily identified.

The following is the formula for the Eigenvalue & Eigenvector:

$Cv = \lambda v$	(10)
Notes: C **= covariance matrix** $\boldsymbol{v}$ **= eigenvector (direction of the principal component)** λ **= eigenvalue (the amount of variance explained by the component)**	

3. Result

The following are the results of lecturer data clustering based on selected attributes analyzed using the X-Means clustering algorithm. Each cluster represents a group of lecturers with similar characteristics or common patterns in the parameters used.

Cluster Group	Cluster Member	Count
C0	1,13,20	3
C1	2,4,6,11.16,18,23,27,28,29,30	11
C2	3,8	2
C3	12,15,24	3
C4	5,7,9,10,14,17,19,21,22,25,26	11

After the clustering process was applied to the lecturer data, each cluster was further analyzed to identify the general characteristics that distinguish one cluster from another.

Table 2. Attribute Contributions to Principal Components (PC1 and PC2)

Attribute Contribution to PC1 & PC2	PC1	PC2
Average Student Score	1.06	8.79
Lecturer Absence	0.14	1.74
Total Credits	0.44	37.51
Number of Publications	0.40	36.04
Question 1	5.65	0.17
Question 2	5.15	0.01
Question 3	7.20	0.06
Question 4	5.31	0.00
Question 5	8.69	0.69
Question 6	7.23	3.75
Question 7	7.47	0.48

Attribute Contribution to PC1 & PC2	PC1	PC2
Question 8	6.62	1.54
Question 9	4.75	0.00
Question 10	6.21	0.01
Question 11	6.46	1.13
Question 12	6.79	3.63
Question 13	8.34	1.45
Question 14	6.50	1.70
Question 15	5.59	1.30

This picture shows the extent to which each variable influences the principal directions in PCA. In PC1, the strongest contributions come from student questionnaire results, particularly questions 5, 13, 7, and 12. This indicates that the PC1 axis primarily represents students' perceptions or evaluations of lecturers. In contrast, PC2 is mainly influenced by numerical variables such as teaching credit load, number of publications, and students' average grades. In other words, PC2 reflects lecturers' workload and academic productivity. Thus, PC1 is more associated with subjective student assessments, while PC2 emphasizes objective performance-related data.

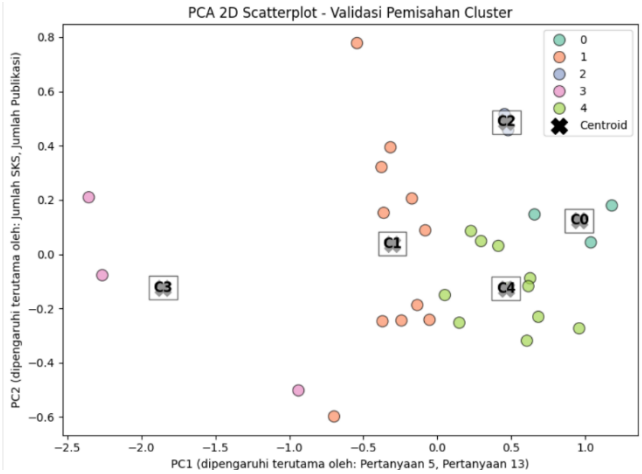

Figure 3. PCA 2D Scatterplot for Cluster Separation Validation

In Figure 3, 2D PCA scatter plot is used to validate the clustering results using two principal components. The first component (PC1) is primarily influenced by student evaluation variables, specifically Question 5 (Teaching Performance) at 8.69% and Question 13 (Commitment to the tridharma) at 8.34%. Meanwhile, the second component (PC2) is more strongly influenced by lecturers' academic performance aspects, namely the number of credit hours taught (37.51%) and the number of publications (36.04%). This indicates that the horizontal axis (PC1) reflects students' perceptions of teaching quality and lecturer commitment, while the vertical axis (PC2) represents lecturers' formal academic achievements.

Values such as 0.2, −0.3, or 0.5 appearing in the PCA results represent loading values (eigenvector coefficients). These values indicate the direction and weight of each variable in forming the principal components. Positive loadings indicate that a variable moves in the same direction as the component, while negative values indicate the opposite direction. Although these values may appear small, they are meaningful because when squared and compared to the total variance, they produce percentage contributions that describe how dominant a variable is in PC1 or PC2.

4. Conclusion

Based on the results of this study entitled "Clustering of Lecturer Performance Using the X-Means Algorithm in the Informatics Engineering Study Program at Universitas Muhammadiyah Jember," several conclusions can be drawn as follows: The implementation of the X-Means algorithm was successful in clustering lecturer performance in the Informatics Engineering Study Program. The implementation process consisted of several stages, including the collection of internal lecturer data and student perception data, data cleaning to eliminate duplicates and missing values, data normalization using the Min–Max method to ensure uniform variable scales, and the application of the X-Means algorithm with Bayesian Information Criterion (BIC) calculation. The results demonstrate that the X-Means algorithm is capable of automatically determining the optimal number of clusters, making it superior to the K-Means method, which requires the number of clusters to be predefined. Based on the data processing results using the X-Means algorithm, the optimal number of clusters obtained was five clusters. Overall, this study indicates that the X-Means algorithm can be effectively applied to cluster lecturer performance. The resulting clusters can serve as a valuable reference for the study program in conducting evaluations and formulating strategies for continuous improvement of lecturer quality.

5. References

[1] Agung, A., Daniswara, A., Kadek, I., & Nuryana, D. (2023). Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru. *Journal of Informatics and Computer Science*, *05*, 97–100.

[2] Agustina, D., & Rahmawati, M. I. (2023). Pengaruh Leverage, Ukuran Perusahaan, dan Profitabilitas Terhadap Ketepatan Waktu Pelaporan Keuangan. *Jurnal Ilmu Dan Riset Akuntansi*, *12*(1), 1–15.

[3] Fathuroh, S. (2022). Metode K-Means Clustering Dalam Optimalisasi Kinerja Dosen Pendamping Akademik Pada Program Kampus Merdeka. *Jurnal Sistim Informasi Dan Teknologi*, *5*, 5–9. https://doi.org/10.37034/jsisfotek.v5i2.172

[4] Li, G., & Qin, Y. (2024). An Exploration of the Application of Principal Component Analysis in Big Data Processing. *Applied Mathematics and Nonlinear Sciences*, *9*(1), 1–24. https://doi.org/10.2478/amns-2024-0664

[5] urfadilla, N., Afdal, M., Permana, I., & Zarnelly, Z. (2023). Comparison of Data Mining Algorithm for Clustering Patient Data Human Infectious Diseases. *Jurnal Teknik Informatika (Jutif)*, *4*(5), 1127–1134. https://doi.org/10.52436/1.jutif.2023.4.5.983

[6] Trianto, R. B., Nugroho, A. S., & Supriyadi, E. (2023). Klasterisasi Menggunakan Algoritma K-Means dan Elbow pada Opini Masyarakat Tentang Kebijakan Sekolah Luring Tahun 2022. *INOVTEK Polbeng - Seri Informatika*, *8*(1), 1. https://doi.org/10.35314/isi.v8i1.2756

[7] Wijayanto, A. (2019). Penggunaan X-Means Clustering Method untuk Mengelompokkan Potensi Sekolah Menengah Unggul di Kabupaten Banyumas. *Journal of Informatics, Information System, Software Engineering and Applications (INISTA)*, *2*(1), 80–88. https://doi.org/10.20895/inista.v2i1.99

[8] Yunita, F. (2018). Penerapan Data Mining Menggunkan Algoritma K-Means Clustring Pada Penerimaan Mahasiswa Baru. *Sistemasi*, *7*(3), 238. https://doi.org/10.32520/stmsi.v7i3.388

[9] Webster, A. J. (2020). *Bayesian information criteria for clustering normally distributed data*. *Icl*, 1–16. http://arxiv.org/abs/2008.03974

[10] Hao, X., Jiang, R., & Chen, T. (2011). Clustering 16S rRNA for OTU prediction: A method of unsupervised Bayesian clustering. *Bioinformatics*, *27*(5), 611–618. https://doi.org/10.1093/bioinformatics/btq725